

# Geographic Co-occurrence as a tool for GIR.

Simon E Overell  
Multimedia & Information Systems  
Dept of Computing, Imperial College London  
London SW7 2AZ, UK  
simon.overell01@imperial.ac.uk

Stefan Rüger  
Multimedia & Information Systems  
Knowledge Media Institute, The Open University  
Milton Keynes MK7 6AA, UK  
s.rueger@open.ac.uk

## ABSTRACT

In this paper we describe the development of a geographic co-occurrence model and how it can be applied to geographic information retrieval. The model consists of mining co-occurrences of placenames from Wikipedia, and then mapping these placenames to locations in the Getty Thesaurus of Geographical Names. We begin by quantifying the accuracy of our model and compute theoretical bounds for the accuracy achievable when applied to placename disambiguation in free text. We conclude with a discussion of the improvement such a model could provide for placename disambiguation and geographic relevance ranking over traditional methods.

## Categories and Subject Descriptors

H.3.1 [Information storage and retrieval]: Content Analysis and Indexing

## General Terms

Algorithms

## 1. INTRODUCTION

Newspapers, television, books and the Internet hold a huge amount of geographic information. A minute proportion of this data is accompanied with machine readable meta-data. It is common to want to browse by placename, when searching for information about a specific event or location [23]; because of this, there is a need for automatic annotation of resources with location data.

Generally in Geographic Information Retrieval (GIR) placenames are extracted from free text and disambiguated to provide geographic meta-data. A geographic relevance ranking algorithm will then approximate how *similar* a user would judge this geographic meta-data to a geographic information need. The most common solutions to the problems of placename disambiguation and geographic relevance ranking use a combination of gazetteers and simple heuristics. Although gazetteers hold a lot of information about locations

such as most common and alternative placenames, vertical topology, population density and coordinates, gazetteers do not tell us how placenames are used in context.

In this paper we propose a geographic co-occurrence model that captures the context placenames are used in and an extensive set of synonyms for each location. We believe such a co-occurrence model is necessary to capture implicit contextual information about the relationship between locations.

In Section 2 we describe the areas of placename disambiguation and geographic relevance ranking. In Section 3 we outline related work describing geographic gazetteers, mining Wikipedia and named entity models. Section 4 describes how we mine Wikipedia for geographic information and map articles to locations in the TGN gazetteer. The geographic information and mappings then form our geographic co-occurrence model. In Section 5 we evaluate our model and attempt to define bounds for the accuracy achievable when applying it to free text. We conclude with Section 6, a discussion on how our model can be applied to GIR.

## 2. BACKGROUND

The motivation behind developing our geographic co-occurrence model is to improve performance in the areas of placename disambiguation and geographic relevance ranking. In this section we describe these areas in detail.

### 2.1 Placename disambiguation

A *geolocation*, or short *location*, is a set of coordinates representing a point, line, polygon or set of polygons on the Earth's surface. A *placename* is a phrase used to refer to a location. Wacholder et al. [28] identified multiple levels of placename ambiguity: the first level of ambiguity is structural ambiguity, where the structure of the words constituting the placename 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?). Approaches from multiple fields have led to varied methods of solving this problem, most of which fit into the categories described below:

**Rule-based methods.** The rule-based disambiguation methods apply simple heuristic rules to placename disambiguation. The most basic disambiguation rules use a specially constructed default gazetteer. A default gazetteer contains only a single location that can be mapped to each recognised

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

GIR'07, November 9, 2007, Lisboa, Portugal.

Copyright 2007 ACM 978-1-59593-828-2/07/0011 ...\$5.00.

placename; these default locations are selected on various criteria including size, population and importance [5, 16].

More complex methods of disambiguation define a geographic scope for a document; this assumes locations close geographically will generally occur close together within a document [2, 22, 32]. 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 [2, 5, 12, 22]. Once a placename has been disambiguated as a location, it can be assumed to refer to the same location until it occurs in a new context [11, 31].

