NTHU at NTCIR-10 CrossLink-2: An Approach toward Semantic Features

Yu-Lan Liu, Joanne Boisson, Jason S. Chang
Natural Language Processing Lab.,
National Tsing Hua University,
Taiwan
{ikulan12, joanne.boisson}@gmail.com, jschang@cs.nthu.edu.tw

ABSTRACT
This paper describes the approaches of NTHU in the NTCIR-10 Cross-Lingual Link Discovery task, also named CrossLink-2. In this task, we aim to discover valuable anchors in Chinese, Japanese or Korean (CJK) articles and to link these anchors to related English Wikipedia pages. To achieve the objective, we do not only depend on Wikipedia's distinguishing features (e.g. anchor links information and language links) but also developed a method that analyzes the semantic features of anchor texts in Chinese Wikipedia. In the linking phase, a Latent Dirichlet Allocation model (LDA) is used for the computation of a text similarity measure among the English Wikipedia articles. This novel approach to address the word-to-links ambiguity issue shows encouraging result in the CrossLink-2 evaluation.

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

Keywords
Cross-lingual Link Discovery, CLLD, CrossLink-2, Wikipedia, Latent Dirichlet Allocations, LDA

Team Name
NTHU

Subtasks
Chinese to English, Japanese to English

1. INTRODUCTION
In this information age, lots of knowledge and new ideas appear at every moment and are accessible in digital format. But the language barriers hold back this progress of information exchange. For example, Wikipedia, the largest online collaboratively human-edited encyclopedia, contains millions of articles in many written languages, but the richness of its content varies among the different languages covered by the Wikipedia.

This imbalance is illustrated by the Wikipedia collection sizes between Chinese, English, Japanese and Korean in CrossLink-2 task (Table 1). The size of the English collection is almost ten times the size of the Chinese collection, which means that most of the English Wikipedia topics do not have a corresponding article in the Chinese Wikipedia. Therefore, in order to extend the accessibility for people, to the growing source of knowledge contained in the English Wikipedia, creating cross-lingual pages automatically becomes a meaningful challenge.

Table 1. The size of CEJK Wikipedia document collections used in CrossLink-2

Language	Number of doc	Size	Dump Date
Chinese	404,620	3.6GB	11/01/2012
English	3,581,772	33.0GB	04/01/2012
Japanese	858,610	9.8GB	04/01/2012
Korean	297,913	2.2GB	22/01/2012

CrossLink-2 is the Cross-Lingual Link Discovery (CLLD) task held by the NTCIR workshop. In this task, NTCIR provides well-formed document collections and an evaluation mechanism to give researchers a convenient research environment. In contrast with the design of the first year's CrossLink task, this year the task focuses on mining anchors in CJK languages and on linking them to English articles.

Because this task involves pairs of languages, one of the important research questions of this work is the determination of the part of the articles that should be translated, the direction of the translation, the translation tools to be used, and the step of the work the translation should take place. We choose to rely on Google Translate for the translation of the anchors. We also apply Google Translate for the global translation of the input Chinese or Japanese page into English in the disambiguation phase.

A second important choice concerns the semantic similarity measurements. An exact match of the extracted and then translated terms with the English Wikipedia titles is not always reached. Even when it is, several titles might be candidate for a link. Those cases necessitate the detection of expressions that have similar senses and the detection of documents that have similar contents. In this work similarity between translated documents and all the documents in the English Wikipedia, considering the content of the documents and not only their title, is computed through a LDA model

We split the cross-link discovery task into three sub-problems, that are also the three main steps of our approach:

(1) *Anchors mining*: We need to decide what word in the article is valuable to be linked out. Because of CJK characters' segmentation specificity, finding anchor text's boundaries is a challenging problem in this phase.

(2) *Cross-lingual linking to related articles*: Then we want to link the anchors to potential related articles in the target language. To overcome the cross-lingual challenges, we may use either machine translation or Wikipedia's human-edited language link information.

