Automatic Keyword Extraction for Wikification of East Asian Language Documents

Kensuke Horita, Fuminori Kimura, and Akira Maeda

***Abstract*—In recent years, research on Wikification, which aims to promote the effective reuse the Wikipedia resources and the understanding of document contents, is attracting much attention. Wikification is a method to automatically extract keywords from a document, and to link them to an appropriate Wikipedia article. Wikification consists of two processes. First, we extract keywords from a document. Second, we identify the appropriate Wikipedia article for each of them. In this paper, we focus on the extraction of keywords from a document for Wikification. Research on Wikification has been conducted for documents in variety of languages. We focus on East Asian language documents and experiment with Japanese documents. Besides, we are planning to do the Wikification not only for documents in the same language but also for other languages (e.g. keywords in Japanese documents are linked to appropriate English Wikipedia articles).**

Our proposed method consists of two steps. First, we extract nouns from a document using a morphological analysis tool, and extract candidate keywords by a method called Top Consecutive Nouns Cohesion (TCNC). The TCNC connects continuous nouns and treat them as one compound word. Second, we rank the extracted candidate keywords using one of two measures for keyword importance, Dice coefficient or Keyphraseness.

In our experiments of extracting appropriate keywords for Wikification in Japanese documents, our proposed method, especially the combination of TCNC and Keyphraseness, achieved the best results.

***Index Terms*—Wikipedia, wikification, keyword extraction, compound word.**

I. Introduction

Wikipedia is a free content Internet encyclopedia. Anyone who can access this site can edit its articles. This site has articles in over 280 languages and 30 million articles in total.

The problem is that it is difficult for readers to understand a document when they cannot understand important words in the document. Wikification aims to promote the effective reuse the Wikipedia resources and to solve this problem. Wikification is a method to automatically extract keywords from a document, and to link them to an appropriate Wikipedia article. An example of Wikification is to extract a keyword "quantitative precipitation forecasting" from a document and link it to the appropriate Wikipedia article "Quantitative precipitation forecast" as shown in Fig. 1. Our final goal is to link the keywords into appropriate Wikipedia articles not only in the same language of the document but also in different languages (e.g. Wikipedia in English). In this paper, we focus on extracting keywords from East Asian language documents for Wikification.

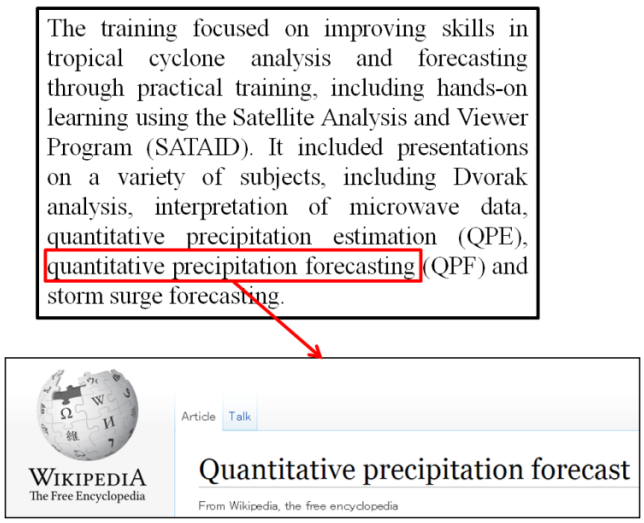

Fig. 1. An example of Wikification.

II. Related Work

The purpose of Wikification [1] is to extract keywords from English documents and to link them to appropriate Wikipedia articles. It is shown to be useful for various situations, such as in education [2].