**Data driven and Hybrid (bootstrapping) methods.** The data driven methods of disambiguation generally apply standard machine learning methods to solve the problem of matching placenames to locations. Hybrid methods of disambiguation apply semi-supervised techniques similar to data driven methods or include a limited set of heuristic rules. A smaller annotated corpus is required than for data driven methods and an additional un-annotated corpus is used to infer further characteristics of the data [5, 15, 25]. As contextual information needed to disambiguate placenames is learned from the data there is no need to manually construct complex sets of heuristic rules.

The problem with these methods is a large accurate corpus of annotated ground truth is required. Small sets of ground truth have been created for the purposes of evaluation or applying supervised learning methods to small domains [5, 15]; however, a large enough corpus does not yet exist in the public domain to apply supervised methods to free text. The co-occurrence model proposed in this paper could be considered an annotated corpus of suitable size.

## 2.2 Geographic relevance ranking

Geographic relevance ranking is the task of assigning a relevance score between documents' geographic meta-data (automatically or manually generated) and the geographic elements of a query with respect to a user's information need. Egenhofer et al. [8] define the concept of *Naive Geography*, which captures and reflects the way people think and reason about geographic space and time. Geographic relevance ranking is generally calculated using heuristic rules based on the elements of Naive Geography.

Egenhofer et al. [9] and Martins et al. [17] apply the "Topology matters, metric refines" premise [8]. Egenhofer et al. capture the topological relationship between locations based on the intersection measures of *splitting* and *closeness*. 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. They calculate the geographic similarity as the weighted sum of these values.

## 3. RELATED WORK

### 3.1 Geographic Gazetteers

Hill [14] identifies gazetteers as a "geospatial dictionaries of geographic names," essentially lists of placenames mapped to geographic coordinates and categories. Schlieder et al. [24] recognise gazetteers as a subset 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. The Getty Thesaurus of Geographical Names (TGN) assigns a

unique identifier to every location and contains approximately 800,000 locations [13]. The unique identifiers, coverage and language independence make annotations with the TGN portable. The most extensive publicly available lists of geographic names (although less accurate than Getty) is the GNIS Gazetteer covering the United States and GNS Gazetteer covering the rest of the world [18, 19]. The GNIS gazetteer currently contains nearly 2 million locations. The GNS contains over 6 million locations.

### 3.2 Mining Wikipedia

Wikipedia is the largest reference website on the Internet. The content is collaboratively written and updated by volunteers [30]; it is extremely useful as a resource due to its size, variation, accuracy and quantity of hyper-links and meta-data [29]. To date there are over two million articles and stubs (short articles) [30].

Weaver et al.'s [29] study checked the accuracy of Wikipedia's relational statements and links. They found these statements and links to be accurate over 97% of the time. Gabrilovich et al. [10] and Strube et al. [27] attempted to mine such semantic relations from Wikipedia. Strube et al. used Wikipedia's category tree as a navigatable folksonomy; the relatedness of pages were considered as the distance between nodes in the tree. Cucerzan [7] and Bunescu et al. [3] extracted information from Wikipedia, which was used to form language models. These language models were then used to disambiguate named entities. Such language models have the advantage of capturing how the same entity can be referred to by different names in different contexts. Overell et al. [20] use a similar model where only the co-occurrence of placenames is recorded and a gazetteer is used as an authoritative source. The model proposed in this paper differs from Overell et al.'s in several ways: we crawl Wikipedia for all synonyms referring to an article before attempting to disambiguate which location an article describes; we do not use the article text for disambiguation; and we enrich our model with further user-generated content from Placeopedia [26] (algorithm described in detail in Section 4).

### 3.3 Named Entity Models

Raghaven et al. [21] 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 summarising systems. Raghaven et al. quantify the information content of their models by measuring the clarity, defined as the Kullback-Leibler divergence between entity model distributions and the global distribution. The clarity for the model built for entity  $e$  is defined thus:

$$\text{Clarity}_e = \sum_{w \in V} P(w|E_e) \log \frac{P(w|E_e)}{P(w|C)}, \quad (1)$$

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  $w$  in the corpus  $C$ .

Agichtein et al. [1] 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' background noise.

