Natural Language Processing for Information Retrieval: Challenges and Opportunities

ChengXiang Zhai

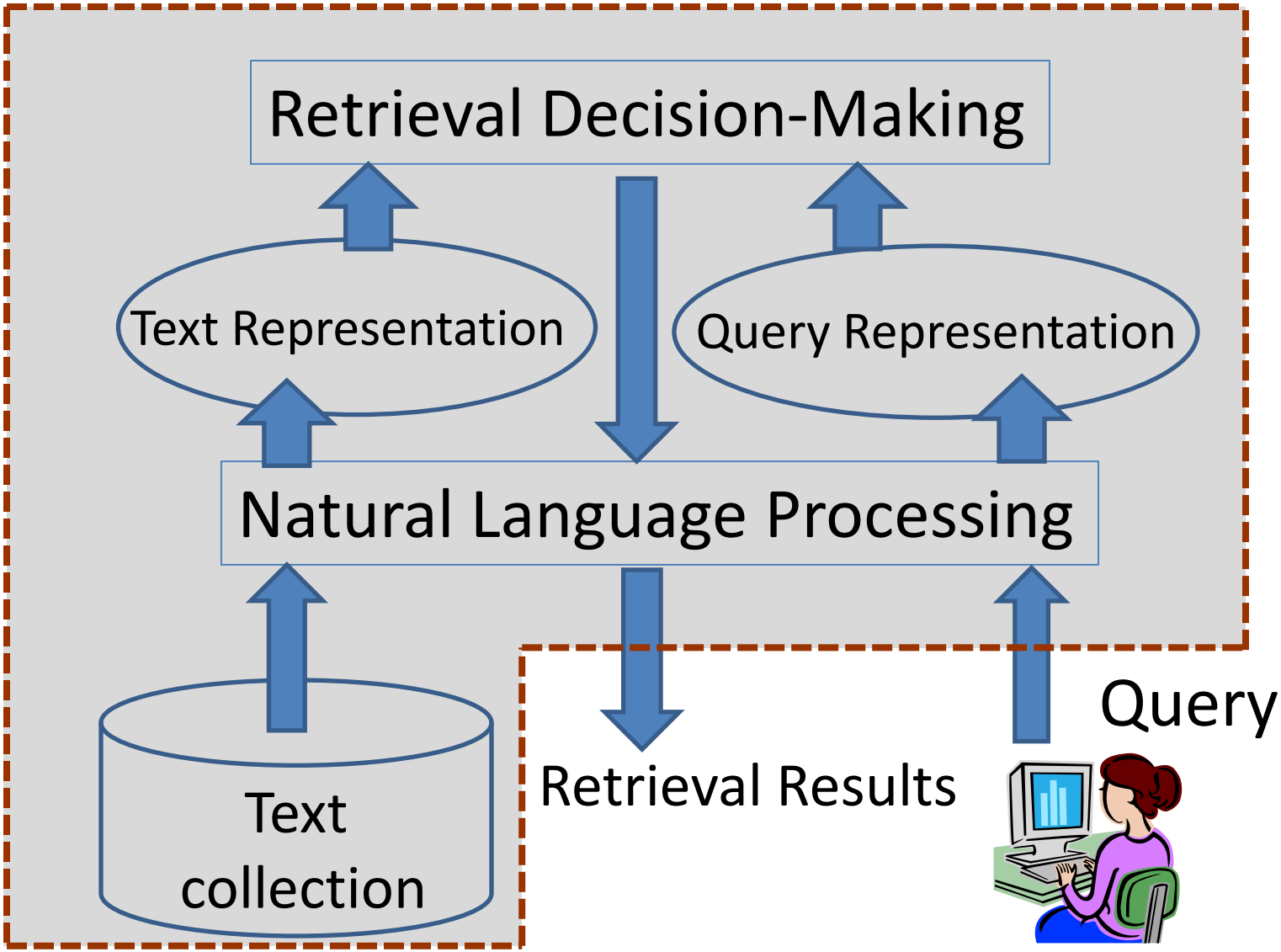
Department of Computer Science
University of Illinois at Urbana-Champaign
http://www.cs.uiuc.edu/homes/czhai

What is Information Retrieval (IR)?

- **Salton's definition (Salton 68): "information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information"**
  - Information: mostly text, but can be anything (e.g., multimedia)
  - Retrieval:
    - **Narrow sense: search/querying**
    - **Broad sense: information access; information analysis**
- **In more general terms**
  - Help people manage and make use of all kinds of information

Users are always an important factor!

NLP as Foundation of IR



IR researchers have been concerned about NLP since day one…

Luhn's idea (1958): automatic indexing based on statistical analysis of text

Hans Peter Luhn (IBM)

"It is here proposed that the frequency of word occurrence in an article furnishes a useful measurement of word significance. It is further proposed that the relative position within a sentence of words having given values of significance furnish a useful measurement for determining the significance of sentences. The significance factor of a sentence will therefore be based on a combination of these two measurements. " (Luhn 58)

LUHN, H.P., 'A statistical approach to mechanised encoding and searching of library information', *IBM Journal of Research and Development,* **1**, 309-317 (1957).
LUHN, H.P., 'The automatic creation of literature abstracts', *IBM Journal of Research and Development*, **2**, 159-165 (1958).

The notion of “resolving power of a word”

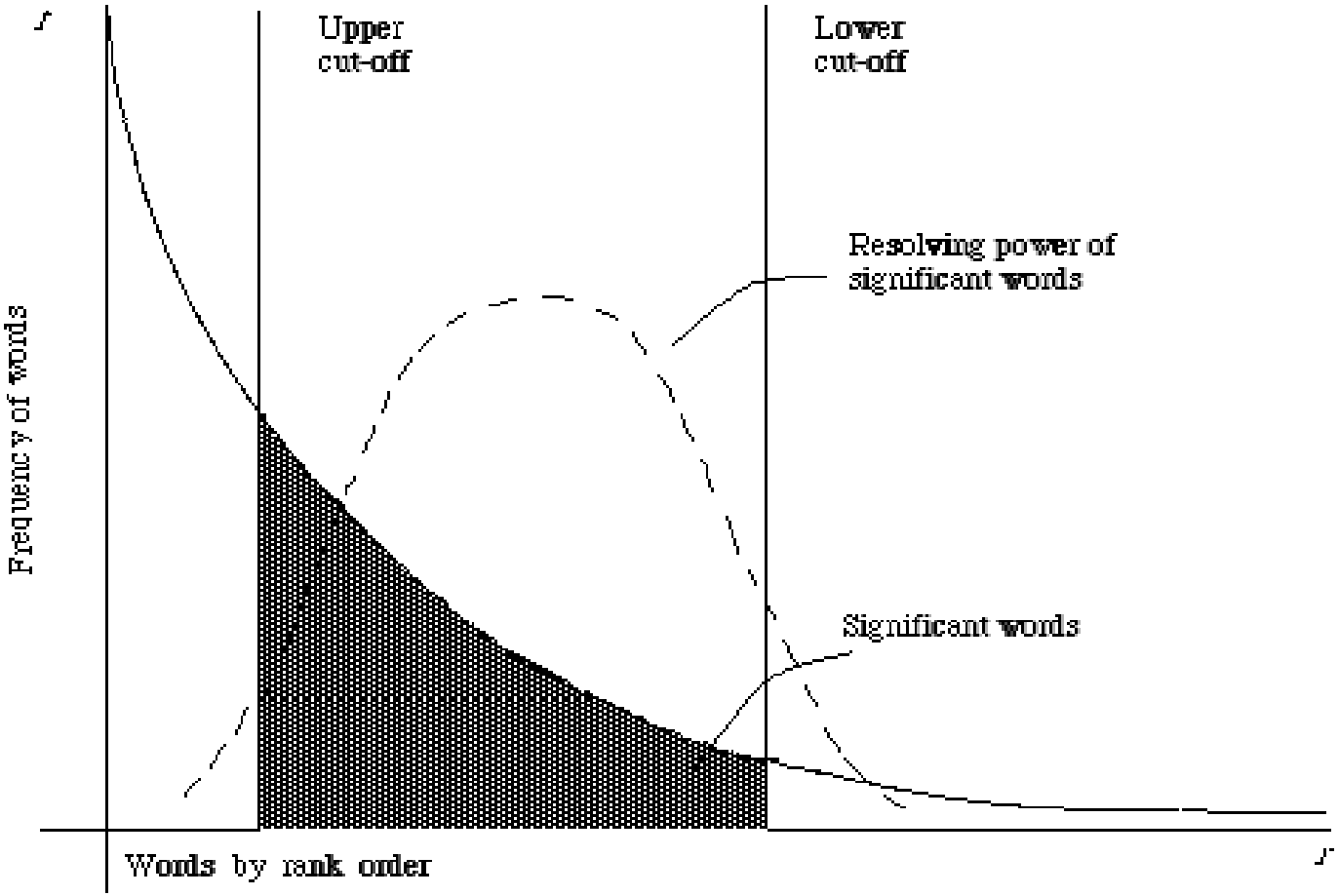


Figure 2.1. A plot of the hyperbolic curve relating f, the frequency of occurrence and r, the rank order (Adapted from Schultz[44] page 120)

