University of Twente

Information Retrieval Modeling

Russian Summer School in Information Retrieval

Djoerd Hiemstra
http://www.cs.utwente.nl/~hiemstra

University of Twente
The Netherlands

University of Twente
The Netherlands

Overview

1. Smoothing methods
2. Translation models
3. Document priors
4. ...

Course Material

- Djoerd Hiemstra, “Language Models, Smoothing, and N-grams'', In M. Tamer Özsu and Ling Liu (eds.) *Encyclopedia of Database Systems, Springer*, 2009

Noisy channel paradigm (Shannon 1948)

- hypothesise all possible input texts I and take the one with the highest probability, symbolically:

$$\hat{I} = \underset{I}{\operatorname{argmax}}\, P(I \mid O)$$
$$= \underset{I}{\operatorname{argmax}}\, P(I) \cdot P(O \mid I)$$

Noisy channel paradigm (Shannon 1948)

- hypothesise all possible documents D and take the one with the highest probability, symbolically:

$$\hat{D} = \underset{D}{\operatorname{argmax}}\, P(D \mid T_1, T_2, \cdots)$$
$$= \underset{D}{\operatorname{argmax}}\, P(D) \cdot P(T_1, T_2, \cdots \mid D)$$

Noisy channel paradigm

- Did you get the picture? Formulate the following systems as a noisy channel:
  - Automatic Speech Recognition
  - Optical Character Recognition
  - Parsing of Natural Language
  - Machine Translation
  - Part-of-speech tagging

Statistical language models

- Given a query $T_1, T_2, \ldots, T_n$, rank the documents according to the following probability measure:

$$P(T_1, T_2, \ldots, T_n | D) = \prod_{i=1}^{n} \left((1 - \lambda_i) P(T_i) + \lambda_i P(T_i | D) \right)$$

λ_i : probability that the term on position i is important

$1 - \lambda_i$: probability that the term is unimportant

$P(T_i | D)$: probability of an important term

$P(T_i)$: probability of an unimportant term

Statistical language models

- Definition of probability measures:

$$P(T_i = t_i \mid D = d) = \frac{tf(t_i, d)}{\sum_t tf(t, d)} \quad \text{(important term)}$$

$$P(T_i = t_i) = \frac{df(t_i)}{\sum_t df(t)} \quad \text{(unimportant term)}$$

$$\lambda_i = 0.5$$

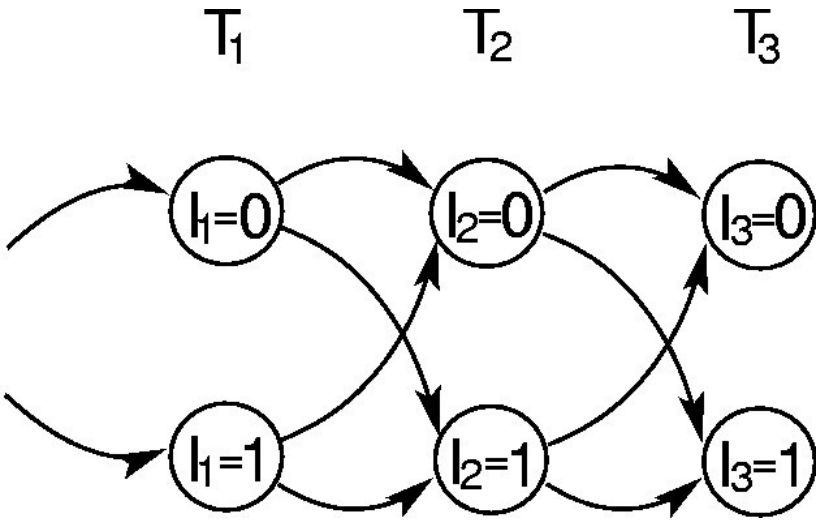
Statistical language models

- How to estimate value of λ_i ?
  - For ad-hoc retrieval (i.e. no previously retrieved documents to guide the search)
    $\lambda_i = constant$ (i.e. each term equally important)
  - Note that for extreme values:
    $\lambda_i = 0$: term does not influence ranking
    $\lambda_i = 1$: term is mandatory in retrieved docs.
    $\lim \lambda_i \rightarrow 1$: docs containing n query terms are ranked above docs containing $n-1$ terms

(Hiemstra 2002)

Statistical language models

- Presentation as hidden Markov model
  - finite state machine: probabilities governing transitions
  - sequence of state transitions cannot be determined from sequence of output symbols (i.e. are hidden)



