Self-Attention Guided Copy Mechanism for Abstractive Summarization

Song Xu*, **Haoran Li***, **Peng Yuan, Youzheng Wu, Xiaodong He and Bowen Zhou**
JD AI Research
{xusong28,lihaoran24,yuanpeng29,wuyouzheng1}@jd.com
{xiaodong.he,bowen.zhou}@jd.com

Abstract

Copy module has been widely equipped in the recent abstractive summarization models, which facilitates the decoder to extract words from the source into the summary. Generally, the encoder-decoder attention is served as the copy distribution, while how to guarantee that important words in the source are copied remains a challenge. In this work, we propose a Transformer-based model to enhance the copy mechanism. Specifically, we identify the importance of each source word based on the degree centrality with a directed graph built by the self-attention layer in the Transformer. We use the centrality of each source word to guide the copy process explicitly. Experimental results show that the self-attention graph provides useful guidance for the copy distribution. Our proposed models significantly outperform the baseline methods on the CNN/Daily Mail dataset and the Gigaword dataset.

1 Introduction

The explosion of information has expedited the rapid development of text summarization technology, which can help us to grasp the key points from miscellaneous information quickly. There are broadly two types of summarization methods: extractive and abstractive. Extractive approaches select the original text segments in the input to form a summary, while abstractive approaches "create" novel sentences based on natural language generation techniques.

In the past few years, recurrent neural networks (RNNs) based architectures (Chopra et al., 2016; Gu et al., 2016; Nallapati et al., 2016, 2017; See et al., 2017; Zhou et al., 2017; Li et al., 2018b,a; Zhu et al., 2019) have obtained state-of-the-art results for text summarization. Benefit from long-term dependency and high scalability, transformer-based networks have shown superiority over RNNs on many NLP tasks, including machine translation (Vaswani et al., 2017; Dehghani et al., 2019), sentence classification (Devlin et al., 2019; Cohan et al., 2019), and text summarization (Song et al., 2019; Zhang et al., 2019).

One of the most successful frameworks for the summarization task is Pointer-Generator Network (See et al., 2017) that combines extractive and abstractive techniques with a pointer (Vinyals et al., 2015) enabling the model to copy words from the source text directly. Although, copy mechanism has been widely used in summarization task, how to guarantee that important tokens in the source are copied remains a challenge. In our experiments, we find that the transformer-based summarization model with the copy mechanism may miss some important words. As shown in Table 1, words like "nominees" and "obama" are ignored by the standard copy mechanism. To tackle this problem, we intend to get some clues about the importance of words from the self-attention graph.

Source: two u.s. senators are blocking 11 of president barack obama 's nominees for senior administration posts at the pentagon and justice department in protest over a proposal to house guantanamo detainees at the fort leavenworth prison in their midwestern home state of kansas
Reference: us senators bar obama nominees protest guantanamo
Transformer: 1 us senators block pentago justice nominees
Transformer + Copy: us senators block 11 from pentago justice posts
Transformer + Guided Copy: us senators block obama nominees over guantanamo
Top Words from Self-attention: nominees, obama, senators, pentagon, guantanamo

Table 1: Yellow shades represent overlap with reference. The above summary generated by standard copy mechanism miss some importance words, such as "obama" and "nominees".

*Equal contribution.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1355–1362
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

We propose a *Self-Attention Guided Copy mechanism* (SAGCopy) that aims to encourage the summarizer to copy important source words. Self-attention layer in the Transformer (Vaswani et al., 2017) builds a directed graph whose vertices represent the source words and edges are defined in terms of the relevance score between each pair of source words by dot-product attention (Vaswani et al., 2017) between the query Q and the key K. We calculate the centrality of each source words based on the adjacency matrices. A straightforward method is using TextRank (Mihalcea and Tarau, 2004) algorithm that assumes a word receiving more relevance score from others are more likely to be important. This measure is known as the in-degree centrality. We also adopt another measure assuming that a word sends out more relevance score to others is likely to be more critical, namely outdegree centrality, to calculate the source word centrality.