(3) *Disambiguation*: Given words or expressions may denote different concepts regarding the varying contexts they occur in. When several candidates have been extracted in the previous steps, we rely on a document similarity measure for the identification of the correct one. The anchor text in an input article is linked to the candidate document that is the most related to the input document.

In the following paper, the related work is reviewed in Section 2. Our approach is presented in Section 3. Section 4 describes the performance of our model in the CrossLink-2 evaluation. The results are discussed in Section 5.

2. RELATED WORK

In NTCIR-9's CrossLink task, HITS[1] presented a system that builds a multilingual concept repository derived from Wikipedia in a first step. The lexicon in the repository is searched to identify candidate anchors. They then apply heuristics and a supervised model to decide whether to retain the anchors as correct or not. They also took advantage of much Wikipedia structural information such as incoming links, outgoing links and topical categories. They are modeled as edge weights in a graph-based disambiguation algorithm. HITS obtained the best performance in the evaluation with the Wikipedia Ground Truth. However, HITS' method relied on the original cross-lingual resources in Wikipedia to address the anchors mining and the cross-lingual linking subproblems. If a concept is in the target knowledge base but is not in the original one, it will probably not be linked with this method. As a consequence, this approach is not appropriate to this year's task. In the case of CJK languages, since English Wikipedia is much larger than CJK Wikipedia, we need to dig out a large amount of concepts that are in the English Wikipedia but that are not in CJK Wikipedia.

Another team running in NTCIR-9, UKP[2], used machine translation and dictionary lookup to help finding documents in target languages. After acquiring anchor text translations, they applied title matching to get possible documents. An IR system, Incoming-link anchor search method was used to rank the target documents. The disadvantage of UKP's approach is that they didn't bring up any disambiguation method. Even though, they still showed competitive result in the task experiment. UKP's evaluation result was ranked in the three top teams in almost every evaluation metrics.

In linking the ambiguous anchor words of the Chinese, Japanese and Korean articles to the English Wikipedia, the title of the articles might not provide enough information. The disambiguation step of our model relies on LDA topic modeling of the English Wikipedia that allow us to compute article similarity based on their content. LDA is a model introduced by Blei et al (2003)[5], designed to automatically induce latent hidden topics from discrete data, and in our case, from a corpus of documents. Each topic is a distribution over the words of the corpus. High frequency words in a given topic tend to be easily interpretable by humans as a coherent semantic domain. Documents are represented as a mixture of topics. This is to say that every topic of the model has a probability in every document, and that the similarity between two documents can then be calculated as a similarity between the topics composing it. Modeling documents as vectors of topics reduces the dimensionality of the vectors in comparison with vectors made of document words. The consequence is a better handling of data sparseness issues. For example, two documents that do not contain the same words can be very similar in the model if the words have high probabilities in similar topics. Once a LDA model is created, it becomes possible to compare new incoming documents to the documents of the model. Applied to our task, after a Wikipedia article has been translated into English, we can compare it with every article of the English Wikipedia to find the most similar ones.

Considering the size of the English Wikipedia, the entire word-document matrix needed to compute the LDA model cannot hold in most of computer's RAM. Moreover, comparing a document with all the documents in the Wikipedia in a reasonable time necessitates a clever indexing scheme. Řehůřek(2010)'s Gensim[1][6] python module addresses all these issues while focusing on efficient memory allocation and fast indexing. Its implementation of the LDA algorithm uses variational inference. Gensim specifically provides a helpful script and a tutorial to create a LDA model of Wikipedia rapidly using the module.

3. METHODOLOGY

In this section, we introduce the framework of our system in CrossLink-2 task. In the following sections, we detailed our method which is formed of three steps: anchors mining, document linking and disambiguation.

