文章编号：2096-1472(2022)-05-01-03
DOI:10.19644/j.cnki.issn2096-1472.2022.005.001

基于BERT的抽取式裁判文书摘要生成方法研究

魏鑫炀[1]，唐向红[1,2]

(1.贵州大学计算机科学与技术学院，贵州 贵阳 550025；
2.贵州大学贵州省公共大数据重点实验室，贵州 贵阳 550025)
✉604564607@qq.com; xhtang@gzu.edu.cn

摘 要：针对民事裁判文书区别于新闻文本的文本结构和重要信息分布的特点，基于BERT提出了一种结合粗粒度和细粒度抽取方法的结构化民事裁判文书摘要生成方法。首先通过粗粒度抽取方法对裁判文书进行重要的模块信息抽取，以保留文本结构；然后采用基于BERT的序列标注方法构建细粒度的抽取式摘要模型，从句子级别对重要模块的信息进行进一步抽取，以构建最终摘要。实验表明，相比于单一的粗粒度抽取或者细粒度抽取，本文方法均获得了更好的摘要生成性能。

关键词：司法领域；裁判文书；抽取式文本摘要；序列标注
中图分类号：TP399 **文献标识码**：A

Research on Extractive Judgment Document Abstract Generation Method based on BERT

WEI Xinyang[1], TANG Xianghong[1,2]

(1.*College of Computer Science and Technology, Guizhou University*, *Guiyang* 550025, *China*;
2.*Guizhou Provincial Key Laboratory of Public Big Data*, *Guizhou University*, *Guiyang* 550025, *China*)
✉604564607@qq.com; xhtang@gzu.edu.cn

Abstract: Aiming at the text structure and important information distribution features of civil judgment documents that are different from news texts, this paper proposes a structured civil judgment document abstract generation method based on BERT (Bidirectional Encoder Representation from Transformers), combining coarse-grained and fine-grained extraction methods. Firstly, important module information is extracted from the judgment documents by the coarse-grained extraction method to preserve the text structure. Then the BERT-based sequence labeling method is used to build a fine-grained extractive abstract model. Information of important modules is further extracted based on the sentence level, so to construct the final abstract. Experiments show that the proposed method has better abstract generation performance than single coarse-grained extraction or fine-grained extraction.

Keywords: judicial field; judgment documents; extractive text abstract; sequence annotation

1 引言(Introduction)

随着国家依法治国的全面推进和互联网、大数据的快速发展，我国各级政府也在积极推进大数据、人工智能与法院司法实践的融合。裁判文书是记录人民法院审理过程和裁判结果的法律文书，是非常重要的司法文本，其中又以民事裁判文书占比最大，最为繁杂。对民事裁判文书进行摘要生成，可以帮助法官、律师及当事人等迅速、有效地简要了解案件审判过程与结果，从而快速找到相关的指导性案例，对我国的司法智能化辅助审判建设也具有重要的现实意义与应用价值。

文本摘要工作旨在从一篇或多篇相同主题的文本中抽取能够反映主题的精简压缩版本[1−2]，解决采用人工进行裁判文书摘要总结导致人力成本高，以及相关司法领域专家缺乏等问题。随着计算机技术的发展和自然语言理解，以及数字人文研究的不断深入，司法领域的自动化文本摘要任务研究逐渐成为一个重要的研究内容。自动文本摘要技术典型的应用场景包括将新闻、社会化短文本和专业领域文献等文本自动生成简短摘要[3]。对于这些应用场景，诸多研究人员已经提出了许多准确且高效的摘要算法，这些算法可以分为抽取式(Extractive)和生成式(Abstractive)。由于抽取式文摘是通过对原文的抽取来形成摘要的，可以保证得到的摘要在语法和事实上的正确性。而法律文本对内容的准确度要求较高，因此本文在探索民事裁判文书的摘要方法时，主要聚焦于抽取式文本摘要方法。

抽取式摘要技术最早能追溯到1958 年LUHN[4]提出的word_significance算法，它基于词频抽取重要的句子组成摘要；2003 年，PADMALAHARI等人[5]提出text_pronouns算法，通过构建句子和词语级别的文本特征来进行抽取式自动

文摘；此外，MIHALCEA等人[6]提出了Textrank算法，利用投票机制对构建好的图模型中的重要成分进行排序来抽取重要句子构成摘要。随着机器学习的发展，许多研究人员也将机器学习算法运用到文本摘要任务中。NALLAPATI等人[7]提出SummaRuNNer方法，LIU[8]提出Bertsum方法，它们均是采用序列标注方法，构建训练模型来抽取重要的句子，相对于传统的统计方法，能够从语义上对生成文摘的质量有较好的把控，并且使新闻数据集具有不错的性能。但是，由于民事裁判文书不同于新闻文本，因此传统的文本摘要方法直接用于民事裁判文书的摘要生成并不能取得很好的效果。

