Overview of the NLPCC 2017 Shared Task: Single Document Summarization

Lifeng Hua[1], Xiaojun Wan[2], Lei Li[1]
[1] Toutiao AI Lab
[2] Institute of Computer Science and Technology, Peking University
{hualifeng, lileilab}@bytedance.com
wanxiaojun@pku.edu.cn

Abstract

In this paper, we give an overview for the shared task at the 6th CCF Conference on Natural Language Processing & Chinese Computing (NLPCC 2017): single document summarization. Document summarization aims at conveying important information and generating significantly short summaries for original long documents. This task focused on summarizing the news articles and released a large corpus, TTNews corpus[1], which was collected for single document summarization in Chinese. In this paper, we will introduce the task, the corpus, the participating teams and the evaluation results.

Keywords: Single Document Summarization, NLPCC 2017, TTNews Corpus

1 Introduction

Document summarization has been an important role in today's fast-grow information time. Now, the Internet products tens of millions of documents everyday and it is impossible for human being to manually summarize them, or even though read them. So the technology of automatic document summarization is necessary for us to obtain and reorganize the information from the Internet.

The methods of document summarization can be defined as extractive and abstractive summarization[1]. The extractive summarization attempts to extract key sentences or key phrases from the original document and then reorders these fragments into a summary. Meanwhile the abstractive summarization focused on generating new fragments and new expressions which are based on the understanding of this document.

Additionally, the document summary can be produced from a single document or multiple documents[2]. In this shared task, we just focus on single document summarization.

[1] TTNews corpus can be downloaded at https://pan.baidu.com/s/1bppQ4z1

2 Task

Traditional news article summarization techniques have been widely explored on the DUC and TAC conferences, and existing corpora for document summarization are mainly focused on western languages, while Chinese news summarization has seldom been explored. In this shared task, we aim to investigate single document summarization techniques for Chinese news articles. It is defined as a task of automatically generating a short summary for a given Chinese news article.

We will provide a large corpus for evaluating and comparing different document summarization techniques. This corpus has a test/training set consisting of a large number of Chinese news articles with reference summaries, together with a large number of news articles without reference summaries (perhaps for semi-supervised methods). Almost these news articles and reference summaries are used for news browsing and propagation at Toutiao.com.

3 Data

TTNews corpus contains test set and training set. For the training set, it contains a large set of news articles browsed on Toutiao.com and corresponding human-written summary which was used on news pushing and other tasks on Toutiao.com. Furthermore it contains another large set of news articles without summary (perhaps for semi-supervised methods). For the test set, it just contains the news articles. The news articles are from lots of different sources and meanwhile contain of different topics, such as sports, foods, entertainments, politics, technology, finance and so on.

As far as we know, TTNews corpus is the largest single document summarization corpus in Chinese. There are 50,000 news articles with summary and 50,000 news articles without summary in training set, and 2000 news articles in test set. As shown in Table 1, the mean length of the short summary is 45 Chinese characters

The example of a news article and its reference summary is shown in Table 2.

Table 1: Statistical information of TTNews corpus

	Number	Mean length of news article	Mean length of news summary
Training(with summary)	50000	1036	45
Training(without summary)	50000	1526	/
Test	2000	1037	45

Table 2: An example of news article and reference summary

News summary:
韩媒称朝鲜外交官在莫桑比克走私犀牛角被抓，将犀牛角放在外交邮袋中运往中国，再在黑市上销售，以赚取外汇
News article:
参考消息网5月30日报道美国之音广播援引韩国驻南非大使馆相关负责人的话报道称，朝鲜外交官被爆在非洲莫桑比克走私濒临灭种野生动物犀牛角。据韩国《中央日报》5月29日报道，大使馆相关负责人匿名表示，“被曝光的人是朝鲜驻南非大使馆参赞朴哲俊（音）和居住在南非的朝鲜跆拳道教练金钟秀”，“(当地)这是从马普托警察厅的奥兰度·木杜马尼（音）发言人那里确认的事实”。据当地警察厅称，朴参赞和金教练5月3日在莫桑比克马普托中部的市场上从当地偷猎者那里购买了4.616公斤犀牛角，在用车辆运输的过程中被抓获。据悉，当地警察厅接到举报出动，将这些人当场抓获。关于濒临灭绝动植物交易的国际公约禁止进行犀牛角相关商业交易。即使以学术研究目的进行国家间贸易，也要出示两国政府颁发的进出口许可证。美国之音报道称，朝鲜不顾公约，利用外交官特权从事犀牛角走私。外交官在通过国境时免搜查，外交邮袋也未经外交官负责人同意不能检查。韩国驻南非大使馆相关负责人表示，莫桑比克是犀牛栖息地，朝鲜驻莫桑比克保健代表部多次走私犀牛角，交给朝鲜驻南非大使馆，大使馆将其放在外交邮袋中运往中国。驻中国的朝鲜相关负责人收到走私货物后在黑市上作为中药材销售。《国家地理》今年3月报道称，犀牛角作为以真犀角为名的中药材在黑市上以每公斤6.5万美元左右的价格进行交易。报道称，朝鲜利用外交官特权赚取外汇的现象最近逐渐频繁起来。上个月，在禁止销售酒水的巴基斯坦，朝鲜外交官夫妇在大街上无执照销售芝华士（Chivas Regal）等洋酒被揭发。今年3月，在孟加拉国，朝鲜外交官拿着装有27公斤金块的行李在入境时被海关查获。报道称，庆南大学教授林乙出（音，朝鲜学系）表示，“因联合国、美国和韩国的对朝制裁，资金来源被切断的朝鲜当局施压要赚取外汇，并指示大使馆运营费也要自行筹措”。报道说，还有分析称，部分外交官尝到了“钱的滋味”，主动犯下这种罪行。东国大学教授（朝鲜学系）金榕炫分析称，“因上级指示而放手赚取外汇的外交官们知道通过交易能够创收的方法，因此表现得更积极”...

