

Imperial College London
Department of Computing

**Geographic Information Retrieval:
Classification, Disambiguation and Modelling**

Simon E Overell

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Computing of the University of London and
the Diploma of Imperial College, July 2009

Abstract

This thesis aims to augment the *Geographic Information Retrieval* process with information extracted from world knowledge. This aim is approached from three directions: *classifying* world knowledge, *disambiguating* placenames and *modelling* users. Geographic information is becoming ubiquitous across the Internet, with a significant proportion of web documents and web searches containing geographic entities, and the proliferation of Internet enabled mobile devices. Traditional information retrieval treats these geographic entities in the same way as any other textual data. In this thesis I augment the retrieval process with geographic information, and show how methods built upon world knowledge outperform methods based on heuristic rules.

The source of world knowledge used in this thesis is Wikipedia. Wikipedia has become a phenomenon of the Internet age and needs little introduction. As a linked corpus of semi-structured data, it is unsurpassed. Two approaches to mining information from Wikipedia are rigorously explored: initially I classify Wikipedia articles into broad categories; this is followed by much finer classification where Wikipedia articles are disambiguated as specific locations. The thesis concludes with the proposal of the Steinberg hypothesis: By analysing a range of wikipedias in different languages I demonstrate that a localised view of the world is ubiquitous and inherently part of human nature. All people perceive closer places as larger and more important than distant ones.

The core contributions of this thesis are in the areas of extracting information from Wikipedia, supervised placename disambiguation, and providing a quantitative model for how people view the world. The findings clearly have a direct impact for applications such as geographically aware search engines, but in a broader context documents can be automatically annotated with machine readable meta-data and dialogue enhanced with a model of how people view the world. This will reduce ambiguity and confusion in dialogue between people or computers.

Acknowledgements

In the same way that far more is produced during a PhD than a thesis, it is a journey taken by many more people than just the author. I would like to take this opportunity to thank the people that have accompanied me on this journey.

- Firstly I would like to thank my supervisor, Stefan Ruger, for his help, advice and support over the past three years.
- My examiners, Mark Sanderson and Julie McCann, for agreeing to scrutinise my work.
- EPSRC for funding my work and helping me remain a student for three more years.
- My colleagues in the Multimedia and Information Systems group, with whom I shared many a cup of tea and lazy afternoon: Joao, Alexei, Peter, Paul, Ed and Daniel from Imperial College London, and Ainhoa, Haiming, Jianhan, Adam and Rui from the Open University.
- Roelof van Zwol who gave me the opportunity to spend a summer in Barcelona where I failed spectacularly to learn any Spanish.
- I would also like to thank Bokur, Georgina, Llus and my colleagues at Yahoo! Research Barcelona, and my flat mates Claudia and Gleb (who, to my knowledge, also failed to learn Spanish).
- As well as Barcelona, I have had the opportunity to travel both nationally and internationally as part of my PhD. I would like to thank the people who have made this possible: Ali Azimi Bolourian for hosting my visit to Glasgow, Cathal Gurrin for hosting my visit to Dublin, Joao Magalhaes and Jose Iria for hosting my visit to Sheffield, MMKM for partially funding my trips to Seattle and Alicante, and Google for partially funding my return visit to Barcelona.
- A chance meeting with Leif Azzopardi at ECIR led me to joining the BCS-IRSG committee. In turn, this led me to a series of fascinating talks and introduced me to many members of the IR community. I would like to thank Leif, Ali, Andy and the rest of the BCS-IRSG committee.
- For the past three years I have lived in a cozy corner of room 433 of Imperial College London’s Huxley building. A series of people have shared this room with me and joined in often quite heated debates. This includes Uri, Sergio, Georgia, Alok, Mohammed, Mark, Adam, Jonathan, Clemens, Daniel and Nick.
- Before terms such as *Credit Crunch*, *Down Turn* and *Recession* were common place, I was approached by Evgeny Shadchnev to found a company. I learnt a huge amount during the next six months and I would like to thank Evgeny for the much needed distraction from my thesis. Unfortunately by the time we were approaching investors, beta-release and business-plan in hand, the credit bubble had burst and the venture-capital ships had sailed.

- PhD, or in full *Philosophiæ Doctor*, litteral translation is “teacher of philosophy”. This brings me to the next great distraction from my thesis: teaching the third-year Robotics course. I would like to thank Keith and Andy, who gave me the opportunity to be paid to teach undergraduates how to play with Lego for three months a year.
- If a PhD is a journey, the person who gave me a map was George Mallen. When I worked at SSL he showed me what industrial research was and introduced me to Information Retrieval.
- I would like to thank the many friends that have surrounded me for the past three years. They have bought me drinks and listened to tales of placename disambiguation. They include Dave, Dom, Steve, Tim, Lewis, Kurt, Alec, Sacha, Mark, Olly, Dan, Ash and far too many more to list.
- During the final three months of my write up, I developed crippling back pain. I would like to thank Stewart and Melanie for helping me walk tall and sit up straight. Without their help I would not have finished this thesis.
- My parents, John and Yvonne, have always supported me to pursue whatever avenue has interested me. I would like to thank them.
- Despite wanting to live in a farmhouse in the country, my girlfriend Jancy has shared a flat with me, slightly too small to be cozy, in Fulham for the past three years. I thank her from the bottom of my heart for her endless love and support.

Contents

Abstract	3
Acknowledgements	5
Contents	7
1 Introduction	15
1.1 GIR, IR and GIS	15
1.2 Placenames and locations	16
1.3 The need for geographic indexing	16
1.4 The need for placename disambiguation	17
1.5 GIR systems	18
1.6 Wikipedia	19
1.7 Scope	20
1.8 Contributions	21
1.9 Publications	22
1.10 Implementations	23
1.11 Organisation	24
1.12 Roadmap	25
2 GIR and Wikipedia	27
2.1 Introduction	27
2.2 From IR to GIR	27
2.2.1 Relevance	27
2.2.2 Phrases and named entities	28
2.3 GIR	29
2.3.1 GIR systems	30
2.3.2 Placename disambiguation	30
2.3.3 Geographic indexing	33
2.3.4 Geographic relevance	35
2.3.5 Combining text and geographic relevance	37
2.4 Geographic resources	39
2.5 Wikipedia	40
2.5.1 Wikipedia articles	41
2.5.2 Wikipedians	41
2.5.3 Accuracy	42

2.5.4	Mining semantic information	43
2.6	Discussion	44
3	Evaluation and metrics	45
3.1	Introduction	45
3.2	Evaluating IR systems	45
3.2.1	Evaluation forums	45
3.3	Evaluation measures	47
3.3.1	Binary classification and unordered retrieval	47
3.3.2	Scored classification and ranked retrieval	48
3.4	Statistical testing	49
3.4.1	The Sign test	50
3.4.2	The Wilcoxon Signed-Rank test	50
3.4.3	The Friedman test	51
3.4.4	The Student's t-test	51
3.4.5	One-tailed vs. two-tailed	52
3.5	Discussion	52
4	Classifying Wikipedia articles	55
4.1	Introduction	55
4.2	Classification and disambiguation of Wikipedia articles	55
4.3	Classifying Wikipedia articles	57
4.3.1	Ground truth	58
4.3.2	Sparsity of data	58
4.3.3	Removing noise	60
4.3.4	System optimisation	60
4.4	Evaluation and comparison to existing methods	62
4.4.1	Experimental setup	63
4.4.2	Results	64
4.5	A case study: Flickr	65
4.5.1	A tag classification system	67
4.5.2	Coverage	69
4.5.3	Summary	70
4.6	Discussion	71
4.6.1	Multilingual classification	71
4.6.2	Context	72
4.6.3	Which Parties party and which Races race?	72
5	Disambiguating locations in Wikipedia	73
5.1	Introduction	73
5.2	Mapping Wikipedia articles to locations	74
5.3	Disambiguating placenames in Wikipedia	75
5.3.1	Classes of evidence considered	75
5.3.2	The three stages	77
5.4	Ground truth	78
5.5	Which class of evidence offers the most information?	78

5.5.1	Building a pipeline	81
5.5.2	Validating the pipeline	83
5.5.3	Enhancing and degrading the ground truth	85
5.6	Testing	86
5.6.1	Naïve baseline methods	86
5.6.2	Results	86
5.6.3	Analysis	87
5.7	Model size and complexity	87
5.7.1	Distribution	87
5.8	Building a geographic co-occurrence model	88
5.8.1	Clarity	89
5.8.2	The model	89
5.9	Discussion	90
6	Placename disambiguation in free text	91
6.1	Introduction	91
6.2	Supervised placename disambiguation	92
6.3	Theoretical performance	94
6.3.1	Computing the bounds	94
6.3.2	Computing the relative performance	95
6.4	Approaches to placename disambiguation	97
6.4.1	Naïve methods	97
6.4.2	Neighbourhoods	98
6.4.3	Support Vector Machines	100
6.5	Direct measurement	101
6.5.1	Building a ground truth	101
6.5.2	Results	102
6.5.3	Analysis	103
6.6	Indirect measurement	104
6.6.1	Forstar	104
6.6.2	Distribution of placenames in the GeoCLEF collection	105
6.6.3	Baseline methods	107
6.6.4	Queries	108
6.6.5	Results	109
6.6.6	Analysis	112
6.7	Discussion	113
6.7.1	Is there more information in a placename or a location?	113
6.7.2	Combining methods – the search for synergy	114
7	The world according to Wikipedia	115
7.1	Introduction	115
7.2	Everything is related to everything else	116
7.3	Alternate language versions of Wikipedia	117
7.4	Disambiguating locations in alternative languages	117
7.5	Bias	124

7.5.1	Quantitative analysis	124
7.5.2	Qualitative analysis	124
7.6	The Steinberg hypothesis	128
7.6.1	Experiment	130
7.6.2	Results	131
7.6.3	Applications of the Steinberg hypothesis	132
7.7	Temporal references in Wikipedia	133
7.8	Discussion	135
7.8.1	Can the different wikipedias be considered independent?	135
7.8.2	Detecting events in Wikipedia	138
7.8.3	Numenore	138
8	Conclusions	141
8.1	Achievements	141
8.1.1	Placename disambiguation	141
8.1.2	The Steinberg hypothesis	142
8.2	Limitations	142
8.2.1	Placename disambiguation	142
8.2.2	Wikipedia as a corpus	143
8.3	Future work	143
8.3.1	Context based placename disambiguation	144
8.3.2	Geographic relevance ranking	144
8.4	Core contributions	144
A	Appendix: History of Wikipedia	145
A.1	Introduction	145
A.2	Anatomy of an article	146
A.3	Clustering Wikipedia articles	148
A.4	Wikipedia in research	148
B	Appendix: Further experiments with the GeoCLEF corpus	151
B.1	Introduction	151
B.2	Query classification	151
B.3	Data fusion	153
B.4	Per-query results	156
C	Appendix: Constructed languages	159
C.1	Introduction	159
C.2	Esperanto	159
	Nomenclature	163
	Glossary	165
	Bibliography	171

List of Tables

1.1	Wikipedia statistics summary	20
3.1	Contingency table	47
4.1	Weighting functions example	61
4.2	Varying feature values	61
4.3	System evaluation results	64
4.4	Ambiguity in mapping from tags to categories	68
4.5	Coverage of Flickr tags	69
4.6	Coverage of the Flickr vocabulary	69
4.7	Examples of tags covered by ClassTag but not covered by WordNet	70
5.1	Contingency table modified for classifying Wikipedia articles as placenames	79
5.2	Information offered by each class of evidence	80
5.3	The pipeline is repeatedly extended by additional evidence	81
5.4	Final pipeline	83
5.5	Greedy pipeline	84
5.6	Comparison of naïve methods and the pipeline with respect to the test set	87
5.7	Sample of the Location Links Table	89
5.8	Samples of the Article Location Table and Placename Frequency Table	90
6.1	D_{KL} between location and placename distributions	95
6.2	D_{JS} between location distributions	97
6.3	Top related placenames in the location neighbourhoods	99
6.4	Weighting functions example	101
6.5	Summary of ground truth collections	102
6.6	Accuracy per collection	103
6.7	Performance across total ground truth	103
6.8	GeoCLEF queries and manually constructed query parts	108
6.9	Percentage overlap of identical classifications	110
6.10	Top locations different from the MR method	110
6.11	Query formulations against MAP(%)	110
6.12	Summary of per-query results	111
6.13	GeoCLEF quartile range results	112
7.1	Language versions of Wikipedia being considered	117
7.2	Summary of co-occurrence models	118

7.3	Co-efficients of the Zipfian distributions and spatial autocorrelation	119
7.4	References per 1m people	125
7.5	D_{JS} between different wikipedias	125
7.6	Top five locations from each language	125
7.7	The symmetric differences between the observed and expected results between different formulations of the Steinberg equation.	131
7.8	Optimal values of α and β for different formulations of the Steinberg equation	132
7.9	Proportion of links extracted from different wikipedias to articles describing years	134
7.10	Co-efficients of the Zipfian distributions for pre-modern and post-modern curves	135
B.1	Tentative classification of geographic topics	151
B.2	Explicit topic difficulties	152
B.3	Topic categorization	152
B.4	Classification by feature type	152
B.5	Classification by location	153
B.6	Penalisation values	154
B.7	GeoCLEF 2008 results	155
B.8	2005 Per-query results and summary	156
B.9	2006 Per-query results and summary	157
B.10	2007 Per-query results and summary	157
B.11	2008 Per-query results and summary	158
C.1	Summary of the Esperanto co-occurrence model	160
C.2	D_{JS} between the Esperanto Wikipedia and other wikipedias	160

List of Figures

1.1	Overlapping retrieval methods for Data, Information and Geographic Meta-data	16
1.2	Generic GIR system architecture	18
2.1	An illustration of how orientation can effect the relationship between MBRs	36
2.2	The counties' and countries' of Great Britain horizontal topology and vertical topology . .	38
4.1	Example category and template network	59
4.2	Threshold – F ₁ -Measure	62
4.3	Threshold – Proportion of articles classified	63
4.4	Per-category precision	64
4.5	Example photo with user-defined tags	66
4.6	Classification of the tags in Figure 4.5	66
4.7	Overview of the ClassTag system	67
4.8	Tag → Category example	68
4.9	Tag → Category example (reduced ambiguity)	68
4.10	Classification of Flickr tags	71
5.1	Grounding and placename recall achieved by different classes of evidence	80
5.2	Pipeline of classes of evidence	81
5.3	Increasing the length of the pipeline against performance measures	82
5.4	Modified pipeline of classes of evidence	83
5.5	Greedy pipeline of classes of evidence	85
5.6	Enhancing and degrading the ground truth	85
5.7	Frequency of location references – rank	88
5.8	2D projection of location references in Wikipedia	88
6.1	Placename disambiguation with a minimum bounding box example	92
6.2	Distribution of how locations co-occur with different Cambridges	96
6.3	Forostar design	105
6.4	Frequency of placename references – rank	106
6.5	2D projection of location references in the GeoCLEF corpus	107
6.6	Overlapping groups of retrieval methods	112
7.1	Alternate language pipeline	118
7.2	Heat maps in different wikipedias	120
7.3	Heat maps in different wikipedias	121
7.4	Distribution of locations in the different wikipedias	122

7.5	Distribution of locations in the different wikipedias	123
7.6	Cartograms of references in different wikipedias	126
7.7	Cartograms of references in different wikipedias	127
7.8	Cover of The New Yorker, March 29, 1976	129
7.9	Histogram sample	132
7.10	Map of which location you are most likely to mean when referring to “London” dependent on your current location	133
7.11	Map of which location you are most likely to mean when referring to “Cambridge” dependent on your current location	133
7.12	Distribution of temporal references in the different wikipedias	136
7.13	Distribution of temporal references in the different wikipedias	137
7.14	Screen shot taken from http://www.numenore.co.uk	139
A.1	Anatomy of a Wikipedia article	149
B.1	Rank example	154
C.1	Distribution of locations in the Esperanto Wikipedia	160
C.2	Maps of the references to locations in the Esperanto Wikipedia	161

Chapter 1

Introduction

Newspapers, television, books and the Internet hold a huge amount of geographic and temporal information (Jones et al. 2008). A minute proportion of this data is accompanied with machine readable meta-data. It is common to want to browse by time and placename, when searching for information about a specific event or location (Sanderson and Kohler 2004; Mishne and de Rijke 2006; Jones et al. 2008; Gan et al. 2008); because of this, there is a need for automatic annotation of resources with time and location data.

The understanding of time and place references in documents generally involves knowledge of the document context and a shared world knowledge between the author and reader. If a document refers to “next month,” or “The Capital,” the time and location the document was authored are necessary contextual clues that help understanding. If a news report refers to “America invading Iraq,” fully understanding the statement relies on a shared knowledge that “America” is a synonym for the nation, the United States of America, situated in the continent of North America; and that Iraq is a country situated in the Middle East.

For the automatic annotation of time and location, both shared world knowledge and document context needs to be captured. It is this problem that I hope to address in this thesis. Put succinctly **to extract geographic world knowledge from Wikipedia and apply it to geographic information retrieval**. Wikipedia is chosen as a source of world knowledge, as it is the largest encyclopædic corpus freely available. I approach this task by exploring methods of disambiguating Wikipedia articles to build a huge corpus of world knowledge, and then evaluate different methods of applying this world knowledge to disambiguating placenames in free text.

1.1 GIR, IR and GIS

Information Retrieval (IR) differs considerably from database retrieval (DB), which is concerned with the retrieval of data as opposed to information (van Rijsbergen 1979). In IR an information need is specified in natural language, and a corpus of unstructured documents is searched for results that will best satisfy the information need; in contrast, in database retrieval queries are formed using a query language where the data required is described, all results matching the query are returned without ranking. The field of XML retrieval is concerned with the retrieval of structured documents; this is the overlap between DB retrieval and IR. Structured documents are made up of clearly defined parts, for example title, abstract and body, and may have associated meta-data, for example cost or author (Lalmas 2000). Geographic Information Systems (GIS) is a field concerned with the efficient storage,

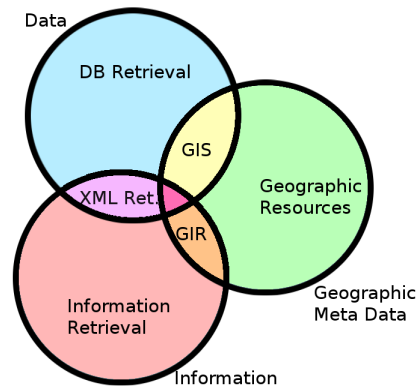


Figure 1.1: Overlapping retrieval methods for Data, Information and Geographic Meta-data

retrieval and display of geographic data; it is the augmentation of DB retrieval with geographic data. Geographic Information Retrieval (GIR) is the augmentation of IR with geographic data: a user will express their geographic information need within their query (Jones and Purves 2006). This relationship is summarised in Figure 1.1.

This thesis is concerned with the area of Geographic Information Retrieval (GIR), more specifically, how world knowledge can be extracted and applied to GIR.

1.2 Placenames and locations

A large part of this thesis is concerned with the mapping from placenames (words referring to places), and locations (places themselves). To add clarity to these discussions, I will define some terminology here. A *location* is a space on the Earth’s surface represented by polygons or points. A *placename* is a phrase used to refer to a location. Note this is a many to many relationship. Where the two may be confused, placenames will be referred to in inverted commas, e.g. “Cambridge” can refer to Cambridge, UK or Cambridge, MA.

1.3 The need for geographic indexing

Historically disambiguation has been performed at indexing time. Professional indexers compiled indexes in a two stage process of conceptual analysis and translation into indexing terms (Lancaster 2003). In traditional IR, conceptual analysis is left out and indexing terms are automatically extracted straight from the document; because of this, the indexing terms are inherently ambiguous, and disambiguation is achieved through manual query expansion. For example, a user dissatisfied with results for “Lincoln” can incrementally expand their query into “Abraham Lincoln, President”.

GIR is a more complex case: information required for disambiguation is often implicit as there are geometric and topological relationships between locations. If the user is interested in all documents relating to areas within “London”; they can reduce the ambiguity in their query by expanding it to search for areas within “Greater London, UK”. This would have to be greatly expanded further to cover all the 33 boroughs and the areas within. Many of these names would also be ambiguous such as “Chelsea”, “Vauxhall” and “Greenwich”. A user presenting this further expanded query will be overwhelmed with false positive results.

One solution to this problem is to assign every geographic phrase or placename in a document a reference to a location. An index can then be built of these semantic representations allowing places to be searched unambiguously.

1.4 The need for placename disambiguation

How significant is the problem of disambiguating placenames? To answer this question I provide an estimate of the accuracy that can be achieved with a trivial method. Assuming we classify every placename as the most referred to location, then the fraction of correctly disambiguated placenames r_{corr} can be estimated as follows (notation detailed in the Nomenclature):

$$r_{\text{corr}} = \frac{\sum_{p \in M} \text{ref}(p, L_1(p))}{|N|} \quad (1.1)$$

Using a model based on a crawl of Wikipedia (detailed in Chapter 5) one can estimate the proportion of placenames that will be correctly matched to locations to be 89.6% (detailed further in Chapter 6). One could easily argue “this is accurate enough!” However, this error is noticeable in the following three circumstances:

- When a location is being searched for and a more commonly referred to location shares its name, the user will be flooded with irrelevant results. To quantify this, for every reference to a location l , where a more commonly referred to location exists for placename p ; the average ratio, r_{ave} , of the frequency of references to l by p divided by the frequency of all references to locations by p can be calculated:

$$r_{\text{ave}} = \frac{1}{|K|} \sum_{p \in K} \sum_{i=2}^{|\mathbf{L}(p)|} \left(\frac{\text{ref}(p, L_i(p))}{\sum_{l \in \mathbf{L}(p)} \text{ref}(p, l)} \right) \quad (1.2)$$

Assuming all locations are equally likely to be searched for, on average only 10% of documents will be relevant when searching for a less common location. Of course not every location is equally likely to be searched for, in the next version of the equation we weight the likelihood a location is to be searched for with respect to how often it is referenced:

$$r_{\text{aveW}} = \left(\sum_{p \in K} \sum_{i=2}^{|\mathbf{L}(p)|} \text{ref}(p, L_i(p)) \right)^{-1} \sum_{p \in K} \sum_{i=2}^{|\mathbf{L}(p)|} \left(\frac{(\text{ref}(p, L_i(p)))^2}{\sum_{l \in \mathbf{L}(p)} \text{ref}(p, l)} \right) \quad (1.3)$$

Assuming all locations are as likely to be searched for as they are referenced, on average only 17% of documents will be relevant when searching for a less common location. Note the r_{aveW} equation is a generalisation of the r_{corr} equation, and if i were initialised to 1 in the two summations they would be equivalent.

When one wishes to display documents relevant to London, Ontario, on a map, the user-interface will quickly become cluttered.

- The second problem is that this error is cumulative when using a GIS that models the relationship between locations. Supposing the user is searching for locations in the continent of North America, the query is expanded to countries, states, counties and towns; at each expansion the error is compounded.

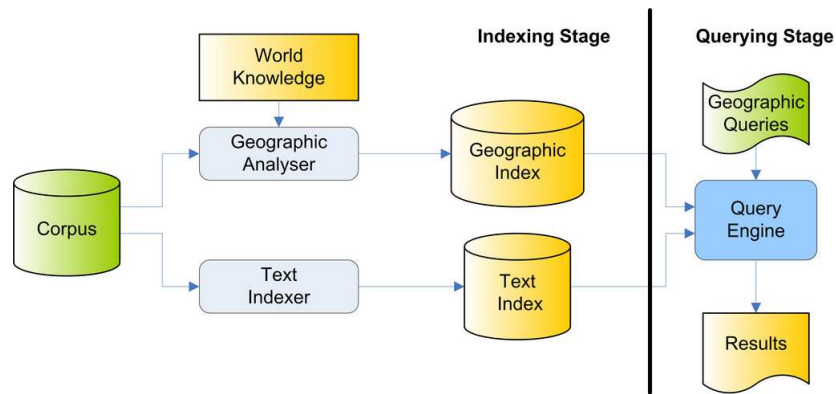


Figure 1.2: Generic GIR system architecture

- The final problem is that context affects the perceived most important locations (Worboys 1996). For example, any local newspaper in Ontario will by default assume that “London” refers to London, Ontario, not London, UK. This is particularly noticeable when two places of roughly equal *importance* exist: for example, Cambridge, UK and Cambridge, MA. Some contextual data, whether specified implicitly with statistical models or explicitly, is needed.

Summarising, with the growing volume of content on the web and presence of Web 2.0 technologies, the amount of annotated and un-annotated geographic information is growing fast; this combined with the growing volume of traditional geographic information sources, and increasingly novel browsing methods shows the growing need for the geographical indexing and browsing of information.

1.5 GIR systems

Microsoft and Google both introduced commercial GIR Systems in 2005, Microsoft Live Local¹ and Google Maps². In the same year the geographic track was added to the CLEF series of evaluation conferences allowing direct comparisons of GIR systems (Gey et al. 2006). Figure 1.2 illustrates the standard components of a GIR system:

- The *Geographic Analyser* processes the corpus identifying references to locations, disambiguating them and building a geographic index. Disambiguation is commonly achieved through the application of geographic world knowledge. The world knowledge used in a GIR system is often a simple geographic gazetteer, a mapping of placenames to geographic co-ordinates.
- The *Text Indexer* is similar to a standard IR system and builds a standard text index.
- The *Query Engine* ranks the documents in the corpus with respect to both their geographic and textual relevance to the query. Generally a score is given for the geographic relevance and a score given for the textual relevance and the two combined.

The components of GIR systems are examined more carefully in Chapter 2. This thesis’s focus is the generation of world knowledge used by the geographic analyser and the geographic analyser itself.

¹<http://www.local.live.com>

²<http://www.maps.google.com>

1.6 Wikipedia

Wikipedia has become one of the big Internet phenomena of the last 5 years (Tapscott and Williams 2008). Launched in 2001 it is now the largest and dominant reference work currently available on the Internet (Wikipedia 2008b; Giles 2005). Wikipedia currently has editions in over 250 languages and by the time this thesis is published, there will be over 10 million articles. Wikipedia has had such success due to the “Wiki” software it is built on (Tapscott and Williams 2008). A wiki is a web-application that allows users to create, edit and link web pages. Wikipedia runs on MediaWiki, the wiki software developed by the Wikimedia Foundation. MediaWiki records a history of every single edit ever made to Wikipedia, all of which is downloadable and viewable. This means information can never be removed from Wikipedia, only added to.

Wikipedia is configured in such a way that anyone can edit almost any article (there are some restrictions on sensitive or vandalism prone articles). It is this collaborative spirit that has led to Wikipedia’s success (Tapscott and Williams 2008). Although it is not required, Wikipedia encourage users to login before editing. This has further allowed them to encourage a community spirit, the phrase “The Wikipedia Community” has been coined to refer to the nearly 6m registered Wikipedia editors (referred to as “Wikipedians”). Wikipedians come from every corner of the globe, from different cultures and religions, and speaking different languages. In an attempt to manage competing opinions Wikipedia have suggested style guidelines and a “neutral point-of-view” policy. Contributors are asked to submit verifiable encyclopædic facts only, no original research or personal opinions.

Wikipedia is now owned and run by the “Wikimedia Foundation,” a charity organisation supported by donations. Described by its creator Jimmy Wales as “an effort to create and distribute a free encyclopedia of the highest possible quality to every single person on the planet in their own language,” (Wales 2005) Wikipedia is beginning to take on a life of its own.

Wikipedia has become a recurring subject in the media receiving both praise and criticism. The journal “Nature” published an article in 2005 claiming it to be as accurate as the Encyclopædia Britannica (Giles 2005). A series of further articles have both backed up and refuted this claim (Encyclopædia Britannica Inc 2006; Nature 2006; Waters 2007; Young 2006). Conservapedia³ was set up in 2006 as a backlash against Wikipedia’s “liberal, anti-Christian, and anti-American bias” — this reported bias is significantly exaggerated. Further attacks on the bias of Wikipedia have been fuelled by evidence that political candidates have edited both their own and opponents pages (Griffith 2007). The problem of vandalism is a noticeable problem downplayed by the Wikimedia Foundation and satirised by *Every topic in the Universe Except Chickens.com* and the Colbert Report, who respectively plead for potential vandals only to edit the Chicken page (as everyone already knows everything they need to know about Chickens), and argue “any user can change any entry, and if enough users agree with them, it becomes true”.

At the forefront of Web 2.0 community driven software, Wikipedia has been of great interest to the academic community. The fact that it is freely available to download and no specialist software or hardware is needed, makes its study very accessible. It is the community aspect of Wikipedia that influenced our choice when searching for a source of “World Knowledge”. The second advantage of Wikipedia is its hyper-linked structure (Mihalcea 2007). Authors are encouraged, in at least the first reference, to hyper-link references to major concepts and themes. These hyper-links help to disambiguate the concept or theme.

³<http://www.conservapedia.com/>

Articles in English Wikipedia	2.1m
Words in English Wikipedia	916m
Internal links in English Wikipedia	100m
Total edits across English Wikipedia	184m
Different language versions of Wikipedia	252
Language versions with over 100k articles	15
Articles across all language versions	9.1m
Words across all language versions	1,410m
Average page requests per second	30k
Proportion of traffic directed at the English Wikipedia	55%
Registered editors	5.9m
Servers needed to serve Wikipedia	100

Table 1.1: Wikipedia statistics summary (correct as of 1 February 2008)

1.7 Scope

This section will outline the objectives, requirements and research questions addressed by this thesis.

Mining world knowledge from Wikipedia

To effectively mine Wikipedia, efficient scalable algorithms are required for extracting how people refer to locations, times and other concepts. Efficiency is essential as the information in Wikipedia is dynamic, therefore the mined world knowledge needs to be easily updatable. Scalability is a priority as Wikipedia is growing at an enormous rate, currently containing millions of articles and predicted soon to contain tens of millions of articles across multiple languages.

World knowledge is a particularly broad term that I will make more precise for the purposes of this thesis. By world knowledge I refer to the shared understanding between people, required for interpreting documents. Specifically the default meaning for words and phrases, and how these are changed by context. I further limit this definition by only considering the understanding of noun phrases; and defining the “meaning” of a word to be either its location on the surface of the Earth, or a time-line if it is a place or a time, or its broad noun class.

The research questions this thesis aims to answer are:

- What meta data is most useful when classifying or disambiguating a Wikipedia article?
- What degree of accuracy can be achieved without a full semantic analysis of every Wikipedia article?
- What proportion of Wikipedia articles can be disambiguated; and what is the distribution of references to locations in Wikipedia?

Applying world knowledge mined from Wikipedia to placename disambiguation

Once a model of how placenames occur in Wikipedia has been built, an evaluation of different methods for applying this model to placename disambiguation will be performed. Heuristic methods for placename disambiguation have been extensively covered (Li et al. 2003; Clough et al. 2004; Rauch et al. 2003; Cardoso et al. 2005; Zong et al. 2005; Leidner et al. 2003). This thesis will only look at supervised learning methods for placename disambiguation, using the world knowledge extracted from Wikipedia to build a model. Three approaches will be used to evaluate the system’s performance: theoretic –

by examining the model; direct – by comparison to a ground truth; and indirect – by comparing how disambiguation affects the retrieval process.

The research questions considered are:

- What are the theoretical bounds of accuracy achievable for placename disambiguation using a co-occurrence model?
- What level of co-occurrence is most useful for placename disambiguation?
- How does the distribution of locations in Wikipedia differ from a standard evaluation corpus?
- How much can placename disambiguation improve the information retrieval process?
- When no in-document context is available, is the location of the author a suitable context for placename disambiguation?

Comparing world knowledge extracted from different language versions of Wikipedia

The methods developed and evaluated for the task of disambiguating locations in the English language Wikipedia will then be applied to alternate language versions of Wikipedia. The motivation for this is to see how quantitatively and qualitatively the distribution of location references vary between different languages.

The research questions this thesis aims to answer are:

- Is Wikipedia truly impartial or do predictable biases occur based on the language of the articles?
- Do all people have the same world view relative to their location? and, can this world view be captured in a simple model?

Scope summary

In the above sections we have summarised all the research sub-questions asked in pursuit of our main goal, which will be reiterated here:

**“to extract geographic world knowledge from Wikipedia
and apply it to geographic information retrieval.”**

1.8 Contributions

The research performed for the purposes of the PhD culminating in this thesis produced a series of contributions to the scientific community. This thesis has:

1. Identified which meta-data in Wikipedia are most useful when disambiguating articles as locations or categorising articles as WordNet broad noun syntactic categories.
2. Presented an efficient scalable supervised method for classifying Wikipedia articles and similar hierarchically structured resources, as WordNet broad noun syntactic categories. This method has been shown to be better than the state of the art.
3. Quantified the ambiguity at each stage when mapping from a tag to a category using Wikipedia, and how to significantly reduce the ambiguity.

4. Produced a publicly available ground truth of 300 randomly selected Wikipedia articles classified by hand as WordNet noun syntactic categories.
5. Presented an efficient scalable heuristic method for disambiguating Wikipedia articles as locations. This method has been shown to be better than the state of the art.
6. Quantified the information contributed by each of the features considered useful for placename disambiguation.
7. Produced a publicly available ground truth of all the links crawled from 1000 randomly selected Wikipedia articles disambiguated as to whether they refer to locations or other concepts and if they do refer to locations matched to relevant entries in the TGN. Note this has already been adopted and used by other research groups for evaluation (Buscaldi and Rosso 2007).
8. Quantified the problem of placename disambiguation based on synonyms extracted from Wikipedia.
9. Produced publicly available geographic co-occurrence models in a variety of languages.
10. Demonstrated supervised placename disambiguation can statistically significantly outperform rule-based placename disambiguation both in direct and indirect evaluation.
11. Provided a macro-level comparison of the distribution of placenames in Wikipedia and the GeocLEF corpus.
12. Demonstrated in a disambiguated geographic query that information useful for the retrieval process is captured in both the location and placename part of the query.
13. Calculated the bias in different language versions of Wikipedia and generated maps to visually represent this.
14. Developed a model to quantify a given person's fish-eye view of the world and validated this model against Wikipedia.
15. Quantified the distribution of references to times and locations in Wikipedia and showed they can both be modelled with a series of Zipfian distributions.

1.9 Publications

This section lists the publications that disseminate the research results obtained with work presented in this thesis. Publications are grouped by chapter.

Chapter 4

The work detailing the classification of Wikipedia articles, comparison to DBpedia, and the use of Flickr as a case study:

- Simon Overell, Bökur Sigurbjörnsson and Roelof van Zwol. *Classifying Tags using Open Content Resources* in WSDM, Barcelona, Spain (2009).
- Simon Overell, Bökur Sigurbjörnsson and Roelof van Zwol. *Classifying Content Resources Using Structured Patterns* Patent Pending (Submitted Nov 2007).

- Simon Overell, Bökur Sigurbjörnsson and Roelof van Zwol. *System and Method for Classifying Tags of Content Using a Hyperlinked Corpus of Classified web pages* Patent Pending (Submitted Dec 2007).

Chapters 5, 6 and Appendix B

The initial work showing that by using only heuristic rules it is possible to achieve a relatively high precision disambiguating Wikipedia articles. This paper also details the creation of a ground truth comprising of disambiguated Wikipedia articles:

- Simon Overell and Stefan Rürger. *Identifying and grounding descriptions of places* in the SIGIR Workshop on Geographic Information Retrieval, Seattle, USA (2006).

Our second paper on disambiguating Wikipedia articles, this time using only meta data and demonstrating that no content is needed from the actual article. We also quantified the bounds and characteristics of the model:

- Simon Overell and Stefan Rürger. *Geographic Co-occurrence as a Tool for GIR* in the CIKM Workshop on Geographic Information Retrieval, Lisbon, Portugal (2007).

Our first article on placename disambiguation. A series of supervised placename disambiguation techniques are compared to naïve methods and traditional IR. The geographic co-occurrence model extracted from Wikipedia forms the training data for the methods:

- Simon Overell and Stefan Rürger. *Using co-occurrence models for placename disambiguation* in the International Journal of Geographic Information Science, Taylor and Francis (2008).

The initial implementation of the GIR system, Forostar, is described in detail and compared to the current state of the art. This paper contains experiments showing the optimal query construction is to use both the placename and location in the query:

- Simon Overell, João Magalhães and Stefan Rürger. *Forostar: A System for GIR* in Evaluation of Multilingual and Multi-modal Information Retrieval, Springer-Verlag LNCS (2007).

Additional work on the implementation of Forostar including experiments showing the optimal data fusion method for textual and geographic data:

- Simon Overell, Adam Rae and Stefan Rürger. *MMIS at GeoCLEF 2008: Experiments in GIR* in CLEF Working notes (2008).
- Simon Overell, Adam Rae and Stefan Rürger. *Geographic and textual data fusion in Forostar* to appear in CLEF LNCS proceedings (2009).

1.10 Implementations

A number of applications and tools have been developed specifically for the experiments or to display results presented in this thesis. In this section I provide a brief overview of these applications grouped by chapter.

Chapter 4 – ClassTag. ClassTag is a tool for classifying Flickr tags using data mined from Wikipedia. It classifies Wikipedia articles as WordNet broad noun categories; anchor texts linking to these articles then form a huge lexicon of classified terms with frequency data, which can be mapped to Flickr tags.

Chapter 5 – PediaCrawler. PediaCrawler is a module of the Forostar GIR system. It disambiguates Wikipedia articles mapping them to their corresponding location. Links to this set of disambiguated Wikipedia articles form a geographic co-occurrence model.

Chapter 6 – Forostar. Forostar is a full GIR system which includes placename disambiguation and data-fusion modules. Experiments with Forostar are performed on the GeoCLEF corpus.

Chapter 7 – Numenore. Numenore is a web application⁴ that displays locations and events mined from different language versions of Wikipedia. The data Numenore is built upon is mined by PediaCrawler and is also available in a machine readable format through the Numenore API.

1.11 Organisation

The purpose of this chapter has been to provide motivation for my research, introduce the field of GIR and identify the scope of this thesis. The organisation of the following chapters will be outlined below:

Chapter 2 – GIR and Wikipedia

A comprehensive literature review of the areas of geographic information retrieval and mining knowledge from Wikipedia. This chapter will concentrate on the tasks closely related to this thesis, specifically placename disambiguation, and the classification and disambiguation of Wikipedia articles.

Chapter 3 – Evaluation and metrics

This chapter contains a brief history and the motivation for the current ad-hoc evaluation framework that has become standard across the IR field. The current metrics used for evaluating IR and GIR systems will be presented with the common statistical techniques.

Chapter 4 – Classifying Wikipedia articles

I begin the body-of-work of this thesis by examining how Wikipedia articles can be classified as the 25 WordNet broad noun categories. A supervised learning approach is adopted with the overlap between WordNet and Wikipedia used as training data. The chapter explores what proportion of Wikipedia can be classified using only associated meta-data and which meta-data is most useful. It concludes by illustrating how the classifiable terms in a sample corpus can be greatly extended over WordNet alone.

Chapter 5 – Disambiguating locations in Wikipedia

Chapter 4 simply classifies whether an article describes a location or geographic object without specifying the specific point on the Earth's surface being referred to. In contrast this chapter uses a heuristic method to match locations in Wikipedia to an authoritative source. I investigate which classes of evidence

⁴<http://www.numenore.co.uk>

provide the greatest information and show, for locations, our heuristic method out-performs the generic supervised method.

Using the disambiguated set of Wikipedia articles a geographic co-occurrence model is generated. The distribution of locations within the model and other characteristics are presented.

Chapter 6 – Placename disambiguation in free text

Using the co-occurrence model developed in Chapter 5, this chapter performs an evaluation of different supervised methods applying these models to placename disambiguation. The supervised methods are compared to a naïve baseline, the theoretical bounds achievable, and the current state of the art. Evaluation is performed both directly on a groundtruth and indirectly in a standard IR ad-hoc evaluation framework.

Chapter 7 – The world according to Wikipedia

In the final body-of-work chapter I apply the methods of disambiguating locations in Wikipedia developed in Chapter 5 to alternate language versions of Wikipedia. The distribution of location references and temporal references between different language wikipedias are compared both quantitatively and qualitatively. Models of people’s world view are fitted to the different wikipedias to demonstrate that all people have much the same fish-eye view of the world.

Chapter 8 – Conclusions

This thesis ends with a summary of the achievements and limitations of the work presented, and by identifying a few of the more interesting research questions that there was no time to answer.

1.12 Roadmap

The scope of this thesis is split in half between the body-of-work chapters (Chapters 4–7); the first two chapters aim to extract world knowledge from Wikipedia, while the second two apply the extracted world knowledge to geographic information retrieval. In Chapter 4, I explore the classification of Wikipedia articles as a precursor to Chapter 5’s more complex task of disambiguating articles as specific locations. A supervised approach is taken to article classification to maximise recall and to see what proportion of Wikipedia is classifiable within a given confidence. Using the WordNet broad noun classification classes as a classification schema allows one to compare achievable accuracy between the classification classes.

Chapter 4 concludes with a case-study attempting to classify tags in Flickr, the popular photo sharing web site. Photo corpora such as Flickr and ImageCLEFphoto are of particular interest to geographic information retrieval as *where* a photograph is taken is integral to its meaning. The case-study of Chapter 4 shows that when classifying the class of entities rather than resolving entities to a gazetteer or ontology, ambiguity can be considerably reduced; however when disambiguating specific entities context based disambiguation is necessary.

Supervised classification is discarded in Chapter 5 in favour of heuristic rules for disambiguating Wikipedia articles as specific locations. This is due to the difficulty in generating a ground truth and the presence of various types of geographic meta-data one can take advantage of. The choice and hierarchy of heuristic rules is determined empirically and benchmarked against naïve baseline methods.

The second aim outlined in the Scope, to apply the extracted world knowledge to geographic information retrieval, is tackled through context based placename disambiguation. In Chapter 7 I return to classifying entities in a corpus; as I am disambiguating entities as specific locations rather than classifying their class, I leave the photo corpus of Chapter 4 to concentrate on the GeoCLEF newspaper corpus. Retrieval of newspaper articles is of similar interest to retrieval of photos as they tend to discuss events which happened in a specific place at a specific time. Newspaper articles have an added complexity as their scope can cover many disjoint locations and time periods, while a photograph covers an instant in time at a single location. This added complexity is a double edged sword: as well as adding more noise to the disambiguation process it provides a far greater context for placename disambiguation. To test performance on other corpora, in direct evaluation three corpora are considered: Newspaper articles, Wikipedia articles and a general text collection.

Chapter 7 extends the concept of context to include the location of the author of documents (when disambiguating locations in a document) and the location of a user (when disambiguating locations in a query). The Steinberg hypothesis is proposed: that all people have the same world view, considering closer locations of greater importance than distant ones. An analysis of how locations are referred to in different language Wikipedia is performed to validate this hypothesis. This analysis is extended to temporal references to see how temporal and spatial references are related. Applying the Steinberg hypothesis to the retrieval process remains future work outlined in Chapter 8.

Chapter 2

GIR and Wikipedia

2.1 Introduction

The ACM-GIS series of symposia (now ACM SIGSPATIAL) began in 1993 covering all aspects of research in the area of geographic information systems, specifically systems based on geo-spatial data and the representation of geographic knowledge. In 2004 a GIR workshop was held for the first time at SIGIR, covering research in geographic information retrieval, specifically information retrieval systems and methods that take advantage of the geographic scope of documents (Jones and Purves 2006). GIR is a fast growing area in the broader IR and GIS disciplines. It involves many of the methods generally associated with IR such as searching, browsing, storing and ranking documents as well as a series of its own problems.

This chapter will explore two of the broad areas covered by this thesis: GIR and Wikipedia. We begin by looking at how GIR has grown from the wider IR discipline, followed by a survey of GIR itself and the geographic resources it requires. The second part of this chapter looks at Wikipedia and its role in research and data mining.

2.2 From IR to GIR

Information retrieval generally views documents as a collection or “bag” of words. In contrast Geographic Information Retrieval requires a small amount of semantic data to be present (namely a location or geographic feature associated with a document). Because of this it is common in GIR to separate the text indexing and analysis from the geographic indexing. In this section I cover a few key IR concepts and methods.

2.2.1 Relevance

In the context of IR, relevance is the measure of how well a document fulfils an information need. One can consider two types of relevance:

- **Subjective:** Whether a document fulfils an information need is ultimately a subjective measure judged by the user. Some work has been done on modelling a user’s anomalous state of knowledge to allow for these subjective judgments (Kuhlthau 1991).

- **Objective:** Arguably some documents can be considered objectively relevant to an information need, i.e. regardless of the user, the information need will be fulfilled. This is particularly true of question answering systems for factual questions.

Despite the fact that different users would judge relevance differently, it is often convenient to assume relevance is independent of the user (van Rijsbergen 1979). Relevance is a key concept to IR and one that separates IR from DB retrieval (as explained in Chapter 1). There are many ways of calculating relevance. Probabilistic methods calculate the likelihood that a term will appear in a relevant document; the relevance is the combination of all the matching terms (Grossman and Frieder 2004). The Vector Space Model (VSM) is the most widely used method, implemented in the popular Lucene¹ and Xapian² IR systems, the MySQL free text search module (Sun Microsystems 2008) and Sphinx IR system for databases (Aksyonoff 2008). This is the method I shall concentrate on in this section. Geographic relevance is often treated separately to textual relevance and is covered in detail later in this chapter.

Vector Space Model

The Vector Space Model was proposed by Salton et al. (1975). Documents and queries are represented as vectors in a multi-dimensional space with one dimension for each term. The relevance measure is the cosine of the angle between the document and query vectors. Salton et al. showed that the VSM could be improved by automatically weighting the terms in a query and that there was little difference between manually assigned terms and automatically generated terms.

Robertson and Sparck-Jones (1976)'s work on probabilistic retrieval proposed that automatically assigned weights for terms should vary with the term frequency (tf), i.e. the number of times term t appears in document D , and the inverse of the document frequency (idf) i.e. the logarithm of the ratio of the total number of documents divided by the number of documents that contain term t . This weighting method is known as the tf-idf weight.

Further work by Robertson and Walker (1994) led to the BM11 weighting scheme, and, in TREC-3 the BM25 weighting scheme (Robertson et al. 1994), which is considered by many to be the most successful weighting scheme to date.

2.2.2 Phrases and named entities

Croft et al. (1991) showed that matching manually constructed query phrases to documents can produce significantly better results than a bag-of-words representation, and that automatic extraction of such phrases gives similarly positive results. More recently these results were repeated by Liu et al. (2004), where using dictionary phrases and proper names significantly improved over a bag-of-words model on the TREC collections. Zhang et al. (2007) went on to show how slight increases in the accuracy of the phrase recognition algorithm could significantly improve retrieval results.

Liu et al. (2004) identify four types of noun-phrases useful to retrieval: proper-names (a.k.a. named entities) – names of people, places and organisations; dictionary phrases – multi-word phrases occurring in dictionaries; simple phrases – a noun phrase of 2–4 words containing no sub noun phrases; and complex phrases – longer phrases composed of one or more shorter noun phrases. In this thesis we are largely concerned with the improvement to retrieval offered by named entities.

Performance in formal evaluation is a useful indicator of how a system will perform against the queries of real users (the merits of formal evaluation will be discussed in more detail in the next chapter).

¹<http://lucene.apache.org/java/docs/>

²<http://xapian.org/docs/>

Sanderson and Kohler (2004), Gan et al. (2008) and Mishne and de Rijke (2006) analysed real user queries taken from query logs and showed that it is common to want to browse by location and named entities as well as terms. Extracting named entities requires an extra level of semantic understanding beyond the bag-of-words model. There are a variety of approaches to this task that range in complexity; the most naïve of which is using regular expressions to describe simple patterns of characters that names, dates and places might be expected to take, as well as the words that will surround them. More complex rules can be formed by first running the document through a part-of-speech tagger. Rules can be specified using a combination of words and part-of-speech patterns, for example a placename could take the form of the word “near” followed by a noun phrase (Brill 1995; Densham and Reid 2003; Croft et al. 1991; Liu et al. 2004). More complex again are named entity recognition (NER) systems such as ANNIE, part of Sheffield University’s GATE toolkit (Cunningham et al. 2001), LingPipe³ or ESpotter (Zhu et al. 2005). These combine complex rules, language models and gazetteers to extract proper nouns, resolve anaphoras (linguistic elements that refer back to other elements) and tag whether they refer to people, locations or organisations.

2.3 GIR

Geographic Information Retrieval (GIR) is the area of Information Retrieval concerned with providing access to information that relates in some way to specific geographic locations (Jones and Purves 2006).

Browsing of structured and semi-structured data has been a task in computer science for years, particularly in the field of DB retrieval, where it is common to have a series of structured fields. Semi-structured data is also becoming more and more common place: for example, an on-line catalogue may have structured fields for a product’s cost and available quantity but unstructured free-text fields for a description. The INEX forum, which has set itself the task of evaluating retrieval methods for semi-structured documents, began in 2001. The motivation of INEX was to encourage research that exploited structured documents and to provide a forum where these methods could be compared (Fuhr et al. 2006). Lalmas (2000) describes a uniform representation of structured documents capturing different fields and tree structures.

In 2008 the CLEF evaluation forum adopted the INEX Wikipedia corpus for a Geographic Question Answering task. Allowing the formal evaluation of retrieval in semi-structured documents augmented with geographic data, examples include “Which Swiss cantons border Germany?” (Santos et al. 2008).

Browsing data by time, location and event has been one of the goals of IR for decades but it is only in recent years that necessary resources have existed. Larson (1996)’s seminal paper, *Geographic Information Retrieval and Spatial Browsing*, identifies the advantages of browsing via location over traditional query-then-browse methods. In a geographic query the user is able to specify that they require documents related to locations falling within a certain locality. Sanderson and Kohler (2004) analysed Excite’s query logs to discover what percentage of queries submitted to a search engine had a geographical term: they found that 18.6% of the queries in their sample had geographical terms, a significant proportion of internet searches. The results of Jones et al. (2008) and Gan et al. (2008) concur with that of Sanderson and Kohler (2004), they found 12% and 13% of web queries contain placenames respectively and a staggering 82% of web documents. Mishne and de Rijke (2006) performed a similar survey of Blog queries. They found that 52% of queries and 74% of filters contained a named entity.

El-Geresy et al. (2002) describe several different methods for representing spatial and temporal in-

³<http://alias-i.com/lingpipe/>

formation. They identify five browsing categories: location-based, object-based, event-based, functional and causal. Location-based models involve simply representing where objects are; this is the simplest representation for a GIR system. Object-based models represent each object separately, where each object holds a reference to its state and position. Event-based models extend the object model temporally, where events are represented as the change between states. The functional-model represents time as a process from one state to another rather than discrete changes. Finally, causal-models explicitly model the link between cause and effect.

From the perspective of search agents and the semantic web it would be ideal for documents to be annotated with meta data by the author (such as associated locations and dates). Unfortunately, the people producing such documents are often not prepared to annotate them, and there already exists a vast amount of un-annotated data. This means to efficiently navigate these vast resources, automated annotation methods are needed.

2.3.1 GIR systems

In the past few years a number of geographic search engines and browsing methods have been developed. These include digital libraries such as the Alexandria Digital Library, the Perseus Digital Library and G-Portal, which annotate objects and documents in their database and provide a map interface to browse them (Smith and Crane 2001; Hill et al. 2004; Lim et al. 2002).

The SPIRIT project was a large scale European project running from 2002 to 2005 based at Sheffield University with the aim of building a geographic search engine (Jones et al. 2004). The Tumba! project from the University of Lisbon's XLDB group is a system to add geographical scope to Portuguese web pages (Silva et al. 2004; Cardoso et al. 2005). Both Google⁴ and Microsoft⁵ are working on their own large scale geographic search engines; these search engines allow users to search for products and services in a defined area and are funded by targeted advertisements based on the search query.

Another application of geographic browsing being explored is navigating large photo collections. The EXIF tags of images can hold time and location meta data added by either the user or a GPS module built into the camera. Collections can then be made browsable either through a map interface or by typing in text queries. MediAssist from the CDVP group at Dublin City University allows the user to browse pictures using a number of dimensions including time, place and even weather by gathering data from the weather station closest to the image's tagged location (O'Hare et al. 2005). Microsoft's WWMX project (PlanetEye⁶ since 2007) allows members of the public to upload their pictures to a huge communal archive browsable with a map interface. Flickr is a popular photo sharing web site; it is free to share photos although additional services are available through subscription. It provides a map interface⁷ through Yahoo!'s web map services. Currently, it holds nearly 10 million geographically tagged images.

2.3.2 Placename disambiguation

The problem of placename disambiguation has been approached from many fields including Natural Language Processing, Topic Detection & Tracking and Geographic Information Systems. Wacholder et al. (1997) identified multiple levels of placename ambiguity: The first type of ambiguity is structural

⁴<http://www.maps.google.com>

⁵<http://www.local.live.com>

⁶<http://www.planeteye.com/>

⁷<http://www.flickr.com/maps>

ambiguity, where the structure of the words constituting the name in the text are ambiguous (e.g. “North Dakota” – is the word “North” part of the placename?). Semantic ambiguity is the next level, where the type of entity being referred to is ambiguous (e.g. “Washington” – is it a placename or a person?). Referent ambiguity is the last level of ambiguity, where the specific entity being referred to is ambiguous (e.g. “Cambridge” – is it Cambridge, UK, or Cambridge, Massachusetts?). Placename disambiguation involves solving multiple levels of ambiguity, either as a semantic interpretation problem or a classification problem.

Placename disambiguation can be considered a classification problem similar to that seen in cross-language information retrieval (CLIR) or machine translation. In CLIR there is a many-to-many mapping between a set of words in language *A* to a set of words in language *B*; this is analogous to the mapping between placenames and locations. As with placename disambiguation this mapping can be unambiguous, for example, English to German, “hippopotamus” becomes “Flusspferd”, or ambiguous, for example, “duck” becomes “sich ducken” (to crouch) or “Ente” (an aquatic bird). The main difference between placename disambiguation and other disambiguation problems is there exists an implicit topological and geographic relationship between locations that can be exploited for disambiguation. Gazetteers are generally used to provide a set of classification classes for placenames. Gazetteers are lists of placenames mapped to latitudes and longitudes (Hill 2000).

Placename disambiguation is a sub-task of the more general problem of word sense disambiguation (WSD), the automatic assignment of semantic senses to ambiguous words. WSD is primarily concerned with semantic ambiguity. Ide and Véronis (1998) concisely define the problem of disambiguation as

“matching the context of the instance of the word to be disambiguated with either information from an external knowledge source, or information about the contexts of the word derived from corpora.”

This captures the two most common approaches to disambiguation, rule-based (or knowledge-driven) and data-driven (or corpus-driven), which will be described below in the context of placename disambiguation. Additionally I shall describe semi-supervised methods, which cover the overlap between the two approaches, and commercial systems, which do not release details of their methods. An extensive literature review of supervised and semi-supervised methods for placename disambiguation is provided at the start of Chapter 6. In Chapter 5, a rule-based method is used to crawl Wikipedia to generate a co-occurrence model, which is applied as a data-driven method to free text in Chapter 6.

Rule-based methods

The rule-based disambiguation methods apply simple heuristic rules to placename disambiguation. The most basic disambiguation rules use a specially constructed gazetteer where there is only a single location for each placename; these default locations are selected on various criteria including size, population and relative importance (Li et al. 2003; Clough et al. 2004).

More complex methods of disambiguation define a geographic scope for a document. The geographic scope can be based on where the document was published or defined by fitting a minimum bounding polygon around locations occurring in the document; this assumes locations close together geographically will generally occur close together within a document (Rauch et al. 2003; Cardoso et al. 2005; Zong et al. 2005). Woodruff (1994) proposes a method of polygonal overlay for placename disambiguation. Brunner and Purves (2008) demonstrates that there is a significant spatial autocorrelation between ambiguous locations, i.e. locations with the same name are more likely to be close together than randomly selected

locations. They argue this reduces the validity of distance based disambiguation methods. Heuristic rules such as, *if a place is mentioned once, until quantified again the same place will be repeatedly referred to*, allow the scope of documents to conditionally change (Leidner et al. 2003; Gale et al. 1992). Note this geographic scope can double up as a document footprint (discussed in the next section).

One of the most accurate methods of disambiguation is to look at the contextual information in the 2-5 words preceding and following a placename. A containing entity will often be mentioned (e.g. London, *Ontario*), as will a feature type (e.g. Orange *County*). Contextual information can provide the type of geographic location being referred to or imply a relationship with another location. Often these elements of description are enough to disambiguate a location (Rauch et al. 2003; Clough et al. 2004; Garbin and Mani 2005; Olligschlaeger and Hauptmann 1999). Buscaldi and Rosso (2008b) perform an evaluation between a map-based approach to disambiguation and a knowledge-based approach. Their map based approach attempts to minimise the distance of ambiguous placenames from the centroid of a document’s scope. Their knowledge based approach maximises the conceptual density in a vertical topology tree (discussed later). They found topographically close locations — such as referent locations — to be more useful for placename disambiguation than physically close locations.