## 4. THE CO-OCCURRENCE MODEL

The co-occurrence model is built in a 4 stage process. In Stage 1 we crawl Wikipedia extracting articles' 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. We can also build a list of synonyms that are used to refer to each article. For example the article 'London' describing the capital of the UK can be referred to by the following placenames: 'Londinium,' 'London,' 'London, England,' 'London, UK,' 'the City' and 28 further names too numerous to list here.

In Stage 2 we build a set of inferences for each article in an attempt to map them to locations in the TGN. An inference is a mapping between a Wikipedia article and a TGN location with supporting evidence. We first build a set of possible locations from the TGN, with placenames matching any synonym of the article. We then search for evidence that will allow us to infer if this article refers to a specific location. There are two types of evidence we search for: Firstly if this article is tagged with coordinates close to one of the possible locations. Latitude and longitude tags are searched for in both the Wikipedia templates and the user-generated content site Placeopedia. Secondly, further evidence is provided if this article mentions a referent location in either the article title or category links. A referent location is a location listed above this location in the TGN's vertical topology tree. For example the article 'Cambridge' is tagged with geographic coordinates (N52.2°, E0.1°) in Wikipedia, coordinates (N52.2°, E0.1°) in Placeopedia, and categories 'Cambridgeshire' and 'England'. It also has as a synonym 'Cambridge, England'. In total 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 we resolve ambiguity caused by inferences to multiple locations for a single article. The inferences for each article are looked at and evaluated in a pipeline. The pipeline checks inference evidence in the following order: coordinates found in the Wikipedia template, coordinates from Placeopedia, matching referent location in the article title, and matching referent location in the article categories. Continuing our example the article 'Cambridge' will be disambiguated as 'Cambridge, Cambridgeshire, UK'.

In Stage 4 all articles in the co-occurrence model not referring to locations are discarded. This reduces the size of the model by two thirds making it easier to manipulate and apply (see Section 4.1) and reduces the background noise (see Section 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 [4, 20] compared to Wikipedia meta-data [20]. This is due to the content of Wikipedia articles being very heterogeneous, while in comparison the meta-data is very structured. We believe Wikipedia is now mature enough that the meta-data alone contains enough information for disambiguation.

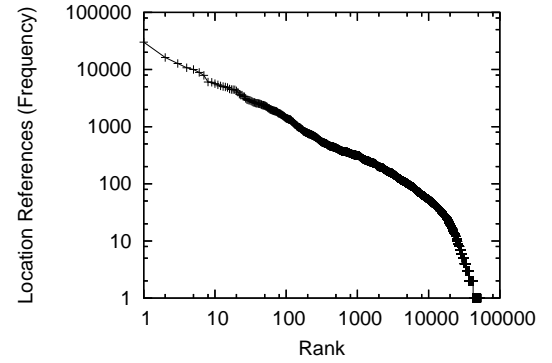


Figure 1: Frequency of Location references – rank

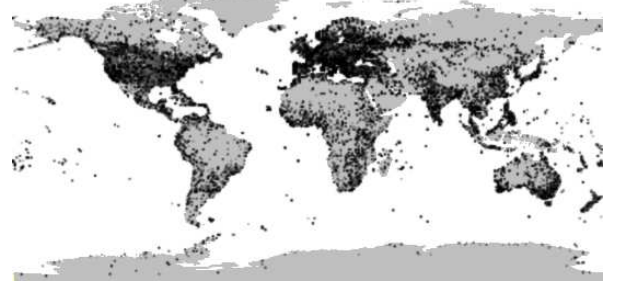


Figure 2: Distribution of locations that are referred to in Wikipedia

### 4.1 Model size and complexity

Currently we have crawled 100,000 randomly selected full length articles. Over 7.5 million links were extracted, 2.3 million of those links to articles describing locations. 329,377 inferences were mined allowing 79,769 articles describing locations to be disambiguated. A total of 75,322 placenames were extracted mapping to 53,643 locations. References to locations in Wikipedia follow a Zipfian distribution: 2,500 locations (~5%) account for 1.1 million links (~50%) and 500 locations (~1%) account for 505,730 links (~25%). This distribution is illustrated in Figure 1 on a log scale.

Figure 2 illustrates this distribution: the darker a pixel the more references to locations falling within that area<sup>1</sup>.