民事裁判文书一般包含首部、事实与理由、裁判依据、裁判主文等部分，如图1所示。民事裁判文书具有层次结构性，文本较长，并且重要信息分散在整个文本中，许多重要信息可能只会在文本中出现一次；而新闻文本的重要信息则主要集中在首部，重要的信息可能会在文中反复出现，并且新闻文本一般不具有特定的结构，因此不能直接使用主流的抽取式文摘方法来进行民事裁判文书的摘要生成。本文根据民事裁判文书的特点，基于BERT提出了一种结合粗粒度和细粒度抽取方法的结构化民事裁判文书摘要生成方法。实验证明，本文提出的方法能够获得较好的性能提升。

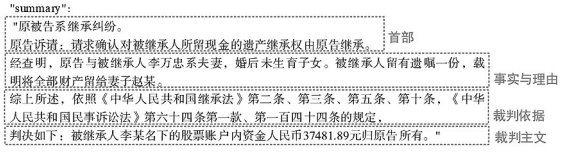

图1 民事裁判文书的结构

Fig.1 Structure of civil judgment documents

2 相关工作研究(Related work research)

2.1 BERT预训练模型

BERT(Bidirectional Encoder Representation from Transformers)是KENTON等人[9]提出的一种预训练模型，其模型结构如图2所示。预训练模型就是通过大量数据训练所得到的一种泛化性很强的模型，在有具体任务需要时，预训练模型可以只使用少量的数据进行一个增量训练，而不再需要大量语料进行从头训练，节约了时间，提高了效率。BERT使用Transformer[10]的编码结构，利用注意力机制对句子进行编码建模，这使得它可以更好地提取上下文语义特征，可以获得更好的句子语义表示，这种语义表示主要是基于“掩模语言模型(Masked Language Model，MLM)”和“下一个句子预测(Next Sentence Prediction，NSP)”的多任务训练来进行预训练提取得到。这样的BERT预训练模型便能用于其他特定任务，将其结合到相应模型的输出层前部，通过微调构建多种相应的任务模型，但是BERT的使用容易受到文本长度的限制，其对于长文本并不能很好地进行处理。

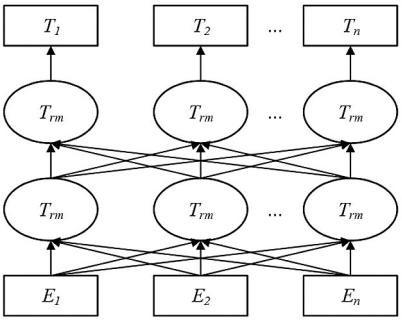

图2 BERT预训练语言模型

Fig.2 BERT pre-trained language models

将BERT用于民事裁判文书摘要生成任务，能够从细粒度方面，如句子之间甚至词之间，更好地获得文本数据中词与词，以及句子与句子之间的语义表示，在构建模型进行训练的过程中，利用这样的语义表示可以提高模型对裁判文书文本语义的捕捉效果。

2.2 序列标注模型

序列标注是指将输入的一串观测序列$x_1x_2x_3...x_n$转化为一串标记序列的$y_1y_2y_3...y_n$过程。要解决序列标记问题，实际上就是要找到一个观测序列到标记序列的映射$f(x_i) \rightarrow y_i (i = 1,2,3,...,n)$。

在自然语言处理中，序列标注模型是最常见的模型，应用广泛，许多自然语言处理的问题都可以想方设法转化为序列标注问题。与一般分类任务不同的是，序列标注模型输出的是一个标签序列。一般情况下，这样的标签之间是有相互关系的，这样的序列构成了标签之间的结构信息。

抽取式摘要同样可以建模为序列标注任务。首先，为原文中的每一个句子添加一个二分类标签(0或1)，0表示该句不属于摘要，1表示该句属于摘要；之后构建相应的序列标注模型进行训练；最终得到的摘要由所有标签为1的句子产生。将文本摘要建模为序列标注任务的关键在于获得句子的表示，即将句子编码为一个向量，根据该向量进行二分类任务。例如SummaRuNNer模型，该模型分别构建词语级别和句子级别的语义向量表示，之后使用得到的句子向量表示进行分类任务训练，以此来构建摘要抽取模型。这同样也是从句子级别的细粒度角度来对文本进行摘要抽取。

