Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6874–6883
Marseille, 20-25 June 2022

A Simple Yet Effective Corpus Construction Method for Chinese Sentence Compression

**Yang Zhao*, Hiroshi Kanayama*, Issei Yoshida*,
Masayasu Muraoka*, Akiko Aizawa†**
*IBM Research - Tokyo
19-21 Nihonbashi Hakozaki-cho, Chuo-ku, Tokyo 103-8510 Japan
yangzhao@ibm.com, {hkana,issei,mmuraoka}@jp.ibm.com
†National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 Japan
aizawa@nii.ac.jp

Abstract

Deletion-based sentence compression in the English language has made significant progress over the past few decades. However, there is a lack of large-scale and high-quality parallel corpus (i.e., (sentence, compression) pairs) for the Chinese language to train an efficient compression system. To remedy this shortcoming, we present a dependency-tree-based method to construct a Chinese corpus with 151k pairs of sentences and compression based on Chinese language-specific characteristics. Subsequently, we trained both extractive and generative neural compression models using the constructed corpus. The experimental results show that our compression model can generate high-quality compressed sentences on both automatic and human evaluation metrics compared with the baselines. The results of the faithfulness evaluation also indicated that the Chinese compression model trained on our constructed corpus can produce more faithful compressed sentences. Furthermore, a dataset with 1,000 pairs of sentences and ground truth compression was manually created for automatic evaluation, which, we believe, will benefit future research on Chinese sentence compression.

Keywords: Chinese Sentence Compression, Corpus Construction, Dependency Tree, Pre-trained Transformer

1. Introduction

Deletion-based sentence compression aims to delete words from a single sentence to produce a shorter version, that is, a compressed sentence that remains readable and faithful to the meaning of the source sentence. Despite the lack of paraphrasing or word order change, this technique has proven useful in many applications, such as compacting subtitles for high-rate speech (Vandeghinste and Pan, 2004), shortening lengthy product titles on online retail platforms (Wang et al., 2018), being used as a pipeline step in multiple document summarization (Banerjee et al., 2015), and improving neural machine translations (Li et al., 2020).

Over the past two decades, English sentence compression has made significant progress in terms of unsupervised and supervised compression systems (Filippova and Strube, 2008a; Filippova et al., 2015), corpora of various genres (Knight and Marcu, 2002; Clarke and Lapata, 2006; Filippova and Altun, 2013), and evaluation (Clarke and Lapata, 2008a; Filippova and Strube, 2008a; Filippova and Altun, 2013). In particular, Filippova and Altun (2013) utilized a dependency tree-based method to construct a large-scale English parallel corpus consisting of pairs of sentences and compressed sentences for the first time by leveraging English-specific dependency tree transformation, thus paving the way for follow-up studies to train their machine-learning models. In contrast, there has been very little progress in sentence compression in Chinese. The main reason for this is the lack of a parallel corpus of sufficient size and quality to develop and evaluate the Chinese compression systems. Also, it is not trivial to evaluate the quality of compressed Chinese sentences in terms of faithfulness. To remedy these shortcomings, we were particularly interested in the following research questions:

(1) How can we adapt the method used for creating an English parallel corpus to the Chinese language based on Chinese-specific characteristics? (2) What types of Chinese dependency tree components should be taken into consideration to generate a grammatical and informative compressed sentence? (3) How can we evaluate whether the compressed Chinese sentence is faithful (not contradictory) to the original sentence?

To answer these questions, we propose a simple yet effective data construction method by first leveraging contextual word embedding to handle paraphrasing and noun abbreviations when aligning Chinese keywords. Then, we rely on statistical induction to transform the four dependency tree components, namely, **particles**, **auxiliary**, **numeric modifiers**, and **negation words**, to make compression more grammatical and faithful. Subsequently, extractive and generative compression (i.e., sequence-to-sequence learning) models were trained using the constructed dataset. We also asked two human annotators to compress 1,000 sentences manually and use them for automatic evaluation. The experimental results demonstrate that our compression model can generate high-quality compression for both automatic and human evaluations. The code and data are presented here[1]. Our contributions are threefold.

- We proposed a data construction method to create 151k pairs of Chinese sentences and compressions, which are useful in training Chinese compression systems. Four Chinese-specific dependency tree transformations were considered, thus enabling our compression systems to produce high-quality compressed sentences.
- We annotated 1,000 gold compressions to overcome the lack of Chinese compression evaluation corpus in news domain and made them publicly available, facilitating future research on Chinese sentence compression.
- We experimented with state-of-the-art extractive and abstractive transformer-based compression models. The results show that our best Chinese compression system can generate more grammatical and faithful compressed sentences, compared to the baseline models.

2. Data Construction

To accelerate research on Chinese sentence compression, a high-quality parallel training corpus[2] and an evaluation corpus with gold compression are needed. We detail the construction of a 151k parallel corpus in Section 2.1 and illustrate the annotation of 1,000 sentences to create ground-truth compression in Section 2.2.