One of the related research of Wikification is CrossLink, which is one of the tasks of NTCIR-9 and 10 [3], [4]. The aim of the task is to automatically find potential links between documents in different languages. Potential links are the link texts that are described in a Wikipedia article in a different language. That link text is called "anchor text" at CrossLink. This task focuses on Wikipedia, and the purpose of the NTCIR-9 CrossLink task is to extract anchor texts from English Wikipedia and link them to appropriate Wikipedia articles in languages of Japanese, Chinese, or Korean. In NTCIR-10, the task included the opposite language direction, i.e., to extract anchor texts from Japanese, Chinese, or Korean Wikipedia articles and link them to appropriate English Wikipedia articles. Kim and Gurevych [5] used word N-gram for extracting anchor texts, and that achieved good results in NTCIR-9. CLEL (Cross-Lingual Entity Linking), which is one of the tracks of KBP (Knowledge Based Population), is held at TAC (Text Analysis Conference) [6]. The purpose of CLEL is to extract PER (person), ORG

Manuscript received September 5, 2014; revised November 18, 2014.

K. Horita is with the Graduate School of Information Science and Engineering, Ritsumeikan University, Shiga, Japan (e-mail: is0038ep@ed.ritsumei.ac.jp).

F. Kimura is with Kinugasa Research Organization, Ritsumeikan University, Kyoto, Japan (e-mail: fkimura@is.ritsumei.ac.jp).

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

(organization) and GPE (geopolitical entity) from Chinese or Spanish documents and to link them to appropriate English documents. We propose a keyword extraction method that are not limited to PER, ORG and GPE.

III. Proposed Method

In this section, we describe our proposed method for extracting keywords from documents written in Japanese. The proposed method of keyword extraction consists of the following two steps.

- Candidate keyword extraction
- Ranking extracted candidate keywords

A. Candidate Keyword Extraction

The keyword extraction methods often use heuristic rules [7]. For example, removing stop words using a stop word list [8], using part of speech tags [9], and using N-grams [10]. Yoshida and Nakagawa [11] proposed a keyword extraction method for Japanese language documents. This method uses perplexity on the term unit's left-side and right-side terms for extracting technical terms. In our proposed method, we extract not only technical terms but also various keywords that will be useful for general users.

We used Top Consecutive Nouns Cohesion (TCNC) method, which has been proposed in our previous work [12], for extracting candidate keywords. 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. For example, when three consecutive nouns appear in a sentence, TCNC obtains three compound words: the first noun of the consecutive nouns, the first and second nouns, and all of them (see Fig. 2). All of these obtained compound words are treated as keyword candidates. Using TCNC, we can reduce significant amount of noise that the word N-gram method derives.

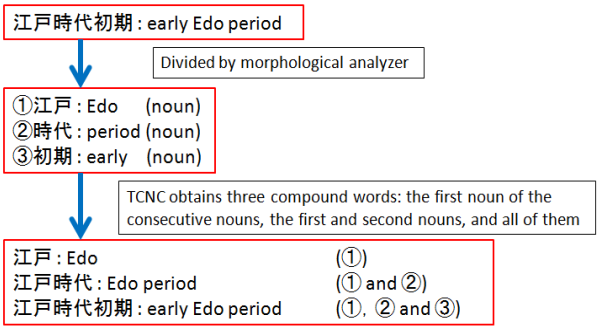

Fig. 2. An example of TCNC.

B. Ranking Extracted Candidate Keywords

TCNC extracts many candidate keywords. If all of them are used to link texts in a document, we may not understand which keyword is important. Therefore, we need to rank them by some importance measure. In this paper, we use one of two importance measures, Dice coefficient or Keyphraseness.

1) Dice coefficient

The Dice coefficient considers how often two words co-occur. The equation of Dice coefficient is shown in Eq. 1. In this paper, the value of the Dice coefficient is higher the more the candidate keyword q in the original article j and the title of the original article T_i appears in the original article j. The purpose of using the Dice coefficient is to give a higher rank to the terms that have a relationship with the title of the original article. The Dice coefficient $S(T_i, q)$ is calculated by the following formula:

$$S(-T_i,q) = \frac{2\sum_{j=1}^{M} w_{ij} w_{qj}}{\sum_{j=1}^{M} w_{ij} + \sum_{j=1}^{M} w_{qj}} \quad (1)$$

$$w_{ij} \begin{cases} 1 & \text{title } T_i \text{ appears in Japanese title of article } j \\ 0 & \text{do not appear} \end{cases}$$

$$w_{qj} \begin{cases} 1 & \text{title } q \text{ appears in Japanese title of article } j \\ 0 & \text{do not appear} \end{cases}$$

where M is the number of all Japanese Wikipedia articles and q is the candidate keyword in the original article. T_i is the title of the original article i.

