2008－NL－184 (16)
2008／3／28

Searching and Computing for Vocabularies with Semantic Correlations from Chinese Wikipedia

Yun LI†, Kaiyan HUANG†, Seiji TSUCHIYA††, Fuji REN††

† Graguate School of Advanced Technology and Science, The University of Tokushima

†† Institute of Technology and Science, The University of Tokushima
2-1 Minamijosanjima, Tokushima 770-8506
(liyun, huangky, tsuchiya, ren)@is.tokushima-u.ac.jp

Abstract

This paper introduces our research and experiment on searching for semantically correlated vocabularies in Chinese Wikipedia pages and computing semantic correlations. Based on the 54,745 structured documents generated from Wikipedia pages, we explore about 400,000 pairs of Wikipedia vocabularies considering of hyperlinks, overlapped text and document positions. Semantic relatedness is calculated based on the relatedness of Wikipedia documents. From comparing experiment we analyze the reliability of our measures and some other properties.

1.Introduction

In natural language, vocabulary is usually used to express a concept or object in the world. As things are associated universally, several semantic relations could be found, such as kind-of, part-of, instance-of, attribute-of or other complicate relations. Semantic information and semantic relations are more and more important in Natural Language Processing (NLP), being applied in text retrieval, information extraction, machine translation and other aspect of NLP applications.

In order to achieve semantic computing for NLP, a semantic knowledge-base is necessary. For the complexity of semantic relations, the majority of famous semantic resources are summarized or constructed artificially. The WordNet produced by Princeton University is a synonyms based Ontology formatted semantic knowledge-base, which is widely employed by NLP researchers. For the Chinese language, Hownet and Cilin are famous ones. Hownet gives the semantic tagging and explanation with units founding according to Chinese characters. Cilin is a collection of synonyms with different semantic distance and relatedness, organized in a multi-level pool of categories. As the creation is time-consuming, automatic acquisition of semantic knowledge is important for research. On the other hand, the web would be a good selection for semantic knowledge exploring because of its amount of human knowledge.

Paying attention to the Wikipedia, a Wiki technology based open encyclopedia with millions of multi-language hyperlinked documents. As the documents are written cooperate by Internet users, the text and hyperlinks could express some semantic relations. Wikipedia also contains a Category Graph to index its vocabularies according to their semantic relations. In NLP applications, the Wikipedia resources could not only act as a big corpus, but also a knowledge base or a semantic resource comparable to artificial constructed ones. It has been evaluated in the researches of Gurevych and Strube. Some researchers have

explored the use of Wikipedia for applications such as semantic relatedness computing (Strube&Ponzetto, 2007), name entity disambiguation (Bunescu&Pasca, 2006), automatic question answering (Ahn, 2004), etc.

In this paper, we work on searching and computing Chinese semantic relatedness from Wikipedia documents massive excavation of human knowledge and information of semantic relations. As the vocabulary is enormous, taking account to our aims, firstly we select about 50,000 words according the word frequency of search engines. Then we download the responding Wikipedia pages to generate a structured Chinese Wikipedia Page Corpus for NLP. On this basis, considering of hyperlinks between documents, as well as text overlaps and the location information, about 400,000 word pairs with semantic correlations are selected in our way. The semantic relatedness is calculated from the text location and word frequency information in the Wikipedia pages. For evaluation, we employ different algorithms and experiment on two 1,000 word-pairs sets between one of high level semantic relatedness and the other with low.

In the second part of paper, the preparing work for a Wikipedia document corpus is briefly introduced. In the third part, we introduce details of finding semantic correlated words and the calculation of relatedness. Different algorithms employed for semantic relatedness computing and compare experiment are described in part 4. In the fifth part we do some further experiment to find the self clustering property in the result set. The final part is conclusion and outlook.

