A Query Translation and Expansion Technique for Cross-Language Information Retrieval Using Wikipedia as Linguistic Resources

Youngjoong Ko

Dept. of Computer Engineering
Dong-A University
(http://web.donga.ac.kr/yjko)

Contents

Introduction (1/2)

Cross-language Information Retrieval (CLIR)

- Traditional IR identifies relevant documents in the same language as the query
- CLIR tries to identify relevant documents in a language different from that of the query

Importance of CLIR

- CLIR research is becoming more and more important for global information exchange and knowledge sharing.

Introduction (2/2)

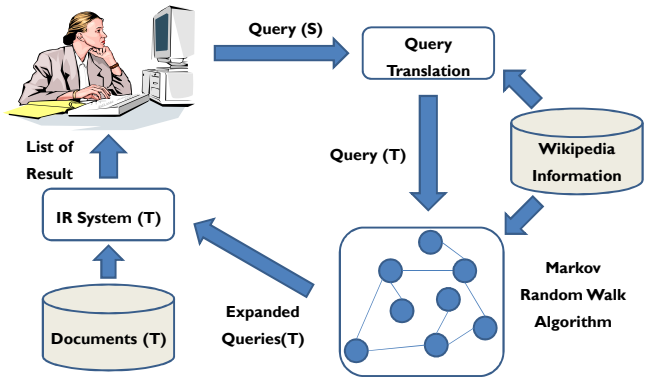
Translation resources are important factors to obtain an effective CLIR system

- Bilingual word lists,
- Parallel and Comparable corpora,
- Machine translation systems
- However, fully developed machine translation systems and even topic-appropriate parallel corpora will sometimes not be available.

Our goal

- Utilize Wikipedia as a bilingual resource for query translation and expansion
- Three lexicons are extracted from Wikipedia for query translation
  - Bilingual pair lexicon, Synonym lexicon and Polysemy lexicon
- A concept link graph is constructed from Wikipedia for query expansion
- Build up a high performance CLIR system without a parallel corpus or with a week parallel corpus

Overview



4 System Developing Stages

1. The Knowledge set from Wikipedia for CLIR → 2. Query Translation for CLIR → 3. Query Expansion for CLIR → 4. Experiment & Summary

1. THE KNOWLEDGE SET FROM WIKIPEDIA FOR CLIR

WBL (Wikipedia Bilingual Lexicon)

WBL (Wikipedia Bilingual Lexicon) consists of three lexicons extracted from Wikipedia

- Bilingual pair lexicon
- Synonymy lexicon
- Polysemy lexicon

Note that all of these lexicons are automatically extracted from Wikipedia.

Bilingual Pair Lexicon

As of January 2012, there are different editions of Wikipedia in 283 languages

- **e.g. Article of English Wikipedia, "*President of the United States*"**
  - **French : "*Président des États-Unis*"**
  - Korean : "*미국의 대통령*"
  - German : "*Präsident der Vereinigten Staaten*"
- **This information is so-called *Inter-wiki links* included in body of each article**

English	Korean
Andre Agassi	안드레 애거시
Apache Software Foundation	아파치 소프트웨어 재단
President of South Korea	대한민국의 대통령

Synonymy Lexicon

Alternative names that can be used to refer to a Wikipedia concept.

- **e.g. The article, "*U.S.A*" redirects to the article "United States" that contains information about the same sense.**
- **This information is so-called *Redirect pages*.**
- **Source language synonymy lexicon and target language synonymy lexicon are operated independently.**

Concept	Synonym	Concept	Synonym
United States	U.S.A	미국	미합중국
United States	USA	미국	아메리카 합중국
United States	UnitedStates	미국	USA

Polysemy Lexicon

Disambiguation pages in Wikipedia are solely intended to allow users to choose among several Wikipedia concepts for an ambiguous word.

- **e.g. The article, "*Washington*" contains three meanings that can be denoted as "George Washington", "Washington (State)", "Washington, D.C."**
- **This information is so-called *Disambiguation pages*.**

English Concept	English Sense	Korean Sense
Washington	George Washington	조지워싱턴
Washington	Washington (State)	워싱턴 주
Washington	Washington, D.C.	워싱턴 D.C.