2.1. Training Data Construction

In this section, we address the training data sparsity issue by creating large-scale pairs of sentences and compressions to be used later for training compression models. Our data construction method is based on the method used for the English language, originally proposed by Filippova and Altun (2013). Their method consists of two major steps: 1) using a word-alignment-based method to identify content words between the first sentence S and headline H of a news article to determine keywords that should be kept in S, and 2) merging keywords with their parent or child nodes in a dependency tree to form phrases or chunks if they share a certain dependency relation, for example, *pobj*. The motivation is that some words have to be kept or deleted together, such as prepositional phrases; *in Afghanistan*, keeping either of them only will lead to an ungrammatical text span. We herein call this the dependency tree transformation (DTT). Following the same philosophy, we adapted this method to the Chinese language using the Chinese Gigaword corpus[3]. The motivation of using Chinese Gigaword corpus is that this data provides abstractive human-written news headline which we can exploit to identify key information in a sentence. However, there are two problems when attempting to align keywords between a Chinese sentence and its corresponding headline. (a) Chinese headlines contain many noun abbreviations, such as 西班牙 ('Spain') → 西 and paraphrases, such as 留守 ('Left-behind') → 剩余 (Remaining). The word-alignment-based method in (Filippova and Altun, 2013) is superficial and cannot capture the semantic similarity of noun abbreviations and paraphrases. (b) When determining which dependency components, such as *pobj*, should be attached to aligned keywords in (a) to form a grammatically-sound and faithful compression, the English dependency rules are not applicable to Chinese because of the unique Chinese language characteristics. To address issue (a), we present a simple yet effective contextual embedding-based method for aligning keywords (detailed in Section 2.1.1). To address issue (b), we propose a data-induced method to empirically determine four Chinese dependency tree components to be transformed (discussed in Section 2.1.2).

2.1.1. BERT-based Alignment

We collected 151k pairs of sentence S and headline H from the news articles in the Chinese Gigaword corpus after pre-processing, which included discarding the cases where either S or H contained a language other than Chinese, S and H were not relevant to each other, the length of H was longer than that of S, etc. We refer readers to Appendix A for more details on preprocessing.

Given each pair of S and H, we first tokenize S and H with the Jieba[4] tokenizer, one of the most widely used Chinese tokenizers, within the framework of Stanza[5] and utilize the BERT Chinese model[6] to convert S and H into embedding sequences $S = e^s_1, e^s_2, ..., e^s_n$ and $H = e^h_1, e^h_2, ..., e^h_m$, where e^s_k and e^h_k refer to the embedding obtained from the BERT model and n and m refer to the length of the sentence and headline, respectively. It is worth noting that the Chinese BERT model outputs only character embedding. For example, in the case of the word 西班牙 (Spain) , BERT model will output three embeddings for each character (西 , 班 , and 牙); thus, we average the three character embeddings to obtain one word embedding of 西班牙 .

Then, we must identify the keywords in S according to H by leveraging the embedding similarity. For this purpose, we consider word-wise cosine similarity, as shown in Fig. 2. For each word embedding e^h_k in headlines H, we match the most similar word embedding e^s_k in sentence S by selecting the index of the maximum cosine value in each **row** of the similarity matrix.

[1] https://github.com/ExperimentCode

[2] The parallel corpus herein refers to a collection of sentence pairs and compressed sentences.

[3] https://catalog.ldc.upenn.edu/LDC2007T38

[4] https://github.com/fxsjy/jieba

[5] https://stanfordnlp.github.io/stanza/

[6] https://huggingface.co/bert-base-chinese

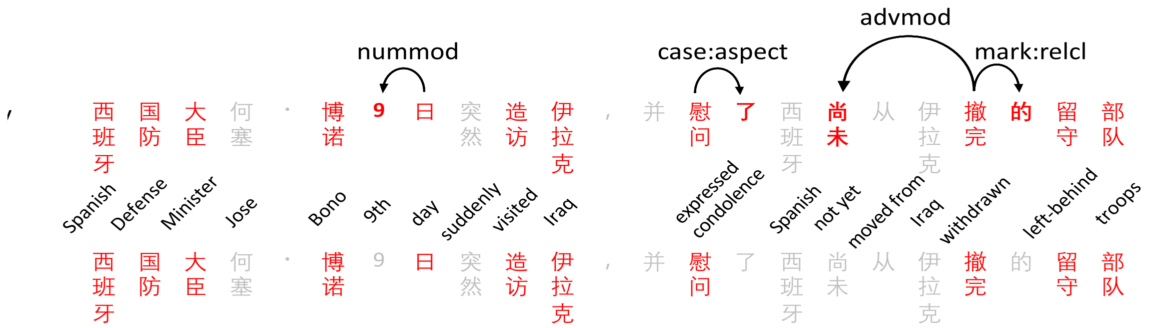


Figure 1: Chinese dependency tree transformation. Red words in the lower of the figure are identified keywords before dependency tree transformation, which forms a disfluent and semantically incorrect compression, while Red words in the upper of the figure are retained words after dependency tree transformation.

UPOS	NOUN	VERB	PROPN	PUNCT	PART	NUM	ADP	ADV	ADJ	CCONJ	AUX	PRON
%	32.85	19.20	13.54	12.71	5.82	5.42	3.62	2.84	1.40	1.05	0.66	0.51

Table 1: Statistics of the top-12 part-of-speech tags of all sentences in collected corpus.

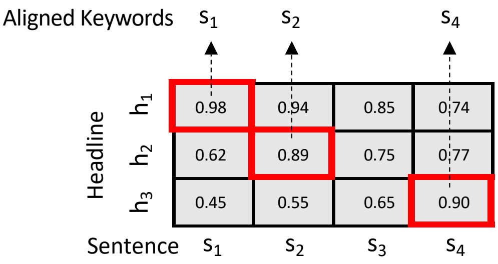


Figure 2: Token-wise cosine similarity matrix. Red boxes indicate the maximum value in each row. The corresponding words in the sentence, i.e., s_1, s_2, and s_4, are selected.