We utilize the centrality score as guidance for copy distribution. Specifically, we extend the dot-product attention to a centrality-aware function. Furthermore, we introduce an auxiliary loss computed by the divergence between the copy distribution and the centrality distribution, which aims to encourage the model to focus on important words.

Our contribution is threefold:

- We present a guided copy mechanism based on source word centrality that is obtained by the indegree or outdegree centrality measures.
- We propose a centrality-aware attention and a guidance loss to encourage the model to pay attention to important source words.
- We achieve state-of-the-art on the public text summarization dataset.

2 Related Work

Neural network based models (Rush et al., 2015; Nallapati et al., 2016; Chopra et al., 2016; Nallapati et al., 2017; Zhou et al., 2017; Tan et al., 2017; Gehrmann et al., 2018; Zhu et al., 2019; Li et al., 2020b,a) achieve promising results for the abstractive text summarization. Copy mechanism (Gulcehre et al., 2016; Gu et al., 2016; See et al., 2017; Zhou et al., 2018) enables the summarizers with the ability to copy from the source into the target via pointing (Vinyals et al., 2015). Recently, pre-training based methods (Devlin et al., 2019; Radford et al., 2018) have attracted growing attention and achieved state-of-the-art performances in many NLP tasks, and pre-training encoder-decoder Transformers (Song et al., 2019; Dong et al., 2019; Lewis et al., 2019; Xiao et al., 2020; Bao et al., 2020) show great successes for the summarization task. In this work, we explore the copy module upon the Transformer-based summarization model.

3 Background

We first introduce the copy mechanism. In Pointer-Generator Networks (PGNet) (See et al., 2017), the source text $\mathbf{x}$ are fed into a bidirectional LSTM (BiLSTM) encoder, producing a sequence of encoding hidden state $\mathbf{h}$:

$$h_i = \text{BiLSTM}(x_i, h_{i-1}) \tag{1}$$

On each step t, a unidirectional LSTM decoder receives the word embedding of the previous word to produce decoder state $\mathbf{s}$:

$$s_t = \text{LSTM}(s_{t-1}, y_{t-1}, c_t) \tag{2}$$

where c_t is a context vector generated based on the attention distribution (Bahdanau et al., 2015):

$$e_{t,i} = v^T tanh(W_h h_i + W_s s_t), \tag{3}$$

$$\alpha_t = \text{softmax}(e_t) \tag{4}$$

$$c_t = \sum\nolimits_i \alpha_{t,i} h_i \tag{5}$$

The vocabulary distribution P_{vocab} over all words in the target vocabulary is calculated as follows:

$$P_{vocab}(w) = \text{softmax}(W_a s_t + V_a c_t) \tag{6}$$

By incorporating a generating-copying switch $p_{gen} \in [0, 1]$, the final probability distribution of the ground-truth target word y_t is:

$$P(y_t) = p_{gen} P_{vocab}(y_t) + (1 - p_{gen}) P_{copy}(y_t) \tag{7}$$

$$p_{gen} = \text{sigmoid}(w_a^T c_t + u_a^T s_t + v_a^T y_{t-1}) \tag{8}$$

Copy distribution P_{copy} determines where to attend in time step t. In the most previous work, encoder-decoder attention weight α_t is serves as the copy distribution (See et al., 2017):

$$P_{copy}(w) = \sum\nolimits_{i:x_i=w} \alpha_{t,i} \tag{9}$$

The loss function $\mathcal{L}$ is the average negative log likelihood of the ground-truth target word y_t for each timestep t:

