Abstract

Automatic text summarization is widely regarded as the highly difficult problem, partially because of the lack of large text summarization data set. Due to the great challenge of constructing the large scale summaries for full text, in this paper, we introduce a large corpus of Chinese short text summarization dataset constructed from the Chinese microblogging website Sina Weibo, which will be released to public soon. This corpus consists of over 2 million real Chinese short texts with short summaries given by the writer of each text. We also manually tagged the relevance of 10,666 short summaries with their corresponding short texts. Based on the corpus, we introduce recurrent neural network for the summary generation and achieve promising results, which not only shows the usefulness of the proposed corpus for short text summarization research, but also provides a baseline for further research on this topic.

1 Introduction

Nowadays, individuals or organizations can share or post information to the public in the social network. Take the popular Chinese microblogging website (Sina Weibo) as example, the People’s Daily, one of the media in China, posts more than tens of weibos (analogous to tweets) each day. Most of these weibos are well-written and highly informative because of the input length constraint (less than 140 Chinese characters). Such data is regarded as naturally annotated web resources (Sun, 2011). If we can mine these high-quality data from these naturally annotated web resources, it will be beneficial to the research that has been hampered by the lack of data.

Figure 1: A Weibo Posted by People’s Daily.

In the Natural Language Processing (NLP) community, automatic text summarization is a hot and difficult task. A good summarization system should understand the whole text and re-organize the information to generate coherent, informative, and significantly short summaries which convey important information of the original text (Hovy and Lin, 1998). (Martins, 2007). Recently, deep learning methods have shown potential abilities to learn representation and generate language from large scale data by utilizing GPUs. Many researchers realize that we are closer to generate abstractive summarizations by using the deep learning methods. However, the publicly available and high-quality large scale summarization data set is still very rare and not easy to be constructed by human writing. For example, the popular document summarization dataset DUC, TAC and TREC have only hundreds of human written English text summarizations. The problem is even worse for Chinese. In this paper, we take one step back and focus on constructing LCSTS, the Large-scale Chinese Short Text Summarization dataset by utilizing the naturally annotated web resources on Sina Weibo. Figure 1 shows one weibo posted by the People’s Daily. In order to convey the import information to the public quickly, it also writes a very informative and short summary (in the blue circle) of the news. Our goal is to mine a large scale, high-quality short text summarization dataset from these texts.

We make the following contributions: (1) We introduce a large scale Chinese short text summarization dataset. To our knowledge, it is the

1http://duc.nist.gov/data.html
2http://www.nist.gov/tac/2015/KBP/
3http://trec.nist.gov/
Figure 2: Diagram of the process for creating the dataset.

largest one to date; (2) We provide standard splits for the dataset into large scale training set and human labeled test set which will be easier for benchmarking the related methods easily; (3) We explore the properties of the dataset and sample 10,666 instances for manually checking the quality of the dataset; (4) we perform recurrent neural network based encoder-decoder method on the dataset to generate summary and get promising results, which can be used as one baseline of the task.

2 Relate Work

Our work is related to recent works automatic text summarization and natural language processing based on naturally annotated web resources, which are briefly introduced as follows.

Automatic Text Summarization in some form has been studied since 1950. Since then, most researches are related to extractive summarizations by analyzing the organization of the words in the document (Nenkova and McKeown, 2011) (Luhn, 1998); Because it needs labeled data sets for supervised machine learning methods and labeling dataset is very intensive. some researches focused on the unsupervised methods (Mihalcea, 2004). The scale of existing data sets are usually very small (most of them are less than 1000). For example, DUC2002 dataset contains 567 document and each document is provided with two 100-words human summaries. Our work is also related to the headline generation, which is a task to generate one sentence of the text it entitles. Colmenares et.al construct a 1.3 million financial news headline dataset written in English for headline generation (Colmenares et al., 2015). However, the data set is not publicly available.

Naturally Annotated Web Resources based Natural Language Processing is proposed by Sun (Sun, 2011). Naturally Annotated Web Resources is the data generated by users for communicative purposes such as web pages, blogs and microblogs. We can mine knowledge or useful data from these raw data by using marks generated by users unintentionally. Jure et.al tracks 1.6 million mainstream media sites and blogs and mine a set of novel and persistent temporal patterns in the news cycle (Leskovec et al., 2009). Sepandar et.al use the users’ naturally annotated pattern ‘we feel’ and ‘i feel’ to extract the ‘Feeling’ sentence collection which is used to collect the world’s emotions. In this work, we use the naturally annotated resources to construct the large scale Chinese short text summarization data to facilitate the research on text summarization.