3.1 Anchors Mining

If some terms occur in a large proportion of articles, others are closely related to the topic in consideration or are simply rare and are the ones that should be anchored. A user is more likely to search for infrequent words that are related to his original query. Given a text, anchor mining is the task of automatically identifying those terms of interest. In the case of CJK languages, this task has a supplementary difficulty due to the absence of word boundaries,

We applied two methods to find valuable anchor texts in input articles. The first one is to rely on the concepts that already exist in the original language version of Wikipedia. In order to mine the concepts that didn't exist in the original language but that are defined in the target language, we utilized natural language processing techniques, such as part-of-speech tagging. The two methods are going to be detailed in the following paragraph.

3.1.1 Case 1, a langlink has been established: Wikipedia Mentions Matching

We preprocessed the CEJK Wikipedia collections and constructed a mention table. We extracted every anchor links in the Wikipedia collections and recorded the anchor text and target linking page of each anchor link. The surface text that is anchored is called the mention in contrast with the target link of the anchor. Mentions may be different from the title of the target topic page. Because anchor links in Wikipedia are edited by humans, they provide useful and trustworthy information concerning the concepts to be linked. The same mention is eventually shared among different concepts. By constructing the mention table, we can know possible forms of surface text of each Wikipedia topic.

Table 2 is an example of the mention table when the mention is

[1] Gensim: http://radimrehurek.com/gensim/

"台灣" (Taiwan). For all anchor links in the Chinese Wikipedia collection, when the anchor text happens to be "台灣", the most likely target page is "台灣". However, the link can also sometimes refer to Taiwan's movie list or Chinese Taipei Baseball Team. We also recorded the referring times of each link and if the mention is the same as the target page's title or redirect. This additional information will be used to rank the target pages in the later process.

In order to verify our method in different languages, we participated in C2E and J2E subtasks and constructed Chinese and Japanese mention tables. For every Chinese or Japanese input article to the system, we discovered anchor candidates by mention matching. Every pair of anchor text and target page will be considered and carried to the anchor link selection stage. We designed a supervised method to compute the keyness of each anchor text and the relevance of the target links. The detail of the anchor link selection process will be given in section 3.4.

3.1.2 Case 2, no langlink is found in the Wikipedia: POS Tag Analysis

While thinking of finding potential anchor texts without any existent information in original language's Wikipedia collection, the part-of-speech tags came to our mind. We decided to use POS tags as clues to discover the anchor texts that didn't exist in the mention table.

We limit the scope of this experience to Chinese and only employ this method in C2E subtask. We randomly chose one hundred

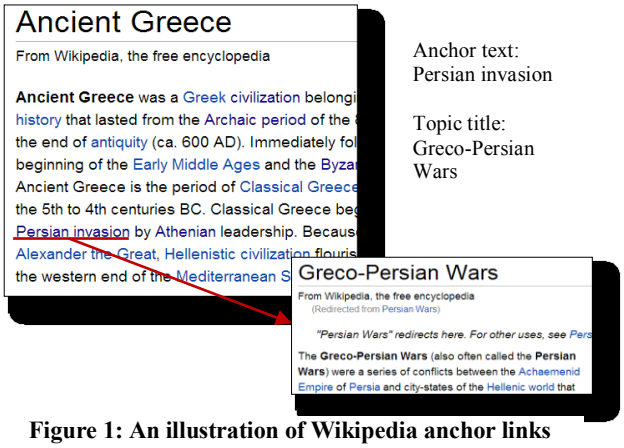

Figure 1: An illustration of Wikipedia anchor links

Table 2: An example of mention table when mention is "台灣"(Taiwan), sorted by show times

Docid	*isTitle*	*isRedirect*	*show times*	*Target Page's Title*
690	T	F	26491	台灣
19865	F	F	91	台灣電影列表 [1]
212347	F	F	69	中華成棒 [2]
271	F	F	65	中華民國 [3]
45653	F	F	12	台灣日治時期 [4]
14125	F	F	12	台灣省 [5]
…	…	…	…	…

Chinese Wikipedia pages in the collection to serve as training data. We extracted the anchor texts in the training data and analyzed all anchor texts' POS tags. Then we retained a list of likely POS tag patterns of Chinese anchor texts. For every input article, we performed POS tagging with a Chinese word segmentation tool and extracted the words that fit in with the patterns as anchor text candidates.

