Improving Pointer-Generator Network with Keywords Information for Chinese Abstractive Summarization

Xiaoping Jiang, Po Hu(✉), Liwei Hou, and Xia Wang

School of Computer Science, Central China Normal University,
Wuhan 430079, China
{jiangxp,houliwei}@mails.ccnu.edu.cn,
phu@mail.ccnu.edu.cn, wangx_cathy@163.com

Abstract. Recently sequence-to-sequence (Seq2Seq) model and its variants are widely used in multiple summarization tasks e.g., sentence compression, headline generation, single document summarization, and have achieved significant performance. However, most of the existing models for abstractive summarization suffer from some undesirable shortcomings such as generating inaccurate contents or insufficient summary. To alleviate the problem, we propose a novel approach to improve the summary's informativeness by explicitly incorporating topical keywords information from the original document into a pointer-generator network via a new attention mechanism so that a topic-oriented summary can be generated in a context-aware manner with guidance. Preliminary experimental results on the NLPCC 2018 Chinese document summarization benchmark dataset have demonstrated the effectiveness and superiority of our approach. We have achieved significant performance close to that of the best performing system in all the participating systems.

Keywords: Abstractive summarization · Sequence to sequence model
Pointer-generator network · Topical keywords · Attention mechanism

1 Introduction

Automatic summarization aims to simplify the long text into a concise and fluent version while conveying the most important information. It can be roughly divided into two types: extraction and abstraction. Extractive methods usually extract important sentences from the original document to generate the summary. However, abstractive methods often need to understand the main content of the original document first and then reorganize even generate the new summary content with natural language generation (NLG). Compared with extractive methods, abstractive methods are more difficult but are closer to human summarization manner.

Recently, the development of deep neural network makes abstractive summarization viable among which attention-based sequence-to-sequence (Seq2Seq) models have increasingly became the benchmark model for abstractive summarization task (Hou et al. 2018a). Seq2Seq models have encoder-decoder architecture with recurrent neural

M. Zhang et al. (Eds.): NLPCC 2018, LNAI 11108, pp. 464–474, 2018.
https://doi.org/10.1007/978-3-319-99495-6_39

network (RNN) configuration. The attention mechanism is recently added to generate more focused summary by referencing salient original context when decoding.

However, the existing models usually face two shortcomings: one is the content inaccuracy and repetition mainly caused by out-of-vocabulary (OOV) words (See et al. 2017), and another is that the existing attention mechanism does not consider topic information of the original document explicitly which may lead to insufficient decoding.

In this work, we propose a novel approach to improve the summary's informativeness by explicitly incorporating topical keywords information from the original document into a pointer-generator network via a new attention mechanism so that a topic-oriented summary can be generated in a context-aware manner with guidance. Specifically, we first adopt a pointer-generator network proposed by See et al. (2017) to improve accuracy of the generated content and alleviate the problem of OOV words. Meanwhile, the coverage mechanism is used to solve content duplication problem. Second, we put topical keywords extracted from the original document into the attention mechanism and incorporate it into the pointer-generator network so that decoder will pay more attention to topic information to better guide the generation of informative summary. In general, our contributions are as follows:

- We adopt a pointer-generator network model to alleviate the problems of inaccurate detail description caused by OOV words (Sect. 3.2).
- We encode topical keywords extracted from the original document into the attention mechanism and incorporate it into the pointer-generator network to enhance the generated summary's topic coverage (Sect. 3.3).
- We applied our proposed method to the Chinese document summarization benchmark dataset provided by NLPCC 2018 shared task3 and ranked the third among all the participating systems (Sect. 4).

2 Related Work

Automatic summarization has always been a classic and hot topic in the field of natural language processing (NLP). Significant progress has been made recently from traditional extractive summarization to more abstractive summarization (Yao et al. 2017).

Earlier research in the last decade is dominated by extractive methods. Extractive methods first score each sentence in the original document. Unsupervised sentence scoring approaches mostly rely on frequency, centrality and probabilistic topic models. Sentence classification, sentence regression and sequence labeling are the supervised approaches commonly used to evaluate the importance of sentences. Having predicted sentences importance score, the next step is to select sentences according to their information richness, redundancy and some constraint rules (such as total summary length, etc.). The popular approaches for sentence selection include maximum marginal relevance (MMR), integer linear programming (ILP) and submodular function maximization. Recently, it has been shown effective to use neural network to directly predict the relative importance of a sentence given a set of selected sentences, under the consideration of importance and redundancy simultaneously (Cao et al. 2017; Narayan et al. 2018).