Statistical Report for Query Matching

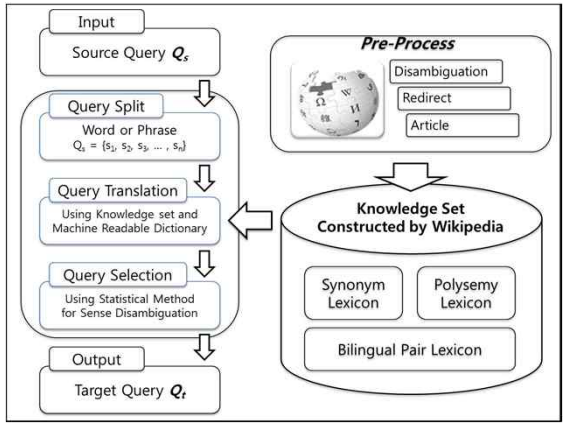
The usefulness of the bilingual word lists (WBL) for query translation is verified on the NTCIR-5 English-Korean CLIR data set.

- **Statistical Report for matching query words in Title**

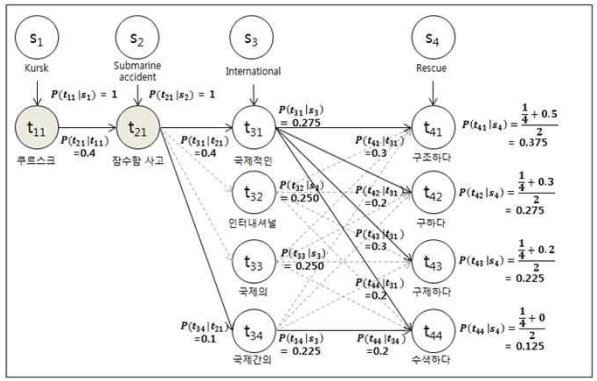
Title	MRD	WBL	CTL (WBL + MRD)
# of topics		50	
# of unique query words		130	
# of translatable query words		240	
# of matched query words	150	163	217
Ratio(%) of matched query words	62.50%	67.92%	90.42%
Total # of equivalents	640	250	559
Avg. of equivalents in each query word	4.26	1.53	2.58

2. QUERY TRANSLATION FOR CLIR

Overview



Example



Translation Candidate Extraction

Step 1. Remove stop-words, extract translatable words, and combine words into a phrase by using the longest first match strategy.

Step 2 . Synonymy information of source words is added by using source synonymy lexicon and it is used in the following two steps.

Step 3. If a source word is ambiguous, it has several translation candidates from polysemy lexicon.

Step 4. The remaining source words are translated by bilingual pair lexicon.

Step 5. Synonymy information of target words is added by using target synonymy lexicon.

Step 6. If a source word was not translated by WBL, the translation method attempts to translate the word by MRD.

Statistic Method for Sense disambiguation(1/5)

The sense disambiguation method assigns the scores to all the candidate target queries for the source query Q_s, and then selects the target query Q_t with the highest score.

$$Q_S = \{s_1, s_2, s_3, \ldots, s_n\}$$

$$Q_t = \arg\max \phi(c_i)$$

$$\phi(c_i) = P(c_{i1} \mid s_1) \prod_{j=1}^{n-1} P(c_{ij+1} \mid c_{ij}) P(c_{ij+1} \mid s_{j+1})$$

Statistic Method for Sense disambiguation(2/4)

Candidate target queries are generated by combining possible translation equivalents of each query word.

$$\boldsymbol{c(Q_s) = \{c_1 = (c_{11}, c_{12}, c_{13}, c_{14}), c_2 = (c_{21}, c_{22}, c_{23}, c_{24}) \ldots\ldots, c_{16} = (c_{16,1}, c_{16,2}, c_{16,3}, c_{16,4})\}}$$