The Chinese word segmentation and POS tagging tool we manipulated is Sinica CKIP Chinese Word Segmentation System[6][12]. Since the system currently only accepts traditional Chinese characters, we need to preprocess the articles and to convert all characters to traditional Chinese. OpenCC[7] is the tool we applied to achieve this goal. In this process, we also recognize an additional problem in dealing with Chinese Wikipedia collections. The document collection is mixed between traditional Chinese and simplified Chinese characters. Moreover, both kinds of characters often co-occur in the same page. Therefore, a system that can take account of both traditional and simplified Chinese characters or at least can transfer the documents into unified characters first is necessary.

3.2 Document Linking

Document linking is the task of creating an edge between the extracted anchor terms and their corresponding document. Cross Language linking requires further steps, implying a translation module or information retrieval techniques to find connections between articles in Wikipedia collection of different languages.

We will introduce our two strategies to solve cross-lingual document linking in this section. One is using the inter-lingual links of Wikipedia which also called langlinks. The other one is applying a machine translation module to overcome the lack of inter-lingual information of Wikipedia.

3.2.1 Wikipedia Langlinks

Langlink is one of Wikipedia's inventive distinguishing features. People can edit a topic's langlink to connect the pages that denote the same concept written in different languages. Consequently, Wikipedia evolves toward a tremendous multilingual encyclopedia. Users can switch languages easily to amend their knowledge of various topics. In this task, langlinks undoubtedly provide the most ideal cross-lingual link solution.

For the anchor links discovered from mention tables, we then check whether the target page has a corresponding English Wikipedia page or not through a langlink. If previous users have marked no link, we need to rely on machine translation.

3.2.2 Anchor Texts Translation

For the anchor texts that we discovered in previous steps, we then translate the surface text to English and find possible linked documents by matching the translated text to titles of English Wikipedia topics. Google Translate API[1] is used as our machine translation system. We have tried to build a machine translation system which focuses on Wikipedia titles. We intended to use the

[1] 台灣電影列表：List of Taiwan Films
[2] 中華成棒：Chinese Taipei Baseball Team
[3] 中華民國：Republic of China
[4] 台灣日治時期：Taiwan Under Japanese Rule
[5] 台灣省：Taiwan Province
[6] Sinica CKIP Chinese Word Segmentation System: http://godel.iis.sinica.edu.tw/CKIP/engversion/wordsegment.htm
[7] OpenCC: Open Chinese Convert https://code.google.com/p/opencc/

cross-lingual title pairs of Wikipedia as parallel corpus. The langlinks were extracted to find corresponding pages between Chinese Wikipedia and English Wikipedia. We obtained about 20,000 title pairs but the data turned out to be too small for a machine translation system. The performance wasn't good because many words were taken as out-of-vocabulary to the system. Accordingly, Google Translate which is trained with large-scale training data is the best machine translation system to use.

At the end of this step, an anchor is associated with a list of English Wikipedia pages that are candidate for final linking. The determination of the best candidate is performed in the disambiguation steps.

3.3 Disambiguation

In this work, the disambiguation step is defined as the task of choosing the best candidate among a list of English pages. Two methods are compared: a keyword method and a topic modeling method based on the relatedness between the original input document and the target candidates documents contents.

The second method relies on the assumption that documents tend to focus about one topic. It follows that the anchors of an input document are likely to be terms that are related to the main topic of this document. If so, the content of the target English documents must also have contents that are related to the source document up to a certain point.

