A Chinese Sentence Compression Method for Opinion Mining

Shi Feng[1], Daling Wang[1], Ge Yu[1], Binyang Li[2], and Kam-Fai Wong[2]

[1] Northeastern University, Shenyang, China
{fengshi,wangdaling,yuge}@ise.neu.edu.cn
[2] The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, China
{byli,kfwong}@se.cuhk.edu.hk

Abstract. The Chinese sentences in news articles are usually very long, which set up obstacles for further opinion mining steps. Sentence compression is the task of producing a brief summary at the sentence level. Conventional compression methods do not distinguish the opinionated information from factual information in each sentence. In this paper, we propose a weakly supervised Chinese sentence compression method which aiming at eliminating the negligible factual parts and preserving the core opinionated parts of the sentence. No parallel corpus is needed during the compression. Experiments that involve both automatic evaluations and human subjective evaluations validate that the proposed method is effective in finding the desired parts from the long Chinese sentences.

Keywords: Sentence Compression, Opinion Mining.

1 Introduction

"What other people think" has always been an important piece of information for most of us during the decision-making process [15]. The goal of opinion mining is to extract and summarize opinionated contents from news, blogs, comments and reviews. It has recently attracted much attention because of its wide range of applications, such as marketing intelligence, government policy making and so on.

Opinion holder and target extraction are fundamental task of opinion mining, and a lot of literatures have been published in this area [1,2,10]. However, the accuracy of this task for Chinese news articles is far from acceptable [17]. This is partially because Chinese sentences are usually very long and often connects two or more self-complete sentences together without any indicating word or punctuation. Therefore, the parsing approach will bring in more errors and noisy for the extraction models and methods.

Sentence compression is a recent framework that aims to select the shortest subsequence of words that yields an informative and grammatical sentence [12]. Most previous studies focus on preserving the critical information of the sentence. However they do not differentiate the opinionate part of the sentence from the factual ones. Table 1 shows a compressed sentence addressing the conventional content-oriented compression task (CS) and an example of opinion-oriented compressed sentence (OOCS).

P.-J. Cheng et al. (Eds.): AIRS 2010, LNCS 6458, pp. 320–329, 2010.

Table 1. The example of conventional and opinion-oriented compressed sentences

Original Sentence (OC)	700多名群众聚集在西南部胡齐斯坦省首府阿瓦士的一家神学院门前，强烈抗议美英对伊拉克发动军事打击。 More than 700 people gathered in front of a theological school in the south-western Khuzestan province, capital of Ahvaz, and strongly protested against the United States and Britain launching a military strike against Iraq.
Compressed Sentence (CS)	群众聚集在神学院门前，抗议美英对伊拉克发动军事打击。 People gathered in front of theological school, and protested against the United States and Britain launching a military strike against Iraq.
Opinion-Oriented Compressed Sentence (OOCS)	群众强烈抗议美英对伊拉克发动军事打击。 People strongly protested against the United States and Britain launching a military strike against Iraq.

From Table 1 we can see that the conventional compression task can get the core information of each sub-sentence. Different from that, the goal of opinion-oriented sentence compression (OOSC) is to eliminate non-opinionated sub-sentence and retain the opinionate part of the sentence. The "gathered" part of the sentence is deleted because it does not express any opinion information. Therefore, the result words in OOCS yield a shortest opinionated and grammatical sentence, which paves the way for the further opinion mining steps.

The traditional sentence compression method could not meet the goal of opinion-oriented sentence compression, because it may retain unnecessary factual part of the sentence. Moreover, it sometimes ignores the opinion words which are very important for OOSC task. On the other hand, an ideal opinion-oriented compressed sentence should not only preserve the opinion holder, target information of the sentence, but also eliminate non-opinionate part of the sentence. It provides a brief summary of the opinion expressed in the sentence, and the shorten sentence brings in a higher parsing accuracy, which will facilitate opinion extraction task in the next mining steps.

Until recently, many papers have been published for sentence compression using both supervised and unsupervised method. However, there are still important challenges to be tackled for opinion-oriented sentence compression:

(1) The Chinese sentences in news articles are usually very long, which brings in errors for parsing based compression methods;
(2) The opinion-oriented sentence compression is lack of parallel corpus;
(3) The compression approach should not only consider the term information importance and grammatical consistency, but also the opinion-related weights of each word.

In this paper, we propose a scoring based opinion-oriented compression method for Chinese news sentences. To best of our knowledge, this is the first paper that seeks to compress Chinese sentences for opinion mining task. The rest of the paper is organized as follows. Section 2 introduces the related work on opinion mining and sentence compression. Section 3 presents the proposed dynamic programming approach for opinion-oriented Chinese sentence compression. Section 4 provides experimental results on Chinese news datasets, including automatic evaluations and human subjective evaluations. Finally we present concluding remarks and future work in Section 5.