2.3 基于BERT的结构化民事裁判文书摘要生成方法

抽取式文本摘要方法主要为建模成序列标注方法，从给定的文本中将适合的句子挑选出来构成文本摘要，其中关键的部分便是获得句子的表示。而BERT则可以很好地将文本中句子的语义信息表示出来，通过“BERT+序列标注”的方式，能够从句子级别对裁判文书进行一个细粒度的抽取，但是由于民事裁判文书过长，冗余信息过多，导致BERT不能很好地处理。如果直接简单截取原文前半部分、后半部分或者中间部分，再进行BERT处理，则会导致裁判文书的文本结构以及重要信息丢失，不能取得很好的效果。而如果只采取粗粒度的抽取方式，虽然能够保留裁判文书的结构信息，却无法做到精细化抽取，会包含大量冗余信息，或者丢失重要信息，从而导致生成的摘要性能降低。

因此，我们提出了一种结合粗粒度和细粒度抽取方法的结构化民事裁判文书摘要生成方法，即先对民事裁判文书进行一个粗粒度的抽取，通过规则将裁判文书中几个主要部分的内容抽取出来，在缩短文本长度的同时，保留了裁判文书中的重要结构信息；之后对粗粒度抽取后的裁判文书文本内容进行一个细粒度的建模抽取，从句子级别进行精细化抽取，以确保最终得到的民事裁判文书在保证简洁明了的基础上最大限度地保留文本的重要信息和文本结构。本文的模型结构如图3所示。

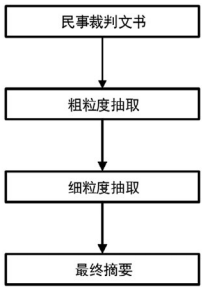

图3 模型结构

Fig.3 The model structure

由于民事裁判文书具有明确的结构，并且每个部分拥有较为明确的标志词，如“经查明”“本院认为”等较为普遍的段落标志词，因此本文采用正则表达式构建抽取规则，对相应的裁判文书内容进行一个粗粒度的抽取。表1为原始裁判文书的文本长度与进行粗粒度抽取后得到的文本长度的统计，从中我们可以看出，经过粗粒度抽取后的裁判文书的文本长度得到大幅度缩减，能够基本满足后续细粒度抽取模型的需要。

表1 文本长度统计

Tab.1 Text length statistics

文本	最多字数	最少字数	平均字数	字数标准差
原始裁判文书	13,064	866	2,568	1,122
粗粒度抽取后的裁判文书	8,599	54	1,395	861

对于细粒度抽取部分，首先是裁判文书中每个句子的标记任务，本文主要采用启发式规则进行文本中摘要句的标记工作，具体的方案为：首先选取原文中ROUGE得分最高的一句话加入候选集合；接着继续从原文中进行选择，保证选出的摘要集合ROUGE得分增加，直至无法满足该条件。得到的候选摘要集合对应的句子设为1标签，其余为0标签。在完成数据集的标注工作之后，就能依托这些标记的数据构建序列标注模型来进行摘要抽取任务了。

本文的抽取模型框架采用的是主流的“BERT+序列标注”框架，如图4所示，将粗粒度抽取得到的裁判文书进行预处理后，通过BERT生成相应的句向量，之后构建序列标注模型作为抽取模型。这里的序列标注模型主要采用三种不同分类方式的构建方法，分别是简单分类器(Simple Classifier)、RNN(Recurrent Neural Network)和Transformer。对于每一个句子，通过分类器得到对应的预测标签P，而整个抽取模型的损失(Loss)便是预测得到的标签与之前所标记的标签之间的二进制分类熵(Binary Classification Entropy)。

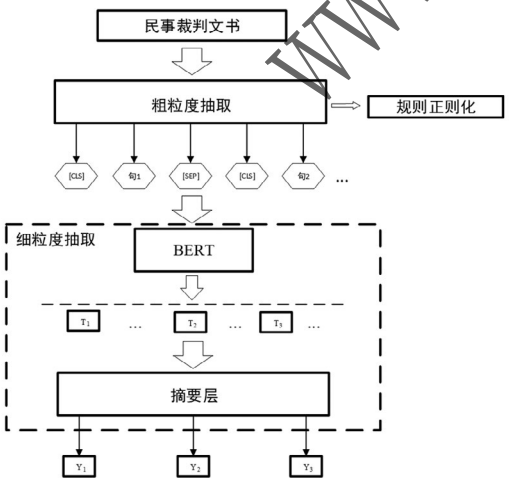

图4 基于BERT的裁判文书摘要模型

Fig.4 Abstract model of judgment documents based on BERT

对于简单分类器，主要是将经过BERT产生的句子向量直接使用线性分类器来获得预测值P：

$$P=\sigma(W_0T_i+b_0) \tag{1}$$

RNN序列标注将BERT的输出经过LSTM(Long Short-Term Memory)层，学习特定的摘要特征，以此判定该输出是否为摘要，并学习到了序列特征，其预测值P的计算方式如下：

