Cross-lingual Information Retrieval
German Query Translation and Expansion for Retrieving English Documents

Julia Suter

Universität Heidelberg
suter@cl.uni-heidelberg.de

Information Retrieval Seminar 2016
PD Dr. Karin Haenelt

February 1, 2016
Outline

1. Introduction to CLIR

2. CLIR methods
   - Approaches
   - Challenges
   - Query Expansion

3. My Implementation
   - VSM and Data
   - Resources
   - Methods
   - Results
   - Discussion

4. Conclusion

5. Demo

6. References
Introduction to CLIR
What is CLIR?

Cross-lingual Information Retrieval
Selecting and ranking documents in a language different from the query language

- also Multi-lingual Information Retrieval (MLIR)
- Keywords: (Multilingual) Web & Machine Translation
IR is a representation problem

- even more evident in cross-language information retrieval (CLIR)
- many ways to express and query same information

Example

Wasser kocht bei 100 Grad.
H₂O beginnt bei einer Temperatur von 373.15 Kelvin zu sieden.
Water boils at 100 degrees Celsius.
L’eau bout à 100°C.
L’acqua bolle a cento gradi.
水は100度で沸騰する。
Why CLIR?

When do we need CLIR...? [Nie, 2010]

- relevant data might not exist in original language
- documents in most languages are accessible thanks to MT
- multimedia information retrieval (pictures, video & audio data)
- documents might contain mix of several languages
- need to retrieve all relevant documents in any language (patent retrieval)
- user might understand target language but does not feel comfortable in formulating queries
- user wants to retrieve documents in several languages (no re-typing of query)
CLIR methods
Three ways to compare query and document representations:

- **document translation approach**
  Mapping the document representation into the query representation space
- **query translation approach**
  Mapping the query representation into the document representation space
- **interlingua / pivot language**
  Mapping both document and query representations to a third space

**Translation**

- Dictionary-based Translation
- Machine Translation
Challenges [Nasharuddin and Abdullah, 2011]

- ambiguity
- inflection
- phrases (*pommes de terre, alles unter einen Hut bringen*)
- compounds (*Australienlangstreckendirektflugstopoverspezialisten*)
- named entities
- technical terms
- unknown words (OOV)

- wrong translations introduce noise
- domain-dependency
Query Expansion in CLIR [Nie, 2010]

Query Expansion

reformulating and extending original query to improve retrieval performance

- initial query of the user is not always the best description of the intended information need
- Query expansion can help finding (more) relevant documents

Methods:
- adding synonyms
- adding topic-related words
- stemming of query
- fixing spelling errors
- ...

Julia Suter
Cross-lingual Information Retrieval
February 1, 2016
Query Expansion in CLIR

Pre-translation Query Expansion
- find documents in source language using original query
- extract relevant words from these documents for Query Expansion
- translate expanded query to find relevant target language documents

Post-translation Query Expansion
- translate query to find relevant target language documents
- extract relevant words from these documents for Query Expansion
- search documents again with expanded query
- Pseudo-relevance Feedback
My Implementation
VSM and Data

Vector Space Model
- good old VSM
- normalized words, tf-idf
- cosine similarity

Mini news corpus
- 50 English news articles
- manually extracted from Google News using 10 keywords
- 5 documents for 10 topics
- title not included in document
- original plan: creating comparable corpus with German news articles...

<table>
<thead>
<tr>
<th>Terrorism</th>
<th>Health</th>
<th>Refugee Crisis</th>
<th>Syria</th>
<th>WEF</th>
<th>US Elections</th>
<th>Australian Open</th>
<th>Blizzard</th>
<th>Earthquake</th>
<th>Pluto Fly-by</th>
</tr>
</thead>
</table>

Julia Suter
Cross-lingual Information Retrieval
February 1, 2016
## Resources

### German
- Lemmatizing and POS-tagging: ParZu
- Compound Splitting: Gertwol morphology tool
- Stemming: Snowball Stemmer

### English
- Stemming: Snowball Stemmer

### Translation
- Query and Doc translation: Google Translate (via TextBlob)
- NE translation: Wikipedia langlinks

### Query Expansion
- Definitions: WordNet
- German reference documents: German Wikipedia
Query Translation and Expansion

Query Translation
- Google Translate
- Wikipedia Langlink Translations
- Fuzzy Matching

Query Expansion
- Pre-trans QE
  - Compound Splitting
  - Relevant words from Wikipedia articles
- Post-trans QE
  - WordNet definition words
  - Pseudo-relevance Feedback
- Cross-lingual Wikipedia word overlap QE
Lemmatization and POS-tagging

ParZu dependency parser [Sennrich, 2009]

Europas Probleme
(Europa, NE)
(Problem, NN)

- use lemmatized query for Wikipedia search
- only if noun (N)
- use unlemmatized query when translating
Compound Splitting

