

Using Information Extraction and Search Engines for Automatic Detection of Inadequate Descriptions and Information Supplements in Japanese Wikipedia

Masaki Murata¹, Naoya Nonami¹, and Qing Ma²

¹ Tottori University, 4-101 Koyama-Minami, Tottori 680-8552, Japan,
{murata,s132043}@ike.tottori-u.ac.jp

² Ryukoku University, Seta, Otsu, Shiga 520-2194, Japan
qma@math.ryukoku.ac.jp

Abstract. When sentences lack important information that the reader wishes to know, they are difficult to read. Thus, correcting inadequate sentences is easier if there is a technique for pointing out inadequate descriptions and adding the missing information. Akano et al. constructed such a technique to detect inadequate descriptions using information extraction; however, the technique did not fill in the missing information. Therefore, this study was conducted, wherein a web search engine was used to provide appropriate information to inadequate sentences. In our method, a search engine gathers webpages and extracts important information from them. In our experiments, the MRR (mean reciprocal rank) and top-five accuracy of measuring important information for inadequate sentences were 0.77 and 0.96, respectively, under a certain experimental condition, which eliminated cases wherein correct answers were not found in webpages. The f-measure for this very difficult task of detecting inadequate descriptions and correct expressions in webpages was 0.77, which indicates that our method was very effective.

Keywords: inadequate description, information supplement, Web search engine, Japanese Wikipedia, information extraction

1 Introduction

When sentences lack important information that the reader wishes to know, they are difficult to read. Thus, correcting inadequate sentences is easier if there is a technique for pointing out inadequate descriptions and adding the missing information.

This study treated items that commonly appear in many entries in the Japanese Wikipedia as important items and information from these entries were extracted, arranged, and displayed in the form of a table. Any blank cells in the table are set as inadequate descriptions wherein important information is missing. Important information is specific information that is included in the important items of the table. For example, in Table 1, the important items,

Table 1. Blanks in the table correspond to missing parts (inadequate descriptions).

	Location	Person	Organization
Uwajima Castle	Uwajima		

Table 2. Missing parts in the table that are filled in

	Location	Person	Organization
Uwajima Castle	Uwajima	(Takatora Todo)	(Uwajima clan)

“personal name” and “organization name,” are missing parts that are not described in the corresponding Wikipedia entry on a castle. By pointing out these missing parts, we can modify the adequate descriptions easily. Although it is useful to point out that there are missing parts in the table, it is more helpful to correct and complement the adequate descriptions if we can fill in the missing parts of the table, as shown in Table 2. The words in parentheses in Table 2 indicate complementary information.

Therefore, in this study, we used a search engine to fill in the appropriate information for the missing parts. In this study, supporting the correction of sentences by filling in the missing parts with the appropriate information corresponds to text correction support. Referring to the information supplements for the missing parts is essential while correcting inadequate sentences with missing parts.

This study has the following characteristics.

- Originality: using information extraction to detect inadequate descriptions in Wikipedia and using a search engine to complement inadequate descriptions. Our purpose is to detect inadequate descriptions in Wikipedia and correct them with a search engine. We extract important information from Wikipedia by extracting clustered words and compiling the extracted information into a table. A blank cell in the table is considered as a missing part. We complement the missing parts using webpages found with the search engine. Our method’s originality lies in our complementing the missing parts in addition to detecting them. Although a previous paper [1] also focused on the detection of missing parts, it did not complement them.
- Information extraction from webpages
Fifty items were acquired using the search engine and important information was extracted from the webpages by clustering. We compiled the extracted information into a table. The top-five accuracy of the extracted information was 0.66.
- Supplementation of missing parts
We checked how accurately we complemented the missing parts of the table. In the experiments, the MRR and top-five accuracy for showing important information for inadequate sentences were 0.77 and 0.96, respectively, in eliminating cases wherein the correct answers were not in the webpages. The f-measure for the very difficult task of detecting inadequate descriptions and detecting correct expressions in the webpages was 0.77; therefore, our method was very effective.