### 4.2 Accuracy

We attempt to quantify the accuracy of our model by comparing it to the ground truth defined in [20] of 1,395 locations and 7,660 non locations extracted from 1000 randomly selected Wikipedia articles. In Table 1 we compare our results found using Wikipedia and Placeopedia meta-data to the naïve methods: assigning a random matching location, and disambiguation using referent locations extracted from the article text (described in detail in [20]). We record 4 values from our experiments: **Placename Recall** (Pn R), the proportion of placenames correctly recognised; **Placename Accuracy** (Pn A), the ratio of locations recognised and non-placenames recognised to the total number of entities; **Placename + Mapping Recall** (Pn+M R), the proportion of recognised placenames matched to the correct location in the TGN; and **Placename + Mapping Accu-**

<sup>1</sup>It is interesting to note some characteristics of this distribution: we crawled the English Language Wikipedia, hence there is a strong European – North American bias. Discounting this, the distribution loosely follows population centers.

**Table 1: Co-occurrence Model Accuracy**

	Pn R	Pn A	Pn+M R	Pn+M A
Random	86.6%	91.7%	53.6%	85.5%
Referents	61.1%	93.0%	93.1%	92.3%
Meta-data	75.7%	95.1%	90.3%	94.1%

**racy** (Pn+M A), the proportion of correctly matched locations and non-placenames recognised to the total number of entities.

## 5. COMPUTING THE BOUNDS

To quantify how useful our tool is for placename disambiguation in free text we have attempted to establish an upper and lower bound of accuracy that can be achieved.

Let  $P(l)$  be the set of placenames that refers to location  $l$ . Similarly, let  $L(p)$  be the set of locations that is referred to by placename  $p$  and let  $|L(p)|$  be its size. Let  $\text{ref}(p, l)$  be the number of references made to location  $l$  by placename  $p$  within a model. Let

$$L_1(p), L_2(p), \dots, L_{|L(p)|}(p) \quad (2)$$

be an enumeration of  $L(p)$  such that

$$\text{ref}(p, L_1(p)) \geq \text{ref}(p, L_2(p)) \geq \dots \geq \text{ref}(p, L_{|L(p)|}(p)). \quad (3)$$

Note that  $L_1(p)$  will be the location most commonly referred to by the placename  $p$ , and that  $p$  is unambiguous when  $|L(p)| = 1$ . Let  $N$  be the multiset of all placenames that appear in the model and  $M$  the set of unique placenames; naturally  $|N| \geq |M|$ . 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 follows:

$$r_{\text{corrLower}} = \frac{\sum_{p \in M} \text{ref}(p, L_1(p))}{|N|}. \quad (4)$$

In our model we calculate  $r_{\text{corrLower}}$  equal to 87.1%. We can take this as 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. We also make the three, later revised assumptions 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 5).

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, where the size of a context is the number of locations co-occurring with  $l$ , then

$$r_{\text{corrUpper}} = 1 - \frac{\sum_{p \in M} \sum_{i=2}^{|L(p)|} \text{ref}(p, L_i(p), 0)}{|N|}. \quad (5)$$

In our model we calculate  $r_{\text{corrUpper}}$  equal to 99.7%. We can take this as the upper bound that could be achieved by a *perfect* classifier. We realise this value is particularly high and attribute this to two causes: Firstly, the placenames

**Table 2: KL-divergence between location and placename distributions**

	KL loc.	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

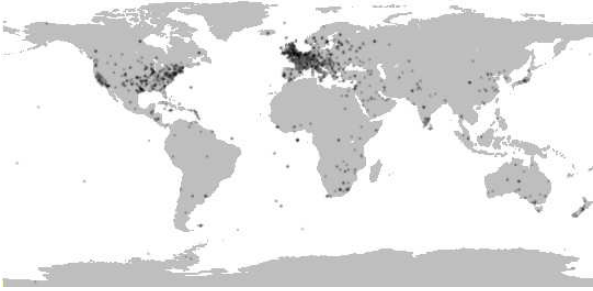
incorrectly classified by the lower bound are ambiguous placenames referring to locations where a more commonly referred to location exists (e.g. ‘London, Ontario’). In general these relatively uncommon placenames will be framed in context. Secondly, the upper and lower bounds quoted above are accurate with respect to the model. As shown in Section 4.2 the placenames in the model are only correctly mapped to locations 90.3% of the time, therefore a more realistic upper and lower bound for achievable accuracy would be **78.7%** and **90.0%** (the product of the bounds and placename+mapping recall) recognising  $\sim$  **75%** of placenames (placename recall). Any classifier built with our model would not be able to correctly classify ambiguous placenames referring to uncommon locations with no contextual information (0.3%).

