UKP at CrossLink: Anchor Text Translation for Cross-lingual Link Discovery

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

ABSTRACT

This paper describes UKP's participation in the cross-lingual link discovery (CLLD) task at NTCIR-9. The given task is to find valid anchor texts from a new English Wikipedia page and retrieve the corresponding target Wiki pages in Chinese, Japanese, and Korean languages. We have developed a CLLD framework consisting of anchor selection, anchor ranking, anchor translation, and target discovery subtasks, and discovered anchor texts from English Wikipedia pages and their corresponding targets in Chinese, Japanese, and Korean languages. For anchor selection, anchor ranking, and target discovery, we have largely utilized the state-of-the-art monolingual approaches. For anchor translation, we utilize a translation resource constructed from Wikipedia itself in addition to exploring a number of methods that have been widely used for short phrase translation. Our formal runs performed very competitively compared to other participants' systems. Our system came first in the English-2-Chinese and the English-2-Korean F2F with manual assessment and A2F with Wikipedia ground truth assessment evaluations using Mean-Average-Precision (MAP) measure.

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, [UKP], [English to Chinese CLLD], [English to Japanese CLLD], [English to Korean CLLD], [English Wikipedia], [Google Translate], [StarDict Dictionary], [Stanford Chinese Segmenter], [Stanford POS Tagger], [Stanford Named Entity Recognizer], [TreeTagger], [MeCab], [KoMA], [DKPro]

1. INTRODUCTION

The web distinguishes itself from other types of document collections by its inter-connectedness; web documents contain hyperlinks that connect to other documents that are related and/or informative with regard to the context and the topic of the source document.

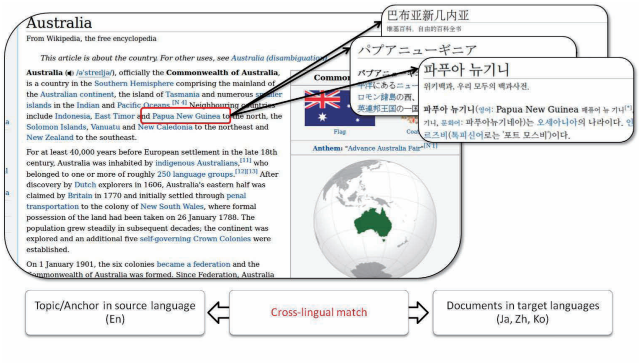
Australia
From Wikipedia, the free encyclopedia
巴布亚新几内亚
パプアニューギニア
파푸아 뉴기니
Topic/Anchor in source language (En)
Cross-lingual match
Documents in target languages (Ja, Zh, Ko)

Figure 1: An illustration of the cross-lingual link discovery task

Wikipedia[1] is such a document collection, where wiki pages are heavily inter-connected to other wiki pages as well as external web resources. Contributors to Wikipedia are encouraged to provide sources of information as detailed as possible and clarify contexts with reference to disambiguated concepts. This characteristic makes Wikipedia a very much valued source of information. However, adding a new wiki page to the collection poses a difficult problem, since the contributor is faced with the challenge of providing a set of informative hyperlinks by searching the entire Wikipedia collection, not to mention the web.

Link discovery emerged as a fairly recent research topic, and the outcome has been successfully utilized in applications such as assisting the authors in the English language with finding potential anchors and target documents. Cross-lingual link discovery (CLLD) further extends this functionality; it aims to find links between wiki pages of different languages and to enable an easier navigation method to the vast amount of multilingual knowledge.

When the scope of the task expands from monolingual to multilingual, it becomes more difficult due to the increase in the document search space and the number of languages to analyze. Above all, the most challenging aspect is due to the additional obstacle of matching across languages, which requires additional step of translation process that results in translation ambiguities (Figure 1).