2.1.2. Dependency Tree Transformation (DTT)

After aligning the keywords, they did not necessarily form a grammatical and faithful to the original sentence. Therefore, some nodes (function words, for the most part) linked by the dependency tree should be retained to maintain a sentence that is both grammatical and accurate, as shown in the upper part of Fig. 1. In this section, we propose a data-induced method to determine the dependency relations used for merging nodes in a dependency tree. To this end, we first analyze the universal part-of-speech (UPoS) tag distribution based on all the sentences in the Chinese Gigaword corpus. As shown in Table 1, Chinese function word particles (PART) and number type (NUM) account for a significant proportion (note that NOUN, VERB, and PROPN are content words instead of function words). Furthermore, AUX is critical to grammatical functions; for example, one type of auxiliary word is negation, which flips the semantic meaning of the whole sentence if it is missed.

After determining the UPoS tag, we analyze the dependency relation distribution under each of the UPoS tags we selected, PART, AUX, and NUM, as shown in Table 2. A single word with each UPoS tag corresponds to several dependency relations. We empirically select the top-n relations until the accumulated proportion exceeds 80%. Finally, seven types of dependency relationships were selected. The linguistic rules for the Chinese dependency tree are summarized as follows:

- **Particles** in Chinese, such as 的, 了, 着, 得, 副, and so forth, are important to the verb tense and phrase structure. If one word is retained in the compression and shares one of the four dependency relations, that is, **mark:relcl**, **case:dec**, **case:aspect**, or **case:pref**, with its child nodes, the child nodes should be kept in compression.

- **Auxiliary** words in Chinese, such as 是, 为, 要, 可, 才能, and so forth, are used to achieve the grammatical functions. If one word is retained in the compression and shares either **aux** or **cop** dependency relations with its child nodes, then the child nodes should be kept in compression.

- **Numeric Modifier** in Chinese refers to words, such as digital numbers, 第一, 多个, 首批, and so forth. These words are crucial for the correctness of the specific and detailed information. An accurate compression system should accurately reflect this information. If one word is retained in the compression and shares the **nummod** dependency relation with its child nodes, the child nodes should be maintained.

- **Negation Words**, as one type of auxiliary words, are important to maintaining the semantics of the sentence and are therefore listed here separately. Omitting negation words flips the semantics of the original sentence. Unlike English, Chinese negation words have a varients of negation words, such

UPOS	Top Dependency Relation (%)	Total (%)
PART	mark:relcl(34.6), case:dec(28.7), case:aspect (10.1), case:pref (6.8)	80.2
AUX	cop(50.3), aux(49.6)	99.9
NUM	nummod (86.6)	86.6

Table 2: Percentage of dependency relation per UPOS.

as, 没, 不, 没有, 不再, and 尚未 . Therefore, we compiled a list of 70 negation words in Chinese. If one word is retained in the compression and shares either **advmod** or **aux** dependency relations with our predefined negation words, the child nodes should be maintained.

-	Sent. (1k)	Compression (1k)	
		Annotation-1	Annotation-2
ave. char. length	42.9	17.3	17.0
inter-agreement	-	0.75	

Table 3: Statistics of annotated evaluation corpus with 1,000 pairs of sentences and compressed sentence. ave. char. length refers to average number of tokens, and inter-agreement refers specifically to Cohen Kappa coefficient (Cohen, 1960).

2.2. Ground-truth Compression Annotation

To construct the gold compression data for evaluation, we asked two native Chinese speakers to manually produce 1,000 compressed sentences selected from a publicly available monolingual news corpus from the machine translation shared task[7]. With respect to the annotation guidelines, we translated the original English annotator instructions for sentence compression in Clarke and Lapata (2008b) into a Chinese version (Appendix B). Prior to annotation, the two annotators underwent a training session to ensure that they understood the compression annotation task correctly.

Table 3 shows the annotation statistics for 1,000 sentences. We follow Napoles et al. (2011a) in using character length to measure the compression rate because it is more practical than word length in real-world applications. The average length of 1,000 sentences was 42.9 Chinese characters. Annotator-1 produced compressions with an average length of 17.3 Chinese characters, whereas annotator-2 produced compressions with an average length of 17.0 Chinese characters. Cohen's unweighted κ was 0.75, indicating a substantial level of agreement[8].

3. Extractive Compression Models

Formally, deletion-based sentence compression converts a word sequence (x_1, x_2, ..., x_n) into a series of ones and zeros (l_1, l_2, ..., l_n), where x_i corresponds to the i-th word pre-tokenized by the Jieba tokenizer, n refers to the number of words in a sentence, and $l_i \in \{0, 1\}$. Here, 1 refers to maintaining x_i and 0 refers to deleting x_i. We exploit a pre-trained encoder, BERT-base-chinese[9], as our compression model to use a word sequence as the input and predict binary labels. However, a BERT tokenizer splits Chinese words x_i into characters. To maintain tokenization, we average the embeddings of all Chinese characters within one word x_i to obtain the word embedding of x_i. A Softmax layer was followed by the BERT model to make a binary prediction.

Furthermore, we are also interested in how the generative summarization model, BART, a denoising autoencoder for pretraining sequence-to-sequence model (Lewis et al., 2019) performs on our constructed datasets. This model delivers state-of-the-art performance on abstractive summarization generation tasks. Therefore, we experimented with BART-base-chinese and BART-large-chinese to investigate how different pre-trained models contribute to the compression performance.