2.Prepare document corpus from Wikipedia

A total of 322,121 Wikipedia vocabularies with hyperlinks to pages are selected from the section of "All-pages" from website of Chinese Wikipedia (zh.wikipedia.org). Given that the majority of professional terms have low frequency in normal text, the selection of vocabulary set is needed. The Wikipedia vocabularies are filtered according to frequency in two Internet wordlists, one of which created by the search engine Sogou (www.sogou.com), the other with 800,000 words and frequencies from Google and Baidu (www.baidu.com) collected by some Internet users. Entries existing in the text corpus of People's Daily 2001 were also selected as candidates.

For the selected 66,725 of Wikipedia vocabularies with URL to Wikipedia pages, we download the pages from Chinese Wikipedia (zh.wikipedia.org) and some mirror sites with speed-limited download tools. In others' works, the Wikipedia image package with Wiki format documents is mainly applied instead of web document, but for research on Chinese the usage is impossible. Chinese Wikipedia allows Internet users from the mainland of China, Singapore, Hong Kong, Macao, Taiwan etc editing the source file in their own custom and Chinese format. So the Simplified Chinese and Traditional Chinese are mixed in the source .Before generate web pages, a localization translation services is called automatically to translate into one of the 6 localized Chinese, such as the "mainland simplified Chinese". In order to get the localized document, it is necessary to use the webpage instead of the source file. Since redirect pages exist, we actually have 54,745 HTML pages.

The next step is to analyze the structure and create a structured corpus in accordance with the need of our study. In Wikipedia document, text of the first part is usually the most fundamental and important explanation for the topic word. Usually it could be found before the outline with less than 3 paragraphs from the document. The following paragraphs of a Wikipedia document are detailed information around the topic with less importance than the first part. Some paragraphs of table or lists with manually grouped vocabularies around some topic may share in more than one document, which were called the Shared Tables.

Saying to the degree of semantic relation to the topic word, they are different for the 3 kinds of text parts, so in the process of reforming the Wikipedia document, they should be separately saved in different fields of the text corpus.

As the document hyperlinks between the text lines are important for our research on searching for semantic relations, for a hyperlink pointed to another Wikipedia document, we get the text words, URL keyword and its count of appearance in each part of paragraphs. In the final structured Wikipedia corpus, the following information for a document are saved, such as the length of raw text and html-text, total count of links and duplicated links, keyword and hyperlinks to the Wikipedia Category Graph, the related hyperlinks to an English or a Japanese Wikipedia document with the same title keyword.

For synonyms, a page redirecting is used to access the same document. The synonyms could be found from the title word and the keyword following a mark of "Redirected from" in the redirected Wikipedia document. As synonyms are important for NLP and even semantic computing, we collect groups of synonyms from the Redirected Pages. In the following tests of our research, all the synonyms should be seen as one word in the computing of semantic relatedness, so this step of collecting synonyms is also important. As it is not enough of synonyms, more were collected from the text paragraphs. For example in a document we can find such kind of text :"China Central Television commonly abbreviated as CCTV…", "An astronaut or cosmonaut…". We collected more than ten thousand groups of candidate synonyms, and after manual check 7440 were left and added combined with the redirect page synonyms.

Taking account of synonyms, the amount of vocabularies in the Wikipedia research corpus is raised to 89,994, following with 54,745 XML formatted structured documents generated from the Wikipedia pages. As synonyms and redirection exists, one page is statistically shared by about 1.6 Wikipedia vocabularies. There are totally 1,823,883 hyperlinks found from all the pages, that is averagely 33.3 in one page. 411,402 pairs of pages hyperlinked to each other, which was seen for pages with more relatedness and being considered more useful in our research.

3.Exploring Semantic Correlated Vocabularies

Many research on NLP referring to the semantic relations between words, or even on computing semantic similarity or relatedness. Semantic similarity is shared among instance with same hyper concept, such as ("apple", "pear", "orange") in semantically similar due to being a kind of "fruit". The degrees of similarity are different between ("apple", "pear") and ("apple", "rice") according to the relation of hyper concepts. Synonymous and semantic similarity is an important property considered in the WordNet. The Chinese Cilin is also a collection of synonymous with more than 70,000 Chinese words grouped in different levels according to semantic relations and the degrees of semantic similarity.

