Information Retrieval Modeling
Russian Summer School in Information Retrieval

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

1. Smoothing methods
2. Translation models
3. Document priors
4. …
Course Material

Noisy channel paradigm (Shannon 1948)

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

$$\hat{I} = \arg\max_I P(I | O)$$

$$= \arg\max_I P(I) \cdot P(O | I)$$
Noisy channel paradigm (Shannon 1948)

\[ \hat{D} = \arg \max_D P(D | T_1, T_2, \ldots) \]

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

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

$\lambda_i$: 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:

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

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

\[ \lambda_i = 0.5 \]
Statistical language models

• How to estimate value of $\lambda_i$?
  – 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.
    $\lim \lambda_i \to 1$: 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

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

\[ P(T_1, T_2, \ldots, T_n|D) \propto \sum_{i=1}^{n} \log \left( 1 + \frac{\lambda_i P(T_i|D)}{(1-\lambda_i)P(T_i)} \right) \]
Statistical language models

- Implementation as vector product:

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

\[
q_k = \text{tf}(k, q)
\]

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

\[ P(T_i = t_i \mid D = d) = \frac{tf(t_i, d)}{\sum_t tf(t, d)} \]

– Documents that do not contain all terms get zero probability (are not retrieved)
Laplace smoothing

• Simply add 1 to every possible event

\[
P(T_i = t_i | D = d) = \frac{tf(t_i, d) + 1}{\sum_t (tf(t, d) + 1)}
\]

– over-estimates probabilities of unseen events
Linear interpolation smoothing

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

\[ P(T_i \mid D) = (1 - \lambda) P(T_i) + \lambda P(T_i \mid D), \text{ where } 0 \leq \lambda \leq 1 \]

– also called “Jelinek-Mercer smoothing”
Dirichlet smoothing

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

\[ \sum_t tf(t, d) + \mu \]

\[
P(T_i = t_i \mid D = d) = \frac{tf(t_i, d) + \mu P(T_i \mid C)}{\hat{c}}
\]

(Zhai & Lafferty 2004)
Cross-language IR

cross-language information retrieval
zoek 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
Search the web for:
cross-language information retrieval

Results containing cross-language information retrieval:

- Cross Language Information Retrieval Resources
- Cross Language Information Retrieval
  - http://www.cis.udel.edu/course/7793sp99/Notes/7009013
- Cross Language Text And Speech Retrieval
  - http://www.cs.uml.edu/other_DOC/slb34.html
- Twenty One - Cross Language Information Retrieval links
  - http://www.txtb.ca/other/other_Co2nd34.html
- Cross Language Information Retrieval
  - http://www.cis.udel.edu/other_DOC/slb34.html
- Links2Go: Information Retrieval
  - http://www.links2go.com/topic/information_retrieval
- ACM Digital Library: CBRT: Implementing a large-scale cross-language text retrieval system
- TMLT 14: Language Technology in Multimedia Information Retrieval
- Cross Language Information Retrieval
  - http://nlp.korea.ac.kr/lbrj/bclicr.html
- HCI Processor4Project Cross Language Information Retrieval
  - http://www.houra.montpellier.univ.fr/ProjetProjet.html
- Cross Language Information Retrieval Resources
  - http://www.unices.snd.edu/research/CURL/other.html
- SIGIR99 papers: Cross Language Information Retrieval with the UMLS Metathesaurus

More Options | Saved Results | Help
Language models & translation

- Noisy channel paradigm

\[ D \text{ (doc.)} \rightarrow \text{noisy channel} \rightarrow T_1, T_2, \ldots \text{(query)} \rightarrow S_1, S_2, \ldots \text{(request)} \]

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

\[
\hat{D} = \arg\max_D P(D | S_1, S_2, \cdots)
\]

\[
= \arg\max_D P(D) \cdot \sum_{T_1, T_2, \cdots} P(T_1, T_2, \cdots; S_1, S_2, \cdots | D)
\]
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 (0.1)

– Structured query: 

((0.8 dangerous ∪ 1.0 hazardous),
Structured query

- Structured query:

  \(((0.8 \text{ dangerous} \cup 1.0 \text{ hazardous}) ,
  (0.3 \text{ fabric} \cup 0.3 \text{ chemicals} \cup 0.1 \text{ dust}))\)
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:
    
    \[
    P(\text{funny} \cup \text{funnies}, \text{table} \cup \text{tables} \cup \text{tabled}) = P(\text{funni}, \text{tabl})
    \]
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)
Prior probabilities
Prior probabilities and static ranking

- Noisy channel paradigm (Shannon 1948)

\[ D \text{ (document)} \xrightarrow{\text{noisy channel}} T_1, T_2, \ldots \text{ (query)} \]

- Hypothesise all possible documents \( D \) and take the one with the highest probability, symbolically:

\[
\hat{D} = \arg\max_D P(D|T_1, T_2, \ldots) \\
= \arg\max_D P(D) \cdot P(T_1, T_2, \ldots|D)
\]
Prior probability of relevance on informational queries

\[ P_{\text{doclen}}(D) = C \cdot \text{doclen}(D) \]
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

\[ P_{\text{inlinks}}(D) = C \cdot \text{inlinkCount}(D) \]
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/
## Priors in Entry Page Search: results

<table>
<thead>
<tr>
<th>method</th>
<th>Content</th>
<th>Anchors</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(Q</td>
<td>D)$</td>
<td>0.3375</td>
</tr>
<tr>
<td>$P(Q</td>
<td>D)P_{doclen}(D)$</td>
<td>0.2634</td>
</tr>
<tr>
<td>$P(Q</td>
<td>D)P_{inlink}(D)$</td>
<td>0.4974</td>
</tr>
<tr>
<td>$P(Q</td>
<td>D)P_{URL}(D)$</td>
<td>0.7705</td>
</tr>
</tbody>
</table>

(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


• Wessel Kraaij, Thijs Westerveld and Djoerd Hiemstra. The Importance of Prior Probabilities for Entry Page Search. In *Proceedings of SIGIR 2002*
