A Comparison of Associated Words and Extracted Knowledge from Documents

Haruhiko Takeuchi   Muneo Kitajima   Motoyuki Akamatsu

National Institute of Bioscience and Human-Technology
1-1, Higashi, Tsukuba, Ibaraki
305-8566, Japan
mailto: takeuchi@nibh.go.jp

Abstract

Latent Semantic Analysis (LSA) is one of the corpus-based statistical methods for extracting the knowledge from documents. We extracted the meaning of words from three sources, using LSA method, and compared the results with human association data. It is shown that there exists some relation between human association and extracted knowledge from documents, when the sources of the documents are selected appropriately.

Keywords

Natural language, association, corpus

1 Introduction

Human association has been mainly studied in the contexts of memory, recall and clustering. Association is one of the innate human ability and plays an important role in human verbal behavior, but it has been paid little attention in the field of natural language understanding. One of the reason is that the association data can only be measured by psychological experiments, and it is difficult to collect large number of associated words.

Recently, various documents, such as newspapers and novels, can easily be utilized in electronic forms. Our major concern is to compare the word relations of human association data and extracted knowledge from the documents. This paper presents the first trial of analyzing human association data from the viewpoint of Latent Semantic Analysis (LSA) approach for inducing the knowledge.

In order to compare the word relations, we collected the linguistic data by two different ways. They were psychological method and corpus-based method. Firstly, We conducted the association test for college students, and gathered human association data. Secondly, We constructed the word relation database, using LSA method for three different types of documents. Then, we picked up some representative words and examined them in detail. Through this process, we aim to clarify the word relations between those two data which were made by different ways.

2 Methods

2.1 Measurement of human association

The association test was conducted to collect the human association data. We used 276 words as stimuli. These words were adjectives and adverbs which express human feeling, emotion and sensation. Subjects were thirteen university students aged from 20 to 25. The association test, we used, was one of the restricted methods (Brown, 1976). We set three constraints. They were the association of synonyms, association of antonyms, and association of objects.

Each subject was given a test booklet. For each trial, the subjects were instructed to look at the target word and then write down three associated words for each constrained association. The subjects were given as much time as they needed to complete the task.

2.2 Extraction of knowledge from documents

We used three different types of documents as sources for extracting word relations. These texts were gathered from songs, novels and newspapers. The song's data were Japanese old popular songs cited from "Omoide no hitto kyoku 500 (Gotou Shoin Henshubu (eds.), 1998)". We used the songs which was produced in 1965 to 1976. The novel's data were cited from one of the children's story book "Gurimu Douwa-syu

(Grimm, B., 1998)". The newspaper's data were cited from "CD-Mainichi-Shinbun '97". On the newspaper's data, currently, we only focussed on the culture and home related pages and used the article which was written in January to April.

In order to induce knowledge from these documents, we adopt the Latent Semantic Analysis (LSA) method (Landauer & Dumais, 1997). The LSA is one of the corpus-based statistical methods for extracting the meaning of words. It is shown that the LSA method can successfully be used to simulate human learning process and some of psycho-linguistic phenomena.

The outline of the LSA method is as follows. Firstly, we divide the source texts into small meaningful passages, which we call contexts. Secondly, we make the matrix, where rows stand for word type and columns stand for text contexts. The cell entries of this matrix are the transformed frequencies of the word's appearance in the context. Thirdly, we apply the singular value decomposition (SVD) method to the matrix. The SVD method works as one of the dimension reduction methods. And it is useful in generalizing the semantic space of word meaning. Finally, we get the matrix where rows stands for word type and columns stand for factors. The result is explained in the multi-dimensional semantic space in which words are represented as vectors. The similarity between the two words can be measured by the cosine of the angle between those vectors.

3 Results and discussion

There are several indices for the multiple-response restricted association tests (Marshall & Coffer, 1963). The index we used was called association rates. The responses were gathered for each stimulus word under each constraint. We counted the number of the subjects whose associated word was the same. This summation was done for each stimulus word. Then the sum was divided by the number of all subjects. This index shows the proportion of association frequency for each associated word. As a result, we got the associated word lists with association rates under three constraints.

