Adapting SMT Query Translation Reranker to New Languages in Cross-Lingual Information Retrieval

Shadi Saleh & Pavel Pecina

Institute of Formal and Applied Linguistics
Charles University

\{saleh,pecina\}@ufal.mff.cuni.cz

21 Jul 2016
Outline

• Introduction
• SMT translation reranker
• Adapting reranker to new languages
• Language-specific vs. language-independent model
• Results
• Q&A
CLIR system in the medical domain.

Based on our work: *Shadi Saleh and Pavel Pecina: Reranking Hypotheses of Machine-Translated Queries for Cross-Lingual Information Retrieval*, which will be published in *CLEF 2016*

Machine learning model that predicts the best query translation (from multiple alternative translations) for CLIR.

The model then is adapted to improve the CLIR system in new languages (Spanish, Hungarian, Polish and Hungarian).
SMT translation reranker
Translation System

- Translate queries into English using SMT systems within Khresmoi
- Trained to translate search queries
- Adapted to translate data in the medical domain
- Returns list of alternative translations for each input sentence
- Refer to it as *n-best-list*
Data & IR model

- CLEF eHealth 2015 IR task collection
- For searching, queries from CLEF eHealth IR tasks 2013–2015, 166 queries in total
- Queries were provided in English and translated into Czech, French and German
- Split queries: 100 for training and 66 for testing
- IR experiments with Terrier’s implementation of Dirichlet model.
IR Evaluation

- IR results are evaluated using TREC-EVAL tool.
- P@10 and MAP.
- CVG@10, the percent of assessed documents in the highest 10 retrieved ones.
- Significance test was performed using paired Wilcoxon signed-rank test, $\alpha$ is set to 0.05.
The single best translation that is returned by SMT system is not selected w.r.t CLIR performance.

- Reranker is trained to select the best translation for CLIR performance
- Generalized linear regression model
- Logit as the link function, response in $[0, 1]$
- P@10 as an objective function
Feature set

- SMT features: Translation model, language model and reordering models
- Rank features: SMT rank and a Boolean feature (1 for best rank, 0 otherwise)
- Features based on Blind relevance feedback, IDF from the collection, translation pool and retrieval state value
- Features that are based on external resources (UMLS, Wikipedia)
100 queries for training, 15-best-list hypotheses for each query. Two approaches for training:

- Language-Specific model
- Language-Independent model

Queries with P@10=0 by all of their hypotheses are excluded from the training.
Language-Specific model

- Model for each language: Czech, French and German
- Query hypotheses from each language used separately to train specific models
Language-Independent model

- One Model for all languages
- Query hypotheses from all languages used separately to train one independent model
Figure: Reranking steps
Adapting reranker to new languages
Queries in new languages

- New SMT systems (Spanish, Hungarian, Polish and Swedish) developed recently also within Khresmoi.
- Human experts translated original English queries into these languages, ”under KConnect project”.
- We want to develop CLIR system for these languages.
Existing reranker

- Using the existing reranker did not help to outperform the baseline.
- Language-specific models for existing/new languages did not help.
Adapting reranker

To adapt the reranker, two sources of data used to create training set:
- Merged data from existing languages (Czech, French and German)
- Data from each new language (Spanish, Hungarian, Polish and Swedish)

The data is used to create language-independent models
## Language-specific model performance

<table>
<thead>
<tr>
<th>system</th>
<th>Spanish P@10</th>
<th>Hungarian P@10</th>
<th>Polish P@10</th>
<th>Swedish P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mono</td>
<td>50.30</td>
<td>50.30</td>
<td>50.30</td>
<td>47.10</td>
</tr>
<tr>
<td>Baseline</td>
<td>44.09</td>
<td>40.76</td>
<td>36.82</td>
<td>36.67</td>
</tr>
<tr>
<td>SMT</td>
<td>43.18</td>
<td>42.58</td>
<td>36.06</td>
<td>37.12</td>
</tr>
<tr>
<td>+Rank</td>
<td>42.88</td>
<td>40.76</td>
<td>38.33</td>
<td>36.52</td>
</tr>
<tr>
<td>ALL</td>
<td>43.33</td>
<td>40.00</td>
<td>37.73</td>
<td>36.21</td>
</tr>
</tbody>
</table>
## Table: Final evaluation results of language-independent models on the test set

<table>
<thead>
<tr>
<th>system</th>
<th>Spanish P@10</th>
<th>Hungarian P@10</th>
<th>Polish P@10</th>
<th>Swedish P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mono</td>
<td>50.30</td>
<td>50.30</td>
<td>50.30</td>
<td>47.10</td>
</tr>
<tr>
<td>Baseline</td>
<td>44.09</td>
<td>40.76</td>
<td>36.82</td>
<td>36.67</td>
</tr>
<tr>
<td>SMT</td>
<td>43.79</td>
<td>40.00</td>
<td>35.61</td>
<td>38.33</td>
</tr>
<tr>
<td>+Rank</td>
<td>43.64</td>
<td>38.94</td>
<td><strong>38.18</strong></td>
<td>36.21</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td><strong>46.36</strong></td>
<td><strong>43.18</strong></td>
<td>36.67</td>
<td><strong>38.79</strong></td>
</tr>
</tbody>
</table>
Conclusion

- Existing data used to build SMT translation reranker for Czech, French and German.
- These languages are fully assessed.
- Queries in new languages: Spanish, Hungarian, Polish and Swedish.
- Retrieval systems for new languages are not fully assessed.
- Existing data used to adapt reranker for new languages.
- Significant improvement over the baselines in Spanish and Hungarian.
- High number of OOVs in Polish and Swedish might be the reason for low reranker performance.
- The effect of assessment level on reranker performance needs further investigation.
Thanks!

Q&A