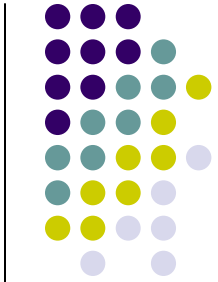
Automatic style classification and transformation

Foaad Khosmood (foaad@soe.ucsc.edu)
Robert Levinson (levinson@soe.ucsc.edu)
Department of Computer Science, School of Engineering
University of California at Santa Cruz

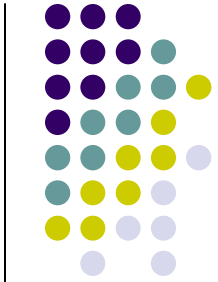
Corpus Profiling Workshop, London
Oct. 2008

Background



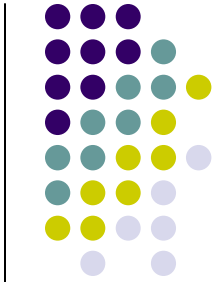
- Work on authorship / source attribution (’05,’06)
  - statistical distribution over selected features
  - machine learning in classifiers
  - “profile” corpus, compare against sample
  - essentially recognize style at best (Bible versions, Shakespeare comedy/tragedies)
- Hypothesis: style can be recognized
- And, if correctly recognized, can be altered
- Should be able to automate using an engine employing heterogeneous methods

Applications



- Searching by style (define by example)
- Improve NLG, or enhance output, using style dimension
- Machine Translation
- HCI, customizable interfaces
- Interactive entertainment
- authorship tools
- plagiarism, copy-right, etc.

But what is “Style” ?

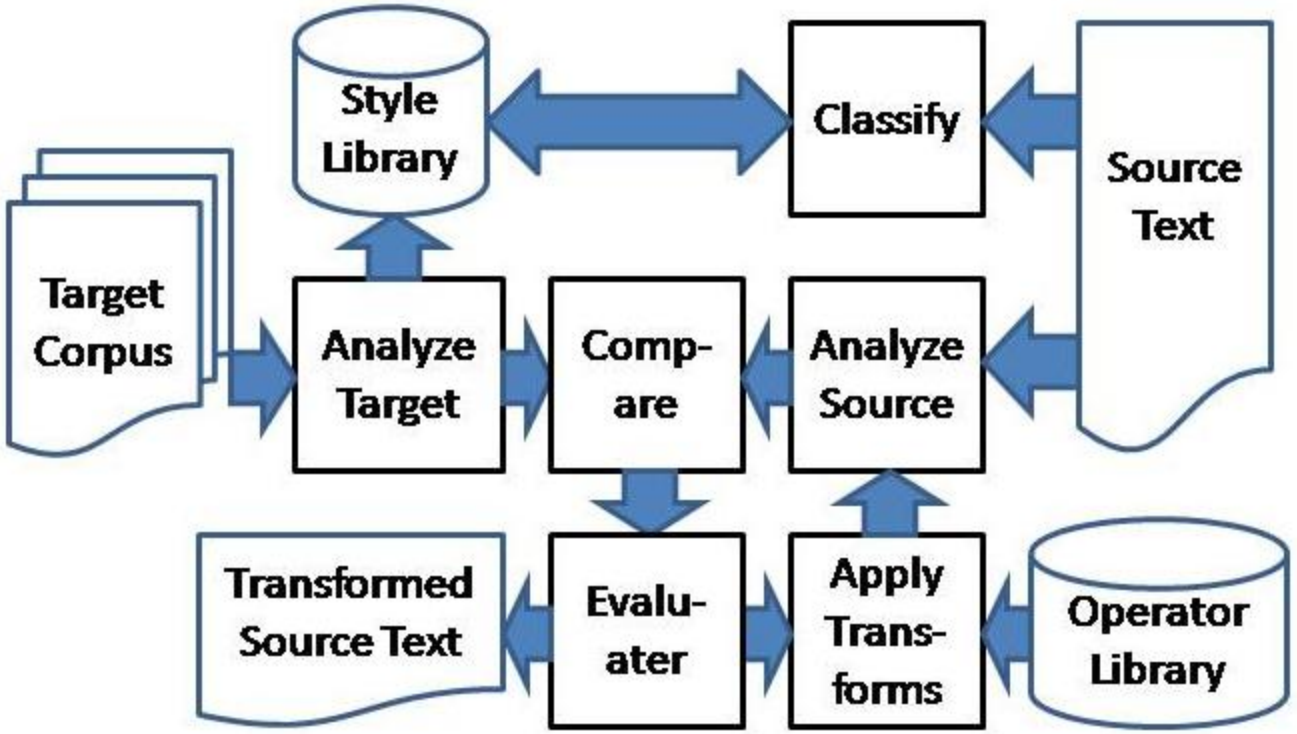


- Difficult to define
- Linguists have basically ignored it in favor of “register”. Some sociolinguistics is relevant.
- Literary studies employs “stylistics” without really defining “style”. Stylistics has had mixed reception.
- Critics: interpretation of style “classification” creates meaning, which many believe to be unjustified.
  - Deriving statistics is tautological unless something is concluded.
  - But nothing can be concluded without creating meaning.

Our philosophy

- Style processing for mimicking and classification purposes only, not for making any literary conclusions.
- One-to-one correspondence of style with a profiled corpus.
- Treat style as any conscious “choice” [Walpole] by originator, where multiple meaning-preserving alternative forms of linguistic expression exists.
- Evolving Detection ←→ Transformation loop leading to ever more sophisticated profiles.

