Cross-language Entity Linking Adapting to User's Language Ability

Jialiang Zhou, Fuminori Kimura, and Akira Maeda

***Abstract*—In this paper, we propose a method to automatically discover valuable keyphrases in Japanese and link these keyphrases to related Chinese Wikipedia pages. The method that we propose has four stages. Firstly, we extract nouns from a Japanese document using a morphological analyzer and extract the candidates of keyphrases using a method called Top Consecutive Nouns Cohesion (TCNC) [1]. Then, we judge the degree of difficulty of the extracted keyphrases and tag them with different linguistic levels. Secondly, we translate extracted Japanese keyphrases into Chinese using a combination of three translation methods. Thirdly, we extract the corresponding Chinese articles of the translated keyphrases. Fourthly, we translate the original Japanese document into Chinese and make a vector of noun frequencies. Then, we calculate the cosine similarities of the translated original document and candidate Chinese Wikipedia articles. Finally, we create links from the Japanese keyphrases to the top-ranking Chinese Wikipedia articles.**

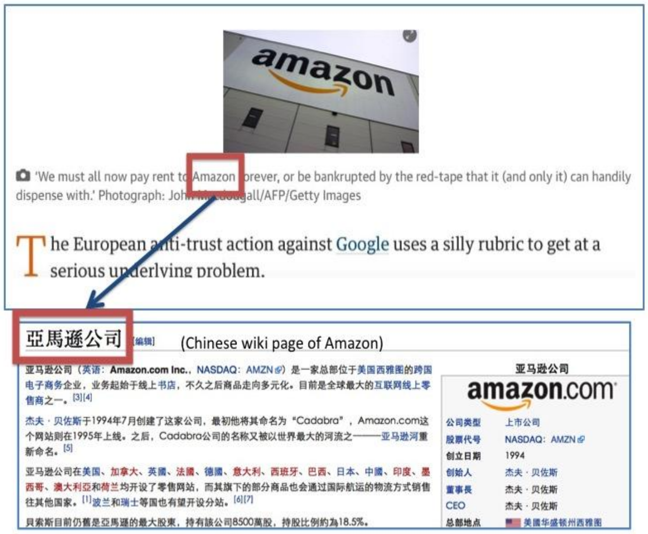
***Keywords*—Entity linking; keyphrase extraction; Wikipedia; Cross-language Link Discovery; linguistic difficulty level estimation**

I. INTRODUCTION

RECENTLY, because of the wide use of tablets and smartphones, the Internet services have become even more popular all over the world. Enormous amount of information is stored in a variety of languages. In addition, the hyperlinks on the Web link to many related information. However, related information is not always available in the native languages of the users. This causes difficulties for users to understand the documents written in foreign languages.

To solve this problem, it is desirable to find potential links automatically from documents written in their native languages (see Figure 1). Therefore, we propose a method to obtain Chinese encyclopedia articles that give the meaning of the keyphrases in a Japanese document.

Such a mechanism is useful for Chinese students studying Japanese, and it could further enhance the utility of the online encyclopedia, such as Wikipedia. Our proposed method aims to support knowledge discovery using the online encyclopedia as a learning tool for Chinese students studying Japanese. Considering the difference of foreign students' language proficiency, the system ought to recommend appropriate keywords. Therefore, the purpose of this research is to perform cross-language entity linking and provide appropriate information, adapting to each user's language ability.



'We must all now pay rent to Amazon forever, or be bankrupted by the red-tape that it (and only it) can handily dispense with.' Photograph: John MacDougall/AFP/Getty Images

The European anti-trust action against Google uses a silly rubric to get at a serious underlying problem.