4. Experiments and Results

To investigate the effects of the proposed BERT-based alignment (Section 2.1) and dependency tree transformation (DTT) (Section 2.2), we compared our methods with the word-alignment-based method adopted by (Filippova and Altun, 2013) for the English language. The word-alignment-based method identifies shared content words (i.e., nouns, verbs, adjectives, and adverbs) between a sentence S and headline H. Accordingly, we utilized four datasets as follows.

- *Dataset-1*: 151k pairs of (S, shared content words in S and H). Shared content words were identified using a word-alignment-based method.
- *Dataset-2*: 151k pairs of (S, shared content words in S and H with dependency tree transformation). In addition to *Dataset-1*, we added a Chinese-dependency tree transformation.
- *Dataset-3*: 151k pairs of (S, BERT-align words in S and H). The shared content words were identified using the BERT-align alignment-based method.

[7] https://www.statmt.org/wmt19/translation-task.html

[8] (Landis and Koch, 1977) characterizes κ values <0 as noagreement, 0 $\sim$ 0.20 as slight, 0.21 $\sim$ 0.40 as fair, 0.41 $\sim$ 0.60 as moderate, 0.61 $\sim$ 0.80 as substantial, and 0.81 $\sim$ 1.0 as almost perfect agreement.)

[9] https://huggingface.co/bert-base-chinese

Extractive model (*Dataset*)	F1		ROUGE (two references)			CR
	Annotation-1	Annotation-2	ROUGE-1	ROUGE-2	ROUGE-L	
Model-1 (*Word-Align*)	76.6	76.7	72.1	58.7	71.8	0.33
Model-2 (*Word-Align+DTT*)	78.2	78.0	74.7	63.4	74.4	0.36
Model-3 (*BERT-Align*)	82.6	82.7	80.0	69.7	79.6	0.34
Model-4 (*BERT-Align+DTT*)	84.1†	83.6	81.2	72.8†	80.9	0.37

Table 4: F1 and ROUGE results of four extractive compression models on the test set. We use two references, i.e., annotation-1 and annotation-2, to compute the ROUGE score. CR refers to the compression rate. Best results are in bold. † refers to results that are significantly better than other results in each column with p = 0.05.

- *Dataset-4*: 151k pairs of (S, BERT-align words in S and H with dependency tree transformation). In addition to *Dataset-3*, we added a Chinese-dependency tree transformation.

We respectively trained (fine-tuned) four BERT-based Chinese compression models using the four datasets and evaluated the results. Model-i was trained using *Dataset-i*. For model fine-tuning, the batch size was set to 200, and the Adam optimizer was used. All the models were run on two Tesla-P100 GPUs with an initial learning rate of 1e-05. For fine-tuning the BART model, the batch size was selected from [10, 32], and we used the same Adam optimizer with different initial learning rates of 2e-05. The number of epochs was set to 5.

4.1. Automatic Evaluation

To evaluate the model performance, we used our annotated evaluation corpus with 1,000 ground-truth compressions. We split the data to use 100 sentences for development and 900 sentences for testing to compute the word-level F1 score and ROUGE score.

Result of Extractive Models

The results listed in Table 4 yield the following observation: (i) The models trained with BERT-based aligned data (Models 3 and 4) significantly outperformed the models trained with word-based aligned data (Models 1 and 2) in terms of both F1 and ROUGE score; (ii) Adding the dependency tree transformation improved both F1 and ROUGE score, when comparing Models 2 and 4 with Models 1 and 3; (iii) The proposed BERT-alignment based method obtained a larger improvement than dependency tree transformation, which is expected because identifying correct keywords is more critical and a prior step to the following dependency tree transformation.

Compression Rate Control through ILP

For a fair comparison, the compression systems should be compared at similar compression ratios (Napoles et al., 2011b). Thus, we provide results across multiple compression rates using the Integer Linear Programming (ILP) framework. Similar to (Wang et al., 2017), which combines LSTMs with ILP to control the compression length, we let α_i denote the probability of the binary label $l_i = 1$, as estimated by the BERT-based compression model. The objective function of ILP with a length constraint is calculated as follows:

$$\max \sum_{i=1}^{n} l_i \alpha_i, \qquad (1)$$

$$\sum_{i=1}^{n} l_i \leq rn$$

where n refers to the number of words in a sentence and r is the desired compression rate[10]. We vary r in the range of [0.2, 0.3, 0.4, 0.5, 0.6, 0.7] to observe the model performance. Note that when we set the desired compression rate to 0.2, the actual compression rate is 0.215, which fluctuates slightly around the desired compression rate owing to the nature of the ILP framework. We compute F1 scores using the datasets annotated by annotator-1 and annotator-2. In Fig. 3, the above observations, that is, (i), (ii), and (iii), to a similar extent, hold true across multiple compression rates.

Result of Generative Model

Table 5 shows ROUGE results of the BART-Base and BART-Large model on ROUGE-1, ROUGE-2, and ROUGE-L metrics. BART-Large/Base-i trained with *Dataset-i*. Overall, our findings show that the BART-Large model performed better than the BART-Base model and that BART-Large-4 achieved the best performance. In addition, increasing the beam size resulted in a slight gain in the ROUGE score. However, as the best ROUGE results of the generative model are not comparable to those of the extractive models, we conducted a human evaluation on the extractive models.

4.2. Human Evaluation

We selected the first 100 generated compressed sentences in the test set for human evaluation to assess readability (fluency of the sentence) and informativeness (how much important information is retained). The results in Table 6 show that adding DTT causes

[10] Compression rate is defined as the character length of compression over the character length of sentence.