In the following paragraphs we estimate how well we would expect classification methods to perform when disambiguating different placenames with respect to each other. As mentioned in Section 3.3, Agichtein et al. suggest one can quantify the difficulty of an information extraction task by measuring the Kullback Leibler divergence between the local and global distributions (Equation 1).

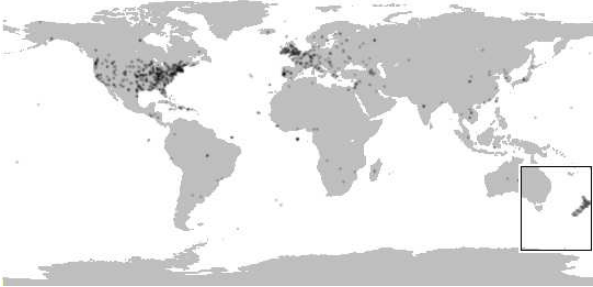
To illustrate this in Table 2 we calculate the KL-divergence between the local ‘Cambridge, MA, USA,’ ‘Cambridge, UK,’ and ‘Cambridge, New Zealand’ distributions and the global ‘Cambridge’ distribution; and the ‘London, UK,’ ‘London, ON, Canada’ and ‘London, CA, USA’ distributions and the global ‘London’ distribution. We compute the KL-divergence in both the co-occurrence model made up of only locations and the model made up of all proper names (skipping Stage 4 in building the co-occurrence model). 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 KL-divergence 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. We can see this trend more easily by computing the average KL-divergence between the local distribution and the 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**. We attribute this difference to contexts in the locations only model being more distinctive. It is because of this trend, and the smaller model being easier to manipulate, that we discard all proper names apart from placenames.

Returning to Table 2, a low KL-divergence 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’.



**Figure 3: Distribution of locations that co-occur with Cambridge, UK**



**Figure 4: Distribution of locations that co-occur with Cambridge, MA and Cambridge, NZ (inset)**

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, New Zealand’ 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, Ontario’.

We illustrate how the location distributions for the placename ‘Cambridge’ differ with similar diagrams to Figure 2. Figures 3 and 4 illustrate how locations co-occur with ‘Cambridge, UK’ and ‘Cambridge, MA’ respectively. Although at first glance the distributions look quite similar one will notice a significant bias toward locations in the United States in the ‘Cambridge, MA’ distribution with an opposite bias to European locations in the ‘Cambridge, UK’ distribution.

The ‘Cambridge, New Zealand’ distribution is inset in Figure 4. Note there are negligible occurrences with places outside of New Zealand, and few outside the North Island.

Mutual Information (MI) is the relative entropy between two distributions [6], it is a measure of the shared information between distributions. We define Mutual Information between locations  $X$  and  $Y$  as

$$MI(X; Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}, \quad (6)$$

where  $x \in \mathcal{X}$  is a location occurring in the distribution of  $X$ ,  $y \in \mathcal{Y}$  is a location occurring in the distribution of  $Y$ ,  $p(x, y)$  is the joint probability of  $x$  and  $y$  occurring in the model, and  $p(x)$  and  $p(y)$  are the individual probabilities of  $x$  and  $y$  occurring in the corpus. The MI between two locations can be considered a similarity measure. The more similar two locations the more shared clues with respect to context and the harder they are to disambiguate.

