English-to-Korean Cross-linking of Wikipedia Articles at KSLP

In-Su Kang
School of Computer Science and Engineering
Kyungsung University
Busan, South Korea
dbaisk@ks.ac.kr

Ralph Marigomen
School of Computer Science and Engineering
Kyungsung University
Busan, South Korea
rmarigomen@gmail.com

ABSTRACT
This paper describes team KSLP's approach for the NTCIR-9 English-to-Korean cross-lingual link detection task. There are three main steps that compose the whole system. Given an English topic document, first is identifying English anchor terms by exploiting both context-less and contextual link evidences from the English Wikipedia corpus. Then, by utilizing English-to-Korean translation dictionaries we obtain Korean translations for each identified anchor terms. Lastly, we disambiguate these translations by computing for the document similarity between its respective Korean document and the English topic document.

Categories and Subject Descriptors
I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *text analysis*.

I.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *linguistic processing*.

General Terms
Experimentation.

Keywords
Wikipedia, Cross-lingual Link Discovery, Anchor Identification, Cross-lingual Link Recommendation.

1. INTRODUCTION
We have participated in this year's NTCIR-9 English-to-Korean Cross-Lingual Link Detection (CLLD) task. The task aims at associating documents written in different languages by linking document terms from a specific language to documents (or smaller units within documents) in another language [1].

CLLD can be tackled using source-language anchor identification followed by target-language link determination. For anchor identification, we exploited anchor evidences previously found in existing English Wikipedia articles in two different ways: context-independent and context-dependent. The basic assumption was that terms such as company names or celebrities are more likely to be anchored regardless of the document topic, while other terms would only show such tendency within a specific context or topic. Based on this assumption, our anchor identifier produces anchor-candidate terms ordered by the combination of their context-less and contextual anchor strengths.

Unlike earlier works [2,3,4] on mono-lingual link detection (MLLD) especially for Wikipedia English articles, CLLD requires a method in obtaining a list of possible translations for each anchor term as well as a translation disambiguation technique. For the former, an English-to-Korean translation dictionary has been used. For the latter, a contextual similarity between a source-language English term and each of its target-language Korean translation is calculated to select the best translation term. Finally, the English anchor term is linked to the Korean article of which title is the same as the disambiguated Korean translation term.

The remainder of this report is organized as follows. Sections 2 and 3 describe the methods for English anchor identification and Korean document link recommendation. Section 4 explains a unified approach to CLLD, which combines anchor finding and link determination. Section 5 shows our submission results. Finally, Section 6 gives concluding remarks.

2. Anchor Identification
Our general procedure for identifying English anchors is to rank all candidate keyword terms from a source document and select the top ranked terms from the list. A step by step discussion of such procedure is explained as follows. Step-1 starts off by generating all possible n-grams from a given source document. Then we utilized a controlled vocabulary list to filter out improper n-grams such as *"specifically to the"* or *"The term is"*. The vocabulary list consisting of title terms and link terms is gleaned from the English Wikipedia corpus, as done in [2].

$$KeywordScore(D,t) = \alpha \times GLS(t) + (1-\alpha) \times LLS(D,t) \quad (1)$$

$$GLS(t) = \frac{af(t)}{cf(t)}$$

$$LLS(D,t) = \frac{|aTopics(t) \cap Topics(D)|}{|Topics(D)|} \approx \delta(Title(D) \in aTopics(t))$$

Step-2 assigns a score to a candidate keyword t appearing in the source document D using the above Formula (1) which combines two kinds of link clues: global and local link evidence. Global link score $GLS(t)$ represents the likelihood that term t would be an anchor term if t is randomly selected from the Wikipedia corpus, while local link score $LLS(D,t)$ means such possibility when t is picked from Wikipedia documents about the topic of D. *GLS* is conceptually similar to the link probability or keyphraseness used in [2]. In this experiment, *GLS* is defined as the fraction of the entire instances of t that are anchored. From the formula (1), $af(t)$ means the anchored frequency of t, and $cf(t)$ collection frequency.

One method to estimate *LLS* is to compute the overlapping ratio between anchored topic set $aTopics(t)$ of t and topic set $Topics(D)$ of D, where $Topics(D)$ might be the set of candidate keyword terms from D and $aTopics(t)$ be the set of anchored terms in a

document titled *t*. To simplify the calculation for *LLS*, we used a binary function that checks if *aTopics*(*t*) includes the title of *D*, *Title*(*D*). δ(*p*) is defined as 1 if *p* is true, 0 otherwise.

In Step-3, we eliminate candidate keywords with score lower than a pre-defined threshold value θ and choose the remaining terms as anchors.

