Research on Chinese Sentence Compression Automatic Evaluation Method based on Sentence Hierarchical Information Extraction

ZhongXian Wang*, WenShuang Du and JiaLu Ren
Hohai University, Jinling North Road, Changzhou City, Jiangsu Province, People's Republic of China

*wzxecho123@163.com

Abstract. People often feel incomprehensible and difficult to understand in the face of huge amounts of information. They instinctively produce the natural need to condense information to save time. Sentence compression technology has been born. The research on sentence compression evaluation methods has developed with the rise of sentence compression technology, and has gradually become a hot issue for sentence compression tasks. The method of sentence compression evaluation is to judge a good or bad cornerstone of compression results. Under the current practical background, the performance evaluation of sentence compression mainly includes two methods: manual evaluation and automatic evaluation. The main problems encountered are: First, most of the research work uses manual evaluation indicators. The compression results under manual evaluation have high accuracy and meet the requirements. People's daily language habits have the disadvantages of inflexibility and poor reusability. Second, there are few automatic evaluation methods for Chinese sentences, and the existing traditional automatic evaluation methods only consider the morphological similarity between the compressed sentence and the original sentence, and cannot well represent the compression effect. Based on the above practical problems, this paper proposes an automatic evaluation method for Chinese sentence compression based on sentence hierarchical information extraction. Compared with the manual evaluation method and the traditional automatic evaluation method, the Chinese sentence compression automatic evaluation method based on sentence hierarchical information extraction overcomes the disadvantages of high manual evaluation cost and low reusability. It uses the similarity of sentence hierarchical information to diversify. Ground characterization and evaluation of compression effects have significantly improved the rationality and accuracy of traditional automatic evaluation methods.

1. Introduction

With the rapid development of sentence compression [1], the evaluation of sentence compression has received increasing attention from researchers. A typical example is the emergence of a new evaluation-driven seminar model in this field, which has greatly promoted related academic research. The evaluation methods of sentence compression mainly include two methods: manual evaluation and automatic evaluation. The compression results under manual evaluation are highly accurate, but they are not flexible and reusable. The automatic evaluation method is flexible and reusable. The accuracy and reliability of the evaluation is worse than manual methods. According to the actual situation of the
above two methods, this paper combines the advantages and shortcomings of the two methods, and proposes an automatic evaluation method of Chinese sentence compression based on sentence hierarchical information extraction and a method with good compatibility with it. The new Chinese sentence compression algorithm has the following contributions: First, the algorithm in this paper comprehensively considers the specific application occasions of compression tasks, and provides different compression strategies according to different application scenarios. The algorithm in this paper has a certain general adaptability, which overcomes the disadvantages of the current compression algorithm deviating from reality, excessive pursuit of high compression ratio and inflexibility of use occasion to a certain extent. Second, the traditional evaluation indexes for sentence compression are mainly compression ratio, grammar, and information retention. Evaluation methods can be divided into manual evaluation and automatic evaluation. Aiming at the disadvantages of previous people's over-emphasis on the compression ratio, and without considering the different needs of the actual application scenarios for information retention, the Chinese sentence compression automatic evaluation method based on sentence hierarchical information extraction improves the traditional automatic evaluation method. From the perspective of multi-layered information extraction, the similarity evaluation of the compressed sentence and the original sentence is considered, which makes the "information amount" index more reasonable and more representative of the general situation of the compression effect.

Section 2 of this paper introduces the work related to sentence compression and sentence compression evaluation. Section 3 mainly introduces the new compression algorithm used in this paper, which considers sentiment information of the sentence, and gives a description of the important steps of the algorithm. Section 4 explains the source of experimental data, explains the experimental steps, and analyzes and evaluates the effect of the algorithm in this paper, combined with the automatic evaluation method of Chinese sentence compression based on sentence hierarchical information extraction. Section 5 gives the full text summary.

