UKP at CrossLink2: CJK-to-English Subtasks

Jungi Kim and Iryna Gurevych
Ubiquitous Knowledge Processing (UKP) Lab
Technische Universität Darmstadt
Hochschulstrasse 10
D-64289 Darmstadt, Germany
http://www.ukp.tu-darmstadt.de
{kim, gurevych}@ukp.informatik.tu-darmstadt.de

ABSTRACT

This paper describes UKP's participation in the cross-lingual link discovery task at NTCIR-10 (CrossLink2). The task addressed in our work is to find valid anchor texts from a Chinese, Japanese, and Korean (CJK) Wikipedia page and retrieve the corresponding target Wiki pages in the English language. The CrossLink framework was developed based on our previous CrossLink system that works on the opposite directions of the language pairs, i.e. discovered anchor texts from English Wikipedia pages and their corresponding targets in CJK languages. The framework consists of anchor selection, anchor ranking, anchor translation, and target discovery sub-modules. Each sub-module in the framework has been shown to work well both in monolingual settings and English to CJK language pairs. We seek to find out whether the approach that worked very well for English to CJK would still work for CJK to English. We use the same experimental settings that were used in our previous participation, and our experimental runs show that the CJK-to-English CrossLink task is a much harder task when using the same resources as the English-to-CJK one.

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, Languages, Algorithms

Keywords

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

Team Name: [UKP]
Subtasks: [Chinese to English], [Japanese to English], [Korean to English]

1. INTRODUCTION

This paper describes Ubiquitous Knowledge Processing (UKP) Lab's methodology for CrossLink2 task at NTCIR-10. The goal of CrossLink tasks at NTCIR is to discover links among Wikipedia pages across different languages. At NTCIR-9, Anchors in English (En) topic pages were linked to their corresponding Chinese, Japanese, and Korean (CJK) target Wiki pages. The task at NTCIR-10 expands to both directions of language pairs, i.e. En-to-CJK and CJK-to-En.

Figure 1: An illustration of the UKP's approach to the cross-lingual link discovery tasks at NTCIR-10

At NTCIR-9, we have developed a CrossLink framework consisting of anchor selection, anchor ranking, anchor translation, and target discovery sub-tasks. We utilized state-of-the-art monolingual anchor selection, anchor ranking, and target discovery approaches. For anchor translation, we utilized a number of methods that have been widely used for short phrase translation, including Wikipedia interlingual links. UKP's CrossLink system at NTCIR-9 worked very well for En-to-CJK task; in fact our runs performed very competitively compared to other participants' systems.

At NTCIR-10, we try to find out whether the same approach would still work for CJK-to-En language direction. Using the same experimental settings that were used in our previous participation, we find out that ... Our experimental runs show that CJK to English is much harder task when using the same resources as English to CJK.

Regarding our experiments, there were two factors that affected the outcome: First, because we were given no training data to train our CJK-to-English subtasks, our official submissions were not optimized but rather we submitted runs based on arbitrarily selected parameters. Secondly, due to a bug in our system, some of our official submissions were erroneous; we present the differences in our official submissions and and re-runs in the experiment section.

2. OUR APPROACH TO CROSS-LINGUAL LINK DISCOVERY

2.1 Overview

Our participating system at CrossLink2 builds upon the English-to-CJK CrossLink system at NTCIR-9 [2].

As it has been done in our previous work, our anchor text translation-based CrossLink approach consists of three steps: anchor discovery in the source language, anchor translation to target languages, and anchor target discovery in the target languages (Figure 1). Anchor candidates produced by *anchor discovery* module are translated into the target lan-

guage, and the translated anchors are used to discover appropriate documents in the target languages. This approach is built upon the existing monolingual link discovery approaches found in the literature (e.g. [1] and [7]. See [2] for complete references).

