LCSTS: A Large Scale Chinese Short Text Summarization Dataset

Baotian Hu **Qingcai Chen** **Fangze Zhu**

Intelligent Computing Research Center
Harbin Institute of Technology, Shenzhen Graduate School
{baotianchina,qingcai.chen, zhufangze123}@gmail.com

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 is released to the public[1]. This corpus consists of over 2 million real Chinese short texts with short summaries given by the author 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 easily share or post information to the public on the social network. Take the popular Chinese microblogging website (Sina Weibo) as an 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 text length limitation (less than140 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.

【江西高考被曝替考 有关考生已被警方控制】人民日报记者吴齐强消息，江西高考被曝光替考，7日中午江西省教育厅发布消息称，接到有人组织替考的举报后，江西省教育厅、江西省教育考试院立即部署南昌市教育考试院，联合南昌市警方开展调查核实，有关考生已被警方控制。调查进展情况将及时向社会公布。

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). Most of traditional abstractive summarization methods divide the process into two phrases (Bing et al., 2015). First, key textual elements are extracted from the original text by using unsupervised methods or linguistic knowledge. And then, unclear extracted components are rewritten or paraphrased to produce a concise summary of the original text by using linguistic rules or language generation techniques. Although extensive researches have been done, the linguistic quality of abstractive summary is still far from satisfactory. Recently, deep learning methods have shown potential abilities to learn representation (Hu et al., 2014; Zhou et al., 2015) and generate language (Bahdanau et al., 2014; Sutskever et al., 2014) 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 manually. For example, the popular document summarization dataset DUC[2], TAC[3] and TREC[4] have only hundreds of human written English text summarizations. The problem is even worse for Chinese. In this pa-

[1] http://icrc.hitsz.edu.cn/Article/show/139.html

[2] http://duc.nist.gov/data.html

[3] http://www.nist.gov/tac/2015/KBP/

[4] http://trec.nist.gov/

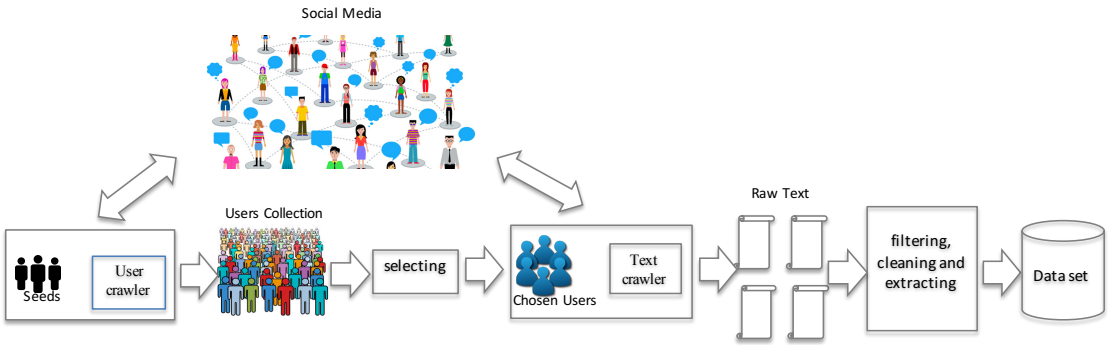

Figure 2: Diagram of the process for creating the dataset.

per, we take one step back and focus on constructing **LCSTS**, the **L**arge-scale **C**hinese **S**hort **T**ext **S**ummarization 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.

This paper makes the following contributions: (1) We introduce a large scale Chinese short text summarization dataset. To our knowledge, it is the 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; (3) We explore the properties of the dataset and sample 10,666 instances for manually checking and scoring 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 Related Work

Our work is related to recent works on 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); Since 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 documents 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 track 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 which are more likely to be clean, formal and informative. There are a lot of human intervention required in each step. 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. They come from the domains of politic, economic, military, movies, game and etc, such as People's Daily, the Economic Observe press, the Ministry of National Defense and etc. 2) We then crawl fusers followed by these seed users and filter them by using human written rules such as the user must be blue verified, the number of followers is more than 1 million and etc. 3) We use the chosen users and text crawler to crawl their weibos. 4) we filter, clean and extract (short text, summary) pairs. About 100 rules are used to extract high quality pairs. These rules are concluded by 5 peoples via carefully investigating of the raw text. We also remove those paris, whose short text length is too short (less than 80 characters) and length of summaries is out of [10,30].