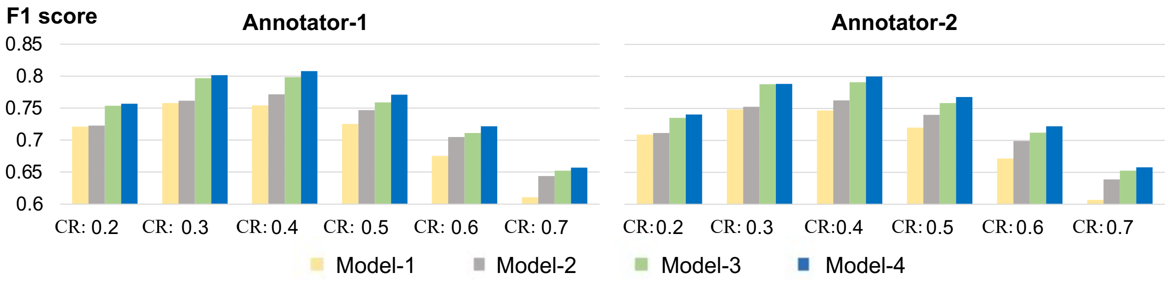


Figure 3: F1 score by varying compression rate from 0.2 to 0.7 through ILP framework.

Generative Model (*Dataset*)	greedy search			beam search (size=2)		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
BART-Base-1 (*Word-Align*)	66.5	53.8	66.1	66.4	53.9	66.0
BART-Base-2 (*Word-Align+DTT*)	68.8	57.6	68.2	68.2	57.5	67.7
BART-Base-3 (*BERT-Align*)	73.4	63.0	72.6	74.1	63.9	73.3
BART-Base-4 (*BERT-Align+DTT*)	74.4	65.3	73.5	74.8	65.9	74.0
BART-Large-1 (*Word-Align*)	66.3	53.2	65.7	66.7	53.8	65.9
BART-Large-2 (*Word-Align+DTT*)	68.6	56.0	67.3	68.6	57.4	67.7
BART-Large-3 (*BERT-Align*)	73.2	61.7	71.5	74.5	64.4	73.7
BART-Large-4 (*BERT-Align+DTT*)	75.1	65.8†	74.2†	75.9	67.0†	75.1

Table 5: ROUGE results of generative (seq2seq) compression models on the test set. CR refers to the compression rate. The greedy search equals to the beam search with size 1. The best scores are in bold. † refers to results that are significantly better than other results in each column with $p = 0.05$.

Extractive Model (*Dataset*)	Read.	Info.
Human Reference	*4.40 (±0.10)*	*3.65 (±0.16)*
Model-1 (*Word-Align*)	2.68 (±0.17)	2.25 (±0.19)
Model-2 (*Word-Align+DTT*)	3.18 (±0.22)	2.67 (±0.21)
Model-3 (*BERT-Align*)	3.77 (±0.14)	3.41 (±0.22)
Model-4 (*BERT-Align+DTT*)	4.08†(±0.15)	3.58 (±0.22)

Table 6: Human evaluation upon readability (± 1.96×SE) and informativeness (± 1.96×SE) metrics. All values in readability column are significantly different from each other at 95% confidence († refers to statistical significance), and so are values in the informativeness column except for Models 3 and 4.

Models 4 and 2 to generate more readable compressions than Models 1 and 3, suggesting that merging particles, auxiliaries, number modifiers, and negation words is important to the grammaticality of compression. There was also a significant improvement in both readability and informativeness scores for Models 3 and 4 compared to Models 1 and 2, which indicates the advantage of the BERT-based alignment method over the word-based alignment method and shows its effectiveness in handling paraphrases and abbreviations in Chinese data.

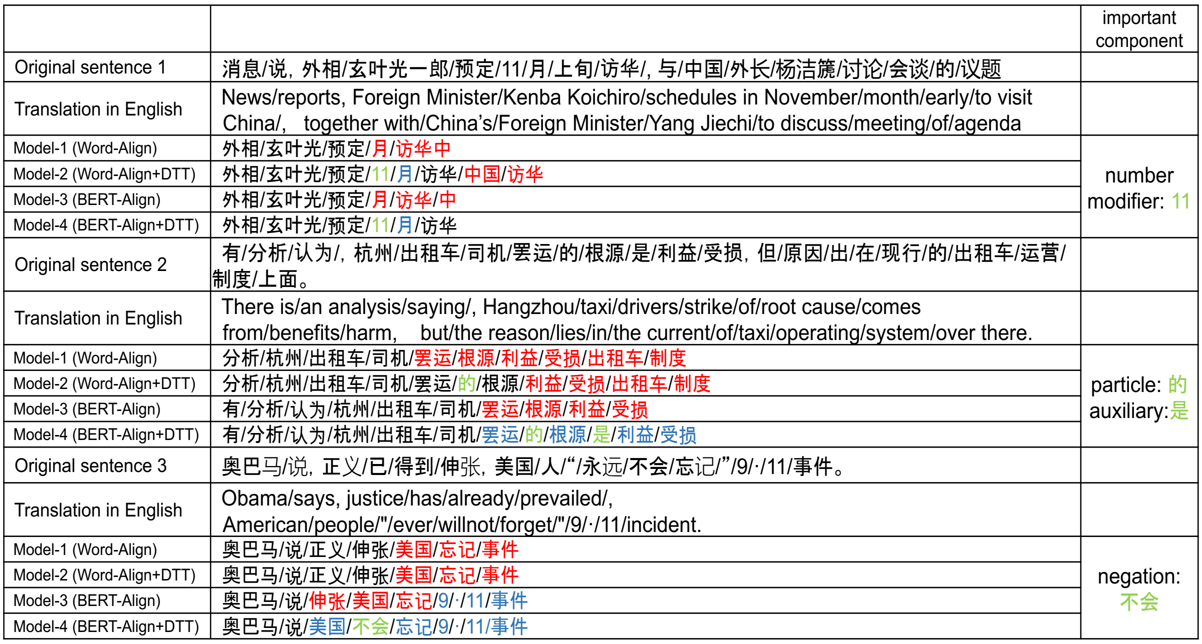
Model (*Dataset*)	entail.	contra.	neutral
Model-1 (*Word-Align*)	780	32	88
Model-2 (*Word-Align+DTT*)	794	23	83
Model-3 (*BERT-Align*)	826	11	63
Model-4 (*BERT-Align+DTT*)	835	8	57

