A Single-step Machine Learning Approach to Link Detection in Wikipedia: NTCIR Crosslink-2 Experiments at KSLP

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

Sin-Jae Kang
School of Computer and Information Technology,
Daegu University
Gyeonsan, Gyeongbuk, South Korea
sjkang@daegu.ac.kr

ABSTRACT
This study describes a link detection method to find relevant cross-lingual links from Korean Wikipedia documents to English ones at term level. Earlier wikification approaches have used two independent steps for link disambiguation and link determination. This study seeks to merge these two separate steps into a single-step machine learning scheme. Our method at NTCIR-10 Korean-to-English CLLD task showed promising results.

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

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

General Terms
Experimentation.

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

Team Name
KSLP

Subtasks
Korean to English

1. INTRODUCTION
Cross-lingual link discovery (CLLD or Crosslink) at NTCIR aims to find relevant links between different-language documents [4]. At NTCIR-10, our team KSLP (Kyungsung University Language Processing Laboratory) has participated at Korean-to-English (K2E) CLLD subtask which handles linking Korean Wikipedia (hereafter *KorWiki*) documents to English ones at term level.

Generally, MLLD (Mono-Lingual Link Discovery) approaches require two major steps: mono-lingual link disambiguation and link determination. In a similar way, CLLD could perform cross-lingual link disambiguation and link determination. However, unlike MLLD where there are rich clues for link disambiguation, the amount of existing cross-lingual links for learning to directly disambiguate target-language translations in CLLD may not be sufficient. To handle this, we have adopted Fahrni's approach [1]. For English to CJK CLLD, Fahrni et al. attempted English MLLD first, and then translated target links (to English documents) into CJK document links using document-level inter-language links gathered mainly from CJKE Wikipedia collections.

Previous MLLD approaches [2, 3] have employed two separate stages to deal with respectively link disambiguation and link determination, mostly relying on machine learning approaches. Unlike earlier methods, this study attempts to combine these two stages into a single-step machine learning method.

2. KOREAN ANCHOR CANDIDATE

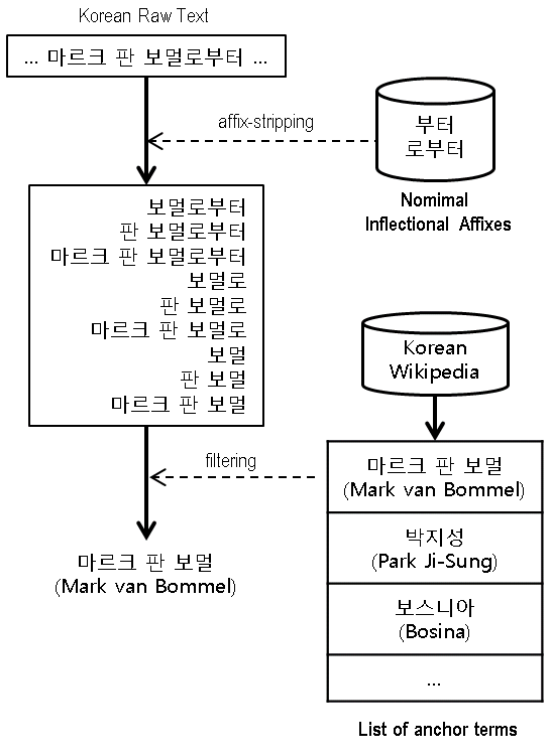

Figure 1. Affix-based multi-word anchor term generation

Unlike English, the Korean spacing unit called *'eojeol'* consists of words (precisely stems) followed by optional inflectional affixes. An eojeol may correspond to a word, two or more words, or a phrase in English. For example, a Korean eojeol '서울로', meaning *'to Seoul'* in English, is composed of two morphemes '서울' (*Seoul*) and '로' (*to*). A certain method is needed to remove inflectional morphemes from eojeols to identify content-bearing

stems which would be anchor candidates. In addition, anchor terms used in Korean text may be comprised of one or more eojeols like '대한민국 임시 정부' ('*Provisional Government of the Republic of Korea*' in English). So, we need to generate multi-word anchor candidates.