Automatic abstracting algorithm [Luhn 58]

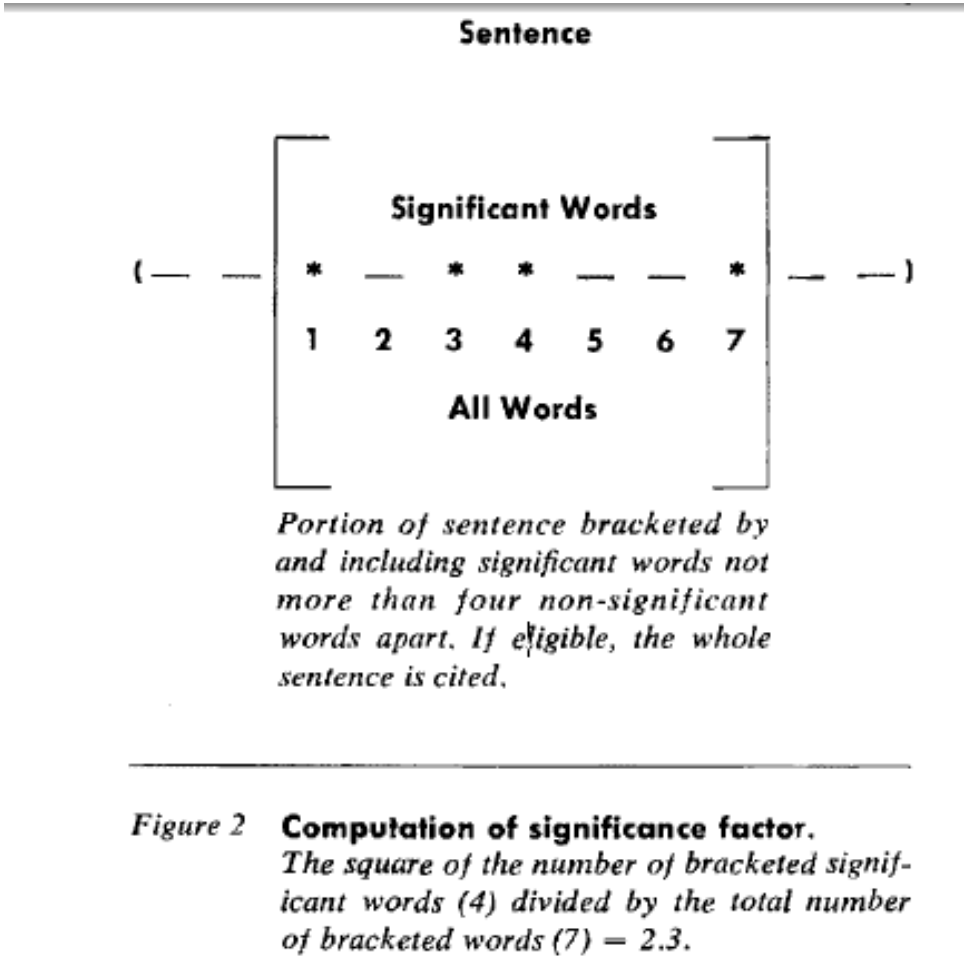


Figure 2 **Computation of significance factor.** *The square of the number of bracketed significant words (4) divided by the total number of bracketed words (7) = 2.3.*

The idea can be adapted for query-specific summarization

"In many instances condensations of documents are made emphasizing the relationship of the information in the document to a special interest or field of investigation. In such cases sentences could be weighted by assigning a premium value to a predetermined class of words."

Cleverdon's Cranfield Project (1957-1966)

Cyril Cleverdon
Cranfield Inst. of Tech, UK

Established rigorous evaluation methodology
Introduced precision & recall
Compared different linguistic units for indexing

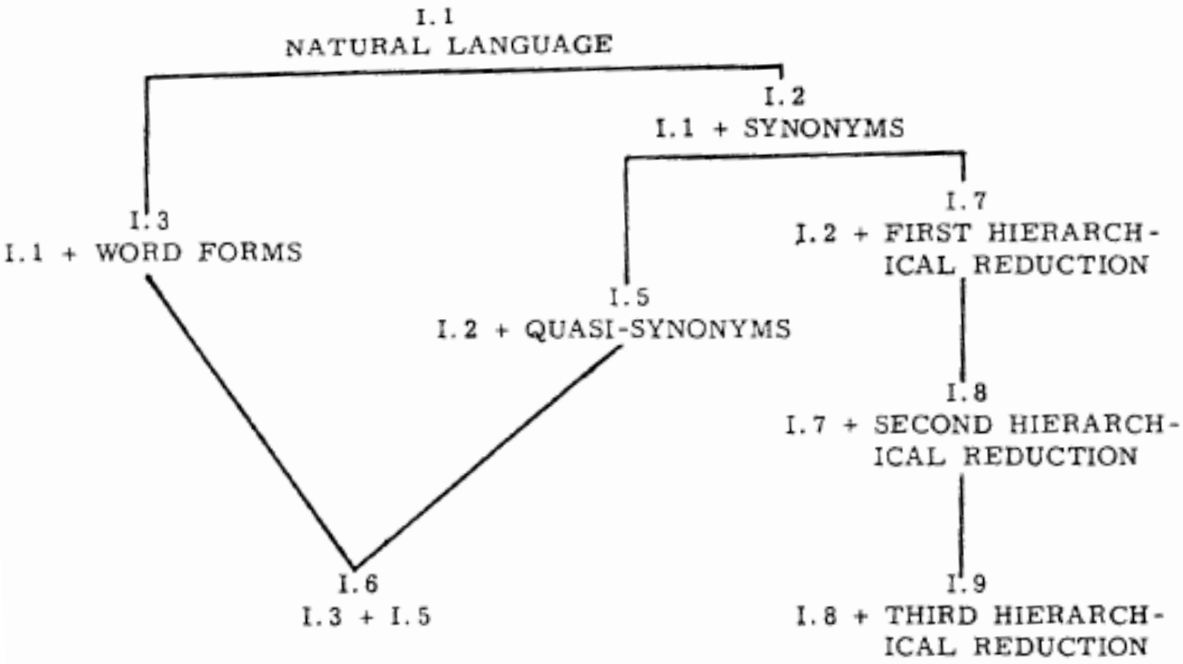


FIGURE 3 SINGLE TERM INDEX LANGUAGES

Indexing and Abstracting by Association

Doyle, Lauren B, *American Documentation (pre-1986);* Oct 1962;

