Exploiting Multiple Translation Resources for English-Persian Cross Language Information Retrieval

Hosein Azarbonyad

Azadeh Shakery

Heshaam Faili

September 2013

Electrical and Computer Engineering Faculty - University of Tehran

Outlines

- Introduction
- Query Translation Approaches
- Experimental Results
- Conclusion and Future Work

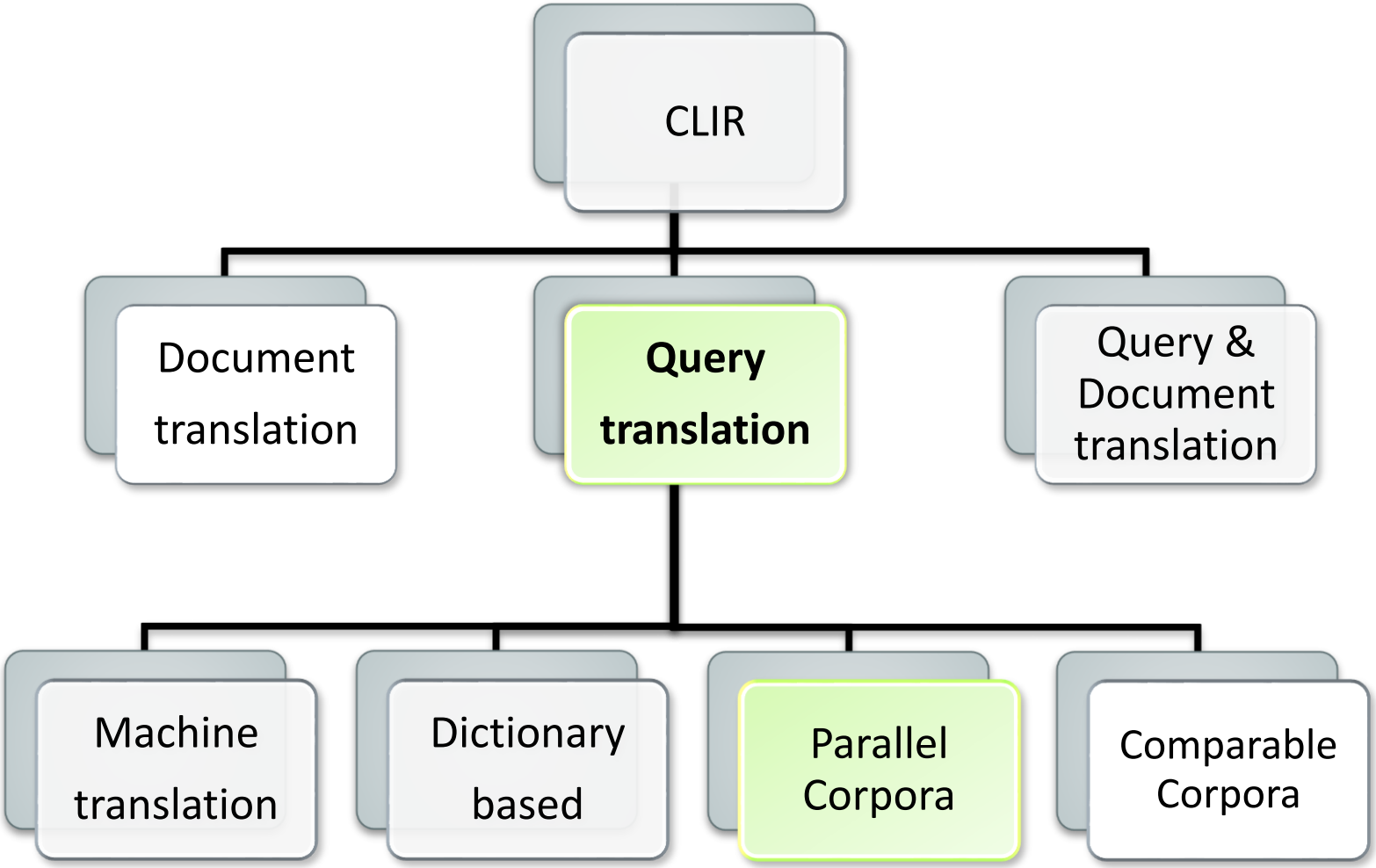
Hosein Azarbonyad h.azarbonyad@ut.ac.ir

Introduction

- CLIR
  - Expressing queries in one language and retrieving documents in another language
- Main problem:
  - Difference between query and documents language
- Solution:
  - Translation

Hosein Azarbonyad h.azarbonyad@ut.ac.ir

Introduction



Hosein Azarbonyad h.azarbonyad@ut.ac.ir

Query Translation Approaches

- The most used method for extracting translation knowledge from parallel corpora
  - IBM model 1
  - Simple and effective
  - This method is used in this research
- Main problem
  - Considering words to be independent

