Contrastive Approach towards Text Source Classification based on Top-Bag-of-Word Similarity

Chu-Ren Huang[1,2], and Lung-Hao Lee [2]
[1]Hong Kong Polytechnic University
[2]Institute of Linguistics, Academia Sinica
churen.huang@polyu.edu.hk
{churen, lunghao}@gate.sinica.edu.tw

Abstract. This paper proposes a method to automatically classify texts from different varieties of the same language. We show that similarity measure is a robust tool for studying comparable corpora of language variations. We take LDC's Chinese Gigaword Corpus composed of three varieties of Chinese from Mainland China, Singapore, and Taiwan, as the comparable corpora. Top-bag-of-word similarity measures reflect distances among the three varieties of the same language. A Top-bag-of-word similarity based contrastive approach was taken to solve the text source classification problem. Our results show that a contrastive approach using similarity to rule out identity of source and to arrive actual source by inference is more robust that directly confirmation of source by similarity. We show that this approach is robust when applied to other texts.

Keywords: Top-bag-of-word similarity, Text Source Classification, Contrastive Approach, Comparable Corpus, Chinese Gigaword.

1. Introduction

Comparable corpora are corpora which select similar texts in more than one language or language variety[1]. These texts are typically gathered during the same time period. Comparable corpora are different from parallel corpora are widely used as resources for statistical machine translation, bilingual lexicons. Comparable corpora overcome the scarcity and limitations of parallel corpora, since sources for original, monolingual texts are much more abundant (Barzilay and Elhadad, 2003; Munteanu et al., 2004; Shao and Ng, 2004; Talvensaari et al., 2007).

The degree and nature of lexical similarity and contrast among Mandarin Chinese used in different Chinese speaking societies were widely observed but not thoroughly studied due to the lack of comparable corpora. Recently, LDC's Chinese Gigaword (2003)contains three sets of monolingual corpora selected according to the same set of criteria but in different language varieties from China, Singapore and Taiwan. We will explore it as a comparable corpus for variations of Chinese in this paper. In particular, we propose a measure of top-bag-of-word similarity for comparing the language variants contained in Chinese GigaWord corpus. Texts from the same period of time from Central News Agency (Taiwan), Xinhua News Agency (PRC) and Lianhe Zaobao (Singapore) are extracted and compared in our study. By comparing these three varieties of Mandarin Chinese, we hope to find the language significant lexical contrasts and meaning variations. We also propose a constrative approach towards automatic text source classification based on co-occurence similarity measures with documents from the same time period of Chinese Gigaword. Experimental resultes indicated that our proposed constrastive approach is reliable and robust.

The rest of this paper is organized as follows. Section 2 investigates related literature in word similarity measures in comparable corpus and a brief introduction to Chinese Gigaword. Section 3 describe the text source classification based on co-occurence similarity. Section 4 presnts experimental results and further discussion. Finally, Section 5 concludes this study.

[1] Definition of comparable corpus according to EAGLES report, accessed at http://www.ilc.cnr.it/EAGLES/corpustyp/node21.html

22nd Pacific Asia Conference on Language, Information and Computation, pages 404–410

2. Literature Review

2.1 Word Similarity Measures in Comparable Corpus

A comparable corpus is one which gathered the similar texts in more than one language or language variety from the same time periods. Comparable corpora were widely used in researches issues consist of machine translation, natural language processing and cross language information retrieval. Fung and Yee (1998) used a comparable-corpus-based approach to estimate the similarity between a word and its translation candidates. Fung and Cheung (2004) used multi-level bootstrapping to iteratively improve alignment for extracting parallel sentences from a quasi-comparable corpus. Cheng et al. (2004) mined bilingual search results obtained from search engines to translate unknown query terms. Their approach was with Web corpora that can alleviate the problem of the lack of large bilingual corpora and benefit cross-language Web search.

Barzilay and Elhadad (2003) focused on monolingual comparable corpus, i.e. texts in the same language to address the task of sentence alignment. They found that context plays an important role to combine with a sentence similarity measure. Shao and Ng (2004) proposed a method by combining both context and transliteration information for the task of mining new word translations. They translated Chinese Words into English and tested it on Chinese and English Gigaword. Munteanu et al. (2004) improved machine translation performance via parallel sentence extraction from comparable corpora that consist of two large monolingual news texts in English and Arabic. Talvensaari et al. (2007) used Relative Average Term Frequency (RATF) valve to create a comparable corpus from articles by a Swedish news agency and a U.S. newspaper.