Although extractive methods have the advantage of preserve the original information more complete, especially ensuring the fluency of each sentence, one of the problems is that they suffer from the secondary or redundant information. More importantly, there is often a lack of coherence between adjacent sentences in an extractive summary. With the rapid development of deep learning technology in recent years, abstractive summarization has gradually become the current research focus.

RNN-based encoder-decoder structure is proposed by Bahdanau et al. (2014) and used in machine translation successfully. Subsequently, this structure has also been successfully applied to other fields of NLP, including but not limited to syntactic parsing (Vinyals and Le 2015), text summarization (Rush et al. 2015) and dialogue systems (Serban et al. 2016). Rush et al. (2015) first introduce the encoder-decoder structure and the attention mechanism into summarization task and achieve good results on DUC-2004 and Gigawords datasets. Later Nallapati et al. (2016) extend their work and combine additional features to the model, which get better results than Rush on DUC-2004 and Gigawords datasets. The graph-based attention mechanism is proposed by Tan and Wan (2017) to improve the adaptability of the model to sentence saliency. Paulus et al. (2017) combine supervised learning with reinforcement learning while training, and their work not only keeps the readability but also ensures the flexibility of summary. For latent structure modeling, Li et al. (2017) add historical dependencies on the latent variables of Variational Autoencoder (VAEs) and propose a deep recurrent generative decoder (DRGD) to distill the complex latent structures implied in the target summaries. Nallapati et al. (2016); Gu et al. (2016); Zeng et al. (2016), See et al. (2017) and Paulus et al. (2017) use copy mechanism to solve the problem of OOV words in the decoding phase. Moreover, See et al. (2017) propose a coverage mechanism to alleviate words repetition. In a recent work, Wang et al. (2018) incorporate topic information into the convolutional sequence-to-sequence (ConvS2S) model.

3 Model

3.1 Attention-Based Seq2Seq Model

Attention-based Seq2Seq model is first used for machine translation tasks. It is also used to generate abstractive summary due to the resemblance between abstractive summarization and machine translation. Attention-based Seq2Seq model mainly consists of three parts: encoder, decoder and the attention mechanism connecting them.

In the encoding phase, the word embedding sequences of the original document are fed into a single bidirectional LSTM to get the encoder hidden states sequence $h = \{h_1, h_2, \ldots, h_n\}$. At each decoding time step, a single unidirectional LSTM reads the previous word embedding to obtain the decoder hidden state s_t, which is used for the output prediction of the current time step. The hidden states of encoder and decoder pass through a linear layer and a softmax function to get the attention distribution a^t. The attention distribution corresponds to a probability distribution of each word in the original document, that tells which words are more important in the current prediction process. It is calculated as follows:

$$e_i^t = v^T \tan h(W_h h_i + W_s s_t + b_{attn}) \tag{1}$$

$$a^t = softmax(\mathrm{e}^t) \tag{2}$$

Where v^T, W_h, W_s and b_{attn} are learnable parameters. Once a^t has been computed, it is used to produce a weighted sum of the encoder hidden states, which is a dynamic representation of the original document called the context vector h_t^*:

$$h_t^* = \sum\nolimits_i a_i^t h_i \tag{3}$$

Finally, the decoder hidden state s_t and the context vector h_t^* pass through two linear layer and a softmax function to produce the vocabulary distribution P_{vocab} of the current time step. The concrete formulas are as follows:

$$P_{vocab} = softmax\left(V'\left(V\left[s_t, h_t^*\right] + b\right) + b'\right) \tag{4}$$

$$P(w) = P_{vocab}(w) \tag{5}$$

Where V', V, b and b' are learnable parameters. $P(w)$ represents the probability of the current prediction for word w. Loss function of the model uses negative log likelihood:

$$\text{loss} = \frac{1}{T}\sum\nolimits_{t=0}^{T} -log\left(P\left(w_t^*\right)\right) \tag{6}$$

Where w_t^* is the target word for the current time step, and T is the total length of the target summary.