2. Related work

2.1. Statement compression
Knight and Marcu[3] (hereinafter referred to as K & M) proposed a method for sentence pruning visual compression using machine learning. The earliest noise channel model was applied to sentence compression tasks, and a decision tree-based compression was also proposed. Their research lays the foundation for many related work. MacDonald[4] uses the maximum edge learning algorithm for sentence compression. This method embodies the feature advantages. Hori and Furui[5] use an unsupervised approach to achieve the task of sentence compression. They consider the deletion of words as a global problem, and naturally use dynamic programming algorithms to gradually optimize the pruning process of sentences, but their method exists This study only considers the lexical factors and does not use the syntax information reasonably. Turner and Chamiak[6] use supervised, unsupervised, and semi-supervised methods to implement sentence compression tasks on the basis of K&M. Chen Jinguang and He Tingting[7] proposed a BI-NC model based on an improved noise channel model in their thesis "Chinese Sentence Pruning Based on Probability and Syntactic Analysis", which has better universality for sentence compression. They adopted an unsupervised method to alleviate the shortcomings of the original Chinese-compressed sentence alignment scarce corpus to a certain extent. However, this paper still does not have a good solution to the problem of non-isomorphism of tree pairs. This problem limits the size of the training corpus that is scarce. Wei Xu and Ralph Grishman[8] attempted to construct a syntax tree based on language heuristic rules, and gradually removed morphemes with low to high risk to obtain a more grammatically compressed sentence, and considered fault tolerance by improving part-of-speech tagging and syntactic analysis for better compression.
2.2. Compression Evaluation
Sentence Compression evaluation mainly has two methods: manual evaluation and automatic evaluation. In the manual evaluation, two indicators proposed by K&M, namely semantic integrity and grammatical correctness, are widely used by researchers. The manual evaluation of the compressed results is more accurate and in line with people's language habits, but the time and artificial cost are too high, the workload is large, and the subjectiveness of manual evaluation is strong, which cannot be compared with the previous research results. In view of the above shortcomings of manual evaluation, the focus of work in recent years has gradually shifted to the research of automatic evaluation methods. Automatic evaluation methods can greatly reduce costs, analyze system performance in a timely manner, and then improve the system in a targeted manner. Common automatic evaluation methods include: n-gram-based automatic evaluation methods[9] (including BLEU and NIST evaluation indicators), GTM automatic evaluation methods based on accuracy and recall, and automatic evaluation methods based on edit distance.

3. Methods

3.1. Affective morpheme capture algorithm
The sentimental morpheme capture algorithm in this article is based on the heuristic rule algorithm[10] to perform preliminary compression on the original sentence. After processing by the heuristic rule algorithm, a preliminary compressed sentence can be obtained from the original sentence, but the preliminary compressed sentence only depends on linguistics. Heuristic rules are obtained by pruning the syntactic tree of the original sentence. The model is single and rigid, lacks generality, and compression performance indicators, including compression ratio, grammatical structure, and information retention need to be improved. Based on the advantages of heuristic rule compression algorithm and its simplicity, and considering the disadvantages of compression only on morphology and syntax, this paper analyzes the hierarchical structure of Chinese sentences and proposes the emotional speech rate based on the deep structure of sentences—sentence scores. Capture algorithm, which can optimize the sentence compression task so that the compression of Chinese sentences can flexibly choose different compression strategies according to the different needs for the sentiment of sentences / text in different application scenarios, and comprehensively improve the performance indicators of compressed sentences to achieve Better compression of Chinese sentences.

Affective morpheme capture is shown in the following algorithm. First, the algorithm can support input as a sentence or text. If the input is an original sentence S, the text category to which the original sentence S belongs is determined. If the input is text, the text is classified into texts, and then the text is segmented. An emotional morpheme capture algorithm is performed on the result of the clause (single sentence S). The text classification work here can use the text classifier that comes with some toolkits, or you can choose deep learning methods such as CNN circular convolutional network for data training. The second step is to determine the emotional demand value of each type based on the existing and established text classification types. The work here is based on mathematical statistics. We perform the calculation of sentiment scores for all articles in each text classification. The sentiment score of each text category is normalized as the sentiment demand value of the category. Calculated as follows:

\[
\text{type}_\text{senti} = \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{1}{n_2} \sum_{j=1}^{n_2} \text{sentence}_\text{senti}
\]

Here \text{type}_\text{senti} is the emotional demand value of a certain type of text, n1 and n2 are the number of documents in the current text category and the number of sentences in the current text, the magnitude of the final sentiment score of \text{type}_\text{senti} and the single text in a category The sentiment score of a single sentence in the same order. And \text{sentence}_\text{senti} is the sentiment score of a single sentence in a single text in the current category, which is calculated by the sentiment score algorithm. Here is the sentiment score algorithm.
The external dictionaries that the algorithm mainly uses are emotion dictionaries, degree adverb dictionaries, and negative word dictionaries. When calculating the sentiment score of a sentence, first set the initial weight $W$ to 1, starting from the first sentiment word, use the weight $W *$ the sentiment value of the sentiment word as the score (recorded with $sentence\_senti$), and then determine whether it is the next sentiment word. If there are degree adverbs and negative words, $W * -1$, if there are degree adverbs, the degree value of $W *$ degree adverbs, and $W$ at this time is used as the weight value to traverse the next emotional word, and iterate until all the sentiment words, the sum of the scores $sentence\_senti$ during each traversal is the sentiment score of this document.

