CNewSum: A Large-Scale Summarization Dataset with Human-Annotated Adequacy and Deducibility Level

Danqing Wang[1], Jiaze Chen[1], Xianze Wu[1], Hao Zhou[1], and Lei Li[2(✉)]

[1] ByteDance AI Lab, Shanghai, China
{wangdanqing.122,chenjiaze,zhouhao.nlp}@bytedance.com
[2] Computer Science Department, University of California, Santa Barbara, Santa Barbara, USA
lilei@cs.ucsb.edu

Abstract. Automatic text summarization aims to produce a brief but crucial summary for the input documents. Both extractive and abstractive methods have witnessed great success in English datasets in recent years. However, there has been a minimal exploration of text summarization in other languages, limited by the lack of large-scale datasets. In this paper, we present a large-scale Chinese news summarization dataset CNewSum, which consists of 304,307 documents and human-written summaries for the news feed. It has long documents with high-abstractive summaries, which encourages document-level understanding and generation for current summarization models. An additional distinguishing feature of CNewSum is that its test set includes adequacy and deducibility annotations for the summaries. The adequacy level measures the degree of summary information covered by the document, and the deducibility indicates the reasoning ability the model needs to generate the summary. These annotations help researchers target their model performance bottleneck. We examine recent methods on CNewSum and will release our dataset after the anonymous period to provide a solid testbed for automatic Chinese summarization research.

Keywords: Automatic text summarization · Chinese summarization dataset · Adequacy and Deducibility

1 Introduction

Text summarization is an important task in natural language processing, which requires the system to understand the long document and generate a short text to summarize its main idea. There are two primary methods to generate summaries: *extractive* and *abstractive*. Extractive methods select semantic units from the source document and reorganize them into a consistent summary, while abstractive models generate summaries using words and phrases freely. Benefiting from pre-trained language models [2,10,14], much process has been made on English summarzation datasets, such as Newsromm [5], CNN/DailyMail [6], and NYT [19].

However, the lack of the high-quality datasets in other languages, such as Chinese, limits further researches on summarization under different language habits and cultural customs. It hinders the application of current summarization models to more languages. Currently, most Chinese summarization datasets are collected from Chinese

L. Wang et al. (Eds.): NLPCC 2021, LNAI 13028, pp. 389–400, 2021.
https://doi.org/10.1007/978-3-030-88480-2_31

Table 1. An example of our CNewSum dataset.

Article [0]图为在广元市朝天区发现的白耳夜鹭。[1]广林局供。[2]中新网广元3月15日电。[3]记者15日从四川省广元市野生动物救治中心获悉：近日，该市朝天区东溪河乡的群众发现一只受伤的“怪鸟”引起多方关注，随后，上报到广元市林业部门。[4]后经当地野生动物保护专家鉴定“怪鸟”为世界最濒危鸟类白耳夜鹭。..... [7]该鸟是我国特有的珍稀鸟类、国家二级保护动物白耳夜鹭，被列为世界最濒危的30种鸟类之一，目前全世界仅存1000余只..... [14]此后一直再没有关于该鸟踪迹的报道。 (*[0]The picture shows the Gorsachius magnificus in Chaotian District of Guangyuan City. [1]Supplied by Guangyuan Forestry Department. [2]Xinhua News Agency, Guangyuan, March 15. [3]Reporters learned from the Wildlife Treatment Center of Guangyuan City, Sichuan Province on the 15th that recently, the discovery of an injured "strange bird" by the local people in Dongxihe village, Chaotian District of the city attracted much attention and was subsequently reported to the forestry department of Guangyuan City. [4]After that, local wildlife protection experts identified the "strange bird" as the world's most endangered bird, the Gorsachius magnificus.....[7]It is a rare bird unique to our country and a national second-class protected animal, the Gorsachius magnificus. It has been listed as one of the world's most endangered 30 species of birds. At present, there are only about 1,000 birds in the world.....[14]There have been no reports of the bird's trace since then.*)
Summary 今日获悉，广元一市民发现受“怪鸟”，经鉴定系世界濒危鸟类白耳夜鹭，全球仅存1000只。 (*It was reported today that a citizen of Guangyuan found an injured "strange bird", which was identified as a world-endangered bird, the white-eared night heron, of which only 1,000 exist worldwide.*)
Sentence-Label:[0,4] **Adequacy**: 1 **Deducibility**: 1

social media Weibo, subject to a 140-word length limit [4,7]. There are also some datasets scraped from news websites, such as Toutiao [8] and ThePaper [12]. However, those datasets are either small-scale or not of high quality.