3 Data Collection

A lot of popular Chinese media and organizations have created accounts on the Sina Weibo. They use their accounts to post news and information. These accounts are verified on the Weibo and labeled by a blue ‘V’. In order to guarantee the quality of the crawled text, we only crawl the verified organizations’ weibos. The process of the data collection is shown as Figure 2 and summarized as follows:

1) We first collect 50 very popular organization users as seeds which come from the domains of politic, economic, military, movies, game and etc, such as People’s Daily, the Economic Observer press, the Ministry of National Defense and etc. 2) We then crawl followed users of these seed users and filter them by conditions such as a) the user must be blue verified, b) the number of followers is more than 1 million. 3) We use the chosen users and text crawler to crawl their weibos, 4) we filter, clean and extract (short text, summary) pairs. We remove those pairs, whose short text length is too short (less than 80 characters) and length of summaries is out of [10,30], from the corpus.
4 Data Properties

The dataset consists of three parts as shown in Table 1 and the length distributions of texts are shown as Figure 3.

**Part I** contains the large scale (short text, summary) pairs. These pairs can be used to train supervised learning model for summarization generation.

| Part | Number of Pairs | Human Score 1 | Human Score 2 | Human Score 3 | Human Score 4 | Human Score 5 |
|------|-----------------|---------------|---------------|---------------|---------------|---------------|
| I    | 2,400,591       | 10,666        | 942           | 1,039         | 3,128         | 3,538         |
| II   |                 |               |               |               |               |               |
| III  |                 |               |               |               |               |               |

**Table 1: Data Statistics**

**Part II** contains the 10,666 human labeled (short text, summary) pairs, the score ranges from 1 to 5 which indicates the relevance between the short text and the corresponding summary. ‘1’ denotes “the least relevant” and ‘5’ denotes “the most relevant”. These data is randomly sampled from Part I and is used to analyze the distribution of pairs in the Part I. Figure 4 illustrates examples of different scores. From the examples we can see that pairs scored by 3, 4 or 5 are very relevant to the corresponding summaries, while the summaries of pairs scored by 1 or 2 are highly abstract and is relatively hard to conclude the summaries from the short text. They are more likely to be headlines or comments instead of summaries.

**Part III** contains 1,106 pairs which are scored by 3 persons simultaneously. This part is independent from Part I and Part II. In this work, we use pairs scored by 3, 4 and 5 of this part as the test set for short text summarization generation task. Part II and Part III can also be used as training and testing set to train a model which can be used to select required part of Part I.

Figure 3: Box plot of lengths for short text(ST), segmented short text(Segmented ST), summary(SUM) and segmented summary(Segmented SUM). The red line denotes the median, and the edges of the box are the quartiles.

Figure 4: Five examples of different scores.
5 Experiment

Recently, recurrent neural network (RNN) have shown powerful abilities on speech recognition (Graves et al., 2013), machine translation (Sutskever et al., 2014) and automatic dialog response (Shang et al., 2015). However, there is rarely research on the automatic text summarization by using deep models. In this section, we use RNN as encoder and decoder to generate the summary of short text. We use the Part I as the training set and the subset of Part III which is scored by 3, 4 and 5, as test set. Two approaches are used to preprocess the data: 1) character-based method, we take the Chinese character as input, which will reduce the vocabulary size to 4,000. 2) word-based method, the text is segmented into Chinese words by using jieba4. The vocabulary is limited to 50,000. We adopt two deep architectures, 1) The local context is not used during decoding. We use the Gated Feedback RNN proposed by Chung et al. (Chung et al., 2015) as encoder and it’s last hidden state as the input of decoder, as shown in Figure 5(a); 2) The context is used during decoding, following (Bahdanau et al., 2014), we use the combination of all the hidden states of encoder as input of the decoder, as shown in Figure 5(b).