Similarly to the link-the-wiki tasks at INEX, CLLD at

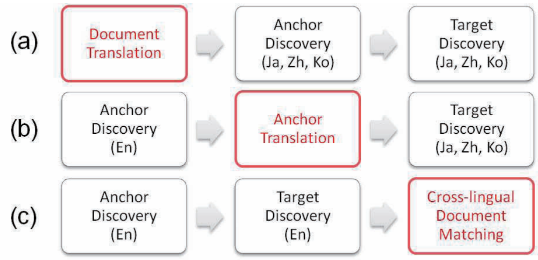
(a)
Document Translation
Anchor Discovery (Ja, Zh, Ko)
Target Discovery (Ja, Zh, Ko)
(b)
Anchor Discovery (En)
Anchor Translation
Target Discovery (Ja, Zh, Ko)
(c)
Anchor Discovery (En)
Target Discovery (En)
Cross-lingual Document Matching

Figure 2: Three different approaches to cross-lingual link discovery with respect to the order in which cross-lingual matching occurs

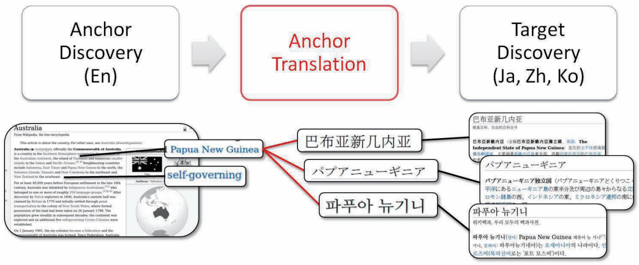

Figure 3: An illustration of the anchor text translation approach to cross-lingual link discovery

NTCIR-9 targets at discovering links among Wiki pages in Wikipedia. Specifically, English topic pages were provided, for which anchors and their corresponding Chinese, Japanese, and Korean target Wiki pages are to be discovered.

This paper describes Ubiquitous Knowledge Processing (UKP) Lab's methodology for CLLD task at NTCIR-9.

2. DIFFERENT APPROACHES TO CROSS-LINGUAL LINK DISCOVERY

UKP's CLLD system builds upon the existing monolingual link discovery approach. To extend the monolingual method to cross-lingual one, it is necessary to introduce a cross-lingual matching step in the "anchor discovery" and "target discovery" stemps of the monolingual link discovery framework.

Considering at which stage to carry out the cross-lingual matching, three different approaches to CLLD are possible (Figure 2).

The first approach is to translate the document in the source language into the target languages, then carry out link discovery tasks in the target languages (Figure 2(a)). The second one is to translate the anchor candidates produced by "anchor discovery" module into the target languages, and discover appropriate documents in the target languages (Figure 2(b)). The third option is to fully carry out the monolingual link discovery task, which generates a list of target documents in the source language, and find documents in the target languages that are equivalent to the retrieved documents (Figure 2(c)).

Different cross-lingual matching methods are needed for each of the three possible CLLD approaches: document translation, anchor translation, and cross-lingual document similarity measurement. These methods have different levels of difficulty and resource requirements.

For translating documents, one can employ a machine translation (MT) system. Though the quality of MT has improved much over time, its quality is yet to catch up with that of the manual translation. Therefore, its output may be insufficient for automatic language analysis. Once the document is translated, the rest of the steps is the same as a monolingual link discovery task, but in the target languages. In such a case, any existing link discovery approach is suitable if necessary resources for the target language are available.

Measuring similarity in cross-lingual settings poses an interesting problem. One may consider it as cross-lingual information retrieval[8] and search for documents in the target languages using the source language document as a query, or one may use various cross-lingual text similarity measures[11, 5, 3, 14].

Given limited resources and time for the task participation, we explore only the second approach, namely the anchor text translation (Figures 2(b) and 3). This approach can easily build upon the previously studied monolingual link discovery approaches[2, 13]. Also, translating a short phrase-like text is an easier and less resource-intensive task than translating an entire document, and allows more translation options such as a bilingual dictionary or a parallel corpus.

3. ANCHOR TEXT TRANSLATION-BASED APPROACH

3.1 Overview

The anchor text translation-based CLLD 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 3). For each component of the framework, we test out a number of methods to evaluate the best configuration for the CLLD task.

A recent survey on link discovery[2] provides detailed descriptions for most approaches used in our system; we limit the scope of this paper to providing the details on new and modified methods. Though we try our best to give reference to related work, we kindly point to the work by [2] for comprehensive overview of the research problem.

3.2 Anchor discovery

Anchor discovery is a two-step process: first, anchor candidates are extracted from the topic document, then the candidates are scored by their "anchorness".

3.2.1 Anchor selection method

Below is the list of anchor selection methods used in our system.

- Noun phrases
- Named entities
- Anchor texts observed in training data
- Titles observed in training data
- Titles observed in topic document
- Word N-grams (N:1∼5)

Noun phrases and named entities in the target documents are automatically identified using text analysis tools (see Section 4.1.1).