Statistical language models

- Implementation

$$P(T_1, T_2, \cdots, T_n|D) = \prod_{i=1}^{n} ((1-\lambda_i)P(T_i) + \lambda_i P(T_i|D))$$

$$\vdots$$

$$P(T_1, T_2, \cdots, T_n|D) \propto \sum_{i=1}^{n} \log\left(1 + \frac{\lambda_i P(T_i|D)}{(1-\lambda_i)P(T_i)}\right)$$

Statistical language models

- Implementation as vector product:

$$score(q,d) = \sum_{k \in \text{ matching terms}} q_k \cdot d_k$$

$$q_k = tf(k,q)$$

$$d_k = \log\left(1 + \frac{tf(k,d)}{df(k)} \frac{\sum_t df(t)}{\sum_t tf(t,d)} \frac{\lambda_k}{1-\lambda_k}\right)$$

Smoothing

- Sparse data problem:
  - many events that are plausible in reality are not found in the data used to estimate probabilities.
  - i.e., documents are short, and do not contain all words that would be good index terms

No smoothing

- Maximum likelihood estimate

$$P(T_i = t_i | D = d) = \frac{tf(t_i, d)}{\sum_t tf(t, d)}$$

  - Documents that do *not* contain all terms get zero probability (are not retrieved)

Laplace smoothing

- Simply add 1 to every possible event

$$P(T_i = t_i \mid D = d) = \frac{tf(t_i, d) + 1}{\sum_t (tf(t, d) + 1)}$$

  - over-estimates probabilities of unseen events

Linear interpolation smoothing

- Linear combination of maximum likelihood and model that is less sparse

$$P(T_i|D) = (1-\lambda)\,P(T_i) + \lambda\,P(T_i|D), \text{ where } 0 \le \lambda \le 1$$

– also called “Jelinek-Mercer smoothing”

Dirichlet smoothing

- Has a relatively big effect on small documents, but a relatively small effect on big documents.

$$\sum_t tf(t,d)+\mu$$

$$P(T_i=t_i|D=d)=\frac{tf(t_i,d)+\mu P(T_i|C)}{¿}¿$$

(Zhai & Lafferty 2004)