3. Cross-lingual Link Recommendation

3.1 Translation Dictionary

To retrieve Korean translation candidates for each generated English anchor terms, we employed two kinds of English-to-Korean (E-K) translation dictionaries as our resource: a Wiki-domain E-K dictionary (WikiEKDic) and a general E-K dictionary (GenEKDic).

WikiEKDic was generated from an English Wikipedia dump file and the CLLD Korean collection using Korean links of English Wikipedia articles and English links of CLLD Korean documents. After removing non-English terms, the total number of translation pairs in WikiEKDic was 123,557.

GenEKDic was created using E-K general translation dictionaries from POSTECH KLE laboratory[1] and the English Wiktionary dump file. GenEKDic consisted of about 1,093,188 translation pairs.

To simplify our cross-lingual link recommendation, translation pairs from the two E-K dictionaries of which Korean translations do not correspond to any titles of the CLLD Korean articles were removed. As a result, the final WikiEKDic and GenEKDic consist of 107,466 and 198,386 translation entries respectively.

These two E-K dictionaries are sequentially accessed in order to retrieve the Korean translation of a given English anchor term t_e. WikiEKDic is first searched to get the possible translations. If no such term exists in WikiEKDic, then we search furthermore in GenEKDic.

3.2 Run Description

This section describes five runs *run1*, *run2*, …, *run5* we have submitted to E-K CLLD task.

3.2.1 Run2

$$TransDisam(D_e, t_e) = \underset{t_k \in EKtrans(t_e)}{\operatorname{argmax}} \left(\max_{D_k \in Postings(t_k)} Sim(D_e, D_k) \right) \quad (2)$$

Translation disambiguation for *run2* was performed using the above Formula (2), where $EKtrans(t_e)$ returns a set of Korean translations for English anchor term t_e using the E-K translation dictionary described in the previous section, and $Postings(t_k)$ retrieves a set of Korean documents in which t_k appeared as an anchor term.

Given an English anchor term t_e and document D_e where t_e appeared, *TransDisam*() thus determines the most probable Korean term t_k for t_e. Some English anchor terms may not have any entries in E-K dictionary. Such terms were ignored in CLLD process.

[1] http://kle.postech.ac.kr/

The basic idea is to find the most appropriate Korean translation with context mostly similar to that of the English anchor term. For *run2*, we used document as the unit for context. For trans-lingual matching between English and Korean documents, the English document is represented as a collection of all possible Korean translations of its English candidate-keyword terms.

As for the Korean indexing unit, overlapping bi-character terms were generated from the entire text of a Korean document. Then, a retrieval model to calculate *Sim*() in Formula (2), BM25 was employed.

3.2.2 Run3

The only difference between *run2* and *run3* is that *run3* indexes only the first paragraph of each Korean document rather than the whole content. This was done for the fact that the first paragraph of a Wikipedia article normally includes the definitional sentence of its topic title.

3.2.3 Run1

$$TransStrength(t_e, t_k, D_e) = Sim\left(D_e, D_{title=t_k}\right) \quad (3)$$

For each English anchor term t_e appearing in topic document D_e, *run1* assigns scores to each of the Korean translations t_k using the above Formula (3), where $D_{title=t_k}$ means the Korean document titled t_k. The representation of D_e for this run is similar to that of *run2*. For *Sim*() in Formula (3), the inner product is performed against two tf-based bi-character vectors.

For each English anchor, *run1* then produces target links of all possible Korean translations ordered by *TransStrength*(). Note that all the other runs, *run2* to *run5*, were created so that every English anchor term has a single Korean document link.

4. A Unified Approach

4.1 Overview

Typical steps for a CLLD system would include anchor identification followed by link recommendation. However, anchor identification may produce anchor terms which don't have a corresponding target-language document. To deal such a problem, we devised a unified approach in CLLD which combines these two typical steps.

4.2 Run Description

4.2.1 Run4

$$\begin{aligned} &AnchorScore(D,t) \\ &= \beta \times KeywordScore(D,t) + (1-\beta) \times TransDisamScore(D,t) \end{aligned} \quad (4)$$

$$TransDisamScore(D_e, t_e) = \max_{D_k \in \bigcup_{t_k \in EKtrans(t_e)} Postings(t_k)} Sim(D_e, D_k)$$

As a unified approach to CLLD, *run4* generates anchor terms with its best links using the above Formula (4) in a single step. *TransDisamScore*() is the same as *TransDisam*() in *run2* except that it returns the similarity score of the best Korean translation for each English anchor term t_e. Note that *run4* produces anchor terms of which *AnchorScore*() is greater than or equal to a threshold θ as mentioned in Section 2.