2 Related works

Akano et al. [1] extracted important information using clustered words and summarized it in a table. Blank spots in the table were considered as missing parts (inadequate descriptions). The user was informed of these missing parts and asked to add descriptions to support the writing preparation. Although the previous study pointed out the missing parts in a table, there has been no study on a method for asking the user to fill in the missing information. Therefore, in this study, we used a search engine to acquire the information to fill in the missing parts and support text correction.

Murata et al. [4] improved the accuracy of the question answering system by a method of using scores in multiple documents. Their study and our study are similar in that they both detect answer expressions using document retrieval but Murata et al. focused on the question-answering system, whereas we handled the writing preparation support.

Okada et al. [6] handled text preparation support by the automatic detection of information that should be present in a paper. They defined the information, such as research results, effectiveness, and necessity, that should be described in papers as items requiring mention (IRM). Okada et al. provided sentence creation support by automatically detecting the missing IRMs. In their study, they supported writing by automatically detecting whether important items were missing. Their system is similar to ours in the use of important items for the detection of inadequate descriptions that lack important information. However, the studies differ in that their research focused on sentences in an academic paper, whereas our research focused on sentences in Wikipedia. Unlike the system of Okada et al., our system implements word clustering for using important items and complementing missing parts.

Fukuda et al. [2] constructed a system that extracts expressions, indicating the effects and trends of technology from research papers, and compiled the extracted information. Ptaszynski et al. [7] developed a system to support the writing of research papers by preparing data and automatically conducting the experiments to obtain accuracies. Using LaTeX templates, the system creates tables containing all the results and graphs these results. Their study is useful for the surveying and preparation of research papers. However, these two systems were not intended to support the correction of inadequate portions of sentences in a document, whereas our system provides support for correcting inadequate descriptions in Wikipedia.

A study by Nadamoto et al. proposed a technique that recognizes missing parts [5]. They observed that discussions in community-type content, such as social networking services or blogs, may concentrate on a small domain, and thereby, miss some viewpoints. They have termed the missed viewpoints as *content holes* and proposed a method of detecting such occurrences by comparing discussions in community-type content and general information such as the sort of material that appears in Wikipedia. Regarding the detection of missing information, the study of Nadamoto et al. and our study are similar but their study focused on community-type contents, whereas ours focused on descriptions in a document.

Tsuda et al. [8] conducted a study of the automatic detection of redundant sentences to support the writing of sentences. They proposed a method of using machine learning to detect redundant sentences automatically. Many other systems support the writing of sentences, but other than those by Akano et al. and Okada et al., few detect inadequate descriptions with important items.

3 Proposed method

Our method detects inadequate descriptions in documents (i.e. Wikipedia pages) and extracts expressions useful for complementing the inadequate descriptions using search engines.

The method proposed by this study comprises two stages: (i) extraction of important information from documents (Wikipedia pages), and (ii) information extraction using search engines.

Using the clustering tool in word2vec [3], we extract important items from entries in the Japanese Wikipedia about castles.

3.1 Extraction of important information from documents (Wikipedia pages)

The method described here is based on that of Akano et al. [1].

We extract important items from the entries in Wikipedia about castles using the clustering tool in word2vec. Extraction is done for each entry. We use “word clustering” in word2vec to select important items related to the extracted data. Word clusters are created by grouping highly similar words. We assign a number to each cluster, select the important items manually, and compile the items into a table. The method of compiling the information into a table is described below.

1. We determine what we want to extract. We extract pages containing things we determine from Wikipedia.
2. Using the clustering function of word2vec, we cluster words in the extracted data. A number is assigned to each cluster. Similar word groups are put into clusters. For example, word groups of locations and personal names are grouped into Clusters No. 1 and 2, respectively. Examples are shown in Tables 3 and 4.
3. A cluster of words corresponds to a column and a page of extracted data corresponds to a row in a table. We place a word belonging to a cluster appearing on a page into the corresponding row and column positions. If more than one word of a cluster appears on a page, we fill all those words in that place of the table.
4. We find the total number of words (called Frequency A) in each column of the table. We sort the columns in the table so that a column with a higher Frequency A is on the left. We delete the columns of clusters with less Frequency A.
5. We manually select columns (important items) that are considered to be important information on castles from the columns of the cluster numbers with