Cross-language IR

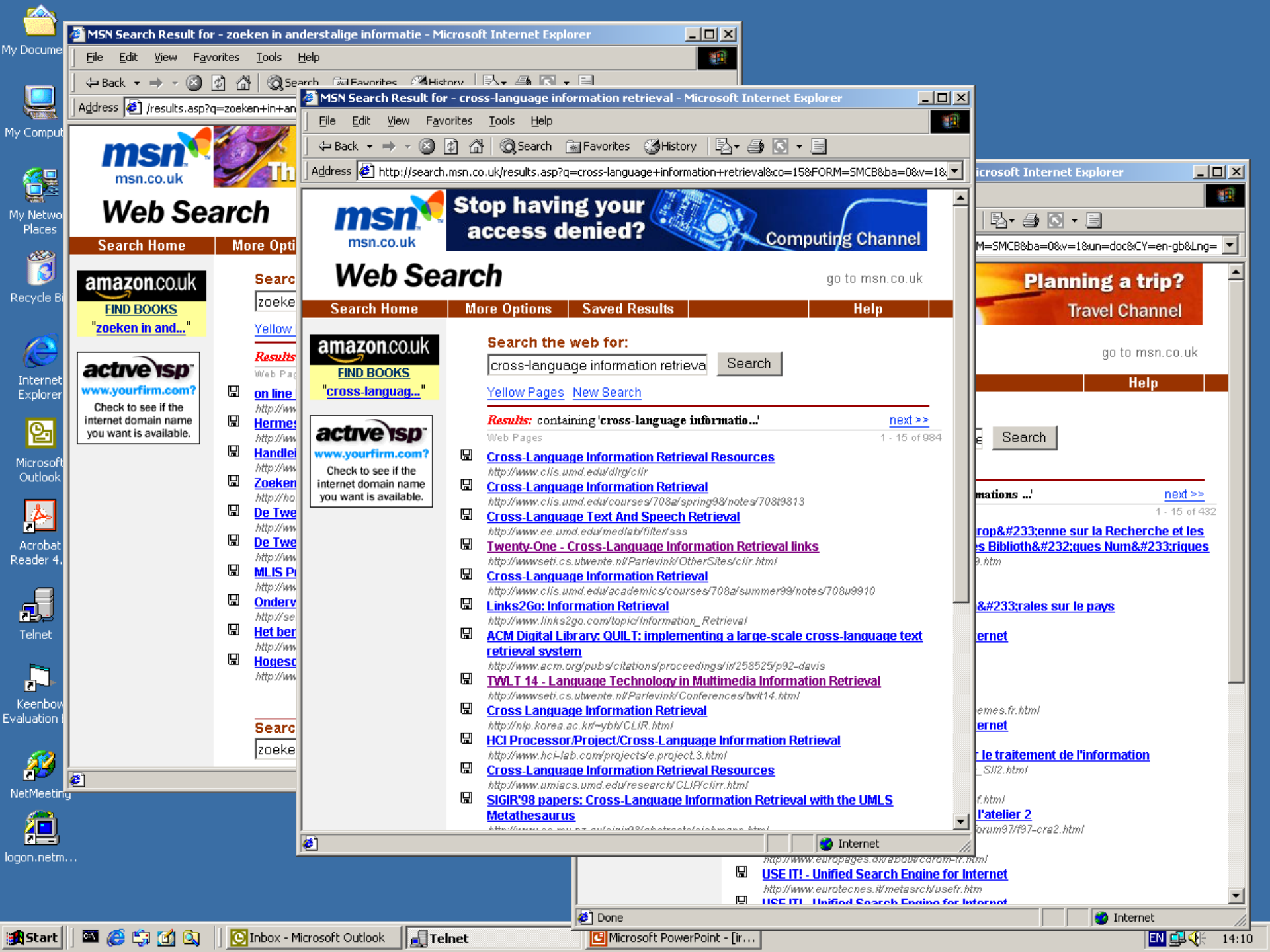
cross-language information retrieval

zoeken in anderstalige informatie

recherche d'informations multilingues

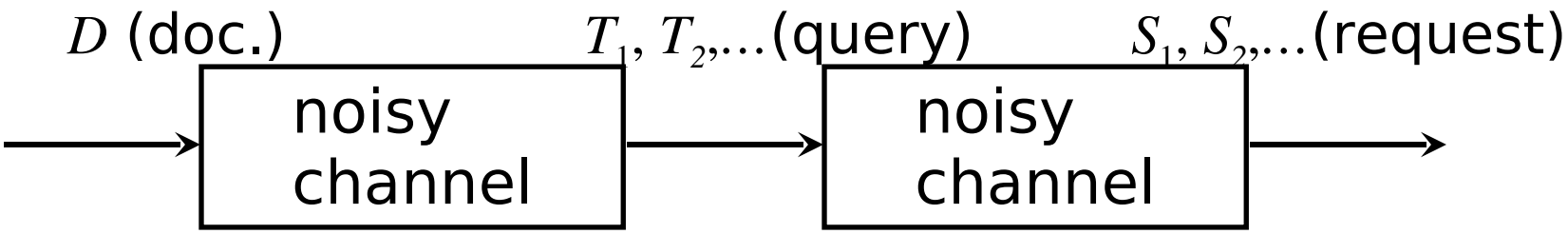
Language models & translation

- Cross-language information retrieval (CLIR):
  - Enter query in one language (language of choice) and retrieve documents in one or more other languages.
  - The system takes care of automatic translation