1) Transition Probabilities

$$\mathrm{P}(c_{ij+1} = t_{j+1,1} \mid c_{ij} = t_{j1}) = \frac{\alpha(t_{j1}, t_{j+1,1})}{\sum_k^m \alpha(t_{j1}, t_{j+1,k})}$$

$$\alpha(\overrightarrow{w_1}, \overrightarrow{w_2}) = 2\log 2 + \sum_{y \in both} \{\overrightarrow{w_1}(y) \log \frac{\overrightarrow{w_1}(y)}{\overrightarrow{w_1}(y)+\overrightarrow{w_2}(y)} + \overrightarrow{w_2}(y) \log \frac{\overrightarrow{w_2}(y)}{\overrightarrow{w_1}(y)+\overrightarrow{w_2}(y)}\}$$

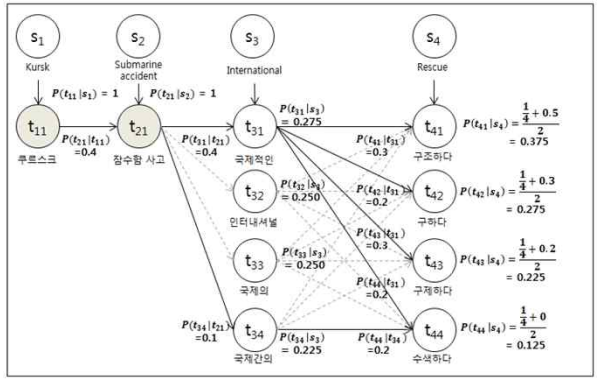
$$\overrightarrow{w_i} = (wt_{i1}, wt_{i2}, \ldots, wt_{in}),\ wt_{ij} = \frac{P(w_i|d_j)}{\sum_{k=1}^{n} P(w_i|d_k)},\quad P(w_i|d_j) = \frac{tf_{ij}}{dl_j}$$

Statistic Method for Sense disambiguation(3/5)

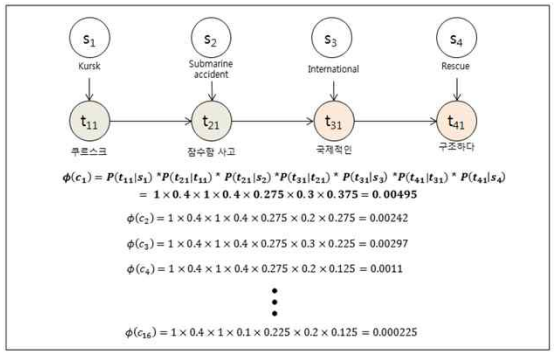
2) Translation Probabilities

$$\mathrm{P}(c_{ij} = t_{jk} \mid s_j) = \frac{\{\frac{1}{m} + P_{giza}(t_{jk} \mid s_j)\}}{2}$$

Statistic Method for Sense disambiguation(4/5)



Statistic Method for Sense disambiguation(5/5)



$$\phi(c_1) = P(t_{11}|s_1) * P(t_{21}|t_{11}) * P(t_{21}|s_2) * P(t_{31}|t_{21}) * P(t_{31}|s_3) * P(t_{41}|t_{31}) * P(t_{41}|s_4)$$
$$= 1 \times 0.4 \times 1 \times 0.4 \times 0.275 \times 0.3 \times 0.375 = 0.00495$$

$$\phi(c_2) = 1 \times 0.4 \times 1 \times 0.4 \times 0.275 \times 0.2 \times 0.275 = 0.00242$$

$$\phi(c_3) = 1 \times 0.4 \times 1 \times 0.4 \times 0.275 \times 0.3 \times 0.225 = 0.00297$$