4 Data Properties

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

Part I is the main content of LCSTS that contains 2,400,591 (short text, summary) pairs. These pairs can be used to train supervised learning model for summary generation.

Part II contains the 10,666 human labeled (short text, summary) pairs with the score ranges from 1 to 5 that indicates the relevance between the short text and the corresponding summary. '1' denotes " the least relevant " and '5' denotes "the most relevant". For annotating this part, we recruit 5 volunteers, each pair is only labeled by one annotator. These pairs are randomly sampled from Part I and are used to analysize 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. These summaries are highly informative, concise and significantly short compared to original text. We can also see that many words in the summary do not appear in the original text, which indicates the significant difference of our dataset from sentence compression datasets. The summaries of pairs scored by 1 or 2 are highly abstractive and relatively hard to conclude the summaries from the short text. They are more likely to be headlines or comments instead of summaries. The statistics show that the percent of score 1 and 2 is less than 20% of the data, which can be filtered by using trained classifier.

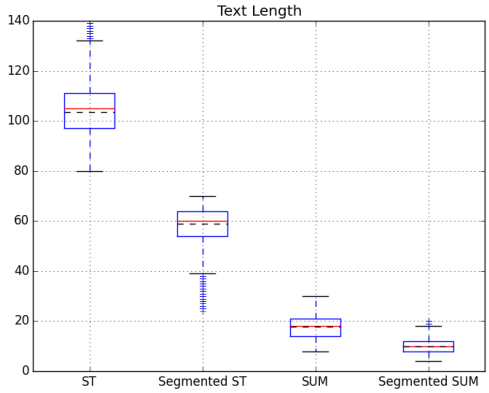

Figure 3: Box plot of lengths for short text(ST), segmented short text(Segmented ST), summary(SUM) and segmented summary(Segmented SUM). The red line denotes the median, and the edges of the box the quartiles.

Part III contains 1,106 pairs. For this part, 3 annotators label the same 2000 texts and we extract the text with common scores. 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 summary generation task.

Part I	2,400,591	
Part II	Number of Pairs	10,666
	Human Score 1	942
	Human Score 2	1,039
	Human Score 3	2,019
	Human Score 4	3,128
	Human Score 5	3,538
Part III	Number of Pairs	1,106
	Human Score 1	165
	Human Score 2	216
	Human Score 3	227
	Human Score 4	301
	Human Score 5	197

Table 1: Data Statistics

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 rare 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

Short Text: 水利部水资源司司长陈明忠今日在新闻发布会上透露，根据刚刚完成的水资源管理制度的考核，有部分省接近了红线的指标，有部分省超过红线的指标。在一些超过红线的地方，将对一些取用水项目进行区域的限批，严格地进行水资源论证和取水许可的批准。

Mingzhong Chen, the Chief Secretary of the Water Devision of the Ministry of Water Resources, revealed today at a press conference, according to the just-completed assessment of water resources management system, some provinces are closed to the red line indicator, some provinces are over the red line indicator. In some places over the red line, It will enforce regional approval restrictions on some water projects, implement strictly water resources assessment and the approval of water licensing.

Summarization: 部分省超过年度用水红线指标 取水项目将被限批

Some provinces exceeds the red line indicator of annual water using, some water project will be. limited approved

Human Score: 5

Short Text: 各团购网站移动端销售额占比均在30%以下，用户通过PC端购物习惯短时间内难以转变。未来中国餐饮O2O市场，移动端将成为餐饮O2O的战略性发展方向，也将由线上驱动转变为线下驱动。一二线城市面临增长窘境，三四线城市O2O市场蕴含机会。

Groupons' sales on mobile terminals are below 30 percent. User's preference of shopping through PCs can not be changed in the short term. In the future Chinese O2O catering market, mobile terminals will become the strategic development direction. And also, it will become off-line driving from on-line driving. The first and second tier cities are facing growth difficulties. However, O2O market in the third and fourth tier cities contains opportunities.