MSN Search Result for - zoeken in anderstalige informatie - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address /results.asp?q=zoeken+in+an
msn.co.uk
Web Search
Search Home
More Opti
amazon.co.uk
FIND BOOKS
"zoeken in and..."
actve isp
www.yourfirm.com?
Check to see if the internet domain name you want is available.
MSN Search Result for - cross-language information retrieval - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://search.msn.co.uk/results.asp?q=cross-language+information+retrieval&co=15&FORM=SMCB&ba=0&v=1&
msn.co.uk
Stop having your access denied?
Computing Channel
Web Search
go to msn.co.uk
Search Home
More Options
Saved Results
Help
amazon.co.uk
FIND BOOKS
"cross-languag..."
actve isp
www.yourfirm.com?
Check to see if the internet domain name you want is available.
Search the web for:
cross-language information retrieva
Search
Yellow Pages
New Search
Results: containing 'cross-language informatio...'
next >>
Web Pages
1 - 15 of 984
Cross-Language Information Retrieval Resources
http://www.clis.umd.edu/dirg/clir
Cross-Language Information Retrieval
http://www.clis.umd.edu/courses/708a/spring98/notes/708t9813
Cross-Language Text And Speech Retrieval
http://www.ee.umd.edu/medlab/filter/sss
Twenty-One - Cross-Language Information Retrieval links
http://wwwseti.cs.utwente.nl/Parlevink/OtherSites/clir.html
Cross-Language Information Retrieval
http://www.clis.umd.edu/academics/courses/708a/summer99/notes/708u9910
Links2Go: Information Retrieval
http://www.links2go.com/topic/Information_Retrieval
ACM Digital Library: QUILT: implementing a large-scale cross-language text retrieval system
http://www.acm.org/pubs/citations/proceedings/ir/258525/p92-davis
TWLT 14 - Language Technology in Multimedia Information Retrieval
http://wwwseti.cs.utwente.nl/Parlevink/Conferences/twlt14.html
Cross Language Information Retrieval
http://nlp.korea.ac.kr/~ybh/CLIR.html
HCI Processor/Project/Cross-Language Information Retrieval
http://www.hci-lab.com/projects/e.project.3.html
Cross-Language Information Retrieval Resources
http://www.umiacs.umd.edu/research/CLIR/clirr.html
SIGIR'98 papers: Cross-Language Information Retrieval with the UMLS Metathesaurus
Internet
Planning a trip?
Travel Channel
go to msn.co.uk
Help
Search
next >>
1 - 15 of 432
USE IT! - Unified Search Engine for Internet
http://www.eurotecnes.it/metasrch/usefr.htm
Done
Internet
Start
Inbox - Microsoft Outlook
Telnet
Microsoft PowerPoint - [ir...
14:10

Language models & translation

- Noisy channel paradigm



- hypothesise all possible documents D and take the one with the highest probability:

$$\hat{D} = \underset{D}{\operatorname{argmax}}\, P(D \mid S_1, S_2, \cdots)$$
$$= \underset{D}{\operatorname{argmax}}\, P(D) \cdot \sum_{T_1, T_2, \cdots} P(T_1, T_2, \cdots; S_1, S_2, \cdots \mid D)$$

Language models & translation

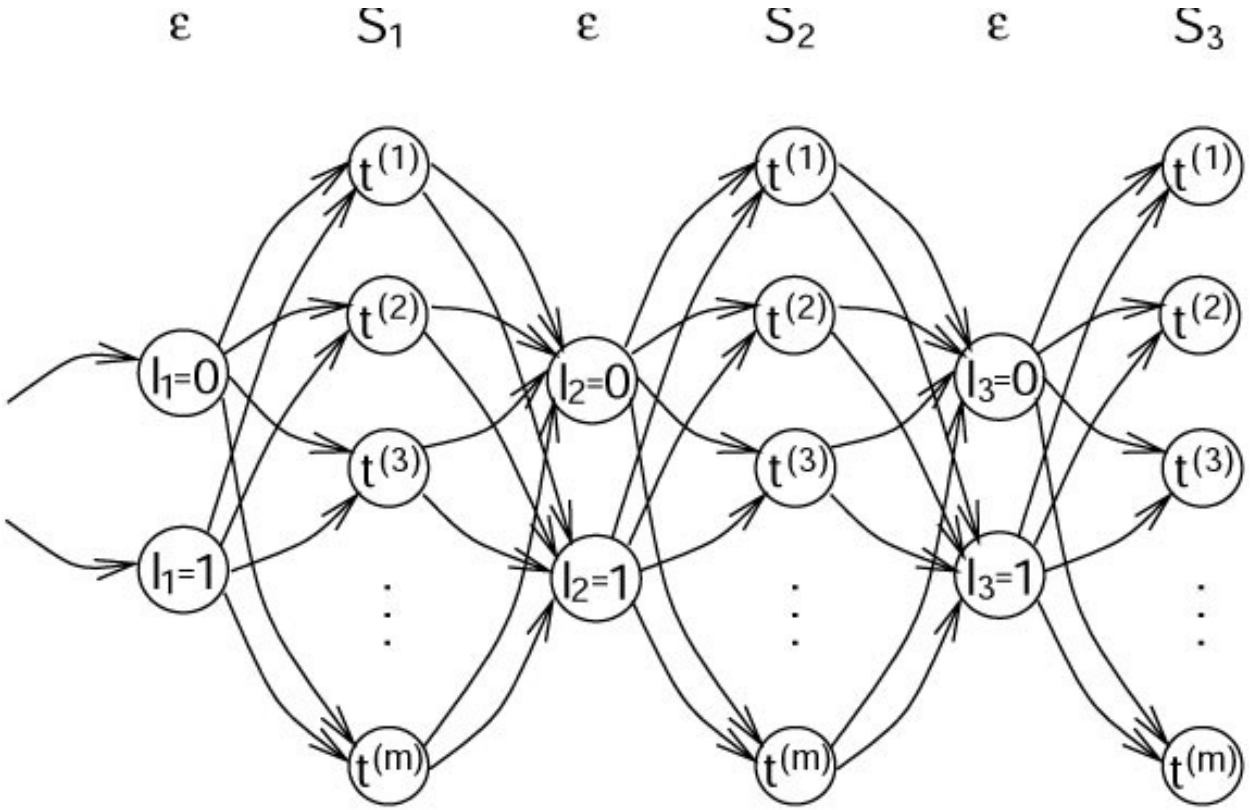
- Cross-language information retrieval :
  - Assume that the translation of a word/term does not depend on the document in which it occurs.
  - if: $S_1, S_2, \ldots, S_n$ is a Dutch query of length n
  - and $t_{i1}, t_{i2}, \ldots, t_{im}$ are m English translations of the Dutch query term S_i