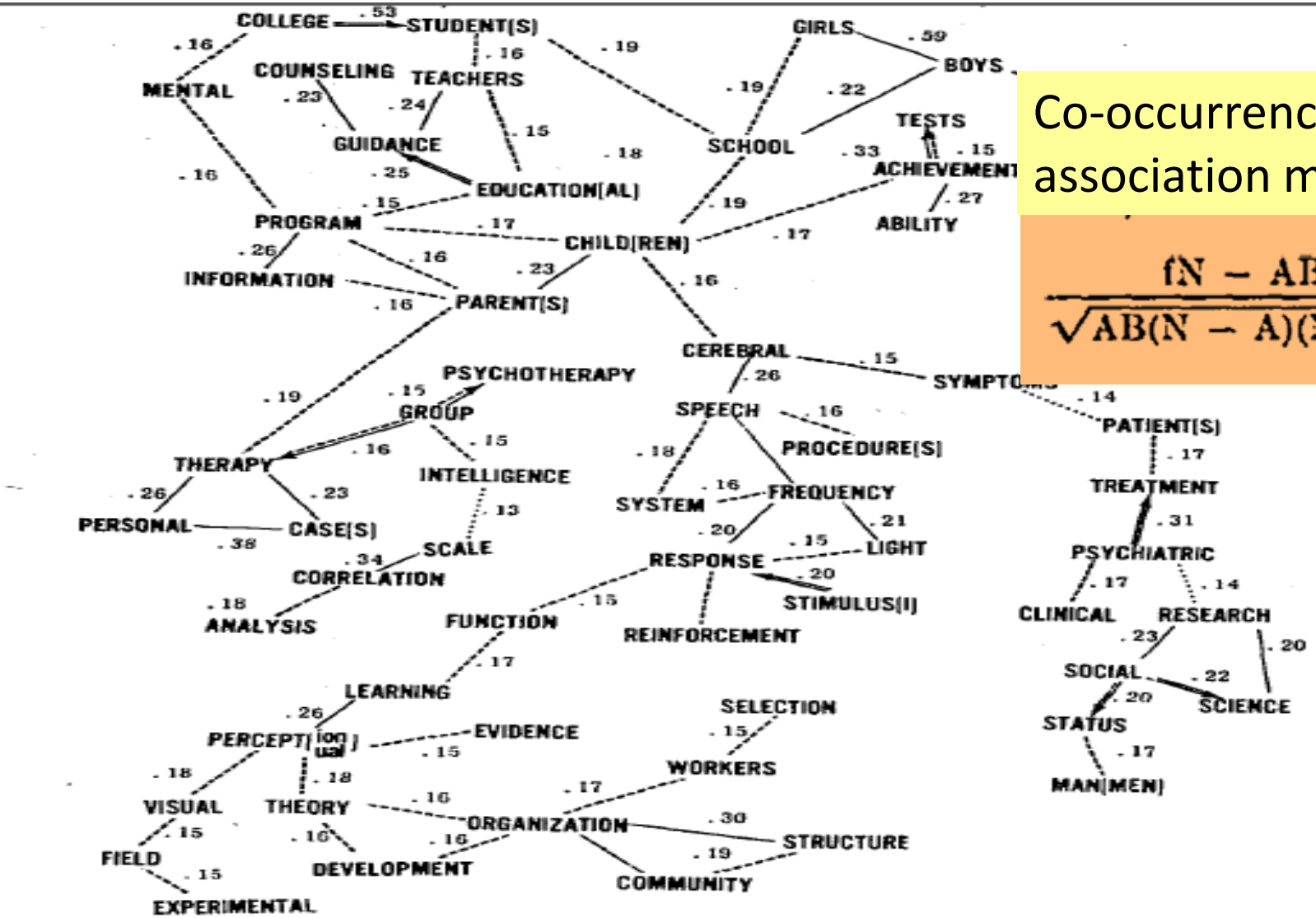


Fig. 1. An Association Map Based on Pearson Correlation Coefficients.

Co-occurrence-based association measure

$$\frac{fN - AB}{\sqrt{AB(N - A)(N - B)}}$$

And many attempts have been made on improving IR with NLP techniques since then…

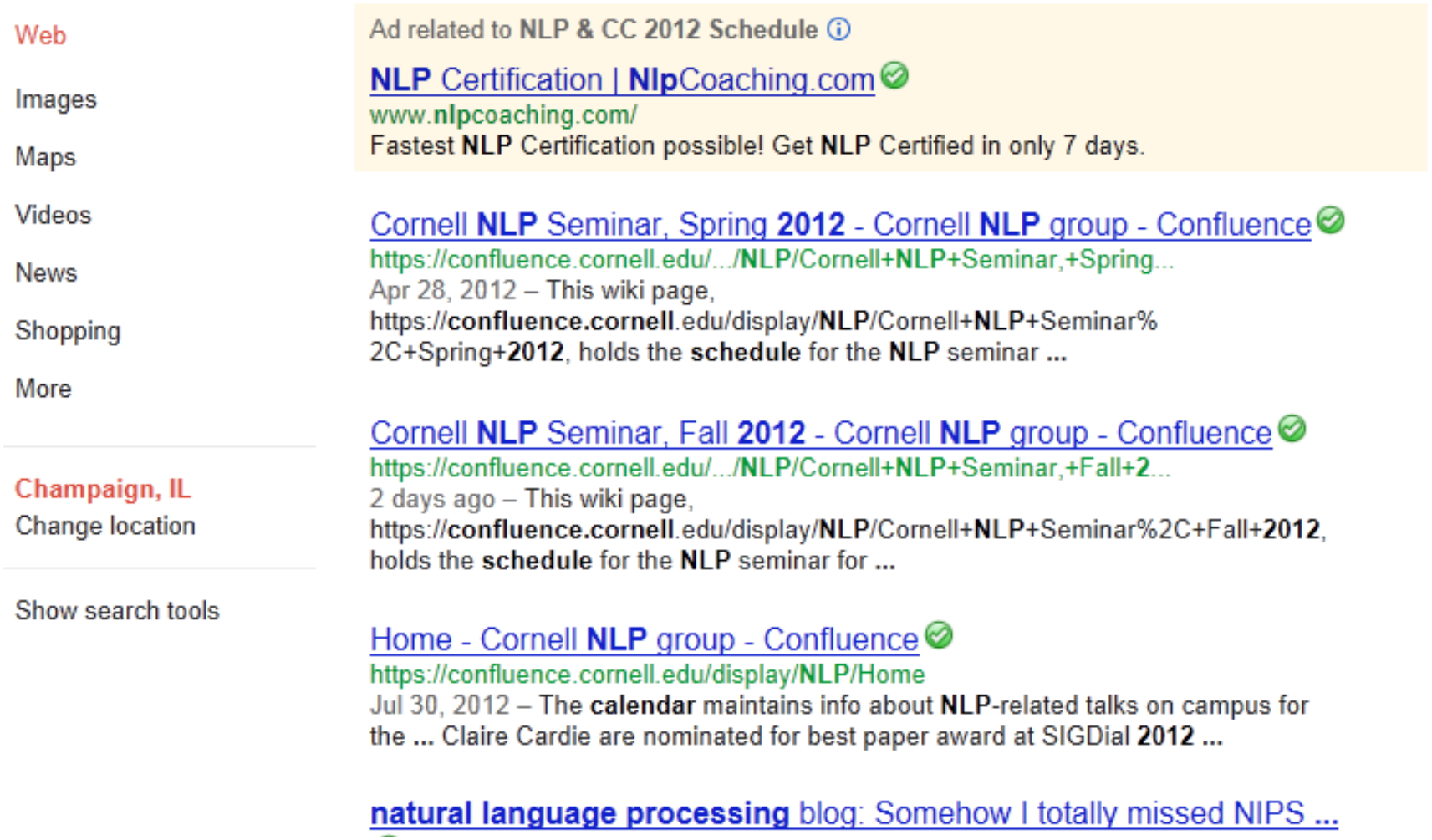
However, today's search engines don't use much NLP!

