

Dealing with Non-Standard Orthography

Sabine Hunsicker

Feb 17th, 2009

- ① Information Retrieval
- ② Problems with Historic Texts
- ③ Approaches in IR for Historic Texts
 - From Modern to Historic
 - From Historic to Modern

What is Information Retrieval?

Definition

For a given query, find the documents d_1 to d_n from a text collection C that are relevant.

Documents are considered relevant, when they contain the query terms.

Internet search engines perform Information Retrieval.


Example



Web

[Text mining](#) - Wikipedia, the free encyclopedia

28 Jan 2009 ... **Text mining**, sometimes alternately referred to as **text data mining**, roughly equivalent to **text analytics**, refers generally to the process of ...

en.wikipedia.org/wiki/Text_mining - 57k - [Cached](#) - [Similar pages](#) - 


[Marti Hearst: What Is Text Mining?](#)

I wrote this essay for people who are curious about the topic of **text mining** after having read the New York Times article by Lisa Guernsey (10/16/2003) or ...

people.ischool.berkeley.edu/~hearst/text-mining.html - 9k - [Cached](#) - [Similar pages](#) - 

[TextMining](#)

You can find info on how to access the svn repository at the url: <http://code.google.com/p/text-mining/source/checkout>. Watch this page for documentation ...

www.textmining.org/ - 3k - [Cached](#) - [Similar pages](#) - 

IR using modern language

IR for modern languages is usually straightforward because:

- Modern written language is normalised with regards to spelling and grammar.
- Where variations in spelling exist, they are well documented and easily accessible.
- Digital resources for English and German exist and can be easily incorporated.

IR using historic language

IR for historic languages is a unique challenge, as there are problems which have not been encountered for modern languages:

- The spelling is usually not normalised: for German this only happened in 1902.
- Depending on which period the texts are from, grammar also varies from today's use.
- Spelling variations are not documented and often irregular.
- Resources are often not available in digital form.

Matching problems

The problems can be categorised as follows:

- *New word forms* - Entirely new word forms have been created.
handeln \mapsto marcken (*to trade*)
- *Latin words* - Latin words were used often by writers to seem educated to their readers.
appellacion, appellation, appelationn (*appellation*)
- *Partial new word forms* - Words have been partially changed by replacing a morpheme or a subword.
Mönchswesen \mapsto Moencherey (*monasticism*), feindselig \mapsto feindlistig (*hostile*)
- *Word splitting* - Compounds which are one word nowadays were often written as two words.
Winterzeit \mapsto Winters zeiten (*wintertime*)

Matching problems, cntd

- *Variation of prefixes/suffixes* - Prefixes and suffixes have changed their form.
Kindchen \mapsto Kindlein (*little child*)
- *Typesetting variations* - When typesetting a text, variations are seen.
When running out of *i*, typesetters used *j* instead.
- *Graphemic-phonetic variations* - Graphemic-phonetic variations happened.
Abenteuer \mapsto Abertheur (*adventure*)
- *Change in Semantics* - The meaning of a word changes over time.
Urlaub \mapsto *permission* \Rightarrow *farewell* \Rightarrow *vacation*
- *New characters* - New characters were introduced.
für \mapsto fûr

Modern to Historic

Hauser et al. (2007) have the following idea:

To find all relevant documents for a modern word, they compile possible historic variants and include those in their query.

Special Dictionaries

- Using modern and historic texts of *Dyff Vlenspiegel* they compile a dictionary that maps between modern and historic word forms.
- This dictionary can help with:
 - entirely or partially new word forms
 - latin words
 - word splitting variants for compounds
 - change of meaning
- But dictionaries are very expensive to create.

Hauser et al. call this the *first layer* of their analysis.

Rule-Based Matching

Some variations are regular:

- the variant forms for prefixes and suffixes
- typesetting alternatives
- the graphemic-phonetic change

These regularities can be described using rules.

This constitutes the *second layer*.

Word Similarity - Levenshtein Distance

Levenshtein Distance

The Levenshtein distance measures the difference between two strings via editing operations (*edit distance*). String A is transformed into string B by using the following operations and the number of operations needed is then the distance.

- Insertion: example: *Kind* (child) \Rightarrow *Kinder* (children): distance: 2
- Deletion: : example: *rund* (round) \Rightarrow *und* (and): distance: 1
- Substitution: example: *lahm* (lame) \Rightarrow *nahm* (took): distance: 1

Example: *cake* \Rightarrow *cook* has the edit distance 3:

- 1 *cake* \Rightarrow *coke*
- 2 *coke* \Rightarrow *cooke*
- 3 *cooke* \Rightarrow *cook*

Matching Based on Word Similarity

They use the standard Levenshtein distance in a slightly modified form.

- The operations are the same: insertion, deletion, substitution.
- Instead of single symbols this distance is defined on a *sequence* of symbols.

