A Hybrid Word-Character Model for Abstractive Summarization

Chieh-Teng Chang, Chi-Chia Huang and **Jane Yung-Jen Hsu**

Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan
{scott820914, chifatty}@gmail.com
yjhsu@csie.ntu.edu.tw

Abstract

Abstractive summarization is the popular research topic nowadays. Due to the difference in language property, Chinese summarization also gains lots of attention. Most of studies use character-based representation instead of word-based to keep out the error introduced by word segmentation and OOV problem. However, we believe that word-based representation can capture the semantics of the articles more accurately. We proposed a hybrid word-character model preserves the advantage of both word-based and character-based representations. Our method also enables us to use larger word vocabulary size than anyone else. We call this new method **HWC** (Hybrid Word-Character). We conduct the experiments on LCSTS Chinese summarization dataset, and outperform the current state-of-the-art by at least 8 ROUGE points.

1 Introduction

Abstractive approach is currently the popular trend in the field of automatic text summarization tasks. It learns an internal language representation from the source text to generate a more human-like summary and thus has wider applications. There is breakthrough in abstractive approach with the development of neural networks. Most studies today are based on Seq2Seq with attention neural networks (Sutskever et al., 2014; Bahdanau et al., 2014), and it has become the baseline model of automatic text summarization research (Rush et al., 2015; Nallapati et al., 2016).

Text summarization in Chinese gains lots of attention in recent years. Chinese is very different from English especially in word segmentation and characters. Chinese words are usually composed of one to four characters, and there is no explicit delimiter between words. Word segmentation is a error-prone process, and it could largely affects the result of summarization. Since Chinese characters usually contain semantic meanings, many studies use character-based representation instead of word-based to keep out the error introduced by word segmentation (Hu et al., 2015; Chen et al., 2016; Li et al., 2017; Ayana et al., 2016).

Character-based representation has its limitation. Chinese characters usually have multiple meanings. Use only character embedding may misunderstand the meaning of the sentences. The difference between character-based and word-based representation are shown in Figure 1. We believe that word-based representation is still needed. It could help the model captures the meaning of the articles more accurately. In this study, we found that the use of hybrid embedding units for encoder and decoder preserves the advantage of both word-based and character-based representations. In addition, character vocabulary size is far less than word vocabulary size. As a side effect, we can distribute more computation to encoder, hence we are able to use larger word vocabulary size than anyone else. We applied the design to current baseline model, Seq2Seq with global attention, and got the result significantly better than the state of the art.

To the best of our knowledge, this is the first study to explore hybrid Chinese representation units in abstractive summarization.

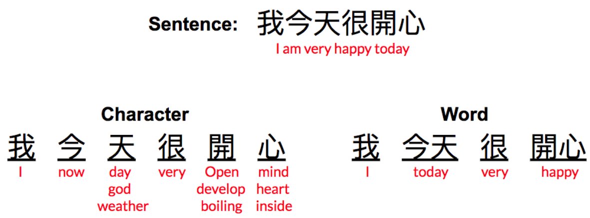

Figure 1: A sentence in different bases.

2 HWC

In this section, we introduce our neural network architecture and the benefits of it in dealing with Chinese summarization.

2.1 Seq2Seq Attention Neural Networks

We follow the work (Rush et al., 2015) using Seq2Seq attention neural networks for abstractive summarization. The details are shown in Figure 2. The tokens of input are fed into the encoder (a bidirectional LSTM) and generate a hidden state h_t. The global attention mechanism (Luong et al., 2015) calculates the context vector c_t. On each time step t in decode, the decoder (a unidirectional LSTM) receives the previous output y_{t-1}, decoder output state s_{t-1} and context vector c_{t-1} to generate the new output state s_t. Finally, the softmax layer generate the next output y_t based on previous word y_{t-1}, context vector c_t and current decoder state s_t.