Table 3. Words corresponding to locations (Cluster No. 1)

Location
Kyoto
Osaka
Miyagi

Table 4. Words corresponding to personal names (Cluster No. 2)

Name
Masamune Date
Ieyasu Tokugawa
Hideyoshi Toyotomi

Table 5. Arranged table

Castle	Location	Person
Osaka Castle	Osaka	Hideyoshi Toyotomi
Nijo Castle	Kyoto	Ieyasu Tokugawa
Sendai Castle	Miyagi	Masamune Date

high Frequency A according to the sorted table. We delete the unselected columns and make a table. Table 5 is an example of a table created in this way.

For example, we assume that the entry on Osaka Castle contains the following sentences.

*oosaka jo wa oosaka ni shozai suru. toyotomi hideyoshi ga kizuita
oosaka jo no ikou wa, genzai subete maibotsu shiteiru.*

(Osaka Castle is located in Osaka. The remains of Osaka Castle, built by Toyotomi Hideyoshi, are now all buried.)

Because the location name, Osaka, and the personal name, Toyotomi Hideyoshi, appear on the page, “Osaka” is extracted and placed into a word cluster of location names and “Toyotomi Hideyoshi” is extracted and placed into a word cluster of personal names. A table such as Table 5 is created.

3.2 Information extraction using search engine

A blank in the table created by the method described in Section 3.1 indicates that the corresponding page does not contain important information. Therefore, we acquire information to fill in the blank using the search engine to complement the table.

We first input the name of a castle as a query to the search engine, which returned 50 webpages. We extracted important information from these using the method described in Section 3.1. Among the extracted information, the five most frequent words were compiled into a table, which was presented to a user to correct the sentences.

Table 6. Table with missing parts provided by extracting information from Wikipedia

Castle	Prefecture	Age	Location	Era name
Nezoe Castle	Miyagi		Sendai	

Table 7. Table complementing missing parts provided by extracting information from webpages

Castle	Prefecture	Age	Location	Era name
Nezoe Castle	Miyagi	(Heian age)	Sendai	(Eisho)

An example of extracting important information from webpages and compiling it in a table is shown in Tables 6 and 7.

Table 6 is constructed by the method of extracting information from Wikipedia pages, as described in Section 3.1. In the Wikipedia entry for Nezoe Castle, no information exists for the age and era name. Therefore, the columns for the age and era name of Nezoe Castle are left blank to indicate the missing parts.

We construct Table 7 from Table 6 by using a method for extracting information from webpages. We input “Nezoe Castle” as a query into the search engine, which extracts “prefecture”, “age”, “location”, and “era name” from the returned webpages and compile the information into a table. The following sentences are excerpts from one page randomly selected from the 50 returned pages.

nezoe jo wa heian jidai ni sakaeta abeshi no yakata to sareru. eisho 6 nen (1051) ni hajimaru zenkunen no ekide, genji no gunzei ga nezoe jo wo kougekishita.

(Nezoe Castle was a castle used by Mr. Abe, who flourished during the Heian age. Genji’s army attacked Nezoe Castle during the Zenkunen War in the sixth year of Eisho (A.D. 1051).)

Because “Heian age,” which is the name of an age, and “Eisho,” which is the name of an era, appear in the document, these words are extracted and output to the table. When comparing Tables 6 and 7, Table 7 can contain “age” and “era name,” whereas Table 6 cannot do so. Since important information not described in Wikipedia can be acquired from webpages in this way, the method can be used for supporting the correction of sentences.

4 Experiments

4.1 Experimental conditions

For this study, from the Japanese Wikipedia (November 2014), we used 2,665 pages whose titles ended in “castle.”

We input the pages and extracted the information using the method described in Section 3.1 and created a table for detecting incomplete descriptions. We used the results of Akano et al. [1].

In the experiments, Clusters 401, 407, and 765 were selected manually from the clustering of the important items. The rows of the table listed the names of the castle and the columns listed the important items. Cluster 401 contained information on battles, 407 contained information on the construction of the castles, and 765 contained information related to traffic.