Heuristic rules can be applied one after another, either in a pipeline or an iterative loop, or the results of different rules can be combined either with weighted voting or probabilistically (Rauch et al. 2003; Li et al. 2003). The Perseus Digital Library use a complex set of rules to classify entities as people or placenames and a further set to disambiguate semantic meaning (Crane and Jones 2005).

Data-driven methods

The data-driven methods of disambiguation generally apply standard machine learning methods to solve the problem of matching placenames to locations. The problem with these methods is that they require a large accurate annotated ground truth; if such a corpus existed naïve methods, e.g. Bayes’ theorem, or more complex methods, e.g. Support Vector Machines, could be applied. Small sets of ground truth have been created for the purposes of evaluation or applying supervised learning methods to small domains (Bucher et al. 2005; Leveling et al. 2005; Nissim et al. 2004). However, a large enough corpus does not yet exist in the public domain to apply supervised methods to free text.

Ide and Véronis (1998) describe the problem of data sparseness with respect to WSD. They observe that enormous amounts of text are required to accumulate enough occurrences of even relatively common words. Three approaches are given to counter this problem:

- **Smoothing.** A non-zero probability is assigned to unseen events.
- **Class-based models.** The granularity of the classification is decreased by grouping terms into classes. This method is seen in Garbin and Mani (2005)’s feature classifier where the *type* of location is classified, and Smith and Mann (2003)’s *back off* classifier where the state or country is classified.
- **Similarity-based methods.** As with the class based methods, the granularity of classification is decreased. However rather than grouping each term into fixed classes, each term is grouped with similar terms, creating per-term classes. For example, one could consider not only the context Cambridge, UK occurs in when constructing a “Cambridge” disambiguater, but the contexts of all references to locations within a 50 mile radius.

Semi-supervised (bootstrapping) methods

Semi-supervised techniques are similar to data-driven methods; a smaller annotated corpus is required than for data-driven methods (however at least one example of each ambiguity) and an additional un-annotated corpus is used to infer further characteristics of the data (Bucher et al. 2005; Leveling et al. 2005). Smith and Mann (2003) train a Naïve Bayes' classifier to annotate texts about the American Civil war. Garbin and Mani (2005) build a training set by disambiguating placenames with coreferents. This training set is then used to learn a decision list using the Ripper algorithm. They classify the feature types of locations using surrounding words and properties of the placename as features. Classifying the feature type of a placename is often enough to disambiguate it. For example knowing if "Victoria" refers to a City or State is enough to discriminate between the capital of British Columbia and the Australian state.

Commercial Systems (Black Boxes)

MetaCarta provide free placename disambiguation tools via a web API. Their GeoParser API provides an interface to the MetaCarta LocationFinder and GeoTagger⁸. It takes a document as input and extracts the locations referenced in the text. The number of documents that can be disambiguated is limited to 100 per day, so it has limited use as a method for annotating a corpus and the methods used are not published.

In 2006 Google Maps released an API to their Geocoding software. Not as powerful as MetaCarta's GeoTagger, Google's Geocoder will provide latitude, longitude and scale when provided with an address. The Geocoder requires placenames and addresses to be extracted before it processes them so clearly it cannot use document context for placename disambiguation.

GeoNames provide a third placename disambiguation API. Their "RSS to GeoRSS" converter takes as input an RSS feed, extracts the placenames and returns a GeoRSS output where placenames have been marked up with latitude and longitude⁹. This is similar functionality to MetaCarta's GeoTagger.

2.3.3 Geographic indexing

The representation and storage of geographic data is an integral part of GIR. The overlapping disciplines shown in Figure 1.1 have differing approaches to indexing geographic and temporal data. Below I list three fields related to GIR and their approaches to indexing spatial data.

- **Spatial databases.** Time and location data have been stored in databases for decades. In 1989 storing these complex data types was recognised as a problem distinct from storing standard structured data when the Symposium for Spatial Databases (SSD)¹⁰ began (Güting et al. 1999). The priorities of a spatial database include: dynamics (data must be able to be inserted or deleted), secondary and tertiary storage management, a broad range of supported operations, simplicity, scalability, and time and space efficiency (Gaede and Günther 1998).

Before multi-dimensional indexes, multiple dimensions of data were indexed with a series of one dimensional indexes with a single index for each dimension, e.g. a B-Tree. Early multi-dimensional access methods, e.g. the K-d-tree and quad-tree, were not optimised for secondary storage management (Gaede and Günther 1998). Guttman (1984) proposed the R-Tree as an efficient way for

⁸<http://developers.metacarta.com/api/geotagger/web-service/1.0.0/>

⁹<http://www.geonames.org/rss-to-georss-converter.html>

¹⁰Now the Symposium for Spatial and Temporal Databases (SSTD)

indexing two dimensional regular data. R-trees allow both polygons and points to be indexed. There are several extensions to the R-tree, e.g. the R⁺-tree or R*-tree, that can make updates and deletion more efficient and alternatives, e.g. the Z-file. However, benchmark tests are yet to show a single multi-dimensional indexing scheme to be superior (Gaede and Günther 1998).

- **GIS.** A Geographic Information System is software that provides access to geometrically structured information based on digital maps (Jones and Purves 2006). The emphasis for GIS is to represent complex geographic data as accurately as possible.

Many GIS take advantage of the efficient speed and storage spatial databases can offer. However, these indexing systems require coordinates to be projected onto a flat surface. Spherical or ellipsoid indexing systems do not require such a projection. Dutton (1996) proposes the Octahedral Quaternary Triangular Mesh (O-QTM) where the Earth is represented as an Octahedron with each face hierarchically split into four triangles. The accuracy of O-QTM is similar to that which can be achieved with a 1:25,000 scale map (about 60m) and allows a single representation for multiple resolutions of data. Lukatela (2000) capture the surface of the Earth to an even greater level of accuracy by splitting the Earth into a Triangular Irregular Network (TIN). The TIN allows not only the Earth to be represented at varying levels of detail in a single model but also elevation to be represented.

- **Application driven.** The field of GIR is a largely practical area of computer science. Often simple solutions that satisfy user requirements will be chosen over more complex systems because of ease of use and implementation. The C-Squares index was developed to represent oceanographic data, where large geographic footprints of varying size and detail need to be represented, searched and exchanged (Rees 2003). The C-Squares index works by decomposing the Earth into a hierarchical grid. The Earth is initially partitioned into 648 squares, which are recursively partitioned into quadrants. The C-Squares index does not have many of the advantages of the O-QTM or the TIN, however, it is easy to conceptualise, implement and use. The advantages of the C-Squares index are: it can be held in a standard text index, the representations produced for each location are unique and portable, and complex shapes can be represented (polygons with holes, multi-part objects, concave polygons etc.) (Rees 2003).

GIR has its own requirements of a geographic index and data-representation. There is a trade-off to be made whether a flat representation or spherical representation is adopted. Egenhofer and Mark (1995) describe the human appreciation of spaces as flat, therefore one could argue a map-like data representation would lend itself to fulfilling an information need. A flat representation also lends itself to being rendered as a map. A spherical representation has the significant advantage that complex shapes can wrap across the international dateline or over poles.

GIR differs from GIS in, amongst other things, the style of query. GIR queries tend to be stated in natural language and vary in the type of geographical phrases and relation used. There are a combination of ambiguous relationships (e.g. “near”) and exact (e.g. “within 10km”) (Gey et al. 2006; Sanderson and Kohler 2004). Gey et al. (2006) present a classification schema for geographic queries demonstrating the relationship between subject and location can vary considerably (see Appendix B). The results of such queries are often required in a ranked list (detailed in the next section) and may need to be combined with the results of other queries such as textual or temporal results.

2.3.4 Geographic relevance

An essential part of IR is assigning a relevance score to documents to represent how well they fulfill a user's information need. This is done by attempting to emulate how a user would judge a document relevant to a query. Section 2.2.1 described how a relevance score can be calculated for a text document with respect to a query; when calculating geographic relevance the motivation is the same, however the methods employed are quite different. Egenhofer and Mark (1995)'s *Naive Geography* captures and reflects the way people think and reason about geographic space and time, both consciously and subconsciously (analogous to the concept of Naïve Physics). They identify fourteen elements of Naive Geography including: People assume the world to be flat; geographical space and time are tightly coupled; geographic information is frequently incomplete; geographic space has multiple conceptualisations and levels of detail; distances are asymmetric and conceptualised at a local scale; and, topology is more important than quantitative distance. There is not currently a consensus about whether it is more appropriate to split textual and geographic relevance or deal with them both simultaneously (Cardoso and Santos 2008), however, the majority of systems take the former approach, and as such, this is what I shall concentrate on in this thesis. Methods of assigning geographic relevance differ on whether they assign a *footprint* to documents or consider the locations referenced as a collection of points or polygons. A document footprint is a polygon or collection of polygons representing all the geographic areas referred to and implied.

Jones et al. (2008) introduce the idea of a document's objective geographic relevance. Once a query has been categorised as a geographic query, more geographically orientated documents can be returned. They identify the number and type of placenames referenced as suitable indicators to the geographic-ness of documents. There are a number of approaches to query-document geographic relevance — that is the relevance between a query and document with respect to geographic entities — GIR systems can either use one or a combination of methods:

Footprint methods

There are various methods of defining the footprint of a document. They generally require a polygon or minimum bounding rectangle (MBR) to be drawn around the locations of interest. A query can then be considered as either a footprint in its own right or as a point (Fu et al. 2005a). If the query is also considered a footprint, relevance can be considered as the area overlapping between the two polygons (Beard and Sharma 1997). Alternatively, the distance between polygons can be measured either by the minimum distance or distance between centroids (Fu et al. 2005a).

Many systems representing footprints as MBRs keep all rectangles aligned to a grid. This has several advantages including that the footprints can be represented with two points (opposite corners) rather than four points (every corner), the footprints can be stored and searched more efficiently in a spatially aware index, and data for this representation is more readily available (although more accurate and more complex data is increasingly being released in the public domain¹¹). The main disadvantage with this system is that dependent on the orientation of locations, the size, shape and relationship between MBRs can vary significantly: in Figure 2.1, MBRs are fitted around two sets of points. Rotating the points 45° significantly changes the relationship between the MBRs.

¹¹<http://code.flickr.com/blog/2008/10/30/the-shape-of-alpha/>

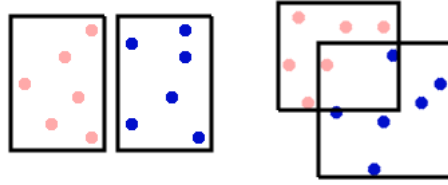


Figure 2.1: An illustration of how orientation can effect the relationship between MBRs

Distance methods

If a document is considered a collection of points, relevance between the query and each point in a document must be measured. These relevance judgments must then be combined to give a single geographic confidence score. More will be discussed about complex methods of combining multiple relevance judgments in the next section, however, in calculating geographic relevance it is common to simply take the location with the greatest relevance score (Hauff et al. 2006). If distance is the sole measure of geographic relevance employed then the relevance score will be inversely proportional to the distance of the closest location (Jones et al. 2008).

The simplest method of computing geographic relevance is to use the distance between points. GIR systems vary in their distance calculations, some systems project the Earth's surface onto a two dimensional plane and apply Euclidean Geometry to calculate distances. Alternatively, the world can be considered a sphere, or more accurately an ellipsoid corresponding to a specific datum, and spherical or elliptic geometry can be applied. In this case distance is considered the length of the geodesic between two points. In practice, the model or projection used for the Earth's surface makes little difference and is largely a speed-versus-accuracy implementation choice.

A drawback of using absolute distance as a method of calculating relevance, is the human appreciation of distance is relative (Montello 1992). As previously noted the aim of assigning relevance to documents is an attempt to model a human user. The human understanding of distance is context dependent and asymmetric, people conceptualise distances differently at the local and global level. For example the appreciation of the distance between London and Edinburgh varies for a user situated in New York or Edinburgh (Worboys 1996). In Chapter 7 we examine how, as the distance increases between a person and a location, its relevance decreases.

Topological methods

The human appreciation of the relationship between locations is asymmetric and inconsistent, and as such does not easily map into a metric space (Egenhofer and Mark 1995; Montello 1992). When geographic relationships are used in queries they are often ambiguous terms such as *near* or *close*. There are many ways of estimating these measures: for example the travel times between locations (in minutes and hours) or the required method of transport (walking, driving or flying) (Egenhofer and Mark 1995).

One method of modelling how people judge relationships between locations is looking at topological distance. Topology is the qualitative properties of how locations are connected. There are multiple ways of representing topology and several factors need to be considered. Physical topology and political topology differ and both are components in judging relevance. For example, the British Isles is made up of two islands: Great Britain and Ireland; politically, the United Kingdom consists of four countries:

Scotland, England, Wales and Northern Ireland, while the Republic of Ireland is a separate nation. These overlapping interpretations of topology add to the complexity of modelling geographic relevance.

Vertical topology represents the hierarchical nature of locations. For example, Figure 2.2¹² illustrates the counties and countries of Great Britain, the hierarchy of which is represented as a connected graph. If all the locations on the same layer of a vertical topology graph are considered tessellating polygons, for example the counties of Great Britain, and all the locations are considered nodes then a horizontal topology graph can be generated by connecting all nodes that share a border, see Figure 2.2 (Schlieder et al. 2001).

The distance and overlap between locations found by navigating vertical and horizontal topology graphs can provide a usable model for how related humans consider locations to be (Egenhofer and Shariff 1998). Egenhofer and Shariff (1998) and Martins et al. (2006) apply Egenhofer and Mark (1995)'s "Topology matters, metric refines" premise. Egenhofer and Shariff capture the topological relationship between locations based on the *splitting* and *closeness* of the intersection between polygons. Martins et al. assign normalised values between 0 and 1 to the vertical topology similarity, adjacency (based on horizontal topology), containment (based on vertical topology) and Euclidean distance. The geographic similarity is then defined to be the convex combination of these values. Rodríguez and Egenhofer (2004) represent classes of geographic object in a hierarchy. They define the Matching-Distance Similarity Measure, a similarity measure that can be applied to a geographic ontology to measure how similar geographic classes are.

2.3.5 Combining text and geographic relevance

Generally in IR it is desirable to browse the results to a query in a single ranked list (in the majority of evaluation forums, this is a requirement). Unless the geographic relevance of a document to a query can be assessed independently of the text relevance, this task requires the combination of geographic and text relevance. The combination of different ranks is a problem often approached in structured document retrieval. Robertson et al. (2004) propose an extension to the BM25 algorithm that allows scores for multiple overlapping fields to be combined non-linearly.

Some experimentation has been done with the graphical representation of different relevance measures as an alternative to combining ranks. For example displaying a 3D grid with each dimension representing a different relevance measure to show the user geographic, temporal and text relevance (Hobona et al. 2005), or representing different relevance measures as bars within a glyph (Beard and Sharma 1997). However, these methods often have a steep learning curve where users have to learn how to interpret the results.

Martins et al. (2006) return a single relevance value as the linear combination between geographic and text relevance. Text relevance is calculated using the vector space model with the BM25 term weights. Geographic relevance is calculated as the convex combination of three normalised measures: horizontal topographic relevance, vertical topographic relevance and distance. Cardoso et al. (2007) continue this work testing how multiple geographic footprints can be combined. They compare three methods of combining relevance scores: the mean, maximum and Boolean score, where Boolean is equivalent to filtering. They found the maximum and Boolean methods to be best.

Wilkins et al. (2006) examine combining scored lists with a weighting dependent on the distribution of the scores within each list. Their hypothesis is that cases where the distribution of scores undergo a rapid initial change correlate with methods that perform well. Currently this hypothesis has only been

¹²I would like to acknowledge the Association of British Counties for use of their map.

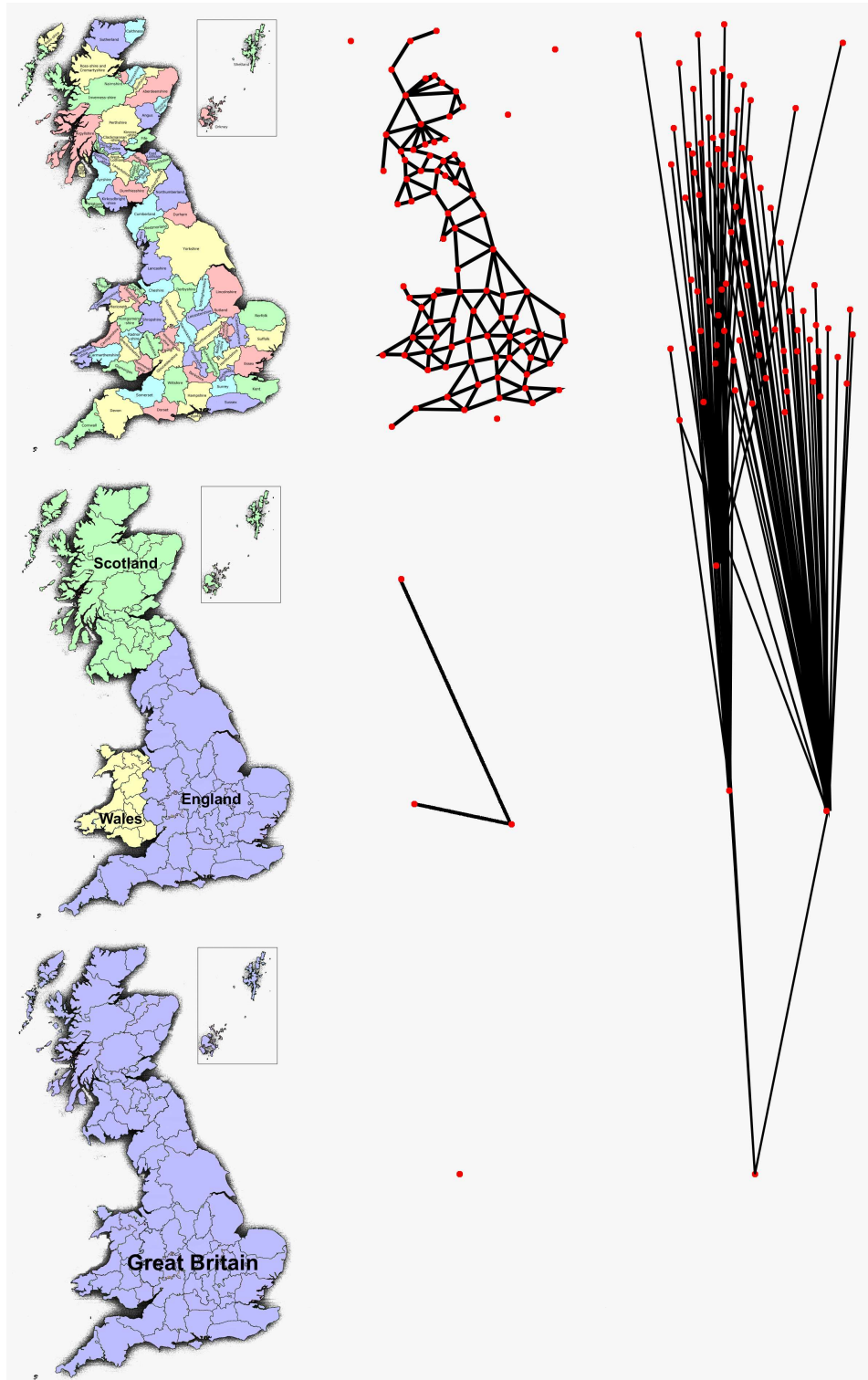


Figure 2.2: The counties' and countries' of Great Britain horizontal topology and vertical topology. The figures from left to right show the political geography of Great Britain, the horizontal topology with areas sharing a land border linked and the vertical topology with sub-areas linked. From top to bottom the figures show counties, countries and landmass of Great Britain. In the topology maps areas are represented as a red dot for their centroid

applied to image retrieval but it is applicable to all score based data-fusion applications.

An alternative to a combination of ranks or scores is to filter one rank against another. A document cut off value is selected for one rank above which documents are considered relevant and below which not-relevant. The other rank can then be filtered against these relevant documents (Overell et al. 2006; Vaid et al. 2005; Hauff et al. 2006). For example, given the GeoCLEF 2006 query “Snowstorms in North America,” documents can be ranked by their text relevance to “Snowstorms.” This text ranking can then have all documents that do not refer to “North America” filtered out. Fusing geographic and text relevance is covered in more detail in Appendix B.

2.4 Geographic resources

Geographic browsing requires several tools and resources to make it efficient and scalable. The development of representations for geographic data is an area of GIR that is currently very active. Many geographical terms and queries represent fuzzy areas, e.g. the American Midwest. Other geographical terms require complex polygons to represent them, e.g. Eurasia (Larson and Frontiera 2004). Zhou and Jones (2003) discuss in depth the storage of geographical data at multiple resolutions. There is a trade off between the accurate representation of geographic areas on one hand and the processing speeds and implementation and storage costs on the other. Another consideration is the information available. An alternative to attempting to accurately represent geographical areas is to assume each area is a point or simple polygon.

To allow documents to be automatically annotated with spatial data two resources are needed: an ontology of possible annotations and a gazetteer of locations. A uniform depository of geographic information has been proposed by Zhou and Jones (2003). Three competing XML schema have emerged: GPX is the Global Positioning system eXchange format¹³, developed to transfer geographic information between devices and over the web. The Geographic Markup Language (GML) was developed by the Open Geospatial Consortium for the modelling, transport and storage of geographic information following the openGIS specification¹⁴. Keyhole Markup Language (KML)¹⁵ is the XML format used by Google for use with their geographic services. The problem of providing a controlled set of geographic annotations has had multiple approaches: Methods of combining gazetteers have been offered as a solution (Axelrod 2003), as have Semantic Web ontologies such as Dublin Core and SPIRIT’s Geographic Ontology (DCMI Usage Board 2006; Rodríguez and Egenhofer 2004; Fu et al. 2005b). However, there has yet to develop a uniformly adopted system of geographic representation.

Gazetteers are an essential part of geographical browsing. Schlieder et al. (2001) identify gazetteers as a sub-set of GIS systems providing a controlled vocabulary of placenames. Gazetteers are often treated as thematic thesauri in GIR systems providing a set of possible annotations for documents. Most gazetteers have some additional information (population, size, feature type, etc.), which can be used to assist disambiguation. The Getty Thesaurus of Geographical Names (TGN) assigns a unique identifier to every location making annotations with this gazetteer portable, it contains approximately 800,000 locations (Harping 2000). The most extensive publicly available lists of geographic names (although less accurate than Getty) are the GNIS and NIMA gazetteers which between them cover approximately 14 million locations¹⁶.

¹³<http://www.topografix.com/GPX/1/1/>

¹⁴<http://www.opengis.net/gml/>

¹⁵<http://code.google.com/apis/kml/documentation/>

¹⁶<http://www.nga.mil/>

To allow users to browse documents geographically they must be presented with a representation of the Earth that can be used to express their query and browse their results. There are currently no available copyright-free maps at a high enough resolution of the Earth's surface to provide such an interface. However, there do exist enough freely available data to generate maps, and through NASA's Blue Marble project, high-resolution satellite images of the Earth are available (Stockli et al. 2005). The Open Street Map Foundation¹⁷ are attempting to change the status quo and produce high quality geographic data in the public domain, but it is a slow process.

An alternative approach to providing users with a map interface is to use one of the growing number of Web 2.0 maps available. MapServer¹⁸ was one of the first applications designed for rendering maps over the internet. Initially developed by NASA and the University of Minnesota in 1997, MapServer is an Open Source map rendering package. Rich Internet Applications (RIA) fall under the umbrella term of Web 2.0; these are web applications where much of the processing is performed on the client side while the bulk of the data is held by the server. This is a convenient structure for internet maps. Microsoft Live is a commercial example of a RIA map. Web APIs are another technology falling under the Web 2.0 umbrella. Web APIs allow external servers to host maps and geographic data allowing users to download the images and overlay their own content. Google provide two APIs: Google Maps and Google Earth¹⁹. Yahoo! also provide a web API²⁰. An Open Source alternative to Google and Yahoo! Maps was launched by MetaCarta in 2006: the OpenLayers project²¹.

2.5 Wikipedia

The core resource used in this thesis for mining information besides geographic gazetteers is Wikipedia. This section provides some background to Wikipedia, Wikipedia articles and the contributors to Wikipedia, known as Wikipedians.

User generated content (UGC) is a fast growing trend on the Internet (Tapscott and Williams 2008). These web sites include community sites (such as MySpace), media sharing sites (such as YouTube and Flickr), blogs (such as Blogger), and wikis (such as Wikipedia). They allow users to contribute their own content in a virtually unmoderated environment. The number of blogs has risen from a few thousand in the late 1990s to tens of millions in 2005 (Mishne and de Rijke 2006). The IR community has only recently started to take advantage of these new data sources with the Blog track being introduced to TREC in 2005, INEX using the Wikipedia collection from 2005 and the "Question Answering using Wikipedia" track (WiQa) added to CLEF in 2006 (Jijkoun and de Rijke 2006; Fuhr et al. 2006).

Wikipedia is the largest reference web site on the Internet. It was launched in 2001 as *The free encyclopedia that anyone can edit* (Wikipedia 2008b). Wikipedia is an example of *Wiki* software, allowing content to easily be authored by multiple people. The content is collaboratively written and updated by volunteers (Wikipedia 2008b); it is extremely useful as a resource due to its size, variation, accuracy and quantity of hyper-links and meta-data (Kinzler 2005; Nakayama et al. 2008).

Appendix A contains a brief history of Wikipedia, a description of Wikipedia articles and Wikipedia in research.

¹⁷<http://www.openstreetmap.org/>

¹⁸<http://mapserver.gis.umn.edu/>

¹⁹<http://www.earth.google.com>

²⁰<http://www.maps.yahoo.com>

²¹<http://www.openlayers.org/>

2.5.1 Wikipedia articles

I will begin this section with some terminology: a *Wikipedia page* is any web page accessible on the Wikipedia web-site. These include pages describing categories, templates, users, administration pages, portals and encyclopædic content. *Wikipedia articles*, or short *articles* refer to a subset of Wikipedia pages of only encyclopædic articles. Each article describes a single unambiguous theme or concept. Unless otherwise stated references to the English Wikipedia use a November 2006 dump, and in references to other wikipedias, Chapter 7 uses March 2008 dumps.

The content of Wikipedia is guided by the Wikipedia Policies and Guidelines, on a macro level this includes Wikipedia's five guiding pillars, while on a micro level, the style, look and tone of individual articles are governed by Wikipedia's Manual of Style. These are covered in detail in Appendix A.

As every article in Wikipedia is required to have a unique title it is possible to unambiguously identify and link to articles. It follows, as each article describes a single concept, concepts can be linked to unambiguously. It is therefore possible to disambiguate polynoms in articles by linking them to the title of the page describing their meaning.

Lüer (2006) identified disambiguation within Wikipedia as the mapping from a word to an article: disambiguation of polynoms is accomplished in Wikipedia by a combination of requiring every article to have a unique guessable name and explicit disambiguation pages; resolution of synonyms is achieved through a network of *redirect* pages. The onus is then on a page author (and editors) to correctly link to intended pages they reference.

Nakayama et al. (2008) identifies the unique-title-for-every-article policy (referred to as “disambiguation by URL”) as one of Wikipedia's most notable characteristics. This keeps references to ambiguous concepts, such as “Apple” that can refer to the fruit or the computer company, semantically separate (the pages are titled *Apple* and *Apple Inc* respectively).

2.5.2 Wikipedians

When mining information from Wikipedia, one should consider for a moment those who contribute this information. Referred to as *Wikipedians*, 6 million users contribute content to Wikipedia articles (Wikimedia 2008). Wikipedians are essentially anonymous, identified by either a user name when logged in or an IP address otherwise. In the discussion of WikiScanner in the next section we will see why people are generally more anonymous when logged in.

Stvilia et al. (2005) define four roles agents can take with respect to the *Information Quality* of Wikipedia:

1. **Editor agents:** Agents that add content to new or existing articles.
2. **Information Quality Assurance agents:** Agents that maintain the information quality of articles, for example minor edits correcting article formatting, spelling or categories.
3. **Malicious agents:** Agents that purposefully degrade articles by vandalism or adding content known to be false.
4. **Environmental agents:** Agents that act on the outside world. These generally cause Wikipedia articles and the state of the world to diverge but can potentially affect the opposite.

Swartz (2006) argues the majority of words contributed to Wikipedia are from *editor agents*, while the majority of edits are from *information quality assurance agents*. Burial et al. (2006) observed the

behaviour of Wikipedians and the changes in Wikipedia over a four year period. The proportion of edits by anonymous (not logged in) editors varied between 20% and 30%. The number of editors an article has follows a Zipfian distribution with 7.5% of articles having only one editor, 50% of articles having more than seven editors and 5% of articles having more than 50 editors. Editors commonly disagree. There are many causes for disagreement, most commonly differing opinions or malicious editors. Each article has a dedicated discussion page for resolving such issues where Wikipedians are invited to back up their arguments with appropriate authoritative sources (Giles 2005). The majority of actions that are reverted²² are vandalism (malicious edits). It is considered rude to revert an addition simply through a difference of opinion and can lead to double reverts and progress to *revert wars* (Burial et al. 2006). 6% of all edits are reverted, a further 5% of those reverts are further reverted (Burial et al. 2006).

2.5.3 Accuracy

Since Wikipedia forked from Nupedia and discarded the editorial/peer review process in favour of a wiki there have been many debates on how accuracy can be maintained. This has led to a number of studies repeatedly testing different aspects of Wikipedia's accuracy, but current debates remain unsolved.

The most notable article on this subject was published in the highly respected journal, Nature, in 2005 and kicked off a heated debate. Giles (2005) compared Wikipedia and the Encyclopædia Britannica. They performed a double blind peer review of 42 entries from a cross section of scientific fields. They found that Wikipedia's accuracy was approaching that of Britannica. 32% more errors were found in Wikipedia than Britannica, however the articles were generally longer. Britannica published a damning open letter in rebuttal to this article entitled "Fatally Flawed" attacking the methods used in the study and the presentation of results (Encyclopædia Britannica Inc 2006). Nature responded refusing to retract the article, defending both the methods used and the presentation of results (Nature 2006).

Wikipedia has come under further attack since the launch of WikiScanner, a website that maps from companies and organisations to the Wikipedia articles they edit via IP-ranges (Griffith 2007). By analysing the anonymous edits in Wikipedia it is possible to see people editing Wikipedia pages with which they have a conflict of interest. Evidence was found of employees of Diebold inc and the Church of Scientology removing criticism from their pages, and employees of Microsoft adding that MSN search was a major competitor to Yahoo! and Google (Griffith 2007).

Despite Wikipedia's obvious popularity, the information it contains comes without authority. Jimmy Wales discourages its use in academic work: "For God sake, you're in college; don't cite the encyclopedia," and advises caution: "It is pretty good, but you have to be careful with it" (Young 2006). Waters (2007) provides a similar view, criticising Wikipedia as a primary source, instead promoting it as a good place to find further reading. Many of Wikipedia's critics argue an encyclopædia requires an editor and the open source model is not appropriate (Waters 2007; Giles 2005; Encyclopædia Britannica Inc 2006).

Stvilia et al. (2005) attempt to indirectly statistically measure the *Information Quality* (IQ) of Wikipedia. The motivation for this is if articles are given an IQ measure, it can aid peoples decisions when acting upon those articles. They propose 19 measures based on article meta-data, which are combined to form seven metrics measuring article IQ. These metrics are Authority, Completeness, Complexity, Informativeness, Consistency, Currency and Volatility. They found a huge variation in the quality of articles. Using a sample of Featured Articles and randomly selected articles, they found using meta-data alone it was possible to classify articles as low or high quality with an accuracy greater than 90%.

²²A revert is a one click undo of another user's edit

The accuracy of relational statements between named entities and links from one article to another within Wikipedia was tested by Weaver et al. (2006). They found these links and relationships to be accurate over 97% of the time. As in Chapter 6 I use Wikipedia as a corpus for supervised placename disambiguation, the accuracy of the corpus provides a ceiling for the accuracy achievable by the classifier using it as training data.

2.5.4 Mining semantic information

From the early stages of Wikipedia's growth people have tried to extract data that a computer can *understand*. We define understand as being able to infer additional information from an article beyond a simple bag-of-words. This section will provide an overview of attempts to infer meaning from Wikipedia articles. Sub tasks of assigning a semantic category to a Wikipedia article and disambiguating Wikipedia articles as specific locations will be discussed further in Chapters 4 and 5.

Wikipedia's suitability for data mining was evaluated in Kinzler's paper *WikiSense — Mining the Wiki*, where the use of the highly formatted template data and links between articles were highlighted as particularly useful. They suggest it should be possible to classify pages, extract properties from pages, and extract relationships between pages. They also suggest that it should be possible to cluster pages based on Wikipedia's hyper-linked structure (Kinzler 2005). More recently, this topic was examined by Nakayama et al. (2008). They identify five aspects of Wikipedia that are particularly useful for data mining: unique article titles, anchor texts, live updates, the link structure, and the link types. They also identify a number of applications of mining Wikipedia including thesaurus generation, word sense disambiguation, ontology construction and bilingual dictionary construction. Medelyan et al. (2008) provide a more in depth review of mining *meaning* from Wikipedia. We further explore these applications in the body-of-work of this thesis.

The lack of machine readable meta-data in Wikipedia is a significant problem for people wishing to mine world knowledge; in fact, Chapters 4 and 5 of this thesis are dedicated to this. The SemWiki workshops began in 2006 to discuss issues on this specific problem. Current opinion is split whether alternative machine readable UGC resources should exist (Semantic-Wikis), whether Wikipedia should be augmented with machine-readable data (Kinzler 2005; Krötzsch et al. 2005) or whether current data-mining techniques should be improved to a point where we can extract machine-readable data from information designed for humans. In this thesis I shall concentrate on extracting information from the human centered resource, Wikipedia.

Extracting relations and facts from Wikipedia

There are a multitude of approaches for extracting relationships from Wikipedia. Kinzler (2005) proposes that relationships between articles could be extracted from Wikipedia categories. Gabrilovich and Markovitch (2007) describes a method to measure the relatedness between Wikipedia articles, terms and text by looking at the distance between articles in a vector space. Strube and Ponzetto (2006) propose a similar method comparing the body text of articles and additionally using Wikipedia's category tree to compare the relatedness of articles. Nakayama et al. (2008) propose the internal link structure as a measure of topic locality. Weaver et al. (2006) propose relational statements in addition to internal links as suitable for mining relationships and useful for named-entity recognition.

Suchanek et al. (2007) extract a variety of semantic relations from Wikipedia. Similarly to Strube and Ponzetto (2006), they use the category structure of Wikipedia: relations are extracted from the category name. In addition to categories, redirects and internal links are used. The category tree of

Wikipedia is discarded as being too inconsistent. Instead, Wikipedia categories are mapped onto the WordNet ontology to allow further inferences.

Auer and Lehmann (2007) extract relational statements from the structured data in article templates and store them as RDF statements. This provides a queryable database of over 8 million entries, which provides the foundation of the DBpedia project (DBpedia 2008). DBpedia links other projects mining Wikipedia, e.g. Yago²³, and dozens of other free data sources such as digital libraries and gazetteers. Powerset²⁴ is a commercial company extracting similar data to DBpedia. They extract what they term *factz* from Wikipedia pages: subject, object, relation triples.

2.6 Discussion

To conclude this chapter I will touch on where Wikipedia has overlapped GIR. This will be revisited in Chapter 5. Two projects are currently underway allowing users to geographically tag Wikipedia: The WikiProject Geographical Coordinates (known as WikiCoords) (Wikipedia 2008a) and Placeopedia (Steinberg 2008). WikiCoords is integrated into the Wikipedia site and allows people to add geographic coordinates to any page. Placeopedia is a Google Maps *mashup* allowing people to locate Wikipedia articles on a google map (Steinberg 2008). In academia Buscaldi et al. (2006), Silva et al. (2004), Hauff et al. (2006), and Overell and R uger (2007) use Wikipedia to generate a gazetteer/geographic ontology and for geographic query expansion.

Despite controversy regarding its validity, Wikipedia is an excellent example of a huge hyper-linked corpus of textual descriptions in the public domain (Wikipedia 2008b; Medelyan et al. 2008). While the debate is still being fought on its validity as a reference resource, I think due to its surge in popularity over the past few years the general public have made up their mind. For the purposes of this thesis I analyse Wikipedia's links and meta-data, for which the accuracy is more than sufficient (Weaver et al. 2006).

This thesis contributes to the already a substantial body of work covering both GIR and Wikipedia.

²³<http://mpi-int.mpg.de/~suchanek/downloads/yago/>

²⁴<http://www.powerset.com/>

Chapter 3

Evaluation and metrics

3.1 Introduction

This chapter outlines the evaluation frameworks and measures used in this thesis. The experiments performed cross the divide between retrieval and classification, and as such use a combination of evaluation measures. This chapter starts by describing the standard IR evaluation framework and the corpora used, this is followed by a description of the evaluation measures used and statistical tests performed.

3.2 Evaluating IR systems

The experimental evaluation of IR systems is a subject that has received a lot of attention over the past 40 years. IR systems are inherently designed to fulfil a user's information need; testing how well this subjective judgement has been fulfilled is not an easy task. Cleverdon et al. (1966) proposed the Cranfield methodology and six measurements. The Cranfield methodology involves a standard triple of a corpus, queries and relevance judgements (C, Q, R) to be provided allowing different IR systems to be compared. The corpus C is a collection of documents, the queries Q a set of requests for information, and the relevance judgements R a set of documents from the collection that fulfil each information request. Cleverdon et al.'s measurable quantities are: coverage of the collection; time lag between the search request and answer; form of presentation; user effort to fulfil their information need; recall (proportion of relevant material that is retrieved); and precision (proportion of retrieved material that is relevant). van Rijsbergen (1979) identifies the first four quantities as easily measurable; precision and recall measure the *effectiveness* of a system and are discussed further below.

3.2.1 Evaluation forums

Evaluation forums are now becoming the accepted method of evaluating IR systems. The Text REtrieval Conference (TREC) laid the foundation for modern evaluation forums. All of the current evaluation forums follow the Cranfield model providing a corpus and set of queries. The relevance judgements are generally not provided until after every group participating in the forum have submitted judgements (Agosti et al. 2006). Pooling was first used as a method of generating relevance judgements by Harman (1992); pooling is where a subset of documents returned by all the IR systems being evaluated are assessed with respect to the query by experts, rather than assessing the whole corpus.

Since TREC began in 1992 a series of other evaluation forums have started, most notably: the NII-NACISIS Test Collection for IR systems (NTCIR) workshop in 1999, the Cross Language Evaluation Forum (CLEF) in 2000, and the INitiative for the Evaluation of XML retrieval (INEX) in 2001. In 2003 the TREC Video track became its own independent workshop, TRECVideo. Evaluation forums are generally split into a series of tasks or tracks (Agosti et al. 2006). Tracks that stop producing interesting results are discarded, while new tasks are added for emerging areas (Mishne and de Rijke 2006).

Word sense disambiguation has a similar culture of evaluation. In 1998 the SenseEval¹ series of workshops began concentrating on annotating words with their semantic senses, and more recently with SemEval, semantic relationships. The Special Interest Group on Natural Language Learning's conference, CoNLL², began an evaluation task in 1999 evaluating often specialist tasks of natural language processing (NLP). Similar to the IR evaluations, WSD and NLP evaluations tend to be split into a number of tasks across a number of tracks. The first notable corpus used for word sense disambiguation was the Semantic Concordance (SemCor), constructed by Miller et al. (1993).

The track of greatest relevance to this thesis is the GeoCLEF track at the CLEF forum, which is specifically designed for the evaluation of GIR systems. Other tracks that have been proposed for GIR evaluations include the TREC Robust track (MacFarlane 2006), the TREC Novelty Track (van Kreveld et al. 2004), and the ImageCLEFphoto track (Clough et al. 2006b).

The ImageCLEF photo track is of particular interest to Geographic Retrieval because it is inherently geographic in nature. Since 2006 ImageCLEFphoto has used the IAPR-TC12 corpus containing 20,000 colour photos with associated meta-data supplied by Viventura, a holiday company (Clough et al. 2006b). As the images are all travel photos the locations that the pictures were taken in is integral to these multimedia documents. There are 60 queries with relevance judgements including 24 queries with geographic constraints. The reason I have decided not to use this collection to evaluate placename disambiguation is because this thesis focuses on placename disambiguation based on textual context, and the context provided by these documents is too small (Overell et al. 2008a).

In an ideal situation one would test each component of a system against a manually constructed standardised ground truth and the whole system in a formal IR evaluation setting. Unfortunately, for most of the experiments conducted in this thesis such ground truth does not exist. In Chapters 4 and 5, I construct my own ground truth by manually annotating Wikipedia articles. Leidner (2004a) and Clough and Sanderson (2004) recognised that a uniform ground truth was needed to compare placename disambiguation systems; currently, one does not exist. Chapter 6 describes how I construct such a ground truth using other annotated corpora. Formal evaluation of the whole system is performed on the GeoCLEF corpus in Chapter 6.

GeoCLEF

GeoCLEF is the Geographic track at the CLEF forum for comparing IR systems augmented with geographic data. It is becoming the de facto standard for evaluating GIR systems. The GeoCLEF 2005-08 English corpus consists of approximately 135,000 news articles, taken from the 1995 Glasgow Herald and the 1994 Los Angeles Times (Gey et al. 2006). The total corpus contains approximately 100M words.

There are 100 GeoCLEF queries from 2005-08 (25 from each year). These topics are generated by hand by the four organising groups. Each query is provided with a title, description and narrative. The title and description contain brief details of the query, while the narrative contains a more detailed

¹<http://www.senseval.org/>

²<http://ifarm.nl/signll/conll/>

	Correct / relevant	Incorrect / irrelevant
Classified as correct / retrieved	true positive (TP)	false positive (FP)
Classified as incorrect / not retrieved	false negative (FN)	true negative (TN)

Table 3.1: Contingency table

description including relevance criteria (Gey et al. 2006). The 2005 queries have additional fields for concept, spatial relation and location. However these fields were discarded in later years as unrealistic and as such are not used in this thesis. Classification of the GeoCLEF queries is discussed in Appendix B.

SemCorr and WordNet

SemCor took the previously existing Brown Corpus, a general text collection constructed in the 1960s containing 500 documents and totalling approximately 1M words, and mapped every word to the corresponding semantic definition (synset) in WordNet (Francis and Kucera 1979; Princeton University 2008).

WordNet is a publicly available English lexicon. 155,000 words (lemmas) are mapped to 118,000 synsets (a many-to-many mapping), each synset representing a distinct concept. Synsets are split into 45 semantic categories. Semantic categories are classified further by part-of-speech into adjective, adverb, verb and noun classes (Fellbaum 1998). WordNet also contains extensive information on the relations between synsets including antonym, hyponym, instance etc.

The 25 WordNet noun semantic categories are used as classification classes for Wikipedia articles in Chapter 4. Instances of the location semantic category are disambiguated as corresponding locations in the TGN in Chapter 6, turning SemCor into a geographically tagged corpus. This is the same task approached by Buscaldi and Rosso (2008c), who have released a mapping of WordNet synsets to geographic co-ordinates: GeoWordNet.

3.3 Evaluation measures

Early evaluation measures for retrieval and classification were based on the contingency table, see Table 3.1 (van Rijsbergen 1979). These measures were based on binary classification or unordered retrieval. Contingency table based methods can be applied to ranked retrieval by fixing a document cut-off value (DCV). This forces every method to return the same number of documents and treats every document above the DCV equally. Hull (1993) criticises using a single DCV method and suggests multiple values should be considered.

This section begins by examining contingency table based methods of evaluation commonly used in classification, followed by score based methods of evaluation preferred for retrieval experiments.

3.3.1 Binary classification and unordered retrieval

A contingency table gives an overall picture of results but is generally broken down further into measures. Cleverdon et al. (1966)’s measures of precision and recall attempt to capture the effectiveness of a retrieval system. Both measures rely on the assumption of relevance (detailed in Section 2.2.1) — that there is a binary measure as to whether documents are relevant or not. Precision is the proportion of retrieved documents that are relevant. Recall is the proportion of relevant documents that are retrieved:

$$\text{precision} = \frac{TP}{TP + FP} \quad (3.1)$$

$$\text{recall} = \frac{TP}{TP + FN} \quad (3.2)$$

Generally, both precision and recall have to be taken into account as there is a trade off between the two. If the threshold at which documents are considered relevant is increased, fewer documents will be retrieved, precision is expected to rise and recall expected to fall. Conversely, if the threshold at which documents are considered relevant is decreased, more documents will be retrieved, recall will rise and precision fall.

There are many ways to combine precision and recall into a single measure that allows the comparison of different IR systems. Dependent on the task and evaluation different measures are viewed as more appropriate. Commonly in retrieval tasks a ranked list will be returned as a result of each query, in which case a *ranked-retrieval* measure may be more appropriate (discussed in the next section). However when the retrieved set is not ranked it is common to use the F-measure.

The F-measure is the weighted harmonic mean of precision and recall. Traditionally, the F_1 measure is used where both are equally weighted. This is calculated as

$$F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \quad (3.3)$$

Generally in retrieval evaluations, measures are chosen that put more emphasis on the retrieval and relevant sets of documents, while in classification-evaluations equal emphasis is put on the objects classified both correctly and incorrectly. The most common measures for binary classification tasks are accuracy, the proportion of correctly classified documents, and its complement error-rate:

$$\text{accuracy} = \frac{TP + TN}{TP + FP + FN + TN}, \quad (3.4)$$

$$\text{error-rate} = \frac{FP + FN}{TP + FP + FN + TN}, \quad (3.5)$$

$$\text{accuracy} + \text{error-rate} = 1. \quad (3.6)$$

Note there is a grey area between retrieval and classification, and it is often unclear whether to put more emphasis on the classified set or the whole corpus. Of course, unless methods are pooled or the whole corpus is evaluated, only the retrieved/classified set of documents can be evaluated. The accuracy of this set is equal to the overall precision.

3.3.2 Scored classification and ranked retrieval

It is common for IR systems to return a ranked list of results rather than assigning documents as relevant or not relevant. This is what users have come to expect from search engines (such as Google or Yahoo). Intuitively it makes sense that some documents will be more relevant to an information need than others.

Precision at n ($P@N$) is a measure that models how a system is used. It can be assumed that in a real system a user will not trawl through page after page of results looking for a relevant document (iProspect 2006). It is assumed a user will only look at the first n documents (where n is 5, 10, 20 etc...); the precision is calculated after the first n documents. $P@N$ is limited as a comparator, as it varies significantly with the number of relevant documents and the selected value of n .

Average precision (AP) is a measure that attempts not to penalise systems for setting the relevance-required-for-acceptance threshold too high or too low. It relies on systems being able to quantify relevance and rank documents by their relevance. AP is the average of precisions computed at each relevant document rank:

$$AP = \frac{\sum_{i=1}^R \text{Precision}(r_i)}{R}, \quad (3.7)$$

where R is the number of relevant documents, and r_i is the rank of the i th relevant document. $\text{Precision}(r_i)$ is the precision at rank r_i and the precision of not-retrieved documents is set to zero. Average precision can be viewed as the P@N value averaged across the ranks of the relevant documents. For a more detailed description of AP cf. Voorhees and Harman (1999).

F-measure, P@N and AP all provide a single per-query effectiveness value of an IR system. However, it is common in evaluation forums to represent the effectiveness of a system executed across all queries with a single number (making comparing systems as easy as possible). The arithmetic mean is the most common method of combining per-query results. The arithmetic mean of the average precision (short *mean average precision* or MAP) is the major evaluation measure for IR systems that produce a ranked set of results. Critics argue the geometric mean of the average precision (short *geometric average precision* or GMAP) is a more appropriate measure as it biases against systems with high variability (Voorhees 2005). This reflects users' preference for systems with consistent performance. In practice systems with a high MAP are likely to have a high GMAP.

Evaluation forums are a driving force behind the development of IR systems. The `trec_eval` software developed for the TREC evaluation processes sets of relevance judgements and ranked results to provide the de facto evaluation measures with MAP being the most commonly quoted (Voorhees and Harman 1999).

Classification has an equivalent measure to AP designed not to penalise systems that set their classification threshold too low: the *Equal Error Rate* (EER) is the error rate at the point where $FP = FN$ and is found by varying the acceptance score. Commonly the arithmetic mean of the EER for a system will be reported.

3.4 Statistical testing

Statistical testing is required in empirical evaluations to test the probability that the observed results could have occurred by chance. By calculating a test statistic it is possible to interpret how significant the results are. In classification it is common to perform per-object significance testing, while in retrieval it is more common to look at per-query results. These are quite different scenarios but involve similar data and similar assumptions.

Hull (1993) provides a summary of statistical tests applicable to IR with their relevant benefits and assumptions. The Sign test, Wilcoxon Signed-Rank test and the Student's t-test are described for comparing two IR systems. The Student's t-test is the most powerful of these tests, however it is a parametric test assuming a normal distribution. The Wilcoxon Signed-Rank test is less restrictive assuming only a continuous distribution, while the Sign test makes no specific assumptions. van Rijsbergen (1979) argues the only valid statistical test in IR is the Sign test as precision and recall are discrete measures. Hull argues that as the sample size increases, although strictly speaking the t-test's assumptions are not met, it provides a useful approximation for computing a test statistic. When interpreting results in these circumstances, the fact that the test statistic is only an approximation must be considered, for example

the test statistic can be compared to a lower α value.

Hull (1993) proposes that the Student’s t-test can be applied to IR despite its assumptions not holding. I agree with this premise, but given the relatively small sample sizes of queries used in the collections in this thesis (see Section 3.2.1), I prefer to err on the side of caution and will use the Wilcoxon Signed-Rank test when comparing two retrieval methods. Intuitively I find it hard to justify the use of parametric tests in retrieval experiments. When testing classification experiments for significance a rank will not be produced, so the Sign test will be the most appropriate test. The Student’s t-test is only applied in Chapter 7, where large samples of data approaching a continuous distribution are compared. The application of all these tests are described below.

While the Wilcoxon Signed-Rank test is applicable when comparing two ranks, it is not appropriate for comparing multiple ranks. This is because as the volume of pairwise comparisons grow the probability of achieving a significant result increases. Hull (1993) suggests two tests in these circumstances: the ANOVA test and the Friedman test. The ANOVA test is parametric and more powerful. In contrast, the Friedman test is non-parametric and has more relaxed assumptions. In this thesis, tests across multiple comparisons only occur where the Wilcoxon Signed-Rank test’s assumptions are met; because of this, I shall use the corresponding test for multiple comparisons: the Friedman test.

3.4.1 The Sign test

van Rijsbergen (1979) describes the Sign test as a statistical test with few assumptions. In fact it makes no assumptions about the form of the data distribution. The calculation of the test statistic is as follows (Hull 1993):

$$T = \frac{2 \sum I[D_i > 0] - n}{\sqrt{n}}, \quad (3.8)$$

where D_i is defined as $Y_i - X_i$, and X_i and Y_i are the scores or classifications of methods X and Y for query i . $I[D_i > 0]$ is 1 if $D_i > 0$, 0 otherwise. The null hypothesis of the Sign test is that X will perform better than Y the same number of times as X will perform worse than Y .

3.4.2 The Wilcoxon Signed-Rank test

The Wilcoxon Signed-Rank test (*a.k.a.* the Wilcoxon matched pairs test) replaces the difference, D_i , between Y_i and X_i with its absolute rank, defined as $\text{rank}|D_i|$. The ranks of D_i where the sign of D is negative are summed (W^-) and the ranks of D_i where D is positive are summed (W^+). The test statistic is the minimum of these.

$$W^+ = \sum_{D_i > 0} \text{rank}|D_i| \quad (3.9)$$

$$W^- = \sum_{D_i < 0} \text{rank}|D_i| \quad (3.10)$$

$$T = \min(W^+, W^-) \quad (3.11)$$

The Wilcoxon Signed-Rank test assumes a symmetric distribution of D_i . van Rijsbergen (1979) argues that this rarely holds in a retrieval scenario, while Hull (1993) again argues this needs to be taken into account.

3.4.3 The Friedman test

The advantage of the Friedman test is that it tests the significance between multiple methods simultaneously, without the chance of finding a significant difference between tests increasing as the number of tests increase. The initial test statistic of the Friedman test, F_N , tests whether there is any significant difference between methods.

$$A = \sum_{i=1}^n \sum_{j=1}^m R_{ij}^2 \quad (3.12)$$

$$B = \frac{1}{n} \sum_{j=1}^m R_j^2 \quad (3.13)$$

where m methods are compared on n queries. R_{ij} is the rank of method j with respect to the other methods for query i . R_j is defined as follows:

$$R_j = \sum_i R_{ij}. \quad (3.14)$$

Finally the test statistic is defined as:

$$F_N = \frac{(n-1)[B - nm(m+1)^2/4]}{A - B}. \quad (3.15)$$

The assumption of the Friedman test is that errors are independent. The null hypothesis is that errors follow an F distribution with $(m-1)$ and $(n-1)(m-1)$ degrees of freedom.

If the Friedman test rejects the null hypothesis, the methods with a significant difference between them can be found by comparing for each methods k and l , the absolute difference between R_k and R_l , to the Friedman multiple comparisons test statistic: If

$$|R_k - R_l| < t_{1-\alpha/2} \left(\frac{2n(A-B)}{(n-1)(m-1)} \right)^{\frac{1}{2}} \quad (3.16)$$

then there is a significant difference between methods k and l at confidence α . Here $t_{1-\alpha/2}$ is the corresponding t statistic for $(n-1)(m-1)$ degrees of freedom at a confidence of $1 - \alpha/2$.

3.4.4 The Student's t-test

The Student's t-test is a paired test similar to the Wilcoxon Signed-Rank test, however, with the assumption that errors are normally distributed. The null hypothesis assumes a normal distribution with $(n-1)$ degrees of freedom. The test statistic t is compared to the Student's t-distribution and is defined thus (Hull 1993):

$$t = \frac{\bar{D}}{s(D_i)/\sqrt{n}} \quad (3.17)$$

where \bar{D} is the average D_i value across i and $s(D_i)$ is defined as follows:

$$s(D_i) = \sqrt{\frac{1}{n-1} \sum_i (D_i - \bar{D})^2} \quad (3.18)$$

The t-test is a parametric test assuming a normal distribution. This only holds for discrete data when large sample sizes are considered.

3.4.5 One-tailed vs. two-tailed

When comparing systems A and B a null hypothesis is constructed that both systems are the same and one tests an alternative hypothesis that A is better than B or that A is different from B . Testing if A is better than B is referred to as a one-tailed test, as it only considers one end of the probability distribution. Testing if A is different from B is referred to a two tailed test, as it looks at the probability distribution where A is significantly worse than B or where A is significantly better. A one-tailed test is appropriate in situations where one only cares about the outcome where A is better than B , i.e. situations where A is the same as B or worse than B , the same conclusions are drawn. In all other situations a two-tailed test is more appropriate.

In this thesis I use one-tailed tests when comparing simple approaches to complex approaches as one would generally only use a complex method if it gave superior results. When comparing two systems of similar complexity I use two-tailed tests. Unless otherwise stated, for two-tailed tests, I use an α value of 5% to compare to the test statistic. A common criticism of one-tailed tests is that they are less discriminative than two tailed tests. Because of this, for one tailed tests I use an α value of 2.5% unless otherwise stated.

3.5 Discussion

I conclude this chapter discussing three questions: What is the difference between information retrieval and classification? Are parametric tests applicable to retrieval experiments? Can evaluation metrics capture how useful a system is?

Classification and retrieval are two different problems that can both be solved with similar tools. Classification is concerned with annotating or labelling documents, while retrieval is concerned with finding documents that fulfil an information need. Classification problems can be rephrased as retrieval problems and vice-versa. Take the classification problem “Classify a set of documents into documents about London and documents not about London,” this could be rephrased as the retrieval problem “Find me documents relevant to London.” This is why retrieval puts more emphasis on the returned set of documents (because this is the set a user would see), while classification is concerned with all documents (because classifying documents as being not about London is an equal part of the defined problem). The line gets even fuzzier when problems are phrased as requiring positive classification only, however this is beyond the scope of this thesis.