$$P(S_1, S_2, \ldots, S_n | D) =$$

$$\prod_{i=1}^{n} \sum_{j=1}^{m_i} P(S_i | T_i = t_{ij}) ((1-\lambda) P(T_i = t_{ij}) + \lambda P(T_i = t_{ij} | D))$$

Language models & translation

- Presentation as hidden Markov model



Language models & translation

- How does it work in practice?
  - Find for each Russian query term N_i the possible translations $t_{i1}, t_{i2}, \ldots, t_{im}$ and translation probabilities
  - Combine them in a structured query
  - Process structured query

Language models & translation

- Example:
  - Russian query: *ОСТОРОЖНО РАДИОАКТИВНЫЕ ОТХОДЫ*
  - Translations of *ОСТОРОЖНО* : *dangerous* (0.8) or *hazardous* (1.0)
  - Translations of *РАДИОАКТИВНЫЕ ОТХОДЫ* : radioactivity (0.3) or *radioactive chemicals* (0.3) or radioactive waste*t* (0.1)
  - Structured query:

    $((0.8\ dangerous \cup 1.0\ hazardous)\ ,$

Structured query

– Structured query:

$$((0.8\ dangerous \cup 1.0\ hazardous)\ ,$$
$$(0.3\ fabric \cup 0.3\ chemicals \cup 0.1\ dust))$$

Language models & translation

- Other applications using the translation model
  - On-line stemming
  - Synonym expansion
  - Spelling correction
  - ‘fuzzy’ matching
  - Extended (ranked) Boolean retrieval

Language models & translation

- Note that:
  - $\lambda_i = 1$, for all $0 \leq i \leq n$: Boolean retrieval
  - Stemming and on-line morphological generation give exact same results:

$$P(\text{funny} \cup \text{funnies}, \text{table} \cup \text{tables} \cup \text{tabled}) =$$
$$P(\texttt{funni}, \texttt{tabl})$$

Experimental Results

- translation language model
  - (source: parallel corpora)
  - average precision: 0.335 (83 % of base line)
- no translation model, using all translations:
  - average precision: 0.308 (76 % of base line)
- manual disambiguated run (take best translation)
  - average precision: 0.315 (78 % of base line)

(Hiemstra and De Jong 1999)