Also, we employ a method where any sequence of words that have been observed as anchor text or title in the training data are considered as anchor candidates, as well as title sections in the source document.

Titles of wiki pages and wiki page sections also make good candidates for anchor texts, and we employ two methods each using the training and the test data.

The last method extracts all word N-grams of size 1∼5. Note that the set of anchor candidates produced by this method subsumes most candidates produced by previous methods. Unlike other anchor selection methods, not all word n-grams are good candidates for anchor texts, and a subsequent filtering step is thus necessary.

3.2.2 Anchor ranking method

Anchor candidates are assigned scores according to their "anchorness". In previous work, a number of approaches utilizing textual, title, and link knowledge have been used. Below is a list of methods we tested in our experiments.

- IR model (BM25 [9]) tf·idf score using statistics from all titles in the English Wikipedia corpus
- IR model (BM25) tf·idf score using statistics from all anchor texts in the English Wikipedia corpus
- IR model (BM25) tf·idf score using statistics from all documents in the English Wikipedia corpus
- Anchor probability [6, 7]: probability of the given text being used as an anchor text in the English Wikipedia corpus
- Anchor strength [4]: probability of the given text being used as an anchor text to its most frequent target in the English Wikipedia corpus

We used the equation below for a BM25 score of a word,

$$BM25(w,d) = ln\frac{N - df(w) + 0.5}{df(w) + 0.5} \cdot \frac{(k_1 + 1)\cdot c(w,d)}{k_1\left((1-b) + b\frac{|d|}{avgdl}\right) + c(w,d)}$$

where $c(w,d)$ is the frequency of w in d, $\mid d \mid$ is the number of unique terms in d, $avgdl$ is the average $\mid d \mid$ of all documents, N is the number of documents in the collection, $df(w)$ is the number of documents with w, C is the entire collection, and k_1 and b are constants 2.0 and 0.75. For multiword anchor candidates, tf·idf scores for each word were added together.

For measures using anchor text statistics, the equations are as follows,

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

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

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}).

There are two differences between the two measures. Unlike *anchor probability*$(\cdot)$, *anchor strength* measure only considers the frequency of the anchor text which links to its most frequent target to take into account how "ambiguous" an anchor text is. Also, *anchor strength*$(\cdot)$ is not a probability as its sum over d does not add up to 1.0.

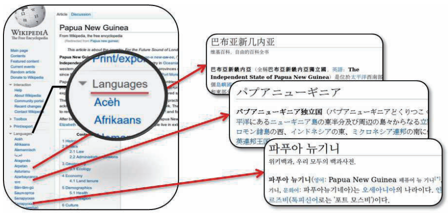

Figure 4: An example of interlingual alignments in Wikipedia

Table 1: Statistics of interlingual alignment between English (En) and Chinese (Zh), Japanese (Ja), and Korean (Ko)

Language direction	Count	Coverage
Ja → En	290,217	40.5% of Ja, 10.9% of En
Zh → En	186,872	59.1% of Zh, 7.01% of En
Ko → En	89,215	44.3% of Ko, 3.35% of En

3.3 Anchor Text Translation

For anchor text translation, we also employ a number of methods of different nature, to investigate their suitability for the task.

- No translation
- Bilingual dictionary
- Machine translation
- Cross-lingual title pairs in Wikipedia
- Cascaded

No translation simply indicates that the anchor text in the source language is used as it is to find the documents in the target languages. For non-English documents, English is often used as an additional description of the topic and within the document for the completeness of the provided information and as a means to disambiguate the concepts. We do not expect this approach to work well, but tested to obtain a performance bottom line.

Bilingual dictionary is a simple look-up method, given a bilingual dictionary between English and the target languages.[1] First, an anchor text is looked up in the dictionary. If it is found, we select the first word in the first sense of the entry as translation. If the anchor text is not in the dictionary, we lemmatize it,[2] then repeat the search. The resources for this method are relatively easy to obtain and even some freely available dictionaries have good coverage. However, we anticipate that coverage for anchor texts that are phrases or named entities will be low.

For *Machine translation* method, we employ the state-of-the-art system available as a web-based service.[3] We expect that MT systems provide better accuracy than dictionary-based, and also have better coverage, as they use contextual information when translating and the state-of-the-art systems are trained based on large-scale training data.