As being noticed in the head or this paper, the semantic relationship is complex and diverse. At least 4 basic types of basic semantic relations (kind-of, part-of, instance-of, attribute-of and other complicate relations) and more functional relations are founded. For example ("Bush", "Clinton", "Lincoln") are related by "President of the USA", ("man", "woman") are relative ones," ("Beijing", "Olympics", "Fuwa") are related due to "The 2008 Olympic Games". Comparing to semantic similarity which shows only the "kind-of" semantic relation, semantic correlation is broad and comprehensive. So in the computing of semantic correlations, any type or aspect of semantic relationship should be in the scope. As the standard of semantic relatedness computing is not unique or even not exists. Different researchers applied their own interpretations. Someone employed the semantic category tree with algorithms of looking for public nodes,

counting the depth of node or finding the shortest path between nodes .Different algorithms on WordNet, Hownet could be find in many NLP related papers.

In the public cooperate edited Wikipedia documents, the semantic correlations between the title keyword and the text paragraphs are higher than many other document from web. In one view, the text was usually seen as the representation for the title keyword. In our Wikipedia research corpus, 1,823,883 hyperlinks 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 pairs of pages which have hyperlinks to each other. After studying tens of the pairs by viewing the pages by reading the lines with hyperlinks, we got a pleasant discovery. Most of the title word pairs are semantically correlated, at least sharing some topic or events. As relatedness is a kind of importance of one to another, if something could be noticed easily or usually, the importance should be high, correspondingly their must be some semantic relations exists. As correlates is for both sides, the relatedness between each other should be exists. According to this view we designed our way of finding semantically correlated words and do the relatedness computing from the co-hyperlinked Wikipedia documents.

For word A and B, they are semantic correlated if one ore more of the following conditions meet. Such as A being used in the definition of description of B; A being noticed in introducing of one aspect of B and with importance; A being frequently or easily thinking of if B noticed ;A exists in a list containing B as the topic or similar data. The condition is not enough but useful.

Experiments for semantic correlated vocabularies are done using the information of document hyperlinks. Firstly in Experiment 1 our aim is to find the most correlated word pairs. This time only the Wikipedia basic definition and description of topic keyword is used, which lies in the first part of document. As during the structure work, we separately saved the main part of text and hyperlink information, we directly used the data. For the document with title word A, we got hyperlinked groups (B, C, F, G), then we search for hyperlinked A from each linked documents in this group. If C and G match the rule, we got two result of (A, C) and (A, G). This experiment was done using a C++ test program using a dataset of word and links with integers IDs. As noticed before the IDs for a group of synonyms are the same, so during the experiment and result bellow, they are considered as a unique word. From Experiment 1, 5,512 pairs were explored. We viewed the whole set of result, most relatedness are obvious. Some of which were listed in Table 1 in the form of (A, B) with the English translations.

A(CN)	A(EN)	B(CN)	B(EN)	A(CN)	A(EN)	B(CN)	B(EN)
健身	Fitness	慢跑	Jogging	按钮	Button	人机交互	HCI
半旗	Halfmast	国旗	Flag	彩蛋	EasterEgg	复活节	Easter
傍晚	Evening	下午	PM	钢琴	Piano	伴奏	Accompaniment
棒球	Baseball	球棒	Bat	挑战者	Challenger	航天飞机	SpaceShuttle
赌博	Gambling	赌徒	Gamblers	汽车	Automobile	车牌	LicensePlate
专科	College	教育	Educate	电池	Battery	记忆效应	MemoryEffect
软件	Software	许可证	License	消费者	Consumer	经济学	Economics
信号	Signals	交通	Traffic	文件夹	Folders	文件系统	File_system
立方	Cubic	立方体	Cube	椅子	Chair	轮椅	Wheelchair
拉面	Ramen	兰州市	Lanzhou	博弈论	GameTheory	演化	Evolutionary
癌症	Cancer	癌基因	Oncogene	行动党	ActionParty	马来西亚	Malaysia