亞馬遜公司 [编辑] (Chinese wiki page of Amazon)

亚马逊公司（英语：Amazon.com Inc.，NASDAQ：AMZN）是一家总部位于美国西雅图的跨国电子商务企业，业务起始于线上书店，不久之后商品走向多元化。目前是全球最大的互联网线上零售商之一。[3][4]

杰夫・贝佐斯于1994年7月创建了这家公司，最初他将其命名为"Cadabra"，Amazon.com这个网站则在1995年上线。之后，Cadabra公司的名称又被以世界最大的河流之——亚马逊河重新命名。[5]

亚马逊公司在美国、加拿大、英國、法國、德國、意大利、西班牙、巴西、日本、中國、印度、墨西哥、澳大利亞和荷兰均开设了零售网站，而其旗下的部分商品也会通过国际航运的物流方式销售往其他国家。[1]波兰和瑞士等国也有望开设分站。[6][7]

貝索斯目前仍舊是亞馬遜的最大股東，持有該公司8500萬股，持股比例約為18.5%。

亚马逊公司

公司类型	上市公司
股票代号	NASDAQ：AMZN
创立日期	1994
创始人	杰夫・贝佐斯
董事長	杰夫・贝佐斯
CEO	杰夫・贝佐斯
总部地点	美國华盛顿州西雅图

Figure 1: An example of cross-language entity linking

II. RELATED WORK

Wikipedia is a multilingual online encyclopedia in which anyone can use and edit on a Web browser. Articles for the same topic in different languages are usually linked via inter-language links. However, for some articles in some languages, there are no appropriate articles in different languages. A variety of previous research has been done so far, dealing with the same problem.

There are studies about Wikification [2] at NTCIR-9 and NTCIR-10 [3][4], which aims to reuse the Wikipedia resources effectively. According to Horita et al. [1], Wikification is a method for automatically extracting keyphrases from a document and linking them with an appropriate Wikipedia article. From NTCIR-9, one of the

Manuscript received January 8, 2017; revised February 7, 2017. This work was supported in part by JSPS KAKENHI Grant Numbers 24500300, 16K00452, and MEXT-Supported Program for the Strategic Research Foundation at Private Universities (S1511026).

J. Zhou is with Graduate School of Information Science and Engineering, Ritsumeikan University, Japan (e-mail: is0095hx@ed.ritsumei.ac.jp).

F. Kimura is with Faculty of Economics, Management and Information Science, Onomichi City University, Japan (e-mail: f-kimura@onomichi-u.ac.jp).

A. Maeda is with College of Information Science and Engineering, Ritsumeikan University, Japan (e-mail: amaeda@is.ritsumei.ac.jp).

related researches of Wikification is called CrossLink. CrossLink is a task aiming to automatically find potential links between online documents in different languages [2]. The linked text is called "anchor text" in the task of CrossLink. This task mainly focuses on extracting anchor texts from English Wikipedia and linking them to appropriate Wikipedia articles in languages of Japanese, Chinese, or Korean.

In the Text Analysis Conference (TAC) [5], the task of Cross-language Entity Linking (CLEL) was being performed. The purpose of this task is to extract PER (person), ORG (organization) and GPE (geopolitical entity) from Chinese or Spanish documents. Then, they link them to appropriate English documents. In this paper, the target entities are not limited to places or personal names, and it is the main difference of our proposed method from this task.

Wang et al. [6] proposed a cross-lingual knowledge linking approach for building cross-lingual links across Wikipedia knowledge bases. Their approach uses only language-independent features of articles and employs a graph model to predict new cross-lingual links.

Chen et al. [7] proposed an approach in which the first step is extracting n-grams from the query source documents as potential anchors. The next step is the anchor expansion and ranking. The final step of the anchor selection process is to re-rank anchors by computing the similarity between the title of the current query Wikipedia page and each element in the vector of expanded potential anchors using Wikipedia Miner [11]. Different from our work, they only use Google Translate to translate the potential anchors. In this paper, we translate the keyphrase using three methods and extract all the Chinese articles of the translated keyphrase. Finally, we make a ranking using cosine similarity comparison of the Japanese document and Chinese Wikipedia articles.

Liu et al. [8] divided their cross-language link discovery task into three sub-problems. Their approach has three steps: anchor mining, cross-lingual linking to related articles, and disambiguation. Similar with Chen et al.'s work, they choose Google Translate as the anchor translation tool. In this paper, we use three translation methods and make a ranking using a cosine similarity comparison of the Japanese document and Chinese Wikipedia articles. The difference between our approach and their work is that they use two different ways, which are Dice coefficient based and LDA model based measures to calculate the similarity of keyphrases. Furthermore, they apply the POS (part of speech) tag analysis module.