In this paper, we present a large-scale Chinese news summarization dataset, CNewSum, to make up for the lack of Chinese document-level summarization, which can become an important supplement to current Chinese understanding and generation tasks. Different from previous summarization datasets crawled from news websites, we called for news articles from over hundreds of thousands press publishers and hired a team of expert editors to provide human-written summaries for the daily news feed. During the summarization process, the editors may perform simple reasoning or add external knowledge to make the summary more reader-friendly. Thus, we further investigate our test set and explore how much knowledge the models need to generate a human-like summary. Specifically, we ask annotators to determine two questions: 1) **Adequacy**: *Is the information of summaries self-contained in the source document?* 2) **Deducibility**: *Can the information be deduced from the source document directly, or needs external knowledge?* We provide these two scores for each example in the test set. Table 1 is an example of our dataset.

Our main contribution are as follows:

(1) We propose a large-scale Chinese news summarization dataset collected from over hundreds of thousands news publishers. We hire a team of expert editors to write summaries for news feed.

(2) In order to figure out how much knowledge the model need to generate a human-like summary, we manually annotate the adequacy and deducibility level for our test set.
(3) We also provide several strong extractive and abstractive baselines, which makes the dataset easy to use as the benchmark for Chinese summarization tasks.

2 Related Work

News Summarization Dataset. Most news summarization datasets focus on English language, and here we give a brief introduction to some popular ones and list the detailed information in the first part of Table 2. NYT is a news summarization dataset constructed from New York Times Annotated Corpus [19]. We tokenize and convert all text to lower-case, follow the split of Paulus et al. [18]. The CNN/DailyMail question answering dataset [6] modified by Nallapati et al. [16] and See et al. [20] is the most commonly-used dataset for single-document summarization. It consists of online news articles with several highlights. Those highlights are concatenated as the summary. Newsroom [5] is a large-scale news dataset scraped from 38 major news publications, ranging from business to sports. These summaries are often provided by editors and journalists for social distribution and search results.

Chinese Summarization Dataset. There are also several Chinese summarization datasets in other domains [3,9,22], but here we only discuss news summarization datasets. The detailed statistics are listed in the second part of Table 2. The LCSTS [7] is a large-scale Chinese social media summarization dataset. It is split into three parts, and the part II and part III are usually used as development and test set after filtering out low-quality examples. RASG [4] collects the document-summary-comments pair data for their reader-aware abstractive summary generation task. It utilizes users' comments to benefit the generation of the abstractive summary of main content. The document is relatively short and has about 9 comments as a complement. TTNews [8] is provided for NLPCC Single Document Summarization competition,[1] including 50,000 training examples with summaries and 50,000 without summaries. CLTS [12] is a Chinese summarization dataset extracted from the news website ThePaper. It contains more than 180,000 long articles and corresponding summaries written by professional editors and authors.

3 The CNewSum Dataset

3.1 Data Collection

We receive news submissions from over hundreds of thousands press publishers.[2] We hire a team of expert editors to provide human-written summaries for the daily news

[1] http://tcci.ccf.org.cn/conference/2018/taskdata.php.

[2] The press publishers include thepaper.cn, wallstreetcn.com, cankaoxiaoxi.com, yicai.com, and so on. They submit their articles in web format to our company. These publishers retain any copyright they may have in their content and grant us a royalty-free, perpetual licence to use, copy, edit and publish their content.

feed. Each example will be double-checked by different experts to ensure its quality. We construct CNewSum by extracting news article from 2015 to 2020[3] and filtering summaries with less than 5 words. We further limit the length of documents to 50–5000. To solve the problem of missing and inaccurate punctuation in web format, we train an extra punctuation tagging model via Bi-LSTM on Chinese articles to correct punctuation.[4]

Finally we obtain a Chinese news corpus with 304,307 document-summary pairs. It is split into training/validation/test by 0.9/0.05/0.05. Besides, we compare document sentences with human-written summaries and use the greedy algorithm following [16] to get the ORACLE sentences with label 1 as the signal for extractive summarization.

