

Natural Language Processing in Information Retrieval

Thorsten Brants
Google Inc.
in *CLIN 2004*

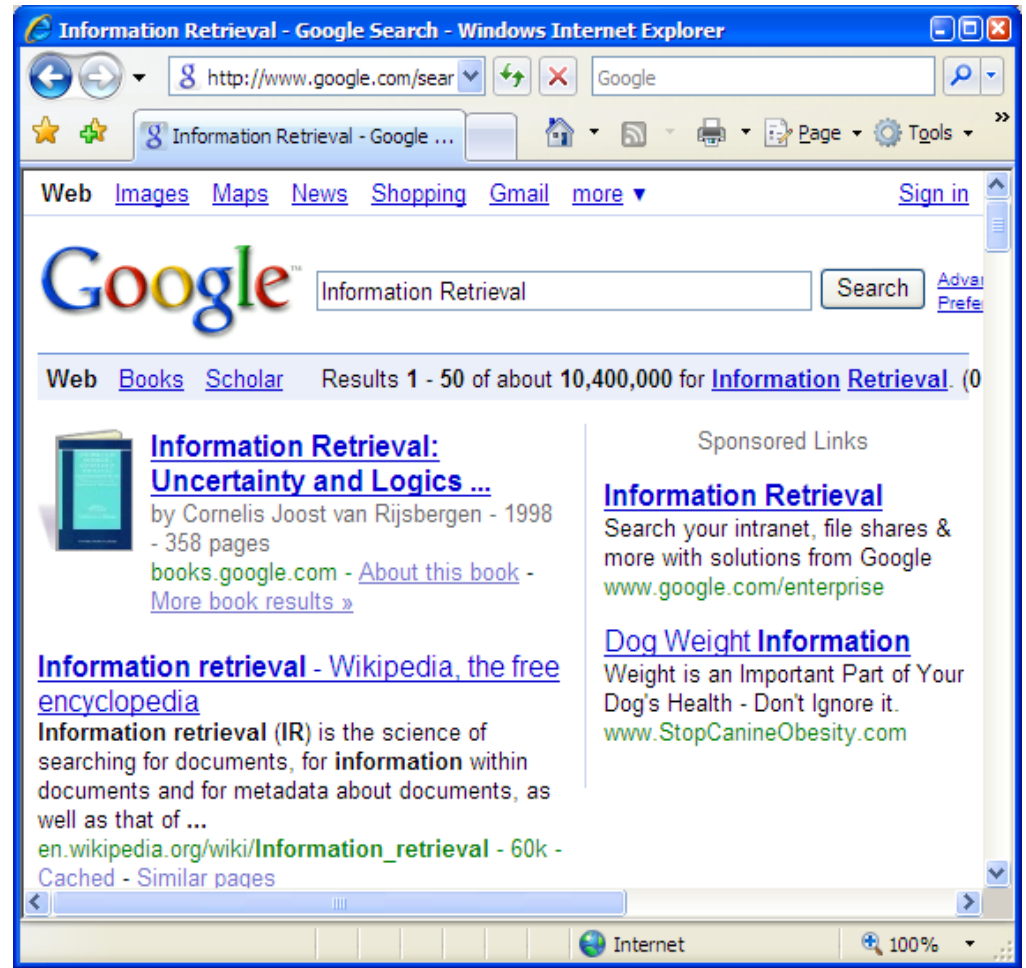
Presented by
Steven P. Crain
s.crain@gatech.edu

Outline

- What is Information Retrieval?
- NLP techniques applied to IR
 - Why expected to be helpful?
 - Why not helpful?
 - Where it might be helpful?
- Future expectations

Information Retrieval

- Document retrieval
- Ranking
- Summarization
- Add placement
- Question answering
- Passage retrieval



NLP for Information Retrieval

- Baseline: Bag of Words model
- Very efficient, but all structure is lost.

What is the topic of the following bag of words?

*Aim commercial discuss diverse engines
information methods queries retrieval search
serving stream submitted such user*

Advertising? Fishing in a stream? Cars?

- NLP should allow us to get better information from the structure and interactions of words.

