Using Concept base and Wikipedia for Cross-Lingual Link Discovery

Pham Huy Anh
Nagaoka University of Technology
anhph@stn.nagaokaut.ac.jp

Takashi Yukawa
Nagaoka University of Technology
yukawa@vos.nagaokaut.ac.jp

Abstract

This paper describes our method for the Cross-Lingual Link Discovery (CLLD). We used English-Japanese document collections in CLLD subtask of NTCIR-9. The topics in our method are translated by Wikipedia. Wikipedia is written by multi-language. In our method, the page written by the target language is retrieved for each topic written in the source language. The topic written in the target language is made from Wikipedia concept part of this page. Cross-language link is retrieved by a TF-IDF model. We use nouns, nouns phrase and adjective to make concept base. Re-ranked result retrieved by TF-IDF model. TF-IDF and concept base are made from the outline part of Wikipidia pages, which are written in the target language extracted in Wikipidia pages collection. Crosslink Evaluation Tool of NTCIR-9 Crosslink Task is utilized for performance evaluation.

Keywords

Cross-Lingual Link Discovery, Concept base, TF-IDF, Wikipedia.

1. Introduction

Cross-referencing documents are an essential part for organizing textual information. However, keeping links in large, quickly growing, document collections up-to-date is problematic due to the number of possible connections. In multilingual document collections, the semantic interlinking of related information in a timely manner becomes even more challenging. Cross-lingual link discovery (CLLD) is a way of automatically finding potential links between documents in different languages. CLLD actively recommends a set of meaningful anchors in the source document and uses them as queries with the contextual information from the text to establish links with documents in other languages.

On the internet, information is written by multi-language. In our methods, Wikipedia online translation is used for making topic written in the target language. A concept part is extracted after taking a Wikipedia page written in the target language. This concept part is a topic written in the target language. Concept base is used for re-rank ranking list obtained by calculating cosine similarity using tfidf vectors. Crosslink Evaluation Tool of NTCIR-9 Crosslink Task is utilized for performance evaluation.

2. Related Work

In the part, many research studies have focused on link-based method and semi-structured method.

Link -based method approaches discover new links by exploiting an existing link graph. Link-based method is used by Itakura and Clarke [4], Lu et al. [5], Jenkinson et al. [6].

Semi-structured approaches try to discover new links using semi-structured information, such as the anchor texts or document titles. Semi-structured method used by Geva, Milne and Witten [7], Mihalcea and Csomai et al. [8], Granitzer et al. [9].

The main disadvantage of the link-based and semi-structured approaches is probable difficulty associated with porting them across different types of document collections. The two well-known solutions to monolingual link detection, the Geva's and Itakura's algorithms (Trotman et al., 2009[10]), fit in these two categories.

In this paper, we present a method for CLLD which use concept base and Wikipedia online translation.

3. Background

3.1 Concept Base

The concept base was proposed by Schüetze and Pederson as a method of automatically constructing a thesaurus with the corpus and using a higher dimension vector space to express the relations between words appearing in a document [3]. Currently, the commonly utilized composition of the concept base is a word × word matrix.

First, in the traditional construction of the concept base, N words occurring with high frequency in the document for retrieval are selected to create a neighborhood co-occurrence matrix of a word with another word in the neighborhood (see **Fig. 1**). Wij is the co-occurrence frequency between word i and word j. Before constructing the neighborhood co-occurrence matrix, it is necessary to perform a morphological analysis and remove the stopwords as a preprocessing step. Stopwords are words such as particles or auxiliary verbs that does not have an important role in the documents.

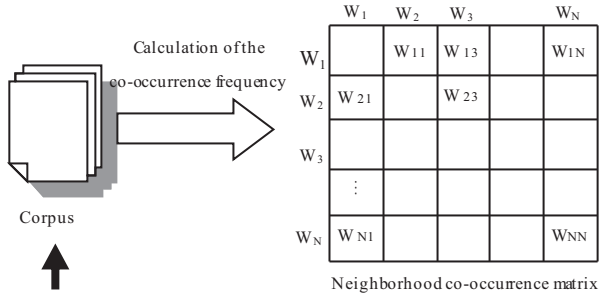


Fig. 1 Making the neighborhood co-occurrence matrix.