Chen and You (2002) proposed using only syntactic related co-occurrences as context vectors and adopted information theoretic methods for measuring word similarity to solve the problem of data sparseness and characteristic precision. Gao et al. (2002) extended the basic co-occurrence model by adding a decaying factor that decreases the mutual information when the distance between the terms increases. The experimental results also indicated that their proposed triple translation model brings further improvements than word-by-word translation. Weeds and Weir (2005) proposed a flexible framework called as co-occurrence retrieval for lexical distributional similarity. Zheng et al. (2007) presented a novel word co-occurrence model based on an ontology representation of word sense and its related applications.

2.2 Introduction to Chinese Gigaword

Automatic annotation is remains a challenging task in Chinese language processing. For instance, ACL SigHan has hosted four bakeoff competition for segmentation, but none for POS tagging. There is only a handful of POS tagging systems and automatic taggers which are widely accepted and accessible. In Taiwan, Academia Sinica's CKIP tagset has been considered the standard and has been used in annotating the Sinica Corpus (CKIP, 1995/1998), which were first annotated in 2006 and contains roughly 10 million words in the latest version (2007). In PRC, the Institute of Computational Linguistics (ICL)'s tagset has been considered the de facto standard and is widely available through the POS tagged People's Daily Corpus (Yu et al., 2002; 2003). However, an even greater challenge occurs with the new demand of very large corpora and the availability of the untagged LDC Gigaword Corpus.

The Chinese Gigaword Corpus (CGW) released in 2003 by Linguistic Data Consortium (LDC). It contains about 1.12 billion Chinese characters, including 735 million characters from Taiwan's Central News Agency (CNA) from 1991 to 2002, and 380 million characters from Mainland China's Xinhua News Agency (XIN) from 1990 to 2002. CNA uses the complex character form and XIN uses the simplified character form. CGW has three major advantages for the corpus-based Chinese linguistic research: (1) It is large enough to reflect the real written language usage in either Taiwan or Mainland China. (2) All text data are presented in a SGML form, using a markup structure to provide each document with rich metadata for further inspecting. (3) CGW is appropriate for the comparison of the Chinese usage between Taiwan

and Mainland China, because it provides the same newswire text type, and these news texts were almost published during the overlapping time period.

LDC's Chinese Gigaword Corpus currently has a segmented and tagged version available (Huang, 2007). This version adopts the CKIP tagset and enhanced Sinica Word Segmenter (Ma and Chen, 2005) to segment the corpus into the words. And they utilized HMM method for POS tagging and morpheme-analysis-based method (Tseng and Chen, 2002) to predict POSs for new words. The annotated Chinese Gigaword Corpus was also performed automatically with automatic and partially manual post-checking (Ma and Huang, 2006). The precision accuracy is estimated to be over 95% for Central New Agency part of data from Taiwan. Quality assurance of automatic annotation of Chinese Gigaword Corpus based on heterogeneous tagging system is also proposed to improve the precision accuracy (Huang et al., 2008).

3. Text Source Classification based on Top-bag-of-word similarity

3.1 Top-bag-of-word similarity Measures

Although there is as yet no agreement on the nature of the similarity of comparable corpora, there were several attempts to use comparable corpora for making similarity measures. In these attempts, top-bag-of-word similarity metric was widely used because it is simple and efficient (Gao et al., 2002; Chen and You, 2002; Cheng et al. 2004; Fung and Cheung, 2004; Weeds and Weir, 2005; Zheng et al, 2007).

In our approach, the corpus is firstly represented as "bags of words" (Baeza-Yates and Ribeiro-Neto, 1999, Manning and Schutze, 1999). The top frequency word types (Weeds and Weir, 2005) are continuously selected as the main features for comparing the language variants. Once the corpus is reformatted as top bags of words, their similarity metric is defined in equation 1:

$$\text{Sim}(C_i, C_j) = \text{Co-Num}(C_i, C_j) \ / \ \text{Num}(C_j) \qquad (1)$$

Where $\text{Sim}(C_i, C_j)$ denotes the similarity between corpus i and corpus j; $\text{Num}(C_j)$ denotes the number of representative words of corpus j; and $\text{Co-Num}(C_i, C_j)$ denotes the number of word types which occurr in both corpus i and j. Obviously, the similarity between a corpus and itself, as well as another corpus which contains exactly the same word types, is 1and the similarity is 0 if there are two corpus i and j have not common word types.