Summarization: 移动端成餐饮O2O的战略性发展方向

The mobile terminals will become catering's strategic development direction.

Human Score: 4

Short Text: 7月百城住宅新建住宅平均价格为10347元/平方米，环比上涨0.87%，自去年6月以来连续14个月环比上涨。其中，广州、北京、深圳、南京涨幅均超过10%。中原张大伟认为，一二线城市因为集聚了过多资源，房价易涨难跌。

In July, 100-cities' average newly-built house prices is 10347 yuan per square, which rose 0.87%. It rises for the 14th consecutive month. Among them, Guangzhou, Beijing, Shenzhen, Nanjing rise more than 10%. Dawei Zhang, from Centaline Property Agency, said that because the first and second-tier city gathers too many resources, the price of house is likely to rise and hard to fall.

Summarization: 百城房价环比“14连涨”一二线城市涨幅扩大

100-cities' house prices gain “14th consecutive rising”, the first and second-tier cities rise more.

Human Score: 3

Short Text: 记者梳理发现，2009年至今有8起福彩开奖延迟事件，至少延迟2小时，2014年5月6日第2014050期延迟开奖达4小时。8起事件中福彩中心对其中3起给出了回应，理由有通讯故障及暴雨导致的数据上传延迟。另5起均未解释原因。

Reporters combed the information and found, from 2009 to now, there are at least 8 lottery delayed events and the delayed time are more than 2 hours. On May 6, 2014, the No. 2014050 delay more than 4 hours. The center of welfare lottery only respond to 3 of the 8 event. Their explanations are that a communications breakdown and heavy rain led to a data upload extension. There are no explanations for other 5 delay events.

Summarization:三问双色球开奖延迟：开奖为何要等数据汇总？

Ask about the lottery delay third times: why lottery should wait data collection?

Human Score: 2

Short Text: 商务部数据显示，中国7月实际利用外资同比大幅下降16.95%至78.1亿美元。外界有分析与近期官方对外资企业的密集反垄断调查有关。沈丹阳回应指出，“不能与对外资的反垄断调查挂钩，或者做其他没有根据的联想”。

According to China's Ministry of Commerce, China's actually utilized foreign capital in July fell sharply about 16.95% to 7.81 billion dollars, comparing to last year. Analysis of the outside world believe that it is related to the recent official intensive antitrust investigation. Danyang Shen responded, “It can not be linked to the antitrust investigation of foreign investment, or do other unfounded association”

Summarization: 商务部表态反垄断：几个案子不会把外商吓回去

China's Ministry of Commerce respond to antitrust investigation: Several cases will not scare foreign investors away.

Human Score: 1

Figure 4: Five examples of different scores.

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

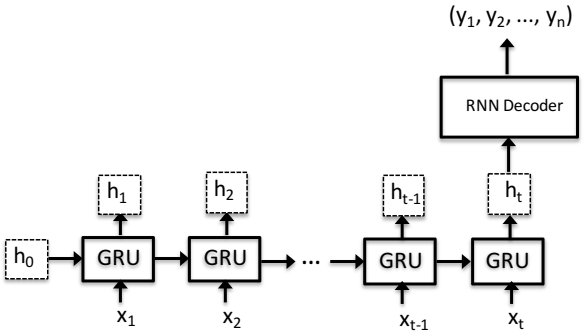

Figure 5: The graphical depiction of RNN encoder and decoder framework without context.