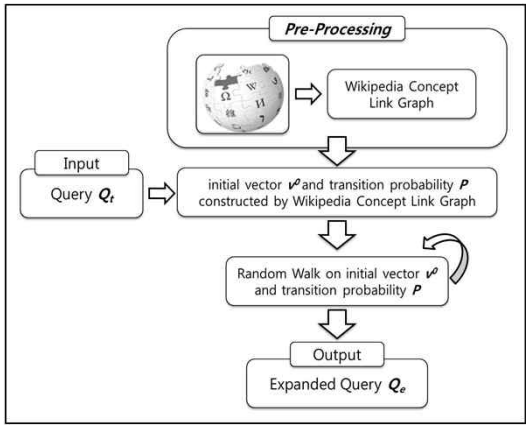
$$\phi(c_4) = 1 \times 0.4 \times 1 \times 0.4 \times 0.275 \times 0.2 \times 0.125 = 0.0011$$

$$\vdots$$

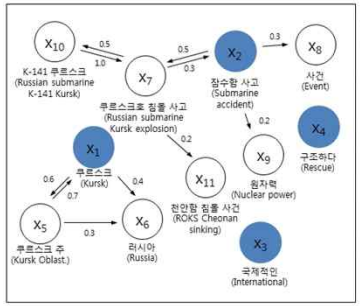
$$\phi(c_{16}) = 1 \times 0.4 \times 1 \times 0.1 \times 0.225 \times 0.2 \times 0.125 = 0.000225$$

3. QUERY EXPANSION FOR CLIR

Overview



Example



P	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11
x1	0	0	0	0	0.6	0.4	0	0	0	0	0
x2	0	0	0	0	0	0	0.5	0.3	0.2	0	0
x3	0	0	0	0	0	0	0	0	0	0	0
x4	0	0	0	0	0	0	0	0	0	0	0
x5	0.7	0	0	0	0	0.3	0	0	0	0	0
x6	0	0	0	0	0	0	0	0	0	0	0
x7	0	0.3	0	0	0	0	0	0	0	0.5	0.2
x8	0	0	0	0	0	0	0	0	0	0	0
x9	0	0	0	0	0	0	0	0	0	0	0
x10	0	0	0	0	0	0	1.0	0	0	0	0
x11	0	0	0	0	0	0	0	0	0	0	0

Wikipedia Concept Link Graph Construction

The correlations between Wikipedia concepts are estimated by using a random walk algorithm on Wikipedia concepts.

- The random walk algorithm is at each step the walk jumps to another site according to some probability distribution [Hu et al. 2009].

Based on the article links of Wikipedia

- Construct a concept link graph *G=(X,E)*, where *X* is a set of Wikipedia articles that is represented as $X = \{x_i\}_{i=1}^{m}$.
- *W* represents an weight matrix based on graph G
  - Element w_{ij} equals the link count associating vertices between x_i and x_j in the matrix *W*.

Random Walk on the concept Link Graph(1/4)

Initialize transition probability matrix *P* based on the weight matrix *W* and vector v_0 on the queries.

- Define transition probabilities $P_{t+1|t}(x_k|x_j)$ from the vertex x_j to x_k (x_j *and* $x_k \in X$) by normalizing the score out of node x_j.

$$P_{t+1|t}(x_k|x_j) = \frac{score_{jk}}{\sum_i score_{ji}}$$

- The initial vector v_0 is an *m*-dimensional vector with binary values

$$v_0 = \{p_0(x_j)\}_{j=1}^{m}$$

$$p_0(x_j) = \begin{cases} 1, if\ x_j \in Q \\ 0, otherwise \end{cases}$$

The probability that a random walk starts from x_j

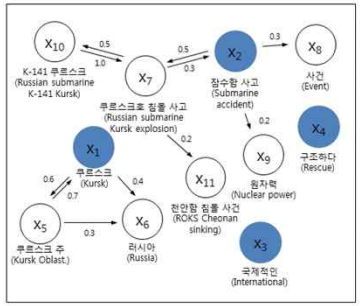
Random Walk on the concept Link Graph(2/4)

Given this definition, the random walk algorithm works as follows:

Input :Transition probability matrix P and initial vector v_0
Output : v_n

1 : for i := 0 to n
2 : compute $v_{i+1} = \alpha P^T v_i + (1-\alpha)v_i$, where $\alpha \in [0,1)$
3 : return v_n
4: Choose top k terms from v_n for query expansion

Random Walk on the concept Link Graph(3/4)