2.2 Anchor discovery

Anchor discovery in the source language extracts and measures appropriate anchors from given documents. Of all approaches that were employed in our last participation, we use *Word N-grams* anchor selection method that extracts all word N-grams of size 1∼5. For ranking anchor candidates, we use *Anchor probability* anchor ranking method that measures probability of the given text being used as an anchor text in the source language Wikipedia corpus [3, 4].

$$anchor\ probability(c) = \frac{|\{d|cnt(c, d_{anchor}) > 0\}|}{|\{d|cnt(c, d) > 0\}|}$$

where $cnt(c, d)$ and $cnt(c, d_{anchor})$ are defined as the count of anchor candidate c appearing in a document d and the count of c being used as an anchor in a document d (d_{anchor}).

We chose to use these two methods because in the previous work the performances using them were unmatched by all other methods.

2.3 Anchor Text Translation

For anchor text translation, we employ a number of methods.

- Cross-lingual title pairs in Wikipedia
- Machine translation
- Cascaded

Cross-lingual title pairs in Wikipedia is a method that utilizes the Wikipedia interlingual alignments. As demonstrated in our previous work, translation pairs from Wikipedia interlingual alignments are a good source of translation knowledge, especially for translating anchors from Wikipedia articles. In our En-to-CJK CrossLink task, it was shown that around 50% of anchors in English Wikipedia articles exactly match the title of its target documents.

For *Machine translation* method, we employ the state-of-the-art system available as a web-based service.[1] MT is a very good alternative approach with high coverage.

Cascaded approach combines the two anchor translation methods in a cascaded way. Figure 2a shows a simple flow chart of *Cascaded* method. The order in which methods are applied is determined heuristically; High precision method is considered first, and successive methods are applied only if the prior method fails.

As some anchor translation methods produce N-best translations, a parameter $NumMaxTrans$ (1∼5) was used to set the upper bound on the number of translation candidates.

[1]Google Translate. http://translate.google.com/

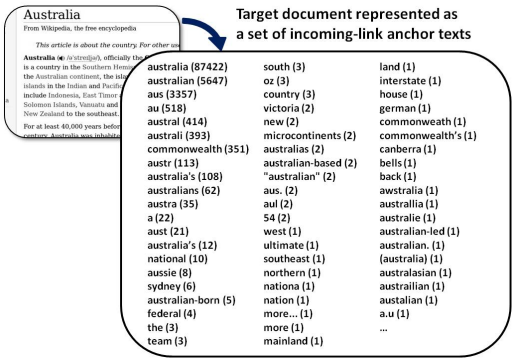


Figure 3: An example of a wiki page represented with incoming link anchor texts. Figure from [2].

2.4 Target discovery

Given anchor candidates in the target language, we use the following methods for target discovery.

- Title match
- Document search
- Incoming link anchor search
- Cascaded

Title match method matches the title of wiki pages in the target language with the translated anchor texts. As in the previous work, no disambiguation is performed When multiple target documents are retrieved.

Document search method utilizes an information retrieval algorithm to rank documents with an anchor text as a query. We used BM25 probabilistic model.

Incoming-link anchor search method also utilizes IR, and rank documents with an anchor text as a query. However, instead of representing documents by the text it contains, the documents are represented with anchor texts of all incoming-links in the Wikipedia corpus. This method is similar to the *target strength* target ranking method [1] that measures the probability of a target document according to the frequency of the anchor text that link to the target documents.

Cascaded method combines *Title match* and *Incoming-link anchor search* methods (Figure 2b).

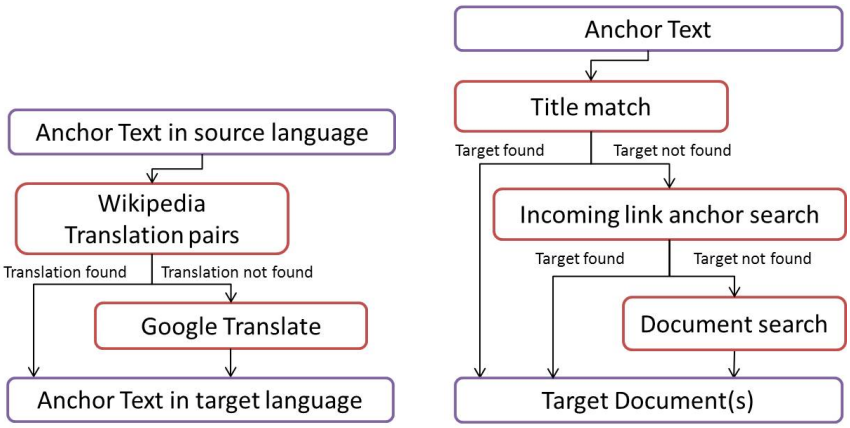
Target discovery methods generate ranked lists of target documents. We use $NumMaxTargets$ parameter to control the number of retrieved target documents.