3.2 Text Source Classification

We further formulate the similarity matrix and determined intervals for text categorization. Table 1 shows a similarity matrix based on the definition of our top-bag-of-word similarity measures in comparable corpus of three kinds of language variations: C_1, C_2 and C_3. The similarity of corpus itself i.e. Sim (C_1, C_1), Sim (C_2, C_2) and Sim (C_3, C_3) is 1. Sim (C_i, C_j) is equal to Sim (C_j, C_i), that is because of the corpus is reformatted as the same size of top bags of words based on this common word type measure. Assuming the similarity of Corpus C_1 and C_2 is *a*, similarity of Corpus C_2 and C_3 is *b* and similarity of Corpus C_2 and C_3 is *c*.

Table 1: A top-bag-of-word similarity matrix of C_1, C_2 and C_3.

	C_1	C_2	C_3
C_1	1	*a*	*c*
C_2	*a*	1	*b*
C_3	*c*	*b*	1

As this similarity formulation, 1 means that two corpora have the same representing word types. Oppositely, 0 means the no representing word type co-occurs in these two corpora. If the similarity, a, is larger than c, it means C_2 is more close to C_1 than C_3. Our hypothesis is that if

C_1 is applied as a baseline for classifying corpus sources, the text of document belongs to C_1 when its similarity falls into the interval [a, 1]; the text of document belongs to C_2 when its similarity falls into the interval [c, a]; if similarity fall into the interval [0, c], the source of document is classified as C_3. The classification process will adopt similar heuristics to generate determined intervals when C_1, C_2 and C_3 are applied as a classification baseline individually. Different from individual corpus based classification, we further use a contrastive elimination algorithm that simple majority voting mechanism is employed for determining the final classification results. For example, if C_2 receives two votes and C_3 receives only one vote, so C_2 wins as the final classified results.

4. Evaluation

4.1 Data Source

In order to compare the use of Chinese, the same time period of tagged Chinese Gigaword (Huang, 2007) is selected as comparable corpus. Three parts of resource is consisting of Taiwan's Central News Agency (CNA), Mainland China's Xinhua News Agency (XIN) and Singapore's Lianhe Zaobao (ZBN). The same time period is October to November in 2000, January 2001 and April to September in 2003. Table 2 shows the document size of these comparable corpora.

Table 2: Number of Document in data source

	# of Doc. in CNA	# of Doc. in XIN	# of Doc. in ZBN
200010	13,010	6,198	2,091
200011	13,123	5,694	2,060
200012	12,736	5,990	1,991
200101	11,518	5,565	1,972
200304	1,028	7,156	5,189
200305	934	6,202	5,897
200306	735	5,786	5,531
200307	787	6,613	5,637
200308	669	5,060	5,478
200309	361	6,461	5,572

In Table 2, it' not difficult to find the document size is significant difference. The size of CNA in 2003 is obviously smaller than that in 2001 and meanwhile smaller than the other two sources, XIN and ZBN. Since the document size is so different, we further analyzed the number of distinct word types in these comparable corpora.

We adopted the CKIP tagset and enhanced Sinica Word Segmenter (Ma & Chen, 2005) to segment the corpus into the words. By these analyzed results shown in Table 3, we finally decided to use the top 5 thousands distinct word types for representing every comparable corpus individually. And we randomly selected a half of the same time period i.e. 200010, 200012, 200304, 200306 and 200308 from three different sources as training data sets and the remaining parts of these corpora as testing data sets for applying similarity measure to text categorization, that is to say that the documents of testing data sets would be blind predicated into only one of CNA, XIN and ZBN.

4.2 Experimental Results

Table 4 shows the co-occurrences similarity of CNA, XIN and ZBN. It is clear that all similarities of these three trained corpora is larger than 0.6 regardless of different sources. The main reason for high similarity is because all documents from Chinese Gigaword with the same time period and similar topics of documents. When we finished calculating this similarity matrix, determined intervals were generated according to our assumptive heuristics.