P	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11
x1	0	0	0	0	0.6	0.4	0	0	0	0	0
x2	0	0	0	0	0	0	0.5	0.3	0.2	0	0
x3	0	0	0	0	0	0	0	0	0	0	0
x4	0	0	0	0	0	0	0	0	0	0	0
x5	0.7	0	0	0	0	0.3	0	0	0	0	0
x6	0	0	0	0	0	0	0	0	0	0	0
x7	0	0.3	0	0	0	0	0	0	0	0.5	0.2
x8	0	0	0	0	0	0	0	0	0	0	0
x9	0	0	0	0	0	0	0	0	0	0	0
x10	0	0	0	0	0	0	1.0	0	0	0	0
x11	0	0	0	0	0	0	0	0	0	0	0

Random Walk on the concept Link Graph(4/4)

v_0

x_1	1
x_2	1
x_3	1
x_4	1
x_5	0
x_6	0
x_7	0
x_8	0
x_9	0
x_{10}	0
x_{11}	0

$P^T v_0$

x_1	0
x_2	0
x_3	0
x_4	0
x_5	0.6
x_6	0.4
x_7	0.5
x_8	0.3
x_9	0.2
x_{10}	0
x_{11}	0

v_1

x_1	0.50
x_2	0.50
x_3	0.50
x_4	0.50
x_5	0.30
x_6	0.20
x_7	0.25
x_8	0.15
x_9	0.10
x_{10}	0
x_{11}	0

$P^T v_1$

x_1	0.210
x_2	0.075
x_3	0
x_4	0
x_5	0.300
x_6	0.290
x_7	0.250
x_8	0.150
x_9	0.100
x_{10}	0.125
x_{11}	0.050

v_2

x_1	0.3550
x_2	0.2875
x_3	0.2500
x_4	0.2500
x_5	0.3000
x_6	0.2450
x_7	0.2500
x_8	0.1500
x_9	0.1000
x_{10}	0.0625
x_{11}	0.0250

$P^T v_2$

x_1	0.2100
x_2	0.0750
x_3	0
x_4	0
x_5	0.2130
x_6	0.2320
x_7	0.2063
x_8	0.0863
x_9	0.0575
x_{10}	0.1250
x_{11}	0.0500

v_3

x_1	0.2825
x_2	0.1813
x_3	0.1250
x_4	0.1250
x_5	0.2565
x_6	0.2385
x_7	0.2281
x_8	0.1181
x_9	0.0798
x_{10}	0.0938
x_{11}	0.0375

4. Experimental Results & Summary

EXPERIMENTAL SETTINGS (1/2)

The proposed query translation and query expansion techniques were tested using the NTCIR-5 English-Korean CLIR test collection.

- **All of the articles are indexed by using the Indri search engine.**

Statistical Report of Wikipedia Dataset

Content	English Wikipedia	Korean Wikipedia
# of Wikipedia articles	8,389,381	273,606
Total # of bilingual pair	105,643	
Total # of synonymy set	1,034,492	131,213
Total # of polysemy set	195,390	16,998

EXPERIMENTAL SETTINGS (2/2)

Baselines

- **Mono-IR: Original Korean Query**
- **Mono-IR+QE: Original Korean Query + QE (by Robertson)**
- **Simple-SQ: All translation candidates of each query word are considered as its synonym**
- **MT: Translation results from Google's free online language translation service**
- **PSQ: Straightforward extension to SQ in which translation probabilities are used.**
- **LF and LS : Lexical Filtering and Smoothing with CTL**
- **WTDM: Proposed query translation technique by (TDM: without translation prob.)**

Baseline Query Expansion Model

- **As a baseline query expansion system (QE), Robertson's query expansion technique is employed in our experiment [Robertson and Walker, 1999].**

$$RW(t) = r_t \log \frac{N}{n_t} - \log \binom{R}{r_t} - \log V$$

EXPERIMENT RESULTS (1/4)

