Improving Semantic Relevance for Sequence-to-Sequence Learning of Chinese Social Media Text Summarization

Shuming Ma[1,2], Xu Sun[1,2], Jingjing Xu[1,2], Houfeng Wang[1,2], Wenjie Li[3], Qi Su[4]
[1]MOE Key Laboratory of Computational Linguistics, Peking University
[2]School of Electronics Engineering and Computer Science, Peking University
[3]Department of Computing, The Hong Kong Polytechnic University
[4]School of Foreign Languages, Peking University
{shumingma, xusun, jingjingxu, wanghf, sukia}@pku.edu.cn
cswjli@comp.polyu.edu.hk

Abstract

Current Chinese social media text summarization models are based on an encoder-decoder framework. Although its generated summaries are similar to source texts literally, they have low semantic relevance. In this work, our goal is to improve semantic relevance between source texts and summaries for Chinese social media summarization. We introduce a Semantic Relevance Based neural model to encourage high semantic similarity between texts and summaries. In our model, the source text is represented by a gated attention encoder, while the summary representation is produced by a decoder. Besides, the similarity score between the representations is maximized during training. Our experiments show that the proposed model outperforms baseline systems on a social media corpus.

1 Introduction

Text summarization is to produce a brief summary of the main ideas of the text. For long and normal documents, extractive summarization achieves satisfying performance by selecting a few sentences from source texts (Radev et al., 2004; Woodsend and Lapata, 2010; Cheng and Lapata, 2016). However, it does not apply to Chinese social media text summarization, where texts are comparatively short and often full of noise. Therefore, abstractive text summarization, which is based on encoder-decoder framework, is a better choice (Rush et al., 2015; Hu et al., 2015).

For extractive summarization, the selected sentences often have high semantic relevance to the text. However, for abstractive text summarization, current models tend to produce grammatical and coherent summaries regardless of its semantic relevance with source texts. Figure 1 shows that the summary generated by a current model (RNN encoder-decoder) is similar to the source text literally, but it has low semantic relevance.

Text: 昨晚，中联航空成都飞北京一架航班被发现有多人吸烟。后因天气原因，飞机备降太原机场。有乘客要求重新安检，机长决定继续飞行，引起机组人员与未吸烟乘客冲突。
Last night, several people were caught to smoke on a flight of China United Airlines from Chendu to Beijing. Later the flight temporarily landed on Taiyuan Airport. Some passengers asked for a security check but were denied by the captain, which led to a collision between crew and passengers.

RNN: 中联航空机场发生爆炸致多人死亡。
China United Airlines exploded in the airport, leaving several people dead.

Gold: 航班多人吸烟机组人员与乘客冲突。
Several people smoked on a flight which led to a collision between crew and passengers.

Figure 1: An example of RNN generated summary. It has high similarity to the text literally, but low semantic relevance.

In this work, our goal is to improve the semantic relevance between source texts and generated summaries for Chinese social media text summarization. To achieve this goal, we propose a Semantic Relevance Based neural model. In our model, a similarity evaluation component is introduced to measure the relevance of source texts and generated summaries. During training, it maximizes the similarity score to encourage high semantic relevance between source texts and sum-

maries. The representation of source texts is produced by an encoder, while that of summaries is computed by a decoder. We introduce a gated attention encoder to better represent the source text. Besides, our decoder generates summaries and provide the summary representation. Experiments show that our proposed model has better performance than baseline systems on the social media corpus.

2 Background: Chinese Abstractive Text Summarization

Current Chinese social media text summarization model is based on encoder-decoder framework. Encoder-decoder model is able to compress source texts $x = \{x_1, x_2, ..., x_N\}$ into continuous vector representation with an encoder, and then generate the summary $y = \{y_1, y_2, ..., y_M\}$ with a decoder. In the previous work (Hu et al., 2015), the encoder is a bi-directional gated recurrent neural network, which maps source texts into sentence vector $\{h_1, h_2, ..., h_N\}$. The decoder is a unidirectional recurrent neural network, which produces the distribution of output words y_t with previous hidden state s_{t-1} and word y_{t-1}:

$$p(y_t|x) = softmaxf(s_{t-1}, y_{t-1}) \quad (1)$$

where f is recurrent neural network output function, and s_0 is the last hidden state of encoder h_N.