To attack the above issues regarding Korean, an affix-based multi-word anchor term generation method is employed to yield all possible anchor candidates from Korean text. The method is as follows. For each eojeol in Korean text, a list[1] of Korean nominal inflectional affixes is applied to discard nominal endings (including null ending) to produce one or more stem candidates ('보멀로부터', '보멀로', '보멀' in Figure 1). For each such stem candidate, multi-word terms are generated as anchor candidates by concatenating up to *n* preceding eojeols[2]. Then, a set of previous real anchor terms extracted from the KorWiki collection is used to filter out incorrect or unexpected anchor candidates (like '*Obama is*' or '*is a president*' in case of English). Figure 1 illustrates this method, which is an multi-word extension of an affix-based index term extraction scheme [5] for Korean information retrieval.

3. MONO-LINGUAL LINK DETECTION

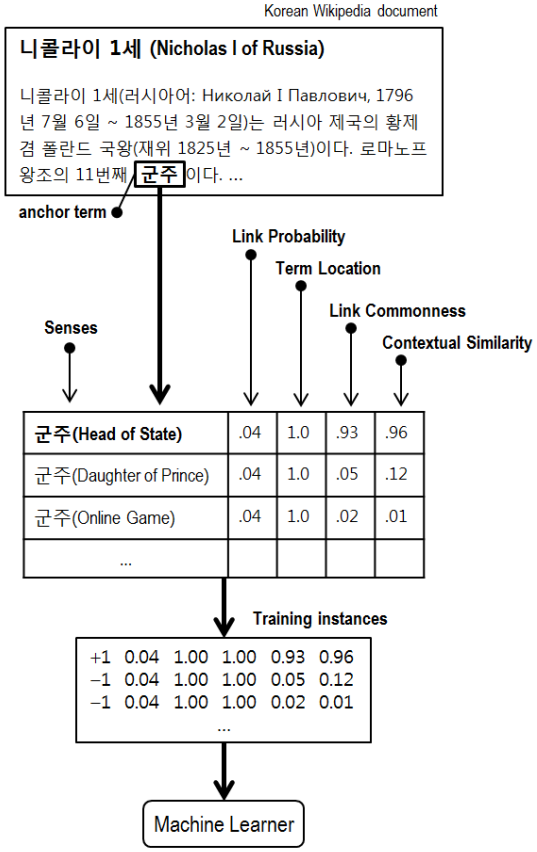

Figure 2. A single-step link detection

For each Korean anchor candidate term, MLLD disambiguates and determines links to KorWiki documents. Figure 2 shows a single-step learning approach to detect links. For each real anchor term *t* (e.g. 군주 in Figure 2) found in training documents sampled from KorWiki, a list of possible Korean links (senses) is retrieved from anchorTerm-to-link index which were created from KorWiki. Next, a correct link of the anchor term is used to generate a positive instance with four features explained in the following, and the other links to produce negative instances.

Four features were chosen: link probability, term location, link commonness, and contextual similarity. The first two features are for determining link relevance and the remaining ones for disambiguating links (senses) of the anchor term. In addition, link probability and link commonness represent global constraints gathered from a collection of Wikipedia documents, while term location and contextual similarity indicate local constraints encoded in the current input document.

Linkness-oriented Global Feature: Link probability [2] is the probability of term *t* being anchored, which is defined below where *cf*(*t*) is a collection frequency of *t*, and *af*(*t*) an anchored frequency of *t* in the collection.

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

Linkness-oriented Local Feature: Term location [3] captures the assumption that terms are more likely anchored in the earlier parts of the document. In its formula below, *DocSize* is the total number of paragraphs in a document where term *t* appears, and *ParagraphNo*(*t*) is a sequence number (starting from zero) of the first paragraph where *t* is found.

$$TermLocation(t) = \frac{DocSize - ParagraphNo(t)}{DocSize}$$

Disambiguation-oriented Global Feature: Link commonness [3] encodes a prior probability for link *l* of term *t*. In its following formula, *LinkSet*(*t*) is a set of all links (senses) of *t*.

$$LinkCommonness(t,l) = \frac{af(l)}{\sum_{l' \in LinkSet(t)} af(l')}$$

Disambiguation-oriented Local Feature: Contextual similarity, first proposed in this study, exploits the assumption that related links (senses) within a natural language context are found more frequently than unrelated ones. For each link (sense) *l* of an anchor candidate term *t*, this similarity seeks to quantify a normalized sum of semantic relatedness between *l* and each of links in a context where *t* occurs. *freq*(*l*,*l'*) in the following formula indicates the count of context units where both links *l* and *l'* are found. Context units could be a sentence, a paragraph, or a document. In our experiments, *freq*(*l*,*l'*) were gathered only for sentential contexts from KorWiki collection. In the formula, *l* is one of links for an anchor candidate term *t*, and *T* indicates a paragraph context where *t* appears. Regarding context representation, a set of terms was used. Thus, *T* corresponds to a set of terms in the *t*-occurring paragraph of the Crosslink-2 topic file.