The question of whether parametric tests are appropriate for retrieval is one that has surfaced repeatedly in this chapter’s discussion of significance testing. van Rijsbergen (1979) criticises the use of the Student’s t-test where its assumptions are not fulfilled. As the sample size of discrete data increases it starts to approach a continuous distribution and can be approximated by one. Hull (1993) suggests inspecting quantile plots of distributions for outliers, and skewness is a suitable test to check if the assumptions are approximately fulfilled. This seems an appropriate method for checking that a distribution can be approximated but relies on a large sample size. The fact that the retrieval experiments in this thesis use the GeoCLEF collection that has a relatively small sample of queries is the reason why I use non-parametric tests.

The final question considered in this chapter is whether evaluation metrics can capture the usefulness of a system. van Rijsbergen (1979) gives the purpose of evaluation as to provide data to a user that allows them to decide whether they want a system, and whether it is worth the cost (for some definition of cost e.g. time, money etc). In practice this is not what performance measures are used for: As Keen

(1992) observes, the primary use of performance measures is to put an ordering across retrieval systems. This raises the question what it means for one system to be better than another. Statistical difference has already been discussed, but Keen suggests systems should also provide a practical difference to the user. Forsyth (2001) criticises current evaluation measures and techniques saying user needs should be the primary concern and the user should be brought into the evaluation loop. Voorhees (2005) criticise methods such as MAP as they are dominated by better performing topics. They argue systems should always provide at least passable results as users only see the results to their queries rather than the average performance. I understand the limitations of the current evaluation framework; however, as long as these limitations are considered during evaluation I am satisfied that they are appropriate. For example, consider a GIR system x , that performs statistically significantly better than any other system on the GeoCLEF collection, one should not interpret from these results that x is better in every situation or that a user will notice the improvements provided by x . It simply shows that on a news corpus with constructed geographic queries x on average provides *some* improvement. This defines the scope of my evaluation.

Chapter 4

Classifying Wikipedia articles

4.1 Introduction

The first body-of-work chapter of this thesis approaches the task of classifying Wikipedia articles. The motivation behind this task is twofold:

- to augment Wikipedia with machine readable meta-data, providing greater browsing and inference ability, and linking to other data sources; and
- to use Wikipedia as an augmented corpus in machine learning tasks.

It is this second motivation that falls within the scope of this thesis. This chapter begins by examining the current work on classifying and disambiguating Wikipedia articles. I then present my own article classification system, *ClassTag*, and compare it to the state of the art. The chapter concludes with a case study where *ClassTag* is extended to classify tags assigned to photos from the popular photo sharing web site, Flickr.

4.2 Classification and disambiguation of Wikipedia articles

I will begin by more formally defining the subtle difference between classification and disambiguation when referred to in this thesis. There is confusion in the literature between these terms, and in some fields such as tag or term classification they are use almost interchangeable. I will resolve this confusion for the purposes of this thesis. In a classification problem a single classifier is built to classify all objects. While disambiguation is a two step process, where a separate classifier is built for every super-class of objects. Consider the problem of matching Wikipedia articles to classes. Approaching as a classification problem one could construct a classifier that given the content and meta-data of a Wikipedia article would classify which type of entity that article is most likely to be describing. On the other hand, approaching this as a disambiguation problem, one would create a separate classifier for each type of article based on simple selection criteria derived from prior knowledge. For example, approaching as a classification problem, one could construct a single classifier *classifying* whether articles describe animal species (cat, dog, monkey, etc.) or people with specific jobs (composer, actor, scientist, etc.). Alternatively, as a disambiguation problem, the selection criteria could be whether the article has a template listing a date of birth or taxonomic data, identifying an article as describing a person or animal respectively. Further classifiers

could *disambiguate* the occupation of the people or species of animal. Classification is commonly used to resolve semantic ambiguity, while disambiguation commonly resolves referent ambiguity.

Before I discuss the task of classifying Wikipedia articles, I will give an overview of a super-task: categorising Wikipedia anchor texts. This task is of particular interest to tag and placename classification, which will be tackled later in the thesis. There are two mappings which are necessary when categorising Wikipedia anchor texts:

1. Anchor text \rightarrow Wikipedia article, and
2. Wikipedia article \rightarrow Category.

The task of mapping an anchor text to a Wikipedia article is studied in several papers. Generally a model of how articles are referred to by specific anchor texts is built from Wikipedia, this model can then be applied to classify entities in an external corpus. This method, taking the links in Wikipedia as ground truth, is proposed by Mihalcea (2007) for word sense disambiguation. The accuracy of the links and relational statements in Wikipedia are quantified in a study by Weaver et al. (2006). They measure the accuracy of internal links in Wikipedia as 99.5% and the accuracy of relational statements as 97.2%. This method assumes Wikipedia is representative of the external corpus.

Bunescu and Paşca (2006) approach this task by learning the textual context of specific categories using a Support Vector Machine. The 24 words preceding and following each anchor text makes up that anchor's context. A mapping is learnt from these 55 word windows to the categories of articles. For example the word “conducted” appearing in an entity's context would provide evidence for the Wikipedia article of the entity being in the category “Composers”. They found substantial improvement taking advantage of the Wikipedia category tree structure over textual features alone. However, their system is not scalable to the whole of Wikipedia due to the volume of features used. Cucerzan (2007) presents an approach that scales to the whole of Wikipedia. They reduce the amount of contextual information extracted from text by only using links occurring in the first paragraph of a Wikipedia article where a reciprocal link is contained in the target page. Contexts are represented in a vector space and compared to ambiguous entities using the scalar product. Their system disambiguates multiple entities by simultaneously maximising their category agreement and contextual similarity. The sparsity of category data is partially solved by using Wikipedia list pages to add additional categories.

Returning to the sub task of classifying Wikipedia articles. Various ontologies and gazetteers have been used to provide a set of possible classifications making comparisons between techniques difficult. We consider the WordNet noun syntactic categories as our classification scheme.

Overell and Rürger (2007) disregard textual context altogether, instead using only the categories and templates of an article for classification. They are concerned only with Wikipedia articles describing locations; entities in the Getty Thesaurus of Geographic Names form their classification classes. They use a series of heuristics to gather evidence supporting a mapping from an article to a location. Buscaldi et al. (2006) also attempt to classify whether a Wikipedia article describes a location. They use Wikipedia as a filter for geographic terms, classifying simply whether an article refers to a location or not. They extract a set of geographic trigger words from WordNet and compare this set to the text of Wikipedia articles using Dice's coefficient. Any article greater than a set threshold is classified as a location. Pu et al. (2008) use a similar method to classify Wikipedia articles as describing locations. They consider the term “coordinates” with digits nearby a trigger. Locations assigned the same name within 3km of each other are resolved to the same entity.

Ruiz-Casado et al. (2005) were the first to map Wikipedia articles to WordNet synsets. Mapping Wikipedia articles to WordNet semantic categories (the focus of this chapter), can be seen as a sub task

of this. They map Wikipedia articles in the Simple English Wikipedia¹ to WordNet lemmas based on string matching the subject of the article. When there is only one lemma for a synset no disambiguation is necessary. However when multiple senses exist, an extended glossary entry for each potential synset is constructed. A synset's extended glossary entry is the original glossary extended with synonyms and hypernyms. These extended glossaries are then mapped into a vector space with tf-idf term weights. The Wikipedia article is mapped into the same feature space and disambiguated as the most similar sense with respect to the dot product of the vectors. In the case of ties, the extended glossary entries are iteratively increased. This method is similar to that presented by Buscaldi et al. (2006): both build a bag-of-words from WordNet for each classification class, which is expected to be similar to the corresponding Wikipedia article.

Cardoso et al. (2008) present their system, *Rembrandt*, for recognising named entities using a knowledge base derived from Wikipedia. In constructing their knowledge base, Rembrandt classifies Wikipedia articles using categories from the *HAREM* ontology as classification classes (Seco et al. 2006). Article categories are used to perform categorisation. Cardoso et al. split the HAREM ontology into a further three parts: categories implying explicit geographic knowledge (locations), categories implying implicit geographic knowledge (entities associated with locations), and categories implying no geographic evidence.

Suchanek et al. (2007) present a method of recognising Wikipedia articles describing entities (referred to as individuals) and relationships between them using the YAGO ontology. YAGO is a semantic knowledge base comprising of entities: an *is-a* taxonomic hierarchy and non-taxonomic relationships (Suchanek et al. 2007). Wikipedia categories are split into administrative, relational, thematic and conceptual classes. The class of a category is identified by parsing its name. Articles in conceptual categories are considered entities, the *type* of the entity is extracted from the conceptual category. WordNet synsets are also mapped into the YAGO ontology with *hypernym* relationships mapping to *subclassof* relationships. The *subclassof* hierarchy is further expanded to include Wikipedia categories using heuristic processing of the titles. The YAGO classifications are not directly comparable with the categorisations made in this chapter because it does not enforce the same strict hierarchy as WordNet. Mika et al. (2008) use a Hidden Markov Model to annotate and classify entities in Wikipedia as classes from the Wall Street Journal Penn Treebank. They use the output of a part-of-speech tagger and named entity recogniser as their features.

The most extensive mapping of Wikipedia articles to WordNet synsets has been built by DBpedia and released under the GNU Free Documentation License (DBpedia 2008). DBpedia stores the structured information contained in Wikipedia templates and uses it as a knowledge base (Auer and Lehmann 2007). The entities and relations are stored in an RDF format and linked to other external knowledge sources such as geographic gazetteers and US census data. Their mapping of Wikipedia articles to WordNet synsets was generated by manually associating individual synsets with specific templates.

4.3 Classifying Wikipedia articles

This chapter aims to build a generic and scalable system to classify Wikipedia articles. It must be generic so that later versions of Wikipedia can be included and full advantage can be taken of new data, also it must be fully applicable to versions of Wikipedia provided in languages other than English and additional open-content resources. Scalability is important because our motivating aim is to maximise coverage. To

¹The Simple English Wikipedia is a version of Wikipedia using simple words and short sentences. It is aimed at all English speakers including children and people learning English: <http://simple.wikipedia.org>.

do this the whole of Wikipedia needs to be processed and updated versions of Wikipedia periodically need to be included. Because of these requirements, a full semantic interpretation of Wikipedia is avoided. Experiments by Buscaldi et al. (2006) have shown Wikipedia articles are too heterogeneous to take advantage of shallow textual features, and Bunescu and Paşca (2006) show representing the context of every link is difficult to scale.

This chapter follows an approach similar to Overell and R ger (2007) and Suchanek et al. (2007), using only Wikipedia article meta-data, specifically the structural patterns of categories and templates. The approach differs from these by using a supervised classifier rather than a set of constructed heuristic rules. This is because we want a scalable approach that will be compatible with future versions of Wikipedia and alternate resources. Articles form the objects, WordNet noun semantic categories form classification classes, and Wikipedia categories and templates form features. The classifier used is a Support Vector Machine (SVM). A SVM is a supervised learning method designed to partition a multi-dimensional space with a linear classifier that is iteratively calculated by the optimisation of a quadratic equation (Joachims 1999). The complexity of the problem is dependent on the number of training examples. The SVM^{light} package has been chosen for learning and classification² (cf. Joachims (1999)). A binary SVM classifier is trained for each class. Each article is classified by each classifier and assigned to the class of the classifier outputting the highest confidence value.

4.3.1 Ground truth

I use the WordNet corpus as a ground truth to train the classifier mapping Wikipedia articles to WordNet semantic categories. WordNet lemmas are matched to Wikipedia article titles and re-directs. The Wikipedia articles are assigned to the class of the matched words. When multiple senses exist for a word, the class of the highest ranked sense is taken. For example the WordNet lemma *Manhattan* is classified as a location in WordNet and is matched to the corresponding Wikipedia article titled *Manhattan*.

The ground truth is formed of all the Wikipedia articles where the titles match WordNet nouns. For each WordNet semantic category the ground truth is partitioned into a training and test set. The test set is made up of 100 articles from each category (or 10% of the articles from a category where less than 1000 examples exist). The final ground truth consists of 63,664 Wikipedia articles matched to WordNet lemmas, 932 of which are partitioned as a test set.

4.3.2 Sparsity of data

With respect to data sparsity, two problems occur. First is WordNet categories that are under represented in the ground truth; second is articles that have very few features.

Under represented categories

The problem of data sparseness is first discussed in Section 2.3.2. Of the three solutions discussed, *smoothing* is the only method that would be appropriate as classes are already partitioned into high level categories (Ide and V ronis 1998). Even with smoothing we would still risk over fitting; because of this under represented classes are discarded.

There are 25 noun syntactic categories in WordNet (not including the “Top” noun category). Of these only 10 are represented with enough articles in Wikipedia matched to WordNet words to train an SVM that will not significantly over fit: act, animal, artifact, food, group, location, object, person, plant

²<http://svmlight.joachims.org/>

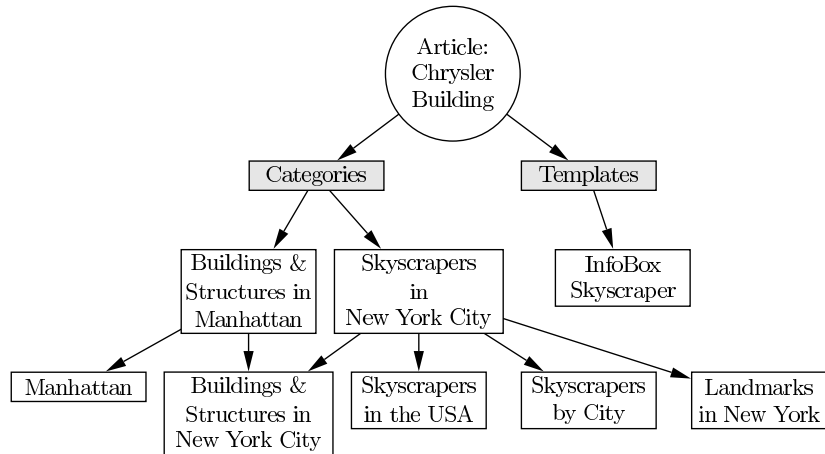


Figure 4.1: Example category and template network

and substance. The Time category can additionally be included by artificially adding to WordNet 457 days and years categorised as times. The 366 days of the year are added in numerical day, full month format (e.g. “01 November”) and 121 years in numerical format (from 1887 – 2007 inclusive).

Sparsity of features

There are a total of 39,516 templates and 167,583 categories in the dump of Wikipedia used in this chapter, the English language November 2006 dump. The majority of these categories or templates occur in less than 10 articles. The categories and templates that occur in more than 50 articles are selected to form our feature list, giving us the 25,000 most commonly occurring categories and templates. This is a small enough number of features to allow relatively fast learning and classification for a SVM.

Most articles in Wikipedia have very few categories and templates (in fact the majority of articles have no templates and only one category). Because of this sparsity of features, I have artificially increased the number of categories and templates each article contains. This is done using the category network and template transclusion. As explained in Section A.2, Wikipedia categories and templates are linked in a directed network. Categories are linked through a category tree, analogous to an ontology. Complex templates inherit (transclude) simpler templates in a similar tree structure. ClassTag navigates backwards through this network to increase the number of categories and templates each article has. The disadvantage of this method of enhancing the number of features is that additional features are not independent. For example consider the Wikipedia article *Chrysler Building* describing the art-deco skyscraper in New York City. Suppose we consider traversing two category arcs and one template arc. The article, *Chrysler Building*, is in categories: *Buildings and Structures in Manhattan* and *Skyscrapers in New York City*; and has one template: *InfoBox Skyscraper*. An additional category arc needs to be traversed adding the parent categories of *Buildings and Structures in Manhattan* and *Skyscrapers in New York City* as second level categories. These additional categories are *Manhattan*, *Buildings and structures in New York City*, *Skyscrapers in the USA*, *Skyscrapers by city* and *Landmarks in New York*. This tree is illustrated in Figure 4.1.

Experiments detailing the choice on how many levels to navigate in these graphs and the weighting function used to determine the scalar values of the features are detailed in Section 4.3.4.

4.3.3 Removing noise

A significant proportion of Wikipedia categories are actually related to Wikipedia administration rather than article content. These categories are identified by navigating every possible path through the category tree back to the root category node for each article. If every path for a category passes through the Wikipedia Administration category, that category is added to a *black list* of categories not considered as features. 12,271 categories were found through this method.

Similarly there exist templates that contain only page formatting information and contribute nothing to article content. These templates are identified by pruning all templates that occur in over 30,000 articles. 11 templates were identified with this method. This is analogous to stop word removal.

4.3.4 System optimisation

As explained in Sections 4.3.1 and 4.3.2, our ground truth consists of WordNet nouns matched to Wikipedia articles and our features for classification are 25,000 categories and templates. This ground truth is partitioned into training and test sets to select the optimum values for variables governing the feature weights. The variables optimised are:

- the number of arcs traversed in the category network;
- the number of arcs traversed in the template network;
- the choice of weighting function.

Between zero and four arcs were considered for both categories and templates. Taking category arcs as an example: zero category arcs means the article's categories are ignored, one category arc means an article's categories are included as features, two category arcs means the article's categories and the categories of the article's categories are included as features etc.

By traversing more arcs we increase the number of features a document contains. The scalar value of each feature is determined by a weighting function. The same weighting function is used by both category and template features. Three weighting functions are considered:

- **Term Frequency (tf):** The scalar value of each feature is the number of times it occurs for this article.
- **Term Frequency – Inverse Document Frequency (tf·idf):** The scalar value of each feature is the number of times it occurs for this article divided by the log of the number of times it occurs in the document collection.
- **Term Frequency – Inverse Layer (tf·il):** The scalar value of each feature is the number of times it occurs for this article divided by the number of arcs that had to be traversed in the category/template network to reach it.

Referring back to the *Chrysler Building* example in Figure 4.1, Table 4.1 shows how the scalar values of the features vary with the choice of weighting function. The *c* or *t* prefix specifies whether a feature is a category or a template. The features added by traversing an additional category arc are shown in *italics*. Notice how the problem of data sparsity has been reduced, as we have added an additional five features to a document that originally had only three.

Feature	tf	tf-idf	tf-il
c:Buildings and Structures in Manhattan	1	0.51	1
c:Skyscrapers in New York City	1	0.53	1
t:InfoBox Skyscraper	1	0.49	1
c: <i>Manhattan</i>	1	0.48	0.5
c: <i>Buildings and structures in New York City</i>	2	1.16	1
c: <i>Skyscrapers in the USA</i>	1	0.59	0.5
c: <i>Skyscrapers by city</i>	1	0.59	0.5
c: <i>Landmarks in New York</i>	1	0.60	0.5

Table 4.1: Weighting functions example

Variable	Value	Prec (%)	F ₁
Category Arcs	0	59.1	0.22
	1	87.1	0.694
	2	87.3	0.694
	3	87.0	0.696
	4	61.1	0.25
Template Arcs	0	86.7	0.695
	1	86.8	0.693
	2	86.9	0.696
	3	87.0	0.696
	4	87.0	0.696
Weighting Function	tf	86.7	0.623
	tf-idf	87.6	0.668
	tf-il	87.0	0.696

Table 4.2: Varying feature values

Selecting variables

An exhaustive search was performed of every combination of variables evaluated against the ground truth test set. The motivation for classifying Wikipedia articles is to build a huge training corpus to classify entities in an external corpus. As we will match entities to classified articles in Wikipedia, the volume of classifiable terms will be dependent on the number of classified Wikipedia articles. To classify as many entities as possible, we must maximise the recall of our article classifier. Conversely the accuracy of a classifier will only ever be as good as its training data.

To manage these competing aims I have decided to maximise the F₁-measure. As a classifier with a classification performance for a specific category below 80% is not useful, an additional caveat is added: only variable combinations producing a precision of more than 80% in each category are considered. The optimal results were achieved traversing three arcs for both categories and templates, and weighting function tf-il.

Table 4.2 shows how, with respect to the *best* method, varying the number of arcs traversed in the category and template networks, and changing the weighting function affects the precision and F₁-measure. Notice that there is in fact minimal difference in performance as template arcs and the weighting function vary. For categories, when no category data is used, the data is too sparse to perform much correct classification. Too sparse in this sense means many objects have few or no features. This is due to many articles having no templates and only one or two categories. Conversely when more than four category arcs are traversed the data becomes far too noisy. Noisy data is just as bad as sparse data and occurs when the additional features no longer add information that distinguishes between classes and instead makes it difficult or impossible to partition classes.

I conclude that the features chosen are fairly robust provided the value selected for the number of

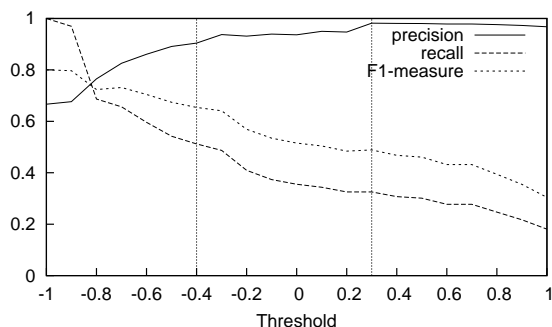


Figure 4.2: Threshold – F₁-Measure

category arcs traversed produces training data that is neither too sparse nor too noisy. Note despite being the *best* method, when compared with a two-tailed Sign test, it is not statistically significantly better than methods with slight variations in the selected parameters.

Selecting a threshold for classifying

The SVM binary classifiers output the values of their decision functions. The decision function output can be interpreted as the confidence with which an article is correctly classified as a member of a category. If there exists no prior knowledge about the distribution of the data one can simply classify articles as the category of the classifier that outputs the greatest positive value. If no classifiers output a positive value, one can consider the article unclassified. However if there exists prior knowledge about the data, for example if one knows a significant proportion of Wikipedia articles can be classified as one of our 11 categories, the threshold could be set lower than zero. On the other hand, if one has prior knowledge that the data is particularly noisy, the threshold could be set greater than zero.

A training experiment was performed where 250 Wikipedia articles were selected at random. Each article was classified as the WordNet semantic category of the classifier outputting the greatest decision function. An assessor then marked each classification as correct or incorrect by hand. The threshold for the minimum acceptable output value was then varied between -1 and 1. Articles where the maximum output value from a classifier were below the threshold were considered unclassified. Figure 4.2 shows how precision, recall and the F₁-measure vary with the threshold value. As this system’s motivation is to maximise the coverage of tags, the method that maximises the recall given a minimum acceptable precision is selected. I have selected the minimum acceptable precision across all categories as 90%; this gives a recall of 51% and a threshold value of -0.4. Alternatively the precision could be maximised instead of recall within an allowable precision range, this would give a threshold of 0.3, a precision of 98% and recall of 33%.

Figure 4.3 shows how varying the threshold affects the proportion of articles classified and the proportion of ambiguous articles (articles with multiple positive classifications). When the threshold is -0.4, 39% of all articles are classified. 5.7% of those classified are ambiguous. When the threshold is 0.3, 21% of all articles are classified, 0.5% of which are ambiguous.

4.4 Evaluation and comparison to existing methods

The following experiment compares the performance of the article classifier developed in this chapter, ClassTag, with the performance of the mapping of Wikipedia articles to WordNet synsets provided for

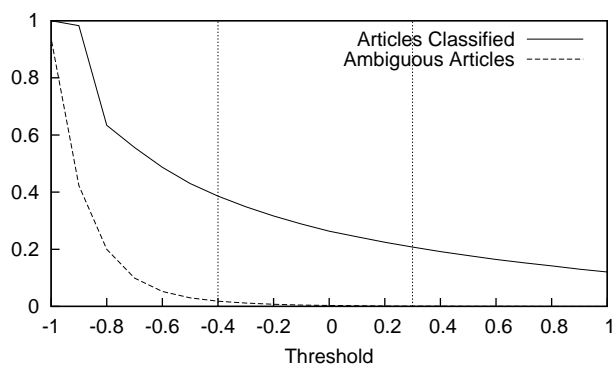


Figure 4.3: Threshold – Proportion of articles classified

download from DBpedia (2008).

4.4.1 Experimental setup

An evaluation set of 300 Wikipedia articles were selected at random from the union of articles classified by DBpedia and articles classified by ClassTag. ClassTag classifies a total of 664,770 Wikipedia articles. DBpedia classifies a total of 338,061 articles, however only the 206,623 articles that also exist in our November 2006 dump of Wikipedia are considered³. ClassTag classifies 258 of the articles in the evaluation set, while DBpedia classifies 88 articles. There is an overlap of 38 articles.

Two configurations of ClassTag were tested, the first referred to simply as ClassTag optimised for recall, and the second referred to as ClassTag⁺, optimised for precision. The difference between these two configurations is the acceptance threshold of the SVM decision function set to -0.4 and 0.3 respectively (cf. Section 4.3.4). By comparing both configurations to DBpedia, I plan to test the extremes of ClassTag’s performance. DBpedia’s classifications are optimised for precision so are directory comparable to ClassTag⁺. ClassTag⁺ classifies a total of 344,539 articles and 125 articles in the evaluation set.

Assessments

Three assessors assessed the Wikipedia articles. The assessors were information retrieval researchers familiar with both WordNet and Wikipedia. A randomly selected 50 articles were assessed by all assessors to measure assessor agreement. All remaining articles were only assessed by a single assessor. Assessments were performed blind. The assessors had no knowledge of which systems had classified the article or what the classifications were. The evaluation interface presented the user with the Wikipedia article that had been classified, a checkbox for each of the 25 semantic categories, and the semantic category brief descriptions taken from the WordNet web site (Princeton University 2008). Assessors were told to select all semantic categories they considered as correct classifications for each article.

Assessor agreement

Two values were measured for assessor agreement: *partial agreement* and *total agreement*. With partial agreement there exists an article classification that all assessors agree on. Total agreement is where assessors agree on all classifications. For **86%** of articles, assessors had partial agreement. For **78%** of articles, assessors had total agreement.

³The copy of DBpedia.org’s data used in this thesis is based on a dump of Wikipedia taken from July 2007. This means DBpedia’s dump of Wikipedia contains 8 months worth of edits missing from our dump.

	ClassTag	DBpedia	ClassTag ⁺
Prec. (%)	72	58	86
Recall (%)	81	17	38
Acc. (%)	62	16	36

Table 4.3: System evaluation results

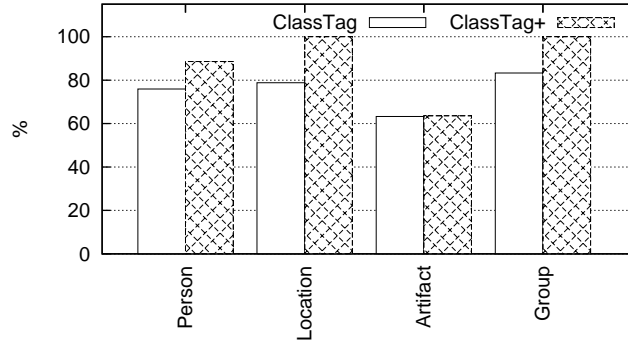


Figure 4.4: Per-category precision

4.4.2 Results

Table 4.3 shows the experimental results. In previous papers classifying Wikipedia articles only the accuracy of the classified set has been reported (Ruiz-Casado et al. 2005; Suchanek et al. 2007). As the sample set presented in this thesis is built from the pool of articles classified by both ClassTag and DBpedia, we can also consider articles not classified. In these circumstances precision can be considered the accuracy of the classified set.

An assessor was selected at random, and their assessments were considered ground truth for the Wikipedia articles with multiple judgements. As we consider a system classification correct if it matches *any* of assessor classifications, the gold standard accuracy can be considered equal to the assessor partial agreement: 86% (This is the point where the judgements provided by the system become as accurate as those provided by a human). ClassTag⁺ reaches the gold standard precision of 86% but at a significant recall trade off, classifying less than half as many articles as ClassTag. ClassTag has a particularly high recall of 81%.

Per category results

The top four most commonly occurring categories in the evaluation set were (in order): person, location, artifact and group. Figure 4.4 shows the per-category precision of ClassTag and ClassTag⁺. *Artifact* is noticeably worse than the other three categories (over 12% lower than the second lowest) with a precision of 63.3%. This difference is even more pronounced for ClassTag⁺ where the precision of the person, location and group categories significantly increases to between 89% and 100%, while the precision of the artifact category barely changes. I attribute the artifact category’s low precision to the huge variation in the types of artifacts in Wikipedia. WordNet defines an artifact as “nouns denoting man-made objects,” this ranges from a paper clip to the Empire State Building.

Summary

In Section 4.3 the goal of ClassTag is identified as to be a generic, scalable system that maximises recall while keeping as high a precision as possible. ClassTag classifies 39% of articles in Wikipedia with a

precision of 72%. The system is flexible enough that it can be optimised for precision, as demonstrated with ClassTag⁺. In the evaluation, both ClassTag and ClassTag⁺ outperformed DBpedia in all our performance measures. For example with respect to precision, ClassTag outperforms DBpedia by 14%, while ClassTag⁺ outperforms DBpedia by 30%.

4.5 A case study: Flickr

Photo corpora are of particular interest to GIR as the location a photograph is taken in is integral to its meaning. The following case study uses the classified corpus of Wikipedia pages generated in the first part of this Chapter to classify tags in Flickr.

The collaborative efforts of users participating in social media services such as Flickr⁴, YouTube⁵, Wikipedia⁶, and Del.icio.us⁷ have led to an explosion in user-generated content. A popular way of organising this content is through folksonomy-style tagging. The flexibility of such a tagging mechanism clearly addresses the user's need to index and navigate the large amount of information that is being generated. As a result, an uncontrolled vocabulary emerges that by far exceeds the semantics of a hierarchical ontology, taxonomy, or controlled vocabulary such as WordNet (Fellbaum 1998). At the same time, it imposes the problem of semantically categorising and exploring a potentially infinite tag space.

This case study addresses the task of classifying tags into semantic categories. Consider this problem in terms of an example. Figure 4.5⁸ shows a Flickr photo annotated with tags. The tags in this example have clear facets: they describe the subject of the photo, indicate where and when it was taken, and what camera was used (this is typical for a tagged photo in Flickr). This case study tackles the task of automatically classifying these tags in order to help users better understand the image annotations. Using WordNet, the tags *skyscraper*, *august* and *vacation* can be classified as representing respectively an artifact, time, and act (See Figure 4.6 (a)). The tags *chrysler building, nyc, 2006, olympus x200* and *william van alen* cannot be matched to WordNet lemmas and as far as WordNet knows their semantic category is thus unknown.

To overcome the limited coverage of WordNet, the ClassTag system is extended to classify tags using its classified set of Wikipedia articles. Flickr tags are mapped to Wikipedia articles using anchor texts in Wikipedia. Since we have classified Wikipedia articles we can thus categorise the Flickr tags using the same classification. For example the tag *nyc*, in Figure 4.5, may be mapped to the anchor text *NYC*. The most common target page for this anchor text is the Wikipedia article *New York City*. The classifier classifies the Wikipedia article *New York City* as a *location*. Consequently, one can argue that the Flickr tag *nyc* is referring to a *location*.

Figure 4.6 (b) illustrates how ClassTag can extend the coverage of WordNet. The tags *chrysler building, nyc, william van alen* and *2006* do not appear in WordNet, however they can be matched to Wikipedia anchor texts and can thus be classified by the system as an artifact, location, person and time, respectively. Finally, the tag *olympus x200* cannot be matched to either a WordNet lemma or a Wikipedia anchor text. Categorising this type of tag is not in the scope of this thesis, but could potentially be covered by incorporating different resources.

⁴<http://www.flickr.com/>

⁵<http://www.youtube.com>

⁶<http://www.wikipedia.org/>

⁷<http://del.icio.us/>

⁸Photo taken by author: <http://tinyurl.com/5cwtq7>



Tags
 chrysler building
 skyscraper
 nyc
 august
 2006
 vacation
 olympus x200
 william van alen

Figure 4.5: Example photo with user-defined tags, extracted from Flickr

<p><i>What:</i> skyscraper (artifact)</p> <p><i>When:</i> august (time) vacation (act)</p> <p><i>Unknown:</i> chrysler bulding nyc 2006 olympus x200 william van alen</p>	<p><i>Where:</i> nyc (location)</p> <p><i>What:</i> chrysler bulding (artifact) skyscraper (artifact) william van alen (person)</p> <p><i>When:</i> august (time) 2006 (time) vacation (act)</p> <p><i>Unknown:</i> olympus x200</p>
(a) WordNet classification	(b) ClassTag classification

Figure 4.6: Classification of the tags in Figure 4.5 using WordNet (a) and ClassTag (b)

This case study uses the corpus built by Sigurbjörnsson and van Zwol (2008): a snapshot of the Flickr database consisting of metadata from 52 million public photos uploaded between 2004 and 2007. The metadata was gathered using the Flickr API⁹.

There are two approaches to categorising tags: corpus-based approaches and knowledge-based approaches. Schmitz (2006) and Rattenbury et al. (2007) follow a corpus based approach, using information inferred from the corpus. Schmitz recognises that people should not have to choose between a hierarchical ontology or unrestricted tags and proposes a probabilistic unsupervised method for inferring an ontology from data. Their results are promising but leave room for improvement. Rattenbury et al. cluster tags from Flickr using temporal and spatial meta data, to assign event and place semantics. Their approach has a high precision, however a large proportion of tags remain unclassified.

This case study follows a knowledge-based approach building on the work of Sigurbjörnsson and van Zwol (2008). They map Flickr tags onto WordNet semantic categories using straightforward string matching between Flickr tags and WordNet lemmas. They found that 51.8% of the tags in Flickr can be assigned a semantic label using this mapping and that the most common semantic categories of Flickr tags were locations, artifacts, objects, people and groups. This case study takes their approach as a baseline and shows that it can be significantly improved upon. This section begins with a description of ClassTag's tag classification system and concludes with an analysis of the tags occurring in Flickr.

⁹<http://www.flickr.com/services/api/>

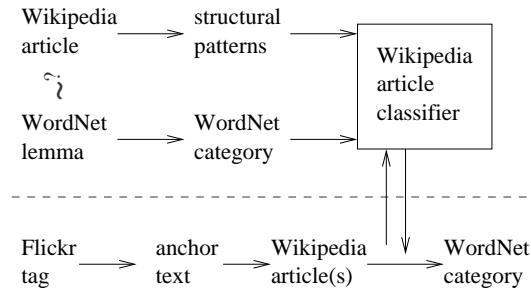


Figure 4.7: Overview of the ClassTag system

4.5.1 A tag classification system

An overview of the tag classification system implemented in ClassTag can be found in Figure 4.7. The system is comprised of two components:

1. A classifier for classifying Wikipedia articles using structural patterns as features and WordNet semantic categories as a classification scheme (top part of Figure 4.7 — described in the body of this chapter). $\sim?$ denotes the string matching of WordNet lemmas to Wikipedia article’s titles to form our ground truth.
2. A pipeline for mapping Flickr tags to WordNet semantic categories, using the classifier (lower part of Figure 4.7 — described below).

I will begin with a high level description of the tag classifier followed by a more detailed description and examples. Having classified Wikipedia articles ClassTag uses the classification results to classify Flickr tags. This is done using a simple pipeline of mappings. First Flickr tags are mapped to Wikipedia anchor texts. Next Wikipedia anchor texts are mapped to Wikipedia articles. This mapping is the same as described by Mihalcea (2007). The lower part of Figure 4.7 displays the steps taken by ClassTag when mapping a tag to a semantic category. The mapping consists of three steps:

1. Tag \rightarrow Anchor text,
2. Anchor text \rightarrow Wikipedia article,
3. Wikipedia article \rightarrow Category.

This is illustrated with an example in Figure 4.8 using the tags from Figure 4.5, the Chrysler Building example. There are four tags that are covered by Wikipedia but not by WordNet: “chrysler building,” “nyc,” “william van alen” and “2006”. “2006” is covered by our extension of the time category of WordNet, leaving “chrysler building,” “nyc” and “william van alen” to be categorised by ClassTag. In the following paragraphs I will demonstrate how the tags “chrysler building” and “nyc” are mapped to semantic categories. Two of the mappings are weighted. The weights on the Anchor text \rightarrow Wikipedia article arcs represent the frequency of the mapping (e.g. the number of times “NYC” refers to *New York City* in Wikipedia). The weights on the Wikipedia article \rightarrow Category arcs represent the output of the SVM decision function.

Mapping from tags to anchors is a straightforward string matching process (Bunescu and Paşca 2006; Cucerzan 2007; Mihalcea 2007; Overell and R uger 2007; Sigurbj rnsson and van Zwol 2008). Some ambiguity is introduced because tags are commonly lower case and often contain no white space or

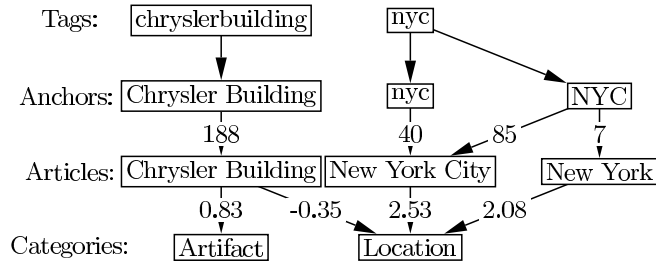


Figure 4.8: Tag \rightarrow Category example

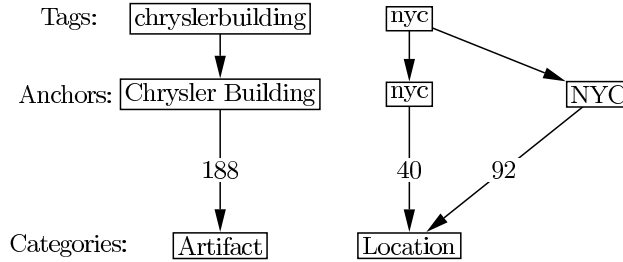


Figure 4.9: Tag \rightarrow Category example (reduced ambiguity)

punctuation. Mapping from Wikipedia articles to categories also introduces relatively little ambiguity. As detailed in Section 4.3.4, only 5.7% of classified articles result in multiple positive classifications. In these cases ClassTag simply classifies the article as the category corresponding to the classifier with the greatest confidence value.

Mapping from anchors to Wikipedia articles is more complex. To reduce the number of mappings considered, we remove *outlier* mappings. An outlier mapping is defined as a mapping from an anchor text to an article that occurs less than five times, or a mapping that makes up less than 5% of the total mappings from a specific anchor. After removing outliers the ambiguity is further reduced by grouping all articles in the same category together.

Continuing with the running example: the *Chrysler Building* article is disambiguated as an artifact, since that mapping has the largest weight. The “Chrysler Building” anchor is unambiguous, and can now map straight to artifact. The “NYC” anchor is ambiguous in Figure 4.8, but both articles map to location so can be combined into a single mapping. The resulting mapping is displayed in Figure 4.9. Observe that the tag “chrysler building” is disambiguated as an artifact and “nyc” a location. Table 4.4 shows the proportion of mappings at each stage that are ambiguous. Notice the reduction in ambiguity achieved by mapping to the category with greatest confidence and combining articles which map to the same category (shown as the Anchor text \rightarrow Category mapping in grey).

Mapping	Prop. Ambiguous (%)
Tag \rightarrow Anchor text	3.0
Anchor text \rightarrow Wikipedia article	13.4
Wikipedia article \rightarrow Category	5.7
Anchor text \rightarrow Category	4.0

Table 4.4: Ambiguity in mapping from tags to categories

	WordNet	ClassTag	Diff. (%)
Vocabulary	89,902	193,444	+115
Vocabulary (%)	2.4	5.2	
Full volume	106,215,397	130,049,982	+22
Full volume (%)	56.5	69.2	

Table 4.5: Coverage of Flickr tags, both in terms of vocabulary coverage and full volume coverage.

	WordNet	ClassTag	Diff. (%)
Act	4,445	8,694	96
Animal	6,480	9,248	43
Artifact	12,648	33,320	163
Food	2,748	3,665	33
Group	2,302	7,096	208
Location	4,035	30,444	654
Object	1,898	7,265	283
Person	15,719	61,696	292
Plant	7,394	7,421	0
Substance	2,342	2,903	24
Time	1,173	5,715	387

Table 4.6: Coverage of the Flickr vocabulary in terms of different semantic categories.

4.5.2 Coverage

At the beginning of this case study Sigurbjörnsson and van Zwol (2008)’s work was introduced as a baseline. They map from Flickr tags to WordNet semantic categories using string matching between Flickr tags and WordNet lemmas. This section will show the results of extending this baseline approach with the ClassTag system to improve the coverage of the semantic labelling¹⁰.

Table 4.5 shows the performance of the WordNet baseline approach and the extension using the ClassTag system, in terms of how many of Flickr tags they are able to classify. Using ClassTag to extend the WordNet baseline, coverage of the vocabulary is increased 115% – from 89,902 to 193,444 unique tags. Measured in terms of the full volume of tags – i.e., taking tag frequency into account – 69.2% of the Flickr tags are now classified. This is an improvement of 22% compared to the WordNet baseline.

Let us now look in more detail at the types of tags the ClassTag system can classify but were not classified by WordNet. Table 4.6 shows the coverage of Flickr tags in terms of different semantic categories. Notice that the ClassTag system improves coverage considerably for all types of tags, except plants. The largest absolute increase in coverage is for the Person category where coverage is increased by almost 46,000 unique tags. For the Location and Artifact categories the coverage is, respectively, increased by over 26,000 and 20,000 unique tags. Having a better coverage of locations, artifacts, and people is certainly useful for any system analysing multimedia annotations. As illustrated in Figure 4.6 the extended coverage of ClassTag enables us to give a more informed presentation of Flickr tags.

Let us now look at some examples of tags covered by ClassTag that were not covered by WordNet; Table 4.7 shows some examples of such tags. Notice the ClassTag system is able to add some frequently photographed artifacts and objects such as Notre Dame, London Eye, Lake Titicaca and Half Dome; as well as some less famous ones such as Hundertwasserhaus and Strokkur. The ClassTag system is able to classify abbreviations of popular locations such as NYC and Philly. Furthermore, some popular tourist locations are added, such as Phuket and Big Island, neither of which were covered by WordNet. Last but not least, ClassTag is able to classify correctly names of famous people such as Norman Foster and

¹⁰Here a slightly improved baseline is used to the one described in Sigurbjörnsson and van Zwol (2008). The improvement includes the categorisation of plural nouns.

Category	Examples
Act	Triathlon, geocaching, mountain biking, kendo.
Animal	Jack Russell Terrier, Australian Shepherd.
Artifact	Notre Dame, London Eye, Sagrada Familia, nikon, nokia, pentax, leica, wii, 4x4.
Food	BBQ, Churrasco, Japanese food, Ramen, Asado.
Group	Live8, G8, NBA, SIGGRAPH, Tate Modern.
Location	NYC, Philly, Phuket, Big Island, Nottingham.
Object	Blue Mountains, Point Reyes, Half Dome, Lake Titicaca, Jungfrau.
Person	Norman Foster, Ronaldinho, Britney Spears, Chris, Alex, Dave, Emily, Laura, Lisa, Jen.
Plant	Guadua, Chelsea Flowershow, red rose.
Substance	Wheatpaste, O2, biodiesel
Time	New Year's Eve, 4th of July, Valentine's day.

Table 4.7: Examples of tags covered by ClassTag but not covered by WordNet

Ronaldinho, as well as a large set of frequent first names such as Chris, Alex, Emily, Laura, etc.

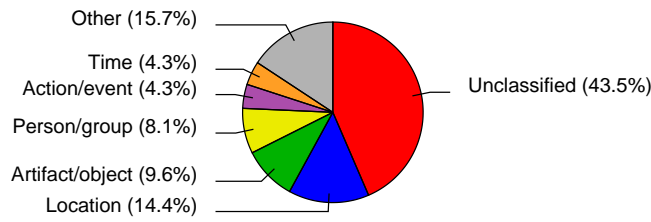
So far this case study has shown that the coverage of Flickr tag classification can be considerably extended using ClassTag; it will now demonstrate how this can be used to provide improved analysis of Flickr tagging behavior. Figure 4.10 shows the distribution of Flickr tags over different semantic categories – both using the baseline system and the ClassTag system. When comparing the two charts notice the effect of being able to classify a larger portion of tags. I believe that this gives us a better understanding of the way people annotate their photos. Using the baseline system the size of the Location, Artifact, and People classes were underestimated since the baseline is not able to recognise locations such as NYC, Phuket and Big Island; artifacts such as Notre Dame, London Eye and Starbucks; and common first names such as Alex, Emily, Laura etc.

4.5.3 Summary

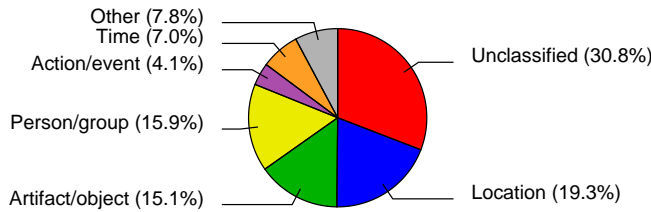
This case study presents a method of categorising Flickr tags as WordNet semantic categories. This is done with the ClassTag system, which first categorises Wikipedia articles and then maps Flickr tags onto these categorised articles. The Wikipedia article categorisation method can be configured to optimise either precision or recall. In either configuration this method outperforms the categorisations provided by DBpedia in a blind evaluation. When optimised for recall nearly **40%** of Wikipedia articles are classified with a precision of **72%**. When optimised for precision **21%** of Wikipedia articles are classified with a precision of **86%**, and a precision of **100%** for certain categories.

I have demonstrated an approach that can categorise a **115%** larger share of the Flickr vocabulary, compared to a baseline using WordNet. Consequently ClassTag is able to categorise **69.2%** of the total volume of Flickr tags – i.e. when tag frequency is taken into account. ClassTag can classify many important entities that are not covered by WordNet, such as, *London Eye*, *Big Island*, *Ronaldinho*, *geocaching* and *wii*.

19.3% of tags in Flickr are classified as locations, this concurs with my initial premise that photographs are inherently location related. I would like to improve the retrieval process through context based placename disambiguation and in this chapter's final discussion, I suggest why photo corpora are not ideally suited for this task.



(a) Using our baseline WordNet based approach.



(b) Using the ClassTag system.

Figure 4.10: Classification of Flickr tags (a) Using our baseline WordNet based approach (b) Using the ClassTag system

4.6 Discussion

I have identified how to use the structural patterns of Wikipedia-article meta-data as features for classification that can produce a relatively high recall while maintaining an acceptable degree of precision. Classification using these structural patterns can easily be applied to other hierarchically structured or networked corpora, such as the Open Directory. A significant proportion of the classifiable Wikipedia articles refer to locations (in fact these are the second most common after articles describing people). Classification and disambiguation of these articles will be further explored in the next chapter.

There are three issues brought up by the case study in this chapter that I would like to discuss further in this final discussion. The first is the benefits a multilingual classifier could provide, followed by whether context based classification is appropriate for tag classification and the effect on data driven methods of differences between training data and the corpus to be classified.

4.6.1 Multilingual classification

Flickr tags are unrestricted and as such appear in a mixture of languages. Clough et al. (2006a) provide an estimate of the language distribution of Flickr tags. They estimate approximately 80% of tags are in English, 7% in German and 6% in Dutch. I expect a large portion of the 21.4% of unclassifiable tags (tags not covered by Wikipedia anchors union WordNet words) fall into this category.

A language specific classification system will clearly fail to classify or will misclassify tags in a foreign language. Whether a multilingual classification system or multiple monolingual classification systems are desirable is dependent on the application. Wikipedia is currently available in 253 languages, the top 14 of which have over 100,000 articles (Wikipedia 2008b). There are no language specific elements to ClassTag. There are two possible methods of creating an alternate language classification of Wikipedia articles:

1. run ClassTag across an alternate language version of Wikipedia with a corresponding lexicon; or

2. the English language classifications generated can be translated into an alternate language using Wikipedia’s Interlanguage links.

This topic will be revisited in Chapter 7.

4.6.2 Context

Bunescu and Paşca (2006) and Cucerzan (2007) have worked on context-aware disambiguation of free-text based on models built from Wikipedia. In this section we consider whether it is appropriate for tagged corpora. Sigurbjörnsson and van Zwol (2008) observe that tagged photos average only 3.6 tags per image, the majority of tagged images in fact have only one tag. Photos are generally annotated with very few tags; because of this, there is often very little context available when classifying images (Overell et al. 2008a). Neither of the methods described by Bunescu and Paşca or Cucerzan are applicable in these circumstances. The method described by Bunescu and Paşca specifies a 55 word window for disambiguation, while Cucerzan attempts to disambiguate all named entities in a document simultaneously, shrinking their window only as small as the sentence level when ties occur. In Section 4.5.1 we observe that when one considers the problem of disambiguating a Wikipedia anchor as a category, instead of disambiguating an anchor as an entity, the number of ambiguous anchor mappings reduce by over 70%.

I consider the area of context-aware disambiguation an essential area for future work in tag classification; when given a tag with multiple possible classifications such as “Java” one would like to be able to classify whether it refers to a location, food or artifact based on context. However, I do not consider this a priority for two reasons: many tags exist with little or no context, and for most tags ambiguity can already be significantly reduced (as shown in Section 4.5.1). I believe the most important area for research in this task is improving categorisation and coverage of tags, unaware of context. Because of this Chapters 5 and 6 of this thesis will focus on building a model and the evaluation of context based placename disambiguation in free text rather than tags, as I believe this is an area context based disambiguation can make significantly more impact.

4.6.3 Which Parties party and which Races race?

When training a classifier one assumes the training data will be representative or at least similar to the corpus being classified. When dramatic differences occur it is near impossible for a classifier to work effectively. Such differences can be seen between Flickr and Wikipedia. While users of Flickr are concerned with documenting events of major and minor importance to them and that are visually interesting, Wikipedians are concerned with encyclopædic knowledge. Articles are often of national or global interest. This difference can easily be seen with the terms *party* and *race*. In Wikipedia “party” most commonly refers to a political party (group in WordNet), while in Flickr it is more commonly an event (act in WordNet). Similarly “race” in Wikipedia most commonly refers to categories of people based on physical traits (group), while in Flickr it is more commonly a competition of speed (act).

The effect differences between corpora have on classification is something that will be considered in Chapter 6.

Chapter 5

Disambiguating locations in Wikipedia

5.1 Introduction

This chapter outlines a method for determining whether a Wikipedia article is describing a location and if it is, grounding that article to a specific location on the Earth’s surface. Our disambiguated Wikipedia articles and the contexts they are referenced in, form a geographic co-occurrence model. The existing methods for mapping Wikipedia articles to locations, discussed in Section 5.2, create inconsistent annotations that are not portable and do not allow for geographic reasoning. Our co-occurrence model has multiple applications in GIR specifically in the areas of placename disambiguation and geographic relevance ranking.

Returning to this thesis’s definition of classification and disambiguation: consider the problem, of matching placenames to locations. Approaching as a classification problem one could construct a classifier that given the context of a placename would classify which country that location is most likely to be in *regardless of the placename itself* (this is the approach taken by Garbin and Mani (2005)). Knowing the likelihood of each country would often be enough to ground a placename as a unique location. On the other hand, approaching this as a disambiguation problem, one would create a separate classifier for each placename with possible locations as classification classes. In this thesis, mapping Wikipedia articles to locations and mapping placenames to locations are both approached as disambiguation problems.

This chapter describes how Wikipedia articles are disambiguated and the classes of evidence considered as implying an article describes a location. It then details the nature and generation of a ground truth. I test which classes of evidence contribute the most information and which are most accurate with respect to implying an article describes a location. Using this information I build a disambiguation pipeline — a pipeline of classes of evidence that are examined in turn when disambiguating an article. The ground truth is then re-examined to check whether it is sufficient to draw conclusions. The pipeline is the disambiguation method used to build our final co-occurrence model, which is compared to four naïve baseline methods of disambiguation.

Finally I present the size, complexity and distribution of the model mined from Wikipedia. An estimate of the model’s clarity is calculated and the size of the model reduced to improve clarity and usability. This chapter concludes with a discussion of the applications and limitations of the co-occurrence model.

5.2 Mapping Wikipedia articles to locations

Assigning locations to Wikipedia articles is not a new task. The first to attempt this were Wikipedia themselves. They established the WikiProject Geographical Coordinates (known as WikiCoords) in 2005 to provide a standard way to handle latitude and longitudes in Wikipedia articles (Wikipedia 2008a). Since then the project has matured. Now to manually geotag an article one simply needs to add “`{{coord | latitude | longitude }}`” into the article source. This creates a coordinate link as shown in Figure A.1:10. The coordinate links point to GeoHack¹, a Wikimedia project linking to a series of geographic information services. Currently there are over 180,000 coordinate links in the English Language Wikipedia across 115,000 articles and over 1 million coordinate links across all languages over nearly 265,000 articles. Pu et al. (2008) mine these coordinate references from the body of articles and from the coordinate templates. This provides a basic gazetteer, which together with locations extracted from the articles, they use for geographic query expansion.

Placeopedia.com was created to make the process of geotagging Wikipedia articles even easier. Placeopedia provides a Google maps mashup to browse and tag Wikipedia articles. To date over 18,000 articles have been tagged on Placeopedia (Steinberg 2008). The advantage of Placeopedia over WikiCoords is that one does not need to know the coordinates of a location in advance and a single interface can be used for tagging and browsing.

The work done by Placeopedia and the WikiCoords project has been successful at achieving their goal: providing additional geographic meta-data about articles in a human readable format. However, there are some problems that occur when trying to apply this data to GIR:

- There are no topological data. Both sources provide only point data and optionally a scale value. This makes geographic reasoning difficult as without topological data one cannot infer such facts as Hawaii is within the United States.
- Coverage is sporadic at best, inconsistent at worst. As the content is generated by a huge volume of users it can be very inconsistent. There are huge variations in the size of places that are and are not geotagged, where the geotag is placed in large or complex locations or even exactly where imprecise or ambiguous regions are, e.g. “the Midlands”.
- Multiple entries exist for single locations. Articles about important locations may be split into separate articles describing different aspects of the location (such as history, geography or politics). Whenever these articles are linked to, it is a non-trivial task to automatically map them to a single location.

These problems are addressed by mapping Wikipedia articles to an authoritative source. Geographic gazetteers, although not perfect, are considerably more consistent than Wikipedia, as in general they are created by a single organisation or relatively small group of organisations. Gazetteers also often contain topological and basic meta-data on the locations they list, making geographic reasoning possible. The disadvantage of using an authoritative source is the same argument that comes up whenever Wikipedia is compared to a top-down data-source: they will have less coverage and will be less current. I believe it is worth trading some coverage and currency for consistency and portability.

¹<http://stable.toolserver.org/geohack/>

5.3 Disambiguating placenames in Wikipedia

This section describes our work on disambiguating placenames in Wikipedia. The previous chapter classified Wikipedia pages as WordNet broad noun categories. This included identifying the subject of pages as locations, objects, artifacts etc. A placename is defined as any word referring to a specific line or polygon on the Earth’s surface. In terms of WordNet categories, this is any specific instance of a location or a geographic object, and some of the larger, fixed position artifacts, e.g. London, the Nile or the Great Wall of China. This does not include the WordNet un-quantified words, e.g., “mountain” or “nation.” For the rest of this thesis when we refer to the term *location* it will be in the gazetteer sense of the word rather than the WordNet sense.

Put succinctly, I am attempting to classify whether a Wikipedia article describes a placename, and, if it does, ground it to a specific location on the Earth’s surface. As with the previous chapter we are going to continue using only article meta-data for disambiguation. However in this chapter we are going to match Wikipedia articles to an authoritative source to maximise the precision of the data. To maintain a high recall we need an authoritative source with a high coverage of placenames. The disadvantage of this is that it significantly increases the ambiguity. We tackle this significant ambiguity problem by only mapping from an article to a location when there exists at least one piece of supporting evidence. Due to the volume of geographic information contained in Wikipedia article meta-data, I have opted for a rule based approach, rather than a supervised classifier approach.

In Section 2.4 some properties of common gazetteers are described. For the experiments in this thesis we have decided to use the Getty Thesaurus of Geographical Names (TGN). It contains approximately 800,000 locations, all with unique identifiers as well as detailed topological information (Harping 2000).

5.3.1 Classes of evidence considered

All of the classes of evidence are based on the meta-data of articles. A piece of evidence must not only identify this article as describing a location but point it to a specific location. Our sources of evidence are:

- article titles,
- article categories,
- anchor texts used to link to articles,
- the content of templates,
- external meta-data.

These sources of evidence are illustrated in Figure A.1; they form the following seven classes of evidence, each class implying a specific article points to a specific location:

- **Default locations**

Important or big locations have articles that are too heterogenous and noisy to disambiguate automatically.