2) Keyphraseness

Keyphraseness measures how often the keyword is used as the link text in Wikipedia articles. The equation of Keyphraseness is shown in Eq. 2. The value of the Keyphraseness is higher the more the candidate keyword q in the Japanese Wikipedia articles. In the equation below, *count* (D_q) is the number of the keyword appeared in Japanese Wikipedia articles. *Count* (D_{lq}) is the number of the keyword appeared as a link text in Japanese Wikipedia articles. Keyphraseness value will be inaccurately large when *count* (D_q) is very few. Therefore, we used only q that appeared at least five times in Japanese Wikipedia articles.

$$P(keyword) \mid q) \approx \frac{count(D_{lq})}{count(D_q)} \quad (2)$$

IV. Experiments

We conducted experiments to verify the effectiveness of the proposed method. In the experiments, we used Japanese Wikipedia articles used in CrossLink-2 of NTCIR-10. We randomly chose 10 articles for test data from that article set. We used "Overview" section and introductory part in the articles. We regarded existing link texts appearing in the articles of test data as correct keywords. The test data contains 296 link texts. It is found that about 6% of the total words are used as link texts in articles of Wikipedia on average [1]. Therefore, we selected top 6% of the extracted candidate keywords as the keywords. We tested the combinations of two keyword extraction methods and two importance calculation methods. Thus, we conducted four

runs, baseline and Dice coefficient, baseline and Keyphraseness, TCNC and Dice coefficient, TCNC and Keyphraseness. The baseline is the keyword extraction method that only uses an extracted noun as a keyword. We judged whether the keywords selected by each run are used as the link texts in the test data. If the keyword is used as the link text, we consider it is correct. We used F-measure for evaluating the effectiveness of each run. F-measure is the harmonic mean of precision and recall. The equation of precision is shown in Eq. 3, and the equation of recall is shown in Eq. 4. In the equations, *count* (*extracted keywords*) is the number of extracted keywords, *count*(*correct keywords*) is the number of extracted correct keywords, and *count*(*all correct keywords*) is the number of all correct keywords in the article.

$$Precision = \frac{count(correct\ keywords)}{count(extracted\ keywords)} \quad (3)$$

$$Recall = \frac{count(correct\ keywords)}{count(all\ correct\ keywords)} \quad (4)$$

TABLE I: AVERAGE F-MEASUARES OF FOUR RUNS

Keyword extraction	Baseline		TCNC	
Importance measures	Dice coefficient	Keyphraseness	Dice coefficient	Keyphraseness
F-measure (average)	0.3338	0.3834	0.36919	0.5886