Hosein Azarbonyad h.azarbonyad@ut.ac.ir

Query Translation Approaches

- Main problem of IBM model 1
  - Considering words to be independent
  - This assumption is not realistic

- Example
  - Query: "Anti cancer drugs"
  - "drugs" has two different senses

Hosein Azarbonyad h.azarbonyad@ut.ac.ir

Query Translation Approaches

- Phrase translation beside word translation
  - Phrase indexing
  - Phrase based translation re-ranking

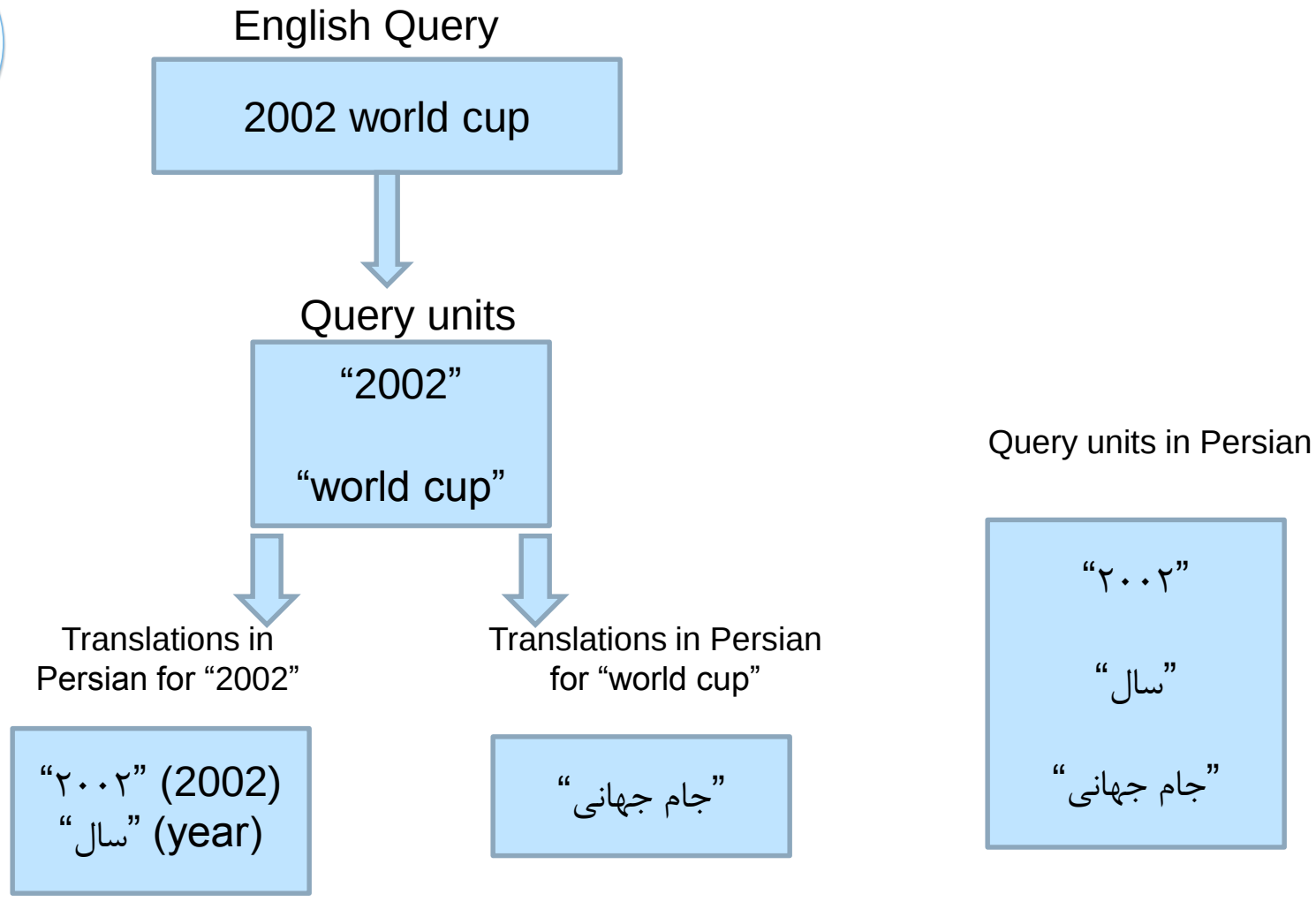
Hosein Azarbonyad h.azarbonyad@ut.ac.ir

Query Translation Approaches

- Phrase indexing
  1. Indexing step
     1. Finding phrases in target language
     2. Considering a phrase as a single unit
     3. Indexing them
  2. Query translation
     1. Considering queries as combinations of phrases and single words
     2. Translation phrases as well as single words
  3. Retrieval
     1. Using BM25 method for calculating similarities of documents and queries