3. EXPERIMENTS

3.1 Dataset

From twenty five test topics were provided for each source language, we extracted the textual data along with title, section, and category annotations.

Also, Wikipedia collections were provided to the task participants. Though the collections contain automatically assigned semantic annotations [5], they were not utilized in our experiments. To ensure the experimental assumptions,



(a) Anchor Translation (b) Target Discovery

Figure 2: Diagrams of the cascaded methods

we removed all test topic wiki pages from all corpora. The English corpora are analyzed with a POS tagger. For complete information, please see the task overview paper [6].

3.2 Experimental Setup

Our experiments were carried on an existing CrossLink framework which was developed for the NTCIR-9 CrossLink task. The framework consists of a set of UIMA-based pipelines.[2]

3.3 Evaluation methods

Given a gold standard for target documents, CrossLink tasks are evaluated using treceval-like measures such as precision at N retrieved documents (P@N, $N = 5, 10, 20 \ldots 250$), precision at R documents, where R is the number of relevant documents (R-prec), and mean average precision (MAP).

Two Gold standards were provided: the Wikipedia-based one as ground-truth and by pooling with subsequent manual annotation. Original topic documents contain links to other Wiki pages and interlingual links to wiki pages in other languages. Wikipedia ground truth is a set of target Wiki pages that can automatically be deduced using the existing links in the topic documents. Manual assessment gold standards are created from merged formal runs from all participating systems by manual evaluations of the task organizers.

3.4 Submission

The Output of our CrossLink system is a ranked list of anchor texts, and a ranked list of target documents for each anchor texts. As specified by the task definition, the submission file was created with at most 250 anchor texts sorted by anchor scores, and for each anchor text either one, three, five target documents were selected. Our submission runs emphasize anchor discovery over target discovery, by first ordering retrieved target documents based on the anchor text scores, then by the target scores.

Because there were no training topics available for CJK-to-En CrossLink task, methods and parameters for official submissions were not optimized, but rather they were selected arbitrarily by hand. Three official runs were created using the following settings: we used cascaded methods for anchor translation and target discovery, $NumMaxTrans$ was fixed to 5, while $NumMaxTargets$ was set to 1∼5.

1. *Word N-gram*, *Anchor Probability*, *Cascaded*, *Cascaded* ($NumMaxTrans$=5, $NumMaxTargets$=1)
2. *Word N-gram*, *Anchor Probability*, *Cascaded*, *Cascaded* ($NumMaxTrans$=5, $NumMaxTargets$=3)
3. *Word N-gram*, *Anchor Probability*, *Cascaded*, *Cascaded* ($NumMaxTrans$=5, $NumMaxTargets$=5)

Details on the results of our official submission, as well as comparison to other competing systems, can be found in the NTCIR-10 CrossLink2 overview paper.[6].