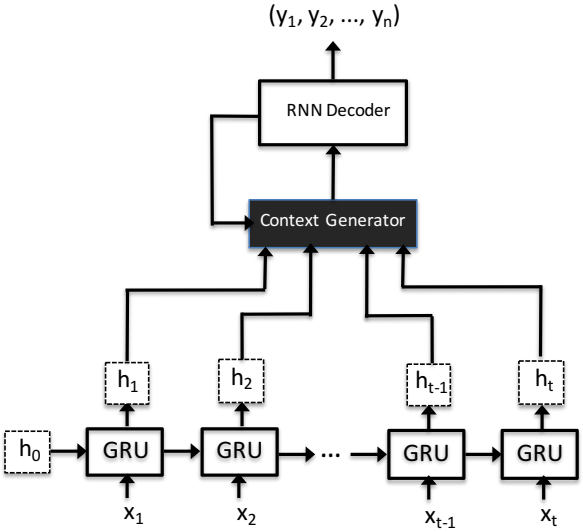

Figure 6: The graphical depiction of the RNN encoder and decoder framework with context.

is segmented into Chinese words by using jieba[5]. The vocabulary is limited to 50,000. We adopt two deep architectures, 1) The local context is not used during decoding. We use the RNN as encoder and it's last hidden state as the input of decoder, as shown in Figure 5; 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 6. For the RNN, we adopt the gated recurrent unit (GRU) which is proposed by (Chung et al., 2015) and has been proved comparable to LSTM (Chung et al., 2014). All the parameters (including the embeddings) of the two architectures are randomly initialized and ADADELTA (Zeiler, 2012) is used to update the learning rate. After the model is trained, the beam search is used to generate the best summaries in the process of decoding and the size of beam is set to 10 in our experiment.

For evaluation, we adopt the ROUGE metric-

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

model	data	R-1	R-2	R-L
RNN	Word	0.177	0.085	0.158
	Char	0.215	0.089	0.186
RNN context	Word	0.268	0.161	0.241
	Char	**0.299**	**0.174**	**0.272**

Table 2: The experiment result: "Word" and "Char" denote the word-based and character-based input respectively.

Short Text: 9月3日，总部位于日内瓦的世界经济论坛发布了《2014-2015年全球竞争力报告》，瑞士连续六年位居榜首，成为全球最具竞争力的国家，新加坡和美国分列第二位和第三位。中国排名第28位，在金砖国家中排名最高。

On September 3, the Geneva-based World Economic Forum released "The Global Competitiveness Report 2014-2015". Switzerland topped the list for six consecutive years, becoming the world 's most competitive country. Singapore and the United States are in the second and third place respectively. China is in the 28th place, ranking highest among the BRIC countries.

Human: 全球竞争力排行榜 中国居28位 居金砖国家首位

The Global competitiveness ranking list, China is in the 28th place, the highest among BRIC countries.

RNN+Char: 全球竞争力报告：瑞士居第六位居第26位第三位居第26位第六位居第26位第六位居第26位第六位居第26位

"The Global Competitiveness Report": Switzerland ranks 6th in 26th ranks 3rd ranks 26th 6th ranks 26th place 6th place ranks 26th place 6th place ranks 26th place

RNN+Word: 2014年全球竞争力排名：瑞士第一北京第第第第名单第第第名单第68位世界第第第名单第68位

2014 Global competitiveness ranking: Switzerland ranks first place, Beijing ranks ranks ranks ranks list ranks ranks ranks list ranks 68th world ranks ranks ranks list ranks 68th .

RNN+Context+Char: 全球竞争力排名：中国排名第28位金砖国最高

The Global competitiveness ranking list: China is in the 28th place which is highest among the BRIC countries.

RNN+Context+Word: 2014-2015年全球竞争力报告：瑞士连续6年居榜首中国居28位(首/3 — — — 访榜首)中国排名第28位

"The Global Competitiveness Report 2014-2015" : Switzerland topped the list for six consecutive years, China is in the 28th place (*can not be translated*) China ranks 28th.

Figure 7: An example of the generated summaries.

s proposed by (Lin and Hovy, 2003), which has been proved strongly correlated with human evaluations. ROUGE-1, ROUGE-2 and ROUGE-L are used. Because the standard Rouge package [6] is used for evaluating English summarization systems, we transform the Chinese words to numerical IDs to adapt to the systems. 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 7, the summaries generated by RNN with context are very close to human written summaries, which indicates that if we feed enough data to the RNN encoder and decoder, it may generate summary almost from scratch.