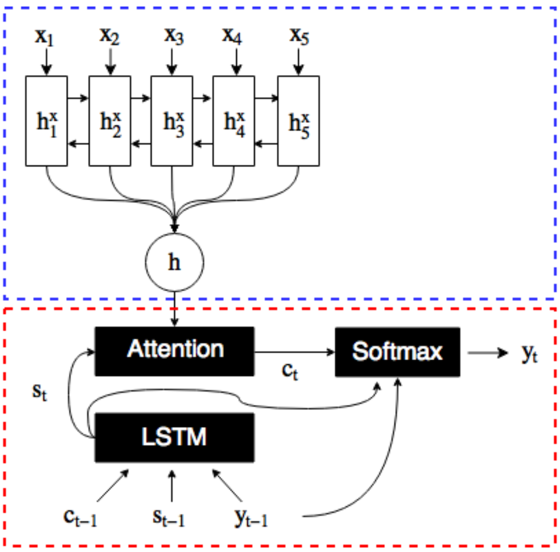

Figure 2: Seq2Seq with attention neural networks.

2.2 Hybrid Word-Character Model

Different from previous studies, which use the same type of embedding unit for both the encoder and the decoder, we propose to use word embedding as the encoder input and character embedding as the decoder learning target. There are few reasons for this. First of all, word as representation unit help the model captures the meaning of the articles more accurately than character as representation unit. Chinese characters have more serious polysemy problem than words. But it is not the ideal representation unit for the learning target. Full word set size is usually too large for training and low frequency words are usually excluded. Since the learning targets are much shorter than input text, unknown words in learning targets can cause dramatic performance drop. Character as representation unit is much fit for the learning targets. Secondly, it is common that long words are replace with shorter words in Chinese summarization. e.g. 奧運委員會 ⇒ 奧委會. Character as representation unit make this kind of transformations more applicable. Thirdly, character set is far small than word set. Take LSCTS dataset as an example. There are around $960k$ words while only around $10k$ characters in source text. Using hybrid embedding units let us distribute more computation to encoder. Thus, we are able to use much larger word bank to capture the meaning of the input text than other approach.

3 Experiment

3.1 Dataset

We conduct the experiments on a large scale Chinese short text summarization dataset(LCSTS) (Hu et al., 2015) to evaluate our method. This dataset is collected from the short news articles with the headlines in Chinese on Sina Weibo, a Chinese social media. This corpus consists of over 2 million real Chinese short texts with short summaries given by the author of each text. It contains 3 parts. Each pair in Part-II and Part-III has a human-labeled score, reflecting the relevance between the short text and the corresponding summary. The pairs in PART-II are labeled by 1 annotator, and pairs in Part-III are labeled by 5 annotators. The scores range from 1 to 5, '1' denotes " the least relevant " and '5' denotes "the most relevant". The statistics of the LCSTS corpus is presented in Table 1.

Following the setting of (Hu et al., 2015) we use Part I as the training set and the subset of Part III, which is scored by 3, 4 and 5, as the testing set.

Dataset	Number of Pairs	Number of score $\geq$3
Part I	2,400,591	-
Part II	10,666	8,685
Part III	1106	725

Table 1: The statistics of the LCSTS dataset.

Method		Encoder		Decoder		Evaluation		
		based	vocab size	based	vocab size	ROUGE-1	ROUGE-2	ROUGE-L
RNN	(Hu et al., 2015)	char	4000	char	4000	21.5	8.9	18.6
		word	50000	word	50000	17.7	8.5	15.8
RNN+context		char	3000	char	3000	29.9	17.4	27.2
		word	10000	word	10000	26.8	16.1	24.1
CopyNet	(Gum et al., 2016)	char	3000	char	3000	34.4	21.6	31.3
		word	10000	word	10000	35	22.3	32
Distraction	(Chen et al., 2016)	char	4000	char	4000	35.2	22.6	32.5
DGRD	(Li et al., 2017)	char	-	char	-	36.99	24.15	34.21
MRT	(Ayana et al., 2016)	char	3500	char	3500	37.87	25.43	35.33
HWC	HWC_1	char	10599	char	8248	**38.81**	**26.01**	**35.95**
	HWC_2	word	50000	char	8248	40.95	28.58	38.34
	HWC_3	word	500000	char	8248	**46.1**	**33.61**	**43.46**
	HWC_4	word	961213	char	8248	44.79	32.49	42.12

Table 2: ROUGE-F1 on LCSTS dataset.