The first subsection presents the keyword method based on a Dice coefficient. The two following subsections describe the creation of a similarity measure for documents, which provides a comparison of any English text with any article of the entire English Wikipedia based on a LDA model. The last subsection relates the application of this model to our disambiguation problem.

3.3.1 Keyword Similarity

Keyword Similarity is a trivial method to compute similarity between two articles. We use the mentions in mention table of English Wikipedia as word bag list to calculate the similarity. We translated the input Chinese or Japanese article to English by machine translation system first. Then apply mention matching to both the translated article and target link page. The score is given by Dice's coefficient.

$$\text{keywordSim} = \frac{2 \times |A \cap B|}{|A| + |B|}$$

A: keywords of input article

B: anchor texts of candidate Wikipedia page

3.3.2 LDA model

Before running the model, a little preprocessing is required. The words of every Wikipedia article（except for the redirect pages）are converted into a sparse vector of word frequencies after filtering out the stop words (determiners, prepositions and very frequent words). Following the guideline of the Gensim module, we first compute the Tfidf of all the vectors. The dimensionality of the vectors can then be reduced with LDA.

Online LDA (Hoffman et al. [13]), implemented in Gensim, is an algorithm that splits the corpus in chunks of documents, updating the model after every chunk is processed. When the topics of the documents remain consistent between the chunks, that model estimation converges faster than the traditional batch LDA model while generating topics of comparable quality. Following Řehůřek's example, the 3770751 Wikipedia documents are split in chunks of 10,000 documents.

racing car engine race cars driver motor formula engines speed
la el mexico spanish puerto san del juan mexican chile
government patrolling court accused police act law clerk defending security
regiment army polish infantry battalion brigade division poland battle
album song chart band track vocals albums songs guitar single
navy ship naval ships hms royal officer vessel uss admiral
river lake antarctic island km park glacier mountain dam mountains
orchestra piano opera composer symphony czech violin dgg jazz
al ottoman khan armenian muhammad pakistan muslim empire afghanistan Israeli
church bishop catholic cathedral rev diocese ordained college parish priest

Table 3: The ten most probable words obtained for ten LDA topics

We choose to run the model for one hundred topics, which is the number frequently encountered in the LDA literature for the computation of models for large collections of texts. The hyperparameters alpha and eta, that affect the sparsity of the document-topic (theta) and the topic-word (lambda) distributions are both set to 0.1. Table 3 displays the headwords of ten of the one hundred obtained topics.

The Chinese and Japanese Wikipedia pages, after being translated into English with Google Translate API, can then be integrated in the LDA model and compared with any article of the English Wikipedia. A new translated English document is first converted in its bag of words vector, and then in the LDA vector of length on hundred that corresponds to the distribution of the new document over the LDA topics.

A different method left for future work would be to build a bilingual LDA model and to compare the final results of this model with our current results on the CrossLink task.

3.3.3 Similarity Computation

The comparison between two documents is done with a cosine similarity between the two documents topic vectors.

$$\cos(A, B) = \frac{\sum_{i=1}^{n} A_i . B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \cdot \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

In this experiment, A is a vector representing a document of the English Wikipedia, B is the vector representing the original Chinese or Japanese input document after its translation into English.

3.3.4 A feature for target link ranking

Coming back to the list of English articles that are candidate links for a mined anchor obtained in the previous step, we can now use a document-document similarity as a feature to disambiguate the list.

Every candidate page is compared with the input document using the cosine similarity of their LDA representation. The link that is finally retained is the one that is the most similar to the input

[1] Google Translate API: https://developers.google.com/translate/

document regarding their distribution over topics.

$$Argmax_{C_i} = \cos(C_i, B) \ , i \in \{1, \dots, n\}$$

C_i is a candidate in the list of n candidates for an anchor link.

In practice, the translated document vector B is first compared to all the documents in the English Wikipedia with a single query. Then, for every anchor, the candidate documents are picked out and ranked from the results of this query.