Table 3 contains the MI between the 6 locations already discussed in this paper. Note the upper right corner compares locations with different placenames while the lower

**Table 3: MI between location distributions**

	Lo.ON	Lo.CA	Ca.MA	Ca.NZ	Ca.UK
Lo.UK	60.2	5.6	143.0	7.9	459.1
Lo.ON		17.4	14.3	55.5	89.4
Lo.CA			1.1	14.3	27.1
Ca.MA				12.7	182.6
Ca.NZ					22.0

and left corners compare locations with the same placename. It is the MI between locations sharing a placename that is of interest to placename disambiguation. Note the MI between ‘Cambridge, UK’ and ‘Cambridge, MA’ is particularly high meaning these locations occur in similar contexts and would be easy to confuse, while the ‘London, UK,’ ‘London, ON’ and ‘London, CA’ distributions are relatively different making ‘London’ easier to disambiguate.

The MI of locations with different placenames (Table 3, upper right) are of interest to geographic relevance ranking. ‘London, UK’ and ‘Cambridge, UK’ have by far the highest MI; they occur in very similar contexts and therefore can be considered the most related. ‘Cambridge, MA’ and ‘London, CA’ have a particularly low MI and therefore occur in very different contexts so can be considered generally unrelated.

## 6. DISCUSSION

In the previous sections we have described a co-occurrence model and quantified its accuracy and expected accuracy when applied to free text. In our final section we provide a discussion of how such a co-occurrence model can be applied as a tool in different areas of GIR and the impact this could have.

### 6.1 Application to placename disambiguation

In Section 5 we showed that a location classifier using our co-occurrence model should recognise  $\sim 75\%$  of placenames and map them to locations with an accuracy of between 78.7% and 90.0%. Where between these bounds a classifier falls is largely dependant on the Mutual Information between locations. The significant advantage of the co-occurrence model over traditional gazetteers is that small, rarely referred to locations such as ‘Cambridge, New Zealand’ or ‘London, CA’ can be included with negligible negative effects on the disambiguation of larger locations. The model also contains significantly more synonyms for locations than a traditional gazetteer. As well as the synonyms the model makes use of the context in which they are used and the frequency of their use. Also as the model contains the probability of each location for every placename (illustrated in Table 2), it is now trivial to build a simple gazetteer of default locations for naïve placename disambiguation.

In summary we believe a co-occurrence model of the nature described in this article to be an essential tool to placename disambiguation. In our future experiments we will compare a supervised learning classifier trained with our co-occurrence model to a default gazetteer and simple heuristic methods.

### 6.2 Application to geographic relevance

We believe geographic relevance ranking is by its nature harder to evaluate than placename disambiguation. When a document is created it can be assumed the author knows which location is intended for each placename. Because of this, it is possible to build a ground truth. Geographic rel-

evance is not only subjective but also transient. A document that was not relevant to a specific user with respect to an information need, may become relevant as their state of knowledge changes (and vice versa). Empirical evaluation such as the CLEF and TREC evaluation forums have become the de facto method for testing relevance methods in Information Retrieval; recently GeoCLEF has become the adopted standard for testing many GIR systems.

We consider the distributions of co-occurring locations contained in our co-occurrence model and illustrated for ‘Cambridge, UK’, ‘Cambridge, MA’ and ‘Cambridge, New Zealand’ in Figures 3 and 4, to be fuzzy sets of relevant collocations. The degree another location is a member of this set measures how likely these locations are to co-occur, while the overlap between such sets is how likely co-occurring locations are to co-occur (this could be considered as latent or mutual information and is illustrated in Table 3).

In our future experiments we will test the hypothesis empirically, that the distributions in our co-occurrence model can provide improvements over heuristic rules when quantifying geographic relevance.