The created neighborhood co-occurrence matrix can be considered as a word vector in which the number of words corresponds to the number of dimensions. However, there is the problem that the number of dimensions increases as the scale of the corpus grows because dimension depends on the number of words. Moreover, because each axis is a word, it is not easy to think of the axes as being mutually orthogonal. Therefore, to create the neighborhood co-occurrence matrix, singular value decomposition (SVD) is implemented. Under SVD, the neighborhood co-occurrence matrix is divided into three matrices: the transposed orthogonal matrix, the diagonal matrix, and the row orthogonal matrix. The row of 100~200 dimensions is extracted from the obtained row orthogonal matrix. The extracted matrix is the concept base (see **Fig. 2).**

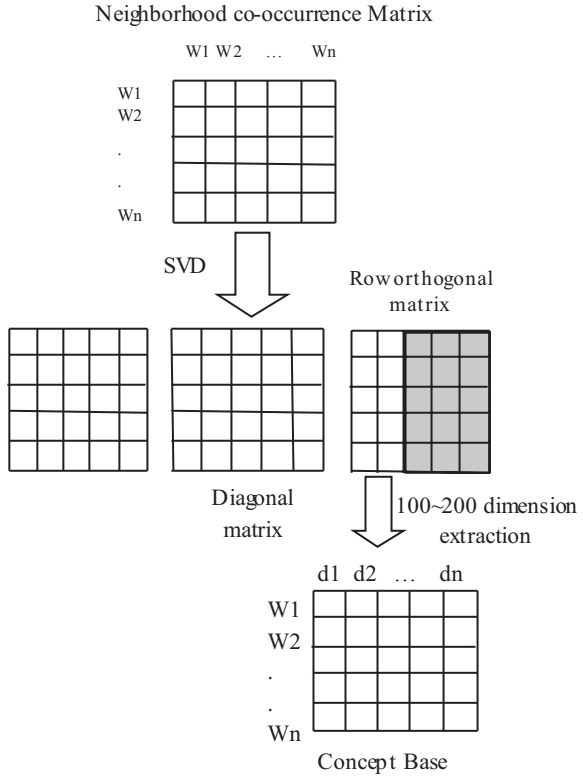


Figure2: Construction of concept base

3.2 Specifically considered concept base

The concept base is made by using a outline part of Wikipedia pages collection. We use the Mecab with Wikipedia dictionary morphological analyzer and discard stopwords, and extract nouns, nouns phrase.

In the past, concept base did not consider the specificity, retrieval performance decreased. We propose to construct a concept base in which this specificity is considered. The IDF is an index showing the specificity of a word. The concept base considers the frequency of a pair of words related to the co-occurrence as a single element. The IDF of a word pair is evaluated and used as a weighting term. For example, since the co-occurrence frequency of (computer, network) is greater than one, the IDF of the word pair is calculated as follows (see **Eq. 1**):

$$idf\{pair(t_1,t_2)\} = \log \frac{N}{df\{pair(t_1,t_2)\}} \qquad (1)$$

Here, ***pair*** (t_1,t_2) is a pair of the words t_1 and t_2 that exist in the co-occurrence relation, **N** is the total number of documents, and $df\{pair(t_1,t_2)\}$ represents the number of documents in which the word pair appears.