The proposed method presents the advantage of addressing ambiguity issues induced both by translation and by language inherent ambiguity, by looking into the content of the articles and modeling the degree to which documents of different languages talk about the same topics.

In the following subsection, we describe how the previously extracted information are combined to produce the final ranking of the links.

3.4 Link Selection

The link selection is performed at the final step because we want to do anchor filtering after computing all the scores including relativeness between articles produced in the disambiguation step. CrossLink-2 sets a limitation that each topic allows no more than 250 anchors and each anchor can have up to 5 targets. The objective is that the system can recommend the relevant terms that deserve to be anchored in the context of the input article. Therefore, we aim to rank the anchors and decide whether an anchor can be kept in the final result in this phase.

In this stage, we rank all the anchor candidates discovered previously, by the combination of many measures. The previously extracted features are listed below:

- Global Keyness (gk): This is a score to see how likely the n-gram is to be an anchor. We counted n-grams (n=1~9) in the Chinese and Japanese Wikipedia collection. The score is given by show times recorded in mention table and its n-gram count.
- Category Probability (cat_p): The topic category is a good information to make a decision about the quality of an anchor. Some categories are very likely to be anchored, e.g. Countries, Movie players. We analyzed all links in the mention table and computed this score by the portion of categories' distribution as the conditional probability of the term to be a link given its category.
- Parenthesis (pp): If a word in an input article is parenthesized, it is very likely to be anchored.
- Keyword Similarity (keywordSim)
- LDA similarity (lda_sim)

We combine these measures in a supervised model, we trained the model with 100 Wikipedia articles that were randomly picked up. Some thresholds were set in the measurement. For example, if gk score is not high enough then the threshold of similarity should be higher.

4. EXPERIMENTS

This section we are going to present the evaluations results in CrossLink-2. NTCIR-10 provides 25 test topics for CEJK language. Each team should discover anchor texts and cross-lingual links from the test dataset using the system they developed. The result will be submitted to the organizers and the submissions of every team will be evaluated under the same evaluation framework.

There are three evaluation metrics, Link Mean Average Precision (LMAP), Precision-at-N (P@N) and R-Prec. LMAP is the mean average precision over all the links in all the test topics. P@N is the average precision for all the test topics at the first N items. R-Prec is the precision over the relevant items in the qrels. The benchmarks include Wikipedia ground truth and manual assessment results. Table 4 and Table 5 describe our performance in C2E and J2E subtask.

5. DISCUSSION

In linking Chinese to English, our system ranked number two at Precision-at-5 metric in file to file evaluation with manual assessment result. Unfortunately, with the other evaluation metrics, our system showed a lot of improvement space in precision in comparison with the other teams. The result can be discussed in two aspects, anchors mining and link selection.

In anchors mining, we aim to discover more anchor text than the mentions that have been showed in Wikipedia. Therefore, we applied the POS tag analysis module. The POS pattern strategy gave satisfying results.

