Exploring Words with Semantic Correlations from Chinese Wikipedia

Yun Li, Kaiyan Huang, Seiji Tsuchiya, Fuji Ren and Yixin Zhong

Abstract In this paper, we work on semantic correlation between Chinese words based on Wikipedia documents. A corpus with about 50,000 structured documents is generated from Wikipedia pages. Then considering of hyper-links, text overlaps and word frequency, about 300,000 word pairs with semantic correlations are explored from these documents. We roughly measure the degree of semantic correlations and find groups with tight semantic correlations by self clustering.

1 Introduction

Semantic information and semantic relations are more and more important in natural language processing (NLP), being applied in applications such as text retrieval, information extraction etc. For semantic computing, semantic knowledge-base is created. For the complexity of semantic relations, majority of them like WordNet are constructed artificially, which is a time-consuming work. Automatic acquisition of semantic knowledge is important for research.

Wikipedia is an open encyclopedia with hyper-linked documents written cooperate by Internet users. In NLP applications, it could not only act as a huge corpus, but also a knowledge base or a semantic resource comparable to artificial constructed ones. It has been evaluated in the researches of Zesch&Gurevych (2007) etc. Someone explored Wikipedia for semantic relatedness computing (Strube&Ponzetto, 2007), name entity disambiguation (Bunescu&Pasca, 2006), automatic question answering (Ahn, 2004), etc.

Yun Li, Yixin Zhong
Beijing University of Posts and Telecommunications, 310#, BUPT, 10 Xitucheng Road, Haidian, Beijing, 100876, China, e-mail: liyun@is.tokushima-u.ac.jp; yxzhong@ieee.org

Yun Li, Kaiyan Huang, Seiji Tsuchiya, Fuji Ren
The University of Tokushima, Ren Lab,2-1 Minamijosanjima-cho,Tokushima,770-8506 e-mail: (liyun,huangky,tsuchiya,ren)@is.tokushima-u.ac.jp

In this paper, we work on word semantic correlation with Chinese Wikipedia documents. A structured Wikipedia document corpus is firstly generated from Wikipedia pages. Considering of hyper-links between documents, as well as text overlaps and the location information, pairs with semantic correlations are selected. Semantic relatedness is calculated from the paragraph locations and word frequency information in the Wikipedia documents. Finally we roughly measure the degree of semantic correlations and find groups with semantic correlations by self clustering.

2 Generate Wikipedia Document Corpus

Totally 322,121 Wikipedia words with hyper-links to pages are selected from Wikipedia. Given that the majority of professional terms have low frequency in normal text, words are filtered according to frequency in two Internet word-lists, one of which created by the search engine Sogou, the other with 800,000 words and frequencies of Google&Baidu collected by Internet users. Entries in corpus of People's Daily 2001 are also selected as candidates.

For the selected 66,725 of Wikipedia words with URLs, we download the pages from Chinese Wikipedia.In other works, the source extracted from image package with Wiki format documents is mainly applied. But for Chinese, the source is a mixture with simplified Chinese and traditional Chinese. In order to get localized documents, it is necessary to use the Html pages instead of the source files, as a localization translation services are called automatically in web pages. As redirect pages exist, we actually get 54,745 pages.

In Wikipedia documents, the basic text part is usually the most fundamental and important explanation for a word, located before the outline with less than 3 paragraphs of a document. Other paragraphs are detailed information with less importance. Some table or lists with manually grouped words are shared among some documents. The three kinds of text parts are not equally related to the topic word. In generating structured document, they are separately saved in different fields of the text corpus. As hyper-links in document are important for our research, meeting a hyper-link to another document, we get the word, URL keyword and count in each document parts. Other information is also collected like the length of raw text and html-text, total count of links and duplicated links, keyword and hyper-links to the category graph etc.

For synonyms, a page redirection is used to access the same document. In the following tests, synonyms should be seen as one word in computing of semantic relatedness. Synonyms could be found from the title word and the keyword following a mark of "Redirected From" in the redirected Wikipedia documents. More are collected from paragraphs such as "China Central Television commonly abbreviated as CCTV", "An astronaut or cosmonaut". Taking account of synonyms, the amount of words in the corpus is raised to 89,994, following with 54,745 XML formatted structured documents generated from Wikipedia pages. As synonyms and redirection exists, one page is statistically shared by 1.6 Wikipedia words. There are

totally 1,823,883 hyper-links found from all the pages, averagely 33.3 in one page. 411,402 pairs of pages are hyper-linked to each other, with more relatedness and being considered more useful in our works.

3 Exploring Words with Semantic Correlations

Many researches on NLP refer to the semantic relations.Comparing to semantic similarity which shows only the "kind-of" semantic relation, semantic correlation is broad and comprehensive. Different researchers applied their own interpretations. Algorithms on WordNet, Hownet could be found from many related works.