阪神	Hanshin	甲子园	Koshien	筷子	Chopsticks	中国烹饪	ChineseCook

Table 1 Some of Semantic Correlated Word Pairs

In the following experiment II, we extend the scope of search to the whole Wikipedia document. As the basic definition or introduction refers only a little part of a topic keyword, the most related materials exists in other paragraphs of the documents. As the result of experiment I show the reliability of semantic information between text lines, the aim of experiment II has been changed 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 hyperlink, they are selected as a candidate pair of Set A, and if one noticed in the main part of another, which is more reliable, we select the pair to Set B. So the rule of Set B is more strict than Set A but looser than what in Experiment I. We get the result descript in Table 2.

Experiment ID	Result Set	Count of Word Pairs	Reviews
I		15,512	Obviously correlated
II	A	79,150	Most obviously correlated
II	B	411,402	With more unrelated pairs

Table 2 Result of Experiment I and II with Manual Evaluation

According to the artificial review, result of Set A is also reliable with most of correlated pairs, but with a high coverage and amount. According to general judgments, part of them has lower relatedness comparing to the set of Experiment I. We also found that for some pairs in Set A, the importance is different from each other, though have correlations. Such as the related words of "春节(Spring Festival)" with "拜年(Say Happy New Year) ", "鞭炮(Firecrackers) ", "年画(New Year Pictures) ".,the "Spring Festival" is more related or more important for "New Year Pictures", but "Spring Festival" could have relatedness with more object such as "Firecracker" , "Say Happy New Year" etc other than "New Year Pictures". For Set B it could be reviewed that though the coverage is large, the accuracy is not acceptable. This count is equal to the co-hyperlinked documents. As text of document except the main part is so free that co-relation of text line may not show enough semantic relations.

As the data of Set B is useful, as a lager percent of valuable correlations are included other than those in Set A, such as words with semantic similarity in lists or tables. A refine work should be down relies on more information not limited to the word frequency, document frequency etc. We checked the most un-related pairs and detected that a common result is that at least one in a pair has a high document frequency. Such as the words of "中国(China)", "公司(Company)", "地区(Regions)", "英语(English)" etc are selected as semantic correlated words for many other keywords but really not. The mistake appears because these words are easily appears in most of the everyday text files, including the Wikipedia documents with a high Document Frequency (DF).

In Experiment III the document frequency and word frequency information are used as filter based on the result of Set B from Experiment II. Firstly by accessing the documents in the corpus, we get the count of document including the object word as a hyperlinked text for all of the selected Wikipedia keywords. The DF value could be calculated with this count divided by the total count of document in our corpus. Then the word pairs with semantic correlations are selected from co-hyperlinks. During this process the Text Frequency (TF) for each keyword is also recognized. For filter, we pay attention on the pairs not appears in Set A of Experiment II. Several times of test are done for getting a proper threshold for a better result with more correlated word pairs and less unnecessary ones. After this work the amount is 360304 pairs, which covers 90%.

4.Semantic Relatedness Computing and comparing experiment

The data of correlated words in our experiments are grouped with a header word and a set of other keywords with different semantic correlations. As it is more meaningful to search and compare between these words. For measuring the semantic relatedness, the location information in documents are mainly considered. According to different level of semantic relatedness, they are ordered in 4 groups.