[1] http://nlp.kookmin.ac.kr/down/data/KorStems.zip

[2] This study set *n* to 5

$$ContextaulSimilarity(l,T) = \frac{\sum_{term \in T} max(\{freq(l,l') | l' \in LinkSet(term)\})}{\sum_{term \in T} max(\{\max(1, freq(l,l')) | l' \in LinkSet(term)\})}$$

4. KOREAN-TO-ENGLISH TRANSLATION

After mono-lingual link detection, a list of Korean anchor terms each with its Korean link (a target KorWiki document) disambiguated and determined is obtained. Then, translating Korean links to English relies on a Korean-to-English dictionary, which was created from inter-language links of the Crosslink-2 English and Korean Wikipedia dump files.

5. SUBMISSION & RESULTS

After downloading KorWiki collection, 500 training documents for mono-lingual link detection were sampled from it, and all later processing has been carried out without the sampled documents. libSVM[3] was used to learn and classify whether or not a link (to a KorWiki document) is relevant for a candidate anchor term in Korean documents.

For each Korean topic file, the affix-based multi-term generation method in Section 2 is applied to produce a list of anchor term candidates. Next, each anchor term is used to produce a list of its possible links (senses), each of which is converted into a test instance comprised of four features described in Section 3. Next, each of such instances is classified into either *to-be-linked class* or *not-to-be-linked class* with its confidence score from SVM. Then, links (to Korean Wikipedia documents) are translated into links to English Wikipedia equivalents by referring to the Korean-to-English dictionary described in Section 4. Finally, translated links are inversely sorted by their confidence scores and grouped by their originating anchor candidate term to create an ordered list of anchor-link recommendations.

Our team KSLP submitted a single run. Its official scores are reported in Table 1. The proposed method showed better performances at two manual evaluations (F2F and A2F) than the other participants' methods, while only achieving less than half of the best performance at ground-truth evaluation.

Ground-truth and manual evaluations would be equally important since the two reflect how well a link detection method matches respectively writers' and readers' thoughts of whether terms in a document should be linked or not. A further analysis would be needed to describe the discrepancies between ground-truth and manual performances our system obtained.

Table 1. Submission Results

	F2F ground-truth		F2F manual		A2F manual	
	LMAP	**R-pre**	**LMAP**	**R-pre**	**LMAP**	**R-pre**
Best	0.322	0.324	0.302	0.301	0.205	0.009
KSLP	0.145	0.215	**0.302**	**0.301**	**0.205**	**0.009**

[3] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

6. CONCLUSION

Our team have participated at NTCIR-10 Korean-to-English CLLD task using an affix-based multi-word generation scheme to produce Korean anchor candidates and a single-step machine learning approach to link detection. The proposed method has reported relatively good performances in manual evaluations among small K2E CLLD participants.

7. 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 (2012R1A1A1011668).

8. REFERENCES

[1] Fahrni, A., Nastase, V., and Strube, M. 2011. HITS' graph-based system at the NTCIR-9 cross-lingual link discovery task. In *Proceedings of the 9th NTCIR Workshop Meeting*. 473-480.

[2] Mihalcea, R. and Csomai, A. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the ACM 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*. 508-518.

[4] Tang, L.X., Kang, I.S., Kimura, F., Lee, Y.H., Trotman, A., Geva, S., and Xu, Y. 2013. Overview of the NTCIR-10 cross-lingual link discovery task. In *Proceedings of the 10th NTCIR Workshop Meeting.*

[5] Yae, Y.H. 1992. Automatic keyword extraction system for Korean documents information retrieval. *Information Management Research*. 23(1):39-62 (in Korean)