Cross-lingual title pairs in Wikipedia is a method that utilizes the interlingual alignments in the Wikipedia (Figure 4).

[1] quick_english-korean, quick_eng-zh_CN, and JMDict from StarDict licensed under GPL and EDRDG. http://stardict.sourceforge.net/

[2] JWI. http://projects.csail.mit.edu/jwi/

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

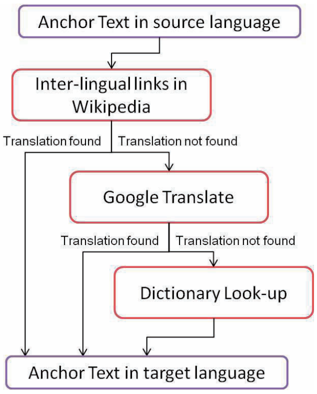

Figure 5: A diagram of the cascaded anchor translation method

Interlingual alignments in Wikipedia is a good source of translation knowledge; the title pairs of Wiki pages from interlingual alignments can be regarded as manual translations. This resource is particularly useful for translating anchor texts that match the exact forms of Wiki page titles to which they anchor. Among 105,375,981 anchors in the English Wikipedia corpora described in Section 4.1.2, 51,961,666 (49.3%) match exactly to the title of its target document. The coverage of interlingual alignment is also substantial (Table 1).

Cascaded approach, as its name suggests, combines the different anchor translation methods in a cascaded way. Figure 5 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 $NumMaxTrans$ was used to set the upper bound on the number of translation candidates.

3.4 Target discovery

To discover Wiki pages in the target languages using the translated anchor texts, we applied the following methods:

- Title match
- Incoming link anchor search

Title match method finds wiki pages in the target languages whose title exactly matches the translated anchor texts. When multiple target documents are retrieved, no disambiguation was carried out for ranking the target document, and the first document is chosen naively.

Incoming-link anchor search method also utilizes an IR system to search and rank target documents with an anchor text as a query; This approach differs from *Document search* in that the target documents are represented not by the text it contains, but with anchor texts of all incoming-links (Figure 6). This method is a variant of the "target strength" target ranking method [2], where the number of occurrences of the anchor text linking to target documents is used for measuring the probability of the target document given the anchor text.

Target discovery methods generate ranked lists of target documents; We set aside a parameter $NumMaxTargets$ to control the maximum number of target documents.

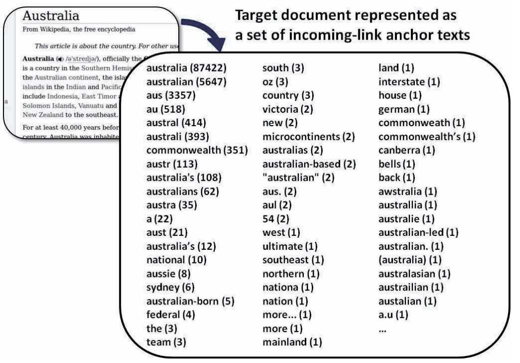

Figure 6: A wiki page represented with incoming link anchor texts. An example shown with a Wiki page in English for the demonstration purpose.

Table 2: Wikipedia collections for Simplified Chinese (Zh), Japanese (Ja), Korean (Ko), and English (En); Numbers of documents (# of docs) were counted before removing topic documents. Size is indicated when compressed.

	Zh	Ja	Ko	En
# of docs	316,251	715,911	201,512	2,666,190
file size	381 MB	1,139 MB	163 MB	5,552 MB

4. EXPERIMENTS

4.1 Dataset

4.1.1 Wikipedia Topics

Three training and twenty five test topics were provided to the participants.

Training (3): Australia, Femme Fatale, Martial arts
Test (25): Ivory, Kim Dae-jung, Croissant, Kiwifruit, Kimchi, Jade, Boot, Cuttlefish, Mohism, Fiat money, Crown prince, Pasta, Zhu Xi, Source code, Sushi, Spam (food), African Wild Ass, Credit risk, Asian Games, Oracle bone script, Cuirassier, Dew point, Cretaceous, Abdominal pain, Puzzle

Though the provided topics were enriched with semantic annotations, we only extracted the textual data along with title, section, and category annotations. Both sets of topics were POS-tagged and chunked with TreeTagger,[4] and named entities were annotated with Stanford Named Entity Recognizer.[5]

4.1.2 Wikipedia Corpora

Chinese, Japanese, and Korean Wikipedia[6] were provided to the task participants (Table 2), which were converted from Wiki to XML format with an automatic semantic annotator [10]. In addition, we utilized English Wikipedia, which was used for the Link-the-Wiki tasks at INEX.[7]

Training and test topic wiki pages were removed from all corpora.