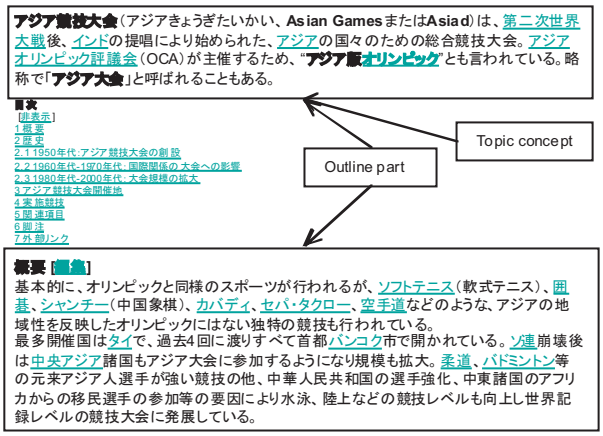
The $idf\{pair(t_1,t_2)\}$ value becomes the origin of the concept base. The weight is calculated by multiplying this value by each element of the neighborhood co-occurrence matrix. Therefore, element W of the neighborhood co-occurrence matrix is determined as follows (see **Eq. 2**).

$$W_{t_1\overline{t_2}} = F_{t_1t_2} \times idf\{pair(t_1,t_2)\} \qquad (2)$$

Here, F_{t1t2} is the co-occurrence frequency of words t_1 and t_2. The element of the neighborhood co-occurrence matrix in **Fig. 1** replaces **Wi (i=1,…,n)** with the value of the above expression. The concept base, composed of the neighborhood co-occurrence matrix with the element of (2), is the specific concept base.

4. The CLLD methods

This section describes the methods used in our experiments. In our experiment used Japanese document collection in English-Japanese subtask of CLLD. The whole process of cross-language link detection is shown in Figure 3. The method takes an input being a topic written in the source language (English). Then, we can automatically get the Wikipedia page written in the target language (Japanese) on the internet. With Wikipedia page written in the target language, we used the regular expression to get one sentence in the page. This sentence is a topic concept. Following is an example:



On the other hand, we used the regular expression to extract the outline part of Wikipidia page, which is written in the target language, form Wikipidia page collection. Above is an example. Mecab with Wikipedia dictionary is used for morphologically analyzing the outline part. Thereafter, the concept base and TF-IDF is made from outputs of the Tokenizer.

The topic concept written in the target language is equivalent to calculating cosine similarity using TF-IDF vectors. Concept base is used for re-rank ranking list obtained by calculating cosine similarity using TF-IDF vectors.

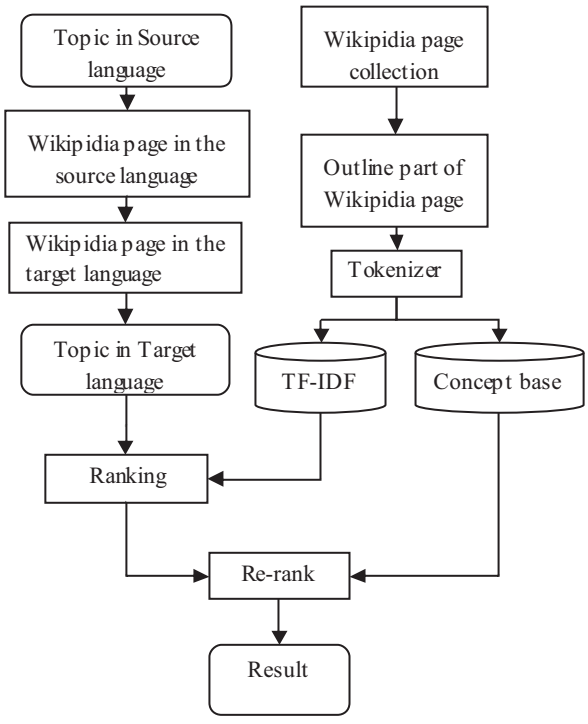


Figure 3: The overall design of our CLLD system

4.1 Using concept base for re-ranking

Wikipedia concept was used for calculating cosine similarity using TF-IDF vectors to obtain various correct links. We use re-ranking to raise correct links to a high rank. We use the concept base re-rank with topics. Generally, when calculation of similarity of topic vector and document vector, as the number of words in the document is much larger than the number of words in the topic, the similarity influence and the result performance decrease.