$$\begin{pmatrix}F_i\\I_i\\O_i\\G_i\end{pmatrix}=LN_h(W_hh_{i-1})+LN_x(W_xT_i) \tag{2}$$

$$C_i=\sigma(F_i)\odot C_{i-1}+\sigma(I_i)\odot\tanh(G_{i-1}) \tag{3}$$

$$h_i=\sigma(O_t)\odot\tanh\left(LN_c(C_t)\right) \tag{4}$$

$$P=\sigma(W_0h_i+b_0) \tag{5}$$

Transformer则是将多个Transformer层只应用于句子表示，从BERT输出中抽取文档级特征，其预测值P的计算方式如下：

$$\widetilde{h}^l=LN\left(h^{l-1}+MHAtt\left(h^{l-1}\right)\right) \tag{6}$$

$$h^l=LN\left(\widetilde{h}^l+FFN\left(\widetilde{h}^l\right)\right) \tag{7}$$

$$P=\sigma(W_0h_i^L+b_0) \tag{8}$$

3 实验与分析(Experiment and analysis)

3.1 实验准备

本文使用CAIL2020提供的数据集，共包括4,047 篇已标注的民事裁判文书，进行数据清洗后的数据集按照6∶2∶2划分，得到训练数据集2,340 条，验证数据集779 条，测试数据集785 条。

本文采用pytorch框架进行模型的搭建，生成的文本摘要最大长度设置为200。在训练过程中，本文设置的学习率为0.001，累加器的初始值设置为0.1，训练的批次大小为20。评价指标则采用通用的ROUGE评价指标进行性能评价，将自动生成的摘要与参考摘要进行比较，其中ROUGE-1衡量一元词匹配情况，ROUGE-2衡量二元词匹配情况，ROUGE-L记录最长的公共子序列，并分别计算每一种指标的召回率(R)、精准率(P)，以及$F1$值进行实验对比。

3.2 实验结果与分析

为了进一步验证模型的有效性，本文进行了一系列对比实验，将我们的方法同经典的抽取式摘要模型Lead-3模型和TextRank模型，以及只进行粗粒度抽取的模型SUM_extract和只进行细粒度抽取的模型BERT_SUM进行对比。其中，抽取式摘要模型中的Lead-3模型是新闻领域的经典模型之一，它抽取文章的前三句文本作为整篇文章的摘要，由于新闻的重要信息通常都集中在文章的前半部分，因此Lead-3模型取得不错的效果。

由表2可以看出，在$F1$指标上，我们的方法在所有对比实验中获得了最好的性能，Lead-3模型则获得了最低的性能。对于抽取式摘要模型，TextRank模型的性能远远高于Lead-3模型，这也进一步说明了民事裁判文书和新闻领域文本的不同之处：新闻领域文本中的重要信息主要集中在文本前半部分，而民事裁判文书的重要信息则是分布在整篇文章之中，因此Lead-3模型用于民事裁判文书获得了极差的性能。从性能结果来看，SUM_extract由于只进行粗粒度的抽取，导致包含过多的冗余信息，因此相对于更加精炼的TextRank方法来说，性能会偏低；但是和只进行细粒度抽取的BERT_SUM相

比，由于BERT_SUM在数据预处理时采取截取前半部分文本的策略，导致文本的众多重要信息丢失，通过BERT得到的句子语义也会不够充分，从而使得最终的模型性能降低，甚至低于SUM_extract。而我们结合粗粒度和细粒度的文本摘要抽取方法在三种分类方法上均优于BERT_SUM，在整体实验性能上也高于其余方法，足以证明我们的方法在民事裁判文书的摘要生成中兼顾了摘要的简短性和信息的全面性，符合实验预期。