[4] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[5] http://nlp.stanford.edu/software/CRF-NER.shtml
[6] Dump date: June, 2010
[7] INEX 2009 Collection. Dump date: October, 2008 http://www.mpi-inf.mpg.de/departments/d5/software/inex/

The corpora are analyzed with POS taggers for each language; English with TreeTagger, Chinese with Stanford Chinese segmenter and POS tagger,[8] Japanese with MeCab,[9] and Korean with KoMA.[10]

4.2 Experimental Setup

Our experiments, from preprocessing training and test data to configuring the subcomponents of the CLLD framework, were carried out with the UIMA-based DKPro framework.[11] The framework already includes the libraries and interfaces to most of the NLP tools mentioned in the paper, as well as easy integration methods for new components.

To find the best configuration of the CLLD system, the methods for each subcomponent as well as parameter values for $NumMaxTrans$ and $NumMaxTargets$ need to be determined.

Methods for each subtask were determined while methods for the rest of the subtasks are fixed. For example, different anchor selection methods were evaluated while methods for anchor ranking, anchor translation, and target discovery are fixed. Values for the parameters $NumMaxTrans$ and $NumMaxTargets$ were tested with 1 and 3.

Parameters tuning and method selection were carried out on the training data for the Japanese target discovery task. The best configuration on the data was used in the formal runs of the Chinese, Japanese, and Korean target discovery tasks.

4.3 Submission

The Output of our CLLD system is a ranked list of anchor texts, and a ranked list of target documents for each anchor texts. As speficied 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 or three target documents were selected.

4.4 Evaluation methods

Systems participating at NTCIR-9 CLLD task were evaluated for finding good target documents. Specifically, given a gold standard for target documents, 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) are used.

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, as illustrated in Figure 7.

After the formal runs from all participating systems had been submitted, the results were merged and manually evaluated by the task organizers. First, anchor texts judged as invalid by humans were filtered out, then target documents for the valid anchor texts were determined.

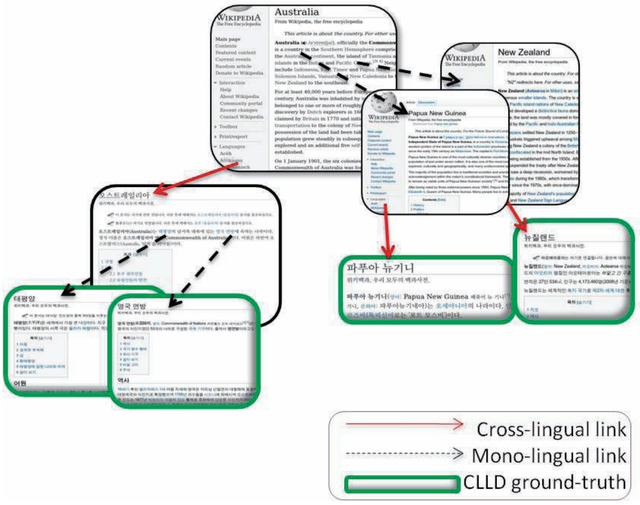

Figure 7: (Ground-truth) Gold standard for automatic evaluation

4.5 Results

4.5.1 Experiments on training data

Due to the paper length constraints, we present our system's results for the English-Japanese language pair only.

Figures 8, 9, 10, 11 and Tables 3, 4, 5, 6 show the results of the methods for each of the CLLD subtasks.

For anchor selection and ranking, the combination of *Word N-gram* and *Anchor Probability* performed the best. *Title Match* target discovery method works better than the *Incoming Link Search*.

Among the anchor translation methods, *Wikipedia translation pairs* out-performed the rest, and *Cascaded* method further improves the performance.