3.5 Post-Submission Experiments

3.5.1 *Bug fix in official runs (Ko-to-En)*

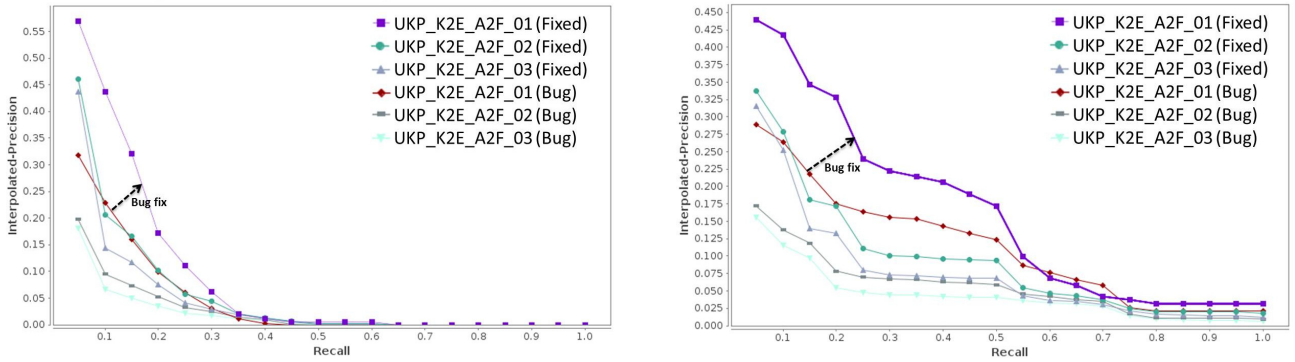
After the submission of official runs, we discovered a bug in the anchor extraction submodule for the Chinese, Japanese, and Korean languages, in which multi-word anchor candidates were not extracted. Figure 4 illustrates the effect of bug fix in the official submissions.

3.5.2 *Method combinations*

To find the best configuration of the CrossLink system, we carried out experiments with different combinations of subtask methods and $NumMaxTrans$ and $NumMaxTargets$ parameter values. Methods for each subtask were determined while methods for the rest of the subtasks are fixed. For example, different anchor translation methods were evaluated while target discovery method is fixed to *Title Match*. $NumMaxTrans$ and $NumMaxTargets$ can have values of 1, 3, or 5. Figures 5 and 6 show the results of the methods for each of the CrossLink subtasks. Note that methods have different effectiveness in each language, due to difference in size and performance of resources.

In general, *Wikipedia Translation Pair* performed the best for anchor translation task, and *Title Match* for target discovery task. Unlike En-to-CJK language direction, in which

[2]DKPro. http://www.ukp.tu-darmstadt.de/research/current-projects/dkpro/



(a) Wikipedia Ground Truth

(b) Manual Assessment

Figure 4: Interpolated precision-recall curve of UKP's Korean-to-English official submissions and bug-fixed runs

Cascaded methods performed on par or better than single methods, *Cascaded* method did not perform well in CJK-to-En direction.

We also present the overall performance of the official submissions and the best parameter configuration in Figures 7 and 8. With *Wikipedia Ground Truth* evaluation, the optimal performances were achieved using the following set of parameters.

- Japanese: *Cascaded*, *Cascaded* (*NumMaxTrans*=5, *NumMaxTargets*=1)
- Korean: *Word N-gram*, *Anchor Probability*, *Cascaded*, *Cascaded* (*NumMaxTrans*=5, *NumMaxTargets*=3)
- Chinese: *Word N-gram*, *Anchor Probability*, *Cascaded*, *Cascaded* (*NumMaxTrans*=5, *NumMaxTargets*=5)

Similar set of parameters are used when evaluating with *Manual Assessment*.

- Japanese: *Cascaded*, *Cascaded* (*NumMaxTrans*=5, *NumMaxTargets*=1)
- Korean: *Word N-gram*, *Anchor Probability*, *Cascaded*, *Cascaded* (*NumMaxTrans*=5, *NumMaxTargets*=3)
- Chinese: *Word N-gram*, *Anchor Probability*, *Cascaded*, *Cascaded* (*NumMaxTrans*=5, *NumMaxTargets*=5)

4. CONCLUSION

For the NTCIR-10 CrossLink2 task, we developed a CJK-to-English cross-lingual link discovery system based on our English-to-CJK system. As demonstrated, our system mostly utilizes language-independent methods and it can be easily adapted to different language pairs and directions. We presented the effectiveness of various approaches of subtasks in CrossLink task using different gold standards. As our experimental results show, link discovery performance for different language pairs and directions vary from task to task.