4 Participants

Each team was allowed to submit at most 5 runs of results in the period of this shared task. The participants were allowed to use any NLP resources and toolkits, but not allowed to use any other news articles with reference summaries.

There were 9 teams submitting their final results in this shared task. The participating teams are shown in Table 3. And they totally submitted 29 runs of result for validation. Both extractive summarization and abstractive summarization was used by the participating teams.

Table 3: Introduction of participating teams

Team Name	Organization Name
NLP_ONE	Central China Normal University
ICDD_Mango	Beijing Information Science And Technology University
NLP@WUST	Wuhan University of Science and Technology
CQUT_AC326	Chongqing University of Technology
HIT_ITNLP_TS	Harbin Institute of Technology
DLUT_NLPer	Dalian University of Technology
AC_Team	Chongqing University of Technology
ECNU_BUAA	Beihang University, East China Normal University
ccnuSYS	Central China Normal University

5 Evaluation

The single document summarization was evaluated automatically.

5.1 Evaluation Metric

We used ROUGE[3] for automatic evaluation metric. ROUGE is the shorthand of Recall-Oriented Understudy for Gisting Evaluation, and contains a set of metrics used for automatic document summarization, machine translation evaluation and other tasks in NLP. We defined the mean value of ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-L, ROUGE-SU4, ROUGE-W-1.2 scores as the overall evaluation score. And we used ROUGE-1.5.5 toolkit to compute the overall score. Note that the length of each summary was limited to 60 Chinese characters at our shared task, so we used -l 60 for truncating longer news summary.

5.2 Results

There are 9 submitted teams in this shared task, and the results are shown in Table 4. As Table 4 given, NLP_ONE, ICCD_Mango and NLP@WUST have better results than others.

5.3 Some Representative Systems

In this section, some representative systems will be brief introduced.

LEAD system is a extractive summarization baseline system. It tasks the first 60 characters one by one from the document as a summary.

ccnuSYS system uses an LSTM encoder-decoder architecture[4] with attention mechanism[5] to generate abstractive summary[6, 7] for this shared task. It uses the article as input sequence and the summary as output sequence.

Table 4: Evaluation results

	ROUGE-2	ROUGE-4	ROUGE-SU4	Overall Score
NLP_ONE	22.89	12.81	21.24	22.10
ICDD_Mango	23.82	12.04	21.19	22.09
NLP@WUST	22.53	10.39	20.81	21.65
CQUT_AC326	19.62	7.83	18.12	19.14
HIT_ITNLP_TS	19.33	8.38	17.86	19.13
DLUT_NLPer	17.64	7.58	16.35	17.54
AC_Team	18.16	7.88	15.92	17.09
ECNU_BUAA	15.73	6.86	14.72	15.99
ccnuSYS	15.58	6.57	14.47	15.79
LEAD	20.91	11.75	19.28	20.31

NLP_ONE system is also focused on abstractive summarization. Due to the shortcoming of traditional attention encoder-decoder models, this work proposes to add an new attention mechanism on output sequence and uses the subword method. And it gets a significant improvement.

NLP@WUST system uses an feature engineering based sentence extraction framework to get extractive summary for this shared task. After the extraction processing, it adds an sentence compression algorithm[8] for compressing shorter summary. And the performance is further improved.

6 Conclusion

This paper briefly introduces the overview of single document summarization shared task at NLPCC 2017. There are 9 participants having submitted final results. And some participants get exciting results in this corpus. Meanwhile, we release a large Chinese news articles and reference summaries corpus (TTNews corpus) for more large-scale research in Chinese document summarization.

Acknowledgement

We are very grateful to the colleagues from our company for their efforts to annotate the data and Knowledge Engineering Group of Tsinghua University for the help of building the submission system. And we also would like to thank the participants for their valuable feedback and results.

References

[1] Ani Nenkova and Kathleen McKeown. A survey of text summarization techniques. *Mining text data*, pages 43–76, 2012.

[2] Dipanjan Das and André FT Martins. A survey on automatic text summarization. *Literature Survey for the Language and Statistics II course at CMU*, 4:192–195, 2007.

[3] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain, 2004.

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

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

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

[7] Jianpeng Cheng and Mirella Lapata. Neural summarization by extracting sentences and words. *arXiv preprint arXiv:1603.07252*, 2016.

[8] Trevor Anthony Cohn and Mirella Lapata. Sentence compression as tree transduction. *Journal of Artificial Intelligence Research*, 34:637–674, 2009.