In our method, the document vector $(\vec{D})$ is divided into sentence vectors $(\vec{S}i)$. The number of words in the topic vector $(\vec{Q})$ and the number of words in the sentence vectors are almost the same. The similarity of the topic vector and the sentence vector is then calculated. The highest similarity is assumed to be the similarity of the query vector and the document vector. The formula is shown below (see **Eqs. 3, 4**).

$$\vec{D} = \vec{S}1 + \vec{S}2 + \ldots + \vec{S}n \qquad (7)$$

$$Sim(\vec{Q}, \vec{D}) = \underset{i=\lceil 1:n \rceil}{Max}(\vec{Q}, \vec{S}i) \qquad (8)$$

where n is the number of sentences in the document.

5. Performance evaluation

5.1 Evaluation Index

For the experiment, file-to-file (F2F) level with Precision-at-N, R-Prec, and Mean Average Precision (MAP) value were used as the evaluation.

Precision, and Recall are defined as:

$$\text{Precision} = \frac{Number\ of\ correct\ links}{Number\ of\ identified\ links} \qquad (3)$$

$$\text{Recall} = \frac{Number\ of\ correct\ links}{Number\ of\ links\ in\ arel} \qquad (4)$$

The precision and recall are computed in link level for each topic.

R-Prec, and MAP are defined as:

$$MAP = \frac{\sum_{t=1}^{n} \frac{\sum_{k=1}^{m} Pkt}{m}}{n} \qquad (5)$$

Where n is the number of topics; m is the number of identified items (links); Pkt is the precision at top K items (links) for topic t.

$$\text{R-Prec} = \frac{\sum_{t=1}^{n} Pt@R}{n} \qquad (6)$$

Where n is the number of topics; m is the number of identified items (links); Pkt is the precision at top k items (links) for topic t; Pt @ R (= number of correct items (links) / number of items (links) in qrel) is the precision calculated using number of links in qrel as denominator for topic t.

5.2 Wikipedia Ground-Truth and Manual Assessment

Wikipedia Ground-Truth: The set of links used as the ground truth is derived from the existing links in the topics, and their counterparts in the target corpus.

Manual Assessment: The Wikipedia ground truth is easy to get, but not necessarily reflecting user preferences optimally. Much of it is automatically generated. Even manually generated Wikipedia links, by the author, may be disliked by most assessors. The NTCIR assessors make the decision abut the quality of both anchors and targets.

5.3 Evaluation Result

For the re-ranking based on concept base system, the MAP valuation was evaluated and compared with TF-IDF system. Table 1 shows the evaluation results of the re-ranking based on concept base system and the TF-IDF system. The re-ranking based on concept base system gives higher precision compared to the TF-IDF system.

Fig. 4 and Fig 5 show Precision-Recall curve of the re-ranking based on concept base system and the TF-IDF system with Wikipedia ground-truth and manual assessment. Precision of the top rank went up by the figure, however recall did not change.

Table 1: Evaluation results of the re-ranking based on concept base system and the TF-IDF system.

	Wikipedia ground-truth		Manual assessment	
	TF-IDF	Re-rank	TF-IDF	Re-rank
MAP	0.041	0.044	0.028	0.03
R-Prec	0.084	0.087	0.075	0.059
P5	0.44	0.456	0.16	0.192
P10	0.364	0.304	0.144	0.116
P20	0.256	0.214	0.094	0.08
P30	0.209	0.181	0.073	0.069
P50	0.151	0.146	0.054	0.054
P250	0.053	0.058	0.019	0.021

The Re-rank using a concept base showed that the correct answer link increased the rank.

6. Conclusions

The proposed system used Wikipedia online to translate topic and use the concept base to re-rank ranking list obtained by tfidf model. The re-ranking system based on concept base gives higher MAP evaluation compared to the TF-IDF system. The top result includes links to the relevant. However, the number of link relevant in the result of proposal system is low. So, CLLD methods do not provide high recall. In the future, if the Recall of the system goes up, the performance of the system will be possibly improved.