4.2.2 Run5

In Section 2, *aTopics*(*t*) was considered as the set of anchored terms in a document titled *t*. Topic terms in *aTopics*(*t*) may be viewed to be directly related to *t* since those terms occurred in a document of which the title is *t*. Then, for topic term *s* in *aTopics*(*t*), topic terms in the document titled *s* can be regarded as indirectly associated to *t*. Topic terms which are directly related to *t* are named as first-order topics while indirectly related topic terms are called second-order or higher-order topics.

Based on the above discussion, *LLS* in Section 2 can be redefined as follows.

$$LLS(D,t) = \max_{L=1}^{N}\left(\lambda_L \times \delta\left(Title(D) \in aTopics_L(t)\right)\right) \quad (5)$$

In Formula (5), *L* indicates the topical association level. Therefore, $aTopics_1(t)$ and $aTopics_2(t)$ would mean first-order topics and second-order topics respectively. λ_L is a tunable parameter which represents the topical relatedness strength between *L*-th order topic terms and *t*. *Run5* is the same as *run4* except that *LLS* is replaced with Formula (5).

5. Submission Results

The parameter values used for creating the final submission runs are as follows. θ=0.2, α=β=0.5, *N*=3, λ_1=1, λ_2=0.5, λ_3=0.25. Unfortunately, their optimal values could not be determined due to time constraints before the run submission deadline.

Tables 1, 2 and 3 shows the official results of our submission runs respectively for F2F automatic, F2F manual, and A2F manual evaluations[1]. In these Tables, *Best* row, which indicates the performances of top-ranked system for each evaluation type, were included for comparison purposes.

Table 1. F2F automatic evaluation result

Run	MAP	R-p	P5	P10	P30	P50	P250
Best	*0.447*	*0.506*	*0.848*	*0.764*	*0.625*	*0.520*	*0.148*
Run1	0.260	0.346	0.632	0.564	0.444	0.362	0.122
Run2	0.316	0.427	**0.696**	**0.660**	**0.547**	**0.448**	**0.116**
Run3	0.318	0.431	0.680	0.652	0.552	0.450	0.117
Run4	0.326	0.439	0.680	0.688	0.548	0.454	0.122
Run5	**0.328**	**0.437**	0.680	0.684	0.544	0.452	0.123

Table 2. F2F manual evaluation result

Run	MAP	R-p	P5	P10	P30	P50	P250
Best	*0.376*	*0.522*	*0.720*	*0.672*	*0.537*	*0.463*	*0.124*
Run1	**0.233**	**0.341**	0.552	0.576	0.539	0.483	0.286
Run2	0.170	0.245	**0.680**	**0.664**	**0.621**	**0.575**	**0.169**
Run3	0.169	0.244	0.672	0.656	0.620	0.578	0.168
Run4	0.177	0.252	0.640	0.656	0.625	0.577	0.175
Run5	0.184	0.264	0.640	0.656	0.631	0.581	0.182

Table 3. A2F manual evaluation result

Run	MAP	R-p	P5	P10	P30	P50	P250
Best	*0.232*	*0.207*	*0.368*	*0.364*	*0.332*	*0.320*	*0.129*
Run1	0.079	0.097	0.184	0.196	0.219	0.231	0.108
Run2	0.072	0.075	0.232	0.252	0.263	0.250	0.083
Run3	0.073	0.076	**0.240**	**0.264**	**0.265**	**0.254**	**0.084**
Run4	0.076	0.080	0.200	0.264	0.265	0.248	0.088
Run5	**0.081**	**0.086**	0.200	0.256	0.268	0.254	0.095

6. Conclusion

Our team KSLP have participated at NTCIR-9 English-to-Korean CLLD task. Among the five runs that we have submitted, A2F manual evaluation using pre@n shows that run3 performed better compared to other runs. Evaluation result using other metrics did not produce good results and such results could not be used for comparing different runs. In the future, we see the need of post-experiments and in-depth analysis for our current approach.

7. REFERENCES

[1] Adafre, S.F., and Rijke, M. 2005. Discovering missing links in Wikipedia. In *Proceedings of the LinkKDD-2005*, 90-97.

[2] Mihalcea, R. and Csomai, A. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 233-242.

[3] Milne, D. and Witten, I.H. 2008. Learning to link with Wikipedia. In *Proceedings of the ACM Conference on Information and Knowledge Management*, 509-518.

[4] Tang, L.X., Geva, S., Trotman, A., Xu Y., and Itakura K. 2011. Overview of the NTCIR-9 Crosslink task: cross-lingual link discovery. In *Proceedings of the 9-th NTCIR Workshop Meeting*.