For evaluation, we adopt the ROUGE metrics (Lin, 2004) proposed by (Lin and Hovy, 2003), which have been proved strongly correlated with human evaluations. ROUGE-1, ROUGE-2 and ROUGE-L are used. All the models are trained on the GPUs tesla M2090 for about one week. Table 2 lists the experiment results. As we can see in Figure 6, the summaries generated by RNN with context is very close to human written summaries, which indicates if we feed enough

4https://pypi.python.org/pypi/jieba/

5 Experiment

Recently, recurrent neural network (RNN) have shown powerful abilities on speech recognition (Graves et al., 2013), machine translation (Sutskever et al., 2014) and automatic dialog response (Shang et al., 2015). However, there is rarely research on the automatic text summarization by using deep models. In this section, we use RNN as encoder and decoder to generate the summary of short text. We use the Part I as the training set and the subset of Part III which is scored by 3, 4 and 5, as test set. Two approaches are used to preprocess the data: 1) character-based method, we take the Chinese character as input, which will reduce the vocabulary size to 4,000. 2) word-based method, the text is segmented into Chinese words by using jieba4. The vocabulary is limited to 50,000. We adopt two deep architectures, 1) The local context is not used during decoding. We use the Gated Feedback RNN proposed by Chung et al. (Chung et al., 2015) as encoder and it’s last hidden state as the input of decoder, as shown in Figure 5(a); 2) The context is used during decoding, following (Bahdanau et al., 2014), we use the combination of all the hidden states of encoder as input of the decoder, as shown in Figure 5(b).

For evaluation, we adopt the ROUGE metrics (Lin, 2004) proposed by (Lin and Hovy, 2003), which have been proved strongly correlated with human evaluations. ROUGE-1, ROUGE-2 and ROUGE-L are used. All the models are trained on the GPUs tesla M2090 for about one week. Table 2 lists the experiment results. As we can see in Figure 6, the summaries generated by RNN with context is very close to human written summaries, which indicates if we feed enough

Table 2: The experiment result: “Word” denotes the text-based word input and “Char” the character-based input.

| model     | data       | R-1  | R-2  | R-L  |
|-----------|------------|------|------|------|
| RNN       | Word       | 0.043| 0.025| 0.043|
|           | Char       | 0.061| 0.028| 0.057|
| RNN context| Word      | 0.087| 0.054| 0.085|
|           | Char       | 0.108| 0.073| 0.107|

6 Conclusion and Future Work

We constructed a large-scale Chinese short text summarization dataset and performed RNN-based methods on that, which achieved some promising results. This is just a start of deep models on this task and there is much room for improvement. We take the whole short text as one sequence, this may be not very reasonable, because one short text consists of many sentences. A hierarchical RNN (Li et al., 2015) is one possible direction. We also plan to construct a large document summarization data set by using naturally annotated web resources.
References

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

[Chung et al.2015] Junyoung Chung, Çağlar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2015. Gated feedback recurrent neural networks. *CoRR*, abs/1502.02367.

[Colmenares et al.2015] Carlos A. Colmenares, Marina Litvak, Amin Mantrach, and Fabrizio Silvestri. 2015. Heads: Headline generation as sequence prediction using an abstract feature-rich space. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics–Human Language Technologies (NAACL HLT 2015).*

[Graves et al.2013] Alex Graves, Abdel-rahman Mohamed, and Geoffrey E. Hinton. 2013. Speech recognition with deep recurrent neural networks. *CoRR*, abs/1303.5778.

[Hovy and Lin1998] Eduard Hovy and Chin-Yew Lin. 1998. Automated text summarization and the summarist system. In *Proceedings of a Workshop on Held at Baltimore, Maryland: October 13-15, 1998, TIPSTER ’98*, pages 197–214, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Leskovec et al.2009] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’09, pages 497–506.

[Li et al.2015] Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. In *Proceedings of ACL*.

[Lin and Hovy2003] Chin-Yew Lin and E.H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003)*, Edmonton, Canada.

[Lin2004] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *In Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004*, Barcelona, Spain.

[Luhn1998] H. P. Luhn. 1998. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

[Mihalcea2004] Rada Mihalcea. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, companion volume*, Spain.

[Nenkova and McKeown2011] Ani Nenkova and Kathleen McKeown. 2011. Automatic summarization. *Foundations and Trend in Information Retrieval*, 5(2-3):103–233.

[Shang et al.2015] Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. *CoRR*, abs/1503.02364.

[Sun2011] Mao Song Sun. 2011. Natural language processing based on naturally annotated web resources. *Journal of Chinese Information Processing*, 25(6):26–32.

[Sutskever et al.2014] Ilya Sutskever, Oriol Vinyals, and Quoc V. V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112.