- Compound splitting with Gertwol Morphology tool
  [Haapalainen and Majorin, 1995]
- Splitting brings ambiguity:

  Staubecken  Staub\|ecken or Stau\|becken
  Abteilungen  Abteilungen or Abteil|lungen

- Algorithm by [Volk, 1999] for selecting best compound
- Use German compound parts and translated compound parts for Query Expansion
Fuzzy Matching

- Idea: treat German query as misspelled English query [Buckley, 2000]
- slight variations are still considered a match
- If vocabulary is similar, stemming can be enough to match query and document words

```
information (DE, EN, FR)
informazione (IT)
informacion (ES)
→ inform-
```

- Watch out for False Friends!

Extension of Fuzzy Matching:

- add transformation rules [Pirkola, 2003]
- e.g. replace every k with c

```
Krokodil → crocodile
Konferenz → conference
Kollege → college
```
First name?
Technical terms?
Transliteration for languages with other alphabets?
Wikipedia Langlink Translation

Barack Obama

From Wikipedia, the free encyclopedia
(Redirected from Obama)

- search for German Wikipedia article with lemmatized nouns from query
- extract title using redirects
  (Obama → Barack Obama)
- get English article title via Wikipedia Langlinks
- use English title as translation (Query Expansion)

and since we already found the Wikipedia articles...
Wikipedia for Pre-translational Query Expansion

- use the Wikipedia title found in last step
- extract introduction of Wikipedia article
- extract whole Wikipedia articles as reference texts (for tf-idf)
- find relevant words in Wikipedia introduction for Query Expansion

Cross-lingual Wikipedia word overlap

- Assumption: if word occurs in both Wiki articles, it must be relevant
- extract introductions from both DE and EN Wikipedia articles
- translate German text
- find overlapping words and compute tf-idf to extract relevant words for QE
Use translated lemmatized query to extract definitions from WordNet
compute tf-idf (English articles as reference) and extract relevant words for Query Expansion

Example for Market
- the world of **commercial** activity where goods and services are bought and **sold**
- the **customers** for a particular product or service
- a marketplace where groceries are **sold**
- the securities markets in the aggregate
- an area in a town where a public mercantile establishment is set up
Use (expanded) query to retrieve relevant documents
Assume that $k$ top documents are relevant
Extract relevant words for QE
Re-search with new expanded query
Double the weights for original expanded query
Comparing baseline to expanded query system

Baseline: results for Google translation of query

How many of the 5 highest ranked documents are relevant?

Results for 30 queries on two systems
- Translation of news article retrieving query
- Related word query
- Multi-word query

Problem: Some topics are overlapping so no automatic evaluation possible
### Results for topic queries

<table>
<thead>
<tr>
<th>Query</th>
<th>Baseline</th>
<th>Query Expansion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terroranschläge</td>
<td>5/5</td>
<td>4/5</td>
</tr>
<tr>
<td>Gesundheit</td>
<td>4/5</td>
<td>5/5</td>
</tr>
<tr>
<td>Flüchtlingskrise</td>
<td>4/5</td>
<td>5/5</td>
</tr>
<tr>
<td>Syrien</td>
<td>5/5</td>
<td>5/5</td>
</tr>
<tr>
<td>Weltwirtschaftsforum</td>
<td>5/5</td>
<td>5/5</td>
</tr>
<tr>
<td>Präsidentschaftswahlen</td>
<td>5/5</td>
<td>5/5</td>
</tr>
<tr>
<td>Schneesturm</td>
<td>5/5</td>
<td>5/5</td>
</tr>
<tr>
<td>Tennis</td>
<td>5/5</td>
<td>5/5</td>
</tr>
<tr>
<td>Erdbeben</td>
<td>5/5</td>
<td>5/5</td>
</tr>
<tr>
<td>Vorbeiflug</td>
<td>5/5</td>
<td>5/5</td>
</tr>
</tbody>
</table>
## Results for related word queries

<table>
<thead>
<tr>
<th>Query</th>
<th>Baseline</th>
<th>Query Expansion</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>[no match]</td>
<td>4/5</td>
</tr>
<tr>
<td>Krankenkasse</td>
<td>4/5</td>
<td>4/5</td>
</tr>
<tr>
<td>Einwanderung</td>
<td>3/5</td>
<td>3/5</td>
</tr>
<tr>
<td>Assad</td>
<td>4/5</td>
<td>5/5</td>
</tr>
<tr>
<td>Konferenz</td>
<td>2/5</td>
<td>5/5</td>
</tr>
<tr>
<td>Wahlkampf</td>
<td>[no match]</td>
<td>5/5</td>
</tr>
<tr>
<td>Wetter</td>
<td>4/5</td>
<td>4/5</td>
</tr>
<tr>
<td>Djokovic</td>
<td>2/5</td>
<td>4/5</td>
</tr>
<tr>
<td>Richterskala</td>
<td>1/5</td>
<td>5/5</td>
</tr>
<tr>
<td>Asteroid</td>
<td>2/5</td>
<td>5/5</td>
</tr>
</tbody>
</table>
Results for multi-word queries