Parameters anchor selection, anchor ranking, anchor translation, target discovery for five official submissions were thus selected as shown below:

1. *Word N-gram*, *Anchor Probability*, *Cascaded*, *Title match* ($NumMaxTrans$=1, $NumMaxTargets$=1)
2. *Word N-gram*, *Anchor Probability*, *Cascaded*, *Title match* ($NumMaxTrans$=3, $NumMaxTargets$=3)
3. *Word N-gram*, *Anchor Probability*, *Wikipedia translation pairs*, *Title match* ($NumMaxTrans$=1, $NumMaxTargets$=1)
4. *Word N-gram*, *Anchor Probability*, *Wikipedia translation pairs*, *Cascaded* ($NumMaxTrans$=3, $NumMaxTargets$=1)
5. *Word N-gram*, *Anchor Probability*, *Cascaded*, *Incoming link search* ($NumMaxTrans$=1, $NumMaxTargets$=1)

4.5.2 Experiments on test data

Details on the results of our official submission, as well as comparison to other competing systems, can be found in the NTCIR-9 CLLD overview paper.[12]

Here we provide only the summary. There were seven groups participating in the English-to-Chinese (E-C) task, four in the English-to-Japanese one (E-J), and six in the English-to-Korean one (E-K). Overall, our submissions turned out to be competitive to those of other systems. For the au-

[8] http://nlp.stanford.edu/software/segmenter.shtml http://nlp.stanford.edu/software/tagger.shtml

[9] http://mecab.sourceforge.net/

[10] Korean Morphological Analyzer. http://kle.postech.ac.kr/

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

Table 3: Performance of anchor selection methods on training data. (Target language: Japanese; Anchor translation: Wikipedia translation pairs; Anchor ranking: anchor probability; Target discovery: Title match.)

	MAP	R-Prec	P5	P10	P20	P30	P50	P250
Word N-gram	0.398	0.458	0.867	0.767	0.817	0.856	0.873	0.571
Anchor in corpus	0.286	0.438	0.067	0.133	0.333	0.400	0.580	0.528
Title in corpus	0.286	0.436	0.067	0.133	0.333	0.411	0.587	0.525
Named Entity	0.232	0.272	0.800	0.900	0.917	0.878	0.833	0.325
Noun Phrase	0.016	0.018	0.267	0.300	0.233	0.156	0.093	0.019
Title in topic	0.001	0.003	0.133	0.067	0.000	0.022	0.013	0.003

Table 4: Performance of anchor ranking methods on training data. (Target language: Japanese; Anchor selection: Word N-gram; Anchor translation: Wikipedia translation pairs; Target discovery: Title match.)

	MAP	R-Prec	P5	P10	P20	P30	P50	P250
Anchor Probability	0.398	0.458	0.867	0.767	0.817	0.856	0.873	0.571
Anchor Strength	0.378	0.446	0.600	0.700	0.800	0.833	0.867	0.537
Text Search (Document)	0.225	0.336	0.867	0.833	0.717	0.700	0.667	0.393
Text Search (Anchor)	0.182	0.321	0.667	0.600	0.633	0.622	0.593	0.384
Text Search (Title)	0.158	0.315	0.533	0.600	0.500	0.522	0.493	0.369

Table 5: Performance of anchor translation methods on training data. (Target language: Japanese; Anchor selection: Word N-gram; Anchor ranking: anchor probability; Target discovery: Title match.)

	MAP	R-Prec	P5	P10	P20	P30	P50	P250
Cascaded	0.403	0.469	0.867	0.800	0.767	0.822	0.847	0.583
Wikipedia translation pairs	0.391	0.458	0.800	0.733	0.800	0.844	0.867	0.573
Machine translation	0.175	0.273	0.733	0.667	0.683	0.600	0.620	0.385
Bilingual Dictionary	0.013	0.024	0.333	0.333	0.383	0.311	0.227	0.047
No Translation	0.003	0.009	0.133	0.167	0.117	0.078	0.060	0.012

Table 6: Performance of target discovery methods on training data. (Target language: Japanese; Anchor selection: Word N-gram; Anchor ranking: anchor probability; Anchor translation: Wikipedia translation pairs.)