2 Related Work

Recently, opinion mining has become a hot topic in research area. For sentiment classification task, documents are classified into positive and negative according to the overall sentiment expressed in them. Turney et al. [18] measured the strength of sentiment by the difference of the Pointwise Mutual Information (PMI) between the given phrase and the seed words. In [14], Pang et al. employed three machine learning approaches (Naive Bayes, Maximum Entropy, and Support Vector Machine) to label the polarity of IMDB movie reviews.

Extracting opinion holders, targets and expressions from documents have attracted many researchers' attentions [3,4,10]. Choi et al. presented an integer linear programming approach for the joint extraction of entities and relations in the sentence. Performance of the system could be further improved when a semantic role labeling algorithm is incorporated [2]. Wu et al. proposed a novel phrase dependency parsing approach for mining opinions from product reviews, where it converted opinion mining task to identify product features, expressions of opinions and relations between them [20]. Although these methods have achieved relatively high extraction accuracy, they are still sensitive to the parsing errors, which set up obstacles for extracting opinions from long Chinese news sentences.

Sentence compression could be usefully employed in wide range of applications. For example, it can be used to automatically generate headline of an article [6]; Other applications include compressing text to be displayed on small screens [5] such as mobile phones or PDAs, and producing audio scanning devices for the blind [7]. However, there is no study on how sentence compression can improve the performance of opinion mining in the previous work.

Most existing studies relied on a parallel corpus to learn the correspondences between original and compressed sentences. Typically sentences are represented by features derived from parsing results, and used to learn the transformation rules or estimate the parameters in the score function of a possible compression. A variety of models have been developed, including but not limited to the noisy-channel model [11], support vector machines [13] and large-margin learning [4]. However, for opinion-oriented Chinese sentence compression, no existing parallel corpus can be directly used to training these models.

An algorithm making limited use of training data was proposed by Clarke and Lapata [3] for English text. Their model searched for the compression with highest score according to the significance of each word, the existence of Subject-Verb-Object structures and the language model probability of the resulting word combination. The weight factors to balance the three measurements were experimentally optimized by a parallel corpus or estimated by experience.

3 Proposed Approach

3.1 Problem Definition

Sentence compression is defined as follows: given a sequence of words $W=w_1w_2\ldots w_N$ (of N words) that constitute a sentence, find a subsequence $V=v_1v_2\ldots v_M$ (of M words, $M<N$), that is a compressed version of W. To take the sentences in Table 1 as an example, we have the following explanations in Figure 1.

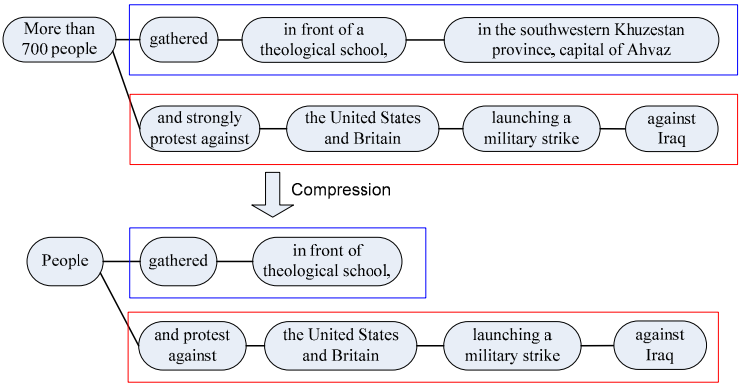

Fig. 1. The traditional sentence compression result for the origin sentence

We can see from Figure 1 that the original sentence can be divided into two self-completed sub-sentence. The conventional compression method can eliminate some modifier words and location words which are less important information for the original sentence. However, since the Chinese sentence are usually very long, not all parts of the sentence contain the opinion-related information that the readers concern. For example, the upper rectangle part of the sentence in Figure 1 describes the fact that "People gathered in front of theological school", which does not include any opinion of "People". For opinion-oriented sentence compression, we intend to get the reduced sentence as shown in Figure 2.