Attention mechanism is introduced to better capture context information of source texts (Bahdanau et al., 2014). Attention vector c_t is represented by the weighted sum of encoder hidden states:

$$c_t = \sum_{i=1}^{N} \alpha_{ti} h_i \quad (2)$$

$$\alpha_{ti} = \frac{e^{g(s_t, h_i)}}{\sum_{j=1}^{N} e^{g(s_t, h_j)}} \quad (3)$$

where $g(s_t, h_i)$ is a relevant score between decoder hidden state s_t and encoder hidden state h_i. When predicting an output word, the decoder takes account of attention vector, which contains the alignment information between source texts and summaries.

3 Proposed Model

Our assumption is that source texts and summaries have high semantic relevance, so our proposed model encourages high similarity between their representations. Figure 2 shows our proposed model. The model consists of three components: encoder, decoder and a similarity function. The encoder compresses source texts into semantic vectors, and the decoder generates summaries and produces semantic vectors of the generated summaries. Finally, the similarity function evaluates the relevance between the sematic vectors of source texts and generated summaries. Our training objective is to maximize the similarity score so that the generated summaries have high semantic relevance with source texts.

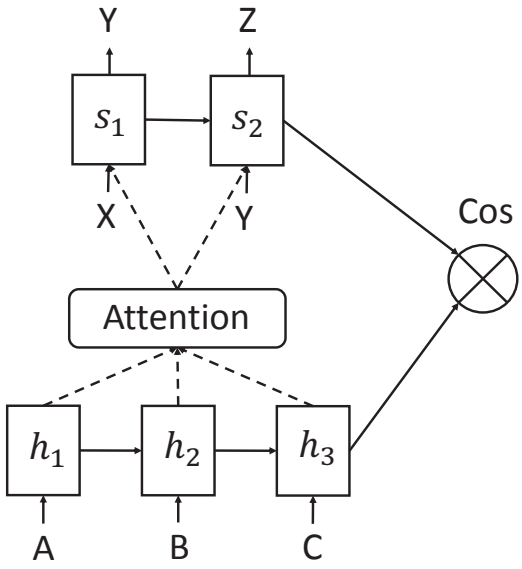

Figure 2: Our Semantic Relevance Based neural model. It consists of decoder (above), encoder (below) and cosine similarity function.

3.1 Text Representation

There are several methods to represent a text or a sentence, such as mean pooling of RNN output or reserving the last state of RNN. In our model, source text is represented by a gated attention encoder (Hahn and Keller, 2016). Every upcoming word is fed into a gated attention network, which measures its importance. The gated attention network outputs the important score with a feedforward network. At each time step, it inputs a word vector e_t and its previous context vector h_t, then outputs the score β_t. Then the word vector e_t is multiplied by the score β_t, and fed into RNN encoder. We select the last output h_N of RNN encoder as the semantic vector of the source text V_t.

A natural idea to get the semantic vector of a summary is to feed it into the encoder as well. However, this method wastes much time because

we encode the same sentence twice. Actually, the last output s_M contains information of both source text and generated summaries. We simply compute the semantic vector of the summary by subtracting h_N from s_M:

$$V_s = s_M - h_N \tag{4}$$

Previous work has proved that it is effective to represent a span of words without encoding them once more (Wang and Chang, 2016).

3.2 Semantic Relevance

Our goal is to compute the semantic relevance of source text and generated summary given semantic vector V_t and V_s. Here, we use cosine similarity to measure the semantic relevance, which is represented with a dot product and magnitude:

$$cos(V_s, V_t) = \frac{V_s \cdot V_t}{\|V_s\|\|V_t\|} \tag{5}$$

Source text and summary share the same language, so it is reasonable to assume that their semantic vectors are distributed in the same space. Cosine similarity is a good way to measure the distance between two vectors in the same space.

3.3 Training

Given the model parameter θ and input text x, the model produces corresponding summary y and semantic vector V_s and V_t. The objective is to minimize the loss function:

$$L = -p(y|x;\theta) - \lambda cos(V_s, V_t) \tag{6}$$

where $p(y|x;\theta)$ is the conditional probability of summaries given source texts, and is computed by the encoder-decoder model. $cos(V_s, V_t)$ is cosine similarity of semantic vectors V_s and V_t. This term tries to maximize the semantic relevance between source input and target output.

4 Experiments

In this section, we present the evaluation of our model and show its performance on a popular social media corpus. Besides, we use a case to explain the semantic relevance between generated summary and source text.

