The Eighteenth International Symposium on Artificial Life and Robotics 2013 (AROB 18th '13),
Daejeon Convention Center, Daejeon, Korea, January 30-February 1, 2013

Application of Natural Language Processing for Information Retrieval

Su Mei Xi[1], Dae Jong Lee[2], Young Im Cho[3]

[1]Department of Information, Shandong Polytechnic University, Jinan, 250353, China
[2]College of Electrical and Computer Engineering, Chungbuk National University, Cheong Ju, 361-763, Korea
[3]College of Information Technology, Suwon University, San 2-2, Bongdam-eup, Hwaseong-si, 445-743, Korea

[1]xiyanzi_79@sina.com.cn
[2]bigbell@chungbuk.ac.kr
[3]ycho@suwon.ac.kr

Abstract: Through a comprehensive analysis of using natural language processing in information retrieval, we compared the effects with the various natural language techniques for information retrieval precision in this paper. This is for the tasks of more suitable as well as accurate results of natural language processing.

Keywords: natural language processing, information retrieval, phrase identification, stemming

1 INTRODUCTION

Many researchers have been trying to use natural language processing (NLP) in information retrieval, but the result is not satisfactory. Less complex basic natural language processing techniques with small calculated consumption and simple implementation help a small for information retrieval, which including stop words removal, word segmentation, stemming etc. But some techniques are still recommended in the information retrieval experimental platform, which can improve the retrieval effect such as stop words removal and stemming etc. Senior high complexity of natural language processing techniques with high calculated consumption and low precision can't help the information retrieval even harmful to it, which including parsing, phrase identification, named entity recognition, concept extraction, anaphora resolution and WSD etc [1].

Therefore, in this paper, through a comprehensive analysis of using natural language processing in information retrieval, we will compare the effects with the various natural languages. This paper is organized as follows: Firstly we introduce the application of natural language processing in information retrieval, and secondly we compare the effectiveness of NLP for IR precision, and third, we discuss and finally we conclude of natural language processing in information retrieval.

2 Application of NLP in Information Retrieval

2.1 Application of Basic NLP Technique

2. 1. 1 Stop Word Removal

Stop word refers to the word that lack actual meaning and appear lot of times in the document, such as most of English prepositions, and articles etc. Because stop word removal technique has no substantial help to improve the retrieval effect actual information retrieval systems such as Web search engines often do not use this technique. Moreover, using this technique could not lead to good results in dealing with some queries. The classic example is the query of "to be or not to be". So stop word also has been reserved as index item in most actual retrieval system.

2.1.2 Word Segmentation

Word segmentation is a special problem in information retrieval of Asian languages such as Chinese and Japanese. Most of European languages need not word segmentation. Word segmentation technique is widely used in Chinese information retrieval systems.

Peng et al have made word segmentation and retrieve experiments in the Chinese data set of TREC5 and TREC6 [2]. Their experiments showed that word segmentation accuracy and retrieval effect is not the monotonous directly proportional. The best retrieval effect can be obtained when word segmentation accuracy is 70% or so. If the Word segmentation accuracy is too high, it may lead to decline in retrieval effect.

2. 1. 3 Stemming

Stemming can make the same stem word match the different form word. Commonly used methods include rule-based stemming (e.g. Porter Stemmer) and dictionary-based stemming (e.g. KSTEM).

Strzalkowski and Vauthey applied the stemming method of dictionary-assisted in their retrieval system. The

unreasonable states were improved in the results of stemming and the retrieval precision has 6% to 8% increase [3]. The corpus-based stemming approach improved the retrieval precision slightly to the Porter stemmer and KSTEM put forward by Xu and Croft [4].

In practice, stemming technique was widely used in information retrieval system for its high availability although it can only improve the retrieval effect a little.

2. 1. 4 Part-of-Speech Tagging

It has no obvious usage about part-of-speech tagging to information retrieval. The biggest problem is that we do not know how to use it in retrieval even if it has a very high accuracy .

One approach is that only index the certain parts-of-speech. Kraaij and Pohlmann studied the importance of the different part-of-speech words to retrieval [5]. Their result is that 58% are noun, 29% are verb and 13% are adjective among the document words that useful to retrieval. It can be found that 84% are noun among the useful words if we only focus on those fronts of documents. Arampatzis et al only used nouns to complete the experiment and the result showed that there was 4% improvement compared with using all words [6].

Another use is to separate the different parts of speech of words. Let the words that have the same part of speech in the query and document can be matched. Using TREC7 and TREC8 data sets, Su Qi and others examined the retrieval effect of this use to the SMART system [7].The results show that if we can tag the words that have the same form but different part of speech we will improve retrieval accuracy and decrease the matching noise. The words that have the same meaning, the same stemming and different parts of speech have no match, leading to the decline of retrieval recall.

2.2 Application of Advanced NLP Technique

Advanced natural language processing technique include parsing, phrase identification, named entity recognition, concept extraction, anaphora resolution and word sense disambiguation etc.