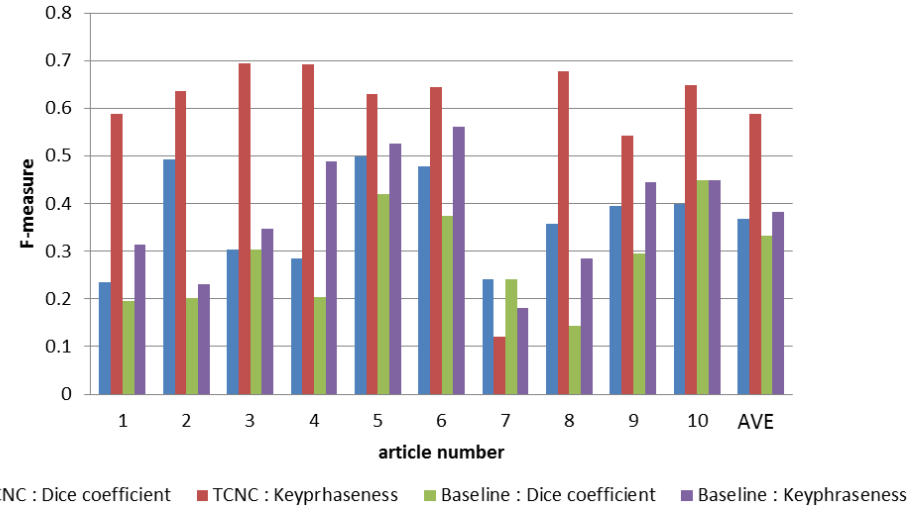

Fig. 3. The results of F-measures of four runs for each article.

TABLE II: PERCENTAGE OF COMPOUND WORDS IN TEST ARTICLES

Article number	Percentage of compound words
1	62%
2	75%
3	60%
4	36%
5	28%
6	26%
7	0%
8	59%
9	35%
10	33%
average	43%

V. DISCUSSION

The results of the average F-measures of four runs are shown in Table I. The percentages of compound words in the test articles are shown in Table II. The results of F-measures of four runs for each article are illustrated in Fig. 3. As can be seen in Fig. 3, the run of the combination of TCNC and Keyphraseness achieved the best score except for article 7. Many extracted keywords are judged as incorrect because article 7 has only a few link texts. Note that all other runs had also low score for article 7. Besides, as shown in Table II, article 7 does not have any link text of compound word at all, therefore TCNC did not work well and was lower than the baseline.

In contrast, in article 2, 75% of the correct keywords are compound words. This percentage is higher than the average, which is 43%. In this case, TCNC achieved much higher score than the baseline that cannot extract compound words. Similarly, articles 1, 2, and 8 have significantly more compound words than the average. Therefore, in these articles, TCNC achieved especially high scores. Conversely, articles 5, 6, and 10 have fewer compound words, therefore there were not much difference between the baseline and TCNC.

In fact, TCNC extracts some redundant keywords, however, it enables TCNC to extract compound words that are regarded as keywords. So precision of TCNC increased than the baseline. In addition, as shown in Table III, some keywords that can be considered important in that article were also extracted at the top rank of the combination of Keyphraseness and TCNC.

In this experiment, we regarded existing link texts in the articles of test data as correct keywords. However, other keywords that are not linked might also be important although they are not contained in the same language version of Wikipedia articles, as shown in Table III. Therefore, we consider that only using existing link texts as correct

keywords might not always be appropriate, and we might consider employing manual judgments for more reliable evaluation.

TABLE III: EXTRACTED KEYWORDS THAT ARE JUDGED INCORRECT BUT CAN BE REGARDED AS CORRECT

Article number	Title of the article	Extracted keywords
7	シミュレーション (simulation)	プログラム (program) デフォルメ (deform) 模倣 (imitation)

VI. CONCLUSION

In this paper, we proposed a method for extracting keywords from Japanese Wikipedia articles for Wikification. Wikification is a method to automatically extract keywords from a document, and to link them to an appropriate Wikipedia article.

Our proposed method consists of two steps. First, we conduct a morphological analysis and extract candidate keywords by Top Consecutive Nouns Cohesion (TCNC) method. The TCNC connects continuous nouns and treat them as one compound word. Second, we rank the extracted candidate keywords by one of two importance measures, Dice coefficient or Keyphraseness.

We conducted experiments to verify the effectiveness of the proposed method using 10 articles in Japanese Wikipedia. As the result, the combination of our proposed TCNC and Keyphraseness achieved the best score in the experiments.

VII. FUTURE WORK

One of the future works is to develop a method to find an appropriate Wikipedia article for the extracted keywords and link to them. It is also important to improve the effectiveness of keyword extraction. Besides, although we applied the proposed method to only Japanese documents in this paper, our proposed method can also be applied to documents written in other East Asian languages than Japanese, so we need to do experiments for other East Asian language documents.