Wikipedia editors have made a list of the most important 1,000 Wikipedia articles, with respect to encyclopædic knowledge². It is a list of articles that every language version of Wikipedia should

²http://meta.wikimedia.org/wiki/List_of_articles_every_Wikipedia_should_have

have. These articles are often very long, with a significant amount of meta-data, and (with respect to automatic disambiguation) a considerable amount of noise. They are also very commonly referred to. There are 150 placenames listed in the geography section of this page, which have been mapped by hand to locations in the gazetteer.

Examples include *Pacific Ocean*, *Europe* and *North America*.

- **Coord in template**

Coordinates in the article template will be enough to disambiguate many pages.

The *coord url* template is one of the most common templates appearing in Wikipedia and has been adopted by the WikiCoords project. As mentioned in Section 5.2 over 115,000 articles in the English Wikipedia directly contain the template or transclude it through more complex templates. If the page title or the anchor text of a link to this page match a placename and the coordinates of the matching placename are in the vicinity of the coordinates listed in the template we consider it evidence.

For example *London (N51.5°, W0.1°)*.

- **Placeopedia.com**

Coordinates mapped to Articles provided by external sources will be enough to disambiguate many pages.

Placeopedia.com is a user generated content web site where Wikipedia can be navigated with a map interface. As with the above evidence, we consider Placeopedia.com as providing evidence when the page title or the anchor text of a link to this page match a placename and the coordinates of the matching placename are in the vicinity of the coordinates listed.

For example *Lands End (N50.1°, W5.7°)*.

- **Title**

As it is required for every Wikipedia article to have a unique title, ambiguous placenames will commonly be disambiguated with a referent location in the article title.

It is common for articles describing locations, particularly ambiguous ones, to have a disambiguating referent placename in the article title. We consider this a piece of evidence when an article title is of one of the following two formats: *placename*, *referent placename* or *placename (referent placename)*, where the referent placename is listed higher in the hierarchical topology tree than the placename in the gazetteer.

For example *London, Ontario*.

- **Title and anchor text**

Placenames will sometimes be referred to by an alternative spelling in the gazetteer than the title of the Wikipedia article. This alternate spelling may be used in an anchor text linking to this article.

As with the previous piece of evidence, we extract a referent location from the article title, however the placename listed in the gazetteer may not necessarily match the article title. The placename must occur in an anchor linking to this article and the article title must be of one of the following formats: *synonym*, *referent placename* or *synonym (referent placename)*, where synonym is a synonym for the placename not contained in our gazetteer.

For example *Chaparral, New Mexico* has the alternative spelling *Chapparal*.

- **Title and category**

Locations are often in categories indicating parent locations.

In the case where the article title matches a placename, or the article title up to the first punctuation character matches a placename and a referent location is found in the name of one of the categories. The names of all the referent locations are looked up from the gazetteer and the names of all the categories are checked to see if any contain a referent location as a sub-string.

For example *Cambridge* is in the category *Cities in Cambridgeshire*.

- **Anchor text and category**

The titles of some articles may not follow standard formats or may have alternative spellings, in which case one must rely on the anchor texts used in references and article categories.

This is a combination of the above two evidences. The article title does not match a placename from the gazetteer however an anchor text linking to this article does, and a referent location for that placename appears as a sub-string in the categories.

For example *Hrodna* has the alternative spelling of *Horodno* and is in the category *Cities and towns in Belarus*.

An additional source of evidence that would be useful, were this evaluation to be repeated, would be the “other uses” templates. These templates came into widespread use at the beginning of 2007 (after the dump of Wikipedia used in this evaluation was taken). It provides disambiguation information and links at the top of an article to distinct articles that this article could easily be confused with.

5.3.2 The three stages

We disambiguate Wikipedia articles as locations in three stages. In Stage 1 Wikipedia is crawled extracting article in-links, out-links and anchor texts. We also record the order in which links occur. This gives us a set of per article language models. Each model captures how different proper names are used to refer to the same article. A list of synonyms used to refer to each article is also built. For example the article *London* describing the capital of the UK can be referred to by the following placenames: “London,” “London, England,” “London, UK,” “the City,” “Central London,” “London’s,” “London, United Kingdom” and “West London”.

In Stage 2 a set of inferences is built for each article in an attempt to map them to locations in the gazetteer. An inference is a mapping between a Wikipedia article and a location with supporting evidence. First a set of possible locations is built from the gazetteer of placenames matching any synonym of the article. Then evidence is searched for that will allow us to infer if this article refers to a specific location. The classes of evidence searched for are listed in the previous section. For example the article *Cambridge* has 23 pieces of evidence implying it refers to the city of Cambridge, UK and 11 pieces of evidence implying it refers to the county of Cambridgeshire, UK.

Stage 2 relies on the assumption that articles not referring to locations that may be referred to by placenames (e.g. *George Washington, China (Porcelain)* or *Texas (Band)*) will not contain any evidence matching them to possible locations.

In Stage 3 ambiguities caused by inferences to multiple locations from a single article are resolved. The inferences for each article are looked at and evaluated in a pipeline. The pipeline checks inferences provided by different classes of evidence in order. Each element of the pipeline has a higher priority

than the following element. Experiments deciding the classes of evidence included in the pipeline and the order they are applied are documented in Section 5.5.

The actual text of Wikipedia articles is not used beyond Stage 1, where article synonyms are extracted. This is due to the limited success found in using the content of Wikipedia articles compared to Wikipedia meta-data (Buscaldi et al. 2006; Overell and R uger 2006). Furthermore, the content of Wikipedia articles are very heterogeneous, while in comparison the meta-data is very structured. In my opinion Wikipedia is now mature enough that the meta-data alone contains enough information for disambiguation.

5.4 Ground truth

The ground truth to evaluate the accuracy of the co-occurrence model takes the form of a list of all the links extracted from 1,000 Wikipedia articles chosen at random. Each link has been manually annotated as to whether it describes a location and matched to a unique identifier in the TGN; this was all done by hand by the author. The ground truth contains 9,108 links: 1,409 to locations and 7,699 to non locations. The 1,409 links are split between 878 unique locations. 99 of these locations are outside of the gazetteer (and therefore unclassifiable). Locations outside of the gazetteer include imprecise regions such as the Scottish Highlands or Upper Canada, and locations at too fine a granularity, for example Raynes Park, a suburb of London.

There are 381 references to pages disambiguated as the 150 *default locations*. As both the ground truth and default locations are annotated by a human they should agree. Disagreement between the ground truth and the default locations can have three causes:

- There exist multiple entries in the gazetteer for a single location, all of which are correct. Large geographic features split across several countries often have multiple entries such as the Andes or the Nile. Also there may be overlapping regions referred to by the same name with different entries where correct disambiguation is difficult and often immaterial, for example the city Brussels, and the administrative region Brussels.
- It is ambiguous whether the location being referred to is the central concept of the article. Some articles do not have a clear single concept and may discuss an ethnic group or an event as well as a geographic location. For example the article *China* describes the Chinese people, while the *People's Republic of China* is the main article on the country.
- Annotator error.

Because of these types of discrepancy, comparisons with the ground truth must be considered as an upper bound of the performance of the system.

We have split the ground truth into a test set and a training set. Each set is made up of the links from 500 articles. The training set will be used in the first set of experiments to quantify the contribution of each type of evidence and build the disambiguation pipeline. The test set will then be used to evaluate the pipeline and compare it to other disambiguation methods. We will revisit this ground truth in Section 5.5.3 and confirm whether or not it is fit for purpose.

5.5 Which class of evidence offers the most information?

In the following section we compare the seven classes of evidence being considered against the test set of the ground truth. Table 5.2 displays six accuracy values measured for each class of evidence and two

	Correct disambig.	Correct class.	Incorrect
Class. placename	true positive (TP)	true classification, false disambiguation (TCFD)	false positive (FP)
Class. non-placename	false negative (FN)		true negative (TN)

Table 5.1: Contingency table modified for classifying Wikipedia articles as placenames

values quantifying the proportion of Wikipedia each class can disambiguate.

All six of these metrics are derived from a modified contingency table. Table 5.1 shows the contingency table from Chapter 3 (Table 3.1) modified for classifying Wikipedia articles as placenames. In this case an article is classified as either a placename or non-placename, and the articles describing placenames are disambiguated as specific locations. Note this splits the True Positive box in half, on the left we have correctly classified and correctly disambiguated locations (TP), and on the right we have correctly classified but incorrectly disambiguated placenames (TCFD).

Below we define the evaluation metrics used for evaluating the different classes of evidence:

- Placename recall (Pn Recall): The proportion of placenames correctly identified as placenames. The number of placenames correctly identified divided by all the placenames in the model.

$$\text{Pn Recall} = \frac{TP + TCFD}{TP + TCFD + FN} \quad (5.1)$$

- Semantic accuracy (Sem Acc): The accuracy with respect to semantic ambiguity. Note I am using Wacholder et al. (1997)’s definition of semantic ambiguity, which is concerned only with the class of an object being correctly identified. The sum of the number of placenames correctly identified and the number of non-placenames correctly identified divided by all the objects considered.

$$\text{Sem Acc} = \frac{TP + TCFD + TN}{TP + TCFD + FP + FN + TN} \quad (5.2)$$

- Grounding: The proportion of placenames recognised correctly matched to the location being referred to in the gazetteer. Note this is an upper bound, as in some places multiple correct locations may exist in the gazetteer. The number of correctly identified placenames correctly matched to locations in the gazetteer divided by all the correctly identified placenames.

$$\text{Grounding} = \frac{TP}{TP + TCFD} \quad (5.3)$$

- Referent accuracy (Ref Acc): The accuracy with respect to referent ambiguity. Here, Wacholder et al. (1997)’s definition of referent ambiguity is used. It is concerned with correctly identifying the specific entity being referred to. The sum of placenames correctly matched to locations and the number of not placenames correctly identified divided by all the objects considered.

$$\text{Ref Acc} = \frac{TP + TN}{TP + TCFD + FP + FN + TN} \quad (5.4)$$

- Semantic F₁ measure (F1 (sem)): The F₁ measure with respect to semantic ambiguity. This can be considered how well the system performs at identifying placenames.
- Referent F₁ measure (F1 (ref)): The F₁ measure with respect to referent ambiguity. This can be considered how well the system performs at identifying placenames and matching them to locations.

Class of evidence	Default locations	Placeopedia .com	Coord in template	Title	Title and anchor text	Title and category	Anchor text and category
Pn Recall	0.280	0.182	0.312	0.158	0.154	0.509	0.594
Sem Acc	0.896	0.882	0.901	0.879	0.876	0.927	0.932
Grounding	0.926	0.887	0.825	0.892	0.841	0.782	0.661
Ref Acc	0.893	0.879	0.893	0.876	0.873	0.911	0.903
F1(sem)	0.438	0.308	0.475	0.273	0.264	0.667	0.716
F1(ref)	0.412	0.278	0.409	0.247	0.227	0.562	0.539
Prop articles	0.005%	0.08%	0.52%	0.93%	0.88%	1.42%	1.61%
Prop links	2.3%	2.3%	10.5%	11.2%	11.6%	19.3%	22.8%

Table 5.2: Information offered by each class of evidence

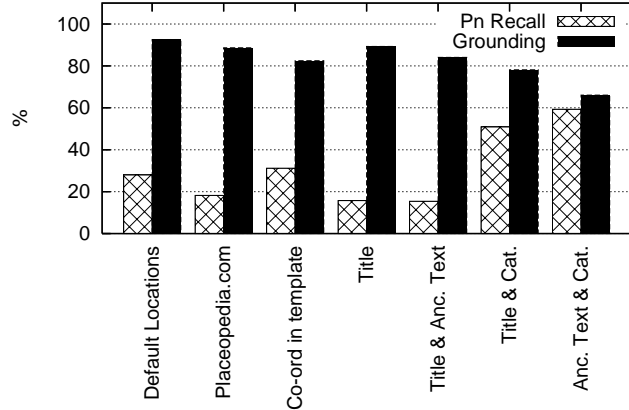


Figure 5.1: Grounding and placename recall achieved by different classes of evidence

- Proportion of articles disambiguated (Prop articles): The proportion of articles this class of evidence matches to locations across the whole of Wikipedia.
- Proportion of links disambiguated (Prop links): The proportion of links to Wikipedia articles this class of evidence matches to locations across the whole of Wikipedia³.

Figure 5.1 illustrates the grounding and placename recall values achieved by the different classes of evidence. There is less than 11% difference between the different classes with respect to grounding except for *default locations*, which is slightly higher, and *anchor text and category*, which is noticeably lower. The two classes using the article category as sources of evidence achieve a much higher placename recall than the others. The placename recall of *default locations* is surprisingly high given how few articles it disambiguates. I attribute this to the distribution of the data, i.e. the skew toward important locations (examined further in Section 5.7).

It is interesting to note in Table 5.2’s last column, the *anchor text and category* class has a noticeable increase in the proportion of articles classified boosting the placename recall at a significant cost to grounding with over a third of the identified placenames incorrectly matched to locations.

The grounding value for the default locations method is 92.6%. One may have expected this value to be 100% as both the default locations and the ground truth were annotated by hand. As the default-locations grounding value is below 100%, this shows a discrepancy between annotations as expected (see Section 5.4). Because of this we can consider the grounding value 92.6% our gold standard to compare the other classes of evidence to.

³Note in the previous chapter only semantic ambiguity was considered so only placename recall, semantic accuracy and semantic F₁ are comparable.

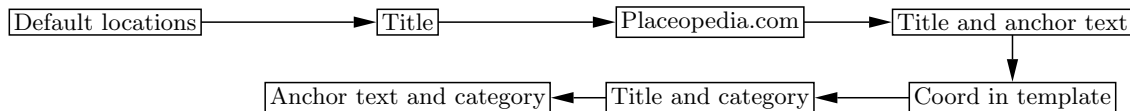


Figure 5.2: Pipeline of classes of evidence

Additional class of evidence	Default locations	Title	Placeopedia .com	Title and anchor text	Coord in template	Title and category	Anchor text and category
Pn Recall	0.280	0.438	0.539	0.539	0.662	0.724	0.748
Sem Acc	0.896	0.919	0.934	0.932	0.949	0.956	0.953
Grounding	0.926	0.914	0.920	0.920	0.889	0.860	0.842
Ref Acc	0.893	0.914	0.928	0.926	0.939	0.941	0.936
F1(sem)	0.438	0.609	0.701	0.695	0.789	0.824	0.820
F1(ref)	0.412	0.572	0.663	0.658	0.734	0.752	0.738
Prop articles	0.005%	0.94%	1.02%	1.08%	1.55%	1.88%	2.13%
Prop links	2.3%	13.6%	15.3%	15.9%	23.9%	26.1%	28.8%

Table 5.3: The pipeline is repeatedly extended by additional evidence (Where additional evidence degrades performance with respect to specific metrics, the degraded values are shown in bold)

In summary the *default locations* method, unsurprisingly, had the greatest grounding as it was manually generated and included only very important locations. The *title* method also has a particularly high grounding as only article titles with a very distinctive format are considered. The *title* and *title and anchor text* methods have very similar results due to the large overlap in articles disambiguated (both methods only disambiguate articles with a referent location in the title). The *anchor text and category* method returns a huge number of a false-positive results. This is reflected in the low grounding value.

5.5.1 Building a pipeline

The classes of evidence are combined in a disambiguation pipeline. Each class disambiguates all the articles it has supporting evidence for. Articles not disambiguated by the first class are passed to the second etc. Articles not disambiguated by any class in the pipeline are classified as non-placenames. The aim being to combine the evidences in such a way that as many placenames as possible are recognised and grounded correctly. With this in mind I decided to order the pipeline by grounding. The motivation for this is all the important locations will be recognised first and grounded with the maximum probability of being correct, as further placenames are recognised they too will be grounded by the class that has the greatest probability of correctly grounding them.

Ordering the classes of evidence by grounding gives us the pipeline pictured in Figure 5.2: *default locations*, *title*, *Placeopedia.com*, *title and anchor text*, *coord in template*, *title and category*, and *anchor text and category*.

In the next experiment, we check what improvement each class of evidence adds to the pipeline. To do this we start with the class of evidence with the greatest grounding (*default locations*), then continually add the next class of evidence. This is illustrated in Table 5.3. Note the first column is the result of only the *Default locations* class of evidence, while the last column is the whole pipeline.

Figure 5.3 shows the effect of adding each additional element to the pipeline. Notice each additional class of evidence recognises additional placenames causing *Pn Recall*, *Prop articles* and *Prop links* to incrementally rise. By combining all the classes of evidence, we classify over 2% of Wikipedia articles as locations and over 28% of the links (note the proportion of these articles that are correctly disambiguated

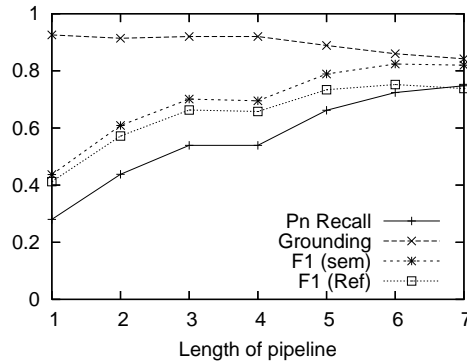


Figure 5.3: Increasing the length of the pipeline against performance measures

is inferred from the ground truth). Also notice as less reliable classes of evidence are added the *Grounding* tends to drop.

As less accurate methods of disambiguation are added to the pipeline and the grounding drops the placename recall can be considered artificially high as some locations outside of the gazetteer will be recognised. As mentioned in Section 5.4, 99 locations in the ground truth are not contained in our gazetteer. Because of this, the gold standard for the Placename recall can be considered **93.0%**.

The semantic accuracy, referent accuracy, semantic F_1 measure and referent F_1 measure rise as every class of evidence is added, except for *title and anchor text* and *anchor text and category*. This means as these two classes are added, more false positives are found than true positives. The following subsections discuss if these classes are actually beneficial to the pipeline.

Analysis

I performed a one-tailed Sign test after each additional class of evidence was added with respect to both semantic accuracy and referent accuracy (Hull 1993). The probability that a statistically significant improvement is provided by adding an additional class is greater than 99.95% for both measures for *title*, *Placeopedia.com* and *coord in template*. The *title and category* method provide a statistically significant improvement with respect to semantic accuracy, however the probability that it provides an improvement with respect to referent accuracy is only 87.8%, which is not significant.

The probability that *title and anchor text* and *anchor text and category* **do not** provide an improvement is over 90% with respect to both performance measures.

Modified pipeline

As the motivation for the pipeline was for both a high number of placenames to be recognised and a large proportion to be correctly grounded, I have decided to manage this trade-off by maximising the semantic F_1 measure and referent F_1 measure. Notice both increase as each class of evidence is added except *title and anchor text* and *anchor text and category*. Both these classes add a significant amount of false positives and many of the placenames they recognise are also recognised by preceding classes. Because of this and the fact that these classes add no significant improvement in accuracy, a new pipeline has been built as illustrated in Figure 5.4. The final pipeline is made up of five classes of evidence: *default locations*, *title*, *Placeopedia.com*, *coord in template* and *title and category*. It has a placename recall of 72.3% (20.7% below gold standard), a grounding of 86% (6.6% below gold standard) and disambiguates over 25% of the links in Wikipedia (shown in Table 5.4). The semantic F_1 and referent F_1 measures are

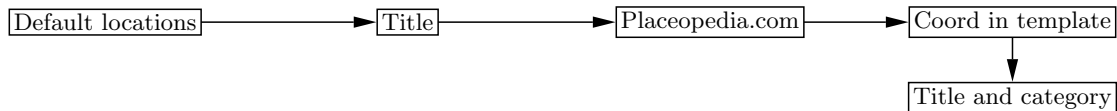


Figure 5.4: Modified pipeline of classes of evidence

Final pipeline	
Pn Recall	0.724
Sem Acc	0.958
Grounding	0.860
Ref Acc	0.943
F1(sem)	0.831
F1(ref)	0.758
Prop articles	1.82%
Prop links	25.5%

Table 5.4: Final pipeline

both higher than the previous pipeline.

5.5.2 Validating the pipeline

To validate the modified pipeline a new pipeline was generated using a greedy algorithm for comparison. I shall refer to this as the greedy pipeline. The greedy pipeline is generated by starting with the class of evidence with the highest grounding value, and at each step adding the class of evidence that gives the greatest increase to the grounding and provides a significant improvement to accuracy. This iterative process stops when no additional classes of evidence can be added to provide a significant improvement with respect to the semantic and referent accuracy.

Table 5.5 shows the results at each stage in the development of the pipeline. Each set of rows shows the classes of evidence considered at each step. Two additional measures are included:

- Probability of statistically significant improvement with respect to semantic accuracy (Prob Sem), and
- Probability of statistically significant improvement with respect to referent accuracy (Prob Ref).

These probabilities are calculated using a one-tailed Sign test. The first sub table is the same as Table 5.2. Notice we start with *default locations* in the first sub table, then we iteratively add other classes of evidence. This continues incrementally growing the pipeline. Notice in the fourth sub table, the class of evidence with the second highest grounding is added. This is because the class of evidence with the greatest grounding provides no significant improvement with respect to semantic or referent accuracy. It is for this reason *title and anchor text* is not considered in the final step and *anchor text and category* is not added.

Comparing the modified pipeline to the greedy pipeline

The greedy pipeline is pictured in Figure 5.5. Notice it is the same as the modified pipeline from Figure 5.4 except the classes *Placeopedia.com* and *title* are transposed. There is no need to do a significance test comparing which of the two pipelines is best as both produce exactly the same classifications (you will notice the bottom left set of results in Table 5.5 are exactly the same as the results in Table 5.4).

If we look closer at the *Placeopedia.com* and *title* classes of evidence, one can see that the majority of articles disambiguated by *Placeopedia.com* are large well known places of interest. In contrast, the

	Default locations	Placeopedia.com	Coord in template	Title	Title and anchor	Title and category	Anchor and cat.
Pn Recall	0.280	0.182	0.312	0.158	0.154	0.509	0.594
Sem Acc	0.896	0.882	0.901	0.879	0.876	0.927	0.932
Grounding	0.926	0.887	0.825	0.892	0.841	0.782	0.661
Ref Acc	0.893	0.879	0.893	0.876	0.873	0.911	0.903
F1(sem)	0.438	0.308	0.475	0.273	0.264	0.667	0.716
F1(ref)	0.412	0.278	0.409	0.247	0.227	0.562	0.539

	Title	Placeopedia.com	Title and anchor	Coord in template	Title and category	Anchor and cat.
Pn Recall	0.438	0.382	0.434	0.471	0.631	0.652
Sem Acc	0.919	0.911	0.917	0.924	0.944	0.940
Grounding	0.914	0.931	0.896	0.880	0.824	0.784
Ref Acc	0.914	0.907	0.910	0.916	0.928	0.920
F1(sem)	0.609	0.552	0.600	0.640	0.765	0.759
F1(ref)	0.571	0.524	0.555	0.586	0.676	0.648
Prob Sem	99.9%	99.9%	99.9%	99.9%	99.9%	99.9%
Prob Ref	99.9%	99.9%	99.9%	99.9%	99.9%	99.9%

	Title	Title and anchor	Coord in template	Title and category	Anchor and cat.
Pn Recall	0.539	0.536	0.504	0.643	0.664
Sem Acc	0.933	0.931	0.928	0.946	0.942
Grounding	0.920	0.905	0.888	0.842	0.842
Ref Acc	0.928	0.924	0.920	0.931	0.931
F1(sem)	0.701	0.692	0.669	0.774	0.773
F1(ref)	0.663	0.647	0.617	0.694	0.694
Prob Sem	99.9%	99.9%	99.9%	99.9%	99.9%
Prob Ref	99.9%	99.9%	99.9%	99.9%	99.9%

	Title and anchor	Coord in template	Title and category	Anchor and cat.
Pn Recall	0.539	0.662	0.684	0.709
Sem Acc	0.932	0.951	0.952	0.949
Grounding	0.920	0.889	0.849	0.824
Ref Acc	0.926	0.941	0.937	0.931
F1(sem)	0.694	0.795	0.804	0.799
F1(ref)	0.658	0.740	0.726	0.708
Prob Sem	1.56%	99.9%	99.9%	99.9%
Prob Ref	1.56%	99.9%	99.9%	99.9%

	Title and category	Anchor and cat.
Pn Recall	0.724	0.748
Sem Acc	0.958	0.954
Grounding	0.860	0.842
Ref Acc	0.943	0.937
F1(sem)	0.831	0.825
F1(ref)	0.758	0.743
Prob Sem	99.9%	78.1%
Prob Ref	87.8%	85.9%

Table 5.5: Greedy pipeline

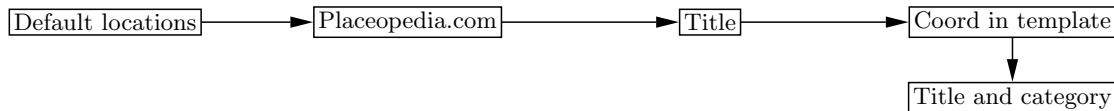


Figure 5.5: Greedy pipeline of classes of evidence

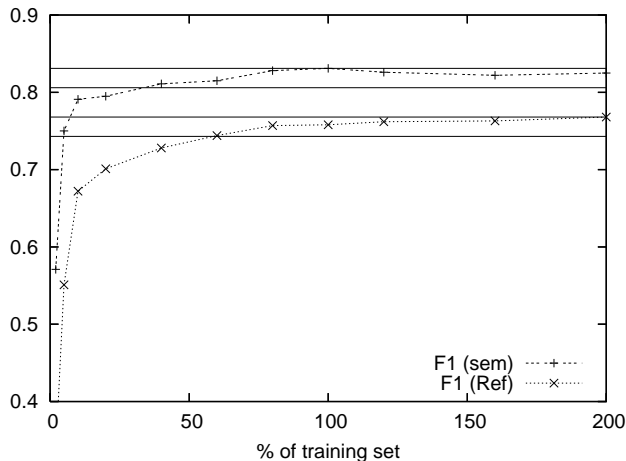


Figure 5.6: Enhancing and degrading the ground truth

majority of articles disambiguated by the *title* method are small places that require a referent location in the article title to make it clear which locations are being described. In fact there are no articles in the ground truth disambiguated by both *Placeopedia.com* and *title* methods.

I conclude that as the *Placeopedia.com* and *title* classes of evidence disambiguate largely disjoint sets of articles, both configurations of the pipeline are equally good and I will use the modified pipeline for further experiments in this thesis.

5.5.3 Enhancing and degrading the ground truth

To confirm that results drawn against the ground truth are valid we need to check that the results seen are stable. In the following experiment, the final pipeline is run across the whole of the ground truth. The training set is both enhanced (up to double its size) and degraded (reduced to one fiftieth its size). The referent F_1 measure and the semantic F_1 measure were chosen as suitable measures to see what size of ground truth is needed to get stable results for disambiguation experiments. Figure 5.6 shows the results.

We consider ground truth a stable size when variation in the referent F_1 measure and the semantic F_1 measure vary by less than 2.5%. The 2.5% window is indicated on the graph in Figure 5.6 using horizontal bars. This point is reached for the semantic F_1 measure once 30% of the training set (15% of the ground truth) has been included (the links extracted from 150 articles). The referent F_1 measure does not stabilise until 60% of the training set (30% ground truth) is included (300 articles). From this I conclude that any ground truth made up of more than 300 articles is large enough to draw conclusions relating to article disambiguation. As our test and training set are made up of 500 articles each, I believe the conclusions are valid.

5.6 Testing

To test the modified pipeline it has been compared to four naïve baseline methods on the test set of the ground truth. The same six measures are used as with the previous experiment. As with the training set, the test set is a subset of the ground truth made up of the links from 500 articles.

5.6.1 Naïve baseline methods

Please refer to the Nomenclature for definitions of the notation used in this section. The first baseline method is *Random*: each placename p is classified randomly as a location from $L(p)$. The intention with Random is to maximise placename recall regardless of grounding and to quantify the amount of error caused by ambiguous placenames.

The second naïve method is *Most Important*: based on the feature type as recorded in the gazetteer, p is classified as the most important location $l \in L(p)$ in the following ordering:

$$\begin{aligned} &\text{as large as or larger than an average nation} \succ \text{large populated area} \succ \text{large geographical feature} \\ &\succ \text{populated place} \succ \text{small geographical feature} \succ \text{small populated place} \end{aligned} \tag{5.5}$$

Any entity not occurring in one of the above categories is deemed too insignificant to return. This method relies on the hypothesis that larger locations are referred to more commonly than smaller ones.

The third naïve method is *Minimum Bounding Box*: the Wikipedia article describing the location is looked at and the first four related placenames (unambiguous if possible) extracted. A minimum bounding box is fitted around every possible combination of locations for the related places. The placename, p , is disambiguated as the location $l \in L(p)$ that is closest to the centre of the box with the smallest area. This method relies on the hypothesis that generally locations referred to within an article are close together.

The final naïve method is *Disambiguate with Co-Referents*: often in articles describing a location, a parent location will be mentioned to provide reference (e.g. when describing a town, mention the county or country). The first paragraph of the article and the article title are searched for names of containing locations listed in the gazetteer. For example if a placename appears in text as “London, Ontario”, Ontario is only mentioned in reference to the disambiguation of London. The gazetteer is then queried for containing objects of locations called “London”: Ontario, Canada and England, United Kingdom. The placename will then be classified as location London, Canada rather than London, United Kingdom.

The intention of this disambiguation method is to maximise placename precision and the proportion of places correctly grounded regardless of placename recall.

5.6.2 Results

Table 5.6 shows, as expected, that to maximise placename recall any article that shares its name with a placename in our gazetteer must be classified as a location (as in Random). To maximise grounding, one must only classify candidate placenames where a referent placename is explicitly mentioned. Our motivation for disambiguating Wikipedia articles referring to locations is to build a co-occurrence model capturing how locations are referred to in context. As we will only be able to reason about locations in our model we need as high a recall as possible. Conversely we can expect the accuracy of annotations or judgments we make using our co-occurrence model to be equal to, or less than, the accuracy of the model itself, thus we need to maximise precision. The pipeline gives a suitable middle ground maximising the

	Random	Most Important	MBB	Referents	Pipeline
Pn Recall	0.855	0.834	0.773	0.612	0.714
Sem Acc	0.919	0.919	0.923	0.929	0.949
Grounding	0.553	0.614	0.693	0.939	0.899
Ref Acc	0.857	0.867	0.885	0.923	0.937
F1(sem)	0.773	0.769	0.765	0.736	0.820
F1(ref)	0.517	0.555	0.601	0.707	0.769

Table 5.6: Comparison of naïve methods and the pipeline with respect to the test set

semantic F_1 measure, referent F_1 measure, semantic accuracy and referent accuracy.

5.6.3 Analysis

As with Section 5.5.1, I performed a one-tailed Sign test comparing the semantic and referent accuracy of the methods (Hull 1993). Four pair-wise statistical significance tests were conducted, in each case the hypothesis was that our pipeline would outperform the naïve methods. The pipeline was significantly better at disambiguating placenames with respect to semantic accuracy and referent accuracy than all four naïve methods with a confidence of more than 99.99%.

5.7 Model size and complexity

The previous sections of this chapter have been concerned with comparing the performance of our algorithms to a small ground truth. The following sections scale the experiments up to cover a significant proportion of Wikipedia. Currently 100,000 randomly selected articles have been crawled. This took approximately five days (Stages 1 and 2 taking approximately two days each and Stage 3 approximately one day) on a 3.2GHz Pentium 4 desktop computer with 1GB of Ram. As the number n , of articles crawled increases, the length of time to complete Stage 1 increases in $\mathcal{O}(n)$ time, and Stages 2 and 3 in $\mathcal{O}(\log(n))$ time.

The current weakness in the efficiency of the implementation is where links are extracted in Stage 1. The links are extracted from Wikipedia using a modified version of the Wikimedia foundation’s MediaWiki application. MediaWiki computes a full parse of every page when extracting links. A more efficient application could simply scan the Wikipedia dump for the relevant markup tags. This would not change the complexity of the algorithm but would increase the speed by an order of magnitude.

Over 7.3 million links were extracted, nearly 2 million of those links to articles describing locations. 385,000 inferences were mined allowing 59,341 articles describing locations to be disambiguated. A total of 63,233 placenames were extracted mapping to 50,694 locations.

5.7.1 Distribution

References to locations in Wikipedia follow a Zipfian distribution⁴: 2,500 locations ($\approx 5\%$) account for 906,000 links ($\approx 50\%$) and 500 locations ($\approx 1\%$) account for 440,000 links ($\approx 25\%$). This distribution is illustrated in Figure 5.7 on a log scale. 25.5% of the links in Wikipedia link to articles describing locations. Figure 5.8 shows the same distribution plotted in two dimensions as a map superimposed on the outline of the Earth’s landmasses. The colour and intensity indicate the number of references to links in a given area. The map is plotted to a granularity of a tenth of a degree, with a one degree radius Gaussian blur.

⁴This is consistent with observations made in Mihalcea (2007) on the distribution of Wikipedia based annotations.

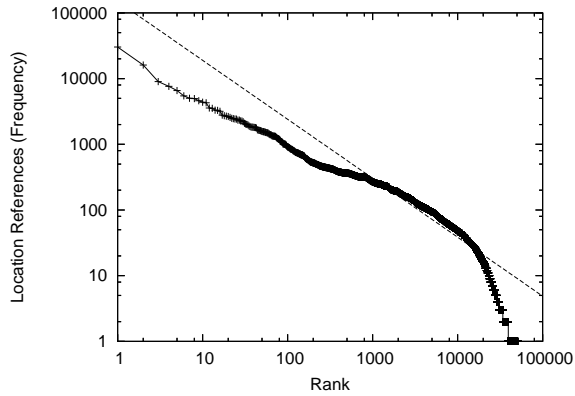


Figure 5.7: Frequency of location references – rank

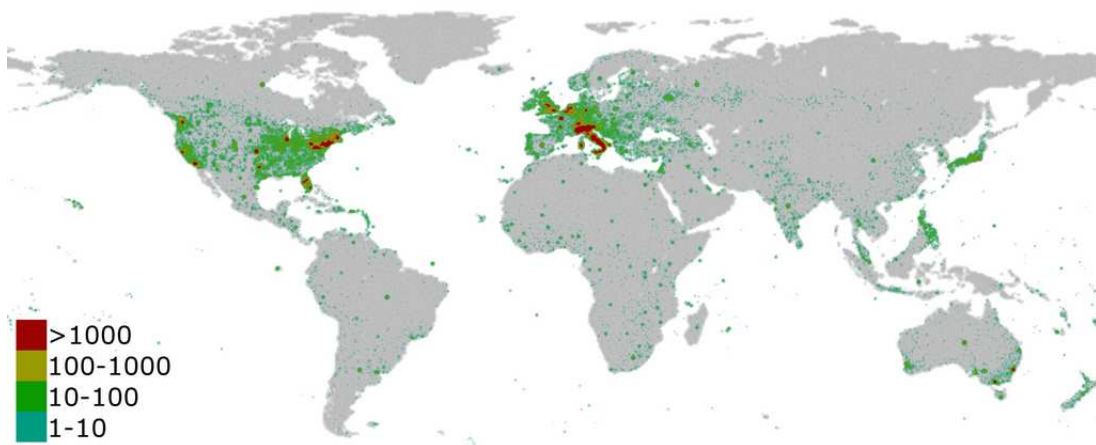


Figure 5.8: 2D projection of location references in Wikipedia – colour indicates number of references to locations within an area

5.8 Building a geographic co-occurrence model

The previous sections of this chapter have described a method for annotating articles in Wikipedia with the locations that they describe. As reported in Section 5.7, this is a large volume of data. To make this annotated portion of Wikipedia a useful resource for geographic information retrieval we need to convert it into a flexible and usable format, and discard any information not strictly necessary.

One approach to discarding information would be to discard the long tail of the distribution. As previously noted, the data follows a Zipfian distribution and by discarding up to 95% of the annotated articles over 50% of references to locations would still be represented. The problem with this method is it discards the main advantage of a model built from this data: small, rarely referred to locations and rarely used synonyms for locations are captured with the context that they are used in.

The alternate approach to discarding data is to discard all references to entities that are not locations. The motivation behind this is the context of locations with respect to named entities is much more heterogeneous than the context of locations with respect to other locations. In the following section I show that discarding references to other entities actually increases the clarity of our model.

Source	Target	Anchor Text	Order	Location
Holland_Park	Greater_London	London	10	7008136
Aberdeen_Airport	Europe	Europe	39	1000003
American_Stock_Exchange	New_York_City	New York	2	7007567
Google_Earth	Mount_Everest	Mount Everest	49	1104571

Table 5.7: Sample of the Location Links Table

5.8.1 Clarity

Raghavan et al. (2004) describe entity models mined from the Web and newswire. These entity models can be considered as per-entity language models that can cluster search results for question answering systems and summarisation systems. Raghaven et al. quantify the information content of their models by measuring the clarity, defined as the Kullback-Leibler divergence (D_{KL}) between entity model distributions and the global distribution. The clarity for the model built for entity e is defined thus:

$$\text{Clarity}_e = \sum_{w \in V} P(w|E_e) \log \frac{P(w|E_e)}{P(w|C)}, \quad (5.6)$$

where E_e is a per entity language model for the entity e . E_e is formed of all the documents where e occurs in the Corpus C . $w \in V$ are the words in the vocabulary V of C , and $P(w|C)$ is the probability of a word w occurring in the corpus C .

Agichtein and Cucerzan (2005) suggest that there is a direct relationship between the accuracy achievable by a classifier trained on a model and the model’s clarity. They observe that as the clarity of a model decreases the accuracy of named entity recognition and relation extraction decreases due to contextual clues being lost in the corpus’s background noise.

Our model is designed for placename disambiguation, geographic relevance ranking and related GIR tasks. These are similar applications of a language model to Agichtein and Cucerzan (2005)’s entity recognition and relation extraction. Therefore, I make the assumption that increasing the clarity of the model will increase the accuracy achieved when it is used as training data for these tasks.

The assumption is that by discarding references to articles not describing locations, it will reduce the corpus’s background noise. This trend can be seen by computing the average Kullback-Leibler divergence between the per-entity distributions and the global model for the 500 most commonly referred to locations. For the locations only model this is **0.75**, while for the model containing all proper names it is **0.35**. I attribute this difference to contexts in the locations only model being more distinctive, and conclude that to increase the clarity references to articles not describing locations should be discarded.

5.8.2 The model

Our final geographic co-occurrence model can be held in a single database table, the *Location Links Table*. This table contains the source and target page of every link to a location article from our crawl. It also contains the anchor text of the link, the order the link occurred in the article, and the unique TGN identifier of the location. Order is defined as the index of the link in the array of all links extracted from a source article, e.g. in Table 5.7 the 10th link in the *Holland Park* article is “London”. The *Location Links Table* contains nearly 7 million records, a sample of which can be seen in Table 5.7.

From the *Location Links Table* two summary tables are generated. The *Article Location Table* contains a many-to-one mapping of Wikipedia articles to their corresponding location unique identifiers. It contains 59,341 records. The *Placename Frequency Table* contains a record for each unique anchor

Article	Location	Anchor Text	Location	Frequency
London	7011781	London	7011781	2273
London,- Ontario	7013066	London	7013066	48
New_York	7007568	New York	7007568	2492
New_York_City	7007567	New York	7007567	142
Europe	1000003	Europe	1000003	2468

Table 5.8: Samples of the Article Location Table and Placename Frequency Table

text – location tuple, and the frequency that a specific anchor text corresponds to a specific location. It contains 75,354 records. A sample of both summary tables is illustrated in Table 5.8.

Rapp (2002) defines two types of relationships that can be inferred from co-occurrence statistics: *syntagmatic* associations are words that co-occur more than one would expect by chance (e.g. coffee and drink); *paradigmatic* relationships are words that can replace one-another in a sentence without affecting the grammaticality or acceptability of the sentence. Clearly both of these relationships are captured in our co-occurrence model. The latter is of less interest as we know all the entities occurring in the model are placenames. The syntagmatic associations are the relationships that will be exploited in the next chapter for placename disambiguation.

5.9 Discussion

This chapter describes the development and evaluation of a pipeline for disambiguating Wikipedia articles and grounding them to locations. This set of disambiguated Wikipedia articles is then converted into a geographic co-occurrence model that captures in what context which placenames refer to which locations. The mapping of articles to locations provided in this chapter is more consistent and has a greater coverage than the current methods of geotagging Wikipedia articles. The annotations are also portable as articles are mapped to unique identifiers from an authoritative source. The chapter has quantified the information offered by different classes of evidence and the information gain as each additional class of evidence is added.

I have developed a ground truth of over 9,000 links extracted from 1,000 articles. The gold standard achieved by a human annotator for grounding locations in this ground truth is 92.6% and the greatest possible recall achievable by matching placenames to locations in the TGN is 93%. The links extracted from 300 articles are all that is needed to make justifiable conclusions.

The co-occurrence model generated is applicable to two areas of GIR: placename disambiguation and geographic relevance ranking. Pu et al. (2008) describe such a co-occurrence model as unsuitable for geographic query expansion as commonly co-occurring locations may be unrelated in a geographic sense. The co-occurrence model captures how locations are referred to by different placenames in different contexts. This allows for fine granularity context based placename disambiguation. To the best of my knowledge this is the largest context model suitable for placename disambiguation that has been created to date. Location contexts are also useful for geographic relevance ranking — when comparing query locations to document locations, a comparison of the location contexts can be considered an additional similarity measure.

Chapter 6

Placename disambiguation in free text

6.1 Introduction

Section 2.3.2 discussed rule based and supervised placename disambiguation. Placename disambiguation based on heuristic rules is an attractive prospect as it requires no training corpus and if the rules are kept simple it can often be done quickly on the fly. The problem with rule-based placename disambiguation is it does not capture the way in which placenames are actually used. Consider a simple example, a document that mentions both “New York City” and “London”. “New York City” can be grounded to New York City, New York through simple string matching. Suppose an algorithm then has to ground “London” as London, Ontario or London, UK. A minimum bounding box approach would clearly choose London, Ontario (Figure 6.1); however if we look at how London and New York City are referred to in context, mining the co-occurrence model from the previous chapter for usage statistics, one can count how often each London occurs with New York City. London, UK and New York City occur in 62 articles together, while London, Ontario and New York City only occur in 6 articles together. This is a highly simplified example but it shows the role data driven placename disambiguation can play. Brunner and Purves (2008) further criticise distance based methods of placename disambiguation (such as MBB methods), noting that placenames with the same name generally occur significantly closer together than random locations. They notice the average distance between Swiss placenames with the same name is smaller than the geographic scope of many newspaper articles, arguably negating distance measures in this context.

The approach taken in this chapter is the same as Mihalcea (2007)’s approach to word sense disambiguation. Wikipedia is used as a disambiguated corpus for training data for classification algorithms. While they employ a Naïve Bayes classifier, this chapter considers two different techniques: neighbourhoods of mined trigger words and a support vector machine. Yarowsky (1993)’s *one sense per collocation* property is assumed to hold: that for every textual context, only one location will be referred to.

This chapter begins by quantifying the problem of placename disambiguation and the theoretical upper and lower bounds of performance achievable for placename disambiguation. Once these bounds have been set we examine where between these bounds we would expect to fall when disambiguating different placenames with respect to each other, answering the question: “Are some placenames easier to disambiguate than others?” Section 6.4 outlines a series of approaches to placename disambiguation.

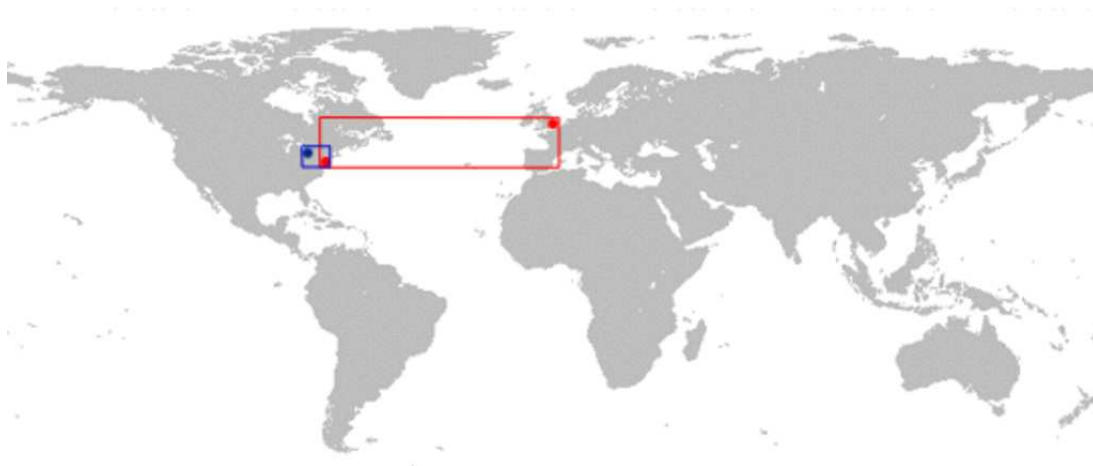


Figure 6.1: Placename disambiguation with a minimum bounding box example

Naïve methods use simple heuristic rules referring to the corpus only for statistics and using standard gazetteers. Supervised methods generate models of how placenames co-occur together, firstly looking at the co-occurrence of single *trigger* placenames, followed by modelling how placenames occur together in a vector space.

I answer a series of questions comparing these approaches in two experimental frameworks. The first experimental framework is a direct comparison of methods on a ground truth. The second experiment integrates these placename disambiguation methods into an ad-hoc geographic information retrieval system to test if disambiguating placenames can improve retrieval results. As well as comparing naïve methods to supervised methods, two baseline methods that perform no disambiguation at all are also introduced. Three different query constructions are tested to allow a comparison of the information contributed by the orthogonal facets of geographic references and text with geographic references removed.

6.2 Supervised placename disambiguation

This section expands on the overview of placename disambiguation provided in Section 2.3.2. The pros and cons of supervised placename disambiguation are discussed in Rauch et al. (2003). They hypothesise that, as much of the contextual information required to disambiguate placenames exists at the document level (and as such placename contexts are very heterogeneous and examples sparse) probabilistic methods are not strictly appropriate. Instead they use a combination of different heuristic methods to provide a confidence that a placename refers to a specific location.

I believe it is possible to extend probabilistic methods beyond simple prior probabilities if the feature space is carefully chosen. Rauch et al. (2003) use supervised methods to learn positive and negative terms which precede and trail named entities implying them to be geographic terms. This allows them to score a context as to whether a geographic entity is expected to occur. Amitay et al. (2003) also use positive and negative examples learnt from contexts at different resolutions for classification (referred to as on and off topic terms). Contexts are scored with respect to a topic based on a weighted sum of terms' $tf \cdot idf$ values.

Both Rauch et al. and Amitay et al. require only a corpus where placenames are annotated to learn corpus statistics. To ground placenames to locations using supervised methods a disambiguated corpus

is necessary. Leidner (2004b) proposes using co-occurrence statistics to learn placename-placename and placename-term relationships. They further propose a disambiguated training and test corpus (Leidner 2004a).

Garbin and Mani (2005) propose a method of learning terms to disambiguate the feature type of placenames¹. Their method is semi-supervised, a heuristic method searches for referent locations (discriminators) within a five word window of ambiguous placenames to build a training corpus. Feature vectors are built for each disambiguated location in the training corpus. The features considered are based on the placename itself, the surrounding k words, the feature types of other placenames occurring in the document, and the feature type of this placename if it has previously occurred in the document. A decision list is then built using the Ripper algorithm to classify placenames as a specific feature type.

This is similar to the approach taken in this thesis. In the previous chapter a training corpus is automatically constructed, it is then applied to placename disambiguation using machine learning techniques in this chapter. The method presented here differs from Garbin and Mani (2005) in two respects:

- the granularity of classification is significantly finer, Garbin and Mani’s classification classes are feature types while the methods presented in this chapter classify placenames as specific locations; and
- while Garbin and Mani build a single classifier for all placenames, I build a separate classifier for each specific placename (classification vs. disambiguation).

I employ the same approach as Mihalcea (2007), where Wikipedia is used to generate a tagged training corpus that can be applied to supervised disambiguation. They apply this method to the domain of supervised word sense disambiguation, solving largely semantic ambiguity. The method presented here further differs as Mihalcea performs the disambiguation of Wikipedia articles manually.

Leveling and Hartrumpf (2006) also use a single classifier to classify aspects of placenames. They use a hand constructed ground truth for training data and classify whether placenames are used in their literal or metonymic sense. A literal reference uses a placename to refer to a location e.g. “the 2012 Olympic games will take place in London,” while a metonymic reference uses a placename to refer to an event or a group of people e.g. “the London Olympics” or “London welcomes a chance to host the Olympics.” Feature vectors contain a collection of context statistics including part-of-speech, word and sentence length etc. They employ the TiMBL classifier.

Smith and Mann (2003) take an approach similar to Garbin and Mani (2005) using heuristic methods to annotate a training corpus and supervised methods to extend these annotations. Their classification classes are the containing state or country of a location. They train per placename classifiers to disambiguate placenames occurring in their training data and story classifiers to classify placenames not seen in their training data. For example if “Oxford” is recognised as a placename and is not in the training data, the document rather than the placename will be classified. If the document can be classified as likely to be describing Mississippi, it is still possible to disambiguate the Oxford reference. Surrounding words are used as features and a Naive Bayes’ classifier for classification.

Similarly to the comparison with Garbin and Mani (2005), the classification granularity presented in this chapter is finer than Smith and Mann (2003)’s. Also they train backoff classifiers to classify placenames not occurring in their training data. This chapter takes the opposite approach, by using as large a ground truth as possible, it is assumed placenames not mentioned in the training set are very

¹The feature type of a placename is the class of the respective location e.g. Region/County, Populated place or Capital.

unlikely to be mentioned in the test set, so their prior probability is set to zero. As Wikipedia is updated in real time and these models use only shallow features it is also possible to keep the models updated.

6.3 Theoretical performance

The following section outlines the expected performance of an arbitrary supervised placename disambiguation algorithm trained with the co-occurrence model described in the previous chapter.

6.3.1 Computing the bounds

To judge the relative performance of the placename disambiguation algorithms compared later in this chapter and to judge whether the gains available with supervised placename disambiguation are worth pursuing, this section quantifies the upper and lower bounds of accuracy that can be achieved.

The following equations use the notation defined in the Nomenclature. Brunner and Purves (2008) observe that dependent on the gazetteer between 10% and 50% of placenames are ambiguous. In the co-occurrence model mined in the previous chapter 7.8% of placenames are ambiguous, but 35.1% of placenames references are ambiguous. Assuming we classify every placename as the most referred to location, then the lower bound for the fraction of correctly disambiguated placenames $r_{\text{corrLower}}$ can be estimated as the proportion of placenames that refer to the most common location with that placename.

$$r_{\text{corrLower}} = \frac{\sum_{p \in M} \text{ref}(p, L_1(p))}{|N|}. \quad (6.1)$$

Using the co-occurrence model, $r_{\text{corrLower}}$ can be calculated as 89.6%. This can be considered a lower bound that can be achieved by classifying every placename as the most referred to location.

To calculate an upper bound we assume that when a placename p is found it has a prior probability of referring to $L_1(p)$ unless the context implies otherwise and that co-occurring placenames provide a suitable context for disambiguation. Three, later revised assumptions are also made: that any context is enough to disambiguate a placename; the co-occurrence model is 100% accurate; and the model represents every context locations occur in. If these assumptions hold we can place the upper bound of performance that a perfect classifier can achieve as the fraction of placenames referring to either the most common location or locations with a context (Equation 6.2).

Let $\text{ref}(p, l, c)$ be the number of references made to location l by placename p with a context of size c within a model, then

$$r_{\text{corrUpper}} = 1 - \frac{\sum_{p \in M} \sum_{i=2}^{L(p)} \text{ref}(p, L_i(p), 0)}{|N|}. \quad (6.2)$$

This is one minus the fraction of locations that occur without a context (i.e. they are the only location referred to in an article) and that do not refer to the most common location ($L_1(p)$). For example an article that mentions “London” in reference to London, Ontario with no context, would be undisambiguatable. $r_{\text{corrUpper}}$ is calculated as equal to 99.98%. This upper bound can be considered what a *perfect* classifier would achieve. I realise this value is particularly high and attribute this to two causes: Firstly, one would expect locations referred to by a placename where a more commonly referred to placename exists to generally be framed in a context. Secondly, the upper and lower bounds quoted above are accurate with respect to the model. As shown in Table 5.6 in the previous chapter, the placenames in the model are only correctly grounded 89.9% of the time, therefore a more realistic

	D_{KL} loc.	D_{KL} Prop. N	$P(l p)$
Camb. MA	0.054	0.26	0.507
Camb. UK	0.592	0.526	0.401
Camb. NZ	0.027	0.015	0.005
Lond. UK	0.004	0.0007	0.961
Lond. ON	0.233	0.11	0.021
Lond. CA	0.036	0.009	0.001

Table 6.1: D_{KL} between location and placename distributions

upper and lower bound for achievable accuracy would be **80.6%** and **89.9%** (found by multiplying the bounds by the grounding) recognising approximately **71%** of placenames.

6.3.2 Computing the relative performance

The following section estimates how well one would expect classification methods to perform when disambiguating different placenames with respect to each other. As mentioned in Section 5.8.1, Agichtein and Cucerzan (2005) suggest one can quantify the difficulty of an information extraction task by measuring the Kullback Leibler divergence (D_{KL}) between the local and global distributions (Equation 5.6). The motivation behind this is, if the D_{KL} is small the local and global distributions are similar, clues for disambiguation will be lost in background noise. If the D_{KL} is high the context individual locations appear in will be more distinctive.

To illustrate this, Table 6.1 shows the D_{KL} between the local Cambridge, MA, Cambridge, UK and Cambridge, New Zealand distributions and the global “Cambridge” distribution; and the London, UK, London, Ontario and London, California distributions and the global “London” distribution. The D_{KL} for the co-occurrence model made up of only locations and the model made up of all proper names² are both calculated. The fourth column is the proportion of times that the placename “Cambridge” or “London” is classified as the corresponding location. This can be considered a prior probability for classification: the probability of location l given placename p ($P(l|p)$).

The higher the D_{KL} the easier a placename should be to disambiguate. Notice in every case except Cambridge, MA there is more noise in the model where all proper names are used rather than only locations. A low D_{KL} can have several causes: If the most referred to location for a placename is referred to significantly more times than any other location it will swamp the global distribution. This is the case for London, UK and to a lesser extent Cambridge, MA. If a location is very rarely referred to there are not enough co-occurring locations to build a suitably large context; this is the case with the relatively small towns of Cambridge, NZ and London, CA. The easiest locations to disambiguate are locations that are commonly referred to in distinctive contexts; this is what we see with Cambridge, UK and London, ON.

To illustrate how the location distributions for the placename “Cambridge” differ, Figure 6.2 illustrates how locations co-occur with Cambridge, UK, Cambridge, MA and Cambridge, NZ respectively. Notice all three distributions are quite biased with Cambridge, MA dominating the U.S., Cambridge, UK dominating Europe and Cambridge, NZ dominating New Zealand.

The Jensen-Shannon divergence (D_{JS}) is a measure of the distance between two probability distributions (Cover and Thomas 1991). The Jensen-Shannon divergence between two locations l_x and l_y can be considered the distance between the distributions of the locations they co-occur with. Let \mathbf{X} and \mathbf{Y}

²The model of all proper-names is the co-occurrence model before non-location entities are removed. This was shown to have a lower clarity in the previous chapter.



Figure 6.2: Distribution of how locations co-occur with Cambridge, UK (red), Cambridge, MA (green) and Cambridge, NZ (blue)

refer to the distribution of co-occurring locations for l_x and l_y respectively and the D_{JS} between \mathbf{X} and \mathbf{Y} defined as

$$D_{JS}(\mathbf{X}, \mathbf{Y}) = H(\alpha \mathbf{p}_X + (1 - \alpha) \mathbf{p}_Y) - \alpha H(\mathbf{p}_X) - (1 - \alpha) H(\mathbf{p}_Y), \quad (6.3)$$

where \mathbf{p}_X is the probability distribution of \mathbf{X} and \mathbf{p}_Y is the probability distribution of \mathbf{Y} . α is a value between 0 and 1. $H(\mathbf{p}_X)$ is the entropy of \mathbf{p}_X defined:

$$H(\mathbf{p}_X) = - \sum_{l \in \mathbf{X}} \mathbf{p}_X(l) \log(\mathbf{p}_X(l)), \quad (6.4)$$

where $l \in \mathbf{X}$ is a location occurring in the distribution of \mathbf{X} and $\mathbf{p}_X(l)$ is the probability of l occurring in \mathbf{X} . The D_{JS} between two locations can be considered a similarity measure. The more similar two locations the less differentiating clues with respect to context and the harder they are to disambiguate.