We will show one of the results, taking the word "red (Akai in Japanese) " as an example. Highly associated words as synonyms were "bright", "hot" , "burning", and "crimson". Highly associated words as antonyms were "blue", "white", "black", "cold", and "dark". Highly associated words as objects are "blood", "fire", "apple" and so on. The indices for these associated words were in between 0.2 to 0.7. Another example for the target word "cute" (Kawaii) is shown in Table 1.

Table 1. Example of human association

(a) Association of synonyms

0.62	small
0.31	childish
0.31	lovely
0.31	pretty
0.23	beatiful
0.15	nice

(b) Association of antonyms

0.62	no charming
0.23	plain
0.23	hateful
0.15	ugly
0.15	big
0.15	bad loooking

(c) Association of objects

0.46	child	0.15	flower
0.31	stuffed	0.15	kitty
0.23	baby	0.15	dog
0.15	girl	0.15	face
0.15	bird	0.15	car

Note. The target word is "cute". The number in front of each word indicats the association rates.

According to the LSA methods, we divided the source texts into small parts. In the case of song's data, we treated each song as separate context. In the case of novel's data, we treated each paragraph as independent context. In the case of newspaper's data, each article was treated as separate context.

Following to the LSA method, we made the matrix of word type by context. In order to make the word type list, we used the Japanese Morphological Analyzer (JTAG) developed by NTT Cyber Space Laboratories (Fuchi,T., 1998). Using the JTAG parser, each sentence was separated into a part of speech. Some useless words, such as alphabet, numbers, conjunction, and prefix, were deleted in this stage. In the case of newspaper's data, we also deleted the proper noun about names and places.

The size of the matrix of word type by context is 966 rows and 349 columns for the song's data. In Table 2, we show the number of words and contexts for each document. Then the matrix was processed, using SVD method.

Table 2. The size of matrix for each documents

Documents	Number of words	Number of contexts
Song's book	966	349
Children's story book	1020	261
Newspaper	8598	1010

Table 3. Example of LSA results

(a) Song's book

0.93	light	0.83	red
0.92	ashamed	0.80	guitar
0.92	candl	0.78	old familiar face
0.89	rose	0.73	nape
0.88	always	0.73	sad
0.88	alone	0.69	thin
0.87	snack bar	0.57	wish

(b) Children's story book

0.87	souvenir	0.77	like
0.86	foot	0.76	youngest child
0.86	thin	0.74	disguise
0.83	bread	0.74	soft
0.83	nice	0.74	gender
0.80	lump	0.73	child
0.77	wall clock	0.72	pure-white

(c) Newspaper

0.59	room	0.52	side
0.57	advise	0.51	angel
0.57	collar	0.50	public domain
0.54	heavy	0.50	intention
0.53	sick child	0.48	framework
0.53	easy	0.47	reform
0.52	noisy	0.47	backward

Note. The target word is "cute". Fourteen near neighbor words are shown for each documents. The number in front of each word indicates the similarity.

Finally we got the matrix of word type by factor for each source of documents. The number of the factors is an important parameter about generalization. Currently, we set the parameter around one fifth to one tenth of original space's dimension. Currently, the parameter was 35 for song's data and novel's data, and 200 for newspaper's data.

Using the matrix of word type by factor, we can calculate the similarity between the two words. By sorting the words according to the similarity, we can get the list of similar words for each target word. We will show one of the result from the children's story book, taking the word "fearful (Osoroshii)" as an example. The similar words are "danger", "pierce", "death", "back", "lively", "rescue", "distant", "thing", "secret", "woman" and so on. The similarity of these words are in between 0.5 to 0.8. We found that both of synonyms and antonyms are highly related in this semantic space. Another example about the target word "cute" (Kawaii) is shown in Table 3.

Although the number of the words which appeared in the source texts were several hundreds to several thousands, all of the stimulus words, which was used in the association test, were not included in the matrix of word type by factor. This is because the large part of the words in the matrix was occupied mainly by nouns. The song's data included comparatively rich adjectives. The children's story book included familiar adjectives to some degree. On the other hand, the newspaper's data was big, but it did not include so many adjectives.