5. ACKNOWLEDGMENTS

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806, and by the German Research Foundation under grant 798/1-5.

6. REFERENCES

[1] N. Erbs, T. Zesch, and I. Gurevych. Link discovery: A comprehensive analysis. In *Proceedings of the 5th IEEE International Conference on Semantic Computing (IEEE-ICSC)*, Palo Alto, CA, USA, Jul 2011.

[2] J. Kim and I. Gurevych. Ukp at crosslink: Anchor text translation for cross-lingual link discovery. In *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*, pages 487–494, NII, Tokyo, December 2011.

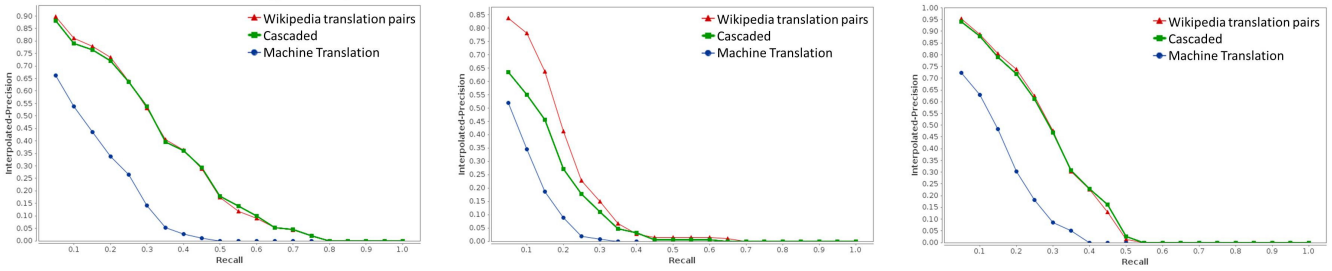
[3] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 233–242, New York, NY, USA, 2007. ACM.

[4] D. Milne and I. H. Witten. Learning to link with wikipedia. In *Proceeding of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 509–518, New York, NY, USA, 2008. ACM.

[5] R. Schenkel, F. Suchanek, and G. Kasneci. YAWN: A semantically annotated Wikipedia XML corpus. In A. Kemper, H. Schöning, T. Rose, M. Jarke, T. Seidl, C. Quix, and C. Brochhaus, editors, *12. GI-Fachtagung für Datenbanksysteme in Business, Technologie und Web (BTW 2007)*, volume 103 of *Lecture Notes in Informatics*, pages 277–291, Aachen, Germany, 2007. Gesellschaft für Informatik.

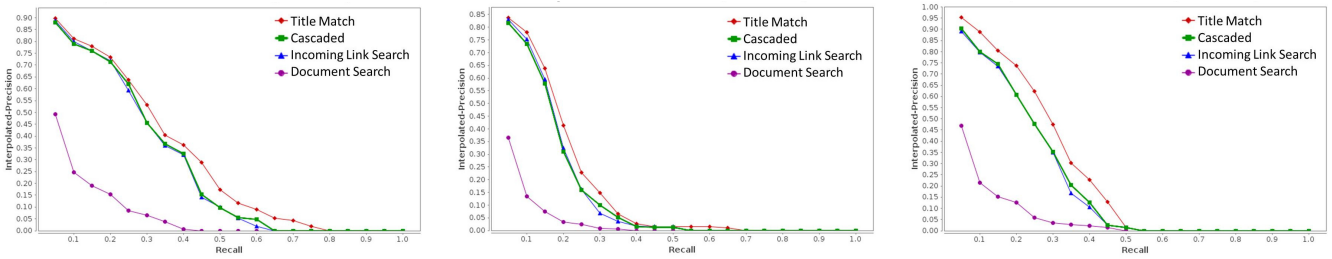
[6] L.-X. Tang, I.-S. Kang, F. Kimura, Y.-H. Lee, A. Trotman, S. Geva, and Y. Xu. Overview of the NTCIR-10 cross-lingual link discovery task. In *Proceedings of the Tenth NTCIR Workshop Meeting*, page to appear, NII, Tokyo, June 2013.