3.2 Pointer-Generator Network

The pointer-generator network proposed by See et al. (2017) is a hybrid model combining both an attention-based Seq2Seq model and a pointer network. It allows the model to generate new words from a fixed vocabulary or copy words from the original document. Therefore, for each original document, it will add the words in it to the fixed vocabulary and get the extended vocabulary. Furthermore, they also adopt the coverage mechanism to solve repetition problem.

The context vector h_t^*, the decoder hidden state s_t and the decoder input x_t pass through a linear layer and a sigmoid function to produce the generation probability P_{gen}, which indicates the probability to generate a new word from the fixed vocabulary:

$$P_{gen} = \sigma\left(w_{h^*}^T h_t^* + w_s^T s_t + w_x^T x_t + b_{ptr}\right) \tag{7}$$

Where w_{h^*}, w_s, w_x and b_{ptr} are learnable parameters. P_{gen} is used to produce a weighted sum of the vocabulary distribution P_{vocab} (referring to formula 4) and the attention distribution a^t (referring to formula 2):

$$\mathrm{P}(w) = p_{gen} P_{vocab}(w) + \left(1 - p_{gen}\right) \sum\nolimits_{i:w_i} a_i^t \tag{8}$$

The coverage mechanism is introduced into the pointer-generator network to alleviate repetition problem. It maintains a coverage vector c^t (i.e., the sum of attention distributions), which records the coverage degree of those words which have received from the attention mechanism so far. The coverage vector in turn affects the attention distribution for the current time step. The coverage mechanism also calculates additional coverage loss to punish repeated attention. The whole computation process is implemented as follows:

$$c^t = \sum\nolimits_{t'=0}^{t-1} a^{t'} \tag{9}$$

$$e_i^t = v^T tanh\left(W_h h_i + W_s s_t + W_c c_i^t + b_{attn}\right) \tag{10}$$

$$covloss_t = \sum\nolimits_i min\left(a_i^t, c_i^t\right) \tag{11}$$

$$\mathrm{loss} = \frac{1}{T} \sum\nolimits_{t=0}^{T} \left[-logP\left(w_t^*\right) + \lambda \sum\nolimits_i covloss_t\right] \tag{12}$$

Where W_c is a new learnable parameter and λ represent the weight of the coverage loss.

3.3 Keywords Attention Mechanism

Recent studies show that the traditional attention mechanism only considers the relationship between the current target word and the original document, so it fails to grasp the main gist of the original document and leads to insufficient information (Lin et al. 2018). Imagine that when people write a summary, they usually have a clear understanding of the topic content, and the summary they write is often centered around topic. Therefore, we propose to adopt topical keywords as the guidance information for the original document and encode them into the attention mechanism to generate summary with better topic coverage.

In this work, TextRank (Mihalcea 2004) algorithm is used to extract topical keywords. We extract d topical keywords from the original documents and get their word embeddings $k = \{k_1, k_2, \ldots, k_d\}$. As shown in Fig. 1, we calculate the sum of word embeddings for d keywords and use it as a part of input for the attention distribution.