<table>
<thead>
<tr>
<th>Query</th>
<th>Baseline</th>
<th>Query Expansion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dschihadistische Angriffe</td>
<td>5/5</td>
<td>2/5 (Post-trans QE)</td>
</tr>
<tr>
<td>Wie bleibt man gesund</td>
<td>1/5</td>
<td>5/5</td>
</tr>
<tr>
<td>Flüchtlingskrisendebatte</td>
<td>4/5</td>
<td>4/5</td>
</tr>
<tr>
<td>Humanitäre Krise</td>
<td>4/5</td>
<td>5/5</td>
</tr>
<tr>
<td>Vierte Industrielle Revolution</td>
<td>2/5</td>
<td>4/5</td>
</tr>
<tr>
<td>USA wählen neuen Präsidenten</td>
<td>5/5</td>
<td>4/5</td>
</tr>
<tr>
<td>Amerika eingeschneit</td>
<td>4/5</td>
<td>4/5</td>
</tr>
<tr>
<td>Wer gewinnt das Achtelfinale</td>
<td>3/5</td>
<td>5/5</td>
</tr>
<tr>
<td>Starkes Beben</td>
<td>3/5</td>
<td>5/5</td>
</tr>
<tr>
<td>Eisfläche in Herzform</td>
<td>2/5</td>
<td>4/5</td>
</tr>
</tbody>
</table>
Expanded query system seems to outperform baseline, especially for less obvious queries.

Some queries in baseline system yield *no matches*.

Possibly too much Query Expansion introduces noise → specific evaluation for various methods and weighting required.

Larger document collection required.

System is independent from German documents could easily be adapted to other languages.
Conclusion
Google Translate works well for query translation
Compound segmentation less important than expected
Wikipedia is a great resource for Query Expansion (langlinks most valuable)
German document collection for tf-idf needed
Fuzzy Matching and WordNet definitions only helpful when other methods for QE failed
Pseudo-relevance Feedback *dangerous* on small corpus with few relevant documents

Many different methods for Query Translation and Expansion
Expanded query system outperforms Google Translate baseline
CLIR is challenging but lotsa fun! :-}
Demo
Example query *Wahlkampf*

original query
[Wahlkampf]

parsed query
[(Wahlkampf, NN)]

Translated query (Google translate)
Electioneering

Compound parts
[wahl, kampf]

Translated compound parts
[choice, fight]
Fuzzy matches:
[wahlcampf, wahl, campf]

Langlinks
[Political campaign]

Pre trans query expansion words:
[citizen, speak, take plac]

QE by wiki comparison:
[campaign, elect, candid]

Definition words:
[persuas, campaign, voter]

Normalized query
[choic, wahlcampf, voter, wahl, campaign, campf, fight, elect, candid, election, citizen, plac, take, speak, persuas, polit]
Post trans QE words
[vote, elector, latino, percent]

Normalized query + post trans QE words
[choic, wahlcampf, voter, wahl, campaign, elector, vote, percent, latino, campf, fight, elect, candid, election, citizen, plac, take, speak, persua, polit]

Final ranking: (5/5)

- US elections 2016: What you need to know
- US Elections 2016: Digital Media, Mobile Ads the Key to Reaching Black, Latino Voters: Study
- Campaign 2016: Understanding the presidential election
- Voices US election 2016: Iowa primary arrives with fringe candidates, but we should let moderation prevail
- Election 2016: The Life of the Party
References
Buckley, Chris, et al.  
Using clustering and superconcepts within SMART: TREC 6.  

Hapalainen, Mariikka and Ari Majorin (1995)  
GERTWOL und morphologische Disambiguierung für das Deutsche.  
*In: Proc. of the 10th Nordic Conference on Computational Linguistics, Helsinki, Finland.*

Nasharuddin, Nurul Amelina, and Muhamad Taufik Abdullah (2011)  
Cross-lingual Information Retrieval - State of the Art.  
*In: Electronic Journal of Computer Science and Information Technology 2.1.*

Nie, Jian-Yun (2010)  
Cross-language information retrieval.  
Fuzzy translation of cross-lingual spelling variants.  
*In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. ACM.*

Sennrich, Rico, et al. (2009)  
A new hybrid dependency parser for German.  

Volk, Martin (1999)  
Choosing the right lemma when analysing German nouns.  
*In: Multilinguale Corpora: Codierung, Strukturierung, Analyse. 11. Jahrestagung.*
Thank you for your attention!