The D_{JS} is a smoothed averaged D_{KL} . D_{JS} and D_{KL} have the following relationship:

$$D_{JS}(\mathbf{X}, \mathbf{Y}) = \alpha D_{KL}(\mathbf{X}|\mathcal{M}) + (1 - \alpha) D_{KL}(\mathbf{Y}|\mathcal{M}), \quad (6.5)$$

where

$$\mathcal{M} = \alpha \mathbf{p}_X + (1 - \alpha) \mathbf{p}_Y \quad (6.6)$$

and α generally equals $1/2$.

Table 6.2 contains the D_{JS} between the three most common locations referred to by the placenames “London” and “Cambridge”. Note the upper left corner compares locations with different placenames while the lower and right corners compare locations with the same placename. It is the D_{JS} between locations sharing a placename that is of interest to placename disambiguation. Note the D_{JS} between London, UK and London, ON is the lowest, followed by Cambridge, UK and Cambridge, MA. This means these locations occur in relatively similar contexts and would be easy to confuse. Cambridge, MA and Cambridge, NZ are the most different making them easier to disambiguate.

Looking at the table as a whole, one can see Cambridge, UK and London, UK have the closest distributions. This is to be expected as these are two commonly referred to cities, 50 miles apart. The two locations with a distribution furthest apart are Cambridge, MA and Cambridge, NZ. This is also

	Ca. UK	Ca. NZ	Ca. MA	Lo. CA	Lo. ON
Lo. UK	0.246	0.661	0.643	0.671	0.503
Lo. ON	0.620	0.675	0.660	0.675	
Lo. CA	0.681	0.677	0.679		
Ca. MA	0.646	0.690			
Ca. NZ	0.675				

Table 6.2: D_{JS} between location distributions

not surprising, despite Cambridge, MA being a small city, it is the home of two prominent Universities: Harvard and MIT. On the other hand, Cambridge, NZ is a medium sized town on the other side of the world.

6.4 Approaches to placename disambiguation

This section describes the three naïve disambiguation methods and two more complex disambiguation methods compared in Sections 6.5 and 6.6. The naïve methods of disambiguation should provide a suitable baseline. This section details how each method works and the motivation behind it.

There are several specific research questions we wish to answer by comparing these disambiguation methods. By comparing the naïve methods we wish to answer:

- Is a well constructed default gazetteer a powerful enough resource for placename disambiguation?
- Can methods employing statistics gathered from a corpus outperform a gazetteer alone?

Our first complex method builds neighbourhoods of trigger words; a single trigger word occurring in the context of a placename is all that is necessary for disambiguation. In contrast our second method of disambiguation builds a vector space of co-occurring placenames and uses all the evidence available to partition this space. In comparing these two methods we hope to answer:

- Is the co-occurrence of single placenames more important than the combined information of all co-occurring placenames (which can introduce noise)?

Finally by comparing the naïve methods to the more complex methods we intend to test:

- Can supervised learning be more effective for placename disambiguation than simple hand constructed rules?

6.4.1 Naïve methods

Three naïve methods are compared: *Most Important*, *Most Referred to* and *Referents*. The *Most Important* method derives information only from the gazetteer, while the other two methods also use information from the co-occurrence model. The *Most Important* and *Most Referred to* methods effectively build default gazetteers; meaning there is only a single possible location for each placename. The *Referents* method allows for some disambiguation.

Most Important (MI)

The *Most Important* method is exactly the same as the method of the same name from the previous chapter for disambiguating Wikipedia articles. Based on the feature type as recorded in the gazetteer, p is classified as the most important location $l \in L(p)$ in the ordering listed in Equation 5.5. Any locations

not an instance of one of the listed feature types is deemed too insignificant to return. As before, this method relies on the hypothesis that larger locations are referred to more commonly than smaller ones.

Most Referred to (MR)

The *Most Referred to* method is similar to the MI method, in that an ordering is given across locations, however the ordering across locations is based on the frequency of references by that placename given in the gazetteer. For example to disambiguate the placename “London,” one would look up how often “London” refers to London, UK (2,273 references) and how often “London” refers to London, Ontario (48 references) and disambiguate as the greatest. The advantage of this method over the MI method is that it allows significant but small ambiguous places to be disambiguated correctly (for example “Washington” referring to Washington D.C. would be confused with Washington State by the MI method), and it solves ties (for example Cambridge, MA and Cambridge, UK are both of a similar size). This method uses the simple hypothesis that the most referred to locations in Wikipedia will be the most likely to appear in a test corpus.

Referents

The *Referents* method is based on the MR method however with an additional heuristic rule. If the placename being disambiguated is immediately followed by another placename that appears higher up in the vertical topology graph (Figure 2.2) it will be disambiguated as that location. For example if “London” is immediately followed by “Ontario” it will be disambiguated as London, Ontario. When the placename is not followed by a referent location, it will default to the MR location. If several possible locations occur, for example “Midtown, USA,” the most commonly referred to of the possible locations will be chosen. The hypothesis behind this method is an extension of the MR hypothesis: the most referred to locations in Wikipedia will be the most likely to appear in a corpus, and when a less commonly referred to location is mentioned it will co-occur with a referent location.

6.4.2 Neighbourhoods

Comparison of the complex methods of disambiguation is designed to test whether the correct disambiguation of a placename is more dependent on individual placenames in the context or the context as a whole. The first method described is borrowed from the field of word sense disambiguation and is described in detail in Guthrie et al. (1991). They describe a method of building *neighbourhoods* of subject-dependent trigger words. Their motivation is that a single sense of a word will generally be used in documents on a single subject area. Neighbourhoods of trigger words can be built based on the context of words. During the disambiguation process, these trigger words can be searched for. This is analogous to Amitay et al. (2003)’s on topic terms and Garbin and Mani (2005)’s learnt discriminative terms.

Trigger words are found based on a relatedness score. Guthrie et al. (1991) describe the relatedness score as the ratio of co-occurrences between words divided by the disjunction of occurrences. This can be defined as:

$$r(x, y) = \frac{f_{xy}}{f_x + f_y - f_{xy}}, \quad (6.7)$$

where f_x denotes the total number of times word x appears.

As an example of this algorithm in action, Guthrie et al. (1991) gives the word “bank,” which appears in different senses in economics and engineering. The top three trigger words in economics are “account,”

Cambridge, UK	Preston	Derby	Sheffield	Lincoln	King's Lynn
Cambridge, MA	Barnstaple	Bristol	Dukes	Essex	Middlesex
Cambridge, NZ	Wuxi	Alexandra	Ashburton	Carterton	Coromandel
London, UK	Europe	France	Spain	China	Manchester
London, ON	Barrie	Gananoque	Orillia	Prescott	Smiths Falls
London, CA	Tulare	Alpaugh	Cutler	Dinuba	Ducor

Table 6.3: Top related placenames in the location neighbourhoods

“cheque,” and “money,” while in engineering the top three words are “river,” “wall,” and “flood.” Please refer to Guthrie et al. (1991) for more details and examples.

The threshold at which point a word is too unrelated to be considered a trigger word is not specified in the algorithm description; so I have chosen 5% as the minimum proportion of ambiguous word occurrences and potential trigger word occurrences that must co-occur for a trigger word to be valid.

When converting this algorithm from word sense disambiguation to placename disambiguation we consider ambiguous locations analogous to ambiguous words. A word’s senses are equivalent to different locations referred to by the same placename. The co-occurrence model described in the previous chapter forms our ground truth from which neighbourhoods are trained. Table 6.3 contains the five most highly related locations to the ambiguous placenames discussed in the previous section. The size and significance of the locations in Table 6.3 is captured rather well in their neighbourhoods. Cambridge, MA and London, ON are related to small and large towns and counties in their state/province. Cambridge, UK is related to large towns and small cities distributed across the middle of England. Cambridge, NZ is most closely related to locations spread across both islands of New Zealand and surprisingly, Wuxi, a chinese city twinned with Hamilton, New Zealand (Cambridge’s closest city). London, CA is a small town in Tulare county. It is most closely related to the county in which it resides and similar small towns in the same county. London, UK is significantly different. Rather than being related to locations in its locality, it is most closely related to cities, countries and continents spread across the world.

In their original paper Guthrie et al. (1991)’s algorithm is as follows: For the sentence the ambiguous word appears in, find the neighbourhood with the highest number of overlapping words. If the number of overlapping words is higher than a threshold (provided in the paper as two) then disambiguate as a sense corresponding to the subject area of this neighbourhood. Otherwise iteratively expand the neighbourhood with words related to the neighbourhood words within the subject area. For example “bank” in the subject area economics would be expanded with the most related words to “account,” “cheque,” and “money,” found from documents only in the economics subject area.

Here neighbourhoods are applied to placename disambiguation in a similar method, however with a few essential differences. The algorithm was originally designed to solve sentence level syntactic/semantic ambiguity. Yarowsky (1994) recognises that a window of 3–4 words (sentence level) is required for this. We, on the other hand, are solving referent ambiguity. Yarowsky (1994) quotes a 20–50 word window (paragraph and above level) for semantic/topic based ambiguity. As this chapter is only working on the co-occurrence of placenames and discarding other word occurrences in the model, I have chosen a window size of ± 10 location references.

Guthrie et al. (1991) set their threshold of overlapping words between the sentence and the neighbourhood as two. I believe this is too high for placename disambiguation. As recognised by Garbin and Mani (2005), one placename is often enough for a positive disambiguation. The final step of Guthrie et al. (1991)’s algorithm is the iterative expansion of the neighbourhoods. I consider individual locations as analogous to subjects; this means that our ambiguous term (the placename) occurs in every document

referring to a specific subject. Because our subjects are so much more narrowly defined than in the initial specification, there is very little information to be gained by expanding them. If there are no overlapping words in any of the neighbourhoods for a specific location the algorithm falls back to the default locations of the MR method.

6.4.3 Support Vector Machines

To take into account higher orders of co-occurrence, the final disambiguation method approaches place-name disambiguation as a vector space classification problem. Florian et al. (2002) describes approaching WSD as a vector space classification problem. In this problem placenames can be considered as objects to be classified and possible locations as classification classes. The choice of features for such a problem is critical.

This method uses a Support Vector Machine (SVM) to classify placenames as locations. The multidimensional space is partitioned with a linear classifier which is iteratively calculated by the optimisation of a quadratic equation (Joachims 1999). The complexity of the problem is dependent on the number of training examples; however, this is not a significant problem as a separate model is generated for every placename and the largest number of occurrences of a single placename in the model is the United States with 30,227 references. The motivation of this method is to see if multiple orders of co-occurrence can improve accuracy.

Building a Vector space

An important part of classification within a Vector Space is the choice of features and weighting function. This was discussed in Section 4.3.4. Placenames co-occurring in a window of size of ± 10 will be considered as features (as discussed in the previous section). Each different placename will be considered a feature (represented as a different dimension). As with Section 4.3.4 different weighting functions will be compared: as the data is not generated by navigating a hierarchical corpus the top performing weighting function from Chapter 4, *tf-il*, is no longer applicable. Instead the commonly accepted *tf-idf* function (which came a close second to *tf-il* in the classifying wikipedia articles experiment) and a proximity preserving feature space are used.

The choice of feature space is also influenced by related work: Amitay et al. (2003) use a *tf-idf* feature space, while both Garbin and Mani (2005) and Leveling and Hartrumpf (2006) have proximity as elements in their feature vectors.

- **Term Frequency · Inverse Document Frequency (SVM-*tf-idf*)**. Each placename co-occurring in the context of the placename being classified forms a feature. The scalar value of each feature is the number of times the respective placename occurs in the window of the placename being classified divided by the log of the number of times it occurs in the whole co-occurrence model.
- **Proximity Preserving (SVM-*prox*)** Again each placename co-occurring in the context of the placename being classified forms a feature. The scalar value of each feature is the inverse of the respective placename's distance from the placename being classified, its sign is governed by whether it appears before or after the placename being classified. This space was chosen because no information on word proximity is lost and the greatest weighting is given to words appearing closest to the placename being classified.

Table 6.4 shows the features the first and second placename in the following excerpt from the "Visit Cambridge" website would have:

Feature	<i>tf-idf</i>		prox	
	London	Cambridge	London	Cambridge
London	0	0.29	0	-1
Cambridge	0.71	0	1	0
England	0.28	0.28	0.5	1
U.K.	0.48	0.48	0.33	0.5
Stanstead	3.32	3.32	0.25	0.33
Harwich	0.70	0.70	0.2	0.25
Europe	0.29	0.29	0.17	0.2

Table 6.4: Weighting functions example

Just 60 miles north of London, Cambridge is located in the heart of the East of England, excellent road and rail links ensure the city is accessible from all parts of the U.K. Stansted Airport and the ferry port of Harwich are both within easy reach of Cambridge. Europe and the rest of the world isn't far away either.

6.5 Direct measurement

The experimental setup in this section is similar to the setup in Section 5.6. A corpus has been constructed from publicly available resources for word sense and named entity disambiguation. The described approaches are compared against this ground truth. The section begins with a description of the ground truth, followed by the experimental results and analysis.

6.5.1 Building a ground truth

Leidner (2004a) and Clough and Sanderson (2004) recognised that a uniform ground truth was needed to compare placename disambiguation systems. At the time of writing this thesis a definitive corpus is yet to emerge. Currently placename disambiguation systems are either tested directly on small bespoke corpora (Garbin and Mani 2005; Smith and Mann 2003) or indirectly on the GeoCLEF corpus (Cardoso et al. 2005; Overell et al. 2007; Martins et al. 2006; Leveling and Veiel 2006) (discussed further in the next section). Two general purpose annotated publicly available corpora are also proposed for placename disambiguation. Buscaldi and Rosso (2008c) proposes the sense tagged SemCor collection as suitable for placename disambiguation evaluation; this is further explored in this chapter. Turton (2008) suggests using a subset of 1871 papers from the PUBMED collection on the subject of avian influenza. These papers have been annotated with locations from the life-sciences MeSH ontology.

The ground truth presented here is a combination of three corpora. Cucerzan (2007) train a model to disambiguate named entities, their model maps from named entities to Wikipedia articles. Their test set is publicly available³ and consists of 350 Wikipedia articles (stripped of markup) and 100 news stories. They use Wikipedia's mapping of anchor texts to Wikipedia articles for the Wikipedia portion of their collection and tag the news stories by hand. As we are only interested in locations, we take a subset of their ground truth – documents where a named entity is mentioned that is mapped to a Wikipedia article disambiguated as a location in the previous chapter. To turn the mapping of named entities → Wikipedia articles into a mapping of placenames → locations, all named entities mapped to disambiguated Wikipedia articles are annotated with the corresponding coordinates.

WordNet is a semantic lexicon of the english language grouping words into synsets and recording the relationships between them (Fellbaum 1998). WordNet was used in Chapter 4 where Wikipedia articles

³released by Microsoft <http://tinyurl.com/olbup9>

Collection	Docs	Loc. Refs	Ambig. Refs	Unique PNs	D_{JS}
Microsoft – Wikipedia	336	951	116	554	0.356
Microsoft – News	20	145	67	80	0.486
SemCor – Brown	121	1054	625	225	0.399
Total	477	2150	808	757	0.388

Table 6.5: Summary of ground truth collections

were classified as WordNet synsets. SemCor provide a subset of the Brown Corpus tagged with WordNet semantic classes⁴ (described in Section 3.2.1).

By mapping from WordNet synsets to locations in the TGN, SemCor’s subset of the Brown Corpus can be converted into a corpus suitable for evaluating placename disambiguation. This mapping of synsets \rightarrow locations is achieved by considering any synset in WordNet with the broad category location and a synonym s matching a placename in the TGN a possible location with possible disambiguations $l \in L(s)$. The glossary entry of the synset is then searched for possible referent locations.

Documents from the Brown Corpus mentioning at least one location with respect to WordNet classes mapped to the TGN form our third and final test corpus. The Corpora are summarised in Table 6.5. The number of documents (Docs), references to locations (Loc. Refs), references to ambiguous placenames (Ambig. Refs), unique placenames (Unique PNs), and Jensen-Shannon divergence with the co-occurrence model are recorded. Ambiguous placenames are placenames ambiguous with respect to the co-occurrence model rather than the gazetteer; this gives a lower bound for the number of placenames that are ambiguous. The D_{JS} from the co-occurrence model is shown because one would expect a corpus that is more similar to the model (lower D_{JS}) to achieve better results.

6.5.2 Results

Each of the six disambiguation methods were run across the ground truth. The entity annotations from the ground truth were used to locate placenames instead of a named entity recognition system, effectively giving perfect semantic accuracy. This allows the referent accuracy to be tested in isolation, as the only task performed by the disambiguation engine is annotating each placename with a location from the TGN. As the TGN contains overlapping entries of very similar locations with the same placename (e.g. the city Brussels, and the administrative region Brussels) the definition of a correct disambiguation is slightly relaxed from matching exact TGN identifiers to matching geographic co-ordinates within 1 kilometre of each other.

Table 6.6 shows the referent accuracy achieved per collection. Notice the best performance occurred on Microsoft’s Wikipedia corpus. This is to be expected as the corpus is most similar to the model. The performance on the Brown Corpus was surprisingly bad compared to the other two. This can be partially attributed to the Microsoft results being slightly inflated due to how the ground truth was generated — the recall was effectively boosted as only locations contained in our co-occurrence model are tagged in the collection.

We can allow for this difference by only considering locations in the corpora that exist in our model. This is justified as our supervised classifiers have only been trained to classify previously seen entities. The final row in Table 6.6 shows the results after removing the 182 references in the Brown Corpus to locations not in the co-occurrence model. Notice that these results, although worse, are more inline with the News corpus results.

⁴<http://www.cs.unt.edu/%7Erada/downloads.html>

	MI	MR	Referents	Neigh.	SVM-tf-idf	SVM-prox
Microsoft – Wikipedia	0.651	0.891	0.888	0.911	0.926	0.915
Microsoft – News	0.507	0.716	0.701	0.754	0.739	0.739
SemCor – Brown	0.459	0.553	0.521	0.471	0.552	0.550
SemCor – Brown (subset)	0.557	0.672	0.633	0.571	0.671	0.668

Table 6.6: Accuracy per collection

	MI	MR	Referents	Neigh.	SVM-tf-idf	SVM-prox
Precision	0.728	0.891	0.864	0.860	0.901	0.894
Recall	0.686	0.780	0.779	0.764	0.792	0.790
Accuracy	0.546	0.712	0.694	0.680	0.728	0.723
F ₁	0.707	0.832	0.819	0.810	0.843	0.839
Recall	0.771	0.864	0.865	0.849	0.877	0.875
Accuracy	0.599	0.781	0.761	0.746	0.799	0.793
F ₁	0.750	0.877	0.864	0.855	0.888	0.884

Table 6.7: Performance across total ground truth and ground truth subset containing only locations in the co-occurrence model

Table 6.7 shows the merged results across the three ground truth corpora for precision, recall and accuracy. All measures are with respect to referent ambiguity as the existing corpus annotations are used to recognise placenames. The lower part of the table includes the subset of the Brown Corpus of only locations contained in the co-occurrence model instead of the whole Brown Corpus. Note this has no effect on precision. With respect to all performance measures and all corpora the SVM with the *tf-idf* vector space is the best performing method.

Referring to our forecast figures in Section 6.3, we expected a recall of approximately 71% of locations being disambiguated with a precision (accuracy of classified placenames) of between 80.6% and 89.9%. The presented methods have performed well against these predictions. The recall is noticeably higher than expected due to the inflated Microsoft score as mentioned previously. Looking at the results using the subset of the Brown Corpus we see this effect to an even greater extent. The precision varies approximately between the bounds.

6.5.3 Analysis

We now return to the questions put forward at the start of Section 6.4.

Is a well constructed default gazetteer a powerful enough resource for placename disambiguation?

The MR method uses only a default gazetteer constructed using statistics from the co-occurrence model. This method performs consistently well with respect to the expected bounds of performance and is always within 3% of the best performing method in all the performance measures. Despite other methods performing better one may question whether a user will notice the improvement (Keen 1992).

Also a two-tailed Sign test compared the MR method to the two other naïve methods and showed it to be statistically significantly better than both with a confidence greater than 99.9%. This shows applying context with simple rules (as with the Referents method) is not necessarily better than using no context at all.

Can methods employing statistics gathered from a corpus outperform a gazetteer alone?

A two-tailed Sign test compared the results of the MR method to the MI method. The MI method uses the feature type of locations rather than usage statistics to form a default gazetteer. The MR method was statistically significantly better than the MI method.

Is the co-occurrence of single placenames more important than the combined information of all co-occurring placenames?

Can supervised learning be more effective for placename disambiguation than simple hand constructed rules?

The SVM-tf-idf method was compared to all the other methods with a series of two-tailed pairwise Sign tests and was shown to be statistically significantly better than all other methods with a confidence greater than 97.5%. The fact that it outperformed the neighbourhood method shows that occurrences of all placenames are more important than the single occurrence of trigger placenames with respect to context. The fact that it outperformed the MI and MR methods shows that context contains a significant amount of information with respect to disambiguation and the fact that it outperformed the Referents method shows this context is difficult to apply in a naïve manner.

6.6 Indirect measurement

Some would argue achieving a statistically significant result disambiguating placenames on a ground truth is immaterial if these results do not translate into improved retrieval performance; and as noted in the previous section, although the presented improvements are statistically significant, they are also small. Zhang et al. (2007) show how a small improvement in performance in noun phrase classification can translate into a significant improvement in retrieval performance. I consider placename disambiguation an analogous task and hope to realise similar results.

This section begins with a description of the GIR system Forostar, developed specifically for this thesis. Forostar contains a disambiguator module that can apply any of the six disambiguation methods already examined as well as an additional *no disambiguation* method, which generates an ambiguous rather than a unique geographic index. GeoCLEF has become the standard corpus for assessing GIR systems; I analyse its distribution of placename references and compare it to the distribution of references in the co-occurrence model.

Continuing with the three naïve and three complex disambiguation methods of the previous section, two baseline methods are introduced, which provide an alternative to disambiguation. The GeoCLEF queries are discussed and the query formulation process described. This section concludes with full statistical analysis of the results.

6.6.1 Forostar

Forostar is the GIR system developed to perform the retrieval experiments shown in this Thesis. Forostar is split into two parts: the *indexing stage* and the *querying stage* (Figure 6.3). The indexing stage requires the corpus and some external resources to generate the geographic and text indexes (a *slow* task). The querying stage requires the generated indexes and the queries; it runs in real time.

The Indexing stage consists of four separate applications: *PediaCrawler* is the name given to the application building the geographic co-occurrence model described in Chapter 5. *Disambiguator* then applies the co-occurrence model, using one of the methods described in Section 6.4, to disambiguate the named entities extracted from the corpus by the *Named Entity Recogniser*. The disambiguated named entities form the geographic index. *Indexer* is used to build the text index.

Named entities are extracted from text using ANNIE, the Information Extraction engine bundled with the General Architecture for Text Engineering (GATE) (Cunningham et al. 2001). Text is indexed

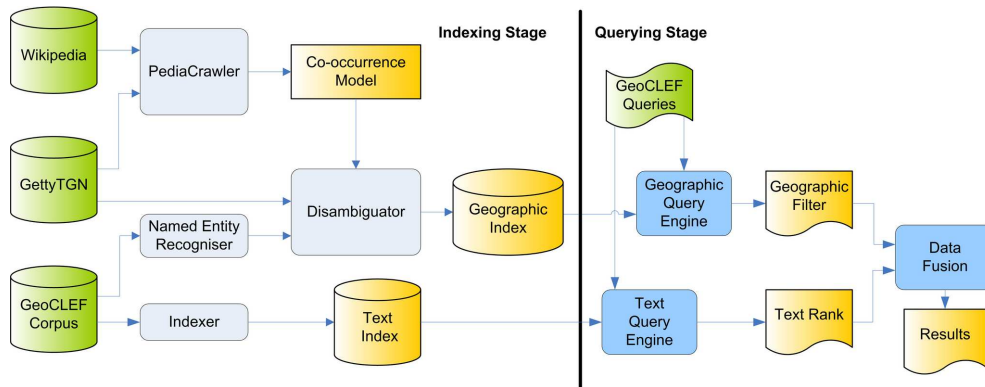


Figure 6.3: Forostar design

using Apache Lucene⁵ using the vector space model, with $tf \cdot idf$ term-weights.

The Querying stage consists of three parts, the *Text Query Engine*, *Geographic Query Engine* and *Data Fusion module*. The text and geographic indexes are queried separately. Placenames are manually extracted from queries and passed to the *Geographic Query Engine*, which produces a geographic filter. The text rank from the *Text Query Engine* is combined with the geographic filter by the *Data Fusion module*.

The *Geographic Query Engine* disambiguates each placename passed to it using the MR method of disambiguation⁶. The locations are indexed in such a way that a location look up includes all child locations. The implementation of this index is not provided here, instead I shall refer the reader to Overell et al. (2008b), suffice to say storing and searching locations in this fashion means all locations contained within a larger location are also included in a query: e.g. a query for “United States” will produce a filter including documents mentioning all the states, counties and towns within the United States as well as references to the country itself.

Forostar uses the standard Lucene *Text Query Engine*. This performs a comparison between the documents and the query in a $tf \cdot idf$ weighted vector space. The cosine distance is taken between the query vector and the document vectors.

The *Data Fusion module* combines the text rank and the geographic filter to produce a single result rank. Documents that match both the geographic filter and the text rank are returned first (as ranked against the text query). This is followed by the documents that hit just the text query. The tail of the results is filled with random documents. This method of filtering text documents against geographic documents is used by Vaid et al. (2005), Hauff et al. (2006) and Cardoso et al. (2007). An alternative, less aggressive form of data fusion based on penalising results not appearing in the geographic filter is examined in Appendix B, however, due to overfitting, the results were found to be significantly worse.

6.6.2 Distribution of placenames in the GeoCLEF collection

The GeoCLEF collection is described in detail in Section 3.2.1; it contains 135,000 news articles, taken from the Glasgow Herald and the Los Angeles Times. In this section we examine the distribution of placenames and locations throughout the collection.

GATE extracted 1.2 million placename references from the GeoCLEF collection. These are made up

⁵<http://lucene.apache.org/java/docs/>

⁶This method is chosen due to the minimal context contained in queries.

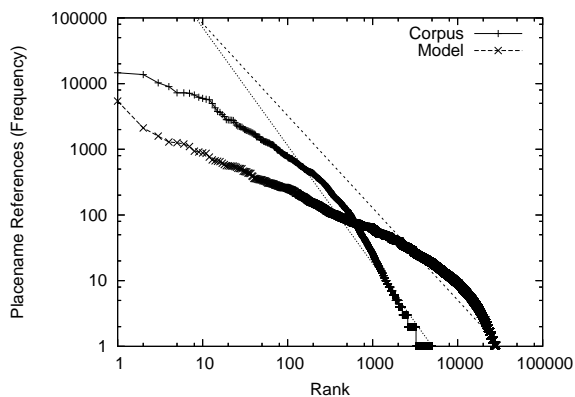


Figure 6.4: Frequency of placename references – rank

of 33,732 unique placenames, 5,361 of which are covered by our co-occurrence model. Although only 15.9% of the vocabulary is covered by our co-occurrence model, 79.7% of the references are. Note this is higher than the expected recall at the start of this chapter of 71% and concurs with the results of Section 6.5. This increase could potentially be inflated as it is measured against GATE’s output rather than a ground truth.

As with the co-occurrence model the placename references in the GeoCLEF corpus follow a Zipfian distribution. However as this is a news corpus, important locations are referred to particularly often and un-eventful locations hardly at all. This makes the Zipfian curve a lot steeper. 100 placenames ($\approx 0.3\%$) account for 610,093 references ($\approx 50\%$) and 20 placenames ($\approx 0.06\%$) accounts for 311,653 references ($\approx 25\%$). 613,727 ($\approx 50\%$) references are to placenames that are ambiguous with respect to our model. Figure 6.4 illustrates this distribution (alongside the normalised model distribution). A curve is fitted using least squares fitting in log-log space to both these distributions (shown as a line in log-log space). This allows us to quantify the difference between distributions by calculating the coefficient in the power law equation:

$$y = ax^k \tag{6.8}$$

For the co-occurrence model the k coefficient is -1.40, while for the corpus the k coefficient is -1.80. Note this graph is not comparable to 5.7 as it plots the placename distribution rather than location distribution.

To convert the placename references to location references and provide a map of the corpus distribution, I have run the MR method of disambiguation over the placename references in the corpus. This produces a location distribution that can be plotted in two dimensions as shown in Figure 6.5. By comparing the maps in Figures 5.8 and 6.5, one can see the GeoCLEF references are significantly more skewed. Note the red spots over Glasgow and Los Angeles (where the newspapers are published), also around the east coast of the United States.

The top 20 placenames (making up approximately 25% of references) are listed here: “Scotland,” “Los Angeles,” “California,” “U.S.,” “Glasgow,” “United States,” “Orange County,” “New York,” “Washington,” “America,” “Britain,” “Edinburgh,” “London,” “England,” “Europe,” “Japan,” “UK,” “Hollywood,” “Wilson” and “San Francisco”. One can see ambiguities and synonyms occurring in this list. “U.S.” and “United States” will generally resolve to the same location. “New York” is ambiguous whether

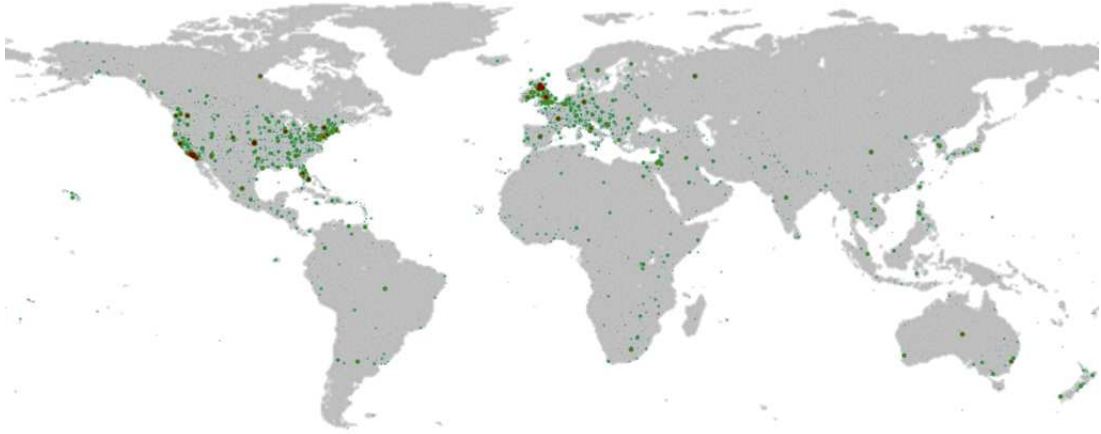


Figure 6.5: 2D projection of location references in the GeoCLEF corpus

it refers to the city or state, although an error in classification will make little material difference. More significant ambiguities in these top occurring locations are “Washington,” which could easily be Washington State or Washington D.C. dependent on context, and “America,” which could refer to the two continents the Americas, the continent North America or the country the United States of America (not to mention towns spread across the globe called America).

It is clear that the publications the GeoCLEF corpus is taken from are regional. Not only is this bias seen in Figure 6.5, but almost all the top 20 placenames are closely related to either Los Angeles or Glasgow.

6.6.3 Baseline methods

This section adds two more disambiguation methods to the six already presented. *Text Only* uses no geographic index at all, while *No Disambiguation* uses an ambiguous geographic index. This allows us to ask the following additional research questions:

- Is disambiguation needed at all?
- Can a unique geographic index outperform an ambiguous geographic index?

Text Only

The text only method is the current solution. Geographic entities remain un-parsed and un-annotated in the text part of the query and are treated as normal text; no geographic index is used. Any disambiguation relies on the user explicitly matching the context of the corpus documents with the query. The motivation of this method is to measure to what extent the geographic index affects the results.

No Disambiguation

In this method no disambiguation is performed. Each placename occurrence is indexed multiple times, once for each matching location in the co-occurrence model. For example if “London” appears in a document it will be added to the geographic index three times: London, UK, London, ON and London, CA. This is analogous to a text only index where extra weight is given to geographic references and

ID	Text part	Geographic part	TextNoGeo part
001	Shark Attacks off Australia and California	Australia, California	Shark Attacks
026	Wine regions around rivers in Europe	Europe	Wine regions around rivers
039	Russian troops in the southern Caucasus	Box((43,50),(39,40))	Russian troops
055	Deaths caused by avalanches occurring in Europe but not in the Alps	!Alps, Europe	Deaths caused by avalanches

Table 6.8: GeoCLEF queries and manually constructed query parts

synonyms expanded. The motivation behind this method is to maximise the recall of the geographic index.

6.6.4 Queries

The experiments are run across the 100 GeoCLEF queries of 2005-08 (25 from each year). Although some ambiguous placenames are referred to in the queries, it is always the most commonly referred to location referenced by said placename that is intended. Queries where a less common location for a placename is intended (such as London, Ontario) or placenames without a clearly dominant location (such as “Cambridge” or “Victoria”) are where an unambiguous geographic index will provide greatest improvement. As no such queries exist in the GeoCLEF query set, we must rely on implied ambiguous placenames to provide an improvement in performance. Implied locations are locations not explicitly mentioned in the query but which are considered relevant through implication (Li et al. 2006). For example London, ON would be considered geographically relevant to the query “Snowstorms in North America,” while London, UK would not.

For simplicity I discard the query narrative and manually split queries using only the title field. Query parsing is outside the scope of this thesis, and since 2007 has been seen as a separate task from the core GeoCLEF problem. Because of this, the geographic parts of the queries in this experiment are manually constructed. An example of some GeoCLEF queries and manually extracted geographic parts can be seen in Table 6.8. Notice that information between the *geographic part* of the query and the *text part* is duplicated. Because of this we have introduced an additional manual query formulation, TextNoGeo, which is the text part of the query with the geographic elements removed. The motivation behind the TextNoGeo formulation is that it is orthogonal from the geographic query and allows us to see the effects of the text and geographic facets of the query independently.

These three query parts are used to construct five query formulations:

- **Text.** This formulation is the standard IR solution at present, it uses the text part of the query only. This provides a baseline for the system.
- **Geo.** The geographic part of the query contains only a list of placenames and bounding boxes. It allows us to quantify the contribution of the geographic part of the query independently from other factors.
- **Text & Geo.** In this formulation the results of the text query are filtered against the geographic query. The ordering remains the same as the original text query results.
- **TextNoGeo.** The TextNoGeo formulation is a standard text query with all geographic elements removed.
- **TextNoGeo & Geo.** The advantage of the TextNoGeo method filtered against the geographic evidence is that the two methods are orthogonal.

Query classification has also been explored with respect to GIR (Gey et al. 2006; Mandl et al. 2007; Andogah and Bouma 2007; Leveling and Veiel 2006). A number of classification schema have been presented for the GeoCLEF query collection, these are summarised in Appendix B. I have searched for a correlation between a particular disambiguation method or query formulation with a particular classification. No correlation was found with respect to the classifications presented in Appendix B and the query formulations or disambiguation methods above. This is further discussed in Section 6.7.

6.6.5 Results

The results of the indirect measurement experiments are presented in three parts. We begin before any queries are executed with a comparison of the different geographic indexes. This is followed by a comparison of the different query formulations and disambiguation methods. The section concludes with the per-query results.

Unique geographic indexes

This section examines the differences in classifications between the disambiguation methods producing a unique geographic index on both a micro and macro level. We begin by looking at the macro scale. 79.2% of placename references in the GeoCLEF collection are covered by the Getty Thesaurus of Geographical Names and hence are classified by the MI method. Slightly fewer locations are covered by the other methods as they only contain locations matched to Wikipedia articles, and therefore contained in the co-occurrence model. 77.5% of placename references are covered by the co-occurrence model. This means the MI method can disambiguate locations not mentioned or mis-classified in Wikipedia.

Table 6.9 shows the percentage overlap between the different classification methods. The most striking observation from this table is how different the MI method is from the other five methods. The reason behind this should become clearer when we look at the micro-scale differences. The similarity between the other four methods is expected; they default to the MR method, classifying non-ambiguous placenames the same.

At the start of this chapter we observed that 64.9% of placename references in the co-occurrence model are unambiguous. As we saw in Figure 6.4, there is a higher proportion of references to *important* places in the GeoCLEF corpus than the co-occurrence model. Due to the fact that people like to name newly established locations after important locations this leads to GeoCLEF having a higher than expected proportion of ambiguous placenames. This problem is compounded by the regional nature of the corpus: references are concentrated in the Los Angeles and Glasgow areas (as seen in Figure 6.5). Both these regions have a high proportion of ambiguous names. In fact only 49.0% of placename references in the GeoCLEF corpus are unambiguous with respect to our model. Given such a high proportion of references to ambiguous placenames one would expect this to be ideal for testing context based disambiguation methods. Unfortunately this intuition does not ring true, as the significant majority of these references are to the most common location with that name.

Essentially the four context based methods are diverging from the MR method. As they are all making classifications based on the same contextual information, one would expect them to diverge in the same direction. As is shown in Table 6.9, this is only true in one case, the two SVM methods. It is an interesting observation here that the different information captured by the two feature spaces do not cause a greater difference.

Table 6.10 shows the differences between placename classifications on a micro-level. The table shows the most commonly occurring classifications made by each method that do not occur in the MR default

	MI	MR	Referents	Neigh.	SVM-tf-idf
SVM-prox	47.6	91.0	86.8	82.9	93.2
SVM-tf-idf	47.9	88.6	84.5	82.0	
Neigh.	44.2	85.5	81.8		
Referents	47.2	93.6			
MR	48.8				

Table 6.9: Percentage overlap of identical classifications

MI	Referents	Neigh.
Los Angeles County	Los Angeles County	United Kingdom
Midlothian, Scotland	New York County	Union Station, Chicago
San Francisco County, California	Aberdeenshire, Scotland	New York City, New York
Aberdeenshire, Scotland	Colorado City, Arizona	Mount Whitney, California
Sacramento County, California	San Francisco County, California	Aberdeen, Scotland
SVM-tf-idf	SVM-prox	
Aberdeen, Scotland	Aberdeen, Scotland	
Pacific Ocean	Pacific Ocean	
Mississippi River	City of Westminster, London	
Milwaukee County, Wisconsin	New York City	
Perth, Scotland	Mississippi River	

Table 6.10: Top locations different from the MR method

gazetteer. Here we can see one of the root causes of the MI method’s differences. Often a county will share the name of its capital city or town, the MI method considers counties of greater importance than cities or towns as they encompass them. Four of the top occurring divergences from the MR method come from this difference in classification. Although in theory this is a significant difference in classification (and arguably often wrong), in practice this will make little difference in a GIR system (as the centre of Los Angeles City and the centre of Los Angeles County are relatively close).

The same difference in classification is also commonly made by the Referents method, but to a significantly lesser degree. The other three context based methods share a number of common locations: Aberdeen, Scotland (confused with Aberdeen, Washington), New York City and Mississippi River (confused with New York State and Mississippi State).

Query formulations and disambiguation methods

In this section we compare the different query formulations against the different disambiguation methods. The motivation for this is to find what information is captured by the different query parts and geographic indexes, and the best way to combine them to achieve synergy. Table 6.11 presents the results. The Text and TextNoGeo methods are the same for every disambiguation method as no geographic index is used.

There are several interesting observations we can make from this table. The MR method has the highest MAP using a geographic index only. However when this is paired with an orthogonal text index (TextNoGeo) to form the TextNoGeo & Geo method, the Neighbourhoods method has the highest MAP.

	MR	NoDis	MI	Referents	Neigh.	SVM-tf-idf	SVM-prox
Text	24.1	24.1	24.1	24.1	24.1	24.1	24.1
Geo	3.1	2.9	2.00	3.0	2.9	2.8	3.0
Text & Geo	24.3	24.9	24.3	24.1	24.4	24.3	24.1
TextNoGeo	11.5	11.5	11.5	11.5	11.5	11.5	11.5
TextNoGeo & Geo	22.6	22.6	20.6	22.4	22.7	22.5	22.4

Table 6.11: Query formulations against MAP(%)

	Text	MR	NoDis	MI	Referents	Neigh.	SVM-tf-idf	SVM-prox
2005	29.2	30.4	30.6	29.3	20.2	30.8	30.7	30.4
2006	23.1	21.0	20.7	17.4	20.2	21.2	20.8	20.3
2007	19.2	20.2	21.7	21.5	23.6	23.7	20.2	21.0
2008	24.1	24.5	26.5	27.8	24.5	24.3	24.4	24.5

Table 6.12: Summary of per-query results — MAP(%)

I have no explanation for this counter-intuitive result, however, as the MAP for the Geo method is so low, it is possible these results are not reliable. In fact when we take a second look at the geographic similarity measure we see it is binary, this means the ordering of results considered positive is arbitrary. Strictly speaking this makes MAP an inappropriate measure, however I present it here for comparison.

There is a strict ordering across the formulations with no result in one formulation performing better than the next (with two exceptions, Text & Geo: Referents and SVM-prox). The ordering is as follows: *Geo*, *TextNoGeo*, *TextNoGeo & Geo*, *Text*, and *Text & Geo*. This is the expected result showing that in general as more information is provided the MAP improves. The only surprise is that the Text method is better than the TextNoGeo & Geo method. This implies more information about a query is captured in a placename than a location; this will be further discussed in Section 6.7. In both query formulations combining textual and geographic information, the Neighbourhood method performs best out of the methods with a unique geographic index. Based on the previous section, I would have expected the SVM-tf-idf method to perform better; in the next section we will see if this difference is significant.

Per-query results

Appendix B details the per-query results, these are summarised in Table 6.12, showing the results for the 25 queries from each year. The summary statistics can be compared to Table 6.13, which shows the per year quartile ranges. The Text formulation is provided for comparison; the Text & Geo query formulation is used for all other runs.

The 2005 results are all between the 3rd quartile and the best result, while 2006-08 results tend to occur between the median and the 3rd quartile. The runs presented in this chapter were not included in the CLEF results pooling, except for the 2008 results. In 2005-07 the best method using a geographic index is the Neighbourhoods method. In 2008 the MI method performs surprisingly well; on close inspection this is caused by queries 76 and 91 – these queries refer to South America and Spanish Islands. My explanation for the poor performance of the other methods for these queries is that they require accurately disambiguating placenames of Spanish origin; a lot of these locations share their names with locations in the United States, so the United States bias of Wikipedia could be coming into effect.

Text only beats the other seven methods in 2006, this is entirely due to its superior performance on query 49, on ETA attacks in France. On closer inspection of this query there are in fact only two documents in the corpus judged relevant. The first of these documents describes ETA attacks in Palma de Majorca, so is arguably a false positive from the assessors. The second document refers to “Southern France,” which Forostar incorrectly recognises as a two-word placename (like “Southern Ocean”) rather than a placename preceded by a modifier. Because of this syntactic ambiguity error, Forostar fails to correctly ground the placename to a region of France. The results returned by the methods including geographic evidence are largely documents referring to “France” and locations in France. The text results included the correct document first giving it an AP of over 50%. This query shows some fragility of Forostar, however I am wary to draw conclusions from a query with only a single correct result.

	Worst	Q1	Median	Q3	Best
2005	5.69	12.99	19.23	27.94	39.36
2006	4.00	15.64	21.62	24.59	32.23
2007	1.13	11.98	17.61	22.92	27.87
2008	16.07	21.39	23.74	26.12	30.37

Table 6.13: GeoCLEF quartile range results — MAP(%)

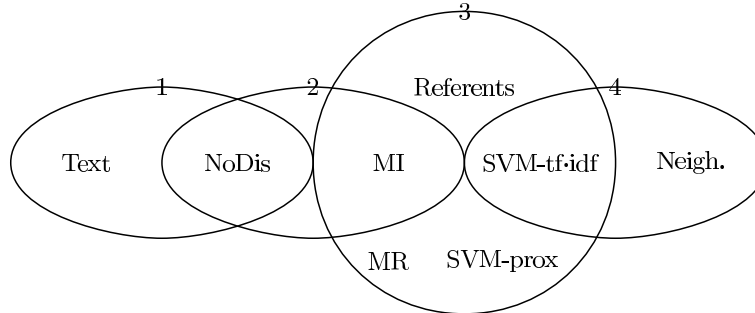


Figure 6.6: Overlapping groups of retrieval methods

6.6.6 Analysis

Two statistical tests were performed on the per-query results. Initially a Friedman test to partition the methods into groups with no infra-group significant difference. This is followed by a two-tailed Wilcoxon Signed-Rank test to provide an ordering across the groups.

The methods were partitioned into four overlapping groups:

1. Text, NoDis
2. NoDis, MI
3. MI, MR, Referents, SVM-tf-idf, SVM-prox
4. SVM-tf-idf, Neigh.

Wilcoxon Signed-Rank tests put the following orderings across the groups:

$$1 \leq 2 \leq 3 \leq 4, \quad (6.9)$$

$$1 < 3, \quad (6.10)$$

$$2 < 4, \quad (6.11)$$

$$3 - 4 < 4 - 3, \quad (6.12)$$

$$2 - 3 < 3 - 2. \quad (6.13)$$

This is illustrated in Figure 6.6. Essentially the worst methods are the Text and NoDis methods and the best methods are the SVM-tf-idf and Neighbourhoods methods. This may seem counter intuitive as the NoDis method had a relatively high MAP. It is the lack of consistency of the NoDis method that accounts for its statistically poor performance.

Now we return to the research questions asked earlier in the chapter:

Is a well constructed default gazetteer a powerful enough resource for placename disambiguation?

The short answer to this question is “yes,” with the caveat “in most circumstances.” Context based disambiguation methods have equalled or statistically significantly improved over the default gazetteer based methods, but the improvement is small. In situations where the best possible performance is critical then context based methods are certainly the most appropriate.

Can methods employing statistics gathered from a corpus outperform a gazetteer alone?

The MI method and NoDis method used a gazetteer with no or minimal corpus statistics respectively. Despite performing well in certain situations they have been consistently beaten by other methods.

Is the co-occurrence of single placenames more important than the combined information of all co-occurring placenames?

Despite the Neighbourhood method performing better than most other methods there was not a statistically significant difference between it and the SVM-tf-idf method. Because of this, the conclusion of the previous experiment still holds (although is not strengthened).

Can supervised learning be more effective for placename disambiguation than simple hand constructed rules?

In both our indirect and direct evaluation experiments, supervised learning methods have statistically significantly beaten other methods, however a single method has not revealed itself as the best. My conclusion is supervised learning can be more effective than hand constructed rules, however there is room for improvement to find the best use of contextual evidence.

Is disambiguation needed at all?

Can a unique geographic index outperform an ambiguous geographic index?

The Text and NoDis methods were the two methods using no disambiguation at all. The NoDis method also used an ambiguous geographic index compared to the other methods’ unique indexes. The text method performed badly in comparison to the other methods apart from a very few queries. The NoDis method, as expected, had a particularly high recall, this gave it very good performance on some queries and a relatively high MAP, but overall it was statistically significantly worse than the data-driven methods.

6.7 Discussion

This chapter shows that data fusion between a text rank and a filter generated with a unique geographic index can significantly improve over text retrieval alone but we are left with some questions. The query formulation experiments showed information is being contributed from both the placename and the location part of a geographic query; also quite an aggressive approach to data fusion is taken here, could a more sophisticated approach allow us to capture the occasions where the Text and NoDis methods perform well but keep the precision of the more complex methods?

6.7.1 Is there more information in a placename or a location?

Referring back to the original definitions in the Chapter 1, a location is a space on the Earth’s surface usually bounded by polygons, and a placename is a phrase used to refer to a location. More pragmatically, in a GIR system (such as Forostar presented here), a location is a unique identifier corresponding to a key in a gazetteer that maps to a unique machine readable description of a location. These identifiers make up the unique geographic index (where they map to documents) and appear in queries. A placename is a list of tokens stored in a text index or appearing in a query, a human would understand these tokens as referring to a location.

The best retrieval results occurred when both the placename and location were included in the query. Intuitively information is provided by both parts. For example consider the placename, location tuple {Islas Malvinas, 7005151}. The placename “Islas Malvinas” refers to two locations, it is the Spanish name for the Falkland Islands (and implies Argentine sovereignty), and a group of rocky islands off the coast of Ibiza. Location identifier 7005151 refers to the archipelago in the South Atlantic Ocean, recognised by its inhabitation as the “Falkland Islands” and under British sovereignty. Provided with both the placename and location, a geographically-aware search engine can return initially, documents referring to the islands as the “Malvinas” (presumably the most relevant documents), followed by references to the islands as the “Falklands” (which may also be relevant but to a lesser degree). This is a particularly extreme example but shows that both a placename and a location capture information about a user’s intention. Although this is a possible explanation for the improved results it seems unlikely this would produce an observable difference. A more plausible explanation would be the placename part of the query softens some of the errors in the geographic index. Consider the location tuple {Cambridge, 7010874}, where 7010874 is the identifier for the city of Cambridge, UK. One would expect a small proportion of references to Cambridge, UK in the corpus to be misclassified as Cambridge, MA. Documents containing these misclassified locations will still be picked up by the placename part of the query and displayed after documents containing the correctly classified locations. If the query contains other disambiguating information (e.g. a reference to punting), one would expect documents referring to Cambridge, UK to appear above references to Cambridge, MA regardless of the geographic index.

The final reason why improved performance is observed when both the placename and location are used is that in the geographic index as implemented in Forostar, all matches to the geographic index are given equal weighting. If the placename is included in the query as well, references to the location mentioned in the query are considered more relevant than subparts of that location and multiple references to that location are given even more relevance due to Lucene’s *tf·idf* vector space. Consider a query for {France, 1000070}, where identifier 1000070 refers to the corresponding European country. A geographic filter will give any reference to France or a location within France equal weighting, e.g. Paris, Corsica or Brittany. Documents ranked with respect to the placename will give a higher ranking to documents mentioning France and higher still to multiple mentions of France. These documents will, potentially, be more relevant to the query.

6.7.2 Combining methods – the search for synergy

Appendix B shows two experiments conducted as part of the development of Forostar in an attempt to fuse geographic and textual retrieval in the most effective way. The results were somewhat disappointing: query classification provided no advantage, and a more sophisticated data fusion technique based on a trained penalisation model actually gave worse results than a text baseline. This cautionary tale of over fitting shows the sensitivity of fusing text and geographic relevance. Due to the current immaturity of geographic relevance ranking techniques it is my conclusion that robust methods such as filtering are the most appropriate for GIR systems. Both text and multimedia retrieval have had much more promising data fusion results (Craswell et al. 1999; Fox and Shaw 1994; Wilkins et al. 2006). Wilkins et al. (2006) propose a method of adjusting the weightings given to different features based on the distribution of scores at query time. Such a method may also be appropriate for geographic information retrieval and could help with the future search for synergy between textual and geographic evidence.

Chapter 7

The world according to Wikipedia

7.1 Introduction

The philosopher Kant maintained that we visualise and reason about space in Euclidean geometry (Montello 1992). This hypothesis has not stood up to experimental analysis. Current thinking suggests that space is reasoned about topographically (Egenhofer and Mark 1995; Montello 1992). Egenhofer and Mark (1995)'s notion of Naive Geography is concerned with formal models of the common-sense geographic world and how this differs from the physical world. People conceptualise their world view on multiple different scales. Worboys (1996) observes that distances must be modelled differently for different people dependent on a person's location. Montello (1992) observes people do not reason about locations in a metric space. They observe that subjects generally have different internal representations of standard distances (a set of *mental yardsticks*) as a contributing factor to a multi-scale world view.

This chapter attempts to quantify the world view of Wikipedians speaking different languages. Initially the combined world views of all the speakers of a language are visualised and systematic biases quantified. This is then generalised to model the world view of a single speaker. The pipeline described in Chapter 5 is applied to versions of Wikipedia in languages other than English. A model is built for each language. The full pipeline is necessary for these experiments rather than just one source of evidence for two reasons:

- By matching locations to an authoritative source (the TGN), we have consistent geographic annotations matched to all locations and can filter out noise. This also allows us to map between the same location occurring in different co-occurrence models.
- By taking advantage of all sources of evidence we can maximise the coverage of the models.

Essentially the disambiguation pipeline maximises the precision and recall of the models produced.

This chapter begins by exploring the background of how people reason about physical space and Wikipedia's multi-lingual aim. I then apply the disambiguation pipeline to versions of Wikipedia in languages other than English. This provides us with a set of per-language co-occurrence models. In Section 7.4 I look at the distribution of locations in each of these co-occurrence models and provide a quantitative comparison. Section 7.5 continues this comparison examining Wikipedia's systematic bias and looking at both quantitative and qualitative differences at the micro and macro levels.

Section 7.6, instead of looking at the *differences* between wikipedias and Wikipedians, attempts to quantify the *invariant* nature of how people refer to locations. This chapter presents the hypothesis

that all people have the same view of the world with respect to their locality. Six plausible quantitative models of varying complexity are compared in an attempt to model the likelihood of a specific person to refer to a specific location. The final experiments of this thesis are in Section 7.7, where methods developed for extracting and analysing spatial data are applied to temporal data. Analysis of spatial data is highly coupled with the analysis of temporal data: many of the same techniques are used for data extraction, as temporal references have to be recognised and grounded to a specific point or period in time; and many applications are similar as when browsing corpora relating to events (e.g. news or photo corpora) both when and where an event takes place are integral to the meaning and relevance of a document.

7.2 Everything is related to everything else...

Waldo Tobler's **first law of geography** is *Everything is related to everything else, but near things are more related than distant things*. This sentiment is common in geographic representation, analysis and modelling people's view of the world. Mehler et al. (2006) analyse local newspapers from across the US showing *people from different places talk about different things*. Essentially people are largely interested in local concerns that effect them. Mei et al. (2006) make similar observations analysing blogs. Liu and Birnbaum (2008) exploit the fact that people's world view varies with locality in their application LocalSavvy, which aggregates news sources published in locations relevant to the respective story. In this chapter I reverse the problem as approached by Mehler et al. and Mei et al. They analyse the distribution of references to a single concept from multiple sources where the source is associated with a known location (the town where a newspaper is published or blog is written). I analyse references from a single source to multiple concepts (placenames), where the concepts are associated with locations.

How people think and reason about the world and how this effects one's world view is a subject that has long been of interest to philosophers, psychologists and geographers, but only more recently to information retrieval and data mining. Montello (1992) provides a review of the theoretical and empirical work on how people reason about space. Historically it was believed that people conceptualised the space around them as a metric space, however empirical evidence shows this is not the case. Egenhofer and Mark (1995)'s premise *topology matters, metrics refine*, consider topology the primary way that people reason about space, resorting to metrics only for refinements. These refinements are often inaccurate, particularly when compared to how well topological information is retained.

We have observed that people are primarily concerned with their surroundings; but how can this be quantified? As well as measuring the skew in the location distributions and the bias of each version of Wikipedia toward and against speakers of its own language, I also consider the spatial autocorrelation. Spatial autocorrelation can be considered a measure of how correlated a collection of geographically distributed values are (Cliff and Ord 1970). Spatial autocorrelation is inherently different from standard statistical testing because points related in a physical space cannot be considered independent, similarly it is different from temporal autocorrelation, which is one-dimensional with a clear division between past and future. Mehler et al. (2006) and Brunner and Purves (2008) observe locations referenced in news corpora exhibit spatial autocorrelation, this concurs with my observations of the GeoCLEF corpus in the previous chapter. I expect Wikipedia to exhibit a similar correlation, however what is of interest in this chapter is the degree of bias and correlation.

Language	English	German	French	Polish	Portuguese	Spanish
No. of Articles	3,264,598	1,131,773	1,397,808	540,851	700,241	483,459
Language	Japanese	Russian	Chinese	Arabic	Hebrew	Welsh
No. of Articles	301,868	376,784	287,873	85,918	133,389	16,343

Table 7.1: Language versions of Wikipedia being considered — note the number of articles includes stubs

7.3 Alternate language versions of Wikipedia

Wikipedia is currently available in over 250 languages. Of these versions 15 have over 100,000 articles¹, 77 have over 10,000 articles and 153 have over 1000 articles². For the experiments in this chapter, I have only considered versions of Wikipedia with over 10,000 articles as these tend to be active encyclopædias that people are using and editing.

Each language edition of Wikipedia is entirely self contained running independent versions of MediaWiki. They are accessible through the URL `XX.wikipedia.org` where `XX` is the corresponding language code, for example the English Wikipedia is situated at `en.wikipedia.org` and the German at `de.wikipedia.org`. All the language versions are accessible from the Wikipedia home page, `www.wikipedia.org`. The language versions are linked through inter-language links. Inter-language links are embedded in the article source and take the form `[[XX:Title]]`, where `Title` is the title of the article in the target language. For example the `Hippopotamus` article in the English Wikipedia links to the German Wikipedia with the following link: `[[de:Flusspferd]]`.