Table 2. The summarization datasets. The top part contains the commonly-used English news summarization and the bottom contains the Chinese summarization datasets. '–' means the original dataset does not provide the standard spit for train/dev/test set. For TTNews, we only take training examples with summaries into consideration. '*' includes 2,000 evaluation examples for NLPCC2017 and 2,000 for NLPCC2018.

Dataset	Train	Dev	Test	Total	Article	Summary	Source
NYT [19]	589,282	32,737	32,739	654,758	552.14	42.77	New York Times
CNNDM [6]	287,227	13,368	11,490	312,085	791.67	55.17	CNN & Daily Mail
Newsroom [5]	995,041	108,837	108,862	1,212,740	765.59	30.22	38 news sites
LCSTS [7]	2,400,591	8,685	725	2,410,001	103.7	17.90	Weibo
RASG [4]	863,826	–	–	863,826	67.08	16.61	Weibo
TTNews [8]	50,000	–	4,000*	54,000	747.20	36.92	Toutiao
CLTS [12]	148,317	20,393	16,687	185,397	1363.69	58.12	ThePaper
CNewSum	275,596	14,356	14,355	304,307	790.55	37.58	News publishers

3.2 Adequacy and Deducibility Annotation

Analyzing our dataset, we find that the expert editors often perform some reasoning or add external knowledge to make the summary more friendly for the readers. For example, the precise figure (2,250) may be summarized as an approximate number (more than two thousand). In another case, a specific date will be converted to a relative time based on the time of publication, e.g. tomorrow. This information is not directly available in the original document. Thus, we wonder how much knowledge the model needs to generate the human-written summary. Inspired by [1], we ask annotators to answer the following two questions for each document-summary pair in our test set:

1) **Adequacy.** *Does necessary information of the summary has been included in the document?* For example, all words in the summary can be directly found in the document, or they have synonyms or detailed descriptions in the original text. Under these circumstances, the summary is labeled as 1.

[3] These data have been checked for legality and can be released for research use.
[4] The accuracy rate is 96.20%.

2) **Deducibility.** *Can the information of the summary be easily inferred from the document?* Unit conversion, number calculation, and name abbreviations that can be inferred are label as 1. In contrast, complex conclusions with no direct mentions in the original document are labeled as 0.

For each question, the annotators should choose 0 or 1. We hired a team of 12 employees to annotate the test set.[5] We first trained these employees on basic annotation rules, and they were required to annotate 100 examples and then be checked and corrected by us. Two voluntary expert annotators were employed to control quality. They were asked to sample 10% from each annotator and recheck the annotation. If one's consistent rate is less than 95%, all annotations of this annotator will be returned and re-annotated. It is consistent only if the two experts and the annotator agree on their answers, otherwise the example will be further discussed.

Table 3. The statistics of news summarization datasets. Cov., Den. and Comp. correspond to the *Coverage*, *Density* and *Compression* introduced by [5]. The Bi., Tri. and 4-gram are the n-gram novelty (%). The novelties of NYT/CNNDM/Newsroom are from [17]. For Chinese data, it is calculated by words.

Dataset	Cov.↓	Den.↓	Comp.↑	Bi.↑	Tri.↑	4-gram↑
NYT	0.83	3.50	24.19	55.59	71.93	80.16
CNNDM	0.85	3.70	13.76	49.70	70.20	79.99
Newsroom	0.82	9.50	36.03	46.80	58.06	62.72
LCSTS	0.54	1.23	6.61	80.29	90.92	94.53
RASG	0.61	2.52	7.27	67.89	76.94	80.15
TTNews	0.76	3.21	22.24	61.09	76.30	83.64
CLTS	0.99	28.73	24.81	5.14	8.08	10.36
CNewSum	0.76	2.77	20.83	63.29	78.54	85.64

3.3 Dataset Analysis

As shown in Table 2, our CNewSum dataset has a similar scale with the most popular English summarization dataset CNNDM, which is suitable for training and evaluating different summarization models. For the Chinese dataset, the average length of the document and the summary are significantly longer than datasets collected from Weibo and similar with TTNews.