Sometimes, they appear to "understand" natural language

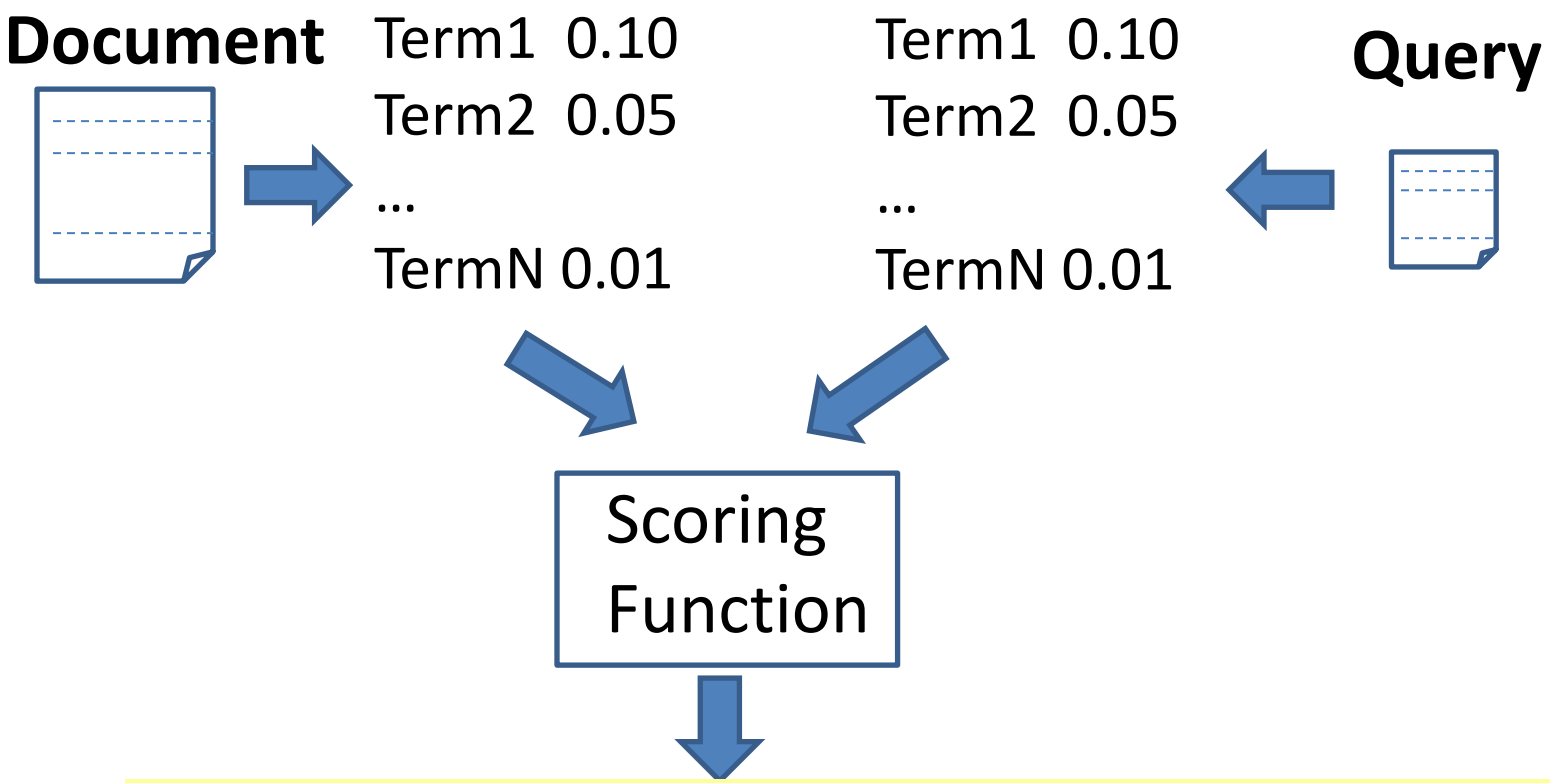
Query: "NLP & CC 2012"

However,…

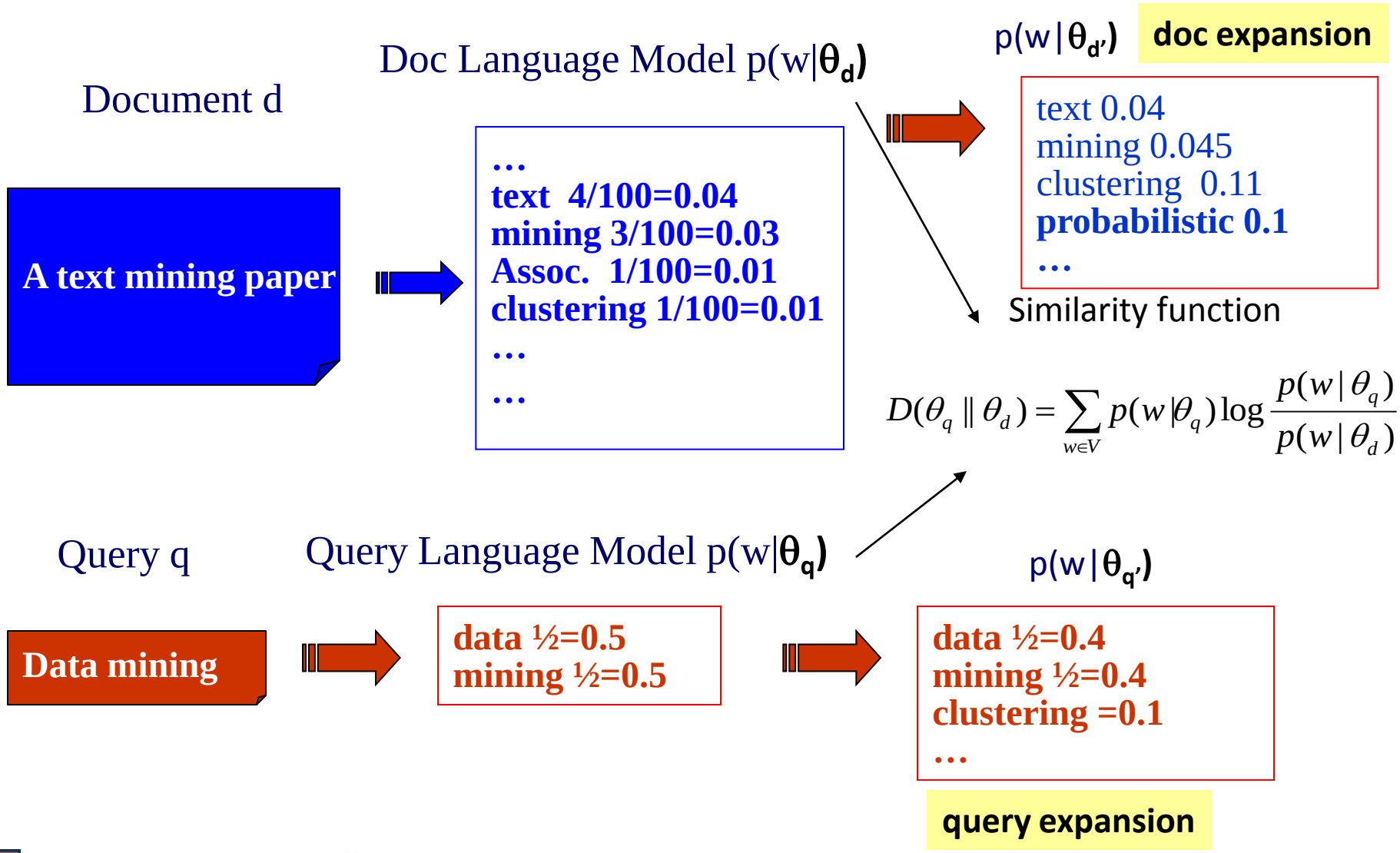
Query: "NLP & CC 2012 schedule"



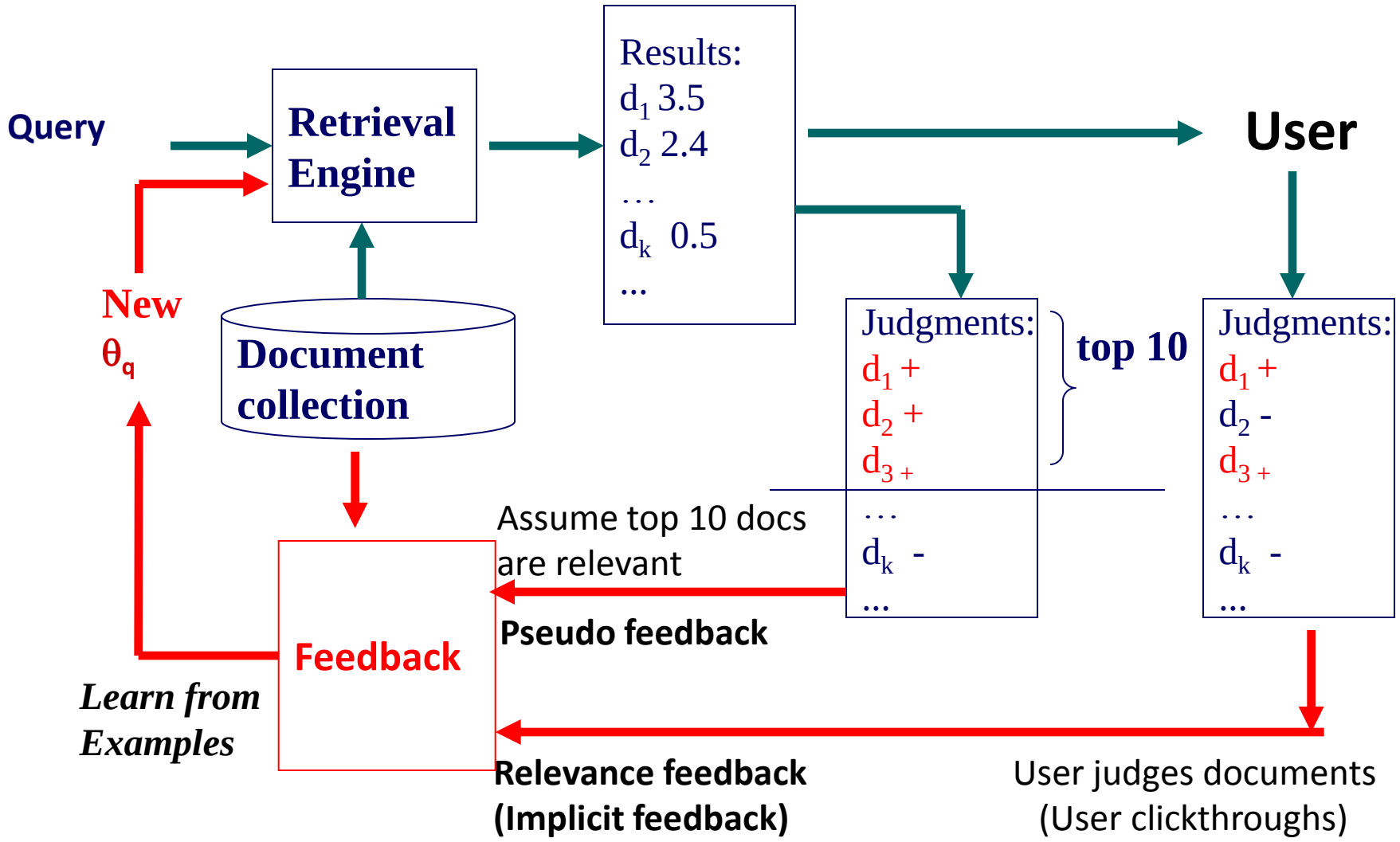
How does a typical search engine work?
Bag of Terms Representation