The results also show that the RNN with context outperforms RNN without context on both character and word based input. This result indicates that the internal hidden states of the RNN encoder can be combined to represent the context of words in summary. And also the performances of the character-based input outperform the word-based input. As shown in Figure 8, the summary generated by RNN with context by inputing the character-based short text is relatively good, while the the summary generated by RNN with context

[6] http://www.berouge.com/Pages/default.aspx

Short Text: 工厂， 大门紧锁， 约20名工人散坐在树荫下。“我们就是普通工人，在这里等工资。”其中一人说道。7月4日上午，记者抵达深圳龙华区清湖路上的深圳愿景光电子有限公司。正如传言一般，愿景光电子倒闭了，大股东刑毅不知所踪。

The door of the factory is locked. About 20 works are scattered to sit under the shade. "We are ordinary workers, we are waiting for our salary here.", one of them said. On the morning of July 4th, reporters arrived at Shenzhen Yuanjing Photoelectron Corporation located on Qinghu Road, Longhua District, Shenzhen. Just as the rumor, Yuanjing Photoelectron Corporation is closed down and the large shareholder Xing Yi is missing.

Human: 深圳亿元级LED企业倒闭烈日下工人苦等老板

Hundred-million--Yuan class LED enterprise is closed down and workers wait for the boss under the scorching sun.

RNN+Context+Char: 深圳愿景光电子倒闭了(图)(组图)

Shenzhen Yuanjing Photoelectron Corporation is closed down (picture) (set of pictures)

RNN+Context+Word: 深圳 "UNK":深圳 UNK， UNK， UNK， UNK

Shenzhen "UNK": Shenzhen UNK , UNK, UNK, UNK, UNK

Figure 8: An example of the generated summaries with UNKs.

on word-based input contains many UNKs. This may attribute to that the segmentation may lead to many UNKs in the vocabulary and text such as the person name and organization name. For example, "愿景光电子" is a company name which is not in the vocabulary of word-based RNN, the RNN summarizer has to use the UNKs to replace the "愿景光电子" in the process of decoding.

6 Conclusion and Future Work

We constructed a large-scale Chinese short text summarization dataset and performed RNN-based methods on it, 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 not be very reasonable, because most of short texts contain several sentences. A hierarchical RNN (Li et al., 2015) is one possible direction. The rare word problem is also very important for the generation of the summaries, especially when the input is word-based instead of character-based. It is also a hot topic in the neural generative models such as neural translation machine(NTM) (Luong et al., 2014), which can benefit to this task. We also plan to construct a large document summarization data set by using naturally annotated web resources.

Acknowledgments

This work is supported by National Natural Science Foundation of China: 61473101, 61173075 and 61272383, Strategic Emerging Industry Development Special Funds of Shenzhen: JCYJ20140417172417105, JCYJ20140508161040764 and JCYJ20140627163809422. We thank to Baolin Peng, Lin Ma, Li Yu and the anonymous reviewers for their insightful comments.

References

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

Lidong Bing, Piji Li, Yi Liao, Wai Lam, Weiwei Guo, and Rebecca Passonneau. 2015. Abstractive multi-document summarization via phrase selection and merging. In *Proceedings of the ACL-IJCNLP*, pages 1587–1597, Beijing, China, July. Association for Computational Linguistics.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2015. Gated feedback recurrent neural networks. *CoRR*, abs/1502.02367.

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

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

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.

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. In *Advances in Neural Information Processing Systems 27*, pages 2042–2050. Curran Associates, Inc.

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.

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

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.

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.*

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

Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, and Wojciech Zaremba. 2014. Addressing the rare word problem in neural machine translation. *CoRR*, abs/1410.8206.

Dipanjan Das Andr F.T. Martins. 2007. A survey on automatic text summarization. Technical report, C-MU.

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.

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

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

Mao Song Sun. 2011. Natural language procesing based on naturaly annotated web resources. *Journal of Chinese Information Processing*, 25(6):26–32.

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.

Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701.

Xiaoqiang Zhou, Baotian Hu, Qingcai Chen, Buzhou Tang, and Xiaolong Wang. 2015. Answer sequence learning with neural networks for answer selection in community question answering. In *Proceedings of the ACL-IJCNLP*, pages 713–718, Beijing, China, July. Association for Computational Linguistics.