Degree of Association between Concepts using Wikipedia Concept-base and Word2Vec

Hirokazu Watabe, Misako Imono and Seiji Tsuchiya
1,2 Dept. of Intelligent Information Engineering and Science, Doshisha University, Kyotanabe, Kyoto, Japan
2 Dept. of Information Systems, Daido University, Nagoya, Japan
hwatabe@mail.doshisha.ac.jp, m-imono@daido-it.ac.jp, stsuchiy@mail.doshisha.ac.jp

Abstract - In this paper, we first build Concept-Base from Japanese Wikipedia as the information source, and Vector space model suitable for measuring similarity between words using Word2Vec. Then, we propose the Calculating Method of Degree of Association considering similarity using the Wikipedia Concept-Base and Vector space model.

Keywords: Wikipedia, Concept-Base, Word2Vec, Degree of Association

1. Wikipedia Concept-Base
A Wikipedia Concept-Base (WCB) is a knowledge base comprised of terms (concepts) mechanically constructed from sources of Japanese Wikipedia (1) and terms (attributes) that express their semantic features. Concepts have been given attributes along with a weighting, which expresses their importance. 1,112,636 concepts (2,095,385 notations) have been compiled in the WCB, with an average of 79 attributes for one concept. A certain concept A has a set of pair of attribute ai and weighting ui as appear below.

\[ A = \{ (a_1, u_1), (a_2, u_2), \ldots, (a_n, u_n) \} \]  

Any primary attribute ai is composed of the terms contained in the set of concept notations in its WCB. Therefore, to ensure that a primary attribute matches a certain concept notation, that primary attribute can be further extracted. This is called a secondary attribute. In WCB, a concept is defined by a chained set of attributes to the n-th order.

2. Degree of Association
The degree of association (DoA) (3,4) is a value ranging from 0.0 to 1.0 that quantifies the strength of the relationship between two concepts (words) registered to the WCB as shown in the equation 3. To find the degree of association, the relevance between two concepts is calculated as a numeric value based on the value found by calculating the degree of matching (DoM) of the secondary attributes of the concepts as shown in the equation 2.

\[ \text{DoM}(A, B) = \sum_{a_i} \min(u_i, v_i) \]  

\[ \text{DoA}(A, B) = \sum_{a_i} \{ \text{DoM}(a_i, b_i) \times (u_i + v_i) \} \times \left( \frac{\min(u_i, v_i)}{\max(u_i, v_i)} \right) / 2 \]  

3. Vector Space Model by Word2Vec
Word2Vec (2) is a group of related models that are used to produce word embeddings. These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words. Word2Vec takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. In this paper, we use Japanese Wikipedia as input of Word2Vec, and 893,656-word vectors are constructed. As parameters, size(dimensions), window, and cbow is 600, 9, and CBOW, respectively.

Cosine similarity is a measure of similarity between two words (concepts). We call this word similarity (Sim). The following is the word similarity between word A and B.

\[ A = (u_1, u_2, \ldots, u_{600}) \]  

\[ B = (v_1, v_2, \ldots, v_{600}) \]  

\[ \text{Sim}(A, B) = \cos(A, B) = A \cdot B / (\|A\| \|B\|) \]  

(4)

4. Extended Degree of Association
By combining degree of association (DoA) using WCB and word similarity (Sim) using Word2Vec form Wikipedia, we construct the extended degree of association (EDoA) between two concepts as the following.

\[ \text{EDoA}(A, B) = \frac{\text{Sim}(A, B) + w \times \text{DoA}(A, B)}{1 + w} \]  

(5)

5. Evaluations
5.1 Hit rate
Figure 1 shows the number of notations (words) of each model. CB is the original Concept-Base (3,4) which was constructed by Japanese dictionaries and newspapers. VSM is the vector space model which is constructed by Word2Vec using Wikipedia. WCB is the Concept-Base which is constructed from Wikipedia.

Using 2345-words from 30 articles of Japanese Mainichi-Newspaper between 16/12/2017 to 17/12/2017, we evaluate hit rate of each model (figure 2). Table 1 shows a part of words using for this hit rate evaluation. From figure 2, the original CB doesn’t have enough vocabularies.
5.2 X-ABC Evaluation

Second evaluations are carried out using an X-ABC evaluation set. This evaluation set is composed of four concepts, $X$, $A$, $B$, and $C$. In Table 2, specific X-ABC evaluation set examples can be seen. Concept $A$ entries have high relations with Concept $X$ entries, and Concept $B$ entries have some relations with Concept $X$ entries, but Concept $C$ entries do not.

This evaluation method calculates $DoA(X, A)$, $DoA(X, B)$ and $DoA(X, C)$. At this time, if the following equations are satisfied, it is assumed that the correct answers are given.

$$DoA(X, A) - DoA(X, B) > AveDoA(X, C)$$  \hspace{1cm} (6)

$$DoA(X, B) - DoA(X, C) > AveDoA(X, C)$$  \hspace{1cm} (7)

$$AveDoA(X, C) = \frac{\sum_{i=1}^{N} DoA(X, C_i)}{N}$$

Where $N$ is the number of evaluation sets. In this paper, the total number of evaluation sets made by plural people is 300.

Figure 3 shows the results of X-ABC evaluations, where $w$ in equation 5 is 1.0. The best result is the case of extended degree of association derived both from Wikipedia Concept-Base and cosine similarity using Word2Vec from Wikipedia.

6. Conclusion

In this paper, we have constructed Concept-Base from Japanese Wikipedia as the information source, and vector space model suitable for measuring similarity between words using Word2Vec. The input sources to Word2Vec are whole text from Japanese Wikipedia. Then, we propose the calculating method of degree of association using both the Wikipedia Concept-Base and vector space model.

Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Number 16K00311.

References