Using the method described in Section 3.2, we created a table to facilitate the correction of incomplete descriptions in the Wikipedia entries. To evaluate the method, we conducted the following two experiments:

- Information extraction using search engines for all parts of the table
- Information extraction using search engines for only missing parts of the table

The experiment for using search engines to extract information from the webpages for all parts of the table was conducted to observe the performance of the information extraction.

Information extraction and table creation were performed on the 2,665 pages in Wikipedia about castles (November 2014) using the methods described in Section 3.1. In addition, we input 30 names of castles randomly selected from the 2,665 pages into the search engine, acquired the webpages, and conducted the experiment with these 30 names to evaluate the method. The search engine used was Microsoft’s Bing Search API.

4.2 Evaluation method

Using top-one accuracy, top-five accuracy, and MRR, we conducted an evaluation. In top n accuracy, the output is judged to be correct when the top n candidate answers contain a correct answer. In MRR, a score of $1/r$ is given when the r -th candidate answer is correct.

4.3 Experimental results of information extraction using search engines for all parts of the table

Using the method proposed in Section 3.2, we entered 30 names of castles randomly selected from the 2,665 pages about castles in Wikipedia (November 2014) into the search engine. Using the webpages acquired from the search engine, a table was created by the method described in Section 3.2.

An example of the results is shown in Table 8. The words in bold font were those judged to be correct. The five most frequent words in the 50 webpages acquired by the search engine were output for each important item. Table 9 shows the results of the top-one accuracy, the top-five accuracy, and MRR evaluations of the constructed table.

The top-five accuracy was 0.66. We found that we could extract information using webpages to a certain extent.

Table 8. Example of results of information extraction

	Cluster 401 (battle)	Cluster 407 (construction of castle)	Cluster 765 (traffic)
Uwajima Castle	opening a castle attack falling castle immediately sortie	main castle keep Nagaya gate relocation palace remains of gate	traffic Kaido highway Setouchi Hokuriku
Chikugo Jugo Castle	great defeat resistance outing falling castle fierce	office main castle keep second government office palace	Kaido traffic contact tie highway
Okazaki Castle	departure break return great defeat battle field	main castle keep second castle keep main gate palace government office	traffic Kaido Tokaido Hokuriku convenience
Sakurao Castle	falling castle opening a castle outing occupation Return	main castle keep second castle keep rising sun office Gate	Setouchi pilgrimage Sanyo traffic convenience
Linderhof Castle	return arson immediately a few	transfer rising sun	traffic highway convenience Kaido romantic
Odawara Castle	opening a castle falling castle retreat return withdrawal	main castle keep second castle keep main gate palace lotus pond	Kaido Tokaido traffic trunk line Hokuriku
Kawata Castle	falling castle approach run surrender ambush	main castle keep second castle keep palace office address	highway traffic strategic stop convenience crossing
Nagamori Castle	falling castle going out offend opening a castle withdrawal	main castle keep second castle keep palace great gate office	traffic Kaido Nakasendo Hokuriku convenience
Shakujii Castle	falling castle defeat arson return great defeats	main castle keep Nagaya gate office second house keep great gate	contact convenient traffic Kaido highway

Table 9. Accuracies in all the parts

Evaluation method	Cluster 401	Cluster 407	Cluster 765	Total
Top-1 Accuracy	0.56 (17/30)	0.50 (15/30)	0.53 (16/30)	0.53 (48/90)
Top-5 Accuracy	0.70 (21/30)	0.60 (18/30)	0.70 (21/30)	0.66 (60/90)
MRR	0.62	0.52	0.57	0.55

Table 10. F-measure of detecting inadequate descriptions (missing parts)

Precision	0.79 (53/67)
Recall	0.96 (53/55)
f-measure	0.84

Table 11. Accuracies in missing parts of the table with no correct answers in Wikipedia