Following Grusky et al. [5], we also use *Coverage*, *Density* and *Compression* to characterize our summarization dataset. *Coverage* measures the overlap degree of the extractive fragment between the article and summary, and *Density* measures the average length of the extractive fragment. *Compression* is the ratio of the article length to the summary length. In Addition, we calculate the n-gram novelty of the summary, which

[5] We paid 1 RMB (0.15 dollar) for each example, and the average hourly wage is 60 RMB (the minimum hourly wage is 24 RMB).

Table 4. The adequacy (A) and deducibility (D) level in our test set.

A = 1 & D = 1	A = 0 & D = 1	A = 0 & D = 0
91.08%	4.11%	4.81%

is the percentage of n-grams that do not appear in the document, as described in [17]. The results are shown in Table 3. We can find that the datasets collected from Weibo usually have lower coverage and density ratio, with high compression and novelty. This indicates that the summaries for these short documents are more abstractive. For news article summarization, CLTS almost copy most words of the summary from the document directly, which is indicated by the highest coverage, density and the lowest novelty. Our CNewSum provides a large-scale document-level summarization dataset with comparable abstractiveness with short social media datasets.

Since all adequacy summaries can be inferred from the document, the A = 1 & D = 0 is meaningless. For the summarization models, the examples with A = 1 & D = 1 is relatively easy to generate, and A = 0 & D = 1 ask for some inference abilities. The A = 0 & D = 0 cannot be solved with the original document and may need the help of external knowledge. From Table 4, we find that more than 80% examples are adequate and deducible, but 20% lack essential information. With $D = 1$, the information can be inferred from the document. For example, "2005–2015" will be summarized as "ten years" which requires the model to do simple calculation. The rest summaries are factual but need external knowledge. News articles from the websites are time-sensitive and are filled with pictures. The editors often write the summary based on the time of the event and the image, which will cause the relative time, such as 'yesterday', and the picture description to appear in the summary. In addition, famous people will be mapped to their position in the summary, such as Obama and the American president of that time. It is difficult for the model to deduce such information from the news text without additional information. We keep these in our dataset to simulate real-world data distribution and let researchers evaluate the model performance from different aspects.

4 Experiment

We train several summarization models on our CNewSum. These systems include both abstractive and extractive methods, and the performance can serve as the baseline for future work.

4.1 Models

Baseline. We calculate three popular summarization baseline for our dataset. LEAD is a common lower bound for news summarization dataset [5,16,20]. For ORACLE, we concatenate the sentences with label 1 in the original order. TextRank [15] is simple unsupervised graph-based extractive methods.

Table 5. Results on the test set of CNewSum. The first part contains the Lead and Oracle baseline. The second and third part are extractive and abstractive summarization models.

Models	ROUGE-1	ROUGE-2	ROUGE-L
LEAD	30.43	17.26	25.33
ORACLE	46.84	30.54	40.08
TextRank [15]	24.04	13.70	20.08
NeuSum [24]	30.61	17.36	25.66
TFExt [13]	32.87	18.85	27.59
BERTExt	34.78	20.33	29.34
PG [20]	25.70	11.05	19.62
TFAbs [13]	37.36	18.62	30.62
BERTAbs	**44.18**	**27.37**	**38.32**

Neural Models. NeuSum [24] jointly score and select sentences for extractive summarization. PG [20] is the pointer-generator network which is a commonly-used encoder-decoder abstractive summarization model with the copy and coverage mechanism. Transformer [21] is a well-known sequence-to-sequence model based on the self-attention mechanism. Following the settings in [13], we employ two Transformer baselines: TFExt and TFAbs. The pre-trained language models such as BERT [2] have improved both abstractive and extractive summarization by a large margin, so we also apply the BERTSum mode [13] to our dataset. We train a Chinese BERT language model with Chinese news articles,[6] which is noted as BERTExt and BERTAbs.