Hosein Azarbonyad h.azarbonyad@ut.ac.ir

Query Translation Approaches



Hosein Azarbonyad h.azarbonyad@ut.ac.ir

Query Translation Approaches

- Phrase based translation re-ranking
  1. Query translation
     1. Finding translation of each query word using the translation knowledge extracted by IBM model 1
  2. Re-weighting and Re-ranking
     1. Finding phrases in query and translating them
     2. Considering phrases as bags of word
     3. Calculating phrasal score of translation candidates

$$S_{ph}(f, Q_e) = \frac{\sum_{\substack{ph \in Q_e, \\ f \in ph}} P(ph|Q_e)}{\sum_{v \in Can} \sum_{\substack{ph \in Q_e, \\ v \in ph}} P(ph|Q_e)}$$

     4. Re-ranking translation candidates based on phrasal scores and translation probabilities

Hosein Azarbonyad h.azarbonyad@ut.ac.ir

Query Translation Approaches

English Query

2002 world cup

Translations of "cup"

Word	Translation probability	Google translation
فنجان	0.46	cup
جام	0.4	cup
لیوان	0.14	glass

Translations of "world cup"

Word	Translation probability	Google translation
جام جهانی	1	World cup

$S_{ph}(\text{جام}) = P(\text{جام جهانی}|world\ cup) = 1$

$S_{ph}(\text{جهانی}) = P(\text{جام جهانی}|world\ cup) = 1$

Re-ranked Translations of "cup"

Word	Translation probability	Google translation
جام	0.6	cup
فنجان	0.31	cup
لیوان	0.09	glass

$S_{ph}(\text{جام}, Q) = \frac{1}{1+1} = 0.5$

Re-weighting: $S(\text{جام}) = 0.4 + 0.5 = 0.9$

Hosein Azarbonyad h.azarbonyad@ut.ac.ir

Query Translation Approaches

- Translation resource combination
  - Weighted linear combination

$$P(f|e) = \lambda * P_{R1}(f|e) + (1 - \lambda) * P_{R2}(f|e)$$

  - R1 and R2 could be either dictionary, parallel corpus or comparable corpus

Hosein Azarbonyad h.azarbonyad@ut.ac.ir

Experimental Results

- Datasets
  - Hamshahri
    - Used in CLEF-2008 and CLEF-2009
    - About 166,000 documents in Persian
    - 100 queries in Persian and English
    - Persian task of CLEF-2008:
      - Retrieving Persian documents from English queries

Hosein Azarbonyad h.azarbonyad@ut.ac.ir

Experimental Results

- Datasets
  - Tehran English Persian (TEP) parallel corpus
    - Constructed from movie subtitles
    - About 4 millions word in English and Persian
  - UTPECC comparable corpus
    - Constructed from news published in Hamshahri and BBC agencies
    - Consists of about 10,000 English documents and 5,000 Persian documents and 15,000 alignments between them
  - Arianpour online English-Persian dictionary
    - http://www.arianpour.com/

Hosein Azarbonyad h.azarbonyad@ut.ac.ir

Experimental Results

- Mean Average Precision (MAP)
- P@5 and P@10
- Comparing with monolingual IR results

Hosein Azarbonyad h.azarbonyad@ut.ac.ir

Experimental Results

- Retrieval method
  - Weighted Okapi BM25 method

$$w(f_i, q_j) = \frac{1}{N} \frac{p(f_i|q_j)}{\sum_{k=1}^{N} p(f_k|q_j)}$$

$$BM25Score(Q, D) = \sum_{q_i \in Q} \sum_{f_j \in Can(q_i)} IDF(f_j) * \frac{(k_1 + 1) * TF(f_j, D)}{k_1((1 - b) + b * (\frac{|D|}{L_{avg}})) + TF(f_j, D)} * w(f_j, q_i)$$