A Typical Ranking Function



Feedback in IR



Search Engines Generally Do Little NLP

- **Bag of words representation remains the pillar of modern IR**
- **Simple lexical processing: stop words removal, stemming, word segmentation**
- **Limited uses of phrases**

Basic Technique = Keyword Matching

+ statistical weighting of terms
+ leveraging clickthroughs (feedback)
+ …

NLP <= Lexical Analysis (?)

IR researchers don't talk much about NLP today either

Assumed Conclusion: NLP isn't useful for IR…

Questions

- **If logically NLP is the foundation of IR, why hasn't NLP made a significant impact on IR?**
- **Is there any way to improve IR significantly with NLP techniques?**
- **What does the future of NLP for IR look like?**

Rest of the Talk

- **Attempts on applying NLP to IR**
- **Why hasn't it be successful?**
- **The future of NLP for IR**

NLP for IR 1: Beyond bag-of-words Representation

- **Motivation: single words have many problems**

Different words, same meaning: car vs. vehicle

Same words, different meaning: Venetian Blinds vs. blind venetians

Different perspectives on single concept:
"The accident" vs. "the unfortunate incident"

Different meanings for the same words in different domains:
"sharp" can mean "pain intensity" or "the quality of a cutting tool"

[Smeaton's ESSIR'95 tutorial]

Many different phrases explored

- **Statistical phrases [Fagan 88]**
  - Phrases are frequent n-grams
- **Linguistic phrases [Fagan 88, Zhai & Evans 96]**
  - Phrases are obtained from parsing
- **Lexical atoms [Zhai et al. 95; Zhai 97]**
  - "Sticky phrases"/non-compositional phrases (e.g., "hot dog", "blue chip")
- **Head-modifier pairs [Strzalkowski & Vauthey 95, Zhai 97]**

  "fast algorithm for parsing context-free languages"

  ➔{"fast algorithm", "parsing algorithm", "parsing language", "context-free language"}

  …

Phrase Indexing: Results

- **Mostly mixed results**
  - Some reported insignificant improvement over single word baseline
  - Others reported degradation of retrieval accuracy
- **While on average, using phrases may help, it doesn't help all queries**
- **Even when adding phrases helps, adding "too many" phrases can hurt the performance**
- **Mixing phrases with single words is generally necessary to improve robustness**

Sample Phrase Indexing Results [Zhai 97]

Experiments	Recall (Ret-Rel)	Init Prec	Avg Prec
WD-SET	0.56(597)	0.4546	0.2208
WD-HM-SET	0.60(638)	0.5162	0.2402
inc over WD-SET	7%	14%	9%
WD-NP-SET	0.58(613)	0.5373	0.2564
inc over WD-SET	4%	18%	16%
WD-HM-NP-SET	0.63(666)	0.4747	0.2285
inc over WD-SET	13%	4%	3%
Total relevant documents: 1064			

Table 1: Effects of Phrases with feedback and TREC-5 topics

Too many phrases hurt performance!

NLP for IR 2: Sense Disambiguation

- **Motivation**
  - Terms are often ambiguous, causing mismatches
  - What about using term disambiguation?
- **Many studies**
  - Krovetz and Croft 1992
  - Voorhees 1993
  - Sanderson 1994
  - Schultz and Pedersen 1995
  - Stokoe et al. 2003
  - …

Disambiguation Results: Non-Promising

- **Manual sense disambiguation [Korvetz & Croft 92]**
  - Very little improvement (<=2% improvement)
  - Possibly degrade performance
  - Explanation: coordination of terms; skewed distribution of senses
- **Automatic sense disambiguation based on WordNet [Voorhees 93]**
  - No improvement
- **"pseudo sense" experiments [Sanderson 94]**

IR performance is very sensitive to erroneous disambiguation …
Only when it gets to 90% accuracy it is as good as no disambiguation…
Beyond that, it yields improvement, but only when the query is short

Disambiguation Results: More Promising

- **Corpus-based senses [Schultz & Pedersen 92]**
  - Senses are acquired by clustering word context
  - Multiple senses are assigned to combat uncertainty
- **Semcor 1.6 + careful weighting [Stokoe et al. 03]**

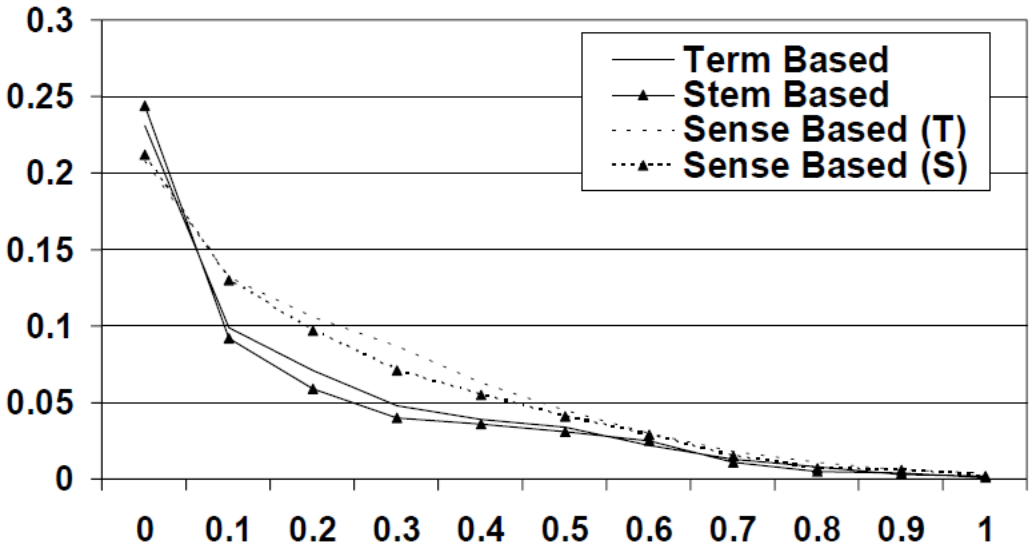
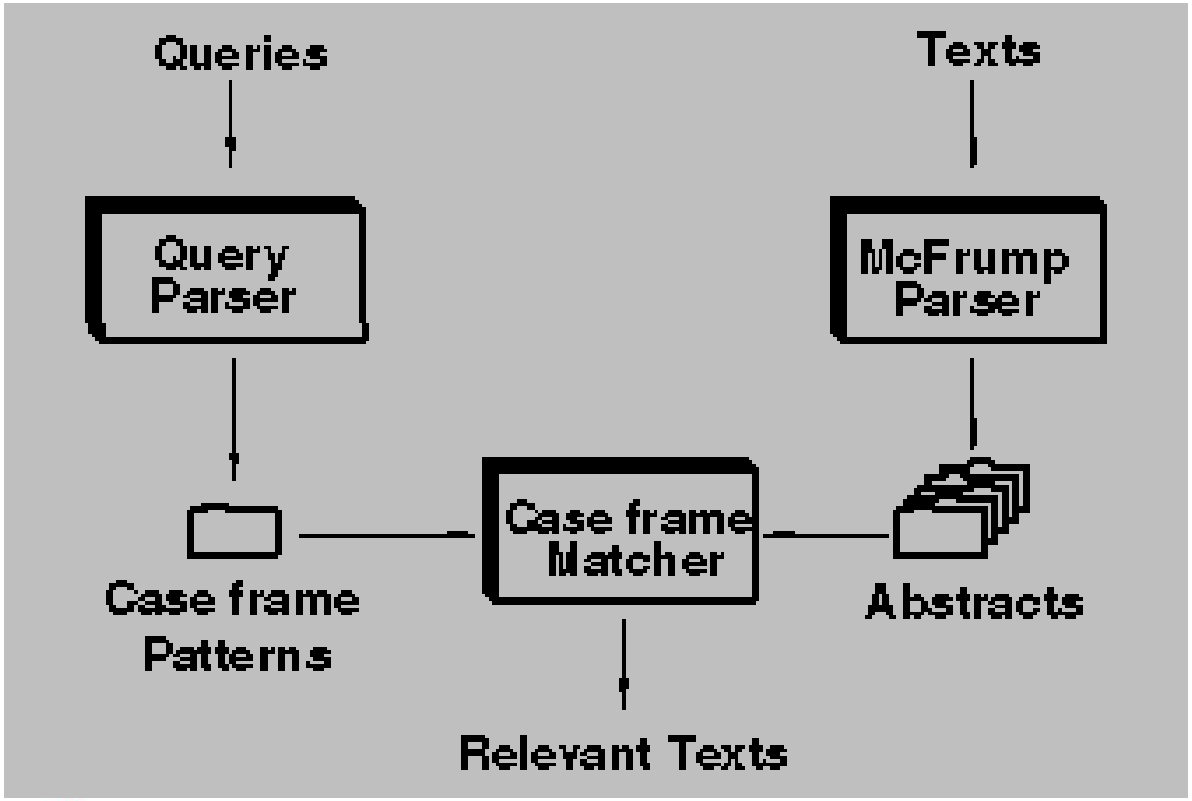