Level	Group Rules	Relatedness
4	First part co-hyperlinked	Very High
3	Hyperlinked from the first part	High
2	Hyperlinked to the first part	Normal
1	Any other related pairs	Low

Table 3 Group with Different Semantic Relatedness

A value of semantic relatedness are calculated with the normally used the Text Overlap based Measure(TO) which measures based on the relatedness between two words defined as a function of text overlap(Lesk,1986). Commonly this measure computes the overlap score by extending the glosses of the concepts under consideration to include the glosses of related concepts in a hierarch. Given two document of D1 and D2, the overlap(D1,D2) is computed as Σ nm for n phrasal m-word overlaps(Banerjee&Pedersen,2003),the length of text is also need to get a normalized result.

$$\mathrm{Re}late(t_1,t_2) = \tanh\left(\frac{overlap(t_1,t_2)}{length(t_1)+length(t_2)}\right) \qquad \text{(Formula 1)}$$

In the case of Wikipedia corpus, text overlap between documents for words is used for calculation. The full document is used in order to consider to more overlapped information. The frequencies of hyperlinked text keywords are considered with a higher weight for text overlap computing.

Another algorithm making use of Shared Document Frequency (SDF) is also selected as one result of our semantic correlations. For word pair (A, B), using a coops large enough, we get the document count which contains A in text lines as Hit(A), for B it is Hit(B). Then a shared document count Hit(AB) is counted with the documents both noticed A and B. The relatedness value is calculated with the Jaccard Formula

$$Jaccard = \frac{Hit(AB)}{Hit(A)+Hit(B)-Hit(AB)} \qquad \text{(Formula 2)}$$

As different algorithms make used of different semantic information, so the result of semantic relatedness is not always following the same order. But a common understanding of semantic relatedness should be shared, that is decided to the real semantic correlations and distance of the words. By comparing the result of different algorithms, we a hoping to find the common information, and in another way the reliability could be evaluated for each other. In this compare experiment, two 1,000 word-pairs sets between one of high level semantic relatedness and the other with low are used. Word pairs in Group A are selected from the result of Experiment I, which are considered to be highly correlated. As a lower-related set, we randomly select pairs from the result of Experiment IV.

Hyperlinked text and frequency are collect by text program into a unified data file with integers as identification. That is a big sparse matrix with 54,745 rows and columns used for the two algorithms to pick up useful information. For the TO algorithm, firstly the two lines representing vectors in document of the

title keywords are selected, then overlapped nodes with frequency are found and summarized into the value of overlap. The length of document is used instead with the count of hyperlinked text keywords. For the SDF way, according to the Jaccard Formula, all the lines are viewed to select each pairs in the test sets where exists in the same document. And also the document count covered each keyword are also fetched. As limitation of length, only part of the experiment data are listed in Table 4 with the inter IDs changed as a keyword in Chinese and English.

Word A	Word B	SDF	TO
一阶逻辑(First order logic)	论域(Domain)	0.0600	0.0185
中世纪(Medieval)	城堡(Castle)	0.0376	0.0288
交响曲(Symphony)	乐章(Movement)	0.0923	0.0185
人际关系(Interpersonal relationships)	朋友(Friends)	0.0232	0.0101
台湾(Taiwan)	新台币(NT)	0.0170	0.0351
夫家(Husband's family)	妻家(Wife's family)	0.0667	0.1818
惠普公司(Hewlett-Packard)	打印机(Printer)	0.0412	0.0206
春节(Spring Festival)	饺子(Dumplings)	0.0160	0.0210
帝国主义(Imperialism)	殖民主义(Colonialism)	0.0506	0.3548
洗发精(Shampoo)	护发素(Conditioner)	0.0500	0.0870
海外华人(Overseas Chinese)	定居(Settlers)	0.0210	0.0106
狮虎(Lion Tiger)	虎狮(Tiger lion)	0.1250	0.1633
瓷器(China)	高岭土(Kaolin)	0.0160	0.0197
芦沟桥(Lugou Bridge)	永定河(Yongding River)	0.0440	0.0305
鸡(Chicken)	家禽(Poultry)	0.0314	0.0357

Table 4 Result of Semantic relatedness Calculation

For the group with 1,00 pairs of high semantic correlated pairs, they get a average score of 0.050 with the SDF algorithm and the result of TO algorithm is 0.110. Compare to the other group with a SDF value of 0.028 and a TO value of 0.037, they are both higher. It means that the standard of our group set is meaningful. In the mixed set with 2,000 pairs of correlated vocabularies, we ordered the result by the result of both algorithms and then give an indexed position ID. Then the following result is generated as Chart 1showing the percentage of pairs from different groups in different position sections.