$$\mathrm{t} = \sum\nolimits_{i=1}^{d} k_i \tag{13}$$

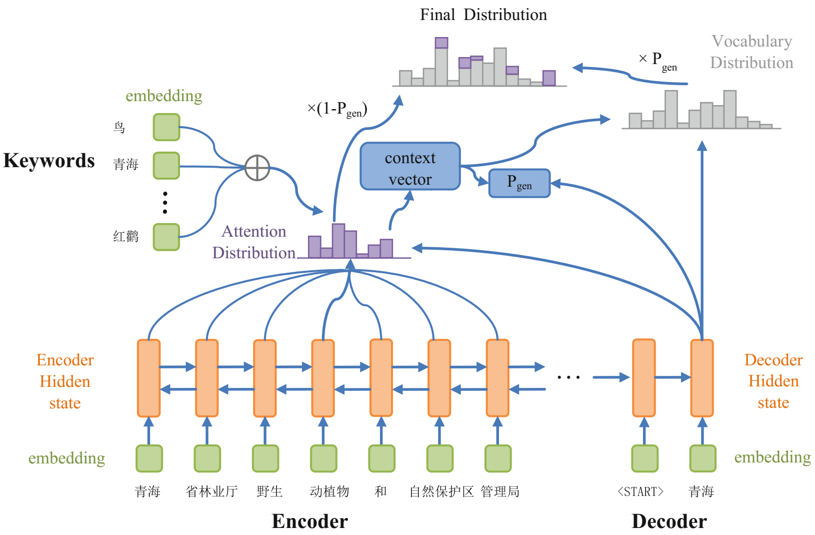

Fig. 1. Pointer-generator model with expanded keywords attention

$$e_i^t = v^T tanh\left(W_h h_i + W_s s_t + W_c c_i^t + W_t t + b_{attn}\right) \tag{14}$$

Where W_t is a new learnable parameter. We change Eqs. (10) to (14).

4 Experiments

4.1 Dataset

We conduct experiments on the open-available dataset provided by NLPCC 2018 shared task3, which is a Chinese document summarization task. The training set contains 50000 document-summary pairs, the validation set contains 2000 document-summary pairs, and the testing set contains 2000 documents without corresponding golden-standard summaries. We compute the length (i.e., the number of words after segmentation using jieba[1] toolkit) of the original documents and the standard summaries whose statistics are shown in Table 1.

4.2 Evaluation

In this work, we use ROUGE[2] toolkit for evaluation. ROUGE is widely used in the summarization research community for content evaluation, which often calculates the recall rate of N-grams or words between the machine-generated summary and the

[1] https://pypi.python.org/pypi/jieba/

[2] https://github.com/andersjo/pyrouge/tree/master/tools/ROUGE-1.5.5

Table 1. The statistics of the NLPCC 2018 document summarization benchmark dataset.

Length	Training set		Validation set		Testing set	
	Document	Summary	Document	Summary	Document	Summary
Min	32	6	35	8	7	–
Max	13186	85	8871	44	7596	–
Average	579.4	25.9	579.9	25.8	426.2	–

golden-standard human summary (Chin 2004). Character-based ROUGE-F score is used as evaluation metric in this work. We conduct evaluation on both the validation set and the testing set due to the lack of golden-standard summaries on the testing set. For the validation set, we calculate and report ROUGE-1, ROUGE-2 and ROUGE-L scores respectively. For the testing set, we present results provided by the NLPCC 2018 official organizer and the results represent the average score of multiple ROUGE metrics including ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-L, ROUGE-W-1.2 and ROUGE-SU4.

4.3 Implementation

In our experiments, we first convert all the datasets into plain texts and use jieba toolkit to conduct word segmentation on both news articles and corresponding summaries. Then, we use TextRank4ZH toolkit[3] to extract keywords from each news articles with the number of keywords set to 8. Furthermore, we use a vocabulary of 50 k words for both original documents and target summaries.

Unlike previous work proposed by Hou et al. (2018b), we do not pretrain the word embeddings in advance and all are learned from scratch during training, and the dimension of the word embeddings is set to 128. In our proposed approach, the encoder is a bidirectional LSTM and the decoder is a unidirectional LSTM with both the hidden layer dimension set to 256. And we set the maximum encoding length (i.e., the maximum length of the input sequence) to 1000, and the decoding length (i.e., the output sequence length) is adjusted from 8 to 40. Our model is trained by Adagrad (Kingma et al. 2014) with learning rate of 0.15 and initial accumulator value of 0.1. We implement all our experiments with Tensorflow on an NVIDIA TITAN XP GPU and the batch size is set to 16.

During the training phase, the initial loss value is 10, and then it drops to 7 after the first 300 times of training, and further drops to 3.5 after 15000 times of training. Finally, with the increase of training time, the loss value gradually becomes stable and converges to 3. When the loss value of the model is stable on the training set, the parameters learned from training phase are used for validation. During the testing phase, we use beam search to get the target summary and set beam size to 6.

[3] https://github.com/letiantian/TextRank4ZH

4.4 Experimental Results and Analysis

The experimental results are shown in Tables 2 and 3. We choose the pointer-generator network model as the baseline, and we also compare with other representative state-of-the-art extractive and abstractive summarization approaches.

