CNewSum: A Large-scale Chinese News Summarization Dataset with Human-annotated Adequacy and Deducibility Level

Danqing Wang¹, Jiaze Chen¹, Xianze Wu¹, Hao Zhou¹ and Lei Li²*

¹ ByteDance AI Lab, Shanghai, China
{wangdanqing.122, chenjiaze, zhouhao.nlp}@bytedance.com
² Computer Science Department, University of California 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 Chinese, 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 can encourage document-level understanding and generation for current summarization models. An additional distinguishing feature of CNewSum is that its test set contains 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 can help researchers analyze and target their model performance bottleneck. We examine recent methods on CNewSum and release our dataset³ 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 methodology. 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 progress has been made on English summarization datasets, such as Newsroom [5], CNN/DailyMail [6], and NYT [19].

* Work is done while the corresponding author was at ByteDance.
³ It is available at https://dqwangl22.github.io/projects/CNewSum/
Weibo, which are limited to a 140-character length [4,7]. Some other datasets are scraped from news websites, such as Toutiao [8] and ThePaper [12]. However, those datasets are either small-scale or of low quality.

Currently, most Chinese summarization datasets are collected from Chinese social media Weib, which are limited to a 140-character length [4,7]. Some other datasets are scraped from news websites, such as Toutiao [8] and ThePaper [12]. However, those datasets are either small-scale or of low 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 hundreds of thousands of 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 signals for extractive models via the greedy algorithm. All information of the summaries self-contained in the source document, so its adequacy and deducibility level are 1.

Table 1. An example of our CNewSum dataset. ‘Sentence Label’ is the id of sentences selected as the supervised signals for extractive models via the greedy algorithm. All information of the summary can be found in the document, so its adequacy and deducibility level are 1.

<table>
<thead>
<tr>
<th>Article</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,4)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 &quot;strange bird&quot; 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 &quot;strange bird&quot; 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.</td>
<td></td>
</tr>
<tr>
<td>(15)今日获悉，广元一市民发现受“怪鸟”，被鉴定为世界濒危鸟类白耳夜鹭，全球仅存1000余只。</td>
<td></td>
</tr>
<tr>
<td>(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.)</td>
<td></td>
</tr>
</tbody>
</table>

Sentence Label: {0,4} Adequacy: 1 Deducibility: 1
knowledge? We provide these two scores for each example in the test set. Table 1 is an example of our dataset.

Our main contributions are as follows:

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

(2) In order to figure out how much knowledge the model needs to generate a human-like summary, we manually annotate the adequacy and deducibility scores for our test set.

(3) We also provide several 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, 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 lowercase, 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 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 [4], 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 summaries written by editors of the website.

3 The CNewSum Dataset

3.1 Data Collection

We receive news submissions from hundreds of thousands of press publishers. These articles do not have corresponding summaries, so we hire a team of expert editors to provide human-written summaries for the daily news feed. Each example will be double-checked by different experts to ensure its quality. We construct CNewSum by extracting news articles from 2015 to 2020 and filtering summaries with less than 5 words. We further limit the length of documents to 50-5000.

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 signals 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. ‘Article’ and ‘Summary’ are the average length of the articles and summaries in each dataset. For English datasets, it is calculated by words and for Chinese, it is calculated by characters. ‘-’ means the original dataset does not provide the standard split 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.

<table>
<thead>
<tr>
<th>Dataet</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Total</th>
<th>Article</th>
<th>Summary</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>NYT [19]</td>
<td>589,282</td>
<td>32,737</td>
<td>32,739</td>
<td>654,758</td>
<td>552.14</td>
<td>42.77</td>
<td>New York Times</td>
</tr>
<tr>
<td>Newsroom</td>
<td>995,041</td>
<td>108,837</td>
<td>108,862</td>
<td>1,212,740</td>
<td>765.59</td>
<td>30.22</td>
<td>38 news sites</td>
</tr>
<tr>
<td>LCSTS [7]</td>
<td>2,400,591</td>
<td>8,685</td>
<td>725</td>
<td>2,410,001</td>
<td>103.7</td>
<td>17.90</td>
<td>Weibo</td>
</tr>
<tr>
<td>RASG [4]</td>
<td>863,826</td>
<td>-</td>
<td>-</td>
<td>863,826</td>
<td>67.08</td>
<td>16.61</td>
<td>Weibo</td>
</tr>
<tr>
<td>TTNews</td>
<td>50,000</td>
<td>-</td>
<td>-</td>
<td>50,000</td>
<td>747.20</td>
<td>36.92</td>
<td>Toutiao</td>
</tr>
<tr>
<td>CLTS [12]</td>
<td>148,317</td>
<td>20,393</td>
<td>16,687</td>
<td>185,397</td>
<td>1363.69</td>
<td>58.12</td>
<td>ThePaper</td>
</tr>
<tr>
<td>CNewSum</td>
<td>275,596</td>
<td>14,356</td>
<td>14,355</td>
<td>304,307</td>
<td>730.42</td>
<td>35.12</td>
<td>News publishers</td>
</tr>
</tbody>
</table>

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, a precise figure (2,250) may be summarized as an approximate number (more

---

5 The press publishers include thepaper.cn, wallstreeetcn.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 license to use, copy, edit and publish their content.