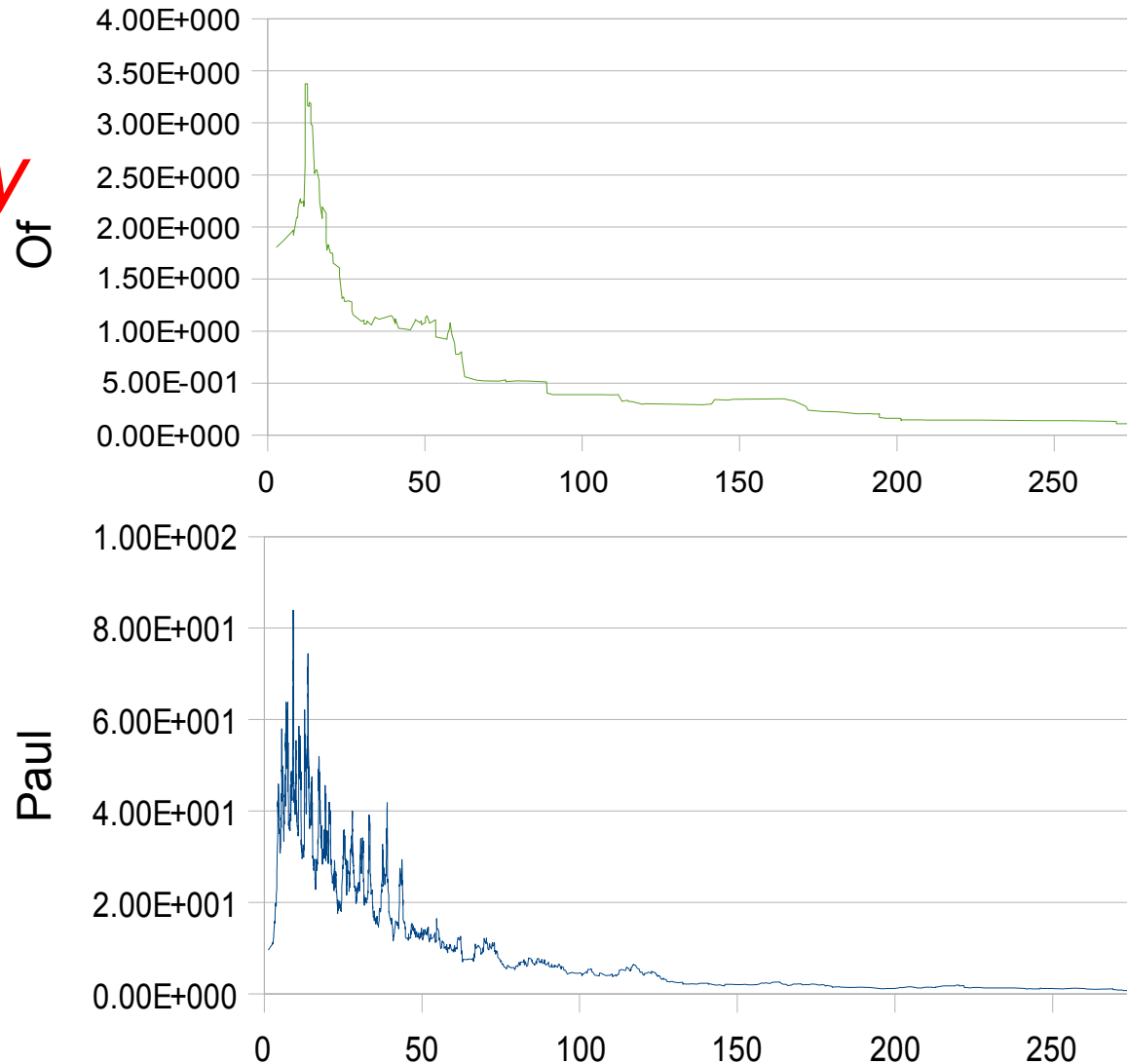
NLP Techniques for IR

Techniques that have been tried:

- Stopwords
- Stemming
- Part-of-speech tagging
- Compounds and statistical phrases
- Compound splitting
- Chunking and shallow parsing
- Head-modifier pairs
- Word sense disambiguation

Stopwords

- Remove common words that are *unlikely* to be important.
- Troublesome queries: “*To be or not to be*”
- NLP can help resolve such issues.



Stemming

- Usually, different morphs of a word are equivalent for IR tasks.
- Porter stemmer: rule-based stemming. *Port stem rule base stem.*
- Generally helpful
- Can introduce confusion
- Unknown how to optimize

Part-of-Speech Tagging

- Part-of-speech can affect semantics/relevance
 - Can improve some results
 - Degrades other results
 - Open question: how to get the improvements without the problems.
- Part-of-speech can affect weighting of query terms
 - Minimal improvement
 - Methodological problems
 - Open question: would this help a state-of-the-art algorithm?

Compounds and Statistical Phrases

- Compounds are groups of words whether the meaning is significantly different from the individual words. *New York*
- Statistical phrases are groups of words that occur frequently.
- Modeling statistical phrases is costly but also very useful (10% relative improvement).
- Open question: how to treat a statistical phrase in a query.

Compound Splitting

- Many (especially Germanic) languages form words by combining other words together.
- Typically (but by no means universally) the semantics of the combination can be inferred from the components.
- Depending on the language, splitting compound words can be very helpful.
- Same open questions as for statistical phrases.

Compound Splitting: Air Plane



<http://www.dailymail.co.uk/news/worldnews/article-1021949/Pictured-Jumbo-jet-splits-tries-off.html>

Chunking and Shallow Parsing

- Identify phrases and index them.
- Results are very good.
- Similar results can be obtained in other ways.
 - *N*-grams
 - Offset indexes

Head-Modifier Pairs

- Identify dependency relations and create features for each head-modifier pair.
- Limited success
- No improvement over n -grams
- Introduces many spurious/useless pairs
- Open question: is there a good way to identify useful pairs?

Word Sense Disambiguation

- Many words can have multiple unrelated senses.
- One attack is to ensure all meanings are reflected in the results.
- Better, NLP can help identify the intended sense and only return appropriate results.
- Works very well in specialized domains (**only**)
- **Makes results worse in general!**
- Not needed for long queries, and does not work well for short ones.

Outline

- What is Information Retrieval?
- NLP techniques applied to IR
 - Why expected to be helpful?
 - Why not helpful?
 - Where it might be helpful?
- Future expectations

Future Expectations

- Consider the 80-20 rule
- Conventional IR techniques are hitting important limitations.
- NLP techniques will be necessary to handle the difficult query situations.

Successful NLP

- Must be able to decide when NLP will help.
- Must be able to amortize computational cost.
- Must thoroughly understand what NLP will contribute, and tune techniques for specific roles.