III. Proposed Method

In this section, we describe our method to detect an appropriate Chinese Wikipedia article for a keyphrase in a Japanese document. The proposed method consists of four processes, 1. Keyphrase extraction, 2. Translation, 3. Obtaining Chinese articles, and 4. Ranking Chinese candidate articles. We used Top Consecutive Nouns Cohesion (TCNC) method [1] for extracting keyphrase candidates. In this method, when consecutive nouns appear in a sentence, we adopt all possible binding patterns starting from the first noun. In other words, TCNC obtains all compound words that are the same in number as the number of consecutive nouns.

Figure 2 shows the overview of the proposed method. When extracting the keyphrase candidates, in order to extract new compound words, we use MeCab [12] to analyze Japanese document and by considering the surrounding contexts of the ketphrase, we can create links among nouns.

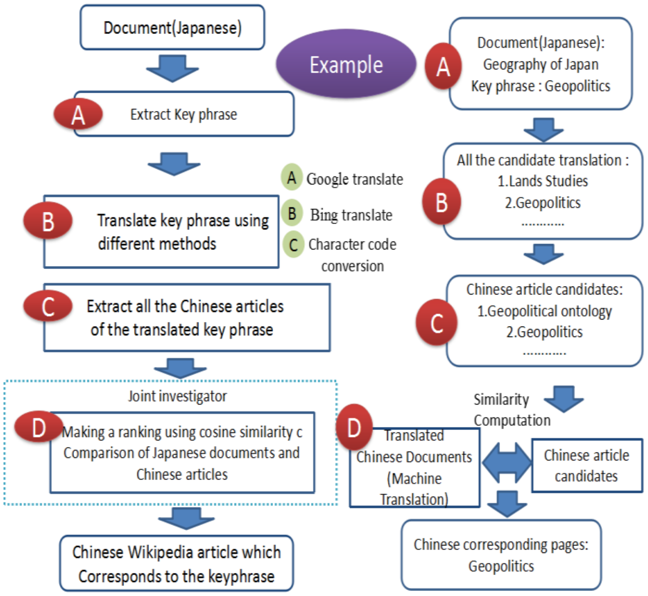

Document(Japanese)
Example
A
Document(Japanese):
Geography of Japan
Key phrase : Geopolitics
A
Extract Key phrase
A Google translate
B Bing translate
C Character code conversion
B
All the candidate translation :
1.Lands Studies
2.Geopolitics
....
B
Translate key phrase using different methods
C
Extract all the Chinese articles of the translated key phrase
C
Chinese article candidates:
1.Geopolitical ontology
2.Geopolitics
....
Joint investigator
D
Making a ranking using cosine similarity c Comparison of Japanese documents and Chinese articles
Similarity Computation
D
Translated Chinese Documents (Machine Translation)
Chinese article candidates
Chinese Wikipedia article which Corresponds to the keyphrase
Chinese corresponding pages: Geopolitics

Figure 2: Overview of the proposed method

A. *Keyphrase extraction*

Horita et al. [1] proposed a method for keyphrase extraction with two steps. Firstly, they conduct a morphological analysis and extract keyphrase candidates by Top Consecutive Nouns Cohesion (TCNC) method, which means combining the two characters and treating them as one compound word. Secondly, they rank the keyphrase candidates by Dice coefficient and Keyphraseness measures [1]　(see Figure 3).

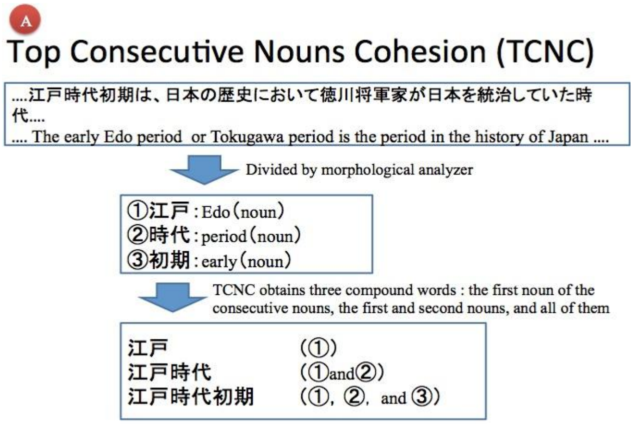

A
Top Consecutive Nouns Cohesion (TCNC)
....江戸時代初期は、日本の歴史において徳川将軍家が日本を統治していた時代....
.... The early Edo period or Tokugawa period is the period in the history of Japan ....
Divided by morphological analyzer
①江戸 : Edo (noun)
②時代 : period (noun)
③初期 : early (noun)
TCNC obtains three compound words : the first noun of the consecutive nouns, the first and second nouns, and all of them
江戸 (①)
江戸時代 (①and②)
江戸時代初期 (①, ②, and ③)