Based on the above process combined with the normalization operation of the emotional demand value of the text type, the emotional demand value of each type of text under the proposed text classification is obtained. Its magnitude is equivalent to the emotional score of a single sentence in a single text in a certain category. The magnitude is the same. The third step of the emotional morpheme capture algorithm is to divide the emotional interval according to the different objective facts of the emotional demand values of different text types. Based on the emotional interval, the thresholds of emotional words and degree adverbs in different emotional intervals are finally removed. The class final removal threshold is compared with the sentiment score of each word in the original sentence’s to-be-deleted word list deleting_wordslist, and those above the corresponding final removal threshold are retained in the final compressed sentence. The sentiment score of each word in the to-be-deleted word list deleting_wordslist is obtained from the sentiment score of the vocabulary belonging to enti_set or adv_set that is most similar to each word * the similarity weight $a$ between the two. The algorithm returns the final compressed sentence $C\_F$. Figure 1 shows steps of the affective morpheme capture algorithm.

Figure 1. Affective morpheme capture algorithm.

3.2. Algorithm evaluation

Based on the complementarity between manual evaluation and automatic evaluation, the experiments in this paper proposed a Chinese sentence compression automatic evaluation method based on sentence hierarchical information extraction. Finally, a combination of manual evaluation and automatic evaluation was used to evaluate the compression algorithm in this paper.

The index of manual evaluation and machine evaluation of sentence compression in this article is as follows: (Syntactically unified by manual evaluation)

1. Importance (information): The semantic integrity of the compressed result requires that the compressed sentence retain as much information as possible in the original sentence. This article uses the percentage system to measure the amount of information. The default amount of information in the original sentence is 1.

2. Grammaticality: The grammatical correctness of the compressed result. The grammatical evaluation only considers the grammatical structure of the sentence. It is divided into 1 to 5 points, a
total of 5 levels, with 5 being the best and 1 being the worst.

3. CompressionRate (compression ratio): The number of characters in the compressed sentence / the number of characters in the original sentence. This article uses a percentage system to measure the compression ratio.

Based on the significance of sentence compression itself, the performance of sentence compression is mainly reflected in two key indicators of information volume and grammatical structure. The evaluation of grammatical structural machines is worse than manual evaluation, so manual evaluation is generally used. And the importance can be automatically measured by the machine to simplify the use of manpower, and also achieve better results. The methods of automatic evaluation of information amount used by previous people mainly include an automatic evaluation method based on edit distance and an automatic evaluation method based on BLEU.

The commonality of the above two commonly used evaluation indicators is to convert the information evaluation to the similarity comparison between the compressed sentence and the original sentence, but this similarity comparison is only performed at the word level, without considering the syntactic level and sentiment tendency of the sentence. The similarity at the level, which makes the similarity between the compressed sentence and the original sentence greatly reduce the effect of information representation, and only considering the lexical similarity comparison method is not in line with reality. The new evaluation method proposed in this paper, the Chinese sentence compression automatic evaluation method based on sentence hierarchical information extraction, effectively improves the above disadvantages.

An automatic evaluation method for Chinese sentence compression based on sentence hierarchical information extraction considers that a Chinese sentence is composed of morphology, syntax, and emotional information, and the emotional information reflects the viewpoints, tendencies, etc. that the sentence wants to express. From the above, in this article, we believe that the theme of the sentence expression and the attitude reflected reflect the entire information of the sentence. After experimental analysis, it is found that sentence features are distributed at different levels of the sentence. The overall information of the sentence can be divided into three levels of information set: surface, middle, and deep [11]. The surface and middle-level information are lexical and syntactic information, respectively. The surface-level and middle-level information of the sentence constitute the main content information of the sentence, which is considered as topic information here. The deep information of a sentence is the sentiment information of the sentence. The evaluation index of the compression method—the importance index can be measured by the similarity between the compressed sentence and the original sentence. The similarity of this experiment uses the similarity of deep analysis combined with the sentence segmentation structure of the sentence, that is, comprehensive consideration of the lexical and syntactic structure of the compressed sentence and the original sentence. And similarity models on emotional information.