Table 7: Number of cases where the compressed sentence is entailed (entail.), contradictory (contra.) or neutral to the original sentence.

4.3. Quantitative Faithfulness Evaluation

To quantitatively assess whether the compressed sentence is faithful to the original sentence, we utilized a Chinese entailment classifier to determine whether the compressed sentence is entailed, contradictory, or neutral to the original sentence. More specifically, we counted how many compressed sentences out of 900 sentences in the test set are NOT contradictory to the original. With respect to the Chinese entailment classifier, we followed Hu et al. (2020) and its hyperparameter setting[11] to train the Chinese RoBERTa-large-based classifier using the original Chinese natural language inference (OCNLI) dataset consisting of 56k instances. The accuracy of the three-class classification is 0.792. Then, we applied the trained classifier, as an out-of-the-box faithfulness evaluator, to the

[11] https://github.com/CLUEbenchmark/OCNLI



		important component
Original sentence 1	消息/说，外相/玄叶光一郎/预定/11/月/上旬/访华/，与/中国/外长/杨洁篪/讨论/会谈/的/议题	
Translation in English	News/reports, Foreign Minister/Kenba Koichiro/schedules in November/month/early/to visit China/, together with/China's/Foreign Minister/Yang Jiechi/to discuss/meeting/of/agenda	
Model-1 (Word-Align)	外相/玄叶光/预定/月/访华中	number modifier: 11
Model-2 (Word-Align+DTT)	外相/玄叶光/预定/11/月/访华/中国/访华	
Model-3 (BERT-Align)	外相/玄叶光/预定/月/访华/中	
Model-4 (BERT-Align+DTT)	外相/玄叶光/预定/11/月/访华	
Original sentence 2	有/分析/认为/，杭州/出租车/司机/罢运/的/根源/是/利益/受损，但/原因/出/在/现行/的/出租车/运营/制度/上面。	
Translation in English	There is/an analysis/saying/, Hangzhou/taxi/drivers/strike/of/root cause/comes from/benefits/harm, but/the reason/lies/in/the current/of/taxi/operating/system/over there.	
Model-1 (Word-Align)	分析/杭州/出租车/司机/罢运/根源/利益/受损/出租车/制度	particle: 的 auxiliary:是
Model-2 (Word-Align+DTT)	分析/杭州/出租车/司机/罢运/的/根源/利益/受损/出租车/制度	
Model-3 (BERT-Align)	有/分析/认为/杭州/出租车/司机/罢运/根源/利益/受损	
Model-4 (BERT-Align+DTT)	有/分析/认为/杭州/出租车/司机/罢运/的/根源/是/利益/受损	
Original sentence 3	奥巴马/说，正义/已/得到/伸张，美国/人/"/永远/不会/忘记/"/9/·/11/事件。	
Translation in English	Obama/says, justice/has/already/prevailed/, American/people/"/ever/willnot/forget/"/9/·/11/incident.	
Model-1 (Word-Align)	奥巴马/说/正义/伸张/美国/忘记/事件	negation: 不会
Model-2 (Word-Align+DTT)	奥巴马/说/正义/伸张/美国/忘记/事件	
Model-3 (BERT-Align)	奥巴马/说/伸张/美国/忘记/9/·/11/事件	
Model-4 (BERT-Align+DTT)	奥巴马/说/美国/不会/忘记/9/·/11/事件	

Figure 4: Three case studies. Words in red color show the grammatically problematic parts in the compression. Words in green color are important components that should be kept with words in blue color to make compressed sentence grammatical, accurate in detail, and be faithful to the original sentence.

four compression models's output. As shown in Table 7, the BERT-alignment-based methods (Models 3 and 4) generate more entailed and less contradictory compressions, indicating that integrating both the BERT-alignment-based method and DTT leads to more faithful compression.

4.4. Case study

These three cases are presented in Fig. 4. Case 1 shows that the numerical details for the month were retrained by Models 2 and 4, which is attributed to the DTT on the **numeric modifier**. Similarly, Case 2 shows that two **particle words** were kept by Model-4 to make the sentence grammatically sound. Case 3 shows that Model-4 kept the important **negation word** to make the compression faithful to the underlying meaning of the original sentence.

