Cross-lingual Link Discovery by Using Link Probability and Bilingual Dictionary

Sin-Jae Kang
School of Computer and Information Technology,
Daegu University
Gyeonsan. Gyeongbuk. 712-714 South Korea
+82-53-850-6584
sjkang@daegu.ac.kr

ABSTRACT
This paper presents a method to discover English to Korean cross-lingual links by using resources such as link probability, title lists of Wikipedia articles, and an English-Korean bilingual dictionary.

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

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

General Terms
Experimentation.

Keywords
DUIIS, English Wikipedia, Korean Wikipedia, English-Korean Bilingual Dictionary, Cross-lingual Link Discovery, Anchor Identification, Link Recommendation

1. INTRODUCTION
Among three NTCIR-9 cross-lingual link discovery (CLLD) subtasks [1], we participated in English to Korean CLLD. For each English document provided as topics, prospective anchors are identified and relevant links for them are recommended in the Korean document collection. Figure 1 shows English to Korean CLLD, outgoing link starting from English source documents and pointing to Korean target documents in Wikipedia.

Generally, CLLD comprises of three phases, anchor extraction, anchor translation, and target link selection. In the phase of anchor extraction, words or phrases relevant to the topic and worthy of being linked to target documents for getting more information are extracted and ranked. In the phase of anchor translation, candidate anchors are translated into target words through the process of word sense disambiguation (WSD). In the final phase of target link selection, for each translated anchor, it is selected as a target cross-lingual link in the case that the translated word exists in title lists of target document collection.

This paper proposed a simple CLLD using link probability for target language instead of WSD.

NTCIR-9 Workshop Meeting, December 6–9, 2011, Tokyo, Japan.

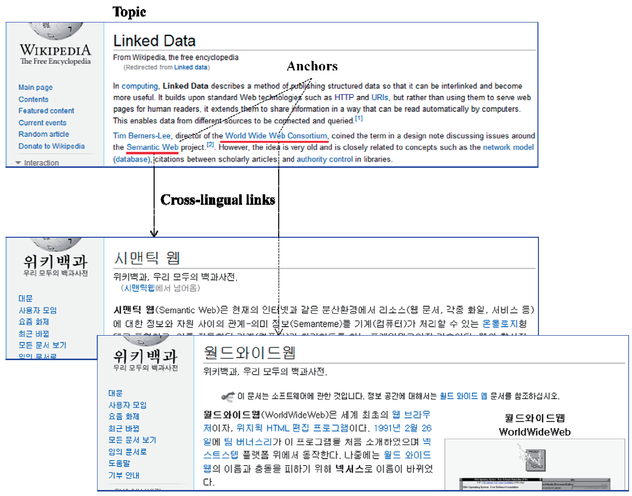

Figure 1. English to Korean Cross-lingual Linking

2. PROPOSED APPROACH

2.1 Resource Construction
First of all, resources for CLLD were extracted from vast information contained in the already annotated articles of Wikipedia[1]. Automatically extracted resources are title lists of English and Korean Wikipedia, link probability for English and Korean, and English-Korean linking list. After extracting the titles of Wikipedia documents, their link probability was calculated. For each English title, its corresponding Korean title was collected as English-Korean linking list if exists.

Mihalcea et al showed *keyphraseness* is the best method than tf-idf and χ^2 in extracting keywords [2]. Keyphraseness is based on the assumption that "the more often a term was selected as a keyword among its total number of occurrences, the more likely it is that it will be selected again." So, the keyphraseness is used to calculate link probability. The probability of a term T to be selected as an anchor is defined as follows:

$$\mathrm{Prob(anchor|T)} = \frac{\mathrm{Freq}_{\ anchor}}{\mathrm{Freq}_{\ T}} \qquad (1)$$

In Wikipedia dump, there are some corresponding links to Korean documents for each English topic. This is valuable information to

[1] http://dumps.wikimedia.org/

translate anchors and select them as cross-lingual links. Figure 2 shows resources extracted from both English and Korean Wikipedia dumps.

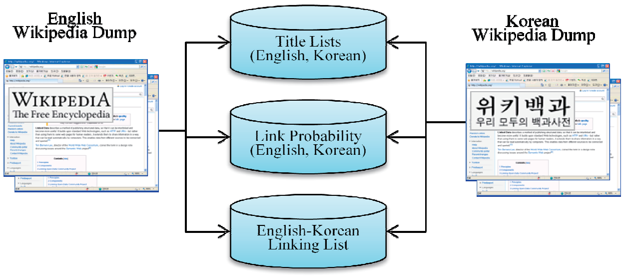

Figure 2. Resources Extracted from Wikipedia Dumps

During the phase of anchor translation, an English-Korean bilingual dictionary is needed to gather possible target Korean words for each English anchor. Table 1 shows some statistics about the constructed resources and the English-Korean bilingual dictionary, which was created from two E-K general translation dictionaries from the POSTECH KLE laboratory[2].

Table 1. Resource size

Resource Type	# of Items
English Title List	11,357,871
Korean Title List	201,486
English-Korean Linking List (English title - Korean title)	141,685
English-Korean Bilingual Dictionary (English word – Korean word)	186,773

2.2 Candidate Anchor Extraction

Words or phrases contained in the English title list are considered as controlled vocabularies for acceptable anchors; nonsense phrases will not appear as candidate anchors. Given an English topic, the longest-match candidate anchors are extracted by considering terms only in the English title list, and ranked by their English link probability, and then up to 250 anchors are selected among top ranked anchors.