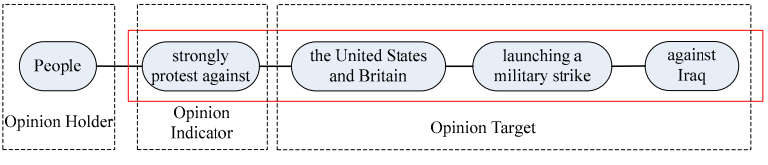

Fig. 2. The opinion-oriented sentence compression result for the origin sentence

In Figure 2, the factual part of the sentence is eliminated, and the opinionated part is retained. We can see that the opinion-oriented compression also preserves the modifier "强烈"(strongly), which is eliminated in the conventional compression method. From the compressed version of the sentence, the opinion is expressed more concise and it will be easier for people to get the opinion holder, target and other related information from the compressed sentence.

Here we give the formal definition of opinion-oriented sentence compression:

Definition 1. (Opinion-Oriented Sentence Compression). Given a sequence of words $W=w_1w_2\ldots w_N$ (of N words) that constitute a sentence, find a subsequence $V=v_1v_2\ldots v_M$ (of M words, $M<N$), that is a compressed version of W. W not only is a informative and grammatical sentence but also eliminates the factual part and preserves the opinionated part as much as possible.

In the Section 3.2, we introduce a score function for opinion-oriented Chinese sentence compression.

3.2 Score Function for Opinion-Oriented Chinese Sentence Compression

Notice that there is no existing parallel corpus for opinion-oriented Chinese sentence compression. To alleviate this problem, we employ a weakly supervised word deletion algorithm based on score function to compress a sentence.

Inspired by the work of Clark and Lapata [3], the score function in our approach is defined as a measure indicating the appropriateness of a compressed sentence. Based on the definition of opinion-oriented sentence compression, a set of words maximizing the score function is extracted from the original sentence using a dynamic programming technique. The score function is defined as the sum of word significance score I, the linguistic score L of the word string in the compressed sentence and the opinion score O of each word in the original sentence.

The score function of the sentence is given by:

$$S(V) = \sum_{i=1}^{M} \{\lambda_I I(v_i) + \lambda_L L(v_i) + \lambda_o O(v_i)\} \tag{1}$$

where $I(v_i)$, $L(v_i)$, $O(v_i)$ is the significance score, linguistic score and opinion score of the word v_i. λ_I, λ_L, λ_O are the weighting factors to balance the dynamic ranges of I, L, O, where the value can be set manually or optimized using a small amount of training data.

Word Significance Score. The word significance score I measures the relative importance of a word in the original sentence. Given a word w_i in the original sentence, the function I is defined as:

$$I(w_i) = f_i \log \frac{F_A}{F_i} \tag{2}$$

where f_i is the frequency of w_i in the document, F_i is the number of occurrences of w_i in all the documents and F_A is the sum of F_i in all the documents ($\sum_i F_i$).

In Clark and Lapata's work, the authors only focused on nouns and verbs as potential significant words. However, adjectives and adverbs are also good indicators for people's opinions. Therefore, in our study, we treat nouns, verbs, adjectives and adverbs equally for calculating the word significance score.

Linguistic Score. Using linguistic score L we can select some function words, thus ensuring that the compression results remain grammatical. We apply n-gram probability estimate the linguistic score of each word in the sentence.

$$L(w_i) = \log P(w_i \mid w_{i-2} w_{i-1}) \tag{3}$$

Opinion Score. Based on the assumption that the opinion words should have more opportunity to be preserved during the compression approach, we assign an opinion score to each word which belongs to a predefined dictionary.

$$O(w_i) = \begin{cases} \log f_i & \text{if } w_i \text{ belongs to opinion dictionary} \\ O_{const} & \text{otherwise} \end{cases} \quad (4)$$

where f_i is the document frequency of the word w_i, O_{const} is a balance weight given to all other words.

3.3 Compression Generation and Selection

Based on the score function, we employ a dynamic programming algorithm to find the best composition of M words in the sentence with N words. Given an original sentence, the dynamic programming algorithm searches different M value to maximize the score function, which generates a list of candidate compressed sentence with different length. Then the final selection of the best candidate compression is a trade-off between sentence and the value of score function. An information density score $D(V_M)$ is used to measure the quality of the candidate compression:

$$D(V_M) = \frac{S(V_M)}{M} \quad (5)$$

The best compression V is selected with the highest density score:

$$V = \arg\max_{V_M} D(V_M) \quad (\alpha N \le M \le \beta N) \quad (6)$$

where α and β are the minimum and maximum compressed ratio of the original sentence, which restrict the final length of the compressed sentence.