	MAP	R-Prec	P5	P10	P20	P30	P50	P250
Title Match	0.403	0.469	0.867	0.800	0.767	0.822	0.847	0.583
Incoming Link Search	0.258	0.407	0.400	0.400	0.450	0.522	0.613	0.496

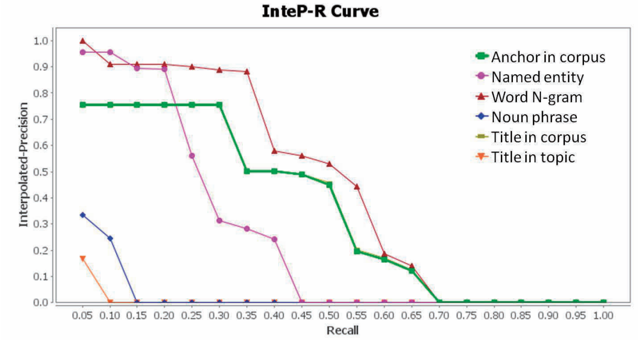

Figure 8: Precision-recall curve of anchor selection methods on training data. (Target language: Japanese; Anchor translation: Wikipedia translation pairs; Anchor ranking: anchor probability; Target discovery: Title match.)

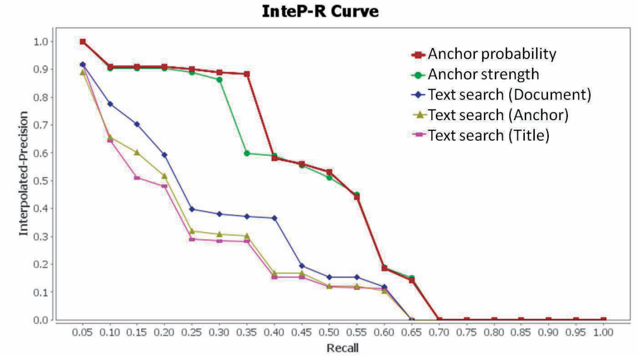

Figure 9: Precision-recall curve of anchor ranking methods on training data. (Target language: Japanese; Anchor selection: Word N-gram; Anchor translation: Wikipedia translation pairs; Target discovery: Title match.)

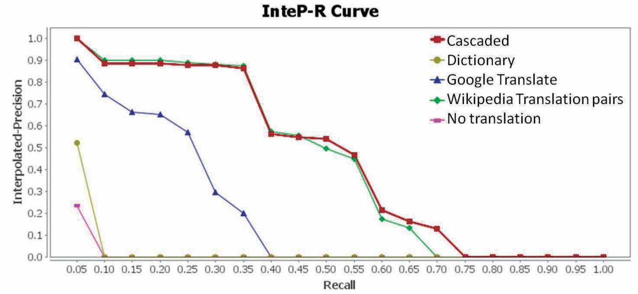

Figure 10: Precision-recall curve of anchor translation methods on training data. (Target language: Japanese; Anchor selection: Word N-gram; Anchor ranking: anchor probability; Target discovery: Title match.)

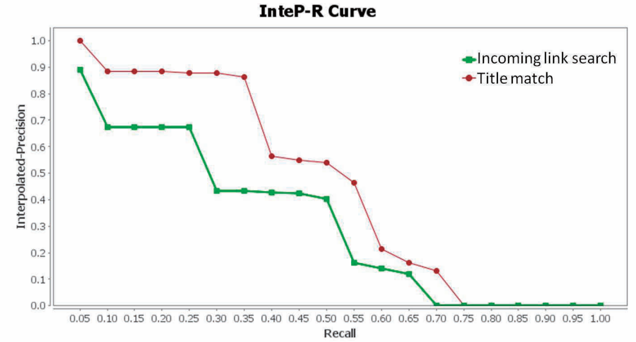

Figure 11: Precision-recall curve of target discovery methods on training data. (Target language: Japanese; Anchor selection: Word N-gram; Anchor ranking: anchor probability; Anchor translation: Wikipedia translation pairs.)

tomatic evaluation of File-to-File task using the Wikipedia-based ground truth, our runs came in 2nd in E-C and E-J and 3rd for E-K in MAP and R-Prec measures. In the File-to-File manual evaluation, our runs came in 1st in E-C and E-K for MAP and R-Prec measures, and 2nd in E-J using MAP, R-Prec and Precision-at-5 measures (P@5). Our runs further improve on the Anchor-to-File manual evaluation; in E-C and E-K, our runs ranked 1st in MAP and R-Prec and 3rd in P@5, and in E-J, 2nd in MAP, R-Prec, and P@5.

5. DISCUSSION

Analysis of method combinations.

In our experimental settings, the combination of *Word N-gram* anchor selection, *Anchor probability* anchor ranking, *cascaded* or *Wikipedia translation pair* anchor translation, and *Title match* target discovery produced the best performance. The observation is consistent across various settings on training and test topics.