2.3 Target Word Selection

Three resources, which are English-Korean linking list, Korean link probability, and English-Korean bilingual dictionary, are used to select target words for each anchor. There is a priority among these resources according to the importance of their information. The English-Korean linking list has higher priority than others; because it contains actual cross-lingual links exist in real Wikipedia. Therefore, first the English-Korean linking list is retrieved for each candidate anchor, if its correspondent exists, then the corresponding Korean title is selected as its target link. Otherwise, the English-Korean bilingual dictionary is retrieved to gather possible Korean words of candidate anchors, and rank the Korean words by using Korean link probability, which plays a role of finding the most frequent word among candidate Korean words. In our approach, to make the CLLD task practical and efficient, WSD process wasn't adopted. Figure 3 shows proposed CLLD process.

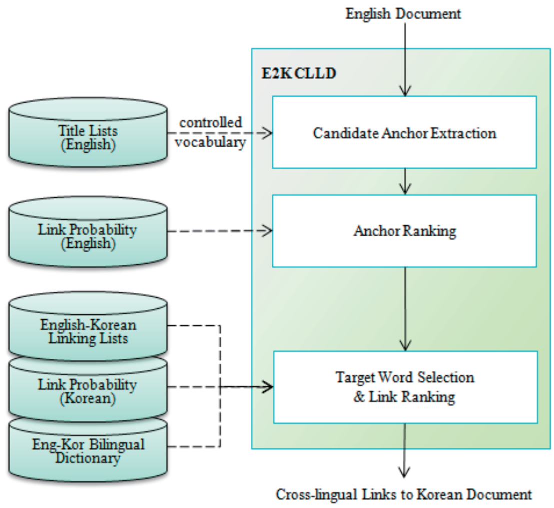

Figure 3. Overall Process for CLLD

3. EXPERIMENTS

In this paper, two experiments were performed, which called *DUIIS_A2F_E2K_4Pre* and *DUIIS_A2F_E2K_4Rec*. During the process of selecting and ranking the translated anchors, *DUIIS_A2F_E2K_4Pre* used three resources described in section 2.3, and *DUIIS_A2F_E2K_4Rec* only used two resources except link probability for Korean. By doing this, the role of link probability for Korean was figured out in CLLD without WSD. Evaluation tool[3] provided by Tang et al was used to evaluate the proposed method. In Table 2, our experimental results are compared with the best scored run in each evaluation. In evaluation with Wikipedia ground-truth, *DUIIS_A2F_E2K_4Pre* shows better result than other, on the other hand *DUIIS_A2F_E2K_4Rec* better in evaluation with manual assessment. Even though both proposed approach for English to Korean CLLD is quite simple and doesn't apply WSD, it ranked as the top three teams in file-to-file evaluation with both Wikipedia ground-truth and manual assessment [1]. That performance could be achieved by using link probability for Korean and English-Korean bilingual dictionary.

4. CONCLUSION

Comparing to general CLLD methods, this paper proposed a method to discover English to Korean cross-lingual links by additionally using link probability for target language and English-Korean dictionary instead of WSD process.

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

[3] http://crosslink.googlecode.com/files/CrosslinkEvaluation-20110907.zip

In the future, a WSD method based on WordNet and Google [3] will be applied to improve the performance of anchor-to-file evaluation.

5. ACKNOWLEDGMENTS
This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0007025).

6. REFERENCES
[1] 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 9th NTCIR Workshop Meeting* (Tokyo, Japan, December 6-9, 2011).

[2] Mihalcea, M. and Csomai, A. 2007. Wikify! Linking Documents to Encyclopedic Knowledge. In *Proceedings of the CIKM'07* (Lisboa, Portugal, November 6-8, 2007).

[3] Kang, S. J. and Kang, I. S. 2009. Generalization of Ontology Instances Based on WordNet and Google. *Journal of Korean Institute of Intelligent Systems*. 19, 3 (June 2009), 363-370. (Written in Korean)

Table 2. Experimental results with evaluation tool provided

Wikipedia Ground-Truth (File to File)

Run ID	MAP	R-Prec	P5	P10	P20	P30	P50	P250
HITS_E2K_A2F_01	0.447	0.509	0.848	0.764	0.720	0.625	0.520	0.148
DUIIS_A2F_E2K_4Pre	0.370	0.442	0.792	0.768	0.674	0.596	0.479	0.124
DUIIS_A2F_E2K_4Rec	0.365	0.438	0.784	0.760	0.760	0.583	0.474	0.126

Manual Assessment (File to File)

Run ID	MAP	R-Prec	P5	P10	P20	P30	P50	P250
UKP_E2K_A2F_02	0.376	0.522	0.544	0.568	0.632	0.656	0.656	0.413
DUIIS_A2F_E2K_4Rec	0.258	0.379	0.632	0.692	0.700	0.687	0.658	0.263
DUIIS_A2F_E2K_4Pre	0.252	0.357	0.632	0.692	0.716	0.705	0.679	0.247

Manual Assessment (Anchor to File)

Run ID	MAP	R-Prec	P5	P10	P20	P30	P50	P250
UKP_E2K_A2F_02	0.232	0.207	0.192	0.248	0.324	0.323	0.326	0.252
DUIIS_A2F_E2K_4Rec	0.043	0.036	0.064	0.076	0.090	0.116	0.115	0.036
DUIIS_A2F_E2K_4Pre	0.041	0.034	0.056	0.076	0.090	0.117	0.118	0.034