5. Related Work

5.1. Sentence Compression

Sentence compression research has made impressive advancements in the past two decades. In this study, we focused on deletion-based (also called extractive) sentence compression. In the early days, rule-based approaches dominated this area, and much attention was focused on leveraging synthetic trees to delete words. For example, Knight and Marcu (2000; Filippova and Strube (2008b; Berg-Kirkpatrick et al. (2011) generated compressions directly by pruning dependency or constituency trees, whereas (Jing, 2000; McDonald, 2006; Clarke and Lapata, 2006; Bingel and Søgaard, 2016) used syntactic information or syntactic features as signals to delete words. Clarke and Lapata (2006) were the first to introduce the ILP optimization framework into sentence compression research, allowing all types of constraints (e.g., length requirements) to be easily added to the objective function. With the ILP framework, Banerjee et al. (2015) defined several linguistic constraints to generate a more grammatically compressed sentence, while Wang et al. (2017) further combined it with a neural network-based approach to address the cross-domain sentence compression issue. During this period, despite the construction of several parallel corpora (e.g., the Ziff-David corpus (Knight and Marcu, 2002) and Broadcast News corpus (Clarke and Lapata, 2006)), the data size remained approximately 1k, which is too small to effectively train a machine-learning compression model.

Filippova and Altun (2013) made an important contribution with their creation of the first relatively large-scale parallel corpus consisting of more than 200k pairs of sentences and compressed sentences. Specifically, the method utilizes content words to align keywords. Many follow-up studies (e.g., (Filippova et al., 2015; Klerke et al., 2016; Wang et al., 2017; Zhao et al., 2017; Hasegawa et al., 2017; Zhao et al., 2018a; Zhao et al., 2018b; Kamigaito and Okumura, 2020)) have used this corpus or their methods to train and evaluate machine learning-based compression systems, further advancing this research field. Undoubtedly, this highlights the importance of benchmark dataset construction. In contrast to (Filippova and Altun, 2013), we replaced the word-based alignment method with a contextual embedding-based method. Despite some advanced word alignment approaches (Nagata et al., 2020; Dou and Neubig, 2021), we found that the simple BERT-based method performs reasonably well and therefore leaves other approach explorations as future work.

5.2. Chinese Sentence Compression

Despite advancements in English sentence compression research, little attention has been paid to Chinese. Zhang et al. (2012) proposed learning a subtree from the source constituency tree of a sentence to generate news titles, while Zhang et al. (2013) exploited a tree-to-tree transduction model based on tree-substitution grammars to conduct compression operation. Recently, Zi et al. (2021) applied a fully neural-network-based method to Chinese sentence compression and evaluated the results using a manually annotated corpus. However, compared to our study, none of the above studies have created a large-scale training corpus set to thoroughly investigate both extractive and generative neural approaches. Furthermore, different from Zi et al. (2021), our constructed training and evaluation corpora are in news domain, which is a primary application domain in compression and summarization research.

6. Conclusion

In this study, we proposed a data construction method to create a large-scale Chinese corpus by introducing two modifications, that is, BERT-based word alignment and dependency tree transformation, based on Chinese-specific characteristics. To investigate the effectiveness of each modification, four compression models were trained using the four constructed datasets in an ablation study. We conducted both quantitative evaluations, that is, the F1 metric, ROUGE metric, human evaluation, and faithfulness measurement, demonstrating the advantages of the proposed simple yet effective method. We believe that both the constructed 151k pairs of Chinese sentences and compressions, as well as the 1,000 annotated gold compressions, will benefit the training and evaluation of Chinese compression systems in the future.

7. Bibliographical References

Banerjee, S., Mitra, P., and Sugiyama, K. (2015). Multi-document abstractive summarization using ilp based multi-sentence compression. In *Proceedings of the 2015th International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 1208–1214.

Berg-Kirkpatrick, T., Gillick, D., and Klein, D. (2011). Jointly learning to extract and compress. In *Proceedings of the 2011th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 481–490.

Bingel, J. and Søgaard, A. (2016). Text simplification as tree labeling. In *Proceedings of the 2016th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 2, pages 337–343.

Clarke, J. and Lapata, M. (2006). Constraint-based sentence compression an integer programming approach. In *Proceedings of the 2006th COLING/ACL on Main conference poster sessions*, pages 144–151.

Clarke, J. and Lapata, M. (2008a). Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research*, 31:399–429.

Clarke, J. and Lapata, M. (2008b). Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research*, 31:399–429.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Dorr, B., Zajic, D., and Schwartz, R. (2003). Hedge trimmer: A parse-and-trim approach to headline generation. In *Proceedings of the HLT-NAACL 03 on Text summarization workshop-Volume 5*, pages 1–8.

Dou, Z.-Y. and Neubig, G. (2021). Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online, April. Association for Computational Linguistics.

Filippova, K. and Altun, Y. (2013). Overcoming the lack of parallel data in sentence compression.

Filippova, K. and Strube, M. (2008a). Dependency tree based sentence compression. *The 2008th International Natural Language Generation Conference (INLG)*, pages 25–32.

Filippova, K. and Strube, M. (2008b). Sentence fusion via dependency graph compression. In *Proceedings of the 2008th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 177–185.

Filippova, K., Alfonseca, E., Colmenares, C. A., Kaiser, L., and Vinyals, O. (2015). Sentence compression by deletion with lstms. In *Proceedings of the 2015th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 360–368.

Hasegawa, S., Kikuchi, Y., Takamura, H., and Okumura, M. (2017). Japanese sentence compression with a large training dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 281–286, Vancouver, Canada, July. Association for Computational Linguistics.

Hu, H., Richardson, K., Xu, L., Li, L., Kuebler, S., and Moss, L. (2020). Ocnli: Original chinese natural language inference. In *Findings of EMNLP*.

Jing, H. (2000). Sentence reduction for automatic text summarization. In *Proceedings of the 6th conference on Applied natural language processing*, pages 310–315.

Kamigaito, H. and Okumura, M. (2020). Syntactically look-ahead attention network for sentence compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8050–8057.