## 7. REFERENCES

- [1] E. Agichtein and S. Cucerzan. Predicting accuracy of extracting information from unstructured text collections. In *Proceedings of CIKM*, pages 567–568, 2005.
- [2] E. Amitay, N. Har’El, R. Silvan, and A. Soffer. Web-a-where: Geotagging web content. In *Proceedings of SIGIR*, pages 273–280, 2004.
- [3] R. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of EACL*, pages 9–16, 2006.
- [4] D. Buscaldi, P. Rosso, and P. Garcia. Inferring geographic ontologies from multiple resources for geographic information retrieval. In *SIGIR Workshop on GIR*, pages 52–55, 2006.
- [5] P. Clough, M. Sanderson, and H. Joho. Extraction of semantic annotations from textual web pages. Technical report, University of Sheffield, 2004.
- [6] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1st edition, 1991.
- [7] S. Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *Proceedings of EMNLP-CoNLL*, 2007.
- [8] M. Egenhofer and D. Mark. Naive geography. In *Proceedings of COSIT*, 1995.
- [9] M. Egenhofer and A. Shariff. Metric details for natural-language spatial relations. *Journal of the ACM TOIS*, 4:295–321, 1998.
- [10] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of IJCAI*, pages 1606–1611, 2007.
- [11] W. Gale, K. Church, and D. Yarowsky. One sense per discourse. In *DARPA Speech and Natural Language Workshop*, pages 233–237, 1992.
- [12] E. Garbin and I. Mani. Disambiguating toponyms in news. In *Proceedings of HLT/EMNLP*, pages 363–370, 2005.
- [13] P. Harping. *User’s Guide to the TGN Data Releases*. The Getty Vocabulary Program, 2.0 edition, 2000.
- [14] L. Hill. Core elements of digital gazetteers: Placenames, categories, and footprints. In *Proceedings of ECDL*, pages 280–290, 2000.
- [15] J. Leveling, S. Hartrumpf, and D. Veiel. University of Hagen at GeoCLEF 2005: Using semantic networks for interpreting geographical queries. In *Working Notes for the CLEF Workshop*, 2005.
- [16] H. Li, R. Srihari, C. Niu, and W. Li. InfoXtract location normalization: A hybrid approach to geographic references in information extraction. In *HLT-NAACL Workshop on Analysis of Geographic References*, pages 39–44, 2003.
- [17] B. Martins, N. Cardoso, M. Chaves, L. Andrade, and M. Silva. The University of Lisbon at GeoCLEF 2006. In *Working Notes for the CLEF Workshop*, 2006.
- [18] National Geospatial-Intelligence Agency. <http://earth-info.nga.mil/gns/html/>. Accessed 15 June 2007.
- [19] National Geospatial-Intelligence Agency. <http://www.nga.mil/>. Accessed 15 June 2007.
- [20] S. Overell and S. Rüger. Identifying and grounding descriptions of places. In *SIGIR Workshop on GIR*, pages 14–16, 2006.
- [21] H. Raghavan, J. Allan, and A. McCallum. An exploration of entity models, collective classification and relation description. In *KDD Workshop on Link Analysis and Group Detection*, pages 1–10, 2004.
- [22] E. Rauch, M. Bukatin, and K. Baker. A confidence-based framework for disambiguating geographic terms. In *HLT-NAACL Workshop on Analysis of Geographic References*, pages 50–54, 2003.
- [23] M. Sanderson and J. Kohler. Analyzing geographic queries. In *SIGIR Workshop on GIR*, 2004.
- [24] C. Schlieder, T. Vögele, and U. Visser. Qualitative spatial representations for information retrieval by gazetteers. In *Proceedings of COSIT*, pages 336–351, 2001.
- [25] D. Smith and G. Mann. Bootstrapping toponym classifiers. In *HLT-NAACL Workshop on Analysis of Geographic References*, pages 45–49, 2003.
- [26] T. Steinberg. <http://www.placeopedia.com/>. Accessed 15 June 2007.
- [27] M. Strube and S. P. Ponzetto. WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of AAAI-06*, pages 1419–1424, 2006.
- [28] N. Wacholder, Y. Ravin, and M. Choi. Disambiguation of proper names in text. In *Proceedings of ANLP*, pages 202–208, 1997.
- [29] G. Weaver, B. Strickland, and G. Crane. Quantifying the accuracy of relational statements in Wikipedia: a methodology. In *Proceedings of JCDL*, pages 358–358, 2006.
- [30] Wikipedia. <http://www.wikipedia.org>. Accessed 15 June 2007.
- [31] D. Yarowsky. One sense per collocation. In *ARPA Human Language and Technology Workshop*, pages 266–271, 1993.
- [32] W. Zong, D. Wu, A. Sun, E. Lim, and D. Goh. On assigning place names to geography related web pages. In *Proceedings of JCDL*, pages 354–362, 2005.