We made the comparison to the specific words which appeared in the both of the data. Example of these words were "cold", "gender", "bright", "brilliant", "fantastic", "warm", "sad", "complicated", "happy" and so on. We picked up several words randomly as representatives for the comparison. We listed twenty most similar words, using the result of LSA method. Then we checked whether these words were included in the human associated word lists. Let us refer to the example about the word "sad" (Kanashii), using the children's story data. We found that the word "happy" was most highly associated as antonym. The word "joyful" was also highly associated. When the target word was "brilliant", the word "dress" and "beautiful" ware included in both of the results. But when the target word is "fine" or "tasty", the human associated words were not included in the results.

Let us consider the effect of source documents, referring the example about the word "cute". The result for the word "cute" is shown in Table 1 and Table 3. In the case of song's data and newspaper's data, no identical word was included in the human association lists. In the case of children's story data, "nice" and “child” appears in both of the lists.

The result was strongly affected by using different types of documents. The knowledge extracted from song’s data sometimes did not fit the association data well. One of the explanation of this result would be that there exists plenty of poetic expressions in the song’s data. The poetic expression often breaks the common sense, for the concepts expressed in these poetic sentences are often new and unusual. On the other hand, the association data is considered to reflect human common sense knowledge.

When the newspaper’s articles are used as a source for LSA method, we found that the number of the adjectives which express human feelings, emotions, and senses were not so many. This is thought to be the characteristics of newspapers. And even if some feeling words were used in the article, they were sometimes used in the specific contexts. So the word meaning extracted from newspaper does not always fit human association data well. For instance, some of the neighbor words for “sad” were “orphan”, “commodities”, “refugee”, “Africa”, “Cambodia” and so on. This result strongly reflect the article written about refugee.

The data cited from the children’s story book showed comparatively better characteristics. The novel’s data included plenty of feeling words, and the knowledge structure extracted from the children’s story book reflected ordinary common sense knowledge.

4 Conclusions

We extracted knowledge from three types of documents based on the LSA method. The knowledge was used to measure the word similarities. We also gathered associative words by conducting psychological experiments. A comparison was done between these two data: human association data and word similarity data extracted from documents. It was shown that several words appeared in both of the data, when the sources of the documents are selected appropriately.

When we apply the LSA method to extract the knowledge from documents, the result is affected by the source documents, the size of context, and the parameter of factors used in SVD method. In this paper, we showed the effect of using different types of documents in extracting the knowledge. The novel’s data showed the best fit, and the newspaper’s data showed the worst fit from the associative viewpoint. Further research on this theme includes increasing the amount of source documents and refining the LSA method especially for Japanese.

Acknowledgments

In this research, we used the Japanese Morphological Analyzer (JTAG) developed by Media Processing Project, NTT Cyber Space Laboratories. We thank NTT Cyber Space Laboratories, particularly Takeshi Fuchi and Hisashi Ohara, for supporting to use JTAG program.

References

Brown, A. S. (1976). Catalog of scaled verbal material. Memory & Cognition, 4(lB), 1S-45S.

CD-Mainichi-Shinbun ’97 (1997). Mainichi-Shinbun-sha.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). Using Latent Semantic Analysis to improve access to textual information, *Proceedings of CHI88*.

Fuchi, T. (1998). Morphological Analyzer System JTAG 2.0 User’s Manual. NTT Cyber Space Laboratories.

Gotou Shoin Henshubu (eds.) (1998) Omoide no hitto-kyoku 500. Gotou-Shoin.

Grimm, B. (1998). Gurimu Douwa-shu (T. Yoshioka & M. Yoshihara, Trans.). Hakusui-sha. (Original work published 1812).

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.

Landauer, T. K., Laham, D., & Foltz, P. W. (1998). Learning human-like knowledge by Singular Value Decomposition: A progress report. IIn M. I. Jordan, M. J. Kearns & S. A. Solla (Eds.), Advances in Neural Information Processing Systems 10, MIT Press.

Marshall, R. R., & Coffer, C. N. (1963). Associative indices as measures of word relatedness: A summary and comparison of ten methods. *Journal of Verbal Learning and Verbal Behavior*, 1, 408-421.