6 These data have been checked for legality and can be released for research use.
than 2000). 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-like summary. Inspired by [1], we ask annotators to answer the 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. Otherwise, the summary is labeled as
0.

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 labeled 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. 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 expert annotators were employed to control quality. They were asked to
sample 10% examples 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.
An example 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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Coverage</th>
<th>Density</th>
<th>Compression</th>
<th>Bigram</th>
<th>Trigram</th>
<th>4-gram</th>
</tr>
</thead>
<tbody>
<tr>
<td>NYT</td>
<td>0.83</td>
<td>3.50</td>
<td>24.19</td>
<td>55.59</td>
<td>71.93</td>
<td>80.16</td>
</tr>
<tr>
<td>CNNDM</td>
<td>0.85</td>
<td>3.70</td>
<td>13.76</td>
<td>49.70</td>
<td>70.20</td>
<td>79.99</td>
</tr>
<tr>
<td>Newsroom</td>
<td>0.82</td>
<td>9.50</td>
<td>36.03</td>
<td>46.80</td>
<td>58.06</td>
<td>62.72</td>
</tr>
<tr>
<td>LCSTS</td>
<td>0.54</td>
<td>1.23</td>
<td>6.61</td>
<td>80.29</td>
<td>90.92</td>
<td>94.53</td>
</tr>
<tr>
<td>RASG</td>
<td>0.61</td>
<td>2.52</td>
<td>7.27</td>
<td>67.89</td>
<td>76.94</td>
<td>80.15</td>
</tr>
<tr>
<td>TTNews</td>
<td>0.76</td>
<td>3.21</td>
<td>22.24</td>
<td>61.09</td>
<td>76.30</td>
<td>83.64</td>
</tr>
<tr>
<td>CLTS</td>
<td>0.99</td>
<td>28.73</td>
<td>24.81</td>
<td>5.14</td>
<td>8.08</td>
<td>10.36</td>
</tr>
<tr>
<td>CNewSum</td>
<td>0.76</td>
<td>2.72</td>
<td>20.85</td>
<td>63.85</td>
<td>78.95</td>
<td>85.97</td>
</tr>
</tbody>
</table>

### 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 to 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 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 copies most words of the summary from the document directly,
which is indicated by the highest coverage, density and the lowest novelty. Our CNew-
Sum 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 are
relatively easy to generate, and the examples with 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.

We find that more than 81.8% examples are adequate and deducible, but the rest lack
essential information. For 7.3% examples with $D = 1$, the information can be inferred
from the document. Typically, “2005-2015” will be summarized as “ten years” which
requires the model to do simple calculations. 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 two popular summarization baselines for our dataset. LEAD is a
common lower bound for news summarization dataset [5, 16, 20], which selects the first
several sentences as the summary. Here, we choose the first two sentences. For ORACLE,
we concatenate the sentences with label 1 with their original order in the document.
Table 4. 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.

<table>
<thead>
<tr>
<th>Models</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEAD</td>
<td>30.12</td>
<td>17.28</td>
<td>25.23</td>
</tr>
<tr>
<td>ORACLE</td>
<td>46.51</td>
<td>30.47</td>
<td>39.99</td>
</tr>
<tr>
<td>Transformer-ext</td>
<td>34.89</td>
<td>19.90</td>
<td>28.60</td>
</tr>
<tr>
<td>BERT-ext</td>
<td>36.49</td>
<td>21.33</td>
<td>30.21</td>
</tr>
<tr>
<td>Pointer Generator [20]</td>
<td>37.60</td>
<td>22.00</td>
<td>31.56</td>
</tr>
<tr>
<td>Transformer-abs</td>
<td>35.30</td>
<td>16.69</td>
<td>28.31</td>
</tr>
<tr>
<td>BERT-abs</td>
<td><strong>41.69</strong></td>
<td><strong>24.70</strong></td>
<td><strong>35.05</strong></td>
</tr>
</tbody>
</table>

**Extractive Models**  TextRank [15] is a simple unsupervised graph-based extractive method. It takes sentences as nodes and calculates the node importance based on eigenvector centrality. NeuSum [24] jointly scores and selects sentences for extractive summarization. Transformer [21] is a well-known sequence-to-sequence model based on the self-attention mechanism, the pre-trained language models such as BERT [2] trained on large corpus have shown great performance. We use the code provided by BERT-Sum [13] and follow the experimental settings to apply the Transformer and BERT to extractive summarization, which are named Transformer-ext and BERT-ext. Both of them use a 6-layer Transformer with hidden size 768 and feed-forward filter size 2048 as the document encoder. The sigmoid layer is put on the top to score the sentences. We choose the top sentence as the summary due to the average sentence number (1.03) of the ground truth summary.

