論文の内容の要旨

Abstract

論文題目：

Title of Dissertation： Linguistic Knowledge-Enhanced Neural Sentence Compression
（言語知識を活用したニューラル文圧縮）

氏　　名: 趙 陽
Name of Author： Yang Zhao

Sentence compression aims to shorten sentences to form a single compressed sentence (compression) that remains readable and preserves important information of the original sentences. With the amount of text data increasing exponentially, there is an urgent need for sentence compression systems capable of automatically and efficiently condensing the significant amount of information such as news content into a short compression. However, yielding, both grammatical and informative compression remains challenging in the area of sentence compression. Despite the progress over the past few years, three important issues have not yet or rarely been explored by previous studies. (i) how to effectively integrate different word-level linguistic features to improve the linguistic and content quality of the compression; (ii) how to define the sentence-level readability of the compression and optimize it explicitly; and (iii) how to yield readable and abstractive compression while maintaining the informative coverage.

To tackle the challenging issues mentioned above, we herein propose the linguistic knowledge-enhanced sentence compressor aiming to yield grammatical and more informative compression efficiently. Concretely, we introduce three essential ideas. The first one is designing a gating mechanism capable of selectively exploiting word-level linguistic features to make it more discriminative. The second idea takes a step forward by considering sentence-level features with a specific focus on improving the

readability of the compression. While the first two ideas focus on single sentence compression, the third idea moves to multiple sentence compression. A coarse-to-fine rewriter is introduced to polish the disfluent compression into a readable one. We will detail each part below.

Firstly, the gating mechanism (thus a gated neural network) is introduced to exploit linguistic features for single sentence compression selectively. The motivation is that in some cases, part of speech of a word determines word deletion, while in other cases, dependency relations such as nominal subject (nsubj) dominate that deletion process. The introduced gated neural network is capable of effectively determining which linguistic feature should be dominant, given different cases. Experimental results show that the proposed gating mechanism leads to better compression upon both automatic metrics and human evaluation, compared to previous competitive compression systems. Also, compression yielded by the proposed method share more grammatical relations in common with the ground-truth compression than the baseline method, indicating that important grammatical relations, such as subject or object of a sentence, are more likely to be kept in the compression by the proposed method.

Despite the improvement obtained by the above approach, which considers word-level features, readability itself is a sentence-level feature. Thus, we introduce the second idea that takes the readability of the whole compression into account via a reinforcement learning framework. A linguistic knowledge-enhanced language model is used as a grammar checker. Subsequently, a series of trial-and-error word deletion operations are conducted on the source sentence via this grammar checker in search of the best compression. The empirical study shows that the proposed model can effectively generate more readable compression, comparable or superior to several strong baselines.

In order to build a more practical sentence compressor, our third idea extends from single sentence compression to multiple sentence compression and deals with longer input text, e.t., multiple sentences. The motivation behind is that deleting unnecessary words in each sentence or some sentences and then concatenating the rest will result in incoherent and disfluent compression. To cope with this problem, we introduce a coarse-to-fine rewriter for multiple sentence compression. Firstly, several sentences were converted into a word graph whose output is coarse-grained compression. Then, the back-translation technique is used to polish coarse-grained compression into a more fluent one. Experimental results show that the proposed method yields more readable and context-aware compression meanwhile introducing a considerable amount of novel n-grams.

To conclude, this thesis proposes incorporating linguistic knowledge into sentence compression via deep neural networks. The proposed approaches deliver state-of-the-art performance on several benchmarks. The resulting sentence compression model will benefit many applications such as displaying compressed text content in a screen with limited size, e.g., mobile phone, shortening a lengthy product title for the E-commerce platform and compressing financial news clusters into a summary. We believe that this thesis takes a further step towards the challenging issues, e.t., readability, and informativeness in this area, yielding a more robust and accurate sentence compressor for the real-world applications.