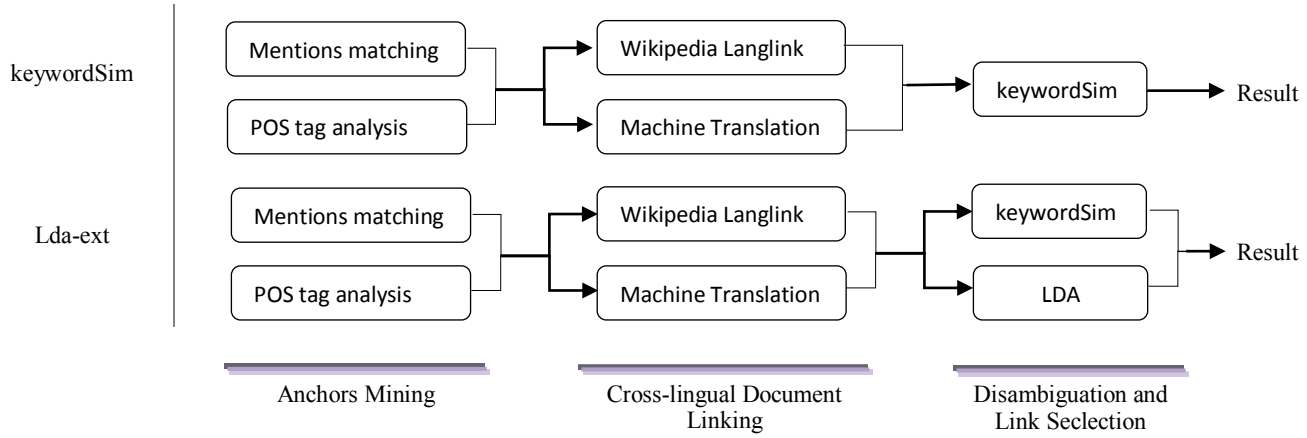

Figure 2. The diagram illustration of the modules employed in our two methods

Table 4. Performance of NTHU's system in Chinese to English subtask

		LMAP	*R-Prec*	*P5*	*P10*	*P30*	*P50*	*P250*
F2F Wikipedia Ground-truth	Best Score	0.517	0.520	1.000	0.972	0.779	0.582	0.123
	keywordSim	0.080	0.192	0.256	0.236	0.221	0.194	0.068
	Lda-ext	0.082	0.194	0.256	0.240	0.224	0.195	0.070
F2F Manual Assessment	Best Score	0.069	0.180	0.384	0.368	0.320	0.266	0.123
	keywordSim	0.025	0.096	0.192	0.136	0.120	0.123	0.051
	Lda-ext	0.034	0.114	0.192	0.136	0.121	0.126	0.078
A2F Manual Assessment	Best Score	0.113	0.147	0.096	0.072	0.083	0.082	0.064
	keywordSim	0.012	0.024	0.040	0.024	0.024	0.028	0.008
	Lda-ext	0.021	0.036	0.040	0.024	0.025	0.029	0.013

Table 5. Performance of NTHU's system in Japanese to English subtask

		LMAP	*R-Prec*	*P5*	*P10*	*P30*	*P50*	*P250*
F2F Wikipedia Ground-truth	Best Score	0.548	0.561	0.946	0.938	0.829	0.657	0.178
	keywordSim	0.083	0.189	0.254	0.246	0.224	0.199	0.084
F2F Manual Assessment	Best Score	0.312	0.418	0.520	0.460	0.357	0.267	0.066
	keywordSim	0.102	0.138	0.184	0.164	0.133	0.123	0.049
A2F Manual Assessment	Best Score	0.270	0.120	0.144	0.120	0.107	0.083	0.037
	keywordSim	0.127	0.074	0.064	0.068	0.064	0.062	0.017

For example, the term “祖师爷” (patron deity, the great founder of certain specialized jobs) appears in one article in the Chinese test set. In our opinion, this term is a good term to be emphasized as an anchor because it is not a general term and its reference may be confusing for readers. However, the page “Patron deity” in English Wikipedia does not have any langlink that links to the same concept in Chinese Wikipedia because there is no such concept of “祖师爷” in Chinese Wikipedia. Consequently, the only way to link “祖师爷” to “Patron deity” is through machine translation. In our system, the anchor text “祖师爷” has been discovered by the POS tag analysis module. And the machine translation result of “祖师爷” also showed “patron deity”. After all at the link selection step, the system picked “Zhang Sanfeng” (A legendary Taoist who is said to be the founder of Tai Chi Chuan or Taijiquan.) as the only correct target link rather than “Patron deity”. The reason is that the term “祖师爷” has a very low global keyness due to the fact that there is no such concept in Chinese Wikipedia. In the situation with low global keyness, the similarity between an original page and the target page should be very high to avoid being filtered out.