For extractive summarization, we choose the top-2 sentences as the summary due to the average sentence number (1.49) of the ground truth summary. The automatic metric ROUGE [11] is used for evaluation. Since the original ROUGE is made only for English, we follow the method of [7] and map the Chinese words to numbers. Specifically, the Chinese text is split by characters and the English words and numbers will be split by space. For example, "Surface Phone 将装载 Windows 10 (*The Surface Phone will be loaded with Windows 10*)" will be transformed to "surface/phone/将/装/载/windows/10" and then mapped to numeral IDs.

4.2 Results

As shown in Table 5, the abstractive models have better results on CNewSum test set, which is consistent with our analysis in Sect. 3.3. The abstractive methods has performed better than extractive models, which means that extractive methods have many performance limitations in CNewSum.

[6] Since the bert-base-chinese model of Google does not perform well in our dataset.

Table 6. The results of models on different adequacy and deducibility level.

Model	Category	ROUGE-1	ROUGE-2	ROUGE-L
TFExt	A = 1 & D = 1	33.16	19.19	27.88
	A = 0 & D = 1	30.89	15.60	25.38
	A = 0 & D = 0	28.92	14.88	23.74
TFAbs	A = 1 & D = 1	37.54	18.85	30.83
	A = 0 & D = 1	36.36	16.70	29.63
	A = 0 & D = 0	34.73	15.95	27.52
BERTExt	A = 1 & D = 1	35.05	20.67	29.62
	A = 0 & D = 1	32.81	16.90	27.05
	A = 0 & D = 0	31.07	16.57	25.72
BERTAbs	A = 1 & D = 1	44.51	27.76	38.70
	A = 0 & D = 1	41.75	23.64	35.34
	A = 0 & D = 0	40.18	23.34	33.60

We further evaluate models based on adequacy and deducibility level. The results shown in Table 6 indicate that this model performs well on A = 1 where all necessary information can be easily found in the source document. However, when it asks for simple deducing or external knowledge, the performance degrade significantly.

4.3 Case Study

We illustrate the differences between abstractive models with a typical example in the appendix. As stated in previous work [20,23], PG tends to copy directly from the original document instead of generating from vocabulary, which makes the output less abstractive. Besides, although it has used the coverage mechanism to avoid repetition, it still suffers the most from the meaningless duplication. For Transformer-based models, the random initialized model TFAbs introduces fake information, while the BERTAbs and TTBERTAbs perform much better in both capturing important information and generating fluent summaries.

Table 7. An example for abstractive summarization models. The text with underline is directly copied from the original article, and the bolded text contains fake information.