Table 4: A top-bag-of-word similarity matrix of CNA, XIN and ZBN

	CNA	XIN	ZBN
CNA	1	0.6068	0.6276
XIN	0.6068	1	0.6814
ZBN	0.6276	0.6814	1

Table 5 shows the experimental results of classifying text sources in testing data sets. In individual source based classification, there is an agreement on classified results excluding CNA based classification. In CNA based classification column, "*" notation denoted incorrect classification results. All incorrect files were classified as XIN by CNA based classification, which is because of the lower top- bag-of-word similarity resulting from variant words for describing topics. Note that although document sizes of comparable corpora vary greatly, they do not distort the expected result from similarity measurement comparison. Hence we are assured of the robustness of the top-bag-of-word similarity measure regardless of corpus size variations.

Table 5: Experimental results of text source classification in Chinese Gigaword

	CNA Based Classification	XIN Based Classification	ZBN Based Classification	Majority Voting
cna_200010	0.9294 (C)	0.5974 (C)	0.6174 (C)	C
cna_200012	0.9246 (C)	0.594 (C)	0.6142 (C)	C
cna_200304	0.4322 (X)*	0.3544 (C)	0.3538 (C)	C (C:2 , X:1)
cna_200306	0.407 (X)*	0.3296 (C)	0.3344 (C)	C (C:2 , X:1)
cna_200308	0.4124 (X)*	0.326 (C)	0.3302 (C)	C (C:2 , X:1)
xin_200010	0.576 (X)	0.8658 (X)	0.6458 (X)	X
xin_200012	0.5828 (X)	0.8632 (X)	0.648 (X)	X
xin_200304	0.5946 (X)	0.8506 (X)	0.6744 (X)	X
xin_200306	0.5898 (X)	0.8626 (X)	0.659 (X)	X
xin_200308	0.5842 (X)	0.8452 (X)	0.6576 (X)	X
zbn_200010	0.6216 (Z)	0.6742 (Z)	0.8138 (Z)	Z
zbn_200012	0.6214 (Z)	0.6626 (Z)	0.802 (Z)	Z
zbn_200304	0.5876 (X)*	0.646 (Z)	0.8704 (Z)	Z (Z:2, X:1)
zbn_200306	0.5982 (X)*	0.648 (Z)	0.8908 (Z)	Z (Z:2, X:1)
zbn_200308	0.5956 (X)*	0.6428 (Z)	0.885 (Z)	Z (Z:2, X:1)

The characters of our proposed contrastive elimination algorithm that simple majority voting mechanism is employed for determining the final classification results are combined with similarity and dissimilarity measures from the suspected text sources. For example, the testing file "cna_200304", if only applied similarity measure from the same text source, that's to say, just CNA based classification was applied, experimental results indicated that this file will be classified as wrong text source XIN. Further analysis found that although the same time period

of news text from PRC, Taiwan and Singapore were selected as parts of Chinese Gigaword. There are still existing greatly differentiated topics. The contents of selected documents in CNA corpus consisted of many weather reports. But if we involved dissimilarity measures from the other sources and majority voting mechanism was adjusted final decision, we can get the accurate prediction. It was proven that our proposed method will be reliable for variant words to describe different topics. We are also assured of the robustness of our contrastive elimination algorithm regardless of corpus topic variations.

5. Conclusion and Future Work

We propose a top-bag-of-word similarity measures for classifying texts from different variants of the same language. We take LDC's Chinese Gigaword Corpus composed of three varieties of Chinese from Mainland China, Singapore, and Taiwan, as the comparable corpora. Top-bag-of-word similarity measures are shown reflect distances among the three varieties of the same language. Our results show that proposed contrastive approach using similarity to rule out identity of source and to arrive actual source by inference is more robust that directly confirmation of source by similarity. And the document size does not influence the prediction. This robust result is notable given the similarity and almost very high degree of mutual intelligibility among these variants.

Ongoing work is focusing on verifying the robustness of the top-bag-of-word similarity measure on outside data and with data from more than three different sources since study is an empirical research in Chinese Gigaword corpus. Other similarity measures for comparable corpus study are also being investigated.

References

Baeza-Yates, Ricardo and Berthier Ribeiro-Neto. 1999. Modern Information Retrieval. *The ACM Press*.