4.1 Dataset

Our dataset is Large Scale Chinese Short Text Summarization Dataset (LCSTS), which is constructed by Hu et al. (2015). The dataset consists of more than 2.4 million text-summary pairs, constructed from a famous Chinese social media website called Sina Weibo[1]. It is split into three parts, with 2,400,591 pairs in PART I, 10,666 pairs in PART II and 1,106 pairs in PART III. All the text-summary pairs in PART II and PART III are manually annotated with relevant scores ranged from 1 to 5, and we only reserve pairs with scores no less than 3. Following the previous work, we use PART I as training set, PART II as development set, and PART III as test set.

4.2 Experiment Setting

To alleviate the risk of word segmentation mistakes (Xu and Sun, 2016), we use Chinese character sequences as both source inputs and target outputs. We limit the model vocabulary size to 4000, which covers most of the common characters. Each character is represented by a random initialized word embedding. We tune our parameter on the development set. In our model, the embedding size is 400, the hidden state size of encoder-decoder is 500, and the size of gated attention network is 1000. We use Adam optimizer to learn the model parameters, and the batch size is set as 32. The parameter λ is 0.0001. Both the encoder and decoder are based on LSTM unit. Following the previous work (Hu et al., 2015), our evaluation metric is F-score of ROUGE: ROUGE-1, ROUGE-2 and ROUGE-L (Lin and Hovy, 2003).

4.3 Baseline Systems

RNN. We denote RNN as the basic sequence-to-sequence model with bi-directional GRU encoder and uni-directional GRU decoder. It is a widely used language generated framework, so it is an important baseline.

RNN context. RNN context is a sequence-to-sequence framework with neural attention. Attention mechanism helps capture the context information of source texts. This model is a stronger baseline system.

4.4 Results and Discussions

We compare our model with above baseline systems, including RNN and RNN context. We refer to our proposed Semantic Relevance Based neural model as **SRB**. Besides, SRB with a gated attention encoder is denoted as **+Attention**. Table 1

[1] weibo.sina.com

Model	ROUGE-1	ROUGE-2	ROUGE-L
RNN (W) (Hu et al., 2015)	17.7	8.5	15.8
RNN (C) (Hu et al., 2015)	21.5	8.9	18.6
RNN context (W) (Hu et al., 2015)	26.8	16.1	24.1
RNN context (C) (Hu et al., 2015)	29.9	17.4	27.2
RNN context + SRB (C)	32.1	18.9	29.2
+Attention (C)	**33.3**	**20.0**	**30.1**

Table 1: Results of our model and baseline systems. Our models achieve substantial improvement of all ROUGE scores over baseline systems. (W: Word level; C: Character level).

Text:仔细一算，上海的互联网公司不乏成功案例，但最终成为BAT一类巨头的几乎没有，这也能解释为何纳税百强的榜单中鲜少互联网公司的身影。有一类是被并购，比如：易趣、土豆网、PPS、PPTV、一号店等；有一类是数年偏安于细分市场。
With careful calculation, there are many successful Internet companies in Shanghai, but few of them becomes giant company like BAT. This is also the reason why few Internet companies are listed in top hundred companies of paying tax. Some of them are merged, such as Ebay, Tudou, PPS, PPTV, Yihaodian and so on. Others are satisfied with segment market for years.

Gold:为什么上海出不了互联网巨头?
Why Shanghai comes out no giant company?

RNN context:上海的互联网巨头。
Shanghai's giant company.

SRB:上海鲜少互联网巨头的身影。
Shanghai has few giant companies.

Figure 3: An Example of RNN generated summary on LCSTS corpus.

shows the results of our models and baseline systems. We can see SRB outperforms both RNN and RNN context in the F-score of ROUGE-1, ROUGE-2 and ROUGE-L. It concludes that SRB generates more key words and phrases. With a gated attention encoder, SRB achieves a better performance with 33.3 F-score of ROUGE-1, 20.0 ROUGE-2 and 30.1 ROUGE-L. It shows that the gated attention reduces noisy and unimportant information, so that the remaining information represents a clear idea of source text. The better representation of encoder leads to a better semantic relevance evaluation by the similarity function. Therefore, SRB with gated attention encoder is able to generate summaries with high semantic relevance to source text.