表2 在测试集上模型的ROUGE得分

Tab.2 The ROUGE score of the model on the test sets

模型	ROUGE-1			ROUGE-2			ROUGE-L		
	P/%	*R*/%	*F*1/%	*P*/%	*R*/%	*F*1/%	*P*/%	*R*/%	*F*1/%
Lead-3	8.21	5.39	1.00	1.08	0.06	0.12	8.84	1.12	1.97
TextRank	36.87	39.77	37.10	18.24	19.34	18.20	31.44	32.05	31.03
SUM_extract	58.05	26.06	32.61	25.92	12.97	15.63	57.08	25.33	31.78
BERT_SUM (class)	35.15	27.62	30.04	12.81	10.45	11.18	23.59	19.60	20.89
BERT_SUM (Trans)	36.03	30.16	31.86	14.03	12.07	12.59	24.85	21.47	22.52
BERT_SUM (RNN)	34.64	27.18	29.58	12.51	10.26	10.95	23.43	19.53	20.78
Our_class	44.19	57.69	49.36	26.39	34.64	29.53	34.93	42.51	38.03
Our_Trans	44.01	57.72	49.26	26.36	34.75	29.56	35.61	43.41	38.81
Our_RNN	44.30	57.62	49.37	26.38	34.53	29.48	35.28	42.94	38.41

4 结论(Conclusion)

本文针对民事裁判文书区别于新闻文本的文本结构和重要信息分布的特点，基于BERT提出了一种结合粗粒度和细粒度抽取方法的结构化民事裁判文书摘要生成方法。我们通过对裁判文书原文依次进行粗粒度抽取和细粒度抽取的方法，以确保生成的摘要能够比较完整地保留裁判文书的结构信息和主要内容。实验结果显示，相比于只进行粗粒度抽取的模型SUM_extract和只进行细粒度抽取的模型BERT_SUM，我们的模型均获得了更好的摘要生成性能，达到了实验预期。不过，抽取式摘要由于是抽取原文句子组成摘要，导致可读性比较差，下一步我们将继续研究如何能够结合生成式文摘方法，在保证摘要生成性能的前提下，提高裁判文书生成摘要的可读性。

参考文献(References)

[1] ERKAN G, RADEV D R. Lexrank: Graph-based lexical centrality as salience in text summarization[J]. Journal of Artificial Intelligence Research, 2004, 22:457-479.

[2] GAMBHIR M, VISHAL V. Recent automatic text summarization techniques: A survey[J]. Artificial Intelligence Review, 2017, 47(1):1-66.

[3] 赵洪.生成式自动文摘的深度学习方法综述[J].情报学报,2020,39(03):330-344.

[4] LUHN H P. The automatic creation of literature abstracts[J]. IBM Journal of Research and Development, 1958, 2(2): 159-165.

[5] PADMALAHARI E, KUMAR D V N S, PRASAD S. Automatic text summarization with statistical and linguistic features using successive thresholds[C]// DHANASEKARAN R, MIJA S J, AMINUL I, et al. 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies. Ramanathapuram, India: IEEE, 2014:1519-1524.

[6] MIHALCEA R, TARAU P. Textrank: Bringing order into text[C]// LIN D, WU D. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain: Association for Computational Linguistics, 2004:404-411.

[7] NALLAPATI R, ZHAI F, ZHOU B. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents[C]// SINGH S P, MARKOVITCH S. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI Publications, 2017:3075-3081.

[8] LIU Y. Fine-tune BERT for extractive summarization[EB/OL]. (2019-9-5) [2021-11-5]. https://arxiv.org/abs/1903.10318.

[9] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2019-5-24) [2021-11-5]. https://arxiv.org/abs/1810.04805.

[10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// LUXBURG U V, GUYON I. The proceedings of 31st Conference on Neural Information Processing Systems (NIPS 2017). NY, USA: Curran Associates Inc., 2017:5998-6008.

作者简介：

魏鑫炀(1993-)，男，硕士生.研究领域：自然语言处理，数据融合分析.

唐向红(1979-)，男，博士，教授.研究领域：大数据技术与智能制造.

(上接第9页)

[6] 总装备部电子信息基础部.军用软件研制能力成熟度模型:GJB 5000A—2008[S].北京:中国人民解放军总装备部军标出版发行部,2008:59-64.

[7] 滕俊元,徐忠宾,高猛.软件产品化在航天领域的应用与管理[J].质量与可靠性,2018,193(1):39-41.

[8] 李军锋,顾滨兵,李海浩.软件测试质量评价方法[J].计算机与现代化,2018,277(9):38-42.

作者简介：

杨勋妲(1985-)，女，硕士，高级工程师.研究领域：质量管理，质量管理体系应用.

段明璐(1982-)，男，硕士，高级工程师.研究领域：自主可控数据库测试，项目管理.