Table 2. Comparison results on the NLPCC 2018 validation dataset using F-measure of ROUGE

Models	R-1	R-2	R-L
TextRank	35.78	23.58	29.60
Pointer-Generator	43.44	29.46	37.66
Pointer-Generator (character)	38.01	24.43	32.32
TextRank + Pointer-Generator	42.79	28.93	37.33
Our Model	**44.60**	**30.46**	**38.83**

Table 3. Examples of the generated summaries set of our approach and the pointer-generator baseline. Here **bold** denotes richer information generated by our model.

Document: 显示图片切尔西官方宣布费利佩离队北京时7月28日，切尔西官方宣布边卫费利佩-路易斯离队 ，巴西人将重返西甲劲旅马德里竞技，马竞官方随后也确认了这一消息。据悉，费利佩的转会费为1500万镑。新浪体育稍后为您带来详细报道

Golden summary: 切尔西官方宣布边卫费利佩 • 路易斯离队 ，巴西人将1500万镑重返 西甲劲旅马德里竞技

Pointer-generator：显示，切尔西官方宣布边卫费利佩-路易斯离队，巴西人将重返西甲劲旅马德里竞技

Our model: 切尔西官方宣布费利佩离队，巴西人将重返西甲劲旅马德里竞技，**转会费为1500万镑**

Document: 据美国媒体报道，美国总统奥巴马7月7日在白宫椭圆形办公室会见了越南共产党总书记阮富仲。白宫发布“美国—越南联合愿景声明”声明说, 两国共同关注南中国海最近提升的紧张局势。声明表示, 越南共产党总书记阮富仲对美国进行了历史性的访问, 两国于2015 年7月1日制定了这份联合愿景声明 ……

Golden summary: 奥马巴会见越共总书记阮富仲，发布美越联合愿景声明，声称共同关注南海紧张局势

Pointer-generator: 两国再次强调将继续“美越防务关系联合愿景声明”中所体现的防务和安全双边合作。

Our model: 奥巴马会见越南共产党总书记阮富仲，发布越南联合愿景声明，**两国共同关注南中国海最近提升的紧张局势。**

TextRank. This approach adopts the open-source TextRank4zh toolkit to extract the most important sentences with highest informatives scores from the original news article to generate the target summary.

Pointer-Generator. This model has been described in Sect. 3.2.

TextRank + Pointer-Generator. Inspired by the work of Tan et al. (2017), we combine the TextRank approach with the pointer-generator model to generate the summary. First, we obtain an 800 words summary extracted by TextRank (Mihalcea 2004). Then, the summary is used as the input of the pointer generator network to generate the final summary.

Pointer-Generator (character). Basically, there are two typical approaches to pre-process Chinese document: character-based and word-based (Wang et al. 2018). In this work, we adopt the word-based approach as we believe that words are more relevant to latent topic of document than characters. Since the official evaluation metrics are Character-based ROUGE F score, we also evaluate pointer-generator network using character-based approach to obtain a comprehensive comparison.

Our Model. The hybrid method combining keywords attention and pointer-generator network introduced in Sect. 3.3.

According to the results shown in Table 2, we can find that our proposed approach outperforms all other methods on ROUGE-1, ROUGE-2 and ROUGE-L. The keywords attention-based pointer-generator network model exceeds the basic pointer-generator network significantly. In addition, word-based approach achieves higher ROUGE performance than character-based approach, and abstractive methods always achieve higher ROUGE performance than TextRank. Furthermore, the method directly Combining TextRank with pointer-generator does not achieve obviously better results.

Table 3 gives two running cases from which it can be observed that our keywords attention-based pointer-generator network model produces more coherent, diverse, and informative summary than the basic pointer-generator network approach.

The testing results are shown in Table 4. The scores are the average of ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-L, ROUGE-SU4 and ROUGE-W1.2 provided by official organizer. Our team ranked the third among all the participating teams, which is 0.11 points slightly below than the first.