4 Experiments

4.1 Experiment Setup

Our intent is to check if the proposed method is effective for opinion-oriented sentence compression. Since there is no existing parallel corpus available, we build a new evaluation dataset for opinion-oriented Chinese sentence compression. The original data come from NTCIR Multilingual Opinion Mining Task, which consist of about 30,000 documents and totally 3,120,000 sentences. From these documents, we randomly pick up 100 opinionate sentences on the topic of "*Iraq War*" for the compression task. One annotator majoring in opinion mining was asked to manually delete the words of sentence, and several rules must be followed during the annotation: (a) the preserved words should not only contain the most important information, but also contain the opinionate content of the original sentence as much as possible; (b) the compressed sentence should remain grammatical. Finally, this annotation result is set to be the gold standard for the opinion-oriented Chinese sentence compression.

Recall that the score function for sentence compression has three components: the significance score, linguistic score and opinion score. The significance score was trained using the NTCIR MOAT corpus which total contain 30,000 documents. The linguistic score was calculated using a trigram language model. In this paper, we use

Google search results to estimate the linguistic score of the give words. Given the word w_i, w_{i-1}, w_{i-2}, the linguistic score is defined as $\log P(w_i|w_{i-1}w_{i-2})$. So we have:

$$L(w_i) = \log P(w_i \mid w_{i-2}w_{i-1}) = \frac{C_{Google}(w_{i-2}w_{i-1}w_i)}{C_{Google}(w_{i-2}w_{i-1})} \tag{7}$$

where $C_{\text{Google}}(w_{i-2}w_{i-1})$ and $C_{\text{Google}}(w_{i-2}w_{i-1}w_i)$ is the number of search results when launch the query words "$w_{i-2}w_{i-1}$" and "$w_{i-2}w_{i-1}w_i$" to Google search engine. Zhu et al have validated that the web search results can effectively estimate the given trigram language model [21].

The sentiment lexicon is the fundamental tools for estimating the opinion score of each word in the sentence. Our sentiment lexicon is built based on following resources: (a) The Lexicon of Chinese Positive Words and the Lexicon of Chinese Negative Words; (b) The opinion word lexicon provided by National Taiwan University (NTU); (c) Sentiment word lexicon and comment word lexicon from HowNet [9]. The lexicon is manually verified. Totally, 14,201 positive words, 17,372 negative words and 478 neutral words are obtained.

4.2 Evaluation Methods

Both automatic evaluation and human subjective evaluation are employed to measure the correctness of the compressed sentences generated by the proposed approach. The parsing based automatic evaluation methods are not appropriate for the long Chinese sentence, because errors may be brought in during the parsing approach for the original and compressed sentence.

For automatic evaluation, we measure each sentence compression method using BLEU scores [16]. BLEU scores are firstly proposed for evaluating machine translation quality and recently it has been used for measuring the quality of compressed sentences [8,19]. A BLEU score is defined as the weighted geometric average of n-gram precisions with length penalties. 4-gram precision and uniform weights are used for the BLEU scores in this paper.

For human subjective evaluation, one native Chinese speakers majoring in opinion mining was asked to rate the grammatically of compressed sentence using the 1 to 5 scale. Notice that the student has placed more emphasis on whether the opinion information is preserved in the final compressed result. At the same time, the semantic and grammatical correctness is also considered during the evaluation.

4.3 Experiment Results

Our goal of the experiments is to answer this question: whether the proposed algorithm can meet the need of the opinion-oriented sentence compression task.

We check the compression rate of the proposed method and the gold standard, as shown in Table 2. In Table 2, gold standard means the human generated compressed sentence; OOSC denotes the opinion-oriented sentence compression method; TSC denotes the conventional sentence compression method. In this paper, for conventional method, we use significance and linguistic score together and do not consider opinion score. The compression rate is the ratio of the number of Chinese characters in a compressed sentence to that in its original sentence. For OOSC method, there is

no compression rate limitation parameter except α and β, which restrict the length of the compressed sentence. In this paper, we set α=0.2 and β=0.5 for the proposed algorithm. In this way, we can limit the range of the compressed sentence range from 20% to 50% of the original sentence. From Table 2, we can see that the human generated results are about 60% size of the original sentences. The compression rate of proposed method OOSC is under fifty percent of the original sentence.

Table 2. Average compression rate

Method	Compression Rate
Gold Standard	63%
OOSC	46%
TSC	47%

To check the effectiveness of the proposed algorithm in compressing Chinese sentence, we evaluate our method using BLUE scores, as shown in Table 3. And the human subjective evaluation results in shown in Table 4. Here we set λ_I=1, λ_L=5, λ_O=2. O_{const}=0.01. We can see from Table 3 and Table 4 that the proposed OOSC method outperforms the TSC method which does not consider the opinion information.