University of Twente
The Netherlands

Prior probabilities

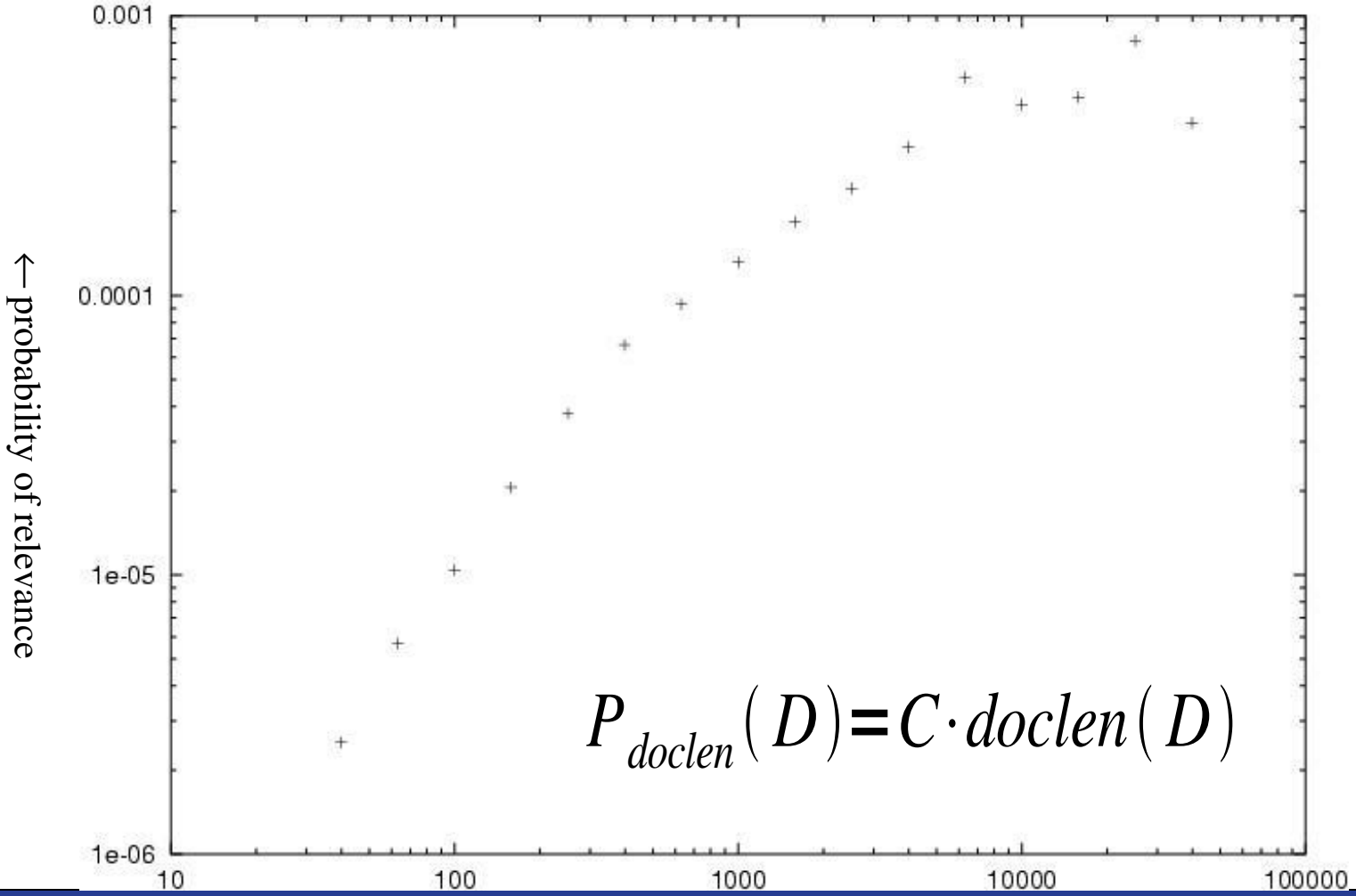
Prior probabilities and static ranking

- Noisy channel paradigm (Shannon 1948)

- hypothesise all possible documents *D* and take the one with the highest probability, symbolically:

$$\hat{D} = \underset{D}{\operatorname{argmax}}\, P(D \mid T_1, T_2, \cdots)$$
$$= \underset{D}{\operatorname{argmax}}\, P(D) \cdot P(T_1, T_2, \cdots \mid D)$$