2. 2. 1 Phrase Identification

Phrase identification technique used in information retrieval mixed the results, largely depending on the specific recognition technology, the phrase type and the matching strategy[8,9,10]. Nie and Dufort made the phrase as an additional unit to combine with the traditional word-based index. They placed the phrases and words in different vectors, calculated the similarity of the query and document and then added their weights[11].

2. 2. 2 Named Entity Extraction

Named entity is a special phrase that identified a concept or entity, such as proper nouns, place names, organization name and so on. Obviously, the named entities express more accurate information than the general phrases.

But the application of named entity in information retrieval has not obtained the direct effect to retrieval. On the one hand it existing error in the named entity recognition technique itself, on the other hand, researchers have also confused that how to match the named entity.

2. 2. 3 Concept Extraction

Concept is a more general special phrase than named entity. Named entity identifies a concept, so we can consider it belong to the concept. Concept also includes other phrases that not belong to the named entity.

2. 2. 4 Anaphora and Co-references

An anaphora and co-references technique is to find the actual things for the pronoun or the unknown phrase that appeared in the document. This technique seems to have contributed to the information retrieval since it is able to eliminate the unclear expression in the document. However, the truth is not the case. Anaphora and co-references also can not improve information retrieval effect. On the one hand anaphora resolution still has more errors, the other hand pronoun and the unknown phrase does not actually affect the results of information retrieval[12].

2. 2. 5 Word Sense Disambiguation

Voorhees used the word sense disambiguation technique in treating word sense as an index item. It is found that some queries can indeed benefit from the word sense index, however more decline after individually analyzing retrieval results of each query. Almost all the declines are because of the match fails of those words should be matched between queries and documents, since word sense index is used.

Then what word sense disambiguation accurate could help the information retrieval? The answer is 90% which is given by Sanderson[13]. In his experiment he found that the improvement of succeeded disambiguation for retrieval would be offset by the negative effects of failed disambiguation if there was 20% to 30% error rate of word sense disambiguation.

Stokoe et al used Semcor corpus which released with WordNet training the word sense disambiguation system. The words of failed disambiguation would be assigned to the meaning of the highest frequency appeared in the WordNet[14]. Although the accuracy of the word sense disambiguation was only 62%, the experimental results which completed in TREC9 data show that this

disambiguation method can relatively improve the retrieval effect 45%. However, their retrieval results are still worse than the best result of TREC9 even if there is more improvement about retrieval accuracy because the performance of the benchmark system used in their experiments is poor.

Kim et al used a particular Word Sense Disambiguation technique. They only consider the 25 most original word meanings of a word in the WordNet and then assign a meaning to a word, which can insure the accuracy of WSD[15].

2.3 Adding Natural Language Processing Technique into Language Model

The nature of the language model is to determine the relevance between the query and document by computing the probability of generating the query from the document.

Kumaran and Allan joined the stemming technique in the language model. They thought that Stemming can be seen as smoothing [16]. They proposed a generative model. In this model, they thought that it can be divided into two steps from d to w. Firstly, d generates c, and then c generates w. This assumption comes from the writer writing. When writers select the words they usually think to a meaning and then select the word with correct form. Experimental results show that the average retrieval accuracy of the new model has increased about 10%.

2.4 The Application Effect of NLP Technique in Information Retrieval

TREC5 NLP evaluation results show that query expansion, phrase identification, terminology, and stemming and other natural language processing Techniques used in information retrieval can obtain better effect than the word-based retrieval system. These systems that applied these technologies are still not better than the systems based on statistical [17].

Manning analyzed the roles of natural language processing techniques in Web Retrieval. In fact, Web retrieval technique has gained more progress through Web link analysis techniques and AnchorText.

2.5 The Application of NLP Resources in Information Retrieval

Natural language processing resources refer to the dictionary like WordNet and HowNet. Smeaton experimented with WordNet after the failure that using natural language technique to retrieval experiment such as parsing. In addition, as mentioned above, WordNet is commonly tools used in word sense disambiguation.

Natural language processing resources are constructed manually or revised manually after generated by machine. It has a very high accuracy and it is suitable to be used in information retrieval. We should use it according to the actual situation for different problems.

3 Comparison of Different NLP Techniques Applied in IR Systems

TREC NLP evaluation results (table1 and table2) show that phrase identification, terminology, and stemming and other natural language processing Techniques used in information retrieval can obtain better effect than the word-based retrieval system.

Table 1. The experimental results of basic NLP techniques used in IR systems

	TREC-5		TREC-6	
	Avg.Prec.	R.Prec.	Avg.Prec.	R.Prec.
Stop word removal	No use	null	No use	null
Word segmentation	0.3721	0.3988	0.5044	0.5072
Stemming	0.328	0.356	0.273	0.304
Part-of-speech tagging	0.4632	0.4804	0.2692	0.2712

Most of the trial had no good effect, which try to use natural language processing in information retrieval, even had some little help, it was not satisfied for people. Researchers have analyzed, and draw conclusions that it was need to be optimized that natural language processing for information retrieval tasks.