Figure 3: An example of TCNC

In this paper, we propose a method to extract appropriate keyphrases to users, considering user's Japanese proficiency. In this method, we first refer to the vocabulary list of Japanese-Language Proficiency Test (JPLT) [13] to find out the nonexistent words or phrases in the list, and then judge the degree of difficulty for these keyphrases. The next procedure is to tag the keyphrases with different linguistic difficulty level.

To be concrete, firstly, we use the vocabulary list of JPLT to categorize keyphrase candidates into different linguistic levels. For keyphrases that are not in the list, we try to estimate the linguistic level of them by using word2vec [9] trained on Japanese Wikipedia articles. Using word2vec, we can get vectors of the keyword's meaning from the surrounding contexts of that keyword. The vectors of the words with closely related meanings also have similarity scores. By using that information, we can estimate the linguistic levels of the keyphrases that are not found in the vocabulary list of JPLT (See Figure 4). For example, we extract the word "オバマ (Obama)" from a news article, to give it a tag, we can make use of words from the vocabulary list of JPLT. By calculating words' vector similarities, we can find out the word with the most similar meaning to the word "オバマ (Obama)", like "アメリカ(America)" and "大統領 (President)" with the score of 0.624 and 0.596. Then we know that Obama is the president of United States of America. Thus, we can label the word "オバマ(Obama)" with the same language proficiency level tag as Intermediate level.

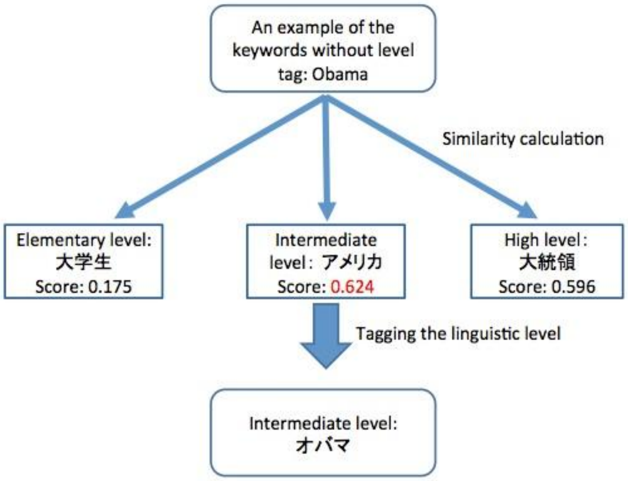

An example of the keywords without level tag: Obama
Similarity calculation
Elementary level: 大学生 Score: 0.175
Intermediate level: アメリカ Score: 0.624
High level: 大統領 Score: 0.596
Tagging the linguistic level
Intermediate level: オバマ

Figure 4: An example of estimating the linguistic level of a keyphrase

B. Translation using multiple methods

In the proposed method, we translate the Japanese keyphrases into Chinese using a combination of three translation methods, i.e., two machine translation services (Google Translate [14], Bing Translator [15]) and the character code conversion method [16] (see Figure 5) [10]. We consider all the obtained Chinese translations from all of three translation methods (including incorrect translations) as the translation candidates for a Japanese keyphrase. The reason of using all translation methods is to prevent missing the appropriate translation for the Japanese keyphrase. Incorrect translations will be eliminated in the ranking process, which will be explained in the Section 3.D.

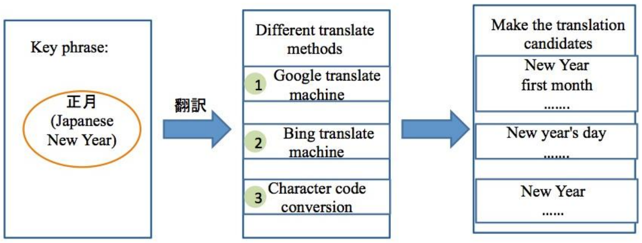

Key phrase:
正月 (Japanese New Year)
翻訳
Different translate methods
1 Google translate machine
2 Bing translate machine
3 Character code conversion
Make the translation candidates
New Year first month ....
New year's day ....
New Year ......