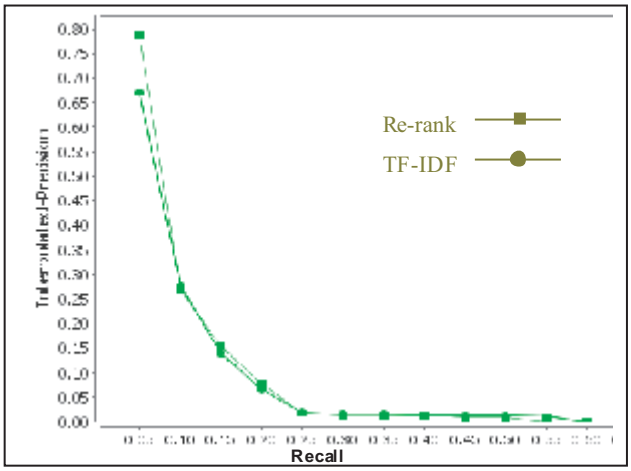


Fig.4: InteP-R Curve: Outgoing (Wikipedia ground-truth)

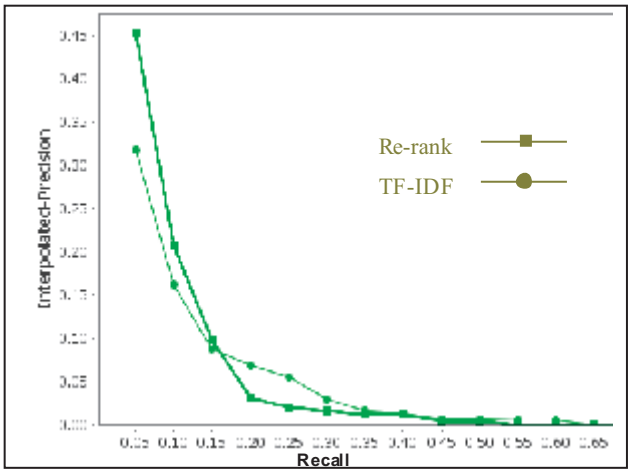


Fig.5: InteP-R Curve: Outgoing (Manual assessment)

7. Acknowledgments

I was granted permission for using the test collection of Wikipedia pages of the Japanese, which were spent by NTCIR9 task, by National Institute of Informatics (NII).

8. References

[1] G.Salton. 1983. Introduciton to Modern Information Retrieval . McGraw-Hill.

[2] A.Yasumune, H.Taiichi, T.Takenobu, T.Hozumi. 2004. Research on cross- language information retrieval using vocabulary extension. Natural Language Processing 10, D3-1, 2004

[3] H.Schüetze, J.Pederson. 1994. Information retrieval Based on Word Senses, In Fourth Annual Symposium on Document Analysis and Information Retrieval, pp.161-175, 1994.

[4] Kelly Y. Itakura and Charles L. A. Clarke. 2008. University of waterloo at inex 2008: Adhoc, book, and link-the-wiki tracks. In Geva et al. (Geva et al., 2009), pages 132–139.

[5] Wei Lu, Dan Liu, and Zhenzhen Fu. 2008. Csir at inex 2008 link-the-wiki track. In Geva et al. (Geva et al.,2009), pages 389–394.

[6] Dylan Jenkinson, Kai-Cheung Leung, and Andrew Trotman. 2008. Wikisearching and wikilinking. In Geva et al. (Geva et al., 2009), pages 374–388.

[7] David Milne and Ian H. Witten. 2008. Learning to link with wikipedia. In James G. Shanahan, Sihem Amer-Yahia, Ioana Manolescu, Yi Zhang, David A. Evans, Aleksander Kolcz, Key-Sun Choi, and Abdur Chowdhury, editors, CIKM, pages 509–518. ACM.

[8] Rada Mihalcea and Andras Csomai. 2007. 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. ACM.

[9] Michael Granitzer, Christin Seifert, and Mario Zechner. 2008. Context based wikipedia linking. In Geva et al. (Geva et al., 2009), pages 354–365.

[10] Shlomo Geva, Jaap Kamps, and Andrew Trotman, editors. 2009. Advances in Focused Retrieval, 7th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2008, Dagstuhl Castle, Germany, December 15-18, 2008. Revised and Selected Papers, Lecture Notes in Computer Science. Springer.