In Wikipedia documents, semantic correlations between the title word and paragraphs are higher than other documents from web. In one view, the text was usually seen as the representation for the title keyword. In our corpus, 1,823,883 hyper-links between lines are explored, which are linked to the corresponding Wikipedia documents, showing relations on semantic meaning of the text lines. We pay more attention on the 411,402 page pairs with hyper-links to each other. By studying some pairs, we found most of them being semantically correlated, at least sharing some topics or events. As relatedness is a kind of importance, if something is noticed easily and usually, the importance should be higher, and their may some semantic relations exists. As correlations are for both sides, the relatedness between each other should be exists. From this view we design our way of finding semantically related words and calculating relatedness from the hyper-linked documents.

Experiments are done using the information of document hyper-links. Firstly Experiment 1 is to find the most word pairs with semantic relations. This time only the Wikipedia basic definition and description part is employed. As during the structure work, we separately saved the main part of text and hyper-link information, we directly used the data. For the document with title word A, we get hyper-linked groups (B, C, F, G), then search for hyper-linked A from each linked documents in this group. If C and G match the rule, two result of (A, C) and (A, G) are selected as candidate pairs. The experiment is done using a C++ program on a data set of word and links with integers IDs. From Experiment 1, 15,512 word pairs were found. It covers 14,290 words that are only 26% of the selected Wikipedia words. During artificial review, most pairs could be accepted by human understanding of semantic correlations. Some of which were listed in Table 1 in the form of (A, B) with English translations.

In Experiment 2, we extend the scope of search to the whole Wikipedia document. As the basic definition or introduction refer only a little part of a topic word, the most related materials exist in other paragraphs. The aim is to find a semantic correlated word set of a bigger coverage of correlates. For word pair (A,B), if each could be find in any position from the other's document as a hyper-link, it is selected as a candidate pair of Set A, and if one noticed in the main part of the other, which is more reliable, we select the pair to Set B. So the rule of Set B is more strict than Set A but being looser than that of Experiment I. We get the result in Table 2. Generally

Table 1 Semantic Correlated Word Pairs.

A(CN)	A(EN)	B(CN)	B(EN)	A(CN)	A(EN)	B(CN)	B(EN)
信号	Signals	交通	Traffic	按钮	Button	人机交互	HCI
椅子	Chair	轮椅	Wheelchair	彩蛋	EasterEgg	复活节	Easter
棒球	Baseball	球棒	Bat	筷子	Chopsticks	中国烹饪	ChineseCook
赌博	Gambling	赌徒	Gamblers	行动党	ActionParty	马来西亚	Malaysia
专科	College	教育	Educate	演化	Evolutionary	博弈论	GameTheory
软件	Software	许可证	License	阪神	Hanshin	甲子园	Koshien

Table 2 Result with Manual Evaluation.

ID	Word Pair Count	Coverage	Reviews
1	15,512	26%	Correlated
2(A)	79,150	65%	Most correlated
2(B)	411,402	73%	Some unrelated

speaking, the result is reliable with semantic correlations. For some pairs in Set A, the importance is different from each other. Such as “春节(Spring Festival)”, “拜年(Say Happy New Year) ”, “鞭炮(Firecrackers) ”, “年画(New Year Pictures) ”.,the “Spring Festival” is more important for “New Year Pictures”, but “Spring Festival” could have relatedness with more other than “New Year Pictures”.

For Set B the coverage is changed to 73% with 39,925 words, but the accuracy is not very good. A refine work should be down relies on more information not limited to the word frequency, document frequency etc. Among most un-related pairs, a common result is that at least one word in a pair has a high document frequency. Such as “中国(China)”, “公司(Company)”, “地区(Regions)”, “英语(English)” are selected related to many keywords. The mistakes appear because these words easily appear in everyday text files with a high document frequency.

In Experiment 3 document frequency and word frequency are used as filter on Set B. By accessing the documents, we get the count of document containing a word as hyper-linked text for all the selected Wikipedia keywords. Document frequency are calculated with the count divided by total count of documents in corpus. Only the pairs not in Set A are filtered. Several tests are done to find a proper threshold leading to more correlated word pairs and less unnecessary ones. Our result set is finally cut to 360,304 semantic related pairs. Here we employ a way of roughly measuring semantic relatedness. We give each part of the Wikipedia document a score: 4 for the main part, 2 for other part of the document text, and 1 for shared tables and text. During the creating of our Wikipedia corpus, hyper linked words and paragraph information have been extracted, which are used directly. For both words in a pair, we add the score in the document of the other’s.