3.2 Evaluation Metric

We adopt ROUGE(Lin, 2004) metrics for automatic evaluation, which has been widely used in abstractive summarization. It measures the quality of a summary by computing overlapping units between an system-generated summary and a reference summary. According to the practice, We report ROUGE-1 (1-gram), ROUGE-2 (bigrams) and ROUGE-L (longest common subsequence) F1 scores in the following experiments, and compare it against the other state-of-the-art models.

3.3 Baselines

- **RNN and RNN-context** (Hu et al., 2015) are two SeqSseq architectures. RNN-context use combination of all the hidden states of encoder as input of the decoder.

- **CopyNet** (Gum et al., 2016) employs a copying mechanism into the Seq2Seq framework.

- **Dirstaction** (Chen et al., 2016) is a Seq2Seq framework and distract the models to different content in order to better grasp the overall meaning of input documents.

- **DRGD** (Li et al., 2017) is a Seq2Seq framework equipped with a latent structure modeling component.

- **MRT** (Ayana et al., 2016) is the state-of-the-art system on both English and Chinese summarization tasks, which employ the minimum risk training strategy.

3.4 Implementation Details

Following the practice, text is segmented into Chinese words by using jieba. The vocabulary sizes of source and target in different bases (character and word) are shown in Table 3. For the experiments on LCSTS dataset, we set the size of embedding dimension and hidden layers to 500. Using dropout with probability $p = 0.5$. We use beam search with beam size 5 to generate the summary. We use $Adagrad$ as the optimizer, and the learning rate is initialized to 0.15. We also decay the learning rate if perplexity does not decrease on the validation set after finish an epoch. We leverage the popular attention-based Seq2Seq framework OpenNMT (Klein et al., 2017) as the foundation.

	source	target
char-based	10,599	8,248
word-based	961,213	427,926

Table 3: The vocabulary sizes of source and target in different bases on LCSTS dataset.

3.5 Result

We set the experiments to verify our method, and the results are shown in Table 2. HWC_1 which uses classic attention-based Seq2Seq with all the size of characters in both encoder and decoder beats the state-of-the-art without integrating other specific mechanisms. HWC_2, HWC_3 and HWC_4 which use the hybrid word-character model with large vocabulary size outperform other state-of-the-art methods with remarkable margins. Among these methods, HWC_3 which uses $500k$ vocabulary size outperforms the current state-of-the-art at least 8 ROUGE points. it is proved that using word-based representation captures the meaning of the articles more accurately.

4 Conclusion and Future Work

In this paper, we proposed a hybrid embedding unit from encoder to decoder method with large vocabulary size. The experiments showed that our method greatly outperforms state-of-the-art methods. For the future, we will work on the following aspects: (1) Adapt our method to other natural language processing(NLP) tasks in Chinese or other language, including dialogue systems, machine translation. (2) Incorporate the part-of-speech(POS) tagging to learn a better representation. (3) Improve the decoder with the coverage mechanism or other trick.

References

[Ayana et al.2016] Ayana, Shiqi Shen, Zhiyuan Liu, and Maosong Sun. 2016. Neural headline generation with minimum risk training. *arXiv preprint*, arXiv:1604.01904.

[Bahdanau et al.2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[Chen et al.2016] Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, and Hui Jiang. 2016. Distraction-based neural networks for document summarization. *arXiv preprint arXiv:1610.08462*.

[Gum et al.2016] Jiatao Gum, Zhengdong Lu, Hang Li, and Victor O. K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint*, arXiv:1603.06393.

[Hu et al.2015] Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A large scale chinese short text summarization dataset. *arXiv preprint*, arXiv:1506.05865.

[Klein et al.2017] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, July.

[Li et al.2017] Piji Li, Wai Lam, Lidong Bing, and Zihao Wang. 2017. Deep recurrent generative decoder for abstractive text summarization. *arXiv preprint*, arXiv:1708.00625.

[Lin2004] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, July.

[Luong et al.2015] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

[Nallapati et al.2016] Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

[Rush et al.2015] Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint*, arXiv:1509.00685.

[Sutskever et al.2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.