Article	英雄联盟神秘预告再现。官方最新发布了一个短片视频，其短片的名称是“他已归来”。而最近更新的巨神峰新故事中就有描述星灵的，难道新英雄是星灵来自银河？今日，国外的LOL官方社交媒体上，放出了一个预告短片，名称为“他已归来”。短片内容为，潘森正在凝视夜空中被星云所围绕的亮光。有人猜测，视频中的场景为潘森故事《巨神之枪》中的末尾内容，也是巨神峰新故事中所描述的《星灵》。歪果仁点评：Gigathor：天啊，下一个新英雄是银河系的！MrBananaHump：跟你们开玩笑呐，这只不过是巴德。SoSaysCory：应该是潘森的兄弟，潘林将会加入峡谷，技能与潘森一样，他们将会成为有史以来最强力的下路组合。Sharjo：将会有全新的巨神峰英雄了！潘森新的背景故事已提到了这个，在《巨神之枪》故事的结尾，指出了新的星灵到来。来自另一个次元的潘森老朋友将会和我们见面了！太酷了！DracCusS：感觉是：a)新英雄。b)潘森模型更新。c)宝石重做？ *League of Legends released a mysterious trailer and the official latest posted a short video. The name of the short film is “He Has Returned”. In the recent new story of Mount Titan, there is a description of the Protoss. Will the new hero be the Protos from the Milky Way? Today, a short trailer was released on the official social media of LOL abroad, titled “He Has Returned.” The content of the video is, Pan Sen stares at the bright light surrounded by nebula in the night sky. Some people guess, the scene in the video is the content of Pan Sen’s story “Gun of the Titans”, and the new story of Mount Titan is also described in the “Protoss”. Foreigns Review: Giant: Oh my god, the next new hero is from the galaxy! BananaHump: I’m joking with you, this happens to be Bud. SoSaysCory: It should be Pan Sen’s brother. Pan Lin will join the canyon. The skills are the same as Pan Sen. They will become the most concentrated bot combination ever. Sharjo: There will be a brand new Giant Mountain soon! Pan Sen’s new background hero story has already mentioned it. At the end of the ”Gun of the Gun” story, I saw the new Protoss. A friend of Pan Sen from another dimension will meet with us soon! so cool! DracCusS: It may be: a) New hero. b) Pan Sen model update. c) Gems reform?*
Gold	英雄联盟官方发布星空主题预告短片，或将推出全新巨神峰英雄 *League of Legends officially released a starry sky theme trailer, or will launch a new Titan Heroes*
PG	LOL官方发布了一个短片视频短片内容为，其短片内容为潘森正在凝视夜空中被星云所围绕亮光（组图） LOL officially released a short video , its short video content is that Pan Sen is gazing at the night sky surrounded by nebula light (Pictures)
TFAbs	英雄联盟官方发布新短片，**称将于明年1月发行**；英雄联盟新短片显示，**新将于9月下旬发布**。（组图） *The League of Legends officially released a new short film, **said it will be released in January next year**. The new League of Legends video shows that **new will be released in late September**. (Pictures)*
BERTAbs	英雄联盟公布新英雄预告：巨神峰新英雄是星灵来自银河？潘森新英雄将加入峡谷，宝石重做巨神之枪（组图） *League of Legends announced the new hero trailer: Is the new hero of Titan Peak from the Milky Way? Pan Sen’s new hero will join the canyon, and the gem will be remade the Titan’s Spear (Pictures)*

5 Conclusion

We present CNewSum, a high-quality summarization dataset composed of human-written summaries to fill up for the lack of news summarization dataset in Chinese. We annotate all test set with adequacy and deducibility levels to help abstractive models solve the unfaithful problem. Finally, we give several popular extractive and abstractive baselines on the dataset for future research.

References

1. Chen, D., Bolton, J., Manning, C.D.: A thorough examination of the CNN/Daily Mail reading comprehension task. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, pp. 2358–2367. Association for Computational Linguistics, August 2016. https://doi.org/10.18653/v1/P16-1223, https://www.aclweb.org/anthology/P16-1223
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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), Minneapolis, Minnesota, pp. 4171–4186. Association for Computational Linguistics, June 2019. https://doi.org/10.18653/v1/N19-1423, https://www.aclweb.org/anthology/N19-1423
3. Gao, S., Chen, X., Li, P., Chan, Z., Zhao, D., Yan, R.: How to write summaries with patterns? Learning towards abstractive summarization through prototype editing. 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), Hong Kong, China, pp. 3741–3751. Association for Computational Linguistics, November 2019. https://doi.org/10.18653/v1/D19-1388, https://www.aclweb.org/anthology/D19-1388
4. Gao, S., et al.: Abstractive text summarization by incorporating reader comments. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, 27 January–1 February 2019, pp. 6399–6406. AAAI Press (2019). https://doi.org/10.1609/aaai.v33i01.33016399, https://doi.org/10.1609/aaai.v33i01.33016399
5. Grusky, M., Naaman, M., Artzi, Y.: NEWSROOM: a dataset of 1.3 million summaries with diverse extractive strategies. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, Louisiana, pp. 708–719. Association for Computational Linguistics, June 2018. https://doi.org/10.18653/v1/N18-1065, https://www.aclweb.org/anthology/N18-1065
6. Hermann, K.M., et al.: Teaching machines to read and comprehend. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, Montreal, Quebec, Canada, 7–12 December 2015, pp. 1693–1701 (2015). https://proceedings.neurips.cc/paper/2015/hash/afdec7005cc9f14302cd0474fd0f3c96-Abstract.html
7. Hu, B., Chen, Q., Zhu, F.: LCSTS: a large scale Chinese short text summarization dataset. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, pp. 1967–1972. Association for Computational Linguistics, September 2015. https://doi.org/10.18653/v1/D15-1229, https://www.aclweb.org/anthology/D15-1229