Figure 5: An example of keyphrase translation using different translation methods

C. Obtaining Chinese articles

In this process, we obtain the corresponding Chinese Wikipedia articles for each translation candidate obtained in the previous process. We use partial string matching between each obtained translation candidate and the titles of Chinese Wikipedia articles [10].

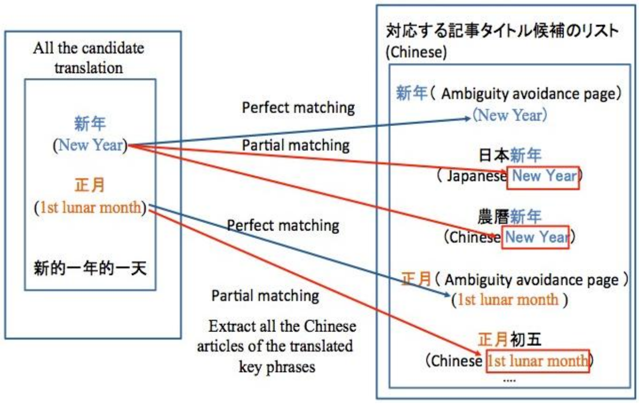

All the candidate translation
新年 (New Year)
正月 (1st lunar month)
新的一年的一天
Perfect matching
Partial matching
Perfect matching
Partial matching
Extract all the Chinese articles of the translated key phrases
対応する記事タイトル候補のリスト (Chinese)
新年 (Ambiguity avoidance page) (New Year)
日本新年 (Japanese New Year)
農曆新年 (Chinese New Year)
正月 (Ambiguity avoidance page) (1st lunar month)
正月初五 (Chinese 1st lunar month)
....

Figure 6: An example of partial string matching of translation candidates and Chinese Wikipedia article titles

The reason of performing partial string matching is that some compound words might be translated into appropriate translation partially, although all of them cannot be translated (see Figure 6). Finally, we obtain the Chinese Wikipedia articles that titles are partially matched to each translation candidate.

D. Ranking of Chinese candidate articles

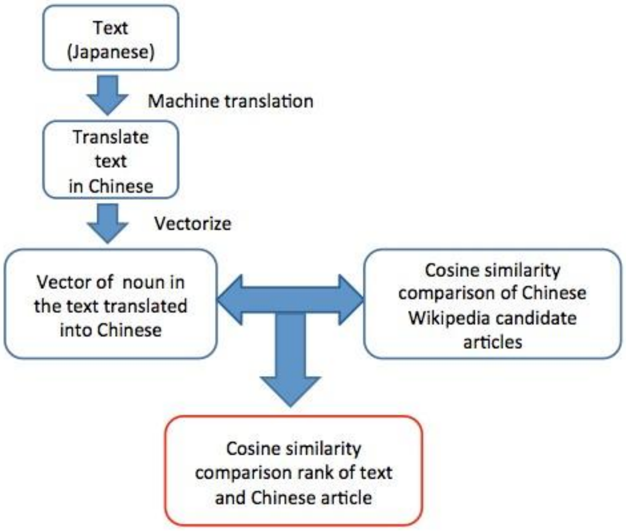


Figure 7: Processing flow of ranking of Chinese candidate articles

In this procedure, we rank obtained Chinese Wikipedia articles in order to figure out the most proper Chinese article for the Japanese keyphrase. Figure 7 shows the flow of this ranking procedure. First, we translate the original Japanese document of the keyphrase into Chinese. Second, we extract nouns from the translated Japanese document and obtain Chinese Wikipedia articles. Third, the frequencies of these nouns are regarded as a vector, and we calculate the cosine similarities of the original document and all the Chinese Wikipedia articles. Fourth, we adopt the highest ranked article as the corresponding Chinese article for the Japanese keyphrase. In this way, we can find the corresponding Chinese article for the Japanese keyphrase even if we obtain many Chinese articles as the candidates.