Evaluation method	Cluster 401	Cluster 407	Cluster 765	Total
Top-1 Accuracy	0.35 (5/14)	0.29 (5/17)	0.36 (8/22)	0.33 (18/53)
Top-5 Accuracy	0.50 (7/14)	0.41 (7/17)	0.59 (13/22)	0.50 (27/53)
MRR	0.42	0.32	0.45	0.40

4.4 Experimental results of information extraction using search engines for only missing parts of the table

In the experiments extracting important information from documents (Wikipedia pages) described in Section 3.1, we detected the missing parts by consulting Akano et. al.’s paper [1]. The method detected 67 missing parts. There were 55 correct missing parts in the data of the 30 castles used in the study. Among the correct missing parts, 53 had been correctly detected. The f-measure, precision, and recall of extracting the correct missing parts are shown in Table 10. The f-measure was 0.84, indicating a very good result.

For the information extraction using a search engine, as described in Section 3.2, we used our method in the experiments.

We conducted experiments to extract information using search engines for only the missing parts of the table with no correct answers in Wikipedia. The results are shown in Table 11. The experiment was performed on the missing parts that had been correctly detected by the method described in Section 3.1. “No correct answers in Wikipedia” means the case where a description in Wikipedia is inadequate and we should complement the inadequate description.

The top-five accuracy of the experiments on the missing parts of the table with no correct answers in Wikipedia was 0.50.

Table 12 is the result of the evaluation we conducted under the following Condition A.

When the correct answer was not in the answer candidates within the first to fifth ranks, we confirmed if the correct answer was in the answer candidates within the sixth to twentieth ranks. If the correct answer was not in the answer candidates, we manually confirmed whether the correct answer was on the Web. If the correct answer was not on the Web, the missing part was the part for

Table 12. Accuracies in the missing parts of the table with no correct answers in Wikipedia (Condition A: eliminating cases when no correct answers exist in the webpages)

Evaluation method	Cluster 401	Cluster 407	Cluster 765	Total
Top-1 Accuracy	0.71 (5/ 7)	0.62 (5/ 8)	0.61 (8/13)	0.64 (18/28)
Top-5 Accuracy	1.00 (7/ 7)	0.87 (7/ 8)	1.00 (13/13)	0.96 (27/28)
MRR	0.85	0.69	0.77	0.77

Table 13. F-measure for detecting inadequate descriptions and detecting correct expressions in webpages

Precision	0.68 (27/40)
Recall	0.90 (27/30)
f-measure	0.77

which we did not have to find correct answers. We conducted experiments by eliminating such missing parts.

Under Condition A (eliminating cases when no correct answer exists in the webpages), the MRR and top-five accuracy for the experiments in all the parts of the table with no correct answers in Wikipedia were 0.77 and 0.96, respectively. The results were very accurate, indicating that with high accuracy, our method could gather information useful for complementing the missing parts when correct answers existed in webpages.

We calculated the f-measures for detecting inadequate descriptions and correct expressions in webpages. The results are shown in Table 13. The recall rate is the rate of dividing the number of detected correct missing parts and outputting correct answers with no correct answers in Wikipedia by the number of correct answers in the case of using the top-five accuracy. The precision rate is obtained by dividing the number of detected correct missing parts and outputted correct answers with no correct answers in Wikipedia by the number of detected missing parts and output answers with no correct answers in Wikipedia in the case of using the top-five accuracy.

The task of detecting inadequate descriptions and detecting correct expressions in webpages is very difficult. Therefore, our method was very effective since it obtained an f-measure of 0.77.

We show an example of the successful results of our method. Using the method described in Section 3.1, Table 14 was created by extracting important information from an entry in Wikipedia. Blanks in the tables are missing parts that correspond to the absence of descriptions of correct answers in Wikipedia. Therefore, the blanks are correct. For Table 14, we used the method proposed in Section 3.2 to complement the missing parts in this table. Table 15 shows what had actually been complemented. The words in parentheses are complementary information, which includes the information taken from the first to fifth candidate answers.