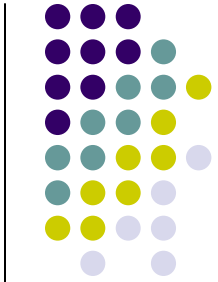
Style processing engine



- Three important pieces:

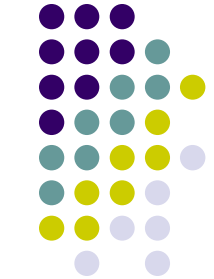
1. Analyzer: gather and process style marker data
2. Transformer: apply a transformation
3. Comparator/Evaluator/Classifier: find metrics, evaluate distance, cluster

Analyzer



- Phase I. Start with a basic set of length-agnostic measurable style markers (ideas from survey)
- Phase II. Plug in more markers as called for by the detection ←→ classification loop
- Phase III. Offer weightings and optimization
- Phase IV. Methods to automatically extract more makers

Transformer



- Two kinds currently envisioned.
  - Lexeme substitution (surface syntax only)
  - Paraphrase / Rewrite (parsing and reassembling)
- Phase I. Apply basic pre-defined transformations
- Phase II. Planning engine for selection and sequencing of transformations
- Phase III. Plug in more transformations
- Phase IV. Explore automatic inferred transformations

Comparator / Evaluator / Classifier

- Use decreasing fuzzy tolerance levels
- What distance value is “close enough” to be considered congruent?

Transformation Exercise (US DOJ legal notice on LEP -> “Animal Farm” by George Orwell)

Second, agencies that have not already published recipient guidance should consider these factors and clarifications in preparing guidance documents.

They should then submit their guidance documents to **[DOJ -> United States Department of Justice]** for approval prior to publication, as is required by the Executive Order. Following approval by the

Department of Justice and before ending its guidance, each agency should obtain public comment on its proposed guidance documents. Those agencies also need to make the determinations regarding the Administrative Procedure Act and Executive Order 12866 as explained above.

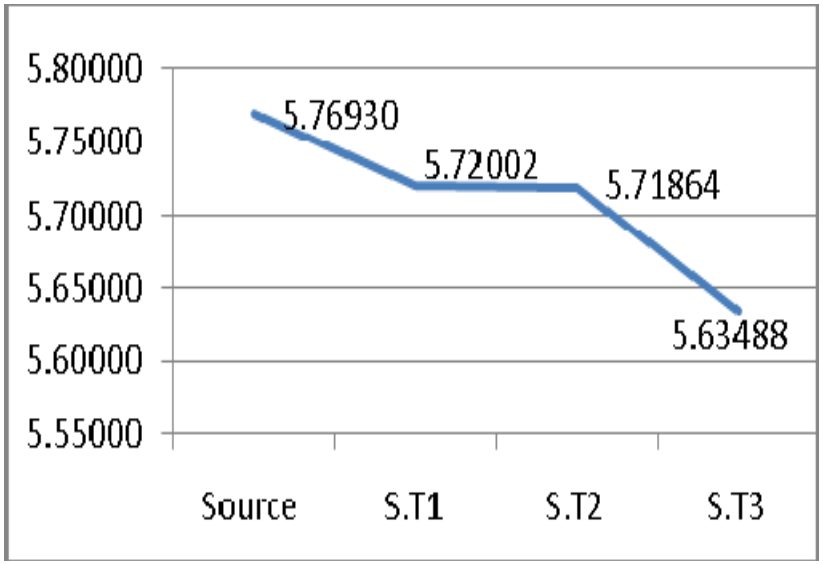
Third, as required by the Executive Order, agencies should continue to design and implement plans for making their own federally conducted programs and activities **[meaningfully -> meaningful]** accessible to **[LEP -> limited English proficiency]** persons, and should consider the **[four-factor -> four factor]** analysis from the **[DOJ -> United States Department of Justice]** guidance and today's memoranda in doing so.

Federal financial assistance includes, but is not limited to, grants and loans of federal funds; grants or donations of federal property; training; details of federal personnel; or any agreement, arrangement, or other contract which has as one of its purposes the provision of assistance. If an agency does not engage in any of those activities, it does not grant federal financial assistance and does not have to issue a recipient guidance document. However, it must still design and implement a federally conducted plan to assure access for **[LEP -> limited English proficiency]** to all of its federally conducted programs and activities (basically, everything that it does).

Phase I example

Marker	Target	Source	S.T1	S.T2	S.T3
avg S per P	5.85700	2.66667	2.66667	2.66667	2.66667
avg W per P	103.43700	77.00000	77.33330	77.33330	81.33330
avg C per P	455.30300	428.33300	428.00000	427.00000	460.33330
avg W per S	17.66000	28.87500	29.00000	29.00000	30.50000
avg C per W	4.40200	5.56277	5.55345	5.52155	5.65980
avg Syl per W	1.33000	1.77000	1.77000	1.75000	1.77000
avg frequency of words	0.00042	0.00794	0.00787	0.00787	0.00775
ratio of W > 6 /W	0.02144	0.02165	0.02155	0.02586	0.02459
ratio of W > 2 Syl	0.09530	0.30303	0.29741	0.29310	0.31147
linux dictionary hits	0.97963	0.99206	1.00000	1.00000	1.00000
RMS Error versus Target	n/a	5.76930	5.72002	5.71864	5.63488

Conclusions



- We can definitely detect a stylistic shift
- We don't have enough makers to truly capture the style the way that is "intuitive" to humans
- We don't have enough transformations
- We don't have enough good example of existing style transformations
  - Will be using Bible, other ancient texts
- Building a system that can handle a large number of markers and transformations and plan the appropriate transformation to minimize distance