IV. EEPERIMENTS

To evaluate the effectiveness of the proposed method, we conducted a questionnaire survey with 15 Chinese students whose Japanese proficiency is N2 or N1 level. We selected 10 Japanese news articles from 5 different genres of Yahoo! Online News (see Table 1) and asked them to pick out words or phrases that they do not know as keyphrases. Next, we conducted an experiment to tag the linguistic level of the chosen keywords. From the results (see Table 2), we found that students whose Japanese proficiency are N1 level selected 63 words, and 6 words from the chosen words are found in the N1 vocabulary list of JPLT. As for the other 57 words, 35 words are tagged with level N1. Meanwhile, students whose Japanese proficiency are N2 level selected 84 words, and 4 of them are from the N1 vocabulary list of JPLT. Out of the other 80 words, 33 words are tagged with level N2. Then, we translated the selected 113 keyphrases and acquired 88 corresponding Chinese Wikipedia articles. Table 3 shows an example that has high similarity scores of tagging the keyphrase with similar meanings from level 1 vocabulary list. To evaluate proposed approach, we conduct experiment. The measurements of our experiment are recall, precision and F-measure. Table 4 shows the results of our experiment (see Figure 8 and Figure 9).

Table 1: Examples of news articles from each genre

The title of the news articles	Genre
ネットストーカートラブル (Internet stalker trouble)	IT
エコカー減税 (Tax reduction of using eco-cars)	地域 (Local News)
バース退団 (Bass left the baseball team)	スポーツ (Sports News)
ＮＺ地震の復興作業 (New Zealand Earthquake disaster reconstruction)	経営 (Business News)
ロヒンギャ迫害悪化 (Rohingya refugee problem)	国際 (International News)

Table 2: The results of the experiment

Chinese students' Japanese proficiency level	N1	N2
The number of Chinese students	15	15
Selected words	63	84
The words that are not in the vocabulary list of JPLT	57	80
The words that are successfully tagged	41	33
The ratio of successful tagging	0.719	0.412

Table 3: The examples of successful tagging of the keyphrase with similar meanings

Keyphrase	Word with similar meaning from level N1	Similarity score
アイドルグループ (Idol Group)	タレント (talent)	0.790
	ソロ (solo)	0.668
	レディー (lady)	0.315
	待望 (expectation)	0.308

Table 4: The results of evaluation experiment

	Recall	Precision	F-measure
Keywords with N1 Level label	0.607	0.561	0.489
Keywords with N2 Level label	0.642	0.490	0.529

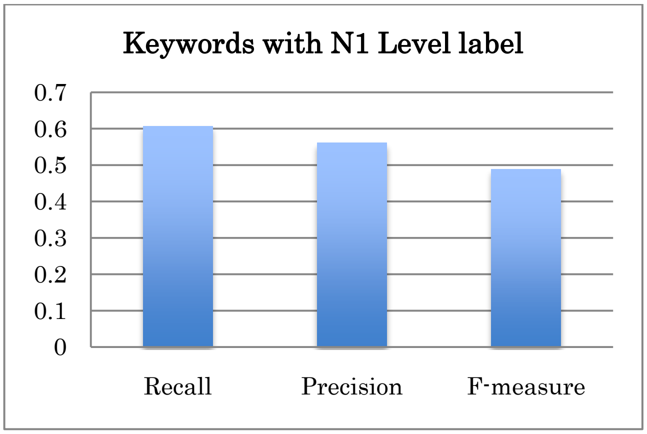


Figure 8: The results of evaluation experiment (1)

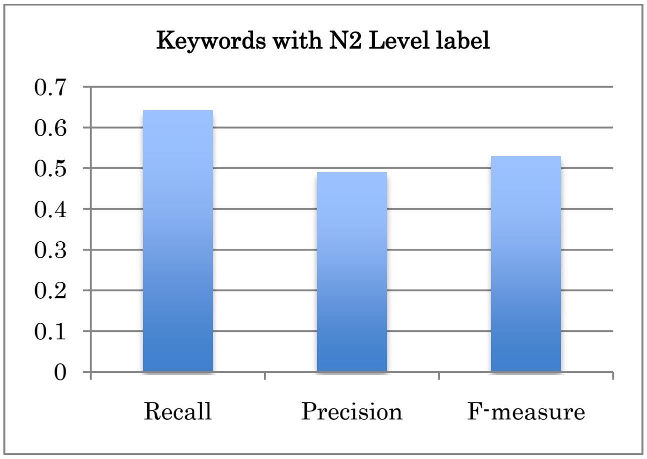


Figure 9: The results of evaluation experiment (2)

We classified the results into four categories based on the results of the translation and the acquired corresponding article candidates as shown in Table 5.