[7] A. Trotman, D. Alexander, and S. Geva. Overview of the INEX 2010 link the wiki track. In S. Geva, J. Kamps, R. Schenkel, and A. Trotman, editors, *Comparative Evaluation of Focused Retrieval*, volume 6932 of *Lecture Notes in Computer Science*, pages 241–249. Springer Berlin / Heidelberg, 2011.



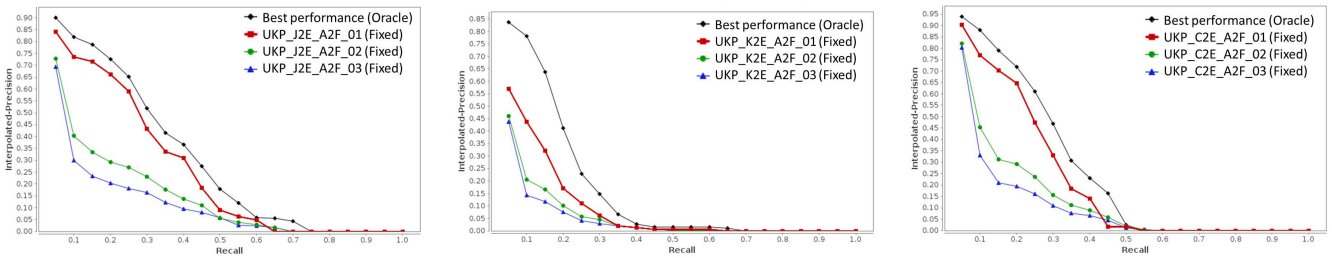
(a) Japanese (b) Korean (c) Chinese

Figure 5: Interpolated precision-recall curves of best parameter configuration for *Anchor Translation* Method, evaluated with *Wikipedia Ground Truth* gold standard. (anchor extraction: *N-gram*, anchor ranking: *Anchor Probability*, target discovery: *Title Match*)



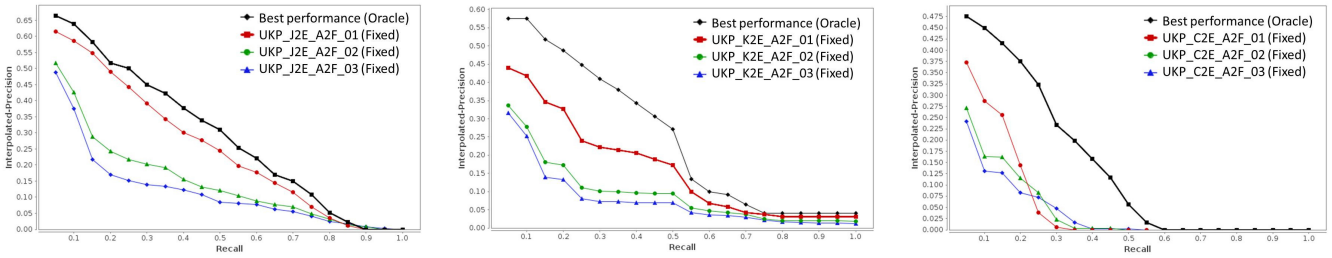
(a) Japanese (b) Korean (c) Chinese

Figure 6: Interpolated precision-recall curves of best parameter configuration for *Target Discovery* Method, evaluated with *Wikipedia Ground Truth* gold standard. (anchor extraction: *N-gram*, anchor ranking: *Anchor Probability*, target discovery: *Wikipedia Translation Pairs*)



(a) Japanese (b) Korean (c) Chinese

Figure 7: Interpolated precision-recall curves comparing official submissions (fixed) vs. best parameter configuration, evaluated with *Wikipedia Ground Truth* gold standard.



(a) Japanese (b) Korean (c) Chinese

Figure 8: Interpolated precision-recall curves comparing official submissions (fixed) vs. best parameter configuration, evaluated with *Manual Assessment* gold standard.