Klerke, S., Goldberg, Y., and Søgaard, A. (2016). Improving sentence compression by learning to predict gaze. *arXiv preprint arXiv:1604.03357*.

Knight, K. and Marcu, D. (2000). Statistics-based summarization-step one: Sentence compression. In *Proceedings of the 7th National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 703–710.

Knight, K. and Marcu, D. (2002). Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Journal of Artificial Intelligence*, 139(1):91–107.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, Z., Wang, R., Chen, K., Utiyama, M., Sumita, E., Zhang, Z., and Zhao, H. (2020). Explicit sentence compression for neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8311–8318.

McDonald, R. (2006). Discriminative sentence compression with soft syntactic evidence. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 297–304.

Nagata, M., Chousa, K., and Nishino, M. (2020). A supervised word alignment method based on cross-language span prediction using multilingual BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 555–565, Online, November. Association for Computational Linguistics.

Napoles, C., Callison-Burch, C., Ganitkevitch, J., and Van Durme, B. (2011a). Paraphrastic sentence compression with a character-based metric: Tightening without deletion. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 84–90.

Napoles, C., Van Durme, B., and Callison-Burch, C. (2011b). Evaluating sentence compression: Pitfalls and suggested remedies. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 91–97.

Vandeghinste, V. and Pan, Y. (2004). Sentence compression for automated subtitling: A hybrid approach. *Text Summarization Branches Out*, pages 89–95.

Wang, L., Jiang, J., Chieu, H. L., Ong, C. H., Song, D., and Liao, L. (2017). Can syntax help? improving an lstm-based sentence compression model for new domains. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 1385–1393.

Wang, J., Tian, J., Qiu, L., Li, S., Lang, J., Si, L., and Lan, M. (2018). A multi-task learning approach for improving product title compression with user search log data. *arXiv preprint arXiv:1801.01725*.

Zhang, Y., Peng, C., and Wang, H. (2012). Research on chinese sentence compression for the title generation. In *Workshop on Chinese Lexical Semantics*, pages 22–31. Springer.

Zhang, C., Hu, M., Xiao, T., Jiang, X., Shi, L., and Zhu, J. (2013). Chinese sentence compression: Corpus and evaluation. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 257–267. Springer.

Zhao, Y., Senuma, H., Shen, X., and Aizawa, A. (2017). Gated neural network for sentence compression using linguistic knowledge. In *The 2017th International Conference on Applications of Natural Language to Information Systems*, pages 480–491.

Zhao, Y., Luo, Z., and Aizawa, A. (2018a). A language model based evaluator for sentence compression. In *Proceedings of the 2018th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 2, pages 170–175.

Zhao, Y., Shen, X., Senuma, H., and Aizawa, A. (2018b). A comprehensive study: Sentence compression with linguistic knowledge-enhanced gated neural network. *Data & Knowledge Engineering*, 117:307–318.

Zi, K., Wang, S., Liu, Y., Li, J., Cao, Y., and Cao, C. (2021). Som-ncscm: An efficient neural chinese sentence compression model enhanced with self-organizing map. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 403–415.

A. Appendix: Pre-processing and Filtering of Chinese Gigaword Corpus

We describe how to yield 151,820 pairs of sentence S and headline H using the Chinese Gigaword Third Edition[12]. First, 3,113,753 pairs of first sentence S and headline H in each article, which are known to be semantically similar (Dorr et al., 2003), were extracted. We tokenized S and H using Jieba[13] and applied the Stanza[14] NLP pipeline to S, yielding a universal part-of-speech (UPOS) sequence $U(S)$ and dependency tree $T(S)$. Next, data alignment, cleansing, and filtering were conducted for the (S, H) pairs, as follows:

- Filter out the pairs where H and S are not aligned. Because H will be utilized later to determine which content in S should be maintained, we identify content words with five POS tags, NOUN, PROPN, VERB, ADJ, and ADV, in both H and S, retaining the S and H pairs that have a significant overlap rate[15] ($R \geq 0.35$) in content words.
- Filter out pairs where either S or H contains English characters, as we observed that most characters other than Chinese are English.
- Filter out the (S, H) pairs where S does not end with a full stop. The two length constraints [t?]s are also added: (a) filter out (S, H) pairs, where H is longer than S, as H should serve as a summary, while S serves as a sentence, and (b) filter out (S, H) pairs, where either H or S is more than 100 or less than five Chinese characters.

After data preprocessing, we excluded all sentences in traditional Chinese because we observed that traditional Chinese sentences were not correctly parsed by the Stanza dependency parser. Finally, we obtained 151,820 aligned Chinese sentence and headline (S, H) pairs.

B. Appendix: Annotator Chinese Sentence Compression Instructions

We herein describe the instructions used for annotating 1,000 Chinese sentences. We modified the original annotator sentence compression instructions in Clarke and Lapata (2008b) and present them in Chinese as follows:

本任务是关于句子压缩的标注任务。您将看到一些来自新闻领域的句子。您的任务是删掉一些中文词，但是不得调整词语的顺序或者增加任何词语。在压缩句子的过程中，您需要注意以下两点:(1）一个理想的压缩句子首先是语法正确和通顺的。(2）一个理想的压缩句子需要忠实于原句子的基本意思并尽可能保留最重要的信息。在保证这两点的基础上，从原句子中尽可能删掉词语。

[12] https://catalog.ldc.upenn.edu/LDC2007T38

[13] https://github.com/fxsjy/jieba

[14] https://github.com/stanfordnlp/stanza

[15] The overlap rate R is the number of overlapping content words over the number of tokens in S.