We give an example of the failures of our method. Our method could not obtain the correct answers in the top-five candidates for Cluster 401 (battles)

Table 14. Examples of a table having missing parts

	Cluster 401 (battle)	Cluster 407 (construction of castle)	Cluster 765 (traffic)
Anozu Castle			
Moji Castle	defeat	castle keep	

Table 15. Successful examples of sentence correction support

	Cluster 401 (battle)	Cluster 407 (construction of castle)	Cluster 765 (traffic)
Anozu Castle	(opening a castle)	(castle keep)	(highway)
Moji Castle	defeat	castle keep	(traffic)

for Tabata Castle. The correct answer, “great defeat,” appears as the 16th candidate in our method. Therefore, the case was judged to be incorrect for top-five accuracy. Table 16 shows 20 candidate answers for Cluster 401 (battles) in Tabata Castle. Although “departure,” “reinforcement,” “falling castle,” etc. in the first to third ranks may seem to be correct answers, they appear in a context unrelated to Tabata Castle and so they are incorrect.

5 Conclusion

The purpose of this study is to support the correction of inadequate descriptions. Important information is extracted from entries in Wikipedia and inadequate descriptions are detected. Information that is useful for correcting the inadequate descriptions is gathered by a search engine and presented to a user to support the correction of the sentences.

In the experiments on the extraction of information using search engines for all parts of a table, our method obtained a top-five accuracy of 0.66. We found that we could extract information using Web pages to a certain extent.

In the experiments on the missing parts of the table with no correct answers in Wikipedia, the top-five accuracy was 0.50. Under Condition A (eliminating cases when no correct answers exist in webpages), experiments on the missing parts of the table with no correct answers in Wikipedia, obtained an MRR and a top-five accuracy of 0.77 and 0.96, respectively, indicating that the results were very accurate. and that our method could gather, with high accuracy, information to complement the missing parts when correct answers existed in webpages.

The f-measure of the very difficult task of detecting inadequate descriptions and detecting correct expressions in webpages was 0.68, indicating that our method was very effective, i.e., it obtained 0.77 in F-measure.

Acknowledgment

This work was supported by a research grant.

Table 16. Words of Cluster 401 with the 1st to 20th ranks for Tabata Castle

Castle	Rank	Cluster 401 (battles)	Number of pages
Tabata Castle	1	departure	4
	2	reinforcement	3
	3	falling castle	2
	4	rushing	2
	5	immediately	2
	6	retreat	2
	7	joining	2
	8	surrender	1
	9	defense	1
	10	arson	1
	11	weapons	1
	12	dispatch	1
	13	rally	1
	14	stationed	1
	15	departure	1
	16	great defense	1
	17	annihilation	1
	18	burned down	1
	19	sortie	1
	20	heading	1

References

1. Akano, H., Murata, M., Ma, Q.: Detection of inadequate descriptions in Wikipedia using information extraction based on word clustering. In: Proceedings of IFSA-SCIS 2017. pp. 1–6 (2017)
2. Fukuda, S., Nanba, H., Takezawa, T.: Extraction and visualization of technical trend information from research papers and patents. D-Lib Magazine (2012)
3. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems 26, 3111–3119 (2013)
4. Murata, M., Utiyama, M., Isahara, H.: Use of multiple documents as evidence with decreased adding in a Japanese question-answering system. Journal of Natural Language Processing 12(2), 209–248 (2005)
5. Nadamoto, A., Aramaki, E., Abekawa, T., Murakami, Y.: Extracting content-holes by comparing community-type content with wikipedia. The International Journal of Web Information Systems 6(3), 248–260 (2010)
6. Okada, T., Murata, M., Ma, Q.: Automatic detection and manual analysis of inadequate descriptions in a thesis. Proceedings of SCIS-ISIS 2016 pp. 916–921 (2016)
7. Ptaszynski, M., Masui, F.: Spass: A scientific paper writing support system. In: The Third International Conference on Informatics Engineering and Information Science (ICIEIS2014). pp. 1–10 (2014)
8. Tsudo, S., Murata, M., Tokuhisa, M., Ma, Q.: Machine learning for analysis and detection of redundant sentences toward development of writing support systems. In: Proceedings of the 13th International Symposium on Advanced Intelligent Systems. pp. 2225–2228 (2012)