In this chapter I compare 12 versions of Wikipedia, listed in the Table 7.1. English, German, French, Polish and Japanese were chosen because they are the five languages with the most full-articles. Portuguese and Spanish were chosen to compare two similar languages that are very widely spoken. Russian, Chinese, Arabic and Hebrew were chosen to compare languages from different character sets. Arabic and Hebrew have the additional interest that they are spoken in countries currently in conflict. Welsh was chosen as an example of a minority wikipedia; Welsh is of additional interest because it is largely spoken in a very localised area by a small community. Additionally, Appendix C considers constructed languages, specifically Esperanto. When referring to versions of Wikipedia I will use Spanish Wikipedia as a shorthand for Spanish Language Wikipedia etc.

When presenting detailed analysis of the invariant nature of placename references, Section 7.6 will be restricted to only eight languages: English, German, French, Portuguese, Spanish, Chinese, Arabic and Hebrew.

7.4 Disambiguating locations in alternative languages — a modified pipeline

To disambiguate placenames in languages other than English the three stage process and disambiguation pipeline described in Chapter 5 is kept with a few slight modifications. An additional class of evidence is considered: inter-language links to the English Wikipedia. For example the article in the German Wikipedia *München* has an inter-language link to the English Wikipedia article *Munich*, which in turn has been disambiguated as location 7004333. In this case we consider this evidence that the German *München* article refers to location 7004333. Links to language versions of Wikipedia other than English are not considered as evidence because errors introduced this way could potentially propagate and compound.

¹For reference the full Encyclopædia Britannica contains approximately 100,000 articles across 32 volumes

²Correct as of November 2008

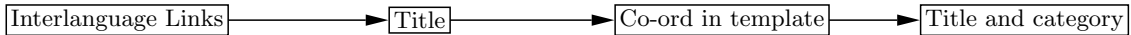


Figure 7.1: Alternate language pipeline

	English	German	French	Polish	Portuguese	Spanish
Links Extracted	7,317,469	2,334,779	2,364,609	3,086,297	1,606,602	2,522,518
prop Loc (%)	26.7	9.1	7.3	6.1	5.4	7.1
prop Ambig (%)	35.1	12.9	9.0	7.8	7.8	18.6
Articles disambig	59,341	23,607	18,332	13,441	13,686	9,664
Unique placenames	63,233	23,779	15,869	13,345	7,882	10,980
prop Ambig (%)	7.8	3.8	2.2	1.2	2.3	2.7
Unique Locations	50,694	17,791	12,261	6,351	6,068	7,244
	Japanese	Russian	Chinese	Arabic	Hebrew	Welsh
Links Extracted	5,419,486	104,019	2,813,221	1,619,460	3,209,710	702,262
prop Loc (%)	1.9	5.0	2.9	3.8	3.2	3.5
prop Ambig (%)	0.1	1.1	0.6	5.2	2.0	1.3
Articles disambig	2,092	5,049	2,259	1,864	1,534	1,086
Unique placenames	3,226	8,628	3,552	2,860	2,162	1,955
prop Ambig (%)	0.3	0.3	0.4	0.6	0.5	0.7
Unique Locations	1,842	3,820	1,953	1,399	826	938

Table 7.2: Summary of co-occurrence models

The default-locations class of evidence is discarded because this would involve annotating 150 articles from each language version of Wikipedia by a user with at least a limited working proficiency in that language. Although this is feasible for one language version of Wikipedia, it does not scale. Placeopedia.com is also discarded because its annotations are only available for the English Wikipedia. The other classes of evidence are applicable to all languages using the Latin character set. The gazetteer used, the TGN, contains the native names of locations and a limited number of foreign terms for locations in European languages (e.g. Munich → München, Lisbon → Lisboa, and Londres → London). This gives us the *Alternate Language Pipeline*, pictured in Figure 7.1. For non-Latin character sets only the inter-language Links are used. Note that it is only our gazetteer that restricts the pipeline to Latin characters. Given gazetteers in Cyrillic, Japanese, Arabic etc. the whole pipeline could be applied to each language.

As with the English Wikipedia, 100,000 randomly selected articles are crawled for each language (except Arabic and Welsh where all the available data is crawled). A summary of the co-occurrence models generated for each language is shown in table 7.2.

Distribution

Figures 7.2 and 7.3 plot these co-occurrence models in two dimensions as heat maps in the same fashion as Figure 5.8³. Locations which get a high proportion of references are shown in red, while locations with less references are shown in yellow and then green, and finally blue for very few references. Grey/white indicates areas with no references. The systematic bias across the different language versions of Wikipedia is plain to see, for example notice in the German and French Wikipedias the clear bias toward Germany and France respectively. The English Wikipedia has noticeably greater coverage than any other language. This can be partially attributed to the fact that English has become the *Lingua Franca* of the Internet

³In fact Figure 7.2:English is the same as Figure 5.8 repeated here for comparison.

Language	English	German	French	Polish	Portuguese	Spanish
k	-0.90	-1.0	-1.10	-1.23	-1.15	-1.12
I	0.326	0.371	0.585	0.633	0.509	0.581
Language	Japanese	Russian	Chinese	Arabic	Hebrew	Welsh
k	-1.23	-1.14	-1.20	-1.24	-1.35	-0.88
I	0.808	0.860	0.948	0.782	0.950	0.746

Table 7.3: Co-efficient k of the Zipfian distributions and spatial autocorrelation of the different wikipeidias

and many non-native English speakers will use the English Wikipedia⁴.

In Section 5.7.1 we observed that location references in the English Wikipedia occur in a Zipfian distribution. This is not surprising as references to locations are analogous to references to terms in a corpus, which have previously been observed to follow a Zipfian distribution (Nakayama et al. 2008; Kucera and Francis 1967). Figures 7.4 and 7.5 show the references against rank plots for the different language versions of Wikipedia.

A power law distribution was fitted in log-log space to the Zipfian distributions:

$$y = ax^k, \quad (7.1)$$

where a is the scaling factor relating to the size of the sample and k is the scaling exponent. This equates to a straight line in log-log space, as shown on the graph. The distribution was fitted using least-squares in log-log space. Locations with less than 10 references were ignored (considered outliers) for the purposes of fitting. The angle of the fitted line (the scaling exponent) is shown in Table 7.3.

As well as measuring the skew in the distributions, we can also measure the correlation. In the following experiment I measure the spatial autocorrelation of the maps presented in Figures 7.2 and 7.3. Initially the data is quantised grouping all references to locations within the same country. The spatial autocorrelation measure used is Moran's I , which varies between -1 and 1:

$$I = \frac{|\mathcal{C}| \sum_{i \in \mathcal{C}} \sum_{j \in \mathcal{C}} W_{i,j} (\text{ref}(i) - \overline{R_{\mathcal{C}}}) (\text{ref}(j) - \overline{R_{\mathcal{C}}})}{(\sum_{i \in \mathcal{C}} \sum_{j \in \mathcal{C}} W_{i,j}) \sum_{i \in \mathcal{C}} (\text{ref}(i) - \overline{R_{\mathcal{C}}})^2} \quad (7.2)$$

where \mathcal{C} is the set of countries, $\overline{R_{\mathcal{C}}}$ is the average number of references to a country and $W_{i,j}$ is the inverse of the distance from i to j ,

$$W_{i,j} = \frac{1}{\text{dist}(i,j)}. \quad (7.3)$$

Table 7.3 shows the degree of spatial autocorrelation for the different language versions of Wikipedia. While the Zipfian scaling exponent illustrates how references are distributed amongst locations of varying importance, I shows to what degree references are randomly distributed across the globe. With respect to Figures 7.2 and 7.3, the exponent can be considered a measure of the ratio of red to green, while I is a measure of the distribution of colour across the maps.

How close the scaling exponent and spatial auto correlation are to zero shows how egalitarian the version of Wikipedia is with respect to locations referenced, i.e. a more egalitarian Wikipedia will have its references more evenly distributed with respect to the size and geographic distribution of locations. Note English and German have the most egalitarian distributions. Hebrew and Japanese have distributions most skewed toward *important* and *spatially autocorrelated* locations.

⁴55% of all traffic directed at Wikipedia is to the English Language site.

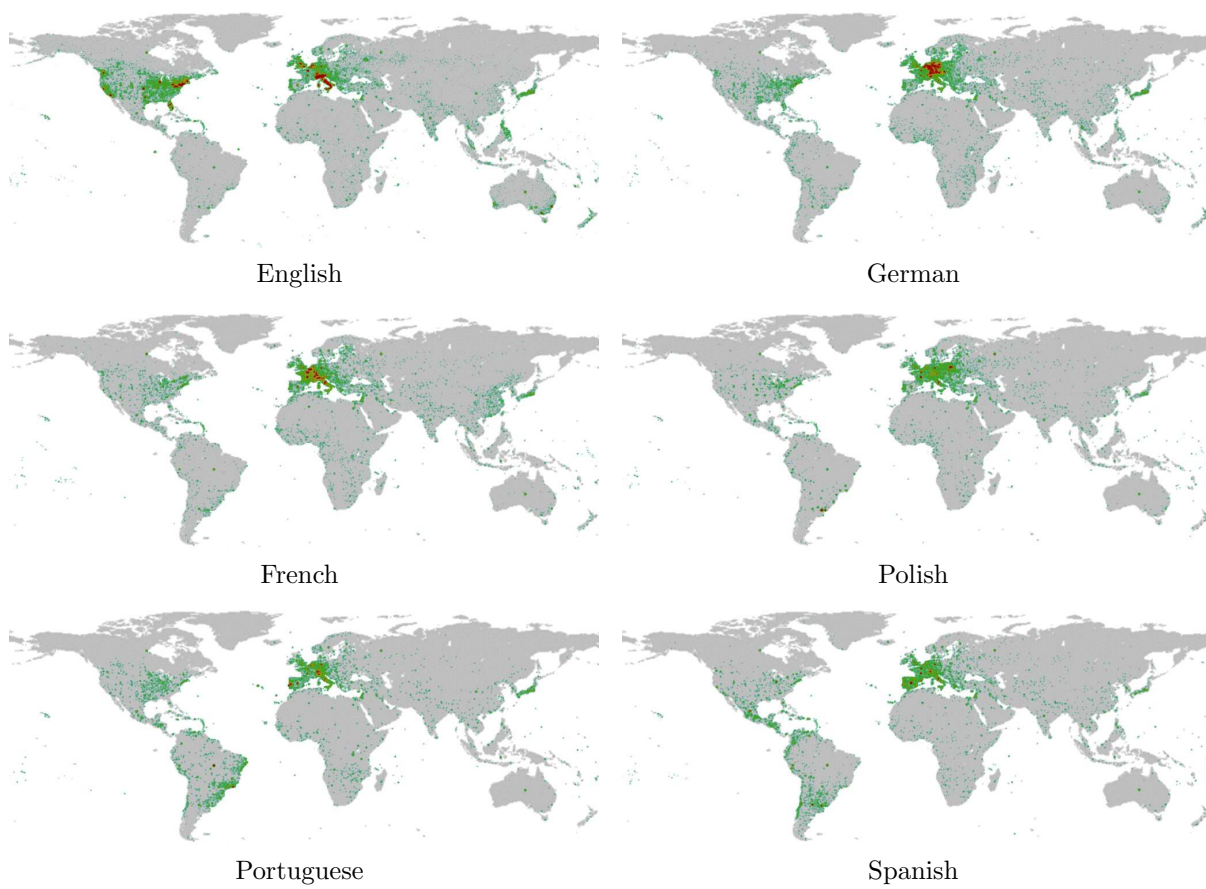


Figure 7.2: Heat maps in different wikipedias

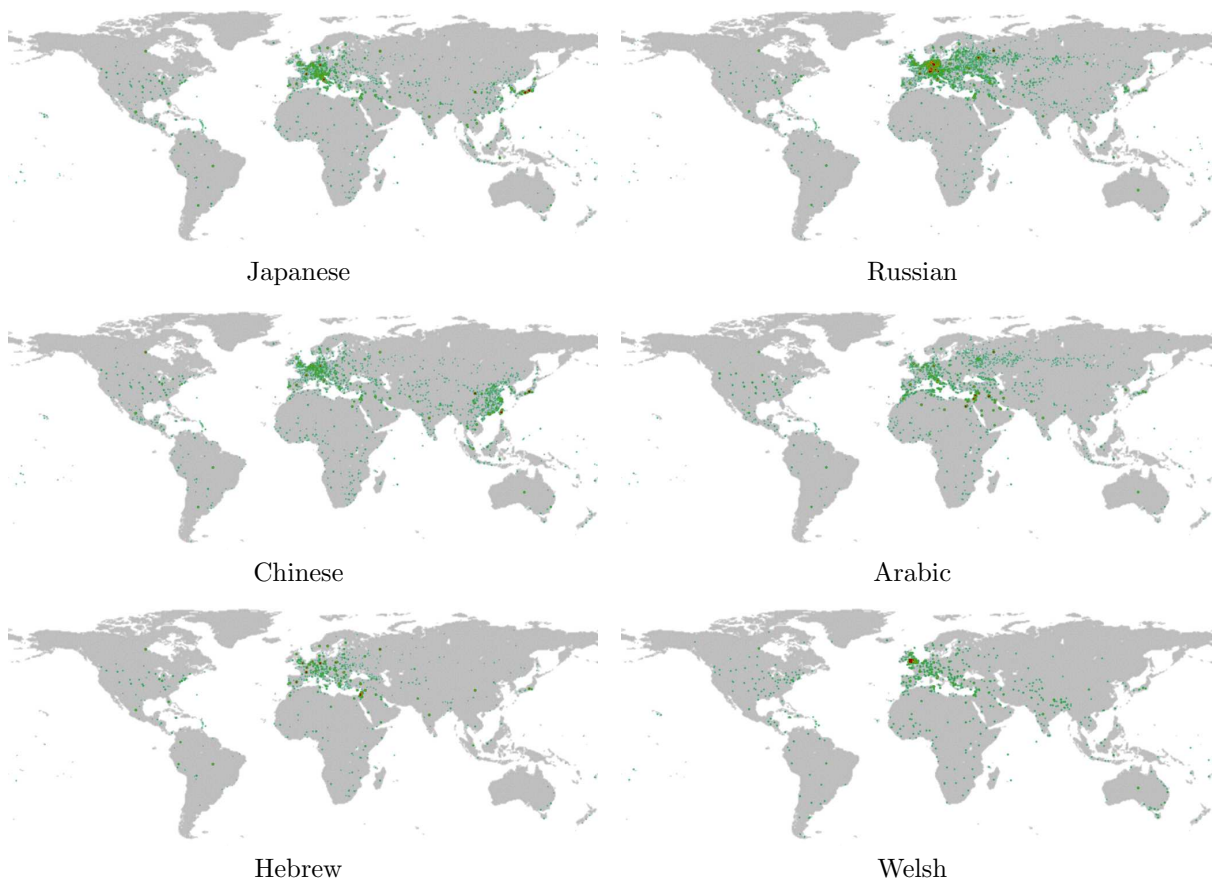


Figure 7.3: Heat maps in different wikis

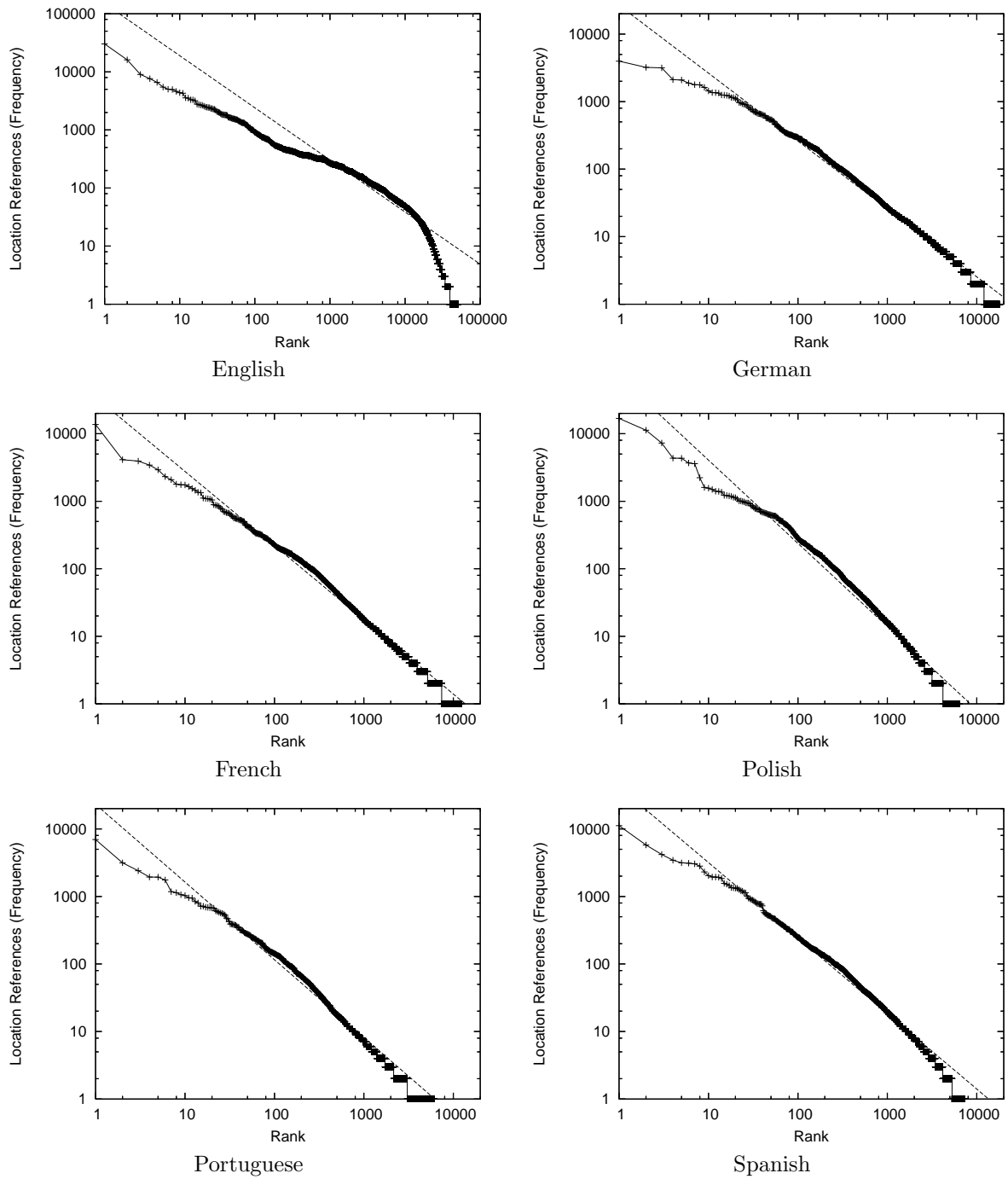


Figure 7.4: Distribution of locations in the different wikis

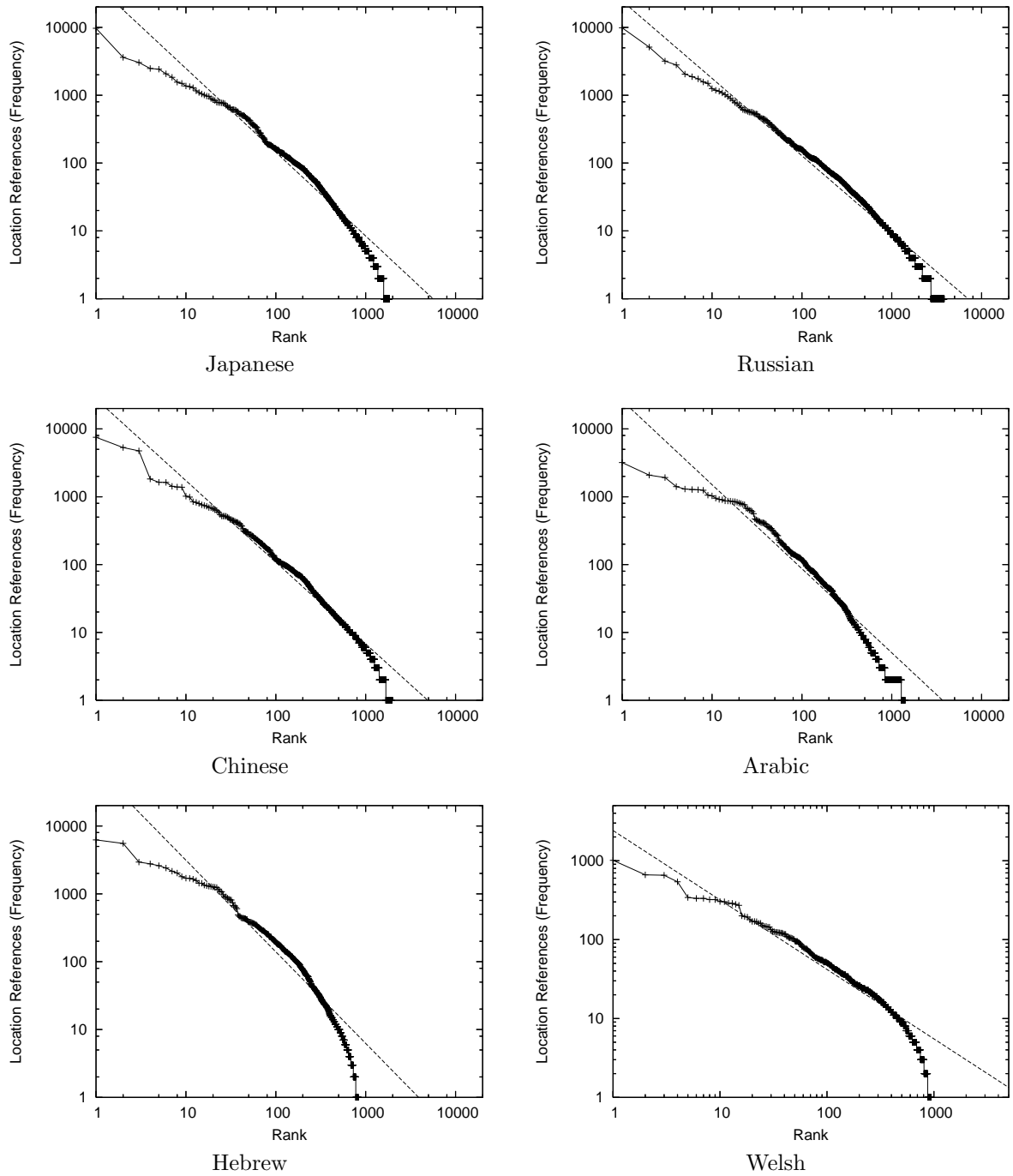


Figure 7.5: Distribution of locations in the different wikis

7.5 Bias

Wikipedia tries to maintain a neutral point of view, in fact this is one of the key policies guiding Wikipedia's content. However you will notice in the previous maps that regions where the respective wikipedia's language is spoken have a disproportionately larger volume of references. This is not contrary to Wikipedia's neutral point of view, so long as only facts rather than opinion are mentioned. This is referred to in Wikipedia as *Systematic Bias*⁵ and occurs where significant omissions across multiple articles exist. Wikipedia accept that the English speaking Wikipedia reflects the concerns of the English speaking world. This is partially due to sources available in English and the views of the editors. A symptom of this is that only 2% of featured articles relate to Africa, a continent that accounts for 14% of the world population and 20% of its area.

7.5.1 Quantitative analysis

I shall now attempt to quantify Wikipedia's systematic bias. The skew in the models can be seen in Table 7.4. For each wikipedia I divide the world into two parts, locations where the respective language is spoken and locations where it is not. I then calculate the number of references per person in each division⁶. How biased a particular wikipedia is toward speakers of its language can be considered the ratio of these two numbers. It indicates how many times more likely a location where the respective language is spoken is to be referenced over a location of equal population where the language is not spoken. For example the English Wikipedia has a bias of 13.15; one could interpret this to mean that given two locations of the same size: X where English is spoken and Y where it is not; location X is 13 times more likely to be referred to in a random Wikipedia article than location Y .

Notice Hebrew has the most bias by far, followed by Welsh, Russian and German. Chinese is the only language with a bias toward non-speakers; I attribute this partially to the fact Wikipedia has been blocked for large periods of time in China and the large expatriate population. Chapter 6 introduced the notion of considering references to locations as a probability distribution. I revisit that concept here and consider the global distribution of references to locations by a wikipedia a probability distribution and as before, measure the distance between these distributions using the Jensen-Shanon Divergence (described in Section 6.3). Table 7.5 shows the divergence between the different distributions. Notice the most similar versions of Wikipedia with respect to their references to locations are Japanese and Chinese, and Spanish and French. The most different languages are English and Welsh, and English and Arabic. The similarity between languages corresponds very loosely to the population distribution of the languages e.g. there is a high concentration of Japanese and Chinese speakers in south east Asia, and a high concentration of Spanish and French speakers in south west Europe. The English and Welsh distributions are both quite different to all the others. Welsh, because the speakers are in a very localised area, and English, because of its unique status as the world's Lingua Franca and the language of the Internet.

7.5.2 Qualitative analysis

In addition to the quantitative differences, I present here the qualitative differences between location references on both a micro and macro scale. I begin at the micro scale looking at the most commonly

⁵http://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view/FAQ#Anglo-American_focus_and_systematic_bias

⁶In countries where multiple languages are spoken the references and population are apportioned in the appropriate ratios. Only people's native language is included.

Language	English	German	French	Polish	Portuguese	Spanish
# Refs. / 1m speakers	2495.7	709.0	461.9	404.6	84.7	102.7
# Refs. / 1m non-speakers	189.8	24.5	23.23	28.4	12.3	25.1
Bias	13.15	28.97	19.88	14.3	6.9	4.1
Language	Japanese	Russian	Chinese	Arabic	Hebrew	Welsh
# Refs. / 1m speakers	85.2	281.8	11.2	33.7	1017.6	136.2
# Refs. / 1m non-speakers	15.2	9.9	14.2	9.3	16.1	4.0
Bias	5.6	28.5	0.8	3.6	63.3	34.1

Table 7.4: References per 1m people in different wikipedias

	Germ.	Fren.	Pol.	Jap.	Port.	Span.	Russ.	Chin.	Arab.	Hebr.	Welsh
Eng.	0.330	0.294	0.343	0.326	0.321	0.345	0.351	0.355	0.393	0.384	0.396
Welsh	0.357	0.296	0.364	0.302	0.354	0.322	0.325	0.330	0.347	0.374	
Hebr.	0.278	0.230	0.286	0.240	0.312	0.272	0.229	0.273	0.309		
Arab.	0.338	0.292	0.357	0.268	0.338	0.303	0.297	0.304			
Chin.	0.284	0.236	0.310	0.129	0.285	0.268	0.255				
Russ.	0.215	0.200	0.237	0.204	0.295	0.242					
Span.	0.248	0.178	0.243	0.222	0.207						
Port.	0.284	0.213	0.265	0.251							
Jap.	0.246	0.191	0.263								
Pol.	0.238	0.194									
Fren.	0.189										

Table 7.5: Jensen-Shanon Divergence between different wikipedias w.r.t. locations

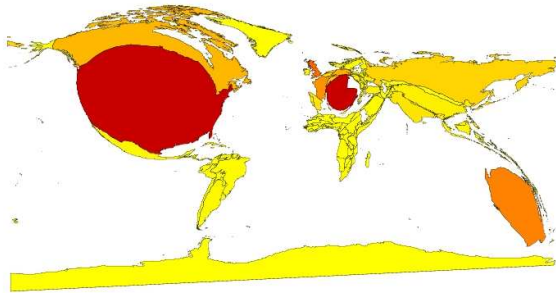
referred to locations in each language and follow with the macro scale by generating per language cartograms.

Table 7.6 shows the top five referred to locations for each of our 12 languages. These locations make up ranks 1-5 (the left extreme) in the distribution graphs of Figures 7.4 and 7.5. Notice for every language except Hebrew, Chinese and Russian the top location is a country where that language is spoken. Only Hebrew is particularly surprising as it is not a native language in any if its top five locations (this could be partially attributed to errors in disambiguation). Berlin, Paris, Warsaw and London are the only locations appearing in Table 7.6 that are not countries. Germany appears in the top five for every language except Welsh. France appears in the top five for every language except Welsh and Arabic. I speculatively attribute the prominence of France and Germany in these distributions to their central roles in European-American politics in the 20th century, most notably in World War II.

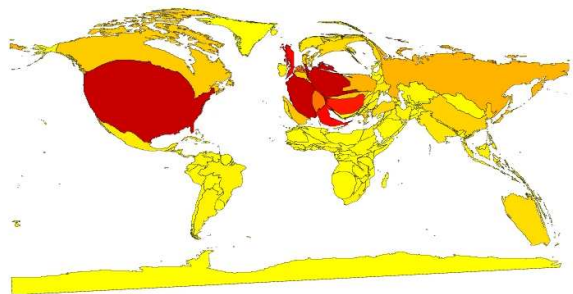
To view the macro differences between the location reference distributions, I have produced car-

Language	Top referred to locations				
English	United States	Italy	Germany	France	Australia
German	Germany	France	Berlin	Austria	Italy
French	France	Paris	Italy	Germany	Spain
Polish	Poland	France	Germany	Italy	Warsaw
Japanese	Japan	France	Germany	Italy	China
Portuguese	Brazil	Portugal	Spain	Germany	France
Spanish	Spain	France	Argentina	Germany	Mexico
Russian	Germany	Russia	Moscow	France	Italy
Chinese	Japan	Taiwan	China	France	Germany
Arabic	Egypt	Russia	Iraq	Germany	Syria
Hebrew	France	Germany	Russia	London	Spain
Welsh	Wales	England	France	Europe	Spain

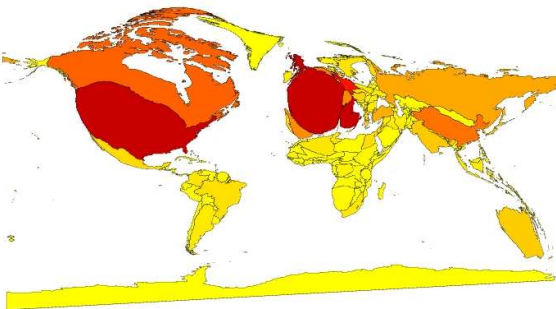
Table 7.6: Top five locations from each language



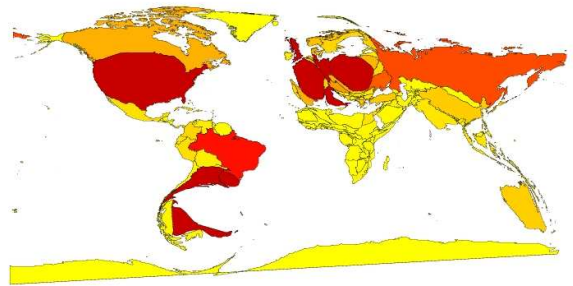
English



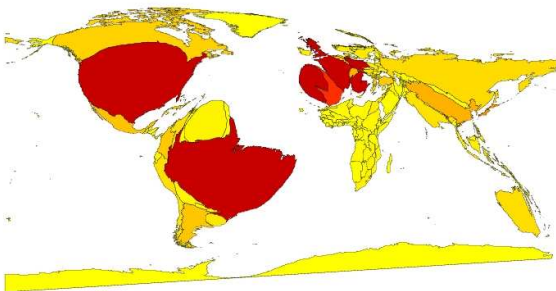
German



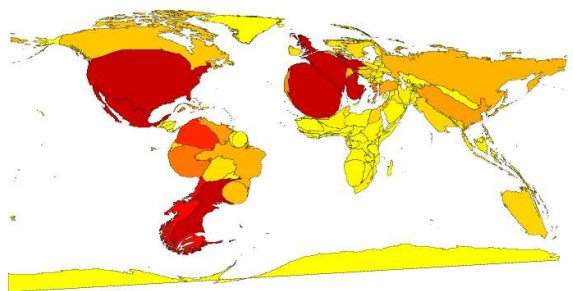
French



Polish

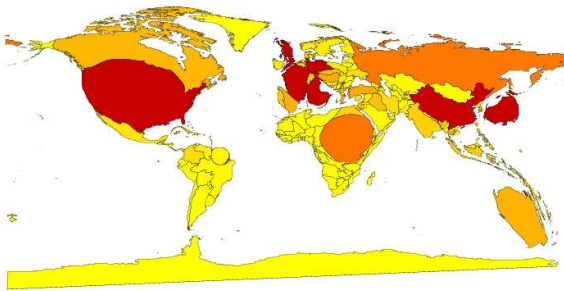


Portuguese

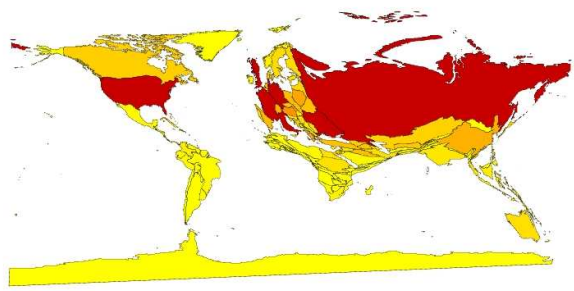


Spanish

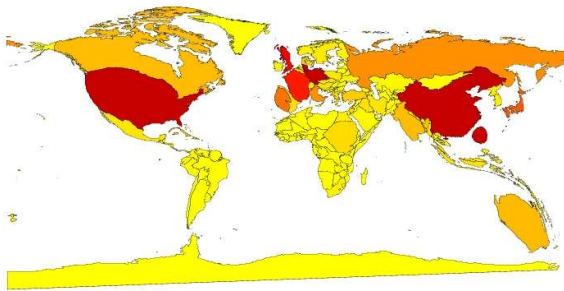
Figure 7.6: Cartograms of references in different wikipeidias



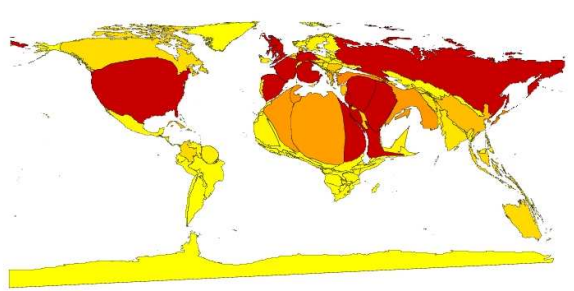
Japanese



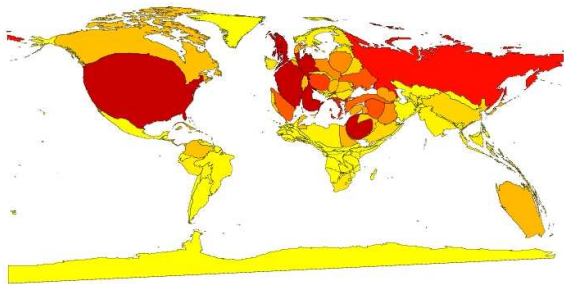
Russian



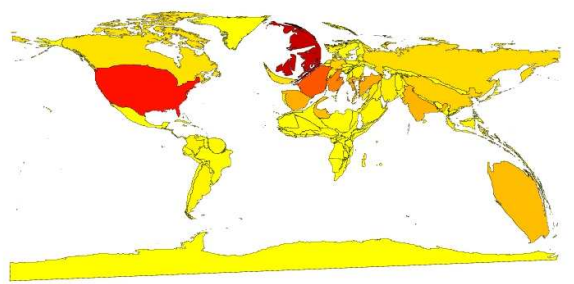
Chinese



Arabic



Hebrew



Welsh

Figure 7.7: Cartograms of references in different wikipedias

tograms of the references made by each language version of Wikipedia to locations in each country (Figures 7.6 and 7.7). The size of the country denotes whether it has a higher or lower than average number of references per inhabitant. The colour of the countries represent the absolute number of references (pale yellow \rightarrow few references, dark red \rightarrow many references). These maps show further bias in how countries are referenced. For example look at how the United States dominates the English map or compare how South America dramatically changes shape between the Spanish and Portuguese maps, reflecting colonial history. In the Spanish map the west of South America is heavily skewed, while the Portuguese Wikipedia is biased to the east. The Iberia Pininsula mirrors this switch on a slightly less dramatic scale. The systematic bias, acknowledged in the English Wikipedia, of omissions related to the African continent is clearly reflected across all of our selected languages. Even the Arabic Wikipedia neglects all but North Africa (Figure 7.7:Arabic).

7.6 The Steinberg hypothesis

Saul Steinberg’s most famous cartoon “View of the world from 9th Avenue” depicts the world as seen by self-absorbed New Yorkers (Figure 7.8⁷). This chapter finds that this particular fish-eye world view is ubiquitous and inherently part of human nature. The *Steinberg hypothesis* proposes that all people have similar world views with respect to their own locality: “We are all little Steinbergs.” The goal for this section is to model how a *single* person sees the world, in the same way as Steinberg does for a single New Yorker. The necessary assumption is that Wikipedia is read and edited by a typical sample of the population. By summing the predicted world views of a population and fitting this combined model to Wikipedia, the validity of the individual models can be tested. To do this I define the relevance of a location to a person. “Relevance” in this context is considered a synonym for “likelihood to use in dialogue”.

I calculate the relevance of a location, l , to a person, ρ , thus:

$$\text{rel}(l, \rho) = \text{subjInt}(l, \rho) \cdot \text{objInt}(l) \quad (7.4)$$

the product of the subjective interestingness of location l to person ρ and the objective interestingness of the location. The subjective interestingness is based on the relationship between ρ and l , while the objective interestingness is based on properties of l . The subjective interestingness is modelled as a function of the distance from ρ to l :

$$\text{rel}(l, \rho) = f(\text{dist}(\rho, l)) \cdot \text{objInt}(l), \quad (7.5)$$

where f is some decay function. This relevance function is analogous to Mehler et al. (2006)’s influence function for calculating the influence of an information source over a city. In the following section we will compare different possible functions for f and $\text{objInt}(l)$. In these experiments the only property taken into account when modelling objective interestingness is population. The distance function used is the geodesic distance, chosen for simplicity.

The relevance of a location l to person ρ can be seen as an estimate of the probability that they will refer to l in a document they author. We can extend this to Wikipedia by averaging across all potential authors. Let \mathcal{P}_v be the set of speakers of language v and $\text{rel}^v(l)$ be the relevance of location l to the speakers of language v , then:

⁷© The Saul Steinberg Foundation/Artists Rights Society (ARS), New York

Mar. 29, 1976

THE NEW YORKER

Price 75 cents

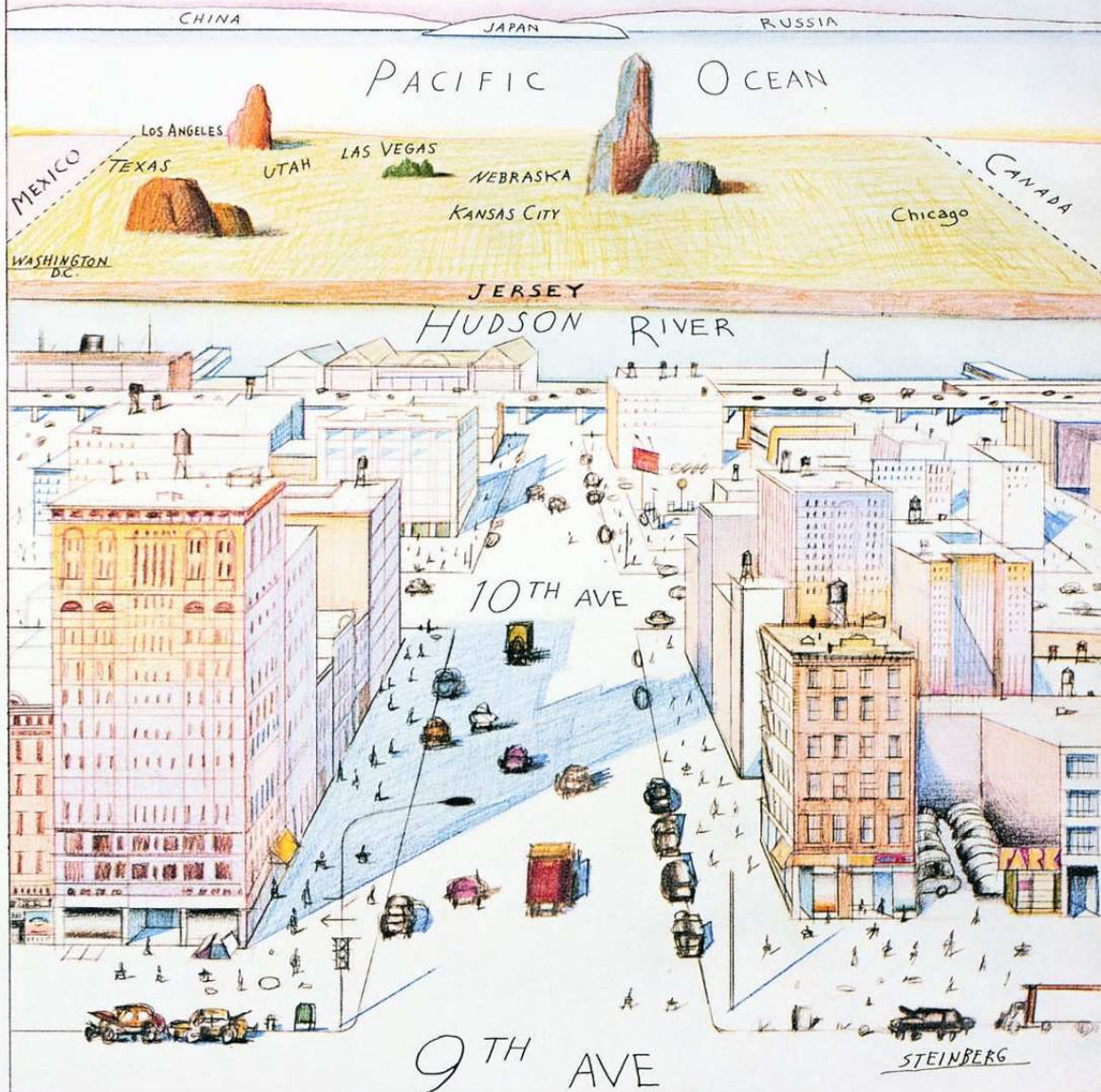


Figure 7.8: Cover of The New Yorker, March 29, 1976

$$\text{rel}^v(l) = \sum_{\rho \in \mathcal{P}_v} (f(\text{dist}(\rho, l))) \cdot \text{objInt}(l). \quad (7.6)$$

This is equivalent to

$$\text{rel}^v(l) = \sum_{a \in A_v} (\text{pop}(a) \cdot f(\text{dist}(a, l))) \cdot \text{objInt}(l), \quad (7.7)$$

where A_v is the set of all non-overlapping locations where v is spoken and $\text{pop}(a)$ is the population of a . We can make one further refinement to this equation to allow for locations where multiple languages are spoken. Let $\text{prop}(a, v)$ be the proportion of people in location a that speak language v .

$$\text{rel}^v(l) = \sum_{a \in A_v} (\text{pop}(a) \cdot \text{prop}(a, v) \cdot f(\text{dist}(a, l))) \cdot \text{objInt}(l) \quad (7.8)$$

7.6.1 Experiment

To test the Steinberg hypothesis I compared different possible equations for $f(d)$ and $\text{objInt}(l)$. The frequencies of links to Wikipedia articles describing locations is considered a histogram, \mathbf{O} . These observed frequencies are normalised to give a unit histogram \mathbf{O}' . A predicted histogram, \mathbf{P} , is generated consisting of the set of predicted frequencies for all locations calculated using $\text{rel}^v(l)$. \mathbf{P} is normalised to give \mathbf{P}' . Population data was taken from the World Gazetteer⁸. To tune the variables in the formulations of $f(d)$ and $\text{objInt}(l)$ an iterative greedy algorithm has been implemented that minimises the symmetric difference between \mathbf{O}' and \mathbf{P}' . Where symmetric difference is defined as follows:

$$\mathbf{O}' \Delta \mathbf{P}' = \sum_{l \in L} (\mathbf{O}'[l] \Delta \mathbf{P}'[l]), \quad (7.9)$$

where $\mathbf{O}'[l]$ is the normalised relevance of location l in \mathbf{O} .

The following methods of calculating relevance are compared:

- **Constant.** Every location is equally likely to be referred to.

$$\text{rel}^v(l) = 1 \quad (7.10)$$

- **$\log(\text{pop}(l))$.** The likelihood of a location being referred to is the log of its population.

$$\text{objInt}(l) = \log(\text{pop}(l)), \quad f(d) = 1 \quad (7.11)$$

- **$\log(\text{pop}(l)) \frac{1}{\log(d)}$.** The likelihood of a location being referred to is the log of its population divided by the log of the distance.

$$\text{objInt}(l) = \log(\text{pop}(l)), \quad f(d) = \frac{1}{\log(d)} \quad (7.12)$$

- **$\text{pop}(l)^\alpha$.** The likelihood of a location being referred to is its population raised to power α .

$$\text{objInt}(l) = \text{pop}(l)^\alpha, \quad f(d) = 1 \quad (7.13)$$

⁸<http://www.world-gazetteer.com/>

Language	English	German	French	Portuguese	Spanish	Chinese	Arabic	Hebrew
Constant	1.05	1.17	1.23	1.19	1.23	1.15	1.32	1.21
$\log(\text{pop}(l))$	1.03	1.15	1.20	1.15	1.19	1.11	1.28	1.18
$\log(\text{pop}(l)) \log(d)^{-1}$	1.02	1.11	1.18	1.14	1.19	1.10	1.27	1.12
$\text{pop}(l)^\alpha$	1.03	1.10	1.11	1.03	1.09	0.94	1.11	1.01
$\text{pop}(l)^\alpha d^{-\beta}$	0.92	0.77	0.87	0.89	1.02	0.92	0.91	0.82
$\text{pop}(l)^\alpha e^{-d\beta}$	0.90	0.81	0.90	0.92	1.04	0.94	0.98	0.89

Table 7.7: The symmetric differences between the observed and expected results between different formulations of the Steinberg equation.

- $\text{pop}(l)^\alpha d^{-\beta}$. The likelihood of a location being referred to is its population raised to power α multiplied by its distance raised to power negative β .

$$\text{objInt}(l) = \text{pop}(l)^\alpha, \quad f(d) = d^{-\beta} \quad (7.14)$$

- $\text{pop}(l)^\alpha e^{-d\beta}$. The likelihood of a location being referred to is its population raised to power α multiplied by e raised to the power of the product of negative β and the distance.

$$\text{objInt}(l) = \text{pop}(l)^\alpha, \quad f(d) = e^{-d\beta} \quad (7.15)$$

7.6.2 Results

Table 7.7 shows the minimised symmetric difference between the observed and predicted histograms. Note the symmetric difference between unit histograms varies between zero for complete overlap (perfect prediction) and two for no overlap (catastrophic prediction).

In all apart from the English Wikipedia the best performing formulation of the equation was $\text{pop}(l)^\alpha d^{-\beta}$. One would expect to see most noise in the English Wikipedia due to the wide distribution of English Speakers across the world and English's position as the main language of the Internet. My conclusion is that $\text{pop}(l)^\alpha d^{-\beta}$ is the most accurate model of the likelihood of a person to refer to a location. As this trend is seen in 7 of the 8 languages, I am confident the only reason it is not observed in the English Wikipedia is the amount of noise. For illustration Figure 7.9 shows the overlap of a sample of the \mathbf{O}' and \mathbf{P}' histograms for the English Wikipedia and $\text{pop}(l)^\alpha d^{-\beta}$ formulation of the model. Note the curve shown by the observed data is the same as that would be seen were Figure 7.4:English plotted on a linear scale.

Table 7.8 shows the optimal values for α and β for the best performing formulations. Note the higher the α value the greater the importance of population and the higher the β value the greater the decay parameter for distance. Also note that β values between rows two and three are not comparable, while α values are comparable across the table. Observe that Chinese has the lowest decay parameter implying the distance a location is from a centre of Chinese population has little effect on how often it is referred to. This observation is consistent with the notably low bias factor of the Chinese Wikipedia. In the rows with a decay parameter for distance, the French and German Wikipedias have the highest α values (approaching one) meaning the likelihood of a location being referenced is almost directly proportional to its size. French, German and Arabic have the highest decay values for distance meaning that they tend to reference locations near centres of population where the respective language is spoken significantly more than places further afield. This observation is consistent with the distribution heat maps: Figures 7.2 and 7.3, and Table 7.4.

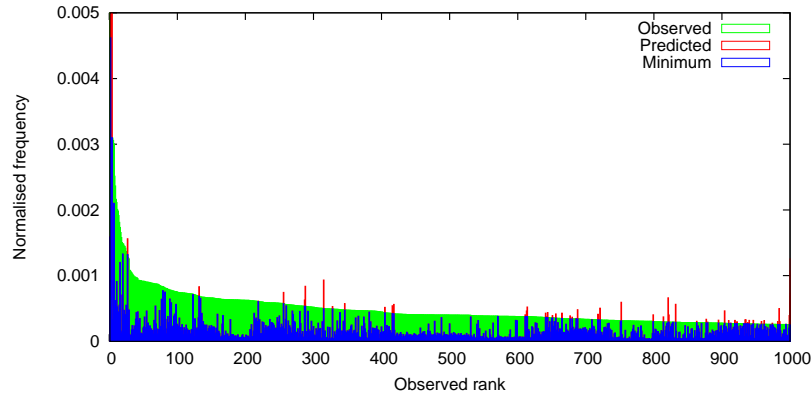


Figure 7.9: First 1000 locations of the normalised histograms for the observed frequencies from the English Wikipedia (green) and the frequencies predicted by the Steinberg hypothesis (red). The minimum of the two values is superimposed in blue making the symmetric difference the sum of the red and green areas

Language		English	German	French	Portuguese	Spanish	Chinese	Arabic	Hebrew
$\text{pop}(l)^\alpha$	α	0.20	0.40	0.54	0.60	0.54	0.72	0.66	0.72
	β	0.72	0.89	0.75	0.57	0.68	0.24	0.79	0.49
$\text{pop}(l)^\alpha d^{-\beta}$	α	0.44	0.91	0.92	0.73	0.57	0.67	0.48	0.63
	β	0.72	0.89	0.75	0.57	0.68	0.24	0.79	0.49
$\text{pop}(l)^\alpha e^{-d\beta}$	α	0.53	1.00	0.9	0.73	0.72	0.72	0.72	0.85
	β	0.00034	0.00187	0.00036	0.00018	0.00016	0.0	0.00072	0.00029

Table 7.8: Optimal values of α and β for different formulations of the Steinberg equation

A series of pairwise statistical significance tests were conducted to test the hypothesis that the $\text{pop}(l)^\alpha d^{-\beta}$ formulation fits the observed data best. Comparisons were done for each language comparing the $\text{pop}(l)^\alpha d^{-\beta}$ formulation to the $\text{pop}(l)^\alpha$, $\text{pop}(l)^\alpha e^{-d\beta}$ and Constant formulations. The significance test chosen was a two-tailed Student's t-test with a significance level of 5%. The $\text{pop}(l)^\alpha d^{-\beta}$ formulation was statistically significantly better than the Constant formulation for all languages, better than the $\text{pop}(l)^\alpha$ formulation for all languages except Spanish, and better than the $\text{pop}(l)^\alpha e^{-d\beta}$ equation only for Chinese, Arabic and Hebrew. Interestingly it is easier to fit a model to languages spoken in smaller areas. These tests demonstrate that the $\text{pop}(l)^\alpha d^{-\beta}$ formulation is statistically significantly better than the Constant and $\text{pop}(l)^\alpha$ method, but it is not conclusively the best method.

7.6.3 Applications of the Steinberg hypothesis

In dialogue between people either face to face, through websites, printed or broadcast media, there is an assumed common shared understanding of the world. The model presented here of how likely people are to refer to different locations and the cartograms of Figures 7.6 and 7.7 help us to appreciate how this shared understanding varies between different people. Part of this shared world understanding is a shared vocabulary. When ambiguous words and terms occur in this vocabulary one relies on contextual clues to assign meaning to these terms. If few or no contextual terms exist we make assumptions of what the most likely semantic meaning is. Here I revisit the problem first tackled in Chapter 6 of placename disambiguation. The Steinberg hypothesis is applicable to placename disambiguation only when one knows the population of the possible locations and the location of the author, for example articles in a local newspaper. Figures 7.10 and 7.11 show how the Steinberg hypothesis could be applied to placename disambiguation. The $\text{pop}(l)^\alpha d^{-\beta}$ formulation of the model is used, and takes the α and β

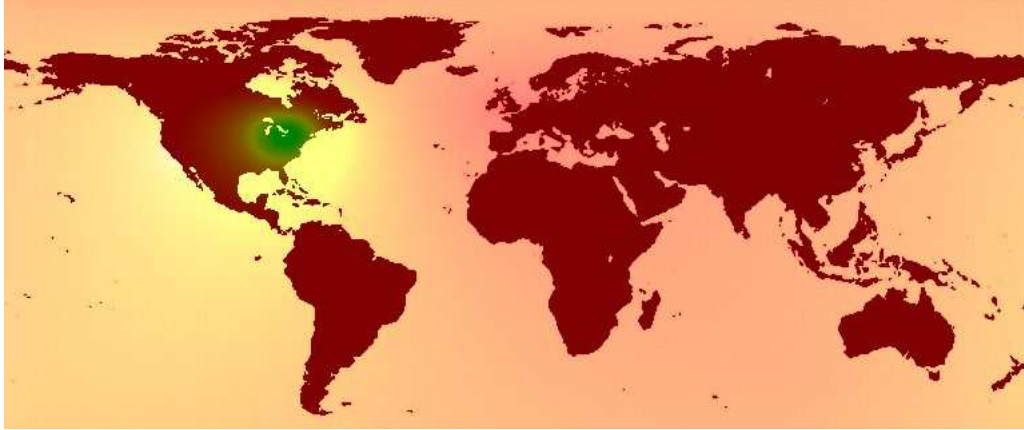


Figure 7.10: Map of which location you are most likely to mean when referring to “London” in English dependent on your current location: London, UK (red), London, Ontario (green), London, California (blue)

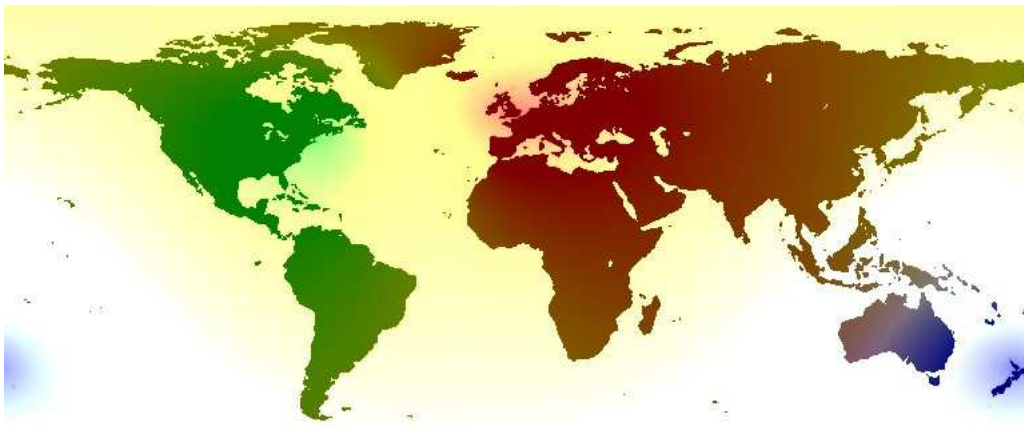


Figure 7.11: Map of which location you are most likely to mean when referring to “Cambridge” in English dependent on your current location: Cambridge, UK (red), Cambridge, Massachusetts (green), Cambridge, New Zealand (blue)

values optimised for the English Wikipedia (Table 7.8). Figure 7.10 shows which location a person is most likely to be referring to when they refer to the placename “London” dependent on where in the world they are. Figure 7.11 is similar however for “Cambridge”. Notice in Figure 7.10, London, California, is completely dominated by London, UK and London, Ontario. Also notice that the ambiguous region is much smaller for “London” than “Cambridge” (although with “Cambridge” much of the ambiguous regions fall in the sea).

7.7 Temporal references in Wikipedia

The previous two chapters have been concerned with geographic co-occurrence models — this section applies the same methods to building a temporal co-occurrence model. The temporal co-occurrence models presented here examine the references to articles describing years in different language versions of Wikipedia.