Model	level	R-1	R-2	R-L
RNN context (Hu et al., 2015)	Word	26.8	16.1	24.1
	Char	29.9	17.4	27.2
COPYNET (Gu et al., 2016)	Word	35.0	22.3	32.0
	Char	34.4	21.6	31.3
this work	Char	33.3	20.0	30.1

Table 2: Results of our model and state-of-the-art systems. COPYNET incorporates copying mechanism to solve out-of-vocabulary problem, so its has higher ROUGE scores. Our model does not incorporate this mechanism currently. In the future work, we will implement this technic to further improve the performance. (Word: Word level; Char: Character level; R-1: F-score of ROUGE-1; R-2: F-score of ROUGE-2; R-L: F-score of ROUGE-L)

Figure 3 is an example to show the semantic relevance between the source text and the summary. It shows that the main idea of the source text is about the reason why Shanghai has few giant company. RNN context produces "Shanghai's giant companies" which is literally similar to the source text, while SRB generates "Shanghai has few giant companies", which is closer to the main idea in semantics. It concludes that SRB produces summaries with higher semantic similarity to texts.

Table 2 summarizes the results of our model and state-of-the-art systems. COPYNET has the highest socres, because it incorporates copying mechanism to deals with out-of-vocabulary word problem. In this paper, we do not implement this mechanism in our model. In the future work, we will try to incorporates copying mechanism to our model to solve the out-of-vocabulary problem.

5 Related Work

Abstractive text summarization has achieved successful performance thanks to the sequence-to-sequence model (Sutskever et al., 2014) and attention mechanism (Bahdanau et al., 2014). Rush et al. (2015) first used an attention-based encoder to compress texts and a neural network language decoder to generate summaries. Following this work, recurrent encoder was introduced to text summarization, and gained better performance (Lopyrev, 2015; Chopra et al., 2016). Towards Chinese texts, Hu et al. (2015) built a large corpus of Chinese short text summarization. To deal with unknown word problem, Nallapati et al. (2016) proposed a generator-pointer model so that the decoder is able to generate words in source texts. Gu et al. (2016) also solved this issue by incorporating copying mechanism. Besides, Ayana et al. (2016) proposes a minimum risk training method which optimizes the parameters with the target of rouge scores.

Our work is also related to neural attention model. Neural attention model is first proposed by Bahdanau et al. (2014). There are many other methods to improve neural attention model (Jean et al., 2015; Luong et al., 2015) and accelerate the training process (Sun, 2016).

6 Conclusion

Our work aims at improving semantic relevance of generated summaries and source texts for Chinese social media text summarization. Our model is able to transform the text and the summary into a dense vector, and encourage high similarity of their representation. Experiments show that our model outperforms baseline systems, and the generated summary has higher semantic relevance.

Acknowledgements

This work was supported in part by National High Technology Research and Development Program of China (863 Program, No. 2015AA015404), and National Natural Science Foundation of China (No. 61673028). Xu Sun is the corresponding author of this paper.

References

Ayana, Shiqi Shen, Zhiyuan Liu, and Maosong Sun. 2016. Neural headline generation with minimum risk training. *CoRR* abs/1604.01904.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*. pages 93–98.

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, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Michael Hahn and Frank Keller. 2016. Modeling human reading with neural attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. pages 85–95.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A large scale chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. pages 1967–1972.

Sébastien Jean, KyungHyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*. pages 1–10.

Chin-Yew Lin and Eduard H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003*.

Konstantin Lopyrev. 2015. Generating news headlines with recurrent neural networks. *CoRR* abs/1512.01712.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based

neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. pages 1412–1421.

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, 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, CoNLL 2016, Berlin, Germany, August 11-12, 2016*. pages 280–290.

Dragomir R. Radev, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Çelebi, Stanko Dimitrov, Elliott Drábek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, Adam Winkel, and Zhu Zhang. 2004. MEAD - A platform for multidocument multilingual text summarization. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*.

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, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. pages 379–389.

Xu Sun. 2016. Asynchronous parallel learning for neural networks and structured models with dense features. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*. pages 192–202.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. pages 3104–3112.

Wenhui Wang and Baobao Chang. 2016. Graph-based dependency parsing with bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Kristian Woodsend and Mirella Lapata. 2010. Automatic generation of story highlights. In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*. pages 565–574.

Jingjing Xu and Xu Sun. 2016. Dependency-based gated recursive neural network for chinese word segmentation. In *Meeting of the Association for Computational Linguistics*. pages 567–572.