Comparisons of Baseline Systems

System	Query Type	Average Precision(%)	MAP (%)	R-Precision (%)
Mono-IR	Title	29.86	36.36	31.58
Mono-IR+QE	Title	32.78	40.29	35.48
Simple-SQ	Title (% mono)	12.41 (41.56%)	15.30 (42.08%)	13.84 (43.83%)
MT	Title (% mono)	22.96 (76.89%)	28.70 (78.93%)	25.47 (80.65%)
PSQ	Title (% mono)	6.56 (21.96%)	8.94 (24.58%)	7.86 (24.88%)
LF-PSQ	Title (% mono)	8.97 (30.04%)	11.74 (32.28%)	9.84 (31.15%)
LS-PSQ	Title (% mono)	18.34 (61.41%)	23.14 (63.64%)	22.65 (71.72%)

EXPERIMENT RESULTS (2/4)

Performances of the proposed query translation technique

System	Query Type	Average Precision(%)	MAP (%)	R-Precision (%)
Mono-IR	Title	29.86	36.36	31.58
MRD-TDM	Title	18.58 (62.22%)	24.01 (66.03%)	21.13 (66.91%)
CTL-TDM	Title (% mono)	24.99 (83.69%)	33.64 (92.52%)	28.02 (88.73%)
CTL-LS-WTDM	Title (% mono)	27.08 (90.69%)	36.19 (99.53%)	30.51 (96.61%)

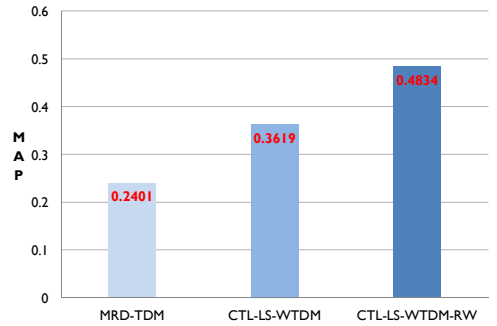
EXPERIMENT RESULTS (3/4)

Performances of the proposed query expansion technique

System	Query Type	Average Precision(%)	MAP (%)	R-Precision (%)
Mono-IR	Title	29.86	36.36	31.58
MRD-TDM-QE	Title	25.17 (84.29%)	33.81 (92.99%)	28.74 (91.01%)
MRD-TDM-WQE	Title (% mono)	29.26 (97.99%)	35.17 (96.73%)	30.66 (97.09%)
CTL-TDM-WLS	Title (% mono)	30.08 (100.74%)	35.73 (98.27%)	30.79 (97.50%)
CTL-TDM-RW	Title (% mono)	36.94 (123.71%)	46.75 (128.58%)	40.08 (126.92%)
CTL-LS-WTDM-RW	Title (% mono)	38.16 (127.80%)	48.34 (132.95%)	41.93 (132.77%)

EXPERIMENT RESULTS (4/4)

Summary of the performances in the proposed technique (WTDM)



Summary

Utilize what may be the world's largest knowledge resource, Wikipedia, to help query translation and expansion in CLIR.

- The query translation technique used bilingual word lists.
- The query expansion technique used a concept link graph from Wikipedia link information and a random walk algorithm

Evaluation results indicate significant improvements over strong baselines, exceeding even monolingual retrieval effectiveness

References

References

- H. Seo, S. Kim, H. Rim, and S. Myaeng. "Improving query translation in English-Korean cross-language information retrieval," *Information Processing and Management*, Vol.41, pp. 507-522, 2005.
- J. Hu, G. Wang, F.Lochovsky, J. tao Sun, and Z. Chen, "Understanding User's Query Intent with Wikipedia," In the *Proceedings of the 18th international conference on World Wide Web (WWW'09)*, pp. 471-480.
- S. E. Robertson and S. Walker, "Okapi/keenbow at TREC-8," In the *Proceedings of the eight Text Retrieval Conference (TREC-8)*, pp. 151-161, 1999.
- S. Kim, Y. Ko and D. W. Oard, "Combining Lexical and Statistical Translation Evidence for Cross-Language Information Retrieval," Journal of the American Society for Information Sciences and Technology. *(Under Revision)*

Thank you so much!

Youngjoong Ko

Dept. of Computer Engineering
Dong-A University
(http://web.donga.ac.kr/yjko)