Figure 3. Plots precision for 11 recall points for the Term Based, Stem Based, Sense Based (T), and Sense Based (S) retrieval runs

NLP for IR 3: Deeper Semantic Representation: FERRET [Mauldin 91]

- using knowledge representation to represent text
- works for a very small data set in astronomy domain
- but, doesn't scale up, possibly not outperforming stronger baseline



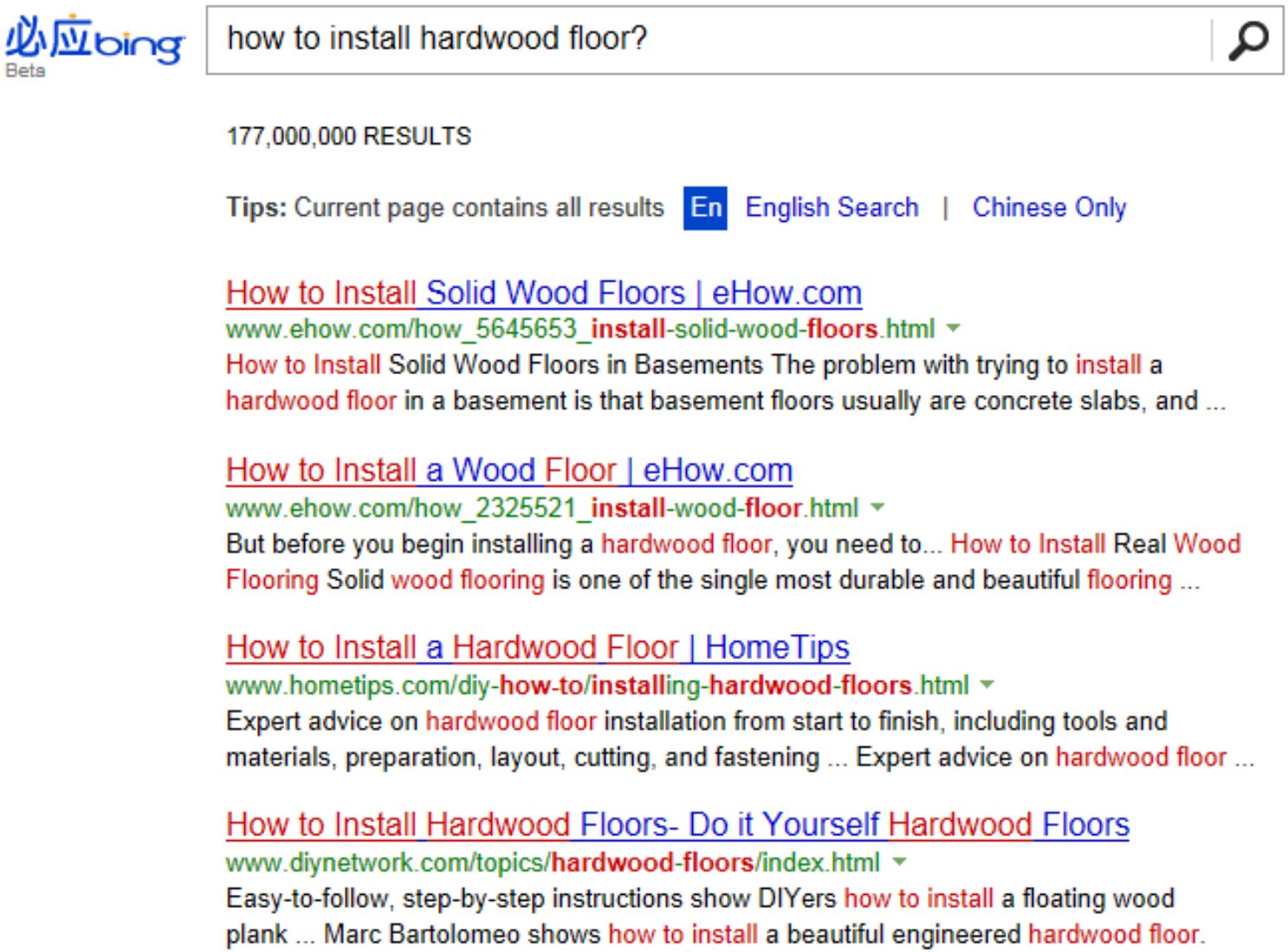
Rest of the Talk

- Attempts on applying NLP to IR
- **Why hasn't it be successful?**
- The future of NLP for IR

Explanation 1: The Power of Bag of Words Representation

- **Retrieval problem is mostly a simple language processing task**
- **"Matching" is sufficiently useful for finding relevant documents**
- **Ideal query hypothesis: given any subset of documents that we assume a user is interested in, there exists a query that would produce near-ideal ranking**
- **Finding an ideal query doesn't necessarily need deep NLP**

Keyword matching may answer questions!



Explanation 2: NLP wasn't used to solve a big pain

feedback & expansion can take care of this

Different words, same meaning: car vs. vehicle

Same words, different meaning: Venetian Blinds vs. blind venetians

Different perspectives on single concept:
"The accident" vs. "the unfortunate incident"

Different meanings for the same words in different domains:
"sharp" can mean "pain intensity" or "the quality of a cutting tool"