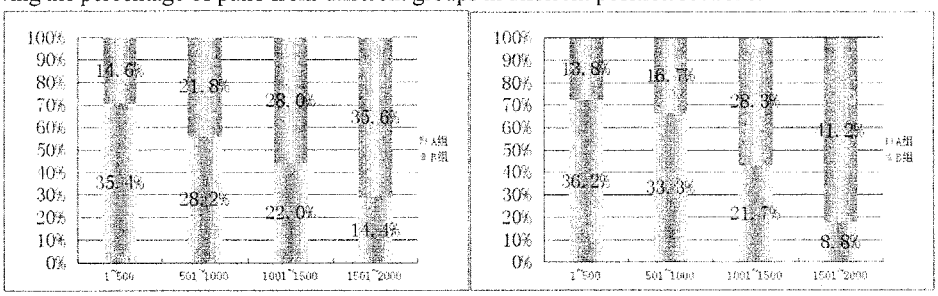

Chart 1 Semantic Relatedness with Different Algorithms on Two Groups

From this experiment, it could be found that for Group A with high relatedness, the result calculated using both algorithm give a acceptable performance, with 35.6% and 41.2% calculated with a high result in the top 25% (with index from 1501 to 2000) among all the 2,000 pairs in this experiment. Most of the pairs selected in the group with high semantic relatedness are supported by both algorithms, with a rate of 69.5%

and 63.6%, which means that our way of exploring words with semantic correlation is generally reliable. It also means that the semantic information in text lines of the Wikipedia document is useful for semantic computing in NLP.

5.Conclusion and Outlook

In this paper, the Chinese Wikipedia pages are used in semantic computing for NLP by searching and Chinese semantic relatedness. We generated our structured Chinese Wikipedia Page Corpus from Wikipedia website. On this basis, considering of hyperlinks between documents, as well as text overlaps and the location information, word pairs with semantic correlations are selected into different groups with different level of semantic correlations. We also employed different algorithms for semantic relatedness calculation, with their result supporting the reliability of our research works.

The hyperlinks between in Wikipedia documents, which help users a lot for research on Chinese word semantic relations, are mostly added by Internet users. As being impossible for authors to know which keyword should be marked as a hyperlinked keyword. Though Wikipedia has programs checking for hyperlinks, the coverage is too less for all the text related informations. Inorder to find more semantic relations, text mining test could be helpfull considering of word frequency or text overlap information in the later research works. For research on semantic relatedness, the Wikipedia Category Graph (WCG) could also be considered as a important resource.

Acknowledgment

This research has been partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B), 19300029,.

References

[1] Ahn, D., V. Jijkoun, G. Mishne, K. M¨uller, M. de Rijke &S. Schlobach (2004). Using Wikipedia at the TREC QA track.In Proc. of TREC-13.

[2] Banerjee, S. & T. Pedersen (2003). Extended gloss overlap as a measure of semantic relatedness. In Proc. of IJCAI-03

[3] Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In Proceedings of the 5th Annual Conference on Systems Documentation, Toronto, Ontario, Canada

[4] M Strube, SP Ponzetto(2006),WikiRelate! Computing semantic relatedness using Wikipedia Proc. of AAAI

[5] Razvan Bunescu ,Marius Pas¸ca (2006),Using Encyclopedic Knowledge for Named Entity Disambiguation Proceedings of the 11th Conference of the European Chapter

[6] SP Ponzetto, M Strube (2007), Deriving a Large Scale Taxonomy from Wikipedia ,Proceedings of the 22nd National Conference on Artificial

[7] T Zesch, I Gurevych(2007),Analysis of the Wikipedia Category Graph for NLP Applications, Proc. of the TextGraphs-2 Workshop, NAACL-HLT,(to appear)