The *Word N-gram* and *Anchor probability* anchor discovery methods and *Title match* target discovery methods have proven their effectiveness in the monolingual link discovery tasks [2]. As our approach builds upon the monolingual approach, it is natural that the best approaches for the monolingual task perform well on top of our cross-lingual framework.

As for anchor translation methods, the outstanding difference in the performance of *Wikipedia translation pair* method seems to be due to the fact that it is a set of high-quality manual translations and that the dataset for our task is Wikipedia, from which the translation resource is extracted. Automatic machine translation method *Google Translate* produces quite good results, but its quality is not as good as the manual translation; with *Wikipedia translation pairs* as gold standard, *Google Translate* achieves 37.55% accuracy and 99.52% coverage on the Japanese data, 37.17% and 99.68% on the Chinese one, and 46.96%/99.91% on translating the Korean anchor texts.

Impact of method combinations.

Though it is evident that the link knowledge-based link discovery approaches out-perform textual knowledge-based approaches, in general, such link knowledge is available for the targeted domain. One may need to discover links for a set of documents that do not have any existing links. In the text knowledge-only approaches, the combination of *Named Entity* and *Document search* anchor discovery with *Title match* target discovery methods performs the best.

Also, the end application of the link discovery may affect on which measure the system needs to be evaluated. If the entire process of link discovery is to be done in an automatic fashion, precision should be the key evaluation measure. But if the application is to provide an interactive user interface for providing suggestions to the writer, recall and the run time may be the most important aspects.

Automatic vesus manual evaluation.

Compared to competitive systems from other task participants, our system performs better on manual evaluation than on the Wikipedia-based ground-truth and on the Anchor-to-File evaluation than the File-to-File one. This may be due to the fact that our system utilizes anchor discovery methods from the state-of-the-art monolingual link discovery. Also, 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.

6. CONCLUSION

For NTCIR-9 CLLD, UKP developed an English-to-Chinese, English-to-Japanese, and English-to-Korean cross-lingual link discovery system on top of the state-of-the-art monolingual link discovery. Our system utilizes language-independent methods other than some preprocessing steps, and it can be easily adapted to different language pairs. We have analyzed the effectiveness of various methods for anchor translation as well as anchor and target discovery based on different evaluation measures and gold standards.

As the participants' results on test topics show, link discovery performance for different language pairs varied from task to task. In future work, it may be interesting to explore how to integrate language-dependent as well as to investigate the opposite direction of cross-lingual links (e.g. English Wiki page discovery with Korean topics) and features to further increase the performance.

7. 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. We thank Nicolai Erbs, Benjamin Herbert and Richard Eckart de Castilho for providing software components utilized in these experiments.

8. REFERENCES

[1] Wikipedia, the free encyclopedia. http://www.wikipedia.org/.

[2] 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.

[3] S. Hassan and R. Mihalcea. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1201, Singapore, August 2009. Association for Computational Linguistics.

[4] K. Y. Itakura and C. L. A. Clarke. University of waterloo at inex2007: Adhoc and link-the-wiki tracks. In *INEX'07*, pages 417–425, 2007.

[5] W. Kim and S. Khudanpur. Cross-lingual latent semantic analysis for LM. In *Proc. of ICASSP*, pages 257–260, 2004.

[6] 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.

[7] 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.

[8] J. Nie. *Cross-Language Information Retrieval.* Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2010.

[9] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006.

[10] 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.

[11] R. Steinberger, B. Pouliquen, and J. Hagman. Cross-lingual document similarity calculation using the multilingual thesaurus EUROVOC. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, pages 415–424, London, UK, UK, 2002. Springer-Verlag.

[12] L.-X. Tang, S. Geva, A. Trotman, Y. Xu, and K. Itakura. Overview of the NTCIR-9 crosslink task: Cross-lingual link discovery. In *Proceedings of the Ninth NTCIR Workshop Meeting*, page to appear, NII, Tokyo, December 2011.

[13] 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.

[14] Y. Xia, T. Zhao, J. Yao, and P. Jin. Measuring Chinese-English cross-lingual word similarity with HowNet and parallel corpus. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 6609 of *Lecture Notes in Computer Science*, pages 221–233. Springer Berlin / Heidelberg, 2011.