Table 4. Official evaluation results for the formal runs of top-10 participating teams

Team	The average score of ROUGE-1, 2, 3, 4, L, SU4, W-1.2
WILWAL	0.29380
Summary++	0.28533
CCNU_NLP (Our Team)	0.28279
Freefolk	0.28149
Kakami	0.27832
Casia-S	0.27395
Felicity_Dream_Team	0.27211
dont lie	0.27078
CQUT_301_1	0.25998
lll_go	0.25611

5 Conclusion

In this work, we propose a novel abstractive summarization approach which incorporates topical keywords information from the original document into a pointer-generator network via a new attention mechanism. The experimental results show that our proposed approach can reduce wording inaccuracy while improve the summary's informativeness. In the future, we will conduct more experiments on other larger-scale datasets like CNN/Daily Mail to verify the effectiveness of our method. Besides, we will also try more advanced keyword extraction algorithm to discover and embed topical keywords information more effectively.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 61402191), the Self-determined Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE (No. CCNU18TS044), and the Thirteen Five-year Research Planning Project of National Language Committee (No. WT135-11).

References

Bahdanau, D., Cho, K., Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate. arXiv preprint arXiv:1409.0473 (2014)

Serban, I.V., Sordoni, A., Bengio, Y., Courville, A.C., Pineau, J.: Building end-to-end dialogue systems using generative hierarchical neural network models. In: AAAI (2016)

Vinyals, O., Le, Q.: A Neural Conversational Model. arXiv preprint arXiv:1506.05869 (2015)

See, A., Liu, P.J., Manning, C.D.: Get to the Point: Summarization with Pointer-Generator Networks. ACL (2017)

Yao, J.-G., Wan, X., Xiao, J.: Recent advances in document summarization. Knowl. Inf. Syst. **53**, 297–336 (2017)

Rush, A.M., Chopra, S., Weston, J.: A Neural Attention Model for Abstractive Sentence Summarization. arXiv preprint arXiv:1509.00685 (2015)

Nallapati, R., Xiang, B., Zhou, B.: Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. arXiv preprint arXiv:1602.06023 (2016)

Tan, J., Wan, X.: Abstractive Document Summarization with a Graph-Based Attentional Neural Model. ACL (2017)

Paulus, R., Xiong, C., Socher, R.: A Deep Reinforced Model for Abstractive Summarization. arXiv preprint arXiv:1705.04304 (2017)

Li, P., Lam, W., Bing, L., Wang, Z.: Deep recurrent generative decoder for abstractive text summarization. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP (2017)

Gu, J., Lu, Z., Li, H., Li, V.O.K.: Incorporating Copying Mechanism in Sequence-to-Sequence Learning. ACL (2016)

Zeng, W., Luo, W., Fidler, S., Urtasun, R.: Efficient Summarization with Read-Again and Copy Mechanism. arXiv preprint arXiv:1611.03382 (2016)

Wang, L., Yao, J., Tao, Y., Zhong, L., Liu, W., Du, Q.: A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization. In: IJCAI-ECAI (2018)

Hou, L., Hu, P., Bei, C.: Abstractive document summarization via neural model with joint attention. In: Huang, X., Jiang, J., Zhao, D., Feng, Y., Hong, Yu. (eds.) NLPCC 2017. LNCS (NLAI), vol. 10619, pp. 329–338. Springer, Cham (2018a). https://doi.org/10.1007/978-3-319-73618-1_28

Cao, Z., Wei, F., Li, W., Li, S.: Faithful to the Original: Fact Aware Neural Abstractive Summarization arXiv:1711.04434 (2017)

Narayan, S., Cohen, S.B., Lapata, M.: Ranking Sentences for Extractive Summarization with Reinforcement Learning. arXiv preprint arXiv:1802.08636 (2018)

Mihalcea, R.: TextRank: bringing order into texts. In: EMNLP (2004)

Chin, Y.L.: Rouge: A Package for Automatic Evaluation of Summaries. ACL (2004)

Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. Computer Science (2014)

Tan, J., Wan, X., Xiao, J.: From neural sentence summarization to headline generation: a coarse-to-fine approach. In: IJCAI (2017)

Hou, L.-W., Hu, P., Cao, W.-L.: Chinese abstractive summarization with topical keywords fusion. Acta Automatica Sinica (2018b)

Lin, J., Sun, X., Ma1, S., Su, Q.: Global Encoding for Abstractive Summarization. ACL (2018)