This example shows the shortcoming of ours link selection procedure. We discovered anchor links from many approaches. But the supervised model of link selection filtered good links out. Because the model is combined by some different measurements, it’s hard to define a really good ranking measure manually. The system can be improved by applying machine learning techniques in the link selection stage in the future, such as a binary classifier. We can improve the model by training a classifier with all the langlinks of the Wikipedia as training data and with our previously extracted features in order to attribute them correct weights.

We submitted two systems that allow us to compare the performance of two different similarity measures. A keyword measure based on a Dice coefficient and a LDA based measure. The results of almost all the evaluation show that the LDA measure performs better than the keyword one in ranking the links. However, this improvement is not significant enough to compete with the best system that has been proposed by the other teams. Nevertheless, the LDA model seems to be promising and many modifications could be imagined in future work. For example, it would be interesting to test the effect of a bilingual LDA model for this task, or to use the model earlier in the process to detect more links by looking at the target links of the English pages that are very similar to the input page. The fact that links exist in pages that are similar to the input page could also become a valuable feature for the final ranking step.

6. CONCLUSION

We presented a system that relies on POS patterns, n-gram computation and document similarity in addition to the Wikipedia immediate features. As suggested in the discussion, in spite of the efficiency of these features taken individually, in its current state, our model is penalized by the absence of a machine-learning frame to combine and weight the features optimally. It appears that document similarity based on topic modeling could become a powerful tool for cross-link discovery. More features based on similarity, making use of existing connections in the English Wikipedia could be tested. Bilingual topic modeling would also be an interesting direction to take for future research.

7. REFERENCES

[1] Tang, L.-X., et al. (2013). Overview of the NTCIR-10 Cross-Lingual Link Discovery Task. Proceedings of NTCIR-10, to appear.

[2] L.-X. Tang, S. Geva, A. Trotman, Y. Xu, and K. Itakura. (2011). Overview of the NTCIR-9 crosslink task:Cross-lingual link discovery. In Proceedings of the 9th NTCIR Workshop Meeting, Tokyo, Japan, 6-9 December 2011.

[3] A. Fahrni, V. Nastase, M. Strube. 2011. HITS' Graph-based System at the NTICR-9 Cross-lingual Link Discovery Task. In Proceedings of the 9th NTCIR Workshop Meeting, Tokyo, Japan, 6-9 December 2011.

[4] J. Kim, I. Gurevych. 2011. UKP at CrossLink: Anchor Text Translation for Cross-lingual Link Discovery. In Proceedings of the 9th NTCIR Workshop Meeting, Tokyo, Japan, 6-9 December 2011.

[5] D. Blei, A. Ng, and M. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

[6] Reh̊uřek, R. and P. SOJKA. Software Framework for Topic modeling with Large Corpora. In Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks. Valletta, Malta: University of Malta, 2010. p. 46--50, 5 pp. ISBN 2-9517408-6-7.

[7] Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.* 3 (Mar. 2003), 1289-1305.

[8] D. Milne and I. H. Witten. 2008. Learning to link with Wikipedia. In Proceedings of the ACM 17th

[9] R. Mihalcea and A. Csomai 2007. Wikify!: Linking documents to encyclopedic knowledge. Proceedings of the 16th ACM conference on conference on information and knowledge management. 233 p.

[10] Y.H. Lin, Y.L. Liu, T.X. Yen, Jason S. Chang. 2012. Context-Aware In-Page Search, Proceedings of ROCLING 2012, Taiwan, September 2012.

[11] Tsai Yu-Fang and Keh-Jiann Chen, 2004, "Reliable and Cost-Effective Pos-Tagging", International Journal of Computational Linguistics & Chinese Language Processing, Vol. 9 #1, pp83-96.

[12] Ma, Wei-Yun and Keh-Jiann Chen, 2003, "Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff", Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing, pp168-171.

[13] M. Hoffman, D. Blei, F. Bach, 2010. "Online Learning for Latent Dirichlet Allocation," in Neural Information Processing Systems (NIPS), Vancouver.