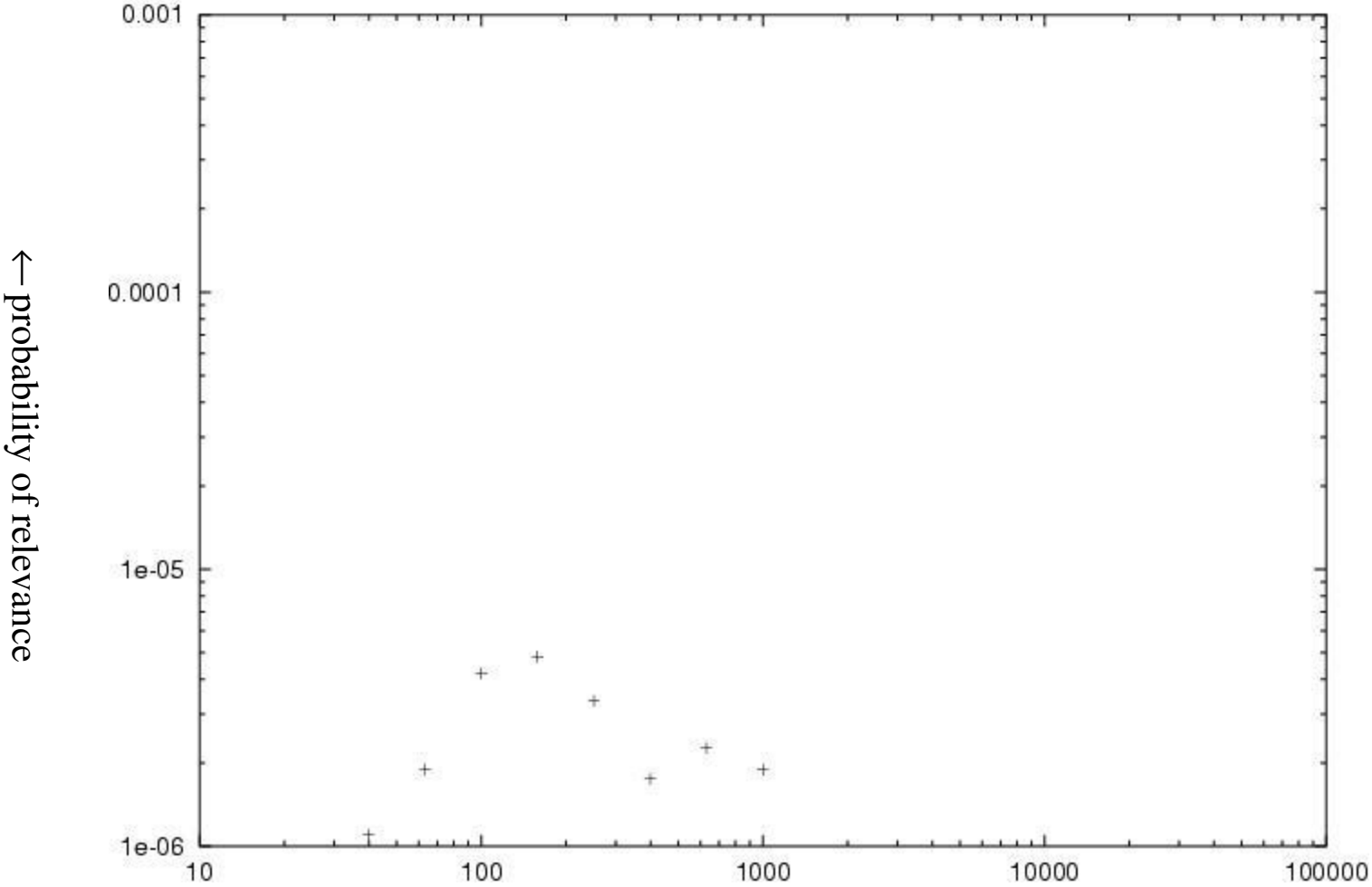
Prior probability of relevance on informational queries



Priors in Entry Page Search

- Sources of Information
  - Document length
  - Number of links pointing to a document
  - The depth of the URL
  - Occurrence of cue words (‘welcome’,’home’)
  - number of links in a document
  - page traffic