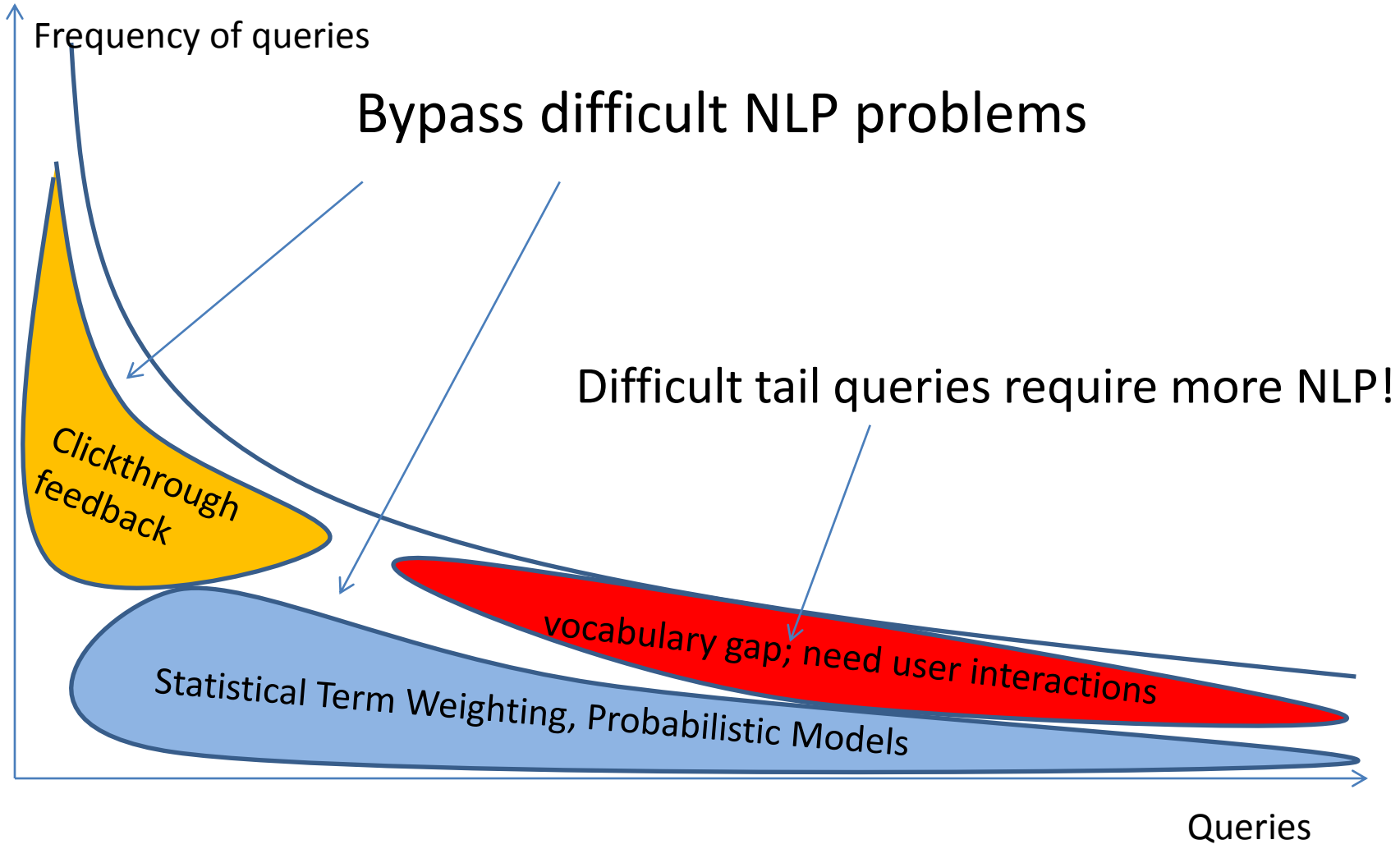
How likely does this happen?

Some times domain restriction solves the problem naturally. Other words in the query help providing disambiguation.

Explanation 3: Lack of consideration of robustness

- **Standard IR models are optimized for bag of terms representation**
- **When incorporating phrases, we no longer have optimal term weighting**
  - e.g., how to optimize phrase weighting when single words are also used for indexing?
- **Need to tolerate NLP errors**

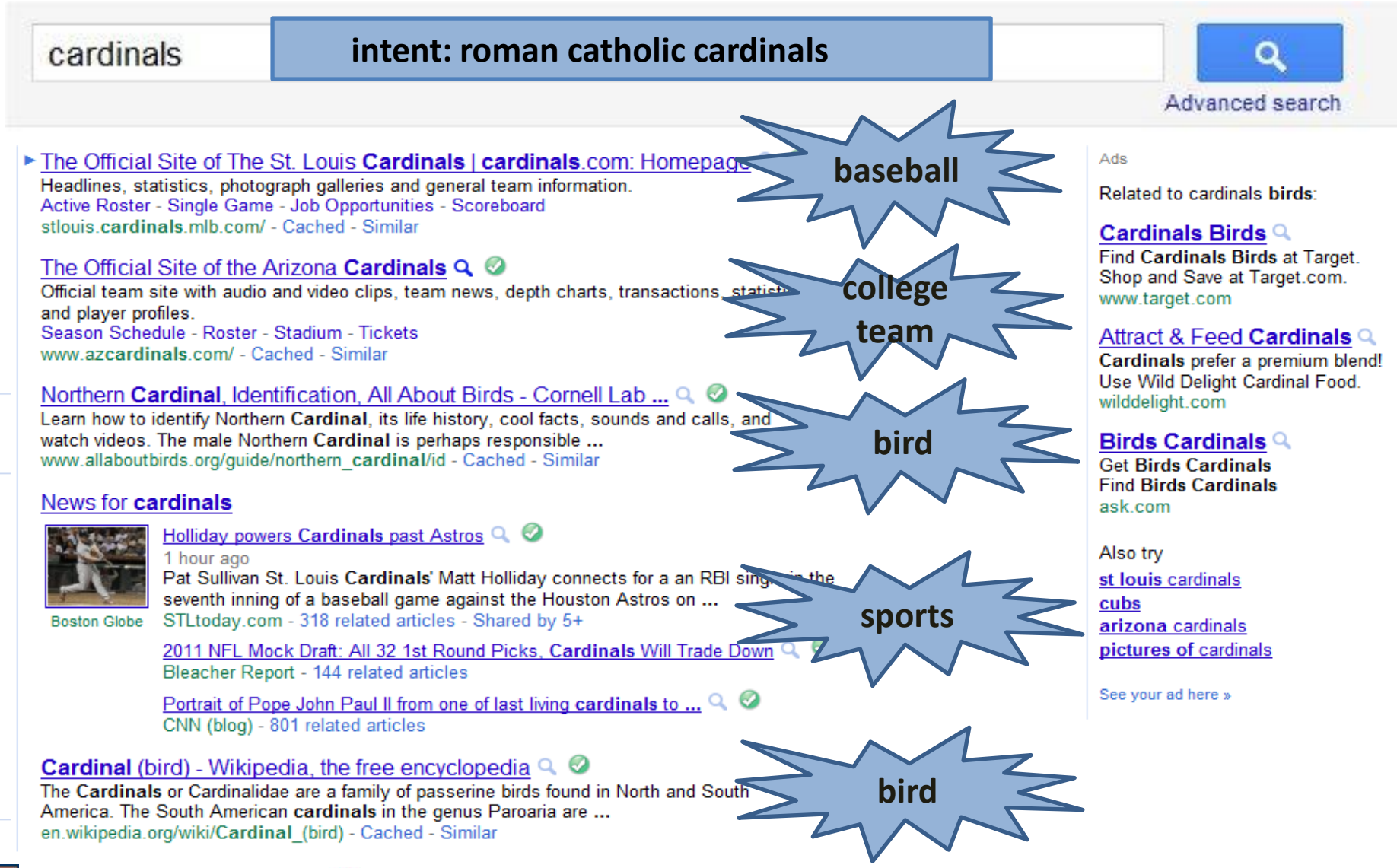
Explanation 4: Workaround possible



Example of NLP for tail queries: sense clarification [Kotov & Zhai 11]

- **Uses global analysis for sense identification:**
  - + does not rely on retrieval results (can be used for **difficult queries**)
  - + identifies **collection-specific senses** and avoids the coverage problem
  - + identifies **both majority and minority senses**
  - + domain independent
- **Presents concise representations of senses to the users:**
  - + eliminates the cognitive burden of scanning the results
- **Allows the users to make the final disambiguation choice:**
  - + **leverages user intelligence** to make the best choice

Query ambiguity



Query ambiguity

Sense feedback improves retrieval accuracy on difficult topics

		KL	KL-PF	SF
AP88-89	**MAP**	0.0346	0.0744	**0.0876**
	P@10	0.0824	0.1412	0.2031
ROBUST04	**MAP**	0.04	0.067	**0.073**
	P@10	0.1527	0.1554	0.2608
AQUAINT	**MAP**	0.0473	0.0371	**0.0888**
	P@10	0.1188	0.0813	0.2375

- Sense feedback outperforms PRF in terms of MAP and particularly in terms of Pr@10 (**boldface** = statistically significant ($p<.05$) w.r.t. KL; underline = w.r.t. to KL-PF)

Rest of the Talk

- Attempts on applying NLP to IR
- Why hasn't it be successful?
- **The future of NLP for IR**

Future of NLP for IR: Challenges

- **Grand Challenge: How can we leverage imperfect NLP to create definite value for IR?**
- **Possible Strategies**
  - Create add-on value: supplement rather than replace existing IR techniques
  - Integrate NLP into a retrieval model (minimize "disruption")
  - Multi-resolution representation
  - Include users into the loop