Table 5: The classification of the results of the acquired translation and article candidates

Category	Whether correct translation was extracted or not	Whether correct article was acquired or not
1	Yes	Yes
2	Yes	No
3	No	Yes
4	No	No

The results of the correct ratio of acquiring relevant Chinese candidate articles for 113 Japanese keyphrases by the proposed method are shown in Table 6. The results are classified into one of the categories in Table 5.

Table 6: Experimental results of acquiring relevant Chinese corresponding articles from Japanese keyphrases

Translation methods	1	2	3	4
MT (Bing)	0.54 (62/113)	0.02 (3/113)	0.06 (7/113)	0.36 (41/113)
MT (Google)	0.59 (67/113)	0.03 (4/113)	0.06 (7/113)	0.30 (35/113)
Character code conversion	0.28 (32/113)	0.03 (4/113)	0.0 (0/113)	0.68 (77/113)
Combination of three translation methods	0.63 (72/113)	0.044 (5/113)	0.08 (9/113)	0.07 (4/55)

In the process of acquiring the corresponding Chinese articles, although we achieved results with high accuracy, we still got a lot of non-relevant articles. It is necessary to explore other methods to reduce the number of non-relevant articles.

V. Conclusion

In this paper, we proposed a method for keyphrase extraction from Japanese documents considering the user's language ability in cross-language entity linking from Japanese to Chinese.

From the questionnaire survey, we found that Chinese students have difficulty in understanding keyphrases from popular phrases and topics in Japanese news articles. This makes it harder to find Chinese Wikipedia articles that are suitable for students with different linguistic proficiency levels. Further research is needed to solve this problem in the future. Furthermore, machine translation services are primarily designed for translating sentences, rather than translating individual keyphrases as our current method does. We are planning to consider the surrounding contexts of the extracted keyphrase for resolving the word ambiguity problem.

REFERENCES

[1] Horita, K., Kimura, F., and Maeda, A., “Automatic Keyword Extraction for Wikification of East Asian Language Documents,” *International Journal of Computer Theory and Engineering*, Vol. 8, No. 1, pp. 32-35, 2016.

[2] Mihalcea, R., and Csomai, A., “Wikify!: linking documents to encyclopedic knowledge,” in *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, pp. 233-242, 2007.

[3] Tang, L.X., Geva, S., Trotman, A., Xu, Y., and Itakura, K.Y., “Overview of the NTCIR-9 Crosslink Task: Cross-lingual Link Discovery,” in *Proceedings of the 9th NTCIR Conference*, pp. 437-463, 2011.

[4] Tang, L.X., Kang. I.S., Kimura, F., Lee, Y.H., Trotman, A., Geva, S., and Xu, Y., “Overview of the NTCIR-10 Cross-lingual Link Discovery Task,” in *Proceedings of the 10th NTCIR Conference*, pp. 8-38, 2013.

[5] Heng, J., Nothman, J., and Hachey, B., “Overview of TAC-KBP2014 entity discovery and linking tasks,” in *Proceedings of TAC2014*, 2014.

[6] Wang, Z., Li, J., Wang, Z., Tang, J., “Cross-lingual knowledge linking across wiki knowledge bases,” in *Proceedings of the 21st International conference on World Wide Web*, pp. 459-468, 2012.

[7] Chen, S., Jones, G.J.F., O’Connor, N.E., “DCU at NTCIR-10 Cross-lingual Link Discovery (CrossLink-2) Task,” in *Proceedings of the 10th NTCIR Conference*, pp. 74-78, 2013.

[8] Liu, Y, Boisson, J, Chang, J.S., “NTHU at NTCIR-10 CrossLink-2: An Approach toward Semantic Features,” in *Proceedings of the 10th NTCIR Conference*, pp. 62-68, 2013.

[9] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J., “Distributed representations of words and phrases and their compositionality,” in *Proceedings of Advances on Neural Information Processing Systems 26 (NIPS 2013)*, pp. 3111-3119, 2013.