4 Self Clustering for Semantic Related Words

In Experiment 3, taking one word as a center word, we could find an average of 7.3 semantic related words. Small group of words sharing a common topic are easy to group together with more correlations among each other. Taking “北京(Beijing)” for example, 242 words are found covering several aspects, from which smaller groups with tight semantic correlations can be found. Such as “奥运会(Olympics)”, “福娃(Fuwa)”, “鸟巢(Bird Nest)” etc form a cluster of the 2008 Olympics, and “天安门(Tiananmen)”, “西单(Xidan)”, “王府井(Wangfujing)” are grouped as famous places of Beijing.

In our experiment, we make a larger group by adding new nodes into a smaller group. If pairs of (AB, AC, BC) could be found in the semantic related word set, a new group of (ABC) are created. Then a candidate group (ABCD) could be extended to if exist the pairs of (AD, BD, CD). Finally a small group should be removed if being sub set of a larger group. In this way, the fully connected word groups are found with semantic correlations. There are still lots of candidate groups with strong correlations but without full node connections. Take the example of (ABCD) and (ABCE), as missing correlations between D and E, they are considered as two small groups with similar topics. The result means that our method is too strict. Wikipedia documents are not able to cover all related words, missing links may exist. In addition during the selection of semantic related pairs, we only pay attention to those appears in documents of each other and the single way links are ignored. A further work is done on the fully connected nod groups. We find groups with one or two different nodes then try to find more information to combine them together. Take D and E in last example, if we can get a single way hyper link from the Wikipedia document, the candidate group of (ABCDE) is accepted. For a group with more words, more missed relations are allowed.

Finally we get a result set with 87,100 groups. For about 1/3 of the groups, the word count is from 3 to 4. Figure 1(a) shows that large groups with more than 5 words cover only 7%, as for more words being difficult to have correlations to each other. As selected by authors for same topics, they are reliable but not so valuable. There are still more than half (59%) of the result being pairs, but not showing a bad result. Considering of the 360,304 semantic related pairs before the experiment, it is only 14.3%. The 3-word groups combine 3 pairs as a unique result, and other large groups with more. There are also some result with more than 10 related words exist in the result set. As being selected from shared tables of Wikipedia meaningfully related to a topic, these result are reliable, though not with such high relatedness. Calculation of degree of semantic correlations is almost the same for a group with more than 2 words. For a m-word group (m¿2), it could also be calculated by summarizing all the part scores in each documents, then divided by $m \times (m-1)/2$ as the result value. Seeing from Figure 1(b), the groups are almost evenly distributed in the range from 2 to 8.

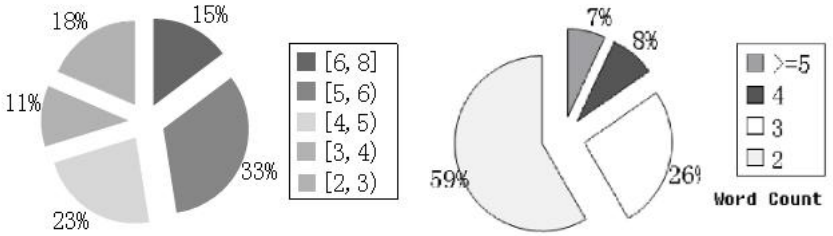

Fig. 1 (a)Groups with Semantic Correlations(b)Distribution of Average Relatedness of Groups.

5 Conclusion and Future Works

In this paper, the Chinese Wikipedia pages are used for semantic related word searching. Considering of hyper-links, text overlaps and word frequency, 360,304 word pairs with semantic correlations are explored from 54,745 structured documents from Wikipedia. We also roughly measured semantic correlations, analyzed the reliability of our measures.

As with similar hierarchical structure, algorithms and applications for WordNet, Hownet may be transplanted to Wikipedia. Semantic Relatedness is used to measuring the degree of semantic correlations, not considering of the difference of relation types. By analyzing the properties of different algorithms based on text overlap or information contents, we are hoping to find a reliable way of searching for groups with semantic correlations and compute the semantic relatedness. For research on semantic relations in NLP, Wikipedia could be employed more in future works.

Acknowledgements This research has been partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B), 19300029. Thanks to Associate Professor Suzuki, and Doctor Matsumoto from The University of Tokushima for instructions.

References

1. D. Ahn, V. Jijkoun etc.: Using Wikipedia at the TREC QA track. In Proc. of TREC-13 (2004)
2. S. Banerjee, T. Pedersen: Extended gloss overlap as a measure of semantic relatedness. In Proc. of IJCAI-03 (2003)
3. M. Strube, SP. Ponzetto: WikiRelate! Computing semantic relatedness using Wikipedia Proc. of AAAI (2006)
4. R. Bunescu ,M. Pasca: Using Encyclopedic Knowledge for Named Entity Disambiguation Proceedings of the 11th Conference of the European Chapter (2006)
5. SP. Ponzetto, M. Strube: Deriving a Large Scale Taxonomy from Wikipedia ,Proceedings of the 22nd National Conference on Artificial (2007)