Language	English	German	French	Polish	Portuguese	Spanish
Links extracted	7,317,469	2,334,779	2,364,609	3,086,297	1,606,602	2,522,518
prop Loc (%)	26.7	9.1	7.3	6.1	5.4	7.1
prop Years (%)	2.7	5.2	7.4	11.1	6.7	9.5
Language	Japanese	Russian	Chinese	Arabic	Hebrew	Welsh
Links Extracted	5,419,486	104,019	2,813,221	1,619,460	3,209,710	702,262
prop Loc (%)	1.9	5.0	2.9	3.8	3.2	3.5
prop Years (%)	12.1	0.6	7.6	2.0	8.1	7.7

Table 7.9: Proportion of links extracted from different wikipeidias to articles describing years

Disambiguating articles describing years in Wikipedia is trivial compared to disambiguating articles describing locations. Titles of articles describing dates and years have a rigid format that varies minimally between the different language versions of Wikipedia (this property was taken advantage of in Section 4.3.2 to enrich WordNet and reduce data sparsity). In the 12 language versions of Wikipedia used in this chapter the article describing the events of a specific year has as its title the respective year in numerical format. This is occasionally followed by the word for *year* in the respective language. For example the article describing the events of 1066 in the English Wikipedia is titled “1066,” while in the Chinese Wikipedia it is titled “1066年”.

Table 7.9 shows the proportion of links in different language versions of Wikipedia that refer to years (the number of links extracted and the proportion referring to locations are repeated from Table 7.2 for comparison). There are a number of observations one can make from this table. For example, the proportion of references to times and places are quite similar (at least within an order of magnitude) with the exceptions of English, Japanese and Russian. English has a relatively high proportion of location references with a low proportion of temporal references, while it is vice-versa in Japanese. Russian has a typical proportion of location references and a notably low number of temporal references — on further analysis this is due to inconsistencies in how editors of the Russian Wikipedia link to years. For comparison one can compare these results to the distribution of tags in Flickr: 7.0% of tags were references to time and 19.3% of tags were references to locations (discussed in Section 4.5 and illustrated in Figure 4.10).

Temporal distribution

As well as temporal references being easier to extract than their spatial counterparts, their distribution is much more straight-forward to model. Temporal references can be modelled with two power-law distributions. In log frequency, log year space, this gives two straight lines, the first of which fitting the late 20th and early 21st centuries, the second fitting prior the 20th century (shown in Figures 7.12 and 7.13). I refer to the period before this switch as *pre-modern* or before living memory and after this point as *post-modern* or during living memory. Table 7.10 shows the scaling exponents of the pre-modern and post-modern curves and the year in which the distribution switches from pre-modern to post-modern.

Notice in all language versions of Wikipedia the pre-modern co-efficient is significantly steeper — a ratio of more than 2:1 — than the post-modern except for Japanese and Chinese. I speculate this is due to the fact that areas where these languages are now spoken were parts of world-powers of greater status pre-20th century than during the 20th century. Taking this hypothesis further, one would expect the Chinese graph to dramatically increase in current references over the next 50 years. The switch from pre-modern to post-modern, found through inspection, in all languages is around 1940. This coincides with World War II, the most widespread war in history. This concurs with Section 7.5.2’s observations

Language	English	German	French	Polish	Portuguese	Spanish
Post-modern	-0.56	-0.24	-0.47	-0.61	-0.65	-0.57
Pre-modern	-1.41	-1.48	-1.26	-1.40	-1.43	-1.30
Switch	1940	1940	1940	1940	1940	1940
Language	Japanese	Russian	Chinese	Arabic	Hebrew	Welsh
Post-modern	-0.72	-0.53	-0.59	-0.37	-0.38	-0.16
Pre-modern	-0.97	-1.68	-0.89	-0.97	-1.28	-0.58
Switch	1940	1940	1940	1940	1940	1940

Table 7.10: Co-efficients of the Zipfian distributions for pre-modern and post-modern curves and the pre-modern and post-modern switching point for each wikipedia

that World War II is a particularly well documented subject in the multiple of versions of Wikipedia.

7.8 Discussion

In this chapter I have presented my work mining location and temporal references from Wikipedia. The Steinberg hypothesis, built on this analysis, allows the world of specific people to be more accurately modelled. This allows greater understanding of a person’s discourse, either by someone else, or automatically by a computer. The approach presented here is sufficiently accurate to support the Steinberg hypothesis: that *everyone* has a localised fish-eye view of the world. I expect that including topographical distance, migration patterns, political, social and economic factors into the modelling process will achieve a more accurate predictive model. Similarly more information could be taken from Wikipedia to improve the model such as in-links and the length of articles. Models such as the ones presented here, despite being facile, can still have significant consequences when it comes to understanding a person’s discourse. To provide a more concrete example, consider a geographically aware search engine with the ability to answer the query “Jobs in Cambridge.” It is not apparent from the query whether Cambridge, UK, Cambridge, Massachusetts, or Cambridge, New Zealand, is intended. The approximate location of the user can be calculated from their IP address; Figure 7.11 shows, according to our model, which Cambridge is most likely to correspond to the user’s intention based on their location.

In a broader context recognising the phenomenon asserted by the Steinberg hypothesis could enrich human dialogue and increase understanding between people. On one level it demonstrates that bias and prejudice toward our own location is part of human nature and to some extent can be excused. On a higher level understanding this phenomenon can help avoid confusion and increase the shared understanding of the world required for any dialogue. I would like to conclude this chapter with discussions on three points: the independence of the different wikipedias, detecting events in Wikipedia, and Numenore, a demonstration application for some of this work.

7.8.1 Can the different wikipedias be considered independent?

In this chapter I have assumed all different language versions of Wikipedia to be independent; that articles written on the same subject are independently authored and later linked. In many cases, this is unlikely to be true. A plausible scenario is articles are first written in a language in which the author is fluent, and later translated into other language versions (often by a different author). Studying this behaviour is outside the scope of this thesis, but could be achieved by comparing the out-links of articles (as a similarity measure) and looking at how articles change over time in the different wikipedias.

This chapter also ignores the possibility that bilingual Wikipedians will edit multiple wikipedias.

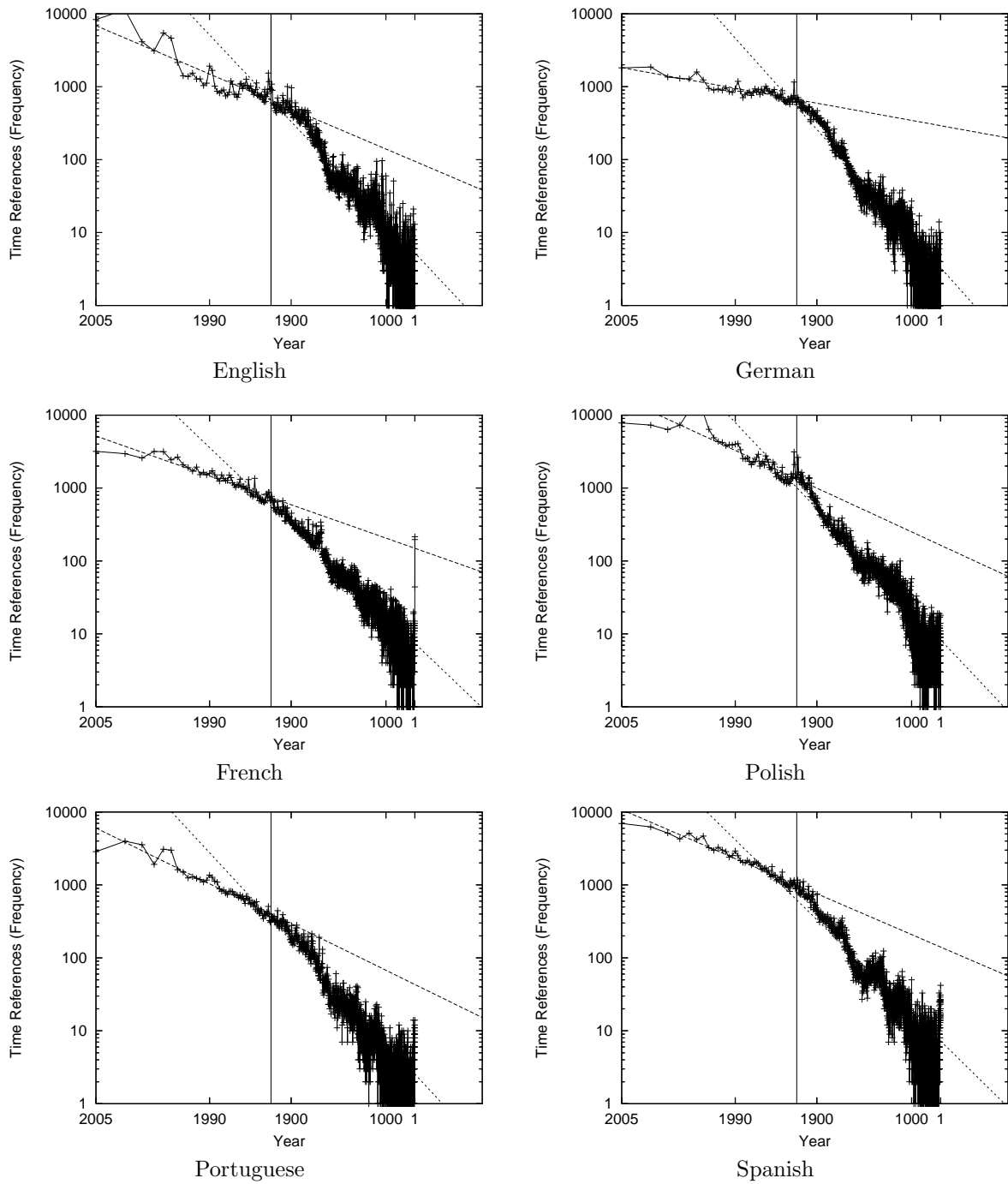


Figure 7.12: Distribution of temporal references in the different wikis

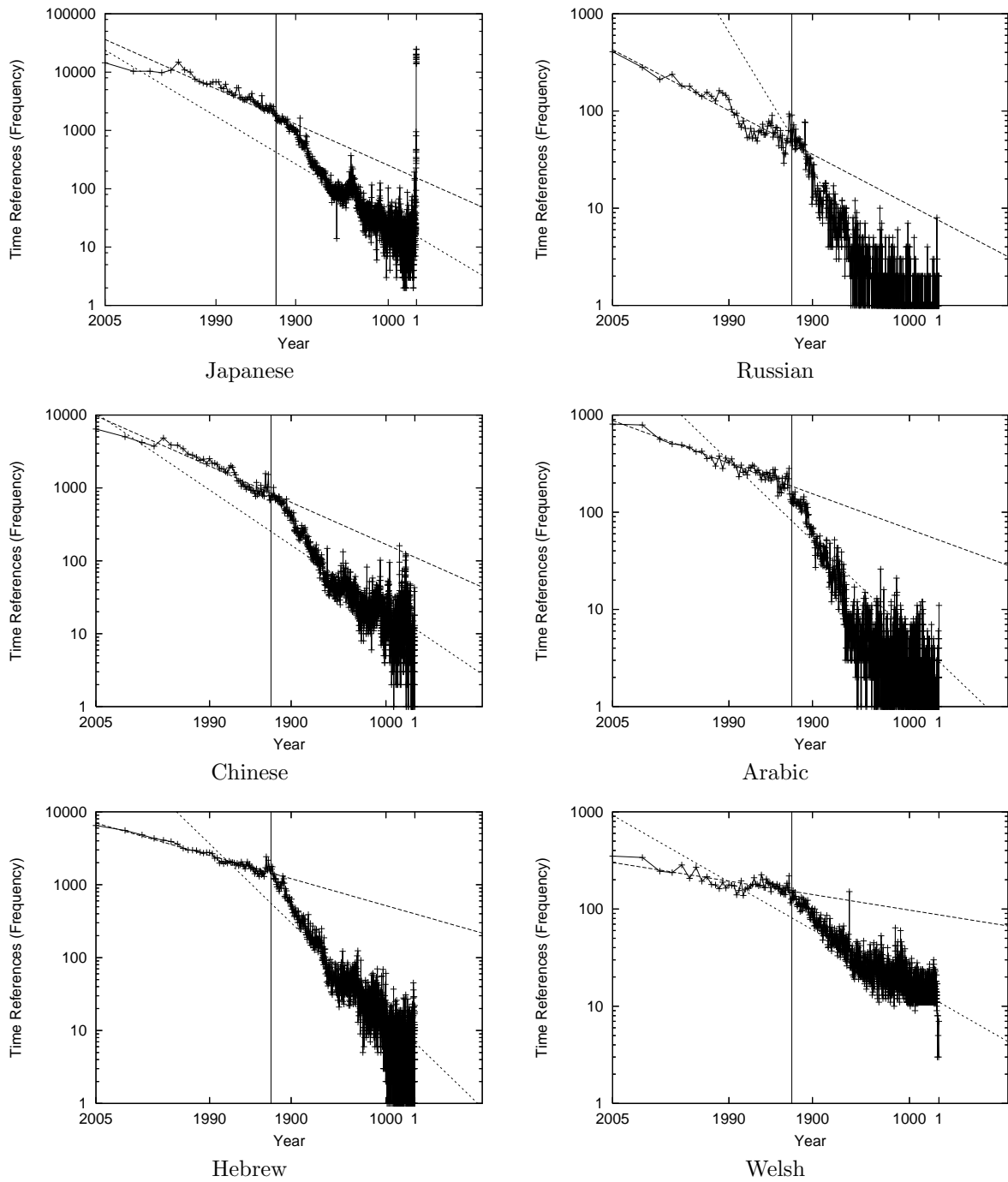


Figure 7.13: Distribution of temporal references in the different wikis

For example, the two most bias wikipedias examined in this chapter were Hebrew and Welsh. The vast majority of Hebrew and Welsh speakers will also speak English. It is a plausible hypothesis that editors of these wikipedias will write in their native language wikipedia on Welsh / Hebrew topics, and the English Wikipedia for more general topics. This could lead to a positive feed back loop as people contribute in areas their native wikipedia is already strong. This falls outside the scope of this thesis but would be testable by matching edits to different wikipedias from the same IP addresses and grounding those IP addresses to locations.

7.8.2 Detecting events in Wikipedia

Event detection is a subject that has been explored for a range of corpora. In this context *event* refers to a *time, place* tuple with an optional *action* component. Events can either be small scale e.g. a sporting event, one day in one town, or global scale e.g. a war spanning years and multiple countries. Both involve different techniques and are appropriate to different corpora. The topic detection and tracking studies tackle small scale events in news corpora, often involving NLP techniques (Wayne 2000). Mehler et al. (2006) and Mei et al. (2006) observe events on a similar scale, however using statistical methods on multiple corpora. Smith (2002) and Rattenbury et al. (2007) look for larger scale events using co-occurrence statistics in huge corpora. Smith uses a date-place contingency table and calculates the log-likelihood of occurrences. Their corpora is a cultural heritage collection covering 4000 years. Rattenbury et al. use a sample of Flickr photos with time and location meta-data; multi-scale burst detection is used to assign semantics to tags.

Event detection is outside the scope of this thesis, however I think it would be an obvious extension of the analysis presented in this chapter. I believe statistical and co-occurrence based event detection similar to that employed by Smith (2002) and Rattenbury et al. (2007) would be particularly appropriate for Wikipedia.

7.8.3 Numenore

Numenore is a web-application built to demonstrate the work presented in this thesis. The heat maps shown in Figures 7.2 and 7.3 are browsable through a Google Maps mashup, and the time distributions shown in Figures 7.12 and 7.13 are viewable as heat time-lines (Figure 7.14). Time and location co-occurrence maps have also been generated allowing users to view only locations that co-occur in documents with specific times. These are quantised in 200 year periods, for example, one could view a heat map of the locations occurring in the German Wikipedia between 1000 A.D. and 1200 A.D. Numenore is online at <http://www.numenore.co.uk>.

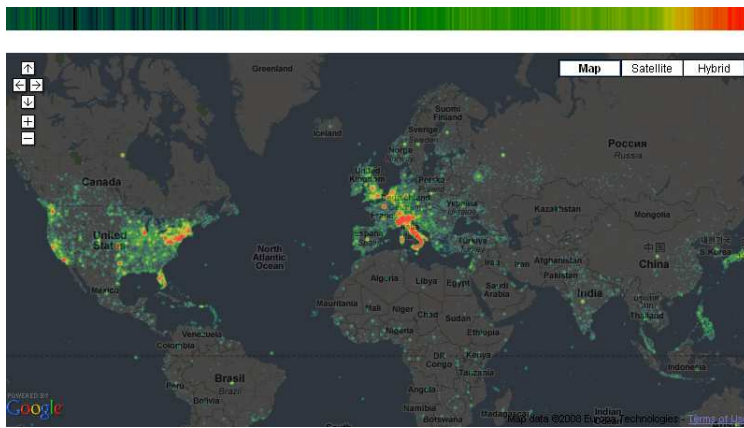


Figure 7.14: Screen shot taken from <http://www.numenore.co.uk>

Chapter 8

Conclusions

Chapter 1 of this thesis identifies its scope as to extract geographic world knowledge from Wikipedia and apply it to geographic information retrieval, and identifies three specific tasks: mining world knowledge from Wikipedia, applying world knowledge mined from Wikipedia to placename disambiguation, and comparing world knowledge extracted from different language versions of Wikipedia. In the next section I shall outline my key achievements in pursuit of these aims followed by the limitations of these methods and future work suggested by this thesis.

8.1 Achievements

A complete list of the contributions made by this thesis to the area of geographic information retrieval is provided in Section 1.8 and noted in the relevant body-of-work chapters. In this section I would like to highlight the key achievements of this thesis in the areas of placename disambiguation and the formulation of the Steinberg hypothesis.

8.1.1 Placename disambiguation

My most significant contribution to placename disambiguation is in Chapter 6 where I demonstrate that supervised placename disambiguation can statistically significantly outperform simple rule-based methods, and that this significant improvement can be observed in retrieval results. The training data for these supervised experiments comes from Chapter 5 where I construct a disambiguation pipeline and demonstrate which classes of evidence provide the greatest information for grounding Wikipedia articles as locations. Essential to interpreting these promising supervised results is the work presented in Section 6.3, where the theoretical bounds of performance are quantified, providing a set of benchmarks.

Mika et al. (2008) observe that models trained on one source typically perform poorly on other sources in language processing tasks. This can be seen to some extent in Section 6.5 where the best results are seen on a Wikipedia based corpus. Despite performing best on Wikipedia corpora, supervised placename disambiguation still performs well on general corpora, with the caveat that locations not seen in the training data will not be recognised. I propose measuring the overlap between locations occurring in the training corpus and a sample of the test corpus as a suitable measure for deciding whether a model is appropriate.

As well as contributions to supervised placename disambiguation, promising results have also been observed performing placename disambiguation with a default gazetteer constructed using statistics from

the annotated corpus. The MR method of placename disambiguation, although statistically significantly worse than the top supervised methods, may not provide noticeably different results to a real user. It performs a unique mapping from placenames to locations. The advantage is it is a fraction of the complexity of the supervised methods. The Steinberg hypothesis may provide further advances in placename disambiguation, this is discussed in future work.

In summary, to achieve the greatest possible accuracy for placename disambiguation, one has to take into consideration how locations are referred to in documents.

8.1.2 The Steinberg hypothesis

The Steinberg hypothesis presented in Chapter 7 demonstrates convincingly that all people have the same localised fish-eye world view, as captured iconically in Saul Steinberg's famous cartoon (Figure 7.8). Similar hypotheses have been put forward by sociologists and geographers in the past, however to my knowledge, it has never been quantified on the scale presented in this thesis. The Steinberg hypothesis has direct applications in the area of geographic information retrieval (context based placename disambiguation is provided as an example), and broader applications increasing shared understanding whenever a dialogue takes place, be it between people or a person and a computer.

8.2 Limitations

Despite the recent advances made by myself and others in the area of GIR, there are still significant limitations. Jones et al. (2008) found that once a query has been identified as a geographic query the feature providing greatest retrieval improvement is the number of placenames that occur in documents. This sort of objective feature requires no query-document similarity measures. The recent success of such simple geographic features suggests the immediate future of GIR may not be in developing complex placename disambiguation algorithms.

I believe placename disambiguation will gain widespread use in the medium term, perhaps the next 3-5 years. The world's biggest search engines, Google and Yahoo, have recently developed scalable robust implementations of the MapReduce framework deployed across 1000 machine clusters¹. MapReduce is a framework that allows large processing problems to be split into chunks and distributed across huge clusters of machines. Such a framework is ideally suited to the computationally expensive task of supervised placename disambiguation, and would allow placename disambiguation to scale to the web. With the proliferation of location aware devices, geographically augmenting search is going to become increasingly important.

8.2.1 Placename disambiguation

In Section 6.3 I quantify the potential improvement supervised placename disambiguation can provide: 80.6%-89.9% is the envelope. Looking at the flip side, there is potential for a 52.6% reduction in error.

Having observed that there are only small gains available in supervised placename disambiguation in English, the gains to be made by multilingual placename disambiguation is significantly smaller. Section 6.3 observes that 35.1% of references to placenames are ambiguous in the English Language. In Section 7.4, we learn that Spanish and German are the only languages where placename disambiguation is even a comparable problem, with ambiguity levels of 18.6% and 12.9% respectively. In the other 10 languages

¹<http://tinyurl.com/5p3he4>, <http://tinyurl.com/4fgok6>

looked at by this thesis less than 10% of the references to placenames are ambiguous. With such a small envelope for improvement, one could argue placename disambiguation in other languages is not worth pursuing.

The fact that the greatest gains for placename disambiguation are in English looks like a significant limitation at first glance; however this is put into context when one considers approximately 78% of websites are in English (Myers and Dyke 2000).

8.2.2 Wikipedia as a corpus

The body of work of this thesis relies on two assumptions: that Wikipedians are a representative sample of the population; and that Wikipedia is a representative corpus for the supervised tasks we apply it to. On the whole these assumptions appear to hold, however there have been instances in this thesis where this is clearly not the case. Mika et al. (2008) note the issues that can occur when test and training data are drawn from different corpora. Despite clear differences between corpora I think the approaches presented at disambiguating known placenames are general enough to apply to free text. This has clearly been observed by positive performance on a subset of Brown placenames in Section 6.5 and the GeoCLEF corpus in Section 6.6. On the other hand corpus differences are still clearly visible: in Section 4.6 where the different uses of the terms *party* and *race* are observed between Wikipedia and Flickr; Section 6.6 where one can see a significantly greater skew in references to placenames in the GeoCLEF corpus than Wikipedia; and finally the most common location referred to by the placename “Aberdeen” (Aberdeen, Scotland in GeoCLEF and Aberdeen, Washington in Wikipedia).

It has been observed by Jimmy Wales that there is a core group of 1000-2000 Wikipedians accounting for approximately 75% of the edits in Wikipedia (Swartz 2006). Wales argues that this means a core group of Wikipedians are contributing the bulk of the content. Swartz (2006) refutes this claim. In their small sample they find that although a small community are contributing the bulk of edits, this is largely protecting against vandalism and bringing pages in-line with Wikipedia’s manual-of-style. In fact, the bulk of the *words* in Wikipedia are from a diverse range of contributors who edit only a few articles.

As with many of the debates surrounding Wikipedia, this is not an argument that can be quickly solved. Medelyan et al. (2008) observe the advantage of using Wikipedia as a corpus rather than a general web or text collection is that it is well-written and well-formulated. I think for the purposes of this thesis, Wikipedians are suitably diverse and Wikipedia is suitably general, however, the systematic bias introduced by its content and contributors (partially observed in Chapter 7), will have to be considered when applying this work in a greater context.

8.3 Future work

Possible continuations and additional experiments highlighted by the work presented in this thesis are discussed at the end of each respective chapter. Notably in Chapter 4, evaluation of the accuracy of Flickr tag disambiguation and content based disambiguation for ambiguous tags; and in Chapter 7, event detection in Wikipedia, and a greater selection of evidences considered when formulating the Steinberg hypothesis than just distance and population. Such directions and experiments are ruled out, either because they fall outside the scope of this thesis, other directions show greater promise, or the diminishing returns in pursuing certain experiments. In this section I shall discuss future work limited only by time: areas that this thesis suggests show promise and lie within the scope.

8.3.1 Context based placename disambiguation

Placename disambiguation, as explored in Chapter 6 of this thesis, is only considered with respect to the content of the documents placenames occur in, however Chapter 7 suggests a great deal of information is implied by the context. Section 6.3's example of a placename that cannot be disambiguated is an occurrence of "London" referring to London, Ontario with no disambiguating content. If the location of where the document is published is known this could aid disambiguation. For example the final experiment of Chapter 6 explores placename disambiguation in the GeoCLEF corpus, a collection of newspaper articles from the Los Angeles Times and Glasgow Herald. Only the co-occurring placenames in these articles are used for disambiguation, yet the contextual information (i.e. where the articles are published) is ignored. Chapter 7 suggests that people's location has a significant bearing on the locations they refer to. Mehler et al. (2006) and Liu and Birnbaum (2008) have observed this in newspapers. It would be interesting to see if combining the methods suggested in these two chapters could produce improved results.

8.3.2 Geographic relevance ranking

The co-occurrence model constructed in Chapter 5 has only been applied to placename disambiguation in this thesis; yet in Section 5.9 two possible applications are identified, the second of which is geographic relevance ranking. By unambiguously indexing locations, we can build a much more accurate model of the latent similarity between locations. This latent similarity can be considered an additional relevance measure. To provide an example consider the locations the United States and the United Kingdom. Both these locations have a huge volume of synonyms ("USA," "U.S.," "America," and "Britain," "UK," "Great Britain," to name a few). Considering overlap a crude measure of latent similarity (equation 6.7). The overlap between the placenames "United States" and "United Kingdom," when considered simple noun phrases, is 4.16%, while the overlap between locations is 9.79%. This more accurate approach to latent similarity could improve relevance ranking. This is part of a much larger trend toward indexing documents based on semantic concepts rather than raw content (Naphade and Huang 2002).

8.4 Core contributions

I shall conclude this thesis by revisiting its core contributions. These are in the areas of extracting information from Wikipedia, supervised placename disambiguation, and providing a quantitative model for how people view the world. The findings clearly have a direct impact for applications such as geographically aware search engines, but in a broader context documents can be automatically annotated with machine readable meta-data and dialogue can be enhanced with a model of how people view the world. This could potentially reduce ambiguity and confusion in dialogue between people or computers.

Appendix A

History of Wikipedia

A.1 Introduction

Wikipedia was launched on 15 January 2001 by Jimmy Wales and Larry Sanger. It was a fork of the free encyclopædia Nupedia, which was peer reviewed and written by experts. In June 2001, when Wikipedia had grown to 5,000 articles, Sanger announced “Wikipedia is now useful”¹. The first event that showed the power of Wikipedia was the 9/11 terrorist attacks on the World Trade Center. Wikipedia was updated in real time from a variety of sources allowing people to visit a single site for aggregated news. Wikipedia’s bottom-up approach allows for large scale live updating (Nakayama et al. 2008). Wikipedia continues to document major disasters in real-time, providing arguably unbiased reporting of events such as the Space Shuttle Columbia disaster, the Iraq war and London’s 7/7 terrorist attack.

Alternate language versions of Wikipedia followed the launch of the English Wikipedia, with the German Wikipedia launched in May 2001². By July 2003 the sum of articles in alternate language versions of Wikipedia had overtaken the English Wikipedia with an all Wikipedia total of 350,000 articles. The importance of Wikipedia being multilingual was shown by Jimmy Wales who described it as *an effort to create and distribute a free encyclopedia of the highest possible quality to every single person on the planet in their own language*.

Since 2003 Wikipedia has continued to grow and gain in popularity, doubling in size every year (Burial et al. 2006). Wikipedia has since become the 7th most popular site on the Internet³. Burial et al.’s paper analyses the WikiGraph, the graph found when all Wikipedia articles are considered nodes and all links considered directed arcs. They generated 17 snapshots of Wikipedia between January 2002 and April 2006 and showed the variation of the WikiGraph over time, with respect to in links, out links and connectedness, is very similar to the evolution of the Web. The WikiGraph forms a scale-free network doubling in size every year with over 98.5% of articles being entirely connected. They observed that when a Wikipedia article is added, the number of words can be expected to grow at a linear rate (although the older the article, the faster this growth). The density of the WikiGraph is also increasing, with each article growing on average by a new link every 100 days. They conclude that although in some respects Wikipedia is starting to mature, [as of 2006] it is still growing at an exponential rate.

Nakayama et al. (2008) also identify Wikipedia’s dense WikiGraph as a notable property of Wikipedia allowing topic locality to be measured. Topic locality states that articles sharing similar links will contain

¹http://meta.wikipedia.org/wiki/History_of_Wikipedia

²http://en.wikipedia.org/wiki/Wikipedia:Multilingual_coordination

³<http://www.alexa.com/> 29 May 2008

similar content. They observe the in links and out links of Wikipedia articles follow a typical Zipfian distribution.

The website Conservapedia⁴ was founded in 2006 as an encyclopædia presenting a right-wing, Christian conservative point of view. Many articles support young Earth creationism, are anti-abortion, anti-gay-rights and anti liberal-teaching. Conservapedia further accuses Wikipedia of a liberal, anti-american and anti-christian systematic bias⁵. Conservapedia's allegations of wide spread systematic bias, particularly against Christians and Americans, have been widely dismissed (Metro 2007).

A.2 Anatomy of an article

A well written description on the structure of Wikipedia can be found in Medelyan et al. (2008). This section will cover the content and structure of articles relevant to this thesis. The content of Wikipedia is guided by the Wikipedia Policies and Guidelines⁶. On a macro level the content of Wikipedia is guided by Wikipedia's five guiding pillars⁷, while on a micro level, the style, look and tone of individual articles are governed by Wikipedia's Manual of Style⁸ (MoS).

The five pillars

Wikipedia's five pillars are: Wikipedia is an encyclopedia; Wikipedia has a Neutral Point of View (NPOV); Wikipedia is free content; Wikipedia has a code of conduct; and Wikipedia does not have firm rules. More specifically:

- **Wikipedia is an encyclopedia:** The fact that Wikipedia is an encyclopædia is what makes it interesting for data-mining. Articles in Wikipedia describe a single concept. This differs from other comparable hyper-linked corpora (e.g. web-collections) where the title of a page may be only partially related to much of the content. Also there is a considerable amount of meta-data that is highly coupled to the content of the article and the title of the page.
- **Wikipedia has a Neutral Point of View:** Wikipedia strives to describe all points of view equally, not to bias through omission or opinion. The NPOV of Wikipedia will be discussed in more detail in Chapter 7.
- **Wikipedia is free content:** The fact that Wikipedia is free and available under the GFDL license makes it ideal for academic research (Free Software Foundation Inc. 2002). The advantage of working with a widely used free corpus is that research is comparable and portable.
- **Wikipedia has a code of conduct:** Part of Wikipedia's code of conduct is to be all-encompassing. Anyone can edit Wikipedia and everyone is welcomed (in principle). It is this all-encompassing nature that makes Wikipedia an ideal corpus for extracting world knowledge.
- **Wikipedia does not have firm rules:** The final pillar of Wikipedia is not a positive point with respect to information extraction. The priority for Wikipedia is to make it as user friendly and readable by humans as possible. If deviating from the standard format makes an article easier for

⁴<http://www.conservapedia.com/>

⁵http://www.conservapedia.com/Wikipedia#Liberal_bias

⁶Only policies and guidelines effecting the thesis scope are discussed here. For full details please visit: http://en.wikipedia.org/wiki/Wikipedia:Policies_and_guidelines

⁷http://en.wikipedia.org/wiki/Wikipedia:Five_pillars

⁸http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style

a person to understand, editors will often do this. The consequence of this is Wikipedia articles are highly heterogenous, meaning any information extraction algorithm needs to be adaptable and robust.

Manual of Style (MoS)

Wikipedia's Manual of Style (MoS) dictates the look and feel of articles. The MoS is extensive specifying distinct guidelines for hundreds of different types of article. Editors can deviate from these guidelines at their discretion. In this section I will describe the components of the majority of Wikipedia articles and the features that are used in this thesis.

Figure A.1 shows a sample of the Wikipedia article for the Chrysler Building, the art-deco skyscraper in New York City. Significant features of the article and features used in this thesis are marked by red numbers defined below:

1. **Title:** Article titles must be unique and conform to naming conventions.
2. **First sentence:** The article title or equivalent name must be the subject of the first sentence and appear in bold-face.
3. **Infobox Template:** Article infobox templates display structured information in a uniform format. The example article contains the Skyscraper template. Wikipedia has a total of 39,516 templates.
4. **Time link:** It is common for articles to link to pages describing periods in which significant events with respect to the subject of the article occurred. In Chapter 7 temporal links are extracted and analysed in detail.
5. **Place link:** Locations related to the article are often linked to. The co-occurrences and the distribution of these links form a running theme through Chapters 5, 6 and 7.
6. **Image:** Many Wikipedia articles contain links to images with a caption. These images often depict the subject of the article.
7. **References:** One of the guiding principles of Wikipedia content is that all information should be verifiable. Citing the sources of statements in Wikipedia is increasingly encouraged. Where possible, links to on-line authoritative sources are included.
8. **Internal related pages:** Related Wikipedia articles that are not linked to within this article, are provided as a list at the end of the article.
9. **External related pages:** External related pages providing additional information about the article subject or associated pages of the article subject are provided after all the main page content.
10. **Coordinate link:** Coordinates are a specific type of template. Geographic coordinates linking to a variety of mapping services are present on most articles about subjects with a fixed position (locations, buildings, geological features etc.).
11. **Additional Template:** Additional templates are provided at the bottom of the page, often these contain less information and have a *softer* association than the infobox template. In the example Supertall Skyscrapers is provided as an additional template.

12. **Categories:** Categories provide soft associations between articles and are often quite broad (e.g. 1930's architecture). Categories can also contain administrative information, e.g. whether this has been a featured article. There are 167,583 categories in total linked in a directed graph known as a *Category tree*.
13. **Navigation links:** All Wikipedia articles have the same navigational links in the upper lefthand side. These link to high quality articles, random articles and administration pages.
14. **Permanent link:** Wikipedia pages are continually edited, some changing minute by minute. The permanent link allows you to always return to the page in the state you currently see it.
15. **Interlanguage links:** Interlanguage links provide links between articles in different language versions of Wikipedia discussing the same subject. The linked-to article will often not be a direct translation, usually it is simply on the same topic. This is further discussed in Chapter 7.

A.3 Clustering Wikipedia articles

There are many reasons one may wish to cluster documents: to gain additional information about a corpus, consolidate information, aid browsing and searching, to name but a few.

Once the relatedness between documents has been quantified, it is relatively straightforward to cluster semantically related documents (Kinzler 2005). Several web-sites offer clustered searching of Wikipedia: Clusty⁹ and Carrot Cluster¹⁰ cluster based on article content, while Powerset use a combination of content and *factz*. Carrot Cluster is the only one of these sites that make their clustering methods public. They use the Lingo algorithm which is based on the standard IR concepts of a VSM and Latent Semantic Indexing (more specifically Singular Value Decomposition) (cf. Osinski and Weiss (2005)).

A.4 Wikipedia in research

Wikipedia has become extremely attractive to researchers from many fields, from school projects to university professors. The use of Wikipedia in peer reviewed conference and journal articles is growing rapidly with over 100 publications in 2007 alone¹¹. The Wikimania conference has been running since 2005. It covers studies and experiments conducted on Wikipedia and the culture and technology behind Wikimedia applications. Medelyan et al. (2008) provide a comprehensive review of the use of Wikipedia in research from its creation until mid-2008.


In the area of Information Retrieval, Wikipedia is increasingly being used as a corpus. In 2006 the Initiative for the Evaluation of XML retrieval (INEX) conference began using Wikipedia as a corpus for both structured document and multimedia retrieval (Fuhr et al. 2006). In 2007 more Wikipedia specific tasks were included such as the "Link the Wiki task", which provides an evaluation forum for algorithms that automatically discover in-links and out-links for articles (Huang et al. 2007). Also in 2006 the WiQA task was introduced at CLEF to test NLP and Question Answering technologies on a multi-lingual corpus. Three languages were used: English, Spanish and Dutch.

Since 2006 further corpora have been developed from Wikipedia: the WikipediaMM corpus, which was originally developed for INEX, then adopted by CLEF, is an image collection developed from Wikipedia.

⁹<http://clusty.com/>

¹⁰<http://carrot2.org/>

¹¹http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_in_academic_studies



WIKIPEDIA
The Free Encyclopedia

[Log in / create account](#)

1 article | [discussion](#) | [edit this page](#) | [history](#)

Chrysler Building¹


Chrysler Building ³

2 The **Chrysler Building** is an Art Deco skyscraper in New York City, located on the east side of Manhattan at the intersection of 42nd Street and Lexington Avenue. Standing at 319 metres (1,047 ft),^[1] it was briefly the world's tallest building before it was surpassed by the Empire State Building in 1931. However, the Chrysler Building remains the world's tallest brick building.^{[2][3]} ...

Contents^[hide]

- 1 History
 - 1.1 Design beginnings
 - 1.2 Construction
 - 1.3 Completion
 - 1.4 Property
- 2 Architecture
 - 2.1 Crown ornamentation
 - 2.2 Observation and broadcasting
 - 2.3 Lighting
 - 2.4 Recognition and appeal
- 3...

History [edit]

6  The Chrysler Building was designed by architect **William Van Alen** to house the Chrysler Corporation. When the ground **breaking** occurred on September 19, 1928, **4** there was an intense competition **5** in New York City to build the world's tallest skyscraper.^{[6][7]} Despite a frantic pace (the building was erected at an average rate of four floors per week), no workers died during the construction of this skyscraper.^[8]

7 **References** [edit]

- 8880952307)
- 8** *The Chrysler Building: Creating a New York Icon Day by Day*, D Stravitz, Prinston Architectural Press Publishers, 2002 (ISBN-10: 1568983549)

Notes [edit]

1. [^] The Chrysler Building – SkyscraperPage.com [↗](#)
2. [^] The World's Tallest Brick Building – SkyscraperPicture.com [↗](#)
3. [^] A view from Above – The Chrysler Building [↗](#)
4. [^] Emporis Data – See Tallest buildings Ranking [↗](#)...

See also [edit]

- 9** Buildings and architecture of New York City ⁸
- 50 Tallest buildings in the U.S.
- Tallest buildings in New York City...


External links [edit]

- 10** Live Webcam of the Chrysler Building [↗](#) ⁹
- New York Architecture Images-the Chrysler Building [↗](#)
- Chrysler Building is at coordinates 40°45′06″N 73°58′31″W﻿ / ﻿40.7517, -73.9753﻿ / 40.7517; -73.9753 [↗](#)Coordinates: 40°45′06″N 73°58′31″W﻿ / ﻿40.7517, -73.9753﻿ / 40.7517; -73.9753 [↗](#)

11 **Supertall skyscrapers** [show]

12 Categories: Skyscrapers in New York City | Art Deco buildings in New York City | Skyscrapers between 300 and 349 meters | National Historic Landmarks in New York City | Registered Historic Places in Manhattan | 1930 architecture | Former world's tallest buildings | Buildings and structures in Manhattan

Chrysler Building ³



Chrysler Building was the world's tallest building from 27 May 1930 to 1931.*

Preceded by 40 Wall Street

Surpassed by Empire State Building

Information

Location 405 Lexington Avenue, New York, New York, U.S.

Status Complete

Constructed 1929-1930...

[edit]

navigation

- 13 Main Page
- Contents
- Featured content
- Current events
- Random article

interaction

- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia
- Donate to Wikipedia
- Help

search

Go Search

toolbox

- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link ¹⁴
- Cite this page

languages

- 15 العربية
- Български
- Català
- Česky
- Dansk
- Deutsch
- Español
- Français
- 한국어
- Italiano
- עברית
- ქართული
- Lietuvių
- Magyar
- Nederlands
- 日本語
- Norsk (bokmål)
- Norsk (nynorsk)
- Polski
- Português
- Română
- Русский
- Simple English...



This page was last modified on 30 May 2008, at 05:33. All text is available under the terms of the GNU Free Documentation License. (See [Copyrights](#) for details.)

Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a U.S. registered 501(c)(3) tax-deductible nonprofit charity.

[Privacy policy](#) | [About Wikipedia](#) | [Disclaimers](#)



Figure A.1: Anatomy of a Wikipedia article

The WikiXML¹² dump (originally developed for INEX and CLEF) is a static dump of Wikipedia that is easier to import into a database and manipulate than the dumps provided by Wikimedia. Not to mention the DBpedia download, the freely available XML dumps downloadable from Wikimedia¹³ and *Wikipedia on DVD*¹⁴ – standardised dumps of Wikipedia with versions of articles checked for correctness and vandalism.

As reviewed in this appendix there is a substantial body of work analysing Wikipedia, which this thesis builds upon.

¹²<http://ilps.science.uva.nl/WikiXML/> updated in 2007

¹³<http://download.wikimedia.org/>

¹⁴<http://www.wikipediaondvd.com/>

Appendix B

Further experiments with the GeoCLEF corpus

B.1 Introduction

This appendix contains experiments performed during the development of Forostar, a GIR application developed for the indirect evaluation of placename disambiguation in Chapter 6. It begins with an analysis of whether query classification can guide the selection of disambiguation methods or query formulation. This is followed by a comparison of data fusion techniques. This appendix concludes with a listing of the per-query results analysed in detail in Chapter 6.

B.2 Query classification

Several attempts have been made to categorise the GeoCLEF queries. This categorisation has two purposes:

1. At query time it allows queries of different types to be treated in different ways; and
2. When interpreting results, it allows one to interpret how different methods perform on different query types.

Freq.	Class	Example
80	Non-geo subject restricted to a place.	<i>Shark Attacks off Australia and California.</i>
6	Geo subject with non-geographic restriction.	<i>Cities near active volcanoes.</i>
6	Geo subject restricted to a place.	<i>Cities along the Danube and the Rhine.</i>
2	Non-geo subject associated to a place.	<i>Independence movement in Quebec.</i>
7	Non-geo subject that is a complex function of a place.	<i>Water quality at the coast of the Mediterranean.</i>
0	Geo relations among places.	<i>How are the Himalayas related to Nepal?</i>
1	Geo relations among (places associated to) events.	<i>Did Waterloo occur more north than Agincourt?</i>
0	Relations between events requiring precise localisation.	<i>Was it the same river that flooded last year in which killings occured in the XVth Century?</i>

Table B.1: Tentative classification of geographic topics

Freq.	Class	Example
1	Ambiguity.	<i>St. Paul’s Cathedral.</i>
19	Vague geographic regions.	<i>Credits to the former Eastern Bloc.</i>
11	Geographic relations beyond IN.	<i>Cities within 100km of Frankfurt.</i>
0	Cross-lingual issues.	<i>Greater Lisbon. Portuguese: Grande Lisboa. German: Großraum Lissabon.</i>
14	Granularity below the country level.	<i>Murders and violence in South-West Scotland.</i>
5	Complex region shapes.	<i>Wine regions around rivers in Europe.</i>

Table B.2: Explicit topic difficulties

Freq.	Class	Example
89	Geographic scopes can easily be resolved to a place.	<i>Shark Attacks off Australia and California.</i>
11	Geographic scopes cannot be resolved to a place.	<i>Cities near active volcanoes.</i>

Table B.3: Topic categorization

Freq.	Class	Example
12	Geographic Feature.	<i>Cities near active volcanoes.</i>
7	Body of water.	<i>Sea rescue in North Sea.</i>
17	Continent.	<i>Trade Unions in Europe.</i>
29	Country.	<i>Japanese Rice Imports.</i>
9	State / County.	<i>Independence movement in Quebec.</i>
6	City.	<i>Cities within 100km of Frankfurt.</i>
0	Smaller than city.	
26	Imprecise region.	<i>Malaria in the tropics.</i>

Table B.4: Classification by feature type

The interpretation of results with respect to query classification guides the GeoCLEF organisers in formulating their queries for the following years. These two purposes for query classification can also be considered an iterative process. Alternative methods can continually be tested for each query type and the best performing methods for each type presented to the user.

Gey et al. (2006) present a “Tentative Classification of Geographic Topics,” (Table B.1). The purpose of this classification schema is to demonstrate that the relationship between subject and location can vary considerably¹. Mandl et al. (2007) present a schema of “Explicit topic difficulties,” which were used when constructing the 2007 queries to keep the queries challenging (Table B.2). Andogah and Bouma (2007) present a simplified version of Gey et al. (2006)’s schema. Their “Topic Categorization” splits queries into topics where a specific location is intended and topics where the intended location is existential (Table B.3).

For comparison I also introduce two low level classification schema. The first classifies the geographic topic by feature type (Table B.4); the second classifies geographic topic by location (Table B.5). The choice of classes for the classification by location schema is governed by the bias of the corpus (illustrated in Figure 6.5). The GeoCLEF English corpus is taken from the LA Times and Glasgow Herald. As such the classification goes to a finer classification for Scotland and California than the rest of the world. Gan et al. (2008) suggest that locations of different size/granularity imply different intentions.

Using the 75 2005-07 queries as training data and the 25 2008 queries as test data, I explored whether there was a correlation between improved average precision and either query formulation (detailed in Section 6.6.4) or disambiguation method (detailed in Section 6.4). The motivation behind this was to be

¹Freq. denotes the frequency that a specific class description occurs; the description is followed by an example in italics. Not every query fits into a class and the classes are overlapping.

Freq.	Class	Example
9	Scotland.	<i>Walking holidays in Scotland.</i>
1	California.	<i>Shark Attacks off Australia and California.</i>
3	USA (excluding California).	<i>Scientific research in New England Universities.</i>
7	UK (excluding Scotland).	<i>Roman cities in the UK and Germany.</i>
46	Europe (excluding the UK).	<i>Trade Unions in Europe.</i>
16	Asia.	<i>Solar or lunar eclipse in Southeast Asia.</i>
7	Africa.	<i>Diamond trade in Angola and South Africa.</i>
1	Australasia.	<i>Shark Attacks off Australia and California.</i>
3	North America (excluding the USA).	<i>Fishing in Newfoundland and Greenland.</i>
2	South America.	<i>Tourism in Northeast Brazil.</i>
8	Other Specific Region.	<i>Shipwrecks in the Atlantic Ocean.</i>
6	Other.	<i>Beaches with sharks.</i>

Table B.5: Classification by location

able to switch query formulation or geographic index based on query classification to provide a synergy between methods. No significant correlation was found. Due to these disappointing results method selection based on query analysis has not been further pursued in this thesis.

B.3 Data fusion

In 2005 when the first GeoCLEF track was run, the structuring of the queries implied one was expected to build a textual retrieval engine, a geographic retrieval engine, and then fuse the results. In fact the first query set was split along these lines into concept, location and spatial relation (Gey et al. 2006). In recent years opinion has become divided whether this is the correct approach and whether all evidence needs to be considered simultaneously (Cardoso and Santos 2008). This split is partially motivated by the relative maturity of textual retrieval compared to geographic retrieval, i.e. it is easier to take a mature system and bolt on a geographic module than build a GIR system from scratch. This section contains a brief review of methods to fuse geographic and textual relevance, followed by an experiment to select the optimal data fusion method to use in the Forstar GIR system (described in Section 6.6).

Traditional data fusion combines a number of ranks or filters at query time. The difference between a rank and a filter is a rank is ordered (assumes scored retrieval), while all documents contained in a filter are considered of equal relevance (assumes binary retrieval). Textual retrieval generally produces ranked results. Geographic relevance ranking is a less mature and less studied area than text retrieval. Some GIR systems have complex geographic relevance ranking methods based on topological distance, geographic distance, overlap etc. (Buscaldi and Rosso 2008a). Other systems employ a more simple filter based approach, where all documents relevant to the geographic query are considered equal (Ferrés and Rodríguez 2007; Hauff et al. 2006). Martins et al. (2005) outline four ways of combining text and geographic relevance: linear combination of similarity², product of similarity, maximum similarity or a step-linear function (equivalent to filtering). The most common ways of combining a geographic and text rank is either as a convex combination of ranks (Overell et al. 2008b) or scores (Cardoso et al. 2008, 2007). The advantage of using ranks rather than scores is that the distribution of scores produced by different relevance methods may differ greatly. This problem can be mitigated by normalising the scores (Martins et al. 2006). Using ranks rather than scores has the disadvantage that information is discarded.

²Where similarity refers to some similarity measure between query and document such as rank or score.

Disambig. Method	Penalisation Values
MR	2.0
SVM-tf.idf	3.0
Neigh.	8.0
NoDis.	3.0

Table B.6: Penalisation values found by using results from GeoCLEF 2005-07 as training data

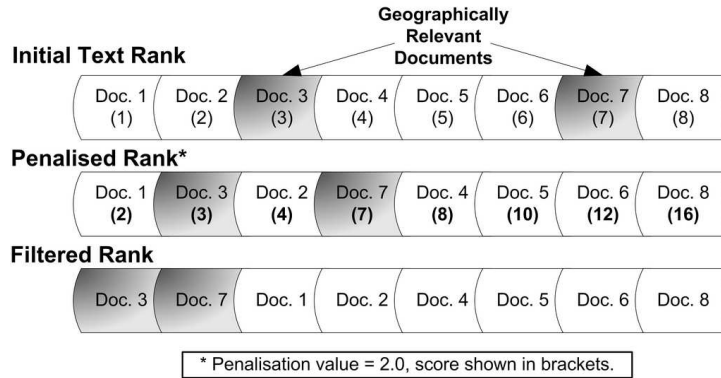


Figure B.1: Rank example

Methods

Two different data fusion methods were considered for Forostar, described below:

Penalisation

The penalisation method multiplies the rank r , of each element in the text rank that is not in the geographic filter by a penalisation value p , to give an intermediate rank r' . This can be described by the filter function $f(r)$

$$f(r) = \begin{cases} r & \text{doc } r \text{ present in filter} \\ rp & \text{doc } r \text{ not present in filter} \end{cases} \quad (\text{B.1})$$

The intermediate rank is sorted by r' to give the final returned rank. The penalisation value p is found using a brute force search, using the 75 queries and relevance judgements from GeoCLEF 2005-07 as training data. The search finds the value of p that maximises MAP. The p values found for the four disambiguation methods considered from Chapter 6 are shown in Table B.6.

Filtering

The filtering method reorders the text rank in a more aggressive way than the penalisation method. All the results of the text rank that are also contained in the geographic filter are returned first, followed by the text results that are not in the geographic filter. The filter method and text baseline are both equivalent to the penalisation method with p values of a high value (e.g. 1000) and one respectively.

An example of these methods and the text baseline are shown in Figure B.1. It shows a hypothetical text rank containing two entries also in the geographic filter. The penalisation method calculates r' , shown in brackets, to re-order the results, while the filter method simply promotes all the documents also in the geographic filter to the top of the rank.

Disambig.	Fusion	MAP	Geo AP
Text Baseline		24.1	6.52
MR	Penalis.	18.9	5.12
SVM-tf.idf	Penalis.	18.9	5.17
Neigh.	Penalis.	19.0	5.24
NoDis.	Penalis.	18.9	5.17
MR	Filter	24.5	11.22
SVM-tf.idf	Filter	24.4	7.85
Neigh.	Filter	24.2	7.88
NoDis.	Filter	26.4	10.98

Table B.7: GeoCLEF 2008 results

Results

The results of the nine runs (four disambiguation methods combined with two data fusion methods plus a text baseline) are displayed in Table B.7. Notice that combining the geographic information using the penalisation filter actually gives worse results than the text baseline. My assumption here is that the penalisation training is over fitting. On the other hand, the filter method outperforms the baseline in every case showing it to be more robust.

Pairwise statistical significance tests were performed on each method with the baseline using a two-tailed Wilcoxon Signed-Rank test rejecting the null hypothesis only when p less than 5%. All the penalisation results were statistically significantly worse than the baseline, and only the NoDis-Filter method was statistically significantly better. Pairwise significance testing was also performed between the penalisation method and filter method runs with the same disambiguation method. This showed that in every case the filter method was statistically significantly better.

Comparison with the rest of the participants at GeoCLEF 2008, can be seen by comparing Table B.7 to the 2008 quartile ranges in Table 6.13. The best result, NoDis-Filter, occurs in the top quartile. The other filtered results and the baseline occur between the Median and Q3. The penalisation results occur in the lower quartile.

Analysis

This appendix has shown that brute force training of a penalisation value to combine text and geographic data is highly sensitive to over fitting. In fact this resulted in a MAP on the test data statistically significantly worse than the baseline or filter methods. These results concur with other studies (Cardoso et al. 2005). Because of these results, the filter method is the data fusion method implemented in Forostar in Chapter 6. I consider these experiments a cautionary tale on how prone the combination of textual and geographic relevance is for over fitting.

Due to the current immaturity of geographic relevance ranking techniques it is my conclusion that more robust training methods or methods such as filtering are the most appropriate for GIR systems. Alternatively post-hoc query analysis could be performed to assess which queries are causing this dramatic over fitting. Wilkins et al. (2006) propose a method of adjusting the weightings given to different features based on the distribution of scores at query time. Such a method may also be appropriate for geographic information retrieval and could help with the search for synergy between textual and geographic evidence.

B.4 Per-query results

Tables B.8-B.11 contain the per-query results for the seven methods evaluated in Section 6.6. A summary of these results is shown in Table 6.12. The entire results are provided here for completeness. Individual queries referenced in Section 6.6.5 are shown in bold.