The experiment is based on the word2vec word vector model method to measure the lexical similarity information (shallow similarity) of the sentence. Word vectors have good semantic characteristics and are a common way to express word features. The value of each dimension of the word vector represents a feature with a certain semantic and grammatical interpretation. Word2vec is used to calculate the average vector of each vocabulary in each sentence. The vector of a single sentence is the sum of the vocabulary vectors. Finally, the cosine similarity between the sentence and the sentence vector is used to calculate the sentence similarity. The experiment calculates the syntactic similarity information (middle-level similarity) of the sentence by analyzing the dependency syntactic tree of the sentence and finding the syntactic dependency of the sentence.

\[
    \text{Simmid} = \text{effect}(S1, S2)/(\text{pairnum1} + \text{pairnum2})
\]  

(2)

Among them, effect (S1, S2) represents the same or similar dependency structure number of the original sentence S1 and the compressed sentence S2, and pairnum1 and pairnum2 respectively represent the number of valid morpheme collocations in S1 and S2. The shallow similarity and middle similarity of a sentence can form the topic similarity of a sentence.
\[ \text{Simtopic} = a \ast \text{Simwordvec} + b \ast \text{Simmid} \] (3)

\( a \) and \( b \) are weights, \( a \cdot b = 1.0 \), and are manually adjusted according to the degree of influence of the two similarities on the topic similarity. In the experiment, \( a = 0.65 \) and \( b = 0.35 \).

The deep structure of the sentence represents the emotional information of the sentence. However, for practical considerations, only when two sentences are similar in topic can their emotional similarity play a role in improving the deep similarity of the sentence. If the two sentences are similar in topic Passing low, no matter how close their emotional tendencies are, there is no practical significance. In the compression task, since the compressed sentence is obtained by pruning the syntactic tree of the original sentence, the compressed sentence and the original sentence have similar subject information, so it is reasonable to consider the similarity of the emotional information of the two sentences before and after.

\[ \text{Simenti} = \partial \ast \left[ 1 - \left( \frac{\text{sentscore}(a)}{\text{avgscore}} - \frac{\text{sentscore}(b)}{\text{avgscore}} \right) \right] \] (4)

The coefficient \( \partial \) here is the normalized emotional demand value mentioned above, \( \text{sentscore} \) is the sentiment score of a sentence, \( \text{avgscore} \) is the average sentiment score of a normal Chinese sentence, and 50,000 sentence examples are used in the experiment. The values of the above four expressions are all in the range \([0, 1]\). The final nonlinear weighted sentence similarity calculation method based on the sentiment similarity and topic similarity of the sentence is:

\[ \text{Simfinal} = \frac{(2 \ast \text{Simtopic} \ast \text{Simenti})}{(\text{Simtopic} + \text{Simenti})} \] (5)

4. Experimental steps and results analysis

4.1. Experimental steps

Due to the lack of research results in the direction of Chinese sentence compression algorithm and no public test set, in this experiment, a relatively authoritative corpus is selected for the data requirements of different modules. For text classification experiments, this article selects the Sogou text classification corpus to construct a RNN neural network for text classification training. The text classification corpus is derived from a large number of edited and categorized news corpora and corresponding classification information saved by the Sohu news website. Its classification system includes dozens of classification nodes, and the scale is about hundreds of thousands of documents. Its news corpus is selected from many fields such as sports, science and technology, finance and entertainment, and has certain representativeness. The external sentiment dictionary used by the sentiment morpheme capture algorithm comes from the BosonNLP sentiment dictionary. BosonNLP is an sentiment dictionary built on data sources such as Weibo, news, and forums. Therefore, it is more suitable for text types related to news forums. Figure 2 shows the partial structure of the affective dictionary.

![Figure 2: Partial structure of the emotion dictionary](image-url)
segmentation) in the above range as the experimental data set. In the experiment, eight types of classification were developed under the emotional morpheme capture algorithm, which comprehensively covered the text category information, respectively: sports, family, technology, entertainment, finance, social, education, and political affairs. The experimental data combined with the proposed classification calculated by the algorithm involved in the previous section, the corresponding sentiment score, and the normalized sentiment demand value are shown in the following table (the results retain five significant digits).