$$\mathcal{L} = -\frac{1}{T} \sum\nolimits_{t=0}^{T} \log P(y_t) \tag{10}$$

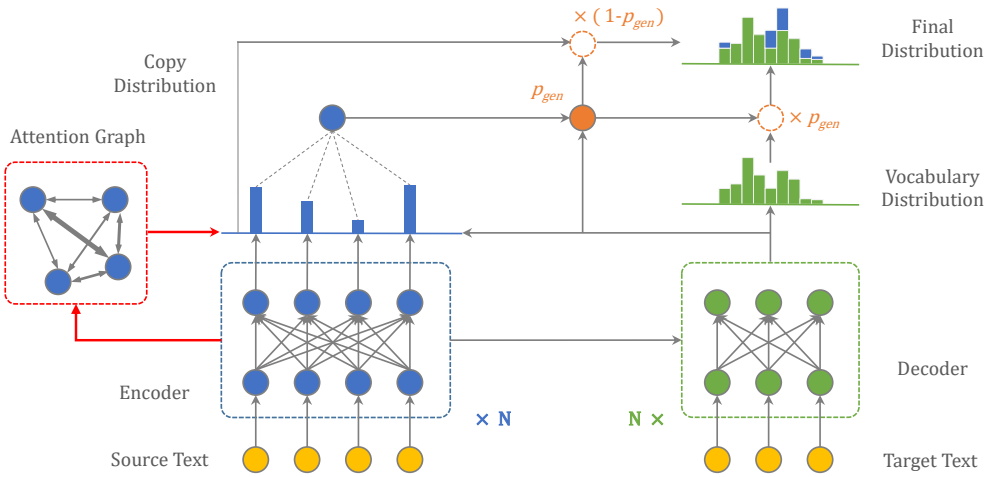

Figure 1: The framework of our proposed model. Based on the encoder self-attention graph, we calculate the centrality score for each source word to guide the copy module.

4 Model

In this section, we present our approach to enhance the copy mechanism. First, we briefly describe the Transformer model with the copy mechanism. Then, we introduce two methods to calculate the centrality scores for the source words based on the encoder self-attention layer. Finally, we incorporate the centrality score into the copy distribution and the loss function. The framework of our model is shown in Figure 1.

4.1 Transformer with the Copy Mechanism

Scaled dot-product attention (Vaswani et al., 2017) is widely used in self-attention networks:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (11)$$

where d_k is the number of columns of query matrix Q, key matrix K and value matrix V.

We take the encoder-decoder attentions in the last decoder layer as the copy distribution:

$$\alpha_{t,i} = \text{softmax}(\frac{(W_s s_t)^T W_h h_i}{\sqrt{d_k}}) \quad (12)$$

Note that for the multi-head attention, we obtain the copy distributions with the sum of multiple heads.

4.2 Self-Attention-Based Centrality

We introduce two approaches, i.e., indegree centrality and outdegree centrality, to calculate the centrality score for each source word based on the last encoder self-attention layer of the Transformer.

Centrality approaches are proposed to investigates the importance of nodes in social networks (Freeman, 1978; Bonacich, 1987; Borgatti and Everett, 2006; Kiss and Bichler, 2008; Li et al., 2011). Degree centrality is one of the simplest centrality measures that can be distinguished as indegree centrality and outdegree centrality (Freeman, 1978), which are determined based on the edges coming into and leaving a node, respectively.

Indegree centrality of a word is proportional to the number of relevance scores incoming from other words, which can be measured by the sum of the indegree scores or by graph-based extractive summarization methods (Mihalcea and Tarau, 2004; Erkan and Radev, 2004; Zheng and Lapata, 2019).

Outdegree centrality of a word is proportional to the number of relevance scores outgoing to other words, which can be computed by the sum of the outdegree scores.

Formally, let $G = (V, D)$ be a directed graph representing self-attention, where vertices V is the word set and edge $D_{i,j}$ is represented by the encoder self-attention weight from the word x_i to the word x_j, where $\sum_i D_{i,j} = 1$. Next, we introduce the approaches to calculate the word centrality with the graph G.