Table 2. The experimental results of advanced NLP techniques used in IR systems

	TREC-5		TREC-6	
	Avg.Prec.	R.Prec.	Avg.Prec.	R.Prec.
Phrase identification	0.2347	0.2939	0.2434	0.2574
Named entity extraction	0.2860	0.2835	0.1961	0.2014
Concept extraction	0.2613	0.2545	0.3346	0.3321
Anaphora co-references	0.2998	0.2876	0.3261	0.3211
Word sense disambiguation	0.3291	0.1994	0.2426	0.1980

4 CONCLUSION

For a long time, the development of natural language processing is to be applied to the tasks which need precise results such as machine translation, so the role of natural

language understanding may be larger in the question answering system, automatic abstract and information extraction. In face, in these tasks, we have achieved good results through the interaction between the NLP and IR. The results of TREC also show that Natural language processing can improve the effect of these tasks.

It can be seen in table 3 that the tasks of longer length of the query are more suitable for using natural language processing, such as information extraction and question answering system etc. It is conceivable that the tasks of shorter length results more need to use natural language processing because of the need for syntactic and semantic processing in order to ensure its accuracy. Therefore, compared to the document retrieval, information extraction and question answering system are more suitable for the use of natural language processing.

Table 3. Classification of Document Retrieval, Passage Retrieval, Question Answering, and Information Retrieval according to query length and result length

Query length / Result length	Short	Medium	Long
Short			QA/ IE
Medium		PR	
Long	DR		

REFERENCES

[1] Ricardo Baeza-Yates, "Challenges in the Interaction of Information Retrieval and Natural Language Processing". In: Proceedings of 5th International Conference on Intelligent Text Processing and Computational Linguistics , CICLing 2004 , Seoul , Korea , February 15-21 , pp.445-456, 2004.

[2] Fuchun Peng , Xiangji Huang , Dale Schuurmans and Nick Cercone, "Investigating the Relationship between Word Segmentation Performance and Retrieval Performance in Chinese IR", In: Proceedings of 19th International Conference on Computational Linguistics, pp.72-78, 2002.

[3] Tomek Strzalkowski and Barbara Vauthey, "Information retrieval using robust natural language processing", In: Proceedings of the 30th annual meeting on Association for Computational Linguistics, pp.104-111,1992.

[4] J Xu and W. B. Croft , "Corpus-based stemming using cooccurrence of word variants", ACM Transactions on Information Systems (TOIS) , vol.16,no.1, pp.61-81, 1998.

[5] W. Kraaij and R. Pohlmann, "Viewing stemming as recall enhancement", In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval ,ACM Press, pp.40-48, 1996.

[6] A. T. Arampatzis , Th. P. van der Weide , C. H. A. Koster and P. van Bommel , "Text Filtering using Linguistically-motivated Indexing Terms". Technical Report CSI-R9901, Computing Science Institute, University of Nijmegen , Nijmegen , The Netherlands ,1999.

[7] Qi Su,Hongying Zan, "Effects of POS Tagging on Performance of IR Systems", Journal of Chinese Information Processing, vol.19,no.2 , pp.58-65,2005.

[8] Thorsten Brants, "Natural Language Processing in Information Retrieval", In: Proceedings of 20th International Conference on Computational Linguistics , Antwerp, Belgium, pp.1-13, 2004.

[9] M. Mit ra , C. Buckley , A. Singhal , and C. Cardie, "An analysis of statistical and syntactic phrases", In : Proceedings of the RIAO97 , pp.200-216,1997.

[10] S. E. Robert son and S. Walker, "Okapi/ Keenbow at TREC28", In: Proceedings of the 8th Text Retrieval Conference, NIST Special Publications 500-246, Gaithersburg , pp.151-162, 1999.

[11] Jian-Yun Nie and J ean-Francois Dufort, "Combining words and compound terms for monolingual and cross-language information retrieval", In: Proceedings of Information. Beijing, pp.453-458, 2002.

[12] James Allan, "Natural Language Processing for Information Retrieval", tutorial presented at the NAACL/ANL P language technology joint conference in Seattle ,Washington , April 29 , 2000.

[13] M. Sanderson, "Word Sense Disambiguation and Information Retrieval", In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval , ACM Press, pp.49-57,1994.

[14] Christopher Stokoe , Michael P. Oakes and John Tait , "Word Sense Disambiguation in Information Retrieval Revisited", In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press , pp.159-166, 2003.

[15] Sang-Bum Kim , Hee-Cheol Seo and Hae-Chang Rim, "Information Retrieval using Word Senses : Root Sense Tagging Approach" , In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,ACM Press, pp.258-265, 2004.

[16] James Allan and Giridhar Kumaran, "Stemming in the Language Modeling Framework", In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (poster) ,ACM Press , pp.455-456,2003.

[17] Canhui Wang , Min Zhang , Shaoping Ma, "A Survey of Natural Language Processing in Information Retrieval", Journal of Chinese Information Processing, vol.21,no. 2,pp.40, 2007.