Table 1. Table of text information

| Text type | Emotional score | Normalized emotional demand |
|-----------|-----------------|-----------------------------|
| sport     | 9.9813          | 1.5569                      |
| family    | 10.6511         | 1.6601                      |
| technology| 7.5571          | 1.1779                      |
| entertainment | 7.2695      | 1.1331                      |
| society   | 7.6844          | 1.1977                      |
| economic  | 6.4158          | 1.0000                      |
| education | 6.8726          | 1.0712                      |
| political | 7.0724          | 1.1023                      |

4.2. Experimental results and analysis

This article evaluates the compression algorithm based on the information volume measurement method that is more suitable for the compression algorithm in this paper. The results are shown in the table below. (Note, the sentiment filtering algorithm here includes regular compression processing and emotional morpheme capture)

Table 2 shows the performance differences of various compression indexes in the compression of Chinese sentences when the compression algorithms and affective morpheme capture algorithms studied by artificial compression, using only heuristic rule compression, and WeiXu et al.

From the overall performance of the four algorithms, it can be seen that the compression ratios of the four compression methods are roughly between 0.54 and 0.6, and the differences are not large, indicating that the four algorithms have similar performance in compression ratio. Considering that the test data comes from the news texts of different classifications of Sina, which has the characteristics of simple language and few modified words, a high compression ratio is in line with reality. Among them, the overall advantage of manual compression over automatic compression is most significant in terms of grammaticality. This is because the sentences produced by manual compression are often free of grammatical structure errors, while the automatic compression algorithm performs the original syntactic tree based on rules Pruning, rules are not all applicable in all cases, and to some extent cannot guarantee the excellent grammatical structure of compressed sentences. In terms of information content (based on BLEU lexical similarity), the artificial compression result has an information content of 0.91384, which retains all important information in the original sentence in a relatively complete manner. The three automatic compression algorithms are inferior to the artificial compression results. WeiXu et al. Human algorithms perform the worst, and emotional morpheme capture algorithms perform better than compression algorithms that use only heuristic rules. In terms of information volume (based on sentence hierarchical information extraction), it can be seen that the artificial compression score is 0.92617, and the emotional morpheme capture algorithm has a score of 0.88642 in this item, which is very close to the artificial compression score, and only artificial compression and emotion The morpheme capture algorithm scores higher on the amount of information (based on sentence hierarchical information extraction) than the corresponding score (based on BLEU lexical similarity), which clearly indicates that the emotional morpheme capture algorithm in this paper is more artificial. The actual effect of compression takes into account the actual needs of the amount of information retention in emotional tendencies. Table 2 shows the results of the evaluation of different algorithms.
5. Conclusion
In this paper, a new method for Chinese sentence compression based on sentence sentiment information is proposed in combination with the automatic evaluation method of Chinese sentence compression based on sentence hierarchical information extraction. This algorithm can be used flexibly and has excellent universality in a variety of application scenarios.

References
[1] Jing H. Sentence reduction for automatic text summarization[C]//Proceedings of the 6th Applied Natural Language Processing Conference, Seattle, WA, USA, 2000:310-315.
[2] Jiang Xue. Research on Key Technologies of Chinese Sentence Compression [D]. Northeastern University, 2014.
[3] Knight K, Marcu D. Summarization beyond sentence extraction: a probabilistic approach to sentence compression[J]. Artificial Intelligence, 2005, 139:91-107.
[4] Mcdonald R T. Discriminative Sentence Compression with Soft Syntactic Evidence[C]// EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy. DBLP, 2006.
[5] Hori C, Furui S. Speech summarization: an approach through word extraction and a method for evaluation[J]. IEICE Transactions on Information and Systems, 2004, E87-D(1):15-25.
[6] Turner J, Charniak E. Supervised and unsupervised learning for sentence compression [C] // Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, 2005:290-297.
[7] Chen Jinguang, He Tingting, Li Fang, et al. Chinese sentence pruning based on probability and syntactic analysis [C] // Proceedings of the 5th National Youth Computing Linguistics Symposium. 2010.
[8] Xu Wei, Grishman R. A parse-and-trim approach with information significance for Chinese sentence compression[C]//Proceedings of the 2009 Workshop on Language Generation and Summarisation, 2009:48-55.
[9] Jiang Xue. Research on Key Technologies of Chinese Sentence Compression [D]. Northeastern University, 2014.
[10] Han Jing, Zhang Dongzhan, HANJing, et al. Heuristic Chinese sentence compression algorithm based on word popularity [J]. CEA, 2014, 50 (4): 132-139.
[11] Chen Xuezhi. Research on similarity of Chinese sentences based on layering [D]. Hunan Normal University, 2014.