We first construct a transition probability matrix T as follows:

$$T_{i,j} = D_{i,j} / \sum_j D_{i,j}. \quad (13)$$

A basic indegree centrality is defined as:

$$score_i = \sum_j T_{j,i} \quad (14)$$

Alternatively, TextRank (Mihalcea and Tarau, 2004) that is inspired by PageRank algorithm (Page et al., 1999) calculates indegree centrality of the source words iteratively based on the Markov chain:

$$score_i = \sum_j T_{j,i} \cdot score_j \tag{15}$$

where $score_i$ is indegree centrality score for vertex V_i with initial score set to $1/|V|$. We can get a stationary indegree centrality distribution by computing $score = T \cdot score$ iteratively, and we take at most three iterations in our implementation.

Outdegree centrality measures how much a word i contributes to other words in the directed graph:

$$score_i = \sum_j D_{i,j} \tag{16}$$

Next, we incorporate the source word centrality score into the decoding process.

4.3 Guided Copy Mechanism

The motivation is that word centrality indicates the salience of the source words, which can provide the copy prior knowledge that can guide the copy module to focus on important source words.

We use word centrality score as an extra input to calculate the copy distribution as follows:

$$\alpha_{t,i} = \text{softmax}(\frac{(W_s s_t)^T (W_h h_i + w_p score_i)}{\sqrt{d_k}}) \tag{17}$$

where $score_i$ is the indegree or outdegree centrality score for the i-th word in source text. The above implementation can be referred to as centrality-aware dot-product attention.

Moreover, we expect that important source words can draw enough encoder-decoder attention. Thus, we adopt a centrality-aware auxiliary loss to encourage the consistency between the overall copy distribution and the word centrality distribution based on the Kullback-Leibler (KL) divergence:

$$\mathcal{L} = -\frac{1}{T}\sum_t \log P(y_t) + \lambda \text{KL}(\frac{1}{T}\sum_t \alpha_t, score) \tag{18}$$

5 Experiments

5.1 Experimental Setting

We evaluate our model in CNN/Daily Mail dataset (Hermann et al., 2015) and Gigaword dataset (Rush et al., 2015). Our experiments are conducted with 4 NVIDIA P40 GPU. We adopt 6 layer encoder and 6 layers decoder with 12 attention heads, and h_{model} = 768. Byte Pair Encoding (BPE) (Sennrich et al., 2016) word segmentation is used for data pre-processing. We warm-start the model parameter with MASS pre-trained base model[1] and trains about 10 epoches for convergence. During decoding, we use beam search with a beam size of 5.

5.2 Experimental Results

We compare our proposed Self-Attention Guided Copy (**SAGCopy**) model with the following comparative models.

Lead-3 uses the first three sentences of the article as its summary.

PGNet (See et al., 2017) is the Pointer-Generator Network.

Bottom-Up (Gehrmann et al., 2018) is a sequence-to-sequence model augmented with a bottom-up content selector.

MASS (Song et al., 2019) is a sequence-to-sequence pre-trained model based on the Transformer.

ABS (Rush et al., 2015) relies on an CNN encoder and a NNLM decoder.

ABS+ (Rush et al., 2015) enhances the ABS model with extractive summarization features.

SEASS (Zhou et al., 2017) controls the information flow from the encoder to the decoder with the selective encoding strategy.

SeqCopyNet (Zhou et al., 2018) extends the copy mechanism that can copy sequences from the source.

We adopt ROUGE (RG) F_1 score (Lin, 2004) as the evaluation metric. As shown in Table 2 and Table 3, **SAGCopy** with both outdegree and indegree centrality based guidance significantly outperform the baseline models, which prove the effectiveness of self-attention guided copy mechanism. The basic indegree centrality (indegree-1) is more favorable, considering the ROUGE score and computation complexity.