Barzilay, Regina and Noemie Elhadad. 2003. Sentence Alignment for Monolingual Comparable Corpora. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Chen, Keh-Jiann and Jia-Ming You. 2002. A Study on Word Similarity using Context Vector Models. *Computational Linguistics and Chinese Language Processing.* 7(2), 37-58.

Cheng, Pu-Jen, Jei-Wen Teng, Ruei-Cheng Chen, Jenq-Huar Wang, Wen-Hsiang Lu, and Lee-Feng Chien. 2004. Traslating Unknown Queries with Web Corpora for Cross-Language Information Retrieval. *In Proceedings of the 27th annual International ACM SIGIR Conference on Research and Development in Informational Retrieval.*

CKIP (Chinese Knowledge Information Processing Group). 1995/1998. The Content and Illustration of Academica Sinica Corpus. *(Technical Report no 95-02/98-04). Taipei: Academia Sinica.*

Fung, Pascale and Lo Yuen Yee. 1998. An IR Approach for Translating News Words from Nonparallel, Comparable Texts. *In Proceedings of International Conference on of Computational Linguistics*

Fung, Pascale and Percy Cheung. 2004. Multi-level Bootstrapping for Extracting Parallel Sentences from a Quasi-Comparable Corpus. *In Proceedings of International Conference on of Computational Linguistics*.

Gao, Jianfeng, Jian-Yun Nie, Hongzhao He, Weijun Chen and Ming Zhou. 2002. Resolving Query Translation Ambiguity using a Decaying Co-occurrence Model and Syntactic Dependence Relations. *In Proceedings of the 25th annual International ACM SIGIR Conference on Research and Development in Informational Retrieval.*

Huang, Chu-Ren, Lung-Hao Lee, Wei-guang Qu and Shiwen Yu. 2008. Quality Assurance of Automatic Annotation of Very Large Corpora: a Study based on Heterogeneous Tagging Systems. *In Proceedings of the 6th International Conference on Language Resources and Evaluation.*

Huang, Chu-Ren. 2007. Tagged Chinese Gigaword. Linguistic Data Consortium, Philadelphia.

Ma, Wei-Yun and Chu-Ren Huang. 2006. Uniform and Effective Tagging of a Heterogeneous Giga-word Corpus. *In Proceedings of the 5th International Conference on Language Resources and Evaluation.*

Ma, Wei-Yun and Keh-Jiann Chen. 2005. Design of CKIP Chinese Word Segmentation System. *Chinese and Oriental Languages Information Processing Society*, 14(3), 235-249.

Manning, Christopher D. and Hinrich Schutze. 1999. Foundations of Statistical Natural Language Processing. *The MIT Press*.

Munteanu, Dragos Stefan, Alexander Fraser and Daniel Marcu. 2004. Improved Machine Translation Performance via Parallel Sentence Extraction from Comparable Corpora. *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Shao, Li and Hwee Tou Ng. 2004. Mining New Word Translations from Comparable Corpora. *In Proceedings of International Conference on of Computational Linguistics*.

Talvensaari, Tuomas, Jorma Laurikkala, Kalervo Jarvelin, Martti Juhola and Heikki Keskustalo. 2007. Creating and Exploiting a Comparable Corpus in Cross-Language Information Retrieval. *ACM Transactions on Information Systems*. 25(1), 1-21.

Tseng, Huihsin and Keh-Jiann Chen. 2002. Design of Chinese Morphological Analyzer. *In Proceedings of 1st SIGHAN Workshop on Chinese Language Processing*.

Yu, Shiwen, Huiming Duan, Xuefeng Zhu, Bin Swen and Baobao Chang. 2003. Specification for Corpus Processing at Peking University: Word Segmentation, POS Tagging and Phonetic Notation. *Journal of Chinese Language and Computing*, 13(2),121-158.

Yu, Shiwen, Huiming Duan, Xuefeng Zhu and BinSun. 2002. The Basic Processing of Contemporary Chinese Corpus at Peking University- Specification. *Journal of Chinese Information Processing*. 16(5&6), 49-64.

Weeds, Julie and David Weir. 2005. Co-occurrence Retrieval: A Flexible Framework for Lexical Distributional Similarity. *Computational Linguistics*. 31(4), 439-475

Zheng, Dequan, Tiejun Zhao, Sheng Li and Hao Yu. 2007. Research on a Novel Word Co-occurrence Model and Its Application. Lecture Notes in Computer Science. 4798, 437-446.