Query	Text	MR	NoDis	MI	Refer.	Neigh.	tf·idf	prox
Shark Attacks off Australia and California	57.1	33.2	37.5	11.9	33.2	41.6	33.2	33.2
Vegetable Exporters of Europe	9.7	20.0	19.7	26.9	19.9	20.0	19.9	19.9
AI in Latin America	26.3	27.4	26.9	27.4	27.3	27.6	27.4	27.4
Actions against the fur industry in Europe...	8.5	8.9	8.9	8.6	8.9	8.9	8.9	8.9
Japanese Rice Imports	48.8	51.9	51.9	51.9	51.9	51.9	51.9	51.9
Oil Accidents and Birds in Europe	25.4	17.2	25.1	16.7	16.9	18.8	25.9	16.9
Trade Unions in Europe	16.6	12.7	12.8	11.3	12.7	12.7	12.8	12.7
Milk Consumption in Europe	4.4	5.8	5.4	6.7	5.8	5.9	5.8	5.8
Child Labor in Asia	34.6	42.7	41.8	40.1	43.1	42.8	42.7	43.1
Flooding in Holland and Germany	69.9	74.4	74.4	74.4	74.4	74.4	74.4	74.4
Roman cities in the UK and Germany	4.4	4.4	4.8	4.9	4.4	4.4	4.9	4.4
Cathedrals in Europe	1.7	1.7	1.6	1.5	1.7	1.7	1.7	1.7
Visits of the American president to Germany	32.5	41.1	41.1	41.1	42.3	41.1	41.1	42.3
Environmentally hazardous Incidents in the...	19.8	27.9	25.7	27.8	27.9	27.9	27.9	28.0
Consequences of the genocide in Rwanda	68.9	72.0	72.0	72.0	72.0	72.0	72.0	72.0
Oil prospecting and ecological problems in...	81.2	75.1	75.1	75.1	75.1	75.1	75.1	75.1
American Troops in Sarajevo, Bosnia...	34.1	35.5	35.5	35.5	35.5	35.5	35.5	35.5
Walking holidays in Scotland	19.3	21.5	21.4	21.7	21.6	21.4	21.5	21.6
Golf tournaments in Europe	10.9	17.8	17.4	18.1	17.5	18.3	17.2	17.5
Wind power in the Scottish Islands	21.4	8.6	8.6	8.5	8.6	8.6	8.6	8.6
Sea rescue in North Sea	29.2	32.9	31.3	31.6	32.9	32.5	32.9	32.5
Restored buildings in Southern Scotland	21.1	37.9	36.5	33.4	38.0	38.1	37.9	38.0
Murders and violence in South-West Scotland	1.7	4.7	4.4	2.7	4.8	4.7	4.6	4.8
Factors influencing tourist industry in...	45.0	44.4	44.1	41.9	44.4	44.6	44.2	44.4
Environmental concerns in and around the...	38.4	40.7	40.1	40.2	40.7	40.7	40.7	40.7
2005 summary	29.2	30.4	30.6	29.3	20.2	30.8	30.7	30.4

Table B.8: 2005 Per-query results and summary — MAP (%)

Query	Text	MR	NoDis	MI	Refer.	Neigh.	tf-idf	prox
Wine regions around rivers in Europe	8.2	9.5	9.0	10.8	9.5	9.6	9.5	9.5
Cities within 100km of Frankfurt	1.2	1.7	1.7	0.7	1.7	1.7	1.7	1.7
Snowstorms in North America	2.6	3.0	2.8	3.3	3.0	3.0	3.0	3.0
Diamond trade in Angola and South Africa	22.2	21.0	21.0	21.0	21.0	21.0	21.0	21.0
Car bombings near Madrid	82.8	93.1	93.1	82.8	78.8	94.8	93.1	78.8
Combats and embargo in the northern part of Iraq	36.1	33.7	33.7	33.7	33.7	33.7	33.7	33.7
Independence movement in Quebec	67.9	68.2	68.2	47.4	68.2	68.2	68.2	68.2
International sports competitions in the Ruhr area	0.1	0.0	0.7	0.0	0.0	0.1	0.0	0.1
Malaria in the tropics	61.0	46.9	46.8	11.5	46.9	46.9	46.9	46.9
Credits to the former Eastern Bloc	0.3	1.0	1.0	1.0	1.0	1.1	1.0	1.1
Automotive industry around the Sea of Japan	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Archeology in the Middle East	3.8	4.1	4.1	4.2	4.1	4.2	4.1	4.1
Solar or lunar eclipse in Southeast Asia	8.3	16.7	16.7	16.7	16.7	16.7	16.7	16.7
Russian troops in the southern Caucasus	6.2	22.4	22.4	18.3	20.8	22.9	22.4	20.8
Cities near active volcanoes	21.1	21.1	21.1	21.1	21.1	21.1	21.1	21.1
Shipwrecks in the Atlantic Ocean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regional elections in Northern Germany	7.8	11.0	11.0	11.1	11.0	11.1	11.0	11.0
Scientific research in New England Universities	2.7	0.5	2.8	0.6	0.5	0.7	0.8	0.5
Arms sales in former Yugoslavia	12.4	18.0	17.8	18.0	18.0	18.0	18.0	18.0
Tourism in Northeast Brazil	0.5	1.1	1.1	1.1	1.1	1.1	1.1	1.1
Forest fires in Northern Portugal	76.7	46.8	46.8	46.8	46.8	46.8	46.8	46.8
Champions League games near the Mediterranean	1.4	3.0	2.9	3.2	3.0	3.5	3.1	3.2
Fishing in Newfoundland and Greenland	70.7	68.1	67.7	56.6	64.3	68.3	64.1	64.3
ETA in France	58.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Cities along the Danube and the Rhine	24.9	33.6	26.3	24.8	33.6	36.6	33.7	36.1
2006 summary	23.1	21.0	20.7	17.4	20.2	21.2	20.8	20.3

Table B.9: 2006 Per-query results and summary — MAP (%)

Query	Text	MR	NoDis	MI	Refer.	Neigh.	tf-idf	prox
Oil and gas extraction found between the UK and ...	49.8	56.1	55.5	56.4	56.1	56.2	55.9	56.1
Crime near St Andrews	0.3	0.2	0.1	0.3	0.2	0.2	0.1	0.3
Scientific research at east coast Scottish Universities	13.7	14.4	14.3	13.7	14.0	14.7	14.6	13.0
Damage from acid rain in northern Europe	7.9	9.2	8.8	10.8	9.2	9.2	9.1	9.2
Deaths caused by avalanches occurring in Europe...	10.1	15.3	14.6	17.8	15.3	15.7	15.3	15.3
Lakes with monsters	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7
Whisky making in the Scottish Islands	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Travel problems at major airports near to London	2.0	4.2	36.2	16.7	4.4	16.8	4.2	16.8
Meetings of the Andean Community of Nations	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Casualties in fights in Nagorno-Karabakh	72.0	46.9	46.9	46.9	46.9	46.9	46.9	46.9
Airplane crashes close to Russian cities	10.5	13.4	20.8	20.8	13.4	13.4	13.4	13.4
OSCE meetings in Eastern Europe	19.3	23.7	23.6	32.2	25.4	32.7	23.7	32.7
Water quality along coastlines of the Mediterranean...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Sport events in the french speaking part of...	1.2	2.1	2.23	2.3	2.3	2.23	2.3	2.3
Free elections in Africa	36.0	37.5	37.4	38.0	37.5	37.5	37.5	37.5
Economy at the Bosphorus	16.8	14.8	14.8	14.8	14.8	14.8	14.8	14.8
F1 circuits where Ayrton Senna competed in 1994	40.3	40.3	40.3	40.3	40.3	40.3	40.3	40.3
Rivers with floods	63.3	63.3	63.3	63.3	63.3	63.3	63.3	63.3
Death on the Himalaya	12.94	10.4	10.4	10.5	10.4	10.4	10.4	10.4
Tourist attractions in Northern Italy	2.6	6.7	6.7	6.8	6.9	71.9	6.7	6.9
Social problems in greater Lisbon	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Beaches with sharks	45.4	45.4	45.4	45.4	45.4	45.4	45.4	45.4
Events at St. Paul's Cathedral	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
Ship traffic around the Portuguese islands	34.5	52.8	52.8	52.8	52.8	52.8	52.8	52.8
Violation of human rights in Burma	36.9	43.0	43.0	42.9	43.0	43.0	43.0	43.0
2007 summary	19.2	20.2	21.7	21.5	23.6	23.7	20.2	21.0

Table B.10: 2007 Per-query results and summary — MAP (%)

Query	Text	MR	NoDis	MI	Refer.	Neigh.	tf-idf	prox
Riots in South American prisons	41.4	6.7	4.0	33.3	6.7	6.7	6.7	6.7
Nobel prize winners from Northern European...	33.2	31.7	36.7	31.96	35.6	31.7	33.2	35.6
Sport events in the Sahara	21.2	29.2	29.0	23.5	23.0	29.2	29.2	23.0
Invasion of Eastern Timor's capital by Indonesia	80.0	80.0	80.0	80.0	80.0	80.0	80.0	80.0
Politicians in exile in Germany	12.6	12.7	12.7	12.7	12.7	12.7	12.7	12.7
G7 summits in Mediterranean countries	4.7	6.9	6.9	3.8	6.9	6.9	6.9	6.9
Agriculture in the Iberian Peninsula	0.05	1.0	0.7	1.0	1.0	1.0	0.8	1.0
Demonstrations against terrorism in Northern...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Documents mentioning bomb attacks in Northern...	49.9	50.0	49.9	50.0	50.0	50.0	50.0	50.0
Nuclear tests in the South Pacific	28.8	28.5	28.5	34.4	28.5	29.1	28.5	28.5
Most visited sights in the capital of France and...	1.63	2.1	2.1	2.3	2.1	2.1	2.1	2.1
Unemployment in the OECD countries	17.4	16.8	16.7	16.60	16.7	16.8	16.8	16.7
Portuguese immigrant communities in the world	20.8	20.8	20.8	20.8	20.8	20.8	20.8	20.8
Trade fairs in Lower Saxony	0.0	8.3	7.7	0.0	10.0	0.0	0.0	10.0
Environmental pollution in European waters	15.4	19.8	18.3	19.7	19.9	19.5	19.6	19.9
Forest fires on Spanish islands	51.8	50.5	100.0	100.0	50.5	50.5	50.5	50.5
Islamic fundamentalists in Western Europe	6.3	10.6	10.1	11.6	10.6	10.6	10.6	10.6
Attacks in Japanese subways	81.2	82.2	82.2	82.2	82.2	82.3	82.2	82.2
Demonstrations in German cities	31.6	18.7	18.62	33.6	18.7	18.7	18.6	18.7
American troops in the Persian Gulf	40.5	49.9	49.9	46.0	49.9	49.9	49.9	49.9
Economic boom in Southeast Asia	24.1	25.3	25.2	25.2	25.4	25.37	25.3	25.4
Foreign aid in Sub-Saharan Africa	8.3	9.0	9.0	10.3	9.0	9.0	9.0	9.0
Tibetan people in the Indian subcontinent	20.7	32.4	32.4	32.4	32.4	32.4	32.4	32.4
Floods in European cities	9.0	15.7	16.4	21.1	15.9	16.1	18.7	15.9
Natural disasters in the Western USA	2.1	4.9	3.9	1.7	5.1	5.8	5.6	5.1
2008 summary	24.1	24.5	26.5	27.8	24.5	24.3	24.4	24.5

Table B.11: 2008 Per-query results and summary — MAP (%)

Appendix C

Constructed languages

C.1 Introduction

Constructed languages are languages that have been consciously devised by an individual or group as opposed to having evolved naturally¹. Constructed languages can be divided into three categories:

- **Engineered Languages.** Languages designed for experimentation in Logic, AI and Linguistics.
- **Auxiliary Languages.** Languages designed for international communication. The first of these was Volapük and the most successful, Esperanto.
- **Artistic Languages.** Languages designed for aesthetic pleasure. The most well known of these is Tolkien's family of related fictional languages, and more recently Klingon² from the Star Trek series of films and television.

Versions of Wikipedia exist in both auxiliary and artistic languages, although recently the latter has been discouraged as against Wikipedia's editorial policy. Although Volapük contains more articles than any other constructed language on Wikipedia, other measures such as number of internal-links, edits per article and active Wikipedians suggest Esperanto is the most *active* constructed language Wikipedia (Wikimedia 2008).

C.2 Esperanto

Esperanto is the most widely used constructed language, developed by Ludwig Lazarus Zamenhof between 1872 and 1885, it is spoken by 200-2000 people natively and around 2,000,000 people as a second language (Gordon 2005). In this section the Esperanto Wikipedia is analysed with the same methods as the 12 natural languages of Chapter 7.

Table C.2 shows a summary of the geographic and temporal co-occurrence models mined from the Esperanto Wikipedia (comparable to Tables 7.2 and 7.9). It has a notably low proportion of ambiguous placenames as one would expect from a constructed language. The proportion of temporal links is particularly high, equalled only by Japanese. Figure C.1 plots the distribution of location references. The scaling exponent is **-1.12** and the spatial autocorrelation of the distribution is **0.890**. This distribution is plotted in two dimensions as a heat map and as a distorted cartogram in Figure C.2. Notice the

¹http://en.wikipedia.org/w/index.php?title=Constructed_language&oldid=252142213

²<http://klingson.wikia.com/wiki/ghItlh'a>

	Esperanto
No. of Articles	154,038
Links Extracted	2,319,434
prop Loc (%)	5.5
prop Ambig (%)	1.3
prop Years (%)	12.1
Articles disambig	10,174
Unique placenames	10,589
prop Ambig (%)	0.3
Unique Locations	7,887

Table C.1: Summary of the Esperanto co-occurrence model

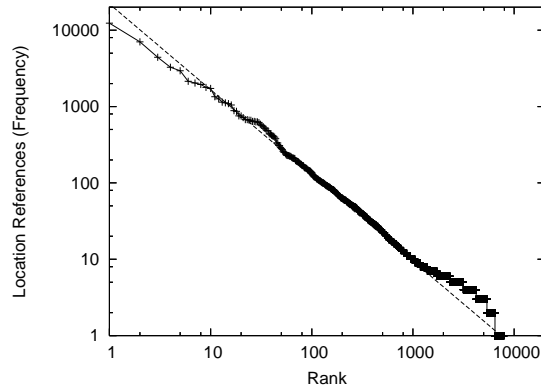


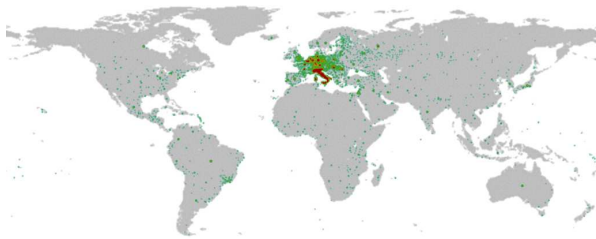
Figure C.1: Distribution of locations in the Esperanto Wikipedia

maps are heavily skewed toward eastern and central Europe. This skew can also be seen when we look at the top five locations referred to in the Esperanto Wikipedia: Germany, Italy, Hungary, France and Romania.

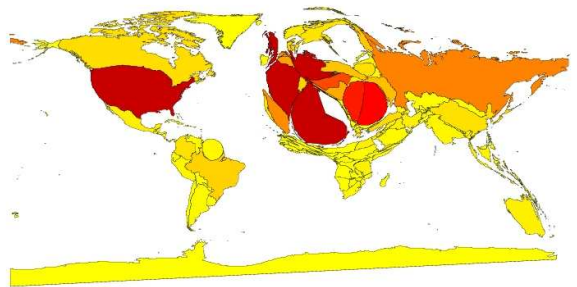
Looking at the Jensen-Shanon Divergence between the co-occurrence models of the Esperanto Wikipedia and those of the languages looked at in Chapter 7 (Table C.2), one can see the distribution of locations referenced is most similar to the Russian Wikipedia, followed by French and German.

	Germ.	Fren.	Pol.	Jap.	Port.	Span.	Russ.	Chin.	Arab.	Hebr.	Welsh	Eng.
Espe.	0.203	0.190	0.227	0.207	0.276	0.231	0.142	0.265	0.308	0.256	0.326	0.339

Table C.2: Jensen-Shanon Divergence between the Esperanto Wikipedia and other wikipedias w.r.t. locations



Heat map



Cartogram

Figure C.2: Maps of the references to locations in the Esperanto Wikipedia

Nomenclature

This nomenclature describes concepts and notation that have been created or assigned specific definitions for the purposes of this thesis. This section is designed to be referred to as needed rather than to be read as a whole.

People, placenames and locations

Let $P(l)$ be the set of placenames (in a corpus or otherwise) that refer to location l . Similarly, let $L(p)$ be the set of locations (in a corpus or otherwise) that are referred to by placename p .

An *annotated model*, or short *model* refers to a *corpus* where all the placenames are annotated as locations. *Free text* refers to unannotated, unstructured text fields where the placenames and locations referred to are unconstrained. Where it is not implicit let the corpus C be specified thus $P^C(l)$ and $L^C(p)$, where the corpus can be a document collection, annotated model, gazetteer or otherwise. A *co-occurrence model* is a model where the only information recorded is the placenames occurring in documents, corresponding locations and the order they occur.

Let $\text{ref}(p, l)$ be the number of references made to location l by placename p within a model. Let

$$L_1(p), L_2(p), \dots, L_{|L(p)|}(p)$$

be an enumeration of $L(p)$ such that

$$\text{ref}(p, L_1(p)) \geq \text{ref}(p, L_2(p)) \geq \dots \geq \text{ref}(p, L_{|L(p)|}(p)).$$

Normally this will be unique. Note that $L_1(p)$ will be the location most commonly referred to by the placename p and that p is unambiguous when $|L(p)|$ is equal to one.

Let N be the multi-set of all location references in a corpus C , M the set of all unique placenames, K the set of placenames referring to more than one location and L the set of unique locations. This will be represented as N^C , M^C , K^C and L^C when C is not implicit. Let $|F|$ denote the size of a general set F ; naturally $|N| \geq |M| \geq |K|$ and $|N| \geq |L|$. Let \mathcal{C} be the set of all countries and $\text{ref}(c)$ refer to the number of references made to locations within country c . Let $\text{pop}(x)$ be the population of a country or location, and $\text{prop}(c, v)$ be the proportion of people living in c that speak language v .

Let $\overline{R_L}$ refer to the average number of references to a location,

$$\overline{R_L} = \frac{|N|}{|L|},$$

and $\overline{R_C}$ refer to the average number of references to a country,

$$\overline{R_C} = \frac{1}{|C|} \sum_{c \in C} \text{ref}(c).$$

Let ρ denote a person, and $\text{dist}(\rho, l)$ denote the geodesic distance between a person and a location. Similarly let the distance between two locations l_x and l_y be denoted $\text{dist}(l_x, l_y)$. Let \mathcal{P}_v be the set of people that speak language v , and A_v the set of locations where v is spoken such that every person is in a location where their language is spoken:

$$\forall \rho \in \mathcal{P}_v \exists l \in A_v \text{dist}(\rho, l) = 0.$$

Consider \mathbf{O} the observed distribution of locations in the co-occurrence model for corpus C . \mathbf{O} can be considered a histogram with corresponding unit histogram \mathbf{O}' . Consider the collection C_l the sub-corpus of documents in C that contain the location l . A co-occurrence model generated from this corpus can be considered a per-location model. Given a location l_x , let \mathbf{X} refer to the distribution of locations in C_{l_x} , and \mathbf{p}_X refer to the probability distribution of \mathbf{X} such that $\mathbf{p}_X(l_y)$ denotes the probability of l_y occurring in \mathbf{X} .

Glossary

In this Glossary acronyms, abbreviations and notation referred to in this thesis are defined.

- Anaphora* A linguistic element that refers back to another element, page 29
- ANNIE* A Nearly-New Information Extraction system, a plugin for the GATE toolkit, page 29
- AP* Average Precision, a retrieval metric, the average of precisions computed at each relevant document rank, page 49
- API* Application Programming Interface. A software interface presented by a program or system, page 40
- Article (Wikipedia)* One of a subset of Wikipedia pages of only encyclopædic content, page 41
- Bag of words* A document model that assumes documents to be a multiset of words, page 27
- BM25* Term weighting scheme for the Vector Space Model developed by S. Robertson, page 28
- Bootstrapping (Machine learning)* A general technique that iteratively trains and evaluates a classifier using features of a larger corpus in order to improve performance achievable with a small training set, page 33
- Category (Wikipedia)* Wikipedia article categories provide soft information about articles and are often quite broad. Every article should have at least one category, page 148
- Category Tree* The directed graph of Wikipedia categories, page 148
- Clarity (Language model)* The information content of an entity model with respect to the corpus, page 89
- ClassTag* A Wikipedia article and tag classification system developed for the experiments in this thesis, page 55
- ClassTag⁺* The ClassTag article-and-tag-classification-system optimised for precision, page 63
- CLIR* Cross Language Information Retrieval, page 31
- Complex phrases* Long phrases composed of one or more shorter noun phrases, page 28
- D_{JS}* Jensen-Shannon divergence, page 96
- D_{KL}* Kullback-Leibler divergence, page 89

<i>DB</i>	Database, page 16
<i>DBpedia</i>	An RDF mapping of Wikipedia articles to WordNet synsets and other linked data sets, page 57
<i>DCV</i>	Document Cut-off Value, page 47
<i>Default Gazetteer</i>	A specially constructed gazetteer where there is only a single location for each placename, page 31
<i>Dictionary phrases</i> ...	Multi-word phrases occurring in dictionaries, page 28
<i>Dublin Core</i>	A meta-data standard, page 39
<i>Dump (Wikipedia)</i> ...	Downloadable version of the Wikipedia database, page 150
<i>EXIF</i>	EXchangeable Image File format – a specification allowing additional meta-data tags to be embedded in images including camera specifications and settings, and, temporal and geographic data, page 30
<i>F-measure</i>	Weighted harmonic mean of precision and recall, page 48
<i>Feature type</i>	The feature type of a placename is the type of the respective location e.g. Region/County, Populated place or Capital, page 93
<i>Flickr</i>	A popular photo sharing web site, page 30
<i>Folksonomy</i>	The bottom-up classification systems that emerge from social tagging, page 65
<i>Footprint</i>	A polygon or collection of polygons representing all the geographic areas referred to and implied by a document or query, page 35
<i>Forostar</i>	Forostar is the GIR system developed to perform the retrieval experiments shown in this thesis, page 104
<i>GATE</i>	Sheffield University’s General Architecture for Text Engineering toolkit – a Natural Language Processing and Information Extraction toolkit, page 29
<i>Gazetteer</i>	A mapping of placenames to geographic coordinates often including other meta-data, page 39
<i>GeoHack</i>	A Wikimedia project linking to a series of Geographic Information Services, page 74
<i>GeoRSS</i>	The RSS XML file format augmented with additional geographic meta-data, page 33
<i>GIR</i>	Geographic Information Retrieval, page 16
<i>GIS</i>	Geographic Information Systems, page 16
<i>GMAP</i>	A retrieval metric, the Geometric Mean of the per-query Average Precision values, page 49
<i>GML</i>	Geographic Markup Language, page 39

<i>GNIS</i>	Geographic Names Information System, a Gazetteer produced by the United States Geological Survey, page 39
<i>Google Maps</i>	Google’s map service and API, page 40
<i>GPX</i>	Global Positioning system eXchange format, page 39
<i>Horizontal topology</i> ..	The qualitative relationships between locations bordering each other, page 37
<i>Information Quality</i> .	The fitness for use of information, page 42
<i>Interlanguage links (Wikipedia)</i>	Links between articles in different language versions of Wikipedia discussing the same subject, page 148
<i>IQ</i>	see Information Quality, page 42
<i>IR</i>	Information Retrieval, page 16
<i>KML</i>	Keyhole Markup Language, page 39
<i>Location</i>	A space on the Earth’s surface usually bounded by polygons, page 16
<i>Lucene</i>	Apache’s open source information retrieval engine implemented in Java, page 28
<i>MAP</i>	Arithmetic Mean of the per-query Average Precision values, page 49
<i>MapReduce</i>	A distributed computing framework, page 142
<i>Mashup</i>	A web application that combines data from more than one source into a single integrated tool, page 44
<i>MBB</i>	(a.k.a MBR) Minimum Bounding Box, page 35
<i>MBR</i>	Minimum Bounding Rectangle, page 35
<i>MeSH</i>	Medical Subject Headings, a controlled vocabulary for life-sciences, page 101
<i>Named entities</i>	(a.k.a. proper-names) Names of people places and organisations, page 28
<i>NER</i>	Named Entity Recognition, page 29
<i>NIMA</i>	National Imagery and Mapping Agency, currently National Geospatial-Intelligence Agency, produces a gazetteer of the United States, page 39
<i>NLP</i>	Natural Language Processing, page 46
<i>Numenore</i>	A web-application built to demonstrate the work presented in Chapter 7 of this thesis, page 138
<i>Nupedia</i>	Precursor to Wikipedia, an encyclopædia written and edited by experts, page 145
<i>Objective relevance</i> ..	A measure of how well a document fulfils an information need regardless of user, page 28
<i>Page (Wikipedia)</i>	A web page accessible on the Wikipedia web-site, page 41

- Paradigmatic association* A relationship between terms that can replace one-another in a sentence without effecting the grammaticality or acceptability of the sentence, page 90
- Partial agreement* When measuring agreement between assessors, when there exists an article classification that all assessors agree on, the assessors are said to be in partial agreement, page 63
- Perseus Digital Library* A digital library covering a wide range of topics particularly classics and 19th century American literature, page 30
- Phrase* Several words treated as a single token by an IR or NLP system, page 28
- Pipeline (Algorithm)* . A set of steps or rules applied one after another, page 32
- Placename* A phrase used to refer to a location, page 16
- Polynym* A word having multiple ambiguous meanings, page 41
- Proper names* see named entities, page 28
- R-tree* An efficient way for indexing two dimensional regular data, a common way of indexing geographic data, page 34
- Referent ambiguity* . . . The specific entity being referred to is ambiguous, page 31
- Relevance* The measure of how well a document fulfils an information need, page 27
- Revert (Wikipedia)* . . . A one click undo of another user’s edit, page 42
- RIA* Rich Internet Application, an Internet application where processing is performed client side and data is held server side, page 40
- RSS* Really Simple Syndication, a standardised XML file format that allows information to be published once and viewed by many different programs, page 33
- Saul Steinberg* A Romanian-born American cartoonist and illustrator, referred to due to his most famous illustration “View of the World from 9th Avenue”, page 128
- Semantic ambiguity* . . . The type of entity being referred to is ambiguous, page 31
- Simple phrases* A noun phrase of 2–4 words containing no sub noun phrases, page 28
- Spatial autocorrelation* A measure of how correlated a collection of geographically distributed values are, page 116
- Spatial database* A standard database with additional features for handling spatial data, page 33
- SPIRIT* Spatially-aware Information Retrieval on the Internet – an EU project to develop geographically aware search engines, page 30
- SSD* Symposium for Spatial Databases, now SSTD – Symposium for Spatial and Temporal Databases, page 33
- SSTD* see SSD, page 33

<i>Steinberg hypothesis</i> .	A hypothesis put forward by this thesis that everyone views the world in a similar way with respect to their locality, page 128
<i>Structural ambiguity</i> .	The structure of the words constituting a named entity are ambiguous, page 31
<i>Subjective relevance</i> . . .	The relevance of how well a document fulfils an information need for a specific user, page 27
<i>SVM</i>	Support Vector Machine, a supervised learning method used to partition a multi-dimensional space, page 58
<i>Synset (WordNet)</i> . . .	Indexed semantic definition, page 47
<i>Syntagmatic association</i>	A relationship between terms that co-occur statistically significantly often, page 90
<i>Systematic Bias (Wikipedia)</i>	Significant omissions across multiple articles, page 124
<i>Template (Wikipedia)</i>	Article templates display structured information in a uniform format, page 147
<i>tf-idf</i>	Term frequency · Inverse Document Frequency, a term weighting scheme, page 28
<i>tf-il</i>	Term Frequency · Inverse Layer, a weighting function proposed in this thesis to be used when classifying hierarchical data, page 60
<i>TGN</i>	The Getty Thesaurus of Geographical Names, a gazetteer produced by the J. Paul Getty Trust, page 39
<i>Title (Wikipedia)</i>	Each Wikipedia article has a unique title conforming to naming conventions, page 147
<i>Total agreement</i>	When measuring agreement between assessors, if assessors agree on all classifications total agreement is said to exist, page 63
<i>UGC</i>	User generated content, page 40
<i>Vertical topology</i>	The hierarchical, qualitative relationships between locations, page 37
<i>View of the World from 9th Avenue</i>	Perspective illustration by Saul Steinberg, cover of the New Yorker, 29 March 1976, page 128
<i>VSM</i>	Vector Space Model, page 28
<i>Web 2.0</i>	An umbrella term for the <i>second generation of internet services</i> , page 40
<i>Weighting function</i> . . .	The function used to determine the scalar values of features in a machine learning classifier, page 60
<i>Wiki</i>	A web application that allows users to add, edit and link web pages, page 19
<i>WikiGraph</i>	The scale-free network found by considering every Wikipedia article a node and every internal link an arc, page 145
<i>Wikipedia</i>	An Internet encyclopædia that anyone can contribute to, page 40

<i>Wikipedia's five guiding pillars</i>	General policies and philosophy guiding the content of Wikipedia articles on a macro scale, page 146
<i>Wikipedia's Manual of Style</i> (a.k.a. MoS)	General policies and philosophy guiding the content of Wikipedia articles on a micro scale, page 146
<i>Wikipedians</i>	The 6 million users who contribute content to Wikipedia articles, page 41
<i>WSD</i>	Word Sense Disambiguation, page 31
<i>XML</i>	eXtensible Markup Language, a common format for structured documents, page 16
<i>Zipfian distribution</i> ..	Discrete power-law distribution, page 146

Bibliography

- E. Agichtein and S. Cucerzan. Predicting accuracy of extracting information from unstructured text collections. In *the Conference on Information and Knowledge Management*, pages 567–568. ACM Press, 2005.
- M. Agosti, G. Nunzio, and N. Ferro. Scientific data of an evaluation campaign: Do we properly deal with them? In *Working Notes for the Cross Language Evaluation Forum Workshop*, 2006.
- A. Aksyonoff. *Sphinx 0.9.8 reference manual*, 2008.
- E. Amitay, R. Nelken, W. Niblack, R. Sivan, and A. Soffer. Multi-resolution disambiguation of term occurrences. In *the Conference on Information and Knowledge Management*, pages 255–262. ACM Press, 2003.
- G. Andogah and G. Bouma. University of Groningen at GeoCLEF 2007. In *Working Notes for the Cross Language Evaluation Forum Workshop*, 2007.
- S. Auer and J. Lehmann. What have Innsbruck and Leipzig in common? extracting semantics from wiki content. In *the Semantic Web: Research and Applications*, pages 503–517. Springer-Verlag, 2007.
- A. Axelrod. On building a high performance gazetteer database. In *the HLT-NAACL Workshop on Analysis of Geographic References*, 2003.
- K. Beard and V. Sharma. Multidimensional ranking in digital spatial libraries. *Journal of Digital Libraries*, 1(2):153–160, 1997.
- E. Brill. Unsupervised learning of disambiguation rules for part of speech tagging. In *the Workshop on Very Large Corpora*, 1995.
- T. Brunner and R. Purves. Spatial autocorrelation and toponym ambiguity. In *the CIKM Workshop on Geographic Information Retrieval*, pages 25–26. ACM Press, 2008.
- B. Bucher, P. Clough, D. Finch, H. Joho, R. Purves, and A. Syed. Evaluation of SPIRIT prototype following integration and testing. Technical report, University of Sheffield, 2005.
- R. Bunescu and M. Paşca. Using encyclopedic knowledge for named entity disambiguation. In *the European Chapter of the Association for Computational Linguistics*, pages 9–16. Association for Computational Linguistics, 2006.
- L. Burial, C. Castillo, D. Donata, S. Leonardi, and S. Millozzi. Temporal analysis of the Wikigraph. In *the Web Intelligence Conference*. IEEE, 2006.

- D. Buscaldi and P. Rosso. A comparison of methods for the automatic identification of locations in Wikipedia. In *the CIKM Workshop on Geographic Information Retrieval*, pages 89–92. ACM Press, 2007.
- D. Buscaldi and P. Rosso. The UPV at GeoCLEF 2008: The GeoWorSE system. In *Working Notes for the Cross Language Evaluation Forum Workshop*, 2008a.
- D. Buscaldi and P. Rosso. Map-based vs. knowledgebased toponym disambiguation. In *the CIKM Workshop on Geographic Information Retrieval*, pages 19–22. ACM Press, 2008b.
- D. Buscaldi and P. Rosso. Geo-WordNet: Automatic georeferencing of WordNet. In *Language Resources and Evaluation Conference*, pages 28–30, 2008c.
- D. Buscaldi, P. Rosso, and P. Garcia. Inferring geographic ontologies from multiple resources for geographic information retrieval. In *the SIGIR Workshop on Geographic Information Retrieval*, pages 52–55, 2006.
- N. Cardoso and D. Santos. To separate or not to separate: reflections about current GIR practice. In *the ECIR Workshop on Novel Methodologies for Evaluation in Information Retrieval*, 2008.
- N. Cardoso, B. Martins, M. Chaves, L. Andrade, and M. Silva. The XLDB group at GeoCLEF 2005. In *Working Notes for the Cross Language Evaluation Forum Workshop*, 2005.
- N. Cardoso, D. Cruz, M. Chaves, L. Andrade, and M. Silva. The University of Lisbon at GeoCLEF 2007. In *Working Notes for the Cross Language Evaluation Forum Workshop*, 2007.
- N. Cardoso, P. Sousa, and M. Silva. The University of Lisbon at GeoCLEF 2008. In *Working Notes for the Cross Language Evaluation Forum Workshop*, 2008.
- C. Cleverdon, J. Mills, and M. Keen. Factors determining the performance of indexing systems. Technical report, Cranfield, 1966.
- A. Cliff and K. Ord. Spatial autocorrelation: A review of existing methods and new measures with applications. *Economic Geography*, 46:269–292, 1970.
- P. Clough and M. Sanderson. A proposal for comparative evaluation of automatic annotation for georeferenced documents. In *the SIGIR Workshop on Geographic Information Retrieval*, 2004.
- P. Clough, M. Sanderson, and H. Joho. Extraction of semantic annotations from textual web pages. Technical report, University of Sheffield, 2004.
- P. Clough, A. Al-Maskari, and K. Darwish. Providing multilingual access to Flickr for Arabic users. In *Working Notes for the Cross Language Evaluation Forum Workshop*, 2006a.
- P. Clough, M. Grubinger, T. Deselaers, A. Hanbury, and H. Müller. Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks. In *Working Notes of the Cross Language Evaluation Forum Workshop*, 2006b.
- T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1st edition, 1991.
- G. Crane and A. Jones. The Perseus american collection 1.0. Technical report, Tufts University, 2005.

- N. Craswell, D. Hawking, and P. Thistlewaite. Merging results from isolated search engines. In *the Australasion Database Conference*, pages 189–200, 1999.
- W. Croft, H. Turtle, and D. Lewis. The use of phrases and structured queries in information retrieval. In *the ACM SIGIR conference on Research and Development in Information Retrieval*, pages 32–45. ACM Press, 1991.
- S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 708–716, 2007.
- H. Cunningham, D. Maynard, V. Tablan, C. Ursu, and K. Bontcheva. Developing language processing components with GATE. Technical report, University of Sheffield, 2001.
- DBpedia. DBpedia. <http://dbpedia.org/> Accessed 1 December, 2008.
- DCMI Usage Board. DCMI metadata terms. Technical report, Dublin Core Metadata Initiative, 2006.
- I. Densham and J. Reid. A geo-coding service encompassing a geo-parsing tool and integrated digital gazetteer service. In *the HLT-NAACL Workshop on Analysis of Geographic References*, 2003.
- G. Dutton. Encoding and handling geospatial data with hierarchical triangular meshes. In *the Symposium on Spatial Data Handling*, pages 34–43, 1996.
- M. Egenhofer and D. Mark. Naive geography. In *Spatial Information Theory: A Theoretical Basis for GIS*, pages 1–15. Springer-Verlag, 1995.
- M. Egenhofer and A. Shariff. Metric details for natural-language spatial relations. *ACM Transactions on Information Systems*, 4:295–321, 1998.
- B. El-Geresy, A. Abdelmoty, and C. Jones. Spatio-temporal geographic information systems: A causal perspective. In *Advances in Databases and Information Systems*, pages 191–203. Springer-Verlag, 2002.
- Encyclopædia Britannica Inc. Fatally flawed. Press release, March 2006.
- C. Fellbaum. *WordNet: An electronic lexical database*. MIT Press, 1998.
- D. Ferrés and H. Rodríguez. TALP at GeoCLEF 2007: Using Terrier with geographical knowledge filtering. In *Working Notes for the Cross Language Evaluation Forum Workshop*, 2007.
- R. Florian, S. Cucerzan, C. Schafer, and D. Yarowsky. Combining classifiers for word sense disambiguation. *Journal of Natural Language Engineering*, 8(4):327–342, 2002.
- D. Forsyth. Benchmarks for storage and retrieval in multimedia databases. Technical report, Computer Science Division, University of California at Berkeley, 2001.
- E. Fox and J. Shaw. Combination of multiple searches. In *the Text REtrieval Conference*, pages 243–252. NIST, 1994.
- W. Francis and H. Kucera. *Brown Corpus Manual*. Brown University, 3rd edition, 1979.
- Free Software Foundation Inc. *GNU Free Documentation License*, 2002.

- G. Fu, C. Jones, and A. Abdelmoty. Ontology-based spatial query expansion in information retrieval. In *On the move to meaningful Internet Systems 2005: CoopIS, DOA and ODBASE*, pages 1466–1482. Springer-Verlag, 2005a.
- G. Fu, C. Jones, and A. Abdelmoty. Building a geographical ontology for intelligent spatial search on the web. In *Databases and Applications*, pages 167–172. Springer-Verlag, 2005b.
- N. Fuhr, M. Lalmas, and A. Trotman. Preface of INEX '06. In *the INitiative for the Evaluation of XML retrieval Workshop*, 2006.
- E. Gabrilovich and S. Markovitch. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *the International Joint Conference for Artificial Intelligence*, pages 1606–1611, 2007.
- V. Gaede and O. Günther. Multidimensional access methods. *ACM Computing Surveys*, 20:170–231, 1998.
- W. Gale, K. Church, and D. Yarowsky. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415–439, 1992.
- Q. Gan, J. Attenberg, A. Markowetz, and T. Suel. Analysis of geographic queries in a search engine log. In *the WWW'08 Workshop on Location and the Web*, pages 49–56. ACM Press, 2008.
- E. Garbin and I. Mani. Disambiguating toponyms in news. In *the Joint Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 363–370, 2005.
- F. Gey, R. Larson, M. Sanderson, K. Bischoff, T. Mandl, C. Womser-Hacker, D. Santos, and P. Rocha. GeoCLEF 2006: the cross-language geographic information retrieval track overview. In *Working Notes for the Cross Language Evaluation Forum Workshop*, 2006.
- J. Giles. Internet encyclopaedias go head to head. *Nature*, 438:900–901, 2005.
- R. Gordon. *Ethnologue: Languages of the World*. SIL International, 15th edition, 2005.
- V. Griffith. WikiScanner: List anonymous Wikipedia edits from interesting organizations. <http://wikiscanner.virgil.gr/> Launched 14 August, 2007.
- D. Grossman and O. Frieder. *Information Retrieval*. Springer-Verlag, 2nd edition, 2004.
- J. Guthrie, L. Guthrie, Y. Wilks, and H. Aidinejad. Subject-dependent co-occurrence and word sense disambiguation. In *the Annual Meeting of the Association for Computational Linguistics*, pages 146–152. Association for Computational Linguistics, 1991.
- R. Güting, D. Papadias, and F. Lochovsky. Preface of SSD. In *Advances in Spatial Databases*, pages 1–2. Springer-Verlag, 1999.
- A. Guttman. R-Trees, A dynamic index structure for spatial searching. In *the International Conference on Management of Data*, pages 47–57. ACM Press, 1984.
- D. Harman. Overview of the first Text REtrieval Conference (TREC-1). In *the Text REtrieval Conference*. NIST, 1992.

- P. Harping. *User's Guide to the TGN Data Releases 2.0*. The Getty Vocabulary Program, 2000.
- C. Hauff, D. Trieschnigg, and H. Rode. University of Twente at GeoCLEF 2006. In *Working Notes for the Cross Language Evaluation Forum Workshop*, 2006.
- L. Hill. Core elements of digital gazetteers: Placenames, categories, and footprints. In *Research and Advanced Technology for Digital Libraries*, pages 280–290. Springer-Verlag, 2000.
- L. Hill, M. Goodchild, and G. Janee. Research directions in georeferenced IR based on the Alexandria digital library. In *the SIGIR Workshop on Geographic Information Retrieval*, 2004.
- G. Hobona, P. James, and D. Fairbairn. An evaluation of a multidimensional visual interface for geographic information retrieval. In *the CIKM Workshop on Geographic Information Retrieval*, pages 5–8. ACM Press, 2005.
- D. Huang, Y. Xu, A. Trotman, and S. Geva. Overview of INEX 2007 link the Wiki track. In *the INitiative for the Evaluation of XML retrieval Workshop*, 2007.
- D. Hull. Using statistical testing in the evaluation of retrieval experiments. In *the ACM SIGIR conference on Research and Development in Information Retrieval*, pages 329–338. ACM Press, 1993.
- N. Ide and J. Véronis. Word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–41, 1998.
- iProspect. iProspect search engine user behavior study. White Paper available online at: <http://www.iprospect.com/about/searchenginemarketingwhitepapers.htm>, 2006.
- V. Jijkoun and M. de Rijke. Overview of WiQA 2006. In *Working notes of the Cross Language Evaluation Forum Workshop*, 2006.
- T. Joachims. *Advances in Kernel Methods – Support Vector Learning*, chapter Making large-scale SVM Learning Practical. MIT-Press, 1999.
- C. Jones and R. Purves. Foreword of GIR'06. In *SIGIR Workshop on Geographic Information Retrieval*, page 2, 2006.
- C. Jones, A. Abdelmoty, D. Finch, G. Fu, and S. Vaid. The SPIRIT spatial search engine: Architecture, ontologies and spatial indexing. In *Geographic Information Science*, pages 125–139. Springer-Verlag, 2004.
- R. Jones, A. Hassan, and F. Diaz. Geographic features in web search retrieval. In *the CIKM Workshop on Geographic Information Retrieval*, pages 57–58. ACM Press, 2008.
- M. Keen. Presenting results of experimental retrieval comparisons. *Information Processing and Management*, 28(4):491–502, 1992.
- D. Kinzler. Wikisense – Mining the Wiki. In *Wikimania*. Wikimedia, 2005.
- M. Krötzsch, D. Vrandečić, and M. Völkel. Wikipedia and the semantic web - the missing links. In *Wikimania*. Wikimedia, 2005.
- H. Kucera and W. Francis. *Computational Analysis of Present-Day American English*. Brown University, 1967.

- C. Kuhlthau. Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science*, 42(5):361–371, 1991.
- M. Lalmas. Uniform representation of content and structure for structured document retrieval. Technical report, Queen Mary and Westfield College, University of London, 2000.
- F. Lancaster. *Indexing and Abstracting in Theory and in Practice*. Facet, 3rd edition, 2003.
- R. Larson. Geographic information retrieval and spatial browsing. In *Geographic Information Systems and Libraries: Patrons, Maps and Spatial Information*, pages 81–124. University of Illinois, 1996.
- R. Larson and P. Frontiera. Ranking and representation for geographic information retrieval. In *the SIGIR Workshop on Geographic Information Retrieval*, 2004.
- J. Leidner. Towards a reference corpus for automatic toponym resolution evaluation. In *the SIGIR Workshop on Geographic Information Retrieval*, 2004a.
- J. Leidner. Toponym resolution in text: “Which Sheffield is it?”. In *the SIGIR Doctoral Consortium*, 2004b.
- J. Leidner, G. Sinclair, and B. Webber. Grounding spatial named entities for information extraction and question answering. In *the HLT-NAACL Workshop on Analysis of Geographic References*, pages 31–38, 2003.
- J. Leveling and S. Hartrumpf. On metonymy recognition for Geographic IR. In *the SIGIR Workshop on Geographic Information Retrieval*, pages 9–13, 2006.
- J. Leveling and D. Veiel. University of Hagen at GeoCLEF 2006: Experiments with metonymy recognition in documents. In *Working Notes of the Cross Language Evaluation Forum Workshop*, 2006.
- J. Leveling, S. Hartrumpf, and D. Veiel. University of Hagen at GeoCLEF 2005: Using semantic networks for interpreting geographical queries. In *Working Notes of the Cross Language Evaluation Forum Workshop*, 2005.
- H. Li, R. Srihari, C. Niu, and W. Li. InfoXtract location normalization: A hybrid approach to geographic references in information extraction. In *the HLT-NAACL Workshop on Analysis of Geographic References*, pages 39–44, 2003.
- Z. Li, C. Wang, X. Xie, X. Wang, and W.-Y. Ma. Indexing implicit locations for geographical information retrieval. In *the SIGIR Workshop on Geographic Information Retrieval*, pages 68–70, 2006.
- E. Lim, D. Goh, Z. Liu, W. Ng, C. Khoo, and S. Higgins. G-Portal: a map-based digital library for distributed geospatial and georeferenced resources. In *the Joint Conference on Digital libraries*, pages 351–358. ACM Press, 2002.
- J. Liu and L. Birnbaum. LocalSavvy: Aggregating local points of view about news issues. In *the WWW'08 Workshop on Location and the Web*, pages 33–40. ACM Press, 2008.
- S. Liu, F. Liu, C. Yu, and W. Meng. An effective approach to document retrieval via utilizing WordNet and recognizing phrases. In *the ACM SIGIR conference on Research and Development in Information Retrieval*, pages 266–272. ACM Press, 2004.

- C. Lüer. Disambiguation: the key to information architecture? Talk presented at Wikimania, August, 2006.
- H. Lukatela. *Discrete Global Grids*, chapter A Seamless Global Terrain Model in the Hipparchus System. University of California, 2000.
- A. MacFarlane. Experiments with TREC data using geographical context. Talk presented at the joint BCS/AGI/City University event, June, 2006.
- T. Mandl, F. Gey, G. Nunzio, N. Ferro, R. Larson, M. Sanderson, D. Santos, C. Womser-Hacker, and X. Xie. GeoCLEF 2007: the cross-language geographic information retrieval track overview. In *Working Notes for the Cross Language Evaluation Forum Workshop*, 2007.
- B. Martins, M. Silva, and L. Andrade. Indexing and ranking in Geo-IR systems. In *the CIKM Workshop on Geographic Information Retrieval*, pages 31–34, 2005.
- B. Martins, N. Cardoso, M. Chaves, L. Andrade, and M. Silva. The University of Lisbon at GeoCLEF 2006. In *Working Notes for the Cross Language Evaluation Forum Workshop*, 2006.
- O. Medelyan, C. Legg, D. Milne, and I. Witten. Mining meaning from Wikipedia. Technical report, University of Waikato, 2008.
- A. Mehler, Y. Bao, X. Li, Y. Wang, and S. Skiena. Spatial analysis of news sources. *IEEE Transactions on visualization and computer graphics*, 12(5):765–772, 2006.
- Q. Mei, C. Liu, H. Su, and C.-X. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW'06*, pages 533–542. ACM Press, 2006.
- Metro. Weird, wild wiki on which anything goes. <http://tinyurl.com/5wgzg4> Published 19 March, 2007.
- R. Mihalcea. Using Wikipedia for automatic word sense disambiguation. In *the Human Language Technologies Conference*, pages 196–203. Association for Computational Linguistics, 2007.
- P. Mika, M. Ciaramita, H. Zaragoza, and J. Atserias. Learning to tag and tagging to learn: A case study on wikipedia. *IEEE Intelligent Systems*, 23(5):26–33, 2008.
- G. Miller, C. Leacock, R. Teng, and R. Bunker. A semantic concordance. In *the Human Language Technology Conference*, pages 303–308. Association for Computational Linguistics, 1993.
- G. Mishne and M. de Rijke. A study of blog search. In *Advances in Information Retrieval*, pages 289–301. Springer-Verlag, 2006.
- D. Montello. The geometry of environmental knowledge. In *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, pages 136–152. Springer-Verlag, 1992.
- W. Myers and C. Dyke. *State of the Internet 2000*. United States Internet Council and ITTA Inc., 2000.
- K. Nakayama, M. Pei, M. Erdmann, M. Ito, M. Shirakawa, T. Hara, and S. Nishio. Wikipedia mining. In *Wikimania*. Wikimedia, 2008.
- M. Naphade and T. Huang. Extracting semantics from audiovisual content: The final frontier in multimedia retrieval. *IEEE Transactions on neural networks*, 13(4):793–810, 2002.

- Nature. Encyclopædia Britannica and Nature: a response. Open letter, March 2006.
- M. Nissim, C. Matheson, and J. Reid. Recognising geographical entities in Scottish historical documents. In *the SIGIR Workshop on Geographic Information Retrieval*, 2004.
- N. O'Hare, C. Gurrin, G. Jones, and A. Smeaton. Combination of content analysis and context features for digital photograph retrieval. In *Integration of Knowledge, Semantic and Digital Media Technology*, pages 323–328, 2005.
- A. Olligschlaeger and A. Hauptmann. Multimodal information systems and GIS: The Informedia digital video library. In *the ESRI User Conference*, 1999.
- S. Osinski and D. Weiss. A concept-driven algorithm for clustering search results. *IEEE Intelligent Systems*, 20(3):48–54, 2005.
- S. Overell and S. Rüger. Geographic co-occurrence as a tool for GIR. In *the CIKM Workshop on Geographic Information Retrieval*, pages 71–76. ACM Press, 2007.
- S. Overell and S. Rüger. Identifying and grounding descriptions of places. In *the SIGIR Workshop on Geographic Information Retrieval*, pages 14–16, 2006.
- S. Overell, J. Magalhães, and S. Rüger. Place disambiguation with co-occurrence models. In *Working Notes of the Cross Language Evaluation Forum Workshop*, 2006.
- S. Overell, J. Magalhães, and S. Rüger. Forostar: A system for GIR. In *Evaluation of Multilingual and Multi-modal Information Retrieval*, pages 930–937. Springer-Verlag, 2007.
- S. Overell, A. Llorente, H.-M. Liu, R. Hu, A. Rae, J. Zhu, D. Song, and S. Rüger. MMIS at ImageCLEF 2008: Experiments combining different evidence sources. In *Working Notes from the Cross Language Evaluation Forum*, 2008a.
- S. Overell, A. Rae, and S. Rüger. MMIS at GeoCLEF 2008: Experiments in GIR. In *Working Notes from the Cross Language Evaluation Forum*, 2008b.
- Princeton University. WordNet, online lexical database. <http://www.cogsci.princeton.edu/~wn/> Accessed 1 December, 2008.
- Q. Pu, D. He, and Q. Li. University of Pittsburgh at GeoCLEF 2008: Towards effective geographic information retrieval. In *Working Notes for the Cross Language Evaluation Forum Workshop*, 2008.
- H. Raghavan, J. Allan, and A. McCallum. An exploration of entity models, collective classification and relation description. In *the KDD Workshop on Link Analysis and Group Detection*, pages 1–10, 2004.
- R. Rapp. The computation of word associations: Comparing syntagmatic and paradigmatic approaches. In *the International Conference on Computational linguistics*, pages 1–7. Association for Computational Linguistics, 2002.
- T. Rattenbury, N. Good, and M. Naaman. Toward automatic extraction of event and place semantics from Flickr tags. In *the ACM SIGIR conference on Research and Development in Information Retrieval*, pages 103–110. ACM Press, 2007.
- E. Rauch, M. Bukatin, and K. Baker. A confidence-based framework for disambiguating geographic terms. In *the HLT-NAACL Workshop on Analysis of Geographic References*, pages 50–54, 2003.

- T. Rees. “C-Squares”, a new spatial indexing system and its applicability to the description of oceanographic datasets. *Oceanography*, 16(1):11–19, 2003.
- S. Robertson and K. Sparck-Jones. Relevance weighting of search terms. *Journal of American Society for Information Systems*, 27:129–146, 1976.
- S. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *the ACM SIGIR conference on Research and Development in Information Retrieval*, pages 345–354. ACM Press, 1994.
- S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *the Text REtrieval Conference*. NIST, 1994.
- S. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. In *the Conference on Information and Knowledge Management*, pages 42–49. ACM Press, 2004.
- A. Rodríguez and M. Egenhofer. Comparing geospatial entity classes: An asymmetric and context-dependent similarity measure. *International Journal of Geographic Information Science*, 3:229–256, 2004.
- M. Ruiz-Casado, E. Alfonseca, and P. Castells. Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. In *Advances in Web Intelligence*, pages 380–386. Springer-Verlag, 2005.
- G. Salton, A. Wong, and C. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.
- M. Sanderson and J. Kohler. Analyzing geographic queries. In *the SIGIR Workshop on Geographic Information Retrieval*, 2004.
- D. Santos, N. Cardoso, P. Carvalho, I. Dornescu, S. Hartrumpf, J. Leveling, and Y. Skalban. Getting geographical answers from Wikipedia: the GikiP pilot at CLEF. In *Working Notes for the Cross Language Evaluation Forum Workshop*, 2008.
- C. Schlieder, T. Vögele, and U. Visser. Qualitative spatial representations for information retrieval by gazetteers. In *Spatial Information Theory*, pages 336–351. Springer-Verlag, 2001.
- P. Schmitz. Inducing an ontology from Flickr tags. In *the WWW’06 Workshop on Collaborative Web Tagging*, 2006.
- N. Seco, D. Santos, N. Cardoso, and R. Vilela. A complex evaluation architecture for HAREM. In *Computational Processing of the Portuguese Language*, pages 260–263. Springer-Verlag, 2006.
- B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *WWW’08*, pages 327–336. ACM Press, 2008.
- M. Silva, B. Martins, M. Chaves, and N. Cardoso. Adding geographical scopes to web resources. In *the SIGIR Workshop on Geographic Information Retrieval*, 2004.
- D. Smith. Detecting and browsing events in unstructured text. In *the ACM SIGIR conference on Research and Development in Information Retrieval*, pages 73–80. ACM Press, 2002.
- D. Smith and G. Crane. Disambiguating geographic names in a historical digital library. In *Research and Advanced Technology for Digital Libraries*, pages 127–136. Springer-Verlag, 2001.

- D. Smith and G. Mann. Bootstrapping toponym classifiers. In *the HLT-NAACL Workshop on Analysis of Geographic References*, pages 45–49, 2003.
- T. Steinberg. Placeopedia.com - Connecting Wikipedia articles with their locations. <http://www.placeopedia.com/> Accessed 1 December, 2008.
- R. Stockli, E. Vermote, N. Saleous, R. Simmon, and D. Herring. The blue marble next generation – a true color Earth dataset including seasonal dynamics from MODIS. Technical report, NASA, 2005.
- M. Strube and S. Ponzetto. WikiRelate! Computing semantic relatedness using Wikipedia. In *the National Conference on Artificial Intelligence*, pages 1419–1424. MIT Press, 2006.
- B. Stvilia, M. Twidale, L. Smith, and L. Gasser. Assessing information quality of a community-based encyclopedia. In *the International Conference on Information Quality*, pages 442–454, 2005.
- F. Suchanek, G. Kasneci, and G. Weikum. YAGO: A core of semantic knowledge unifying WordNet and Wikipedia. In *WWW'07*, pages 697–706. ACM Press, 2007.
- Sun Microsystems. *MySQL 6.0 Reference Manual*, 2008.
- A. Swartz. Raw thought: Who writes Wikipedia? Blog article available at <http://www.aaronsw.com/weblog/whowriteswikipedia> Published 4 September, 2006.
- D. Tapscott and A. Williams. *Wikinomics*. Atlantic Books, 2nd edition, 2008.
- I. Turton. A system for the automatic comparison of machine and human geocoded documents. In *the CIKM Workshop on Geographic Information Retrieval*, pages 23–24. ACM Press, 2008.
- S. Vaid, C. Jones, H. Joho, and M. Sanderson. Spatio-textual indexing for geographical search on the web. In *the International Symposium on Spatial and Temporal Databases*, pages 218–235, 2005.
- M. van Kreveld, I. Reinbacher, A. Arampatzis, and R. van Zwol. Distributed ranking methods for geographic information retrieval. Technical report, Utrecht University, 2004.
- C. van Rijsbergen. *Information Retrieval*. Butterworths, 2nd edition, 1979.
- E. Voorhees. The TREC robust retrieval track. *SIGIR Forum*, 39(1):11–20, 2005.
- E. Voorhees and D. Harman. Overview of the 8th Text REtrieval Conference (TREC-8). In *the Text REtrieval Conference*, pages 1–33. NIST, 1999.
- N. Wacholder, Y. Ravin, and M. Choi. Disambiguation of proper names in text. In *Applied Natural Language Processing*, pages 202–208. Association for Computational Linguistics, 1997.
- J. Wales. Wikipedia is an encyclopedia. E-mail available at <http://tinyurl.com/622poc> Sent 8 March, 2005.
- N. Waters. Why you can't cite Wikipedia in my class. *Communications of the ACM*, 50(9):15–17, 2007.
- C. Wayne. Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. In *Language Resources and Evaluation Conference*, pages 1487–1494, 2000.
- G. Weaver, B. Strickland, and G. Crane. Quantifying the accuracy of relational statements in Wikipedia: a methodology. In *the Joint Conference on Digital libraries*, page 358. ACM Press, 2006.

- Wikimedia. Statistics. <http://meta.wikimedia.org/wiki/Statistics> Accessed 1 December, 2008.
- Wikipedia. WikiProject geographical coordinates. http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Geographical_coordinates Accessed 1 December, 2008a.
- Wikipedia. Wikipedia, the free encyclopedia. <http://www.wikipedia.org> Accessed 1 December, 2008b.
- P. Wilkins, P. Ferguson, and A. Smeaton. Using score distributions for query time fusion in multimedia retrieval. In *the SIGMM International Workshop on Multimedia Information Retrieval*, pages 51–60. ACM Press, 2006.
- A. Woodruff. Gipsy: Georeferenced information processing system. Technical report, University of California at Berkeley, 1994.
- M. Worboys. Metrics and topologies for geographic space. In *Advances in GIS Research II*, pages 365–376. Taylor and Francis, 1996.
- D. Yarowsky. One sense per collocation. In *ARPA Human Language and Technology Workshop*, pages 266–271, 1993.
- D. Yarowsky. Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. In *the Annual Meeting of the Association for Computational Linguistics*, pages 88–95. Association for Computational Linguistics, 1994.
- J. Young. Wikipedia founder discourages academic use of his creation. *The Chronical of Higher Education: The Wired Campus*, June 2006. Available at <http://tinyurl.com/lxhxo>.
- W. Zhang, S. Liu, C. Yu, C. Sun, and W. Meng. Recognition and classification of noun phrases in queries for effective retrieval. In *the Conference on Information and Knowledge Management*, pages 711–720. ACM Press, 2007.
- S. Zhou and C. Jones. A multi-representation spatial data model. In *Advances in Spatial and Temporal Databases*, pages 394–411. Springer-Verlag, 2003.
- J. Zhu, V. Uren, and E. Motta. ESpotter: Adaptive named entity recognition for web browsing. In *the Professional Knowledge Management Conference*, pages 518–529, 2005.
- W. Zong, D. Wu, A. Sun, E. Lim, and D. Goh. On assigning place names to geography related web pages. In *the Joint Conference on Digital Libraries*, pages 354–362. ACM Press, 2005.