Besides ROUGE evaluation, we further investigate the guidance from the view of the loss function. For each sample in the Gigaword test set, we measure the KL divergence between the centrality score and the copy distribution, and we calculate the ROUGE-1 and ROUGE-2 scores. Figure 2 demonstrates that lower KL divergence yields a

[1] https://github.com/microsoft/MASS

Models	RG-1	RG-2	RG-L
Lead-3*	40.34	17.70	36.57
PGNet*	39.53	17.28	36.38
Bottom-Up*	41.22	18.68	38.34
MASS	41.38	19.11	38.42
MASS+Copy	41.71	19.41	38.66
SAGCopy Outdegree	42.53	**19.92**	**39.44**
SAGCopy Indegree-1	42.30	19.75	39.23
SAGCopy Indegree-2	**42.56**	19.89	39.40
SAGCopy Indegree-3	42.34	19.72	39.29

Table 2: ROUGE F_1 scores on the CNN/Daily Mail dataset. Results with * mark are taken from the corresponding papers. **Indegree-i** denote indegree centrality obtained by TextRank with i-iteration. Note that **Indegree-1** is the basic indegree centrality that is equivalent to TextRank with 1-iteration.

Models	RG-1	RG-2	RG-L
ABS*	29.55	11.32	26.42
ABS+*	29.76	11.88	26.96
SEASS*	36.15	17.54	33.63
SeqCopyNet*	35.93	17.51	33.35
MASS*	38.73	19.71	35.96
MASS+Copy	38.53	19.93	35.86
SAGCopy Outdegree	**38.86**	19.91	36.06
SAGCopy Indegree-1	38.84	**20.39**	**36.27**
SAGCopy Indegree-2	38.70	20.16	36.09
SAGCopy Indegree-3	38.69	19.83	35.98

Table 3: Experimental result on the Gigaword dataset.

higher ROUGE score, showing that our loss function is reasonable.

Additionally, we visualize the self-attention weights learned from our model in Figure 3, which demonstrates the guidance process.

5.3 Human Evaluation

We conduct human evaluations to measure the quantify of the summaries for *importance* and *readability*.

We randomly selected 100 samples from the Gigaword test set. The annotators are required to give a comparison between two model summaries that are presented anonymously. The results in Table 4 show that *SAGCopy* significantly outperforms *MASS+Copy* in terms of *Importance* and is comparative in terms of *Readability*.

	Win	Loss	Tie	kappa
Importance	20.67%	13.67%	65.67%	0.473
Readability	6.67%	3.67%	89.67%	0.637

Table 4: Human evaluation results on the Gigaword dataset. "Win" denotes the generated summary of *SAGCopy* is better than that of *MASS+Copy*. We evaluate the agreement by Fleiss' kappa (Fleiss, 1971).

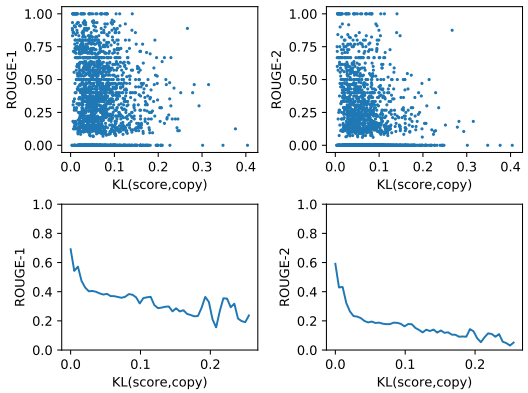

Figure 2: KL divergence with ROUGE F_1 in the Gigaword test set for *SAGCopy Indegree-1* model. Each point in the above plots represents an sample. The bottom plots show the average ROUGE score for different KL values.