Table 3. Automatic evaluation for the compressed results of Chinese sentence

Method	BLEU score
OOSC	0.281
TSC	0.210

Table 4. Human evaluation for the compressed results of Chinese sentence

Method	Score
OOSC	3.2
TSC	2.6

5 Conclusion and Future Work

Compared to Web comments and reviews, the Chinese news articles usually have very long sentences, which bring in new challenge for opinion mining in this area. Conventional sentence compression methods can compress a sentence without changing its meaning. However, these methods do not distinguished opinionated information from factual information. In this paper, an opinion-oriented Chinese sentence compression method is proposed. The weight of each word in original sentence is measured by its information significance, linguistic consistence and opinion score. A dynamic programming algorithm is employed to find the best combination of the words in the original sentence. Automatic evaluations and human subjective evaluations demonstrate that the proposed method is effective in finding the desired parts from the long Chinese sentences.

Future work will consider more linguistic features and constraints to improve the grammatical consistence of the compressed sentence. We also intend to further evaluate the compressed results by applying opinion holder and target extraction algorithms.

Acknowledgments. This work is partially supported by National Natural Science Foundation of China (No.60973019, 60973021), National 863 High Technology Development Program of China (2009AA01Z131, 2009AA01Z150) and HKSAR ITF (No. GHP/036/09SZ). The authors would like to thank Donghao Fang for developing parts of algorithms in this paper.

References

1. Choi, Y., Breck, E., Cardie, C.: Joint Extraction of Entities and Relations for Opinion Recognition. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 431–439 (2006)
2. Choi, Y., Cardie, C., Riloff, E., Patwardhan, S.: Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pp. 355–362 (2005)
3. Clarke, J., Lapata, M.: Models for Sentence Compression: A Comparison across Domains, Training Requirements and Evaluation Measures. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pp. 377–384 (2006)
4. Cohn, T., Lapata, M.: Large Margin Synchronous Generation and its Application to Sentence Compression. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and on Computational Natural Language Learning, pp. 73–82 (2007)
5. Corston-Oliver, S.: Text Compaction for Display on Very Small Screens. In: Proceedings of the NAACL Workshop on Automatic Summarization, pp. 89–98 (2001)
6. Dorr, B., Zajic, D., Schwartz, R.: Hedge trimmer: A Parse-and-Trim Approach to Headline Generation. In: Proceedings of HLT-NAACL Text Summarization Workshop and DUC, pp. 1–8 (2003)
7. Grefenstette, G.: Producing Intelligent Telegraphic Text Reduction to Provide an Audio Scanning Service for the Blind. In: Proceedings of the AAAI Symposium on Intelligent Text Summarization, pp. 111–117 (1998)
8. Hirao, T., Suzuki, J., Isozaki, H.: A Syntax-Free Approach to Japanese Sentence Compression. In: Proceedings of Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, pp. 826–833 (2009)
9. HowNet, `http://www.keenage.com`
10. Kim, S., Hovy, E.: Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In: Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text, pp. 1–8 (2006)
11. Knight, K., Daniel, M.: Summarization Beyond Sentence Extraction: A Probabilistic Approach to Sentence Compression. Artificial Intelligence 139(1), 91–107 (2002)
12. Martins,T., Smith, A.: Summarization with A Joint Model for Sentence Extraction and Compression. In: Proceedings of NAACL-HLT Workshop on Integer Linear Programming for NLP, pp. 1–9 (2009)

13. Nguyen, M., Akira, S., Susumu, H., Tu, B., Masaru, F.: Probabilistic Sentence Reduction Using Support Vector Machines. In: Proceedings of the 20th International Conference on Computational Linguistics, pp. 743–749 (2004)
14. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification Using Machine Learning Techniques. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 79–86 (2002)
15. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Foundations and Trends in IR 2(1-2), 131–135 (2008)
16. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: A Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistic, pp. 311–318 (2002)
17. Seki, Y., Evans, D., Ku, L., Sun, L., Chen, H., Kando, N.: Overview of Multilingual Opinion Analysis Task at NTCIR-7. In: Proceedings of the Seventh NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access (2008)
18. Turney, P.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: 40th Annual Meeting of the Association for Computational Linguistics, pp. 417–424 (2002)
19. Unno, Y., Ninomiya, T., Miyao, Y., Tsujii, J.: Trimming CFG Parse Trees for Sentence Compression Using Machine Learning Approach. In: Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics, pp. 850–857 (2006)
20. Wu, Y., Zhang, Q., Huang, X., Wu, L.: Phrase Dependency Parsing for Opinion Mining. In: Proceedings of Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, pp. 1533–1541 (2009)
21. Zhu, X., Rosenfeld, R.: Improving Trigram Language Modeling with the World Wide Web. In: Proceedings of the International Conference on Acoustics Speech and Signal Processing, pp. 533–536 (2001)