REFERENCES

[1] R. Mihalcea and A. Csomai, "Wikify!: Linking documents to encyclopedic knowledge," in *Proc. the 16th ACM Conference on Information and Knowledge management*, Portugal, 2007, pp. 233-242.

[2] A. Csomai and R. Mihalcea, "Linking educational materials to encyclopedic knowledge," *Frontiers in Artificial Intelligence and Applications*, IOS Press, vol. 158, pp. 557-559, USA, 2007.

[3] L.-X. Tang, S. Geva, A. Trotman, Y. Xu, and K. Y. Itakura, "Overview of the NTCIR-9 crosslink task: Cross-lingual link discovery, " in *Proc. the 9th NTCIR Conference on Evaluation of Information Access Technologies*, Japan, 2011, pp. 437-463.

[4] L.-X. Tang, I. S. Kang, F. Kimura, Y. H. Lee, A. Trotman, S. Geva, and Y. Xu, "Overview of the NTCIR-10 cross-lingual link discovery task," in *Proc. the 10th NTCIR Conference on Evaluation of Information Access Technologies*, Japan, 2013, pp. 8-38.

[5] J. Kim and I. Gurevych, "UKP at CrossLink: Anchor text translation for cross-lingual link discovery," in *Proc. NTCIR-9 Workshop Meeting*, Japan, 2011, volume 9, pp. 487-494.

[6] TAC KBP 2013 entity linking track. [Online]. Available: http://www.nist.gov/tac/2013/KBP/EntityLinking/index.html

[7] K. S. Hasan and V. Ng, "Automatic Keyphrase extraction: A survey of the state of the art," in *Proc. the 52nd Annual Meeting of the Association for Computational Linguistics*, pp.1262-1273, USA, 2014.

[8] Z. Liu, P. Li, Y. Zheng, and M. Sun, "Clustering to find exemplar terms for Keyphrase extraction," in *Proc. the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, 2009, pp. 257-266.

[9] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," in *Proc. the 2004 Conference on Empirical Methods in Natural Language Processing*, Spain, 2004, pp. 404-411.

[10] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning, "KEA: Practical automatic Keyphrase extraction," in *Proc. the 4th ACM Conference on Digital Libraries*, ACM Press, 1999, pp. 254-255.

[11] M. Yoshida and H. Nakagawa, "Automatic term extraction based on perplexity of compound words," in *Proc. IJCNLP*, 2005, pp. 269-279.

[12] F. Kimura, K. Horita, Y. Konishi, H. Harada, and A. Maeda, "RDLL at CROSSLINK anchor extraction considering ambiguity in CLLD," in *Proc. the 10th NTCIR Conference on Evaluation of Information Access Technologies*, Japan, 2013, pp. 82-86.

Kensuke Horita is a master's student at the Graduate School of Information Science and Engineering, Ritsumeikan University, Shiga, Japan. He is interested in extraction keywords in Asian language documents.

Fuminori Kimura is a senior researcher in Art Research Center, Ritsumeikan University. He obtained the Ph.D. degree in engineering from Nara Institute of Science and Technology in 2007. His research interests include information retrieval, and text mining and multilingual information processing

Akira Maeda is a professor at the Department of Media Technology, College of Information Science and Engineering, Ritsumeikan University. He received the B.A. and M.A. degrees in library and information science from University of Library and Information Science (ULIS) in 1995 and 1997, respectively, and received the Ph.D. degree in engineering from Nara Institute of Science and Technology (NAIST) in 2000. He has visited Virginia Polytechnic Institute and State University (Virginia Tech) from October 2000 through March 2001 as a postdoctoral visiting scholar. He has worked as a postdoctoral researcher of the CREST program, Japan Science and Technology Corporation (JST) from April 2001 through March 2002. He was a visiting professor at King's College London from September 2011 through September 2012. His research interests include digital libraries, digital humanities, information retrieval, and multilingual information processing.