NLP for IR Opportunity 1: long-tail queries

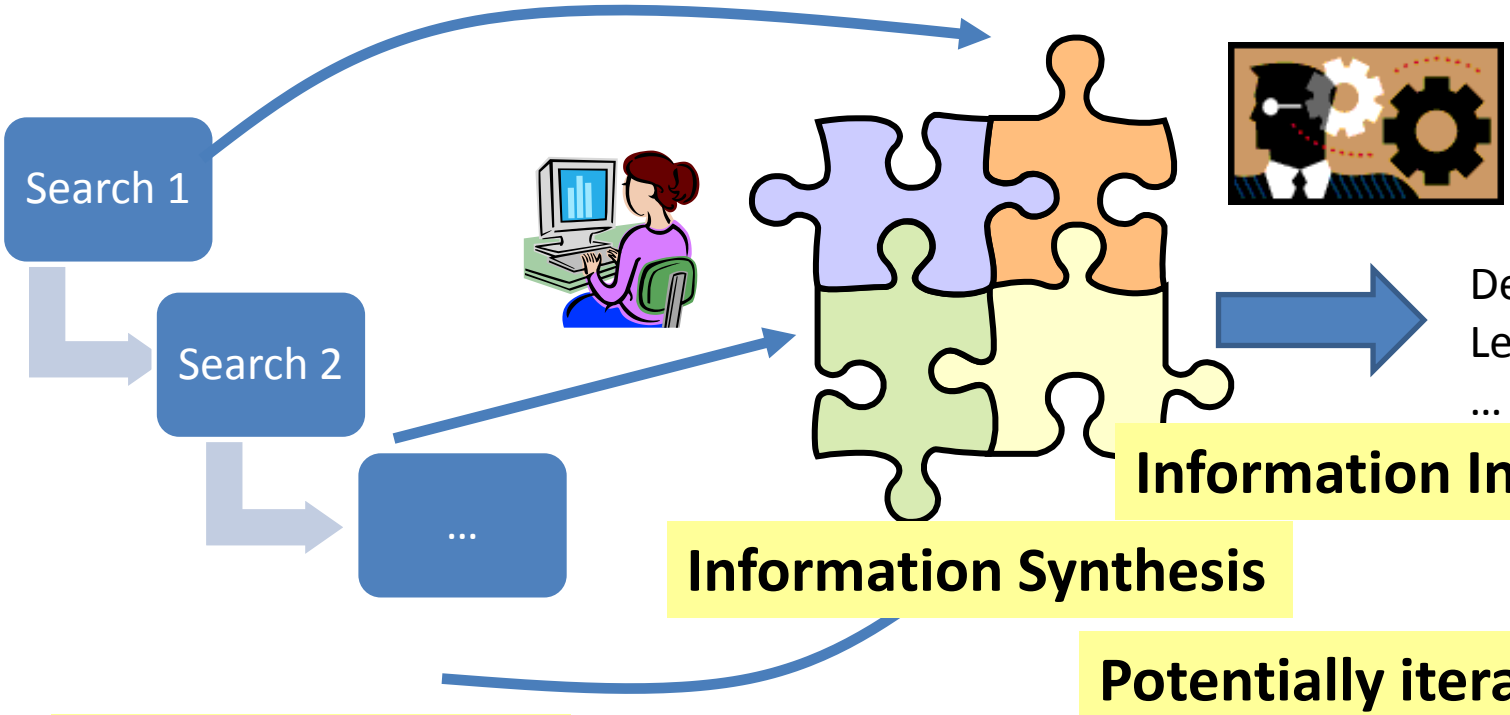
- **NLP for query understanding in context**
  - Query segmentation
  - Query parsing
  - Query interpretation
  - Do all these in the context of search session and user interaction history
- **NLP for document browsing**
  - When querying fails, browsing helps
  - How to create a multiresolution topic map?
- **NLP for interactive search**
  - How to generate clarification questions?

NLP for IR Opportunity 2: Beyond Topical Relevance

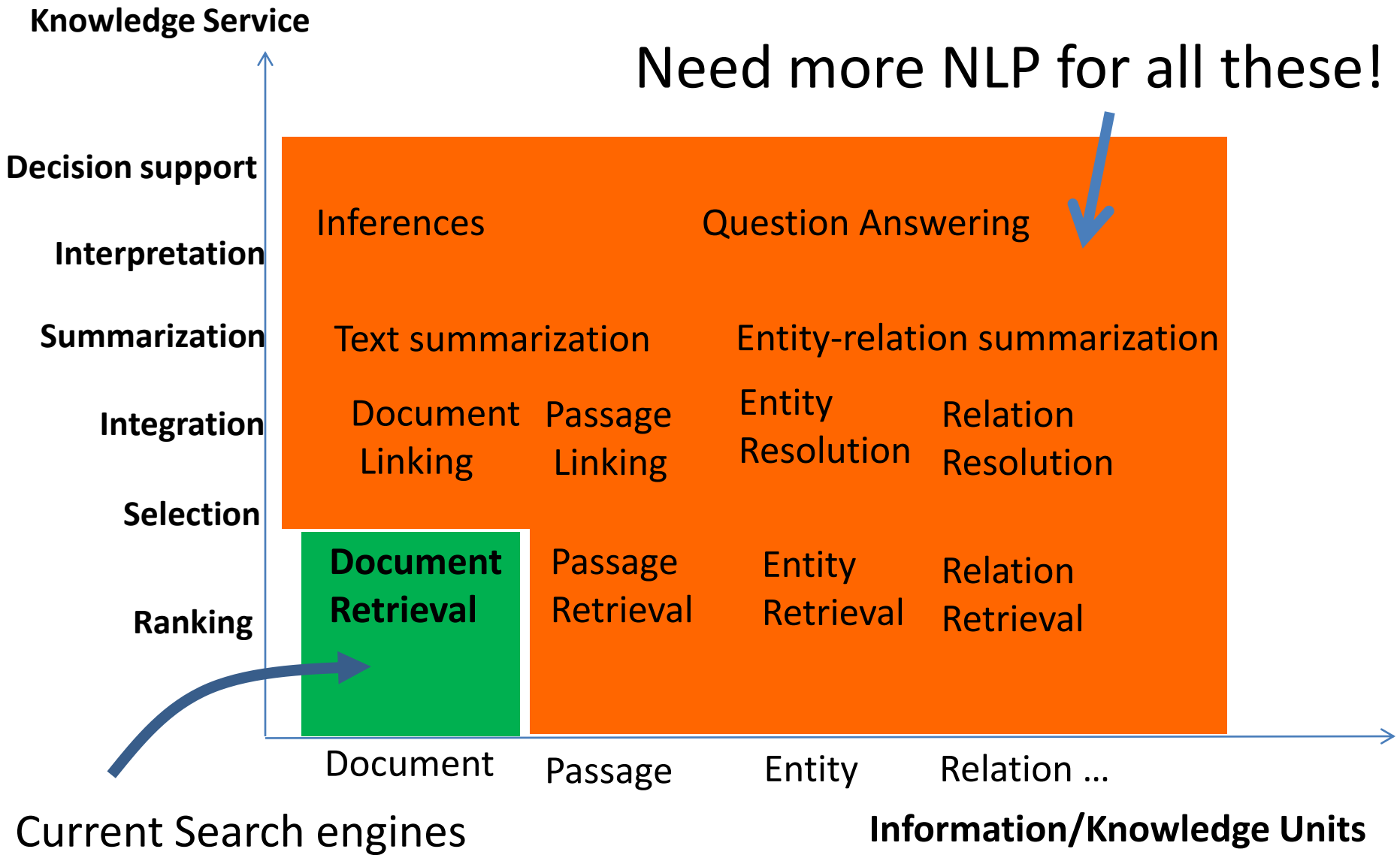
- **Traditional IR work has focused on exclusively topical relevance**
- **Real users care about other dimensions of relevance as well**
  - Sentiment/Opinion retrieval: find positive opinions about X
  - Readability: find documents with readability level of elementary school students
  - Trustworthiness
  - Genre
  - …

NLP for IR Opportunity 3+: Beyond Search

Search → **Information Synthesis & Analysis** → **Task Completion**



Towards an intelligent knowledge service system



Summary

- **NLP is the foundation of IR, but keyword matching is quite powerful**
- **NLP for IR hasn't been so successful because of the focus on document retrieval (narrow sense of IR)**
- **Many more opportunities in applying NLP to IR in the future**
  - Need to supplement, rather than replace existing IR techniques
  - Aim at more intelligent,interactive knowledge service system

Thank You!

Questions/Comments?