Hosein Azarbonyad h.azarbonyad@ut.ac.ir

Experimental Results

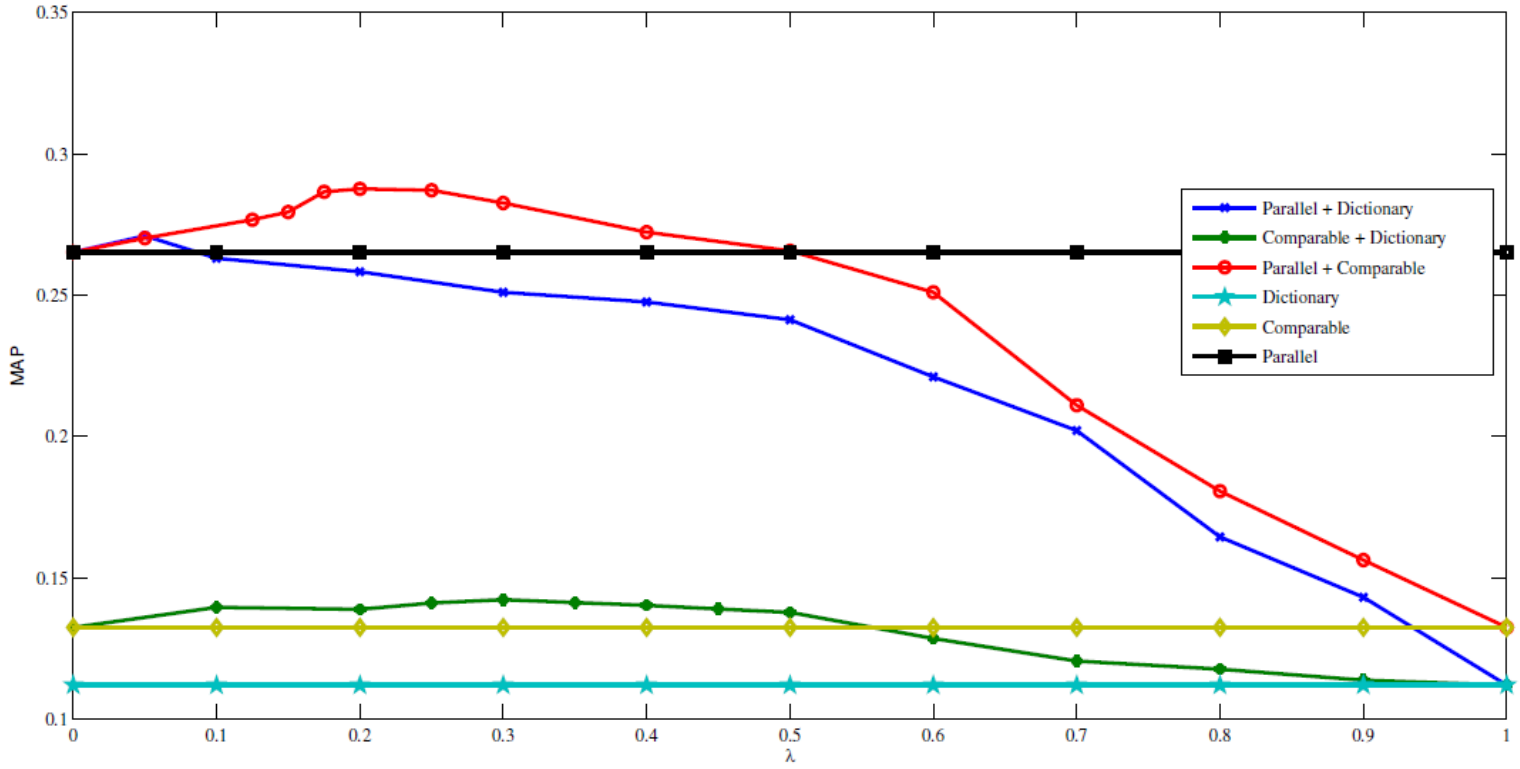
- Results of different methods

Method	MAP	%Mono	P@5	%Mono	P@10	%Mono
Monolingual IR	0.412	-	0.702	-	0.643	-
Dictionary based CLIR	0.138	34	0.216	31	0.204	32
Comparable corpus based CLIR	0.148	36	0.288	40	0.266	41
Parallel corpus based CLIR	0.265	64	0.44	62	0.418	65
Phrase indexing	0.272	66	0.446	64	0.429	67
Phrase based translation re-ranking	0.281	68	0.452	64	0.431	67

Hosein Azarbonyad h.azarbonyad@ut.ac.ir

Experimental Results

- Results of combinational method



Hosein Azarbonyad h.azarbonyad@ut.ac.ir

Conclusion

- Parallel corpus has higher accuracy than other resources in English-Persian CLIR
- Using phrases improves the accuracy of CLIR
- Combining translation resources could improve the accuracy of CLIR

Hosein Azarbonyad h.azarbonyad@ut.ac.ir

Future Work

- Exploiting more and high quality resources for query translation
- Using other contextual information such as correlations of translation candidates in query translation process
- Employing other methods such as Learning to Rank for combining translation resources

Hosein Azarbonyad h.azarbonyad@ut.ac.ir

Thank you