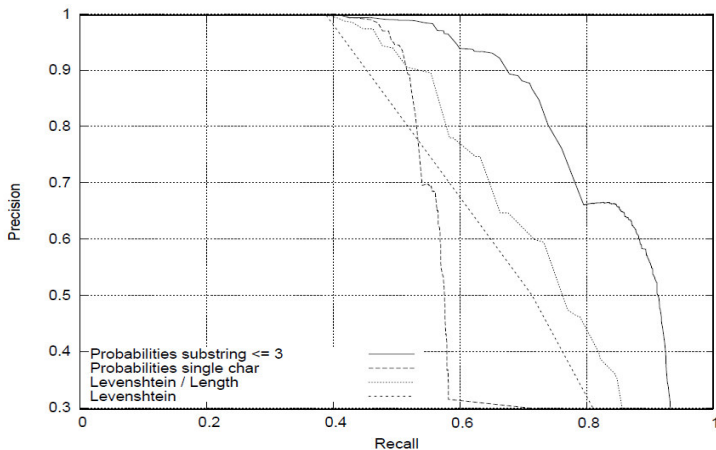
This is the *third layer*.

Ranking

The candidates from all three layers are ranked using:

- a word similarity measure
- frequency information
- heuristics

Evaluation



$$Precision = \frac{\#relevant + retrieved}{\#retrieved} \quad Recall = \frac{\#relevant + retrieved}{\#relevant}$$

From Historic to Modern

Braun et al. start from a different point of view than Hauser et al. did: Their idea is to take a historic word and map it to the corresponding modern word. They do not want to transform the old word completely into the modern one, but still make use of a stemmer for Dutch that reduces words to their stems based on their affixes.

The historic affixes, however, might not be recognisable by the stemmer, so they need to be transformed into their modern counterparts.

Heuristics

To start the mapping Braun et al. define several sets of heuristics that are going to be used to transform the historic word form.

There are sets for:

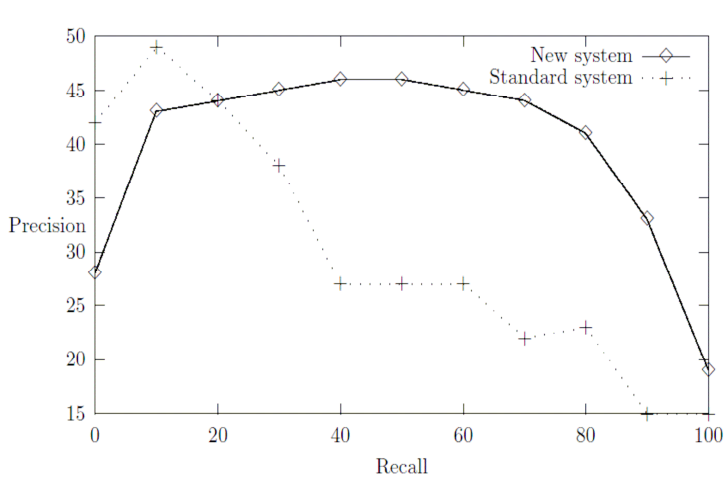
- Prefixes: **uuy**trichinghe (**ver**binden): uuyt \mapsto uit
- Infixes: uuy**trich**inghe (**ver**binden): richt \mapsto richt
- Suffixes: uuytrich**inghe** (**ver**binden): inghe \mapsto ingen

There has to be a set for each kind of affix as the transformation of an affix depends on whether it is a suffix, prefix or infix.

Subsets of Heuristics

- The suffixes and prefixes had to be divided into two subsets. After an initial transformation, additional heuristics might be required in order to map the affix to a modern one.
- Example: *lijck* → *lijk* & *ck* → *k*
- *mogelijck* (*possible/möglich*) can be transformed to either *mogelijk* or *mogelijck*.
- If the *ck* is first transformed, the rule for *lijck* cannot be applied anymore, but *mogelijck* is **not** the modern word form.
- To prevent such erroneous applications, an order of application must be enforced → subsets.

Evaluation



Conclusion

- Information Retrieval in historic texts has unique problems that differ much from the difficulties encountered in modern text.
- Words and language can change in many aspects: orthography, syntax, semantics, ...
- If you want to use IR in historic texts, you need to have some kind of mapping between historic and modern wordforms:
 - from modern to historic wordforms
 - from historic to modern wordforms
- Compiling resources manually is expensive both with regards to the time required as well as the cost of an expert writing the resource.
- Compiling resources automatically is much less expensive.

References

Andreas Hauser, Markus Heller, Elisabeth Leiss, Klaus U. Schulz, Christiane Wanzeck. *Information Access to Historical Documents from the Early New High German Period*. In: L. Burnard, M. Dobрева, N. Fuhr, A. Ldeling (eds): *Digital Historical Corpora - Architecture, Annotation, and Retrieval*. Dagstuhl Seminar Proceedings, 2007.

Loes Braun, Floris Wiesman, Ida Sprinkhuizen-Kuyper. *Information Retrieval from Historical Corpora*. Proceedings of the 3rd Dutch-Belgian Information Retrieval Workshop (DIR), Leuven, Belgium, pp. 106-112.