**Abstractive Models**  Pointer Generator [20] is the pointer-generator network which is a commonly-used encoder-decoder abstractive summarization model with the copy and coverage mechanism. We also use the Transformer encoder and decoder for abstractive summarization. They are called Transformer-abs and BERT-abs to distinguish them from the above extractive models. These Transformer-based abstractive models use the same transformer encoder as the extractive ones and a transformer decoder with 6 layers for generation.

4.2 Results

Since the original summarization metric ROUGE [11] 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

---

8 Since the bert-base-chinese model of Google does not perform well in our dataset, we train a Chinese BERT language model with Chinese news articles.

9 https://github.com/nlpyang/PreSumm
Table 5. The results of models on different adequacy and deducibility level.

<table>
<thead>
<tr>
<th>Model</th>
<th>Category</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pointer Generator</td>
<td>A=1&amp;D=1</td>
<td>38.17</td>
<td>21.92</td>
<td>38.16</td>
</tr>
<tr>
<td></td>
<td>A=0&amp;D=1</td>
<td>36.03</td>
<td>17.38</td>
<td>36.02</td>
</tr>
<tr>
<td></td>
<td>A=0&amp;D=0</td>
<td>32.61</td>
<td>17.38</td>
<td>36.02</td>
</tr>
<tr>
<td>Transformer-abs</td>
<td>A=1&amp;D=1</td>
<td>37.80</td>
<td>17.88</td>
<td>31.10</td>
</tr>
<tr>
<td></td>
<td>A=0&amp;D=1</td>
<td>36.83</td>
<td>14.91</td>
<td>29.44</td>
</tr>
<tr>
<td></td>
<td>A=0&amp;D=0</td>
<td>34.26</td>
<td>14.28</td>
<td>27.61</td>
</tr>
<tr>
<td>BERT-abs</td>
<td>A=1&amp;D=1</td>
<td>44.32</td>
<td>26.27</td>
<td>38.00</td>
</tr>
<tr>
<td></td>
<td>A=0&amp;D=1</td>
<td>42.25</td>
<td>21.44</td>
<td>35.17</td>
</tr>
<tr>
<td></td>
<td>A=0&amp;D=0</td>
<td>38.70</td>
<td>20.26</td>
<td>32.31</td>
</tr>
</tbody>
</table>

loaded with Windows 10” will be transformed to “surface/phone/将/装/载/windows/10” and then mapped to numeral IDs.

As shown in Table 4, the abstractive models have better results on CNewSum test set, which is consistent with our analysis in Section 3.3. The simple abstractive baseline, pointer generator, has performed better than BERT-based extractive models, which means that extractive methods have many performance limitations in CNewSum.

We further evaluate abstractive models based on adequacy and deducibility level. The results shown in Table 5 indicate that this model performs well on examples with A=1 where all necessary information can easily be found in the source document. However, on examples that ask for simple deducing or external knowledge, the performance degrades significantly.

4.3 Case study

We illustrate the differences between abstractive models with a typical example in Table 6. As stated in previous work [20,23], the pointer generator 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 meaningless duplication. For Transformer-based models, the random initialized model Transformer-abs introduces fake information, while the BERT-abs performs much better in both capturing important information and generating fluent summaries.

5 Conclusion

We present CNewSum, a high-quality summarization dataset composed of human-written summaries to fill up the lack of news summarization dataset in Chinese. We annotate all test set with adequacy and deducibility scores to help abstractive models figure out how to generate a more human-friendly summary. Finally, we report results of several popular extractive and abstractive baselines on the dataset for future research.
Table 6. An example for abstractive summarization models. The text with wavy underline is directly copied from the original article, and the text with wavy underline contains fake information.

| Article | English: Chinese secrets are revealed. The latest short video released by the official media of League of Legends shows that Pan Sen is joining the canyon. The new hero is the Titan Peak, a new Giant Mountain will be released on the official social media of LOl abroad, titled "He Has Returned." The name of the short film is "Gun of the Titans." The events in the video are the content of Pan Sen's story "Gun of the Titans," and the new story of Mount Titan is also described in the "Protoss." 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." The content of the video is, Pan Sen staries at the bright light surrounded by nebula in the night sky.

| Gold | English: League of Legends officially 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 Protoss 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 staries at the bright light surrounded by nebula in the night sky.

| Pointer Generator | English: League of Legends officially released a mysterious trailer, or will launch a new Titan Heroes.

| Transformer-abs | English: 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.

| BERT-abs | English: 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.
Acknowledgments

The authors would like to thank Huiying Lin and many language annotators for help on preparing the data. Lei Li is not supported by any funding during this project.

References