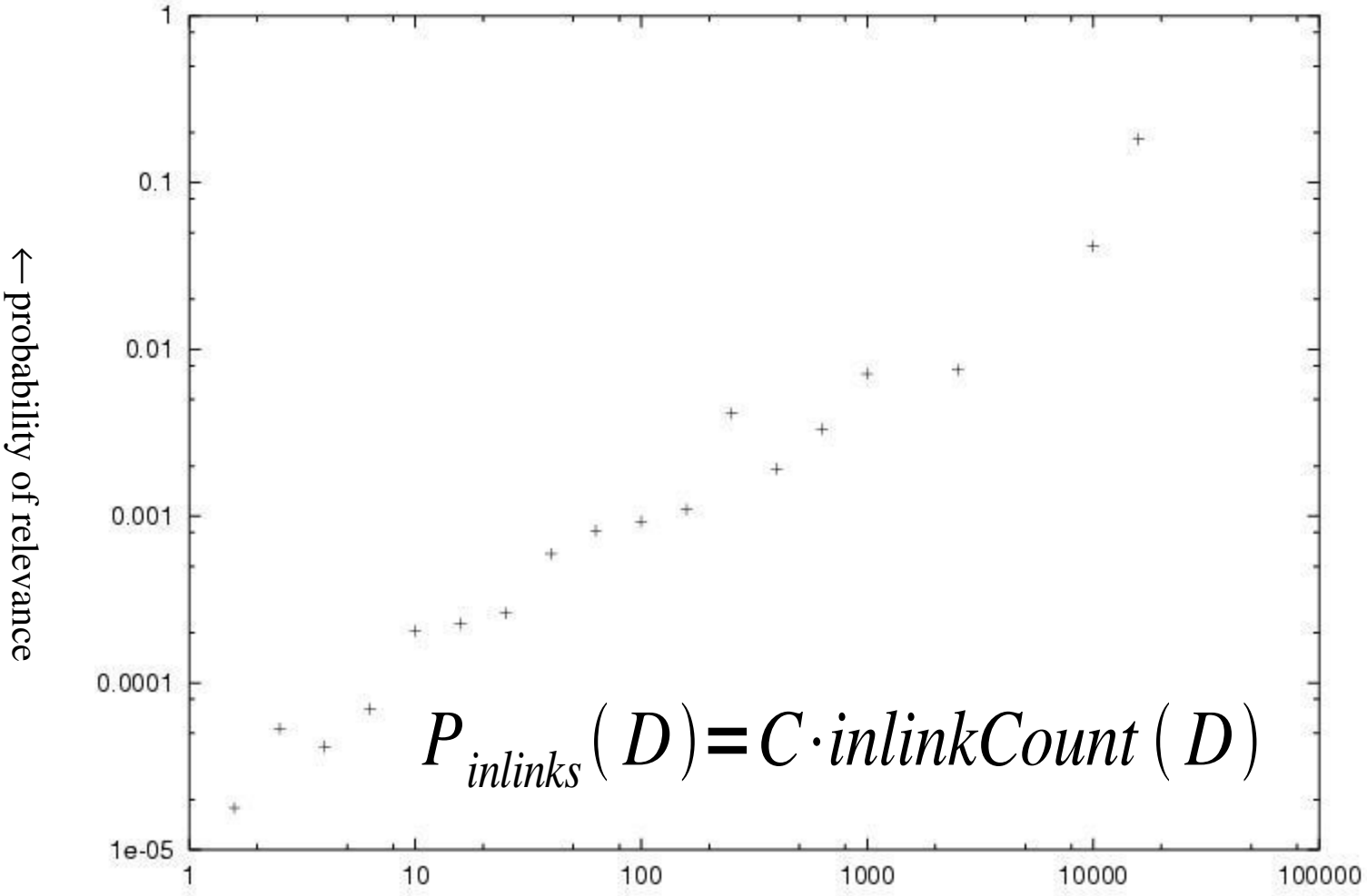
Prior probability of relevance on navigational queries



Priors in Entry Page Search

- Assumption
  - Entry pages referenced more often
- Different types of inlinks
  - From other hosts (recommendation)
  - From same host (navigational)
- Both types point often to entry pages

Priors in Entry Page Search



Priors in Entry Page Search: URL depth

- Top level documents are often entry pages
- Four types of URLs
  - root: **`www.romip.ru/`**
  - subroot: **`www.romip.ru/russir2009/`**
  - path: **`www.romip.ru/russir2009/en/`**
  - file: **`www.romip.ru/russir2009/en/venue.html`**

Priors in Entry Page Search: results

method	Content	Anchors
$P(Q\|D)$	0.3375	0.4188
$P(Q\|D)P_{doclen}(D)$	0.2634	0.5600
$P(Q\|D)P_{inlink}(D)$	0.4974	0.5365
$P(Q\|D)P_{URL}(D)$	0.7705	0.6301

(Kraaij, Westerveld and Hiemstra 2002)

Language Models conclusion

- Smoothing: accounts for sparse documents, and bad queries
- Translation model: accounts for multiple query representations (e.g. CLIR or stemming)
- Document priors: account for "non-content" information

References

- Djoerd Hiemstra and Franciska de Jong. Disambiguation strategies for cross-language information retrieval., *Lecture Notes in Computer Science 1696: Research and Advanced Technology for Digital Libraries, Springer-Verlag*, 1999
- Wessel Kraaij, Thijs Westerveld and Djoerd Hiemstra. The Importance of Prior Probabilities for Entry Page Search. In *Proceedings of SIGIR 2002*
- Claude Shannon. A mathematical theory of communication. *Bell System Technical Journal 27*, 1948
- ChengXiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. ACM Transactions on Information Systems 22(2), pages 179-214, 2004.