Statistical Machine Translation for Query Expansion in Answer Retrieval

Stefan Riezler, Alexander Vasserman, Ioannis Tsochantaridis, Vibhu Mittal and Yi Liu
2007

Kätriin Kukk & Zhe Han
Information Retrieval
Uppsala University
Spring 2021
Agenda

- The problem
- Method
- Evaluation
- Results
- Summary
The problem

- The goal in QA is to find relevant answers to questions
- There is a mismatch between question and answer vocabularies
- Techniques for question reformulation
  - Rule-based syntactic and semantic patterns
  - Reformulations based on shared dependency parses
  - Use of WordNet to close the lexical gap word-by-word
  - SMT models for ranking candidate answers
The problem

- In Information Retrieval, techniques for *query expansion*
  - Adding strongly related terms from a thesaurus such as WordNet
  - Global techniques based on corpus statistics and that take the entire query into account
  - Local techniques that use the top ranked documents for expanding the original query
- Query expansion has also been used in QA with similar results
- The task is answer retrieval from FAQ pages using SMT models for query (question) expansion
Method - Data Collection

- A precision-oriented FAQ and question-answer pair extraction
- Filtered 4 billion web pages to extract those that had “faq” or “faqs” in the URL -> 2.6 million web pages
- Manually labelled 1000 -> classifier -> 795 483 FAQ pages
- Question-answer pairs from the FAQ pages were collected using feature functions on punctuation, HTML tags, listing markers and lexical cues
- 10 568 160 QA pairs collected in total
- Manual evaluation of 100 documents showed 98 % precision and 82 % recall for extracting question-answer pairs
Method - SMT-based Query Expansion

- Implementation based on phrase-based SMT framework
- The probability of translating a foreign sentence $f$ into English $e$:
  \[
  \arg\max_e p(e|f) = \arg\max_e p(f|e)p(e)
  \]
- Translation model + language model
- The translation model can be decomposed into ($I$ is a sequence of phrases)
  \[
  p(f^I_e | e^I_i) = \prod_{i=1}^{I} p(f_i | e_i)
  \]
Method - Question-Answer Translation

- Questions & answers as two distinct languages -> SMT training pipeline
- Preprocessing, sentence, chunk & word alignment, and phrase extraction
- The goal: learning associations between question words & synonymous answer words
- Phrase extraction based on the alignments from both translation directions
Method - SMT-based Paraphrasing

- Identify paraphrases or synonyms at the phrase level by pivoting on another language.
- Given an English para-phrase pair (trg, syn), the probability $p(syn|trg)$ that trg translates into syn is defined as the joint probability that the English phrase trg translates into the foreign phrase src, and that the foreign phrase src translates into the English phrase syn:

$$p(syn|trg) = \max_{src} p(src|trg)p(syn|src)$$
$$p(trg|syn) = \max_{src} p(src|syn)p(trg|src)$$

- Query expansion is done by adding terms introduced in n-best paraphrases of the query.
Evaluation

- Baseline: tf-idf retrieval model of Jijkoun and de Rijke (2005)
- Combines vector similarity scores between the query and question-answer pair fields
- Expansion terms taken from n-best translations (50 for paraphrasing, 10 for QA translation) - average 7.8 for paraphrasing & 3.1 for QA translation
- Local expansion - Expansion terms taken from top n answers retrieved by baseline - average 9.25
- Test queries from query logs of MetaCrawler search engine - 60 chosen
- Manual labeling of the top 20 answers for each (0,1,2)
- Evaluation measure: success rate at 10 or 20 answers
## Results

<table>
<thead>
<tr>
<th>Method</th>
<th>$S_2@10$</th>
<th>$S_2@20$</th>
<th>$S_{1,2}@10$</th>
<th>$S_{1,2}@20$</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline <strong>tfidf</strong></td>
<td>27</td>
<td>35</td>
<td>58</td>
<td>65</td>
</tr>
<tr>
<td>local expansion</td>
<td>30 (+11.1)</td>
<td>40 (+14.2)</td>
<td>57 (-1)</td>
<td>63 (-3)</td>
</tr>
<tr>
<td>SMT-based expansion</td>
<td>38 (+40.7)</td>
<td>43 (+22.8)</td>
<td>58</td>
<td>65</td>
</tr>
<tr>
<td>Query</td>
<td>Local Expansion</td>
<td>QA-Translation</td>
<td>Paraphrasing</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-----------------</td>
<td>----------------</td>
<td>--------------</td>
<td></td>
</tr>
<tr>
<td>(1)</td>
<td>how to live with cat allergies</td>
<td>allergens allergic infections filter plasmacluster rhinitis introduction effective replacement</td>
<td>allergy cats pet food</td>
<td>way allergens life allergy feline ways living allergen</td>
</tr>
<tr>
<td>(2)</td>
<td>how to design model rockets</td>
<td>models represented orientation drawings analysis element environment different structure</td>
<td>models rocket</td>
<td>missiles missile rocket grenades arrow designing prototype models ways paradigm</td>
</tr>
<tr>
<td>(3)</td>
<td>what is dna hybridization</td>
<td>instructions individual blueprint characteristics chromosomes deoxyribonucleic information biological genetic molecule</td>
<td>slides clone cdna sitting sequences</td>
<td>hibridization hybrids hybridation anything hibridacion hybridising adn hybridisation nothing</td>
</tr>
<tr>
<td>(4)</td>
<td>how to enhance competitiveness of indian industries</td>
<td>resources production quality processing established investment development facilities institutional increase industry</td>
<td>promote raise improve increase industry strengthen</td>
<td></td>
</tr>
<tr>
<td>(5)</td>
<td>how to induce labour</td>
<td>experience induction practice imagination concentration information consciousness different meditation relaxation</td>
<td>birth industrial induced induces</td>
<td>way workers inducing employment ways labor working child work job action unions</td>
</tr>
</tbody>
</table>
Summary

● The problem: answer retrieval in QA - there is a mismatch between question and answer vocabularies

● The suggested solution: use techniques from statistical machine translation for query (question) expansion in answer retrieval

● The result: higher success rates for finding contained answers compared to the baseline tf-idf system and a local expansion system
Thank you for listening!