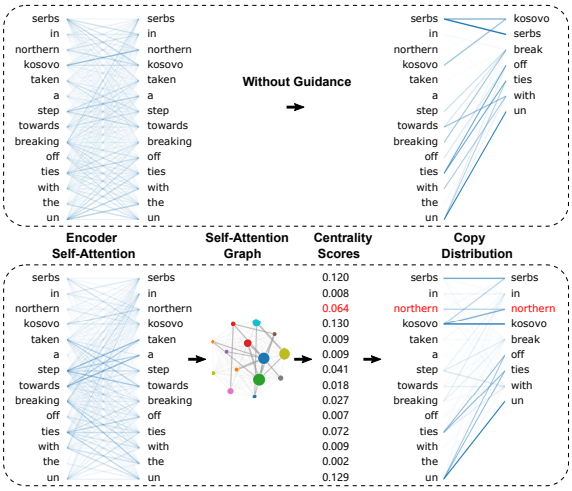

Figure 3: The guidance process for *SAGCopy Indegree* model, showing that the keyword "northern" is correctly copied for our model.

6 Conclusion

In this paper, we propose the SAGCopy summarization model that acquires guidance signals for the copy mechanism from the encoder self-attention graph. We first calculate the centrality score for each source word. Then, we incorporate the importance score into the copy module. The experimental results show the effectiveness of our model. For future work, we intend to apply our method to other Transformer-based summarization models.

Acknowledgments

This work is partially supported by Beijing Academy of Artificial Intelligence (BAAI).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Songhao Piao, Jianfeng Gao, Ming Zhou, et al. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. *arXiv preprint arXiv:2002.12804*.

Phillip Bonacich. 1987. Power and centrality: A family of measures. *American journal of sociology*, 92(5):1170–1182.

Stephen P. Borgatti and Martin G. Everett. 2006. A graph-theoretic perspective on centrality. *Soc. Networks*, 28(4):466–484.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California. Association for Computational Linguistics.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained language models for sequential sentence classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3693–3699, Hong Kong, China. Association for Computational Linguistics.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. 2019. Universal transformers. In *Proceedings of ICLR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 13042–13054.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.*, 22:457–479.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Linton C Freeman. 1978. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. 2016. Pointing the unknown words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 140–149, Berlin, Germany. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

Christine Kiss and Martin Bichler. 2008. Identification of influencers—measuring influence in customer networks. *Decision Support Systems*, 46(1):233–253.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020a. Aspect-aware multimodal summarization for chinese e-commerce products. In *Proceedings of the Thirty-Forth AAAI Conference on Artificial Intelligence (AAAI)*.

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, and Chengqing Zong. 2018a. Multi-modal sentence

summarization with modality attention and image filtering. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4152–4158.

Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2018b. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1430–1441, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Haoran Li, Junnan Zhu, Jiajun Zhang, Chengqing Zong, and Xiaodong He. 2020b. Keywords-guided abstractive sentence summarization. In *Proceedings of the Thirty-Forth AAAI Conference on Artificial Intelligence (AAAI)*.

Yung-Ming Li, Cheng-Yang Lai, and Ching-Wen Chen. 2011. Discovering influencers for marketing in the blogosphere. *Information Sciences*, 181(23):5143–5157.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3075–3081.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 5926–5936.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. Abstractive document summarization with a graph-based attentional neural model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1171–1181, Vancouver, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2692–2700.

Dongling Xiao, Han Zhang, Yu-Kun Li, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE-GEN: an enhanced multi-flow pre-training and fine-tuning framework for natural language generation. *CoRR*, abs/2001.11314.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5059–5069, Florence, Italy. Association for Computational Linguistics.

Hao Zheng and Mirella Lapata. 2019. Sentence centrality revisited for unsupervised summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6236–6247, Florence, Italy. Association for Computational Linguistics.

Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. 2017. Selective encoding for abstractive sentence summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1095–1104, Vancouver, Canada. Association for Computational Linguistics.

Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. 2018. Sequential copying networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. NCLS: Neural cross-lingual summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3054–3064, Hong Kong, China. Association for Computational Linguistics.