8. Hua, L., Wan, X., Li, L.: Overview of the NLPCC 2017 shared task: single document summarization. In: Huang, X., Jiang, J., Zhao, D., Feng, Y., Hong, Yu. (eds.) NLPCC 2017. LNCS (LNAI), vol. 10619, pp. 942–947. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73618-1_84
9. Huang, K.H., Li, C., Chang, K.W.: Generating sports news from live commentary: a Chinese dataset for sports game summarization. In: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, Suzhou, China, pp. 609–615. Association for Computational Linguistics, December 2020. https://www.aclweb.org/anthology/2020.aacl-main.61
10. Lewis, M., et al.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7871–7880 (2020)
11. Lin, C.Y.: ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out (2004)
12. Liu, X., Zhang, C., Chen, X., Cao, Y., Li, J.: CLTS: a new Chinese long text summarization dataset. In: Zhu, X., Zhang, M., Hong, Yu., He, R. (eds.) NLPCC 2020. LNCS (LNAI), vol. 12430, pp. 531–542. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60450-9_42
13. Liu, Y., Lapata, M.: Text summarization with pretrained encoders. 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), Hong Kong, China, pp. 3730–3740. Association for Computational Linguistics, November 2019. https://doi.org/10.18653/v1/D19-1387, https://www.aclweb.org/anthology/D19-1387
14. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach (2019)
15. Mihalcea, R., Tarau, P.: TextRank: bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, pp. 404–411. Association for Computational Linguistics, July 2004, https://www.aclweb.org/anthology/W04-3252
16. Nallapati, R., Zhai, F., Zhou, B.: SummaRuNNer: a recurrent neural network based sequence model for extractive summarization of documents. In: Singh, S.P., Markovitch, S. (eds.) Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, California, USA, 4–9 February 2017, pp. 3075–3081. AAAI Press (2017). http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14636
17. Narayan, S., Cohen, S.B., Lapata, M.: What is this article about? Extreme summarization with topic-aware convolutional neural networks. J. Artif. Intell. Res. **66**, 243–278 (2019)
18. Paulus, R., Xiong, C., Socher, R.: A deep reinforced model for abstractive summarization. In: 6th International Conference on Learning Representations, ICLR 2018, Conference Track Proceedings, Vancouver, BC, Canada, 30 April–3 May 2018. OpenReview.net (2018). https://openreview.net/forum?id=HkAClQgA-
19. Sandhaus, E.: The New York times annotated corpus. In: Linguistic Data Consortium, Philadelphia, vol. 6, no. 12, p. e26752 (2008)
20. See, A., Liu, P.J., Manning, C.D.: 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), Vancouver, Canada, pp. 1073–1083. Association for Computational Linguistics, July 2017. https://doi.org/10.18653/v1/P17-1099, https://www.aclweb.org/anthology/P17-1099
21. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017, pp. 5998–6008 (2017). https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

22. Xi, X., Pi, Z., Zhou, G.: Global encoding for long Chinese text summarization. ACM Trans. Asian Low-Resour. Lang. Inf. Process. **19**(6), 1–17 (2020). https://doi.org/10.1145/3407911
23. Zhang, F., Yao, J.G., Yan, R.: On the abstractiveness of neural document summarization. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, pp. 785–790. Association for Computational Linguistics, October–November 2018. https://doi.org/10.18653/v1/D18-1089, https://www.aclweb.org/anthology/D18-1089
24. Zhou, Q., Yang, N., Wei, F., Huang, S., Zhou, M., Zhao, T.: Neural document summarization by jointly learning to score and select sentences. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, pp. 654–663. Association for Computational Linguistics, July 2018. https://doi.org/10.18653/v1/P18-1061, https://www.aclweb.org/anthology/P18-1061