文章编号：2095-2163(2019)02-0200-05 中图分类号：TP391.41 文献标志码：A

基于回归的抽取式摘要模型

赵怀鹏，车万翔，刘 挺
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：文本摘要就是一个高度概括原文重要信息的过程。摘要算法大致可以分为 2 类：抽取式摘要和生成式摘要。抽取式摘要的目的是从原文中选择一些重要的短语或句子来组成摘要。生成式摘要是利用算法生成文本的另一种表达，所用到的词汇表述并不一定来自于原文。自动文本摘要能够帮助很多下游任务（例如新闻摘要，社会媒体等）。近些年一些基于神经网络的工作大都将抽取式摘要任务当成序列标注来建模。这就存在训练和测试的不一致性问题：训练时当成分类任务，测试时当成排序任务。研究提出一种基于神经网络的回归模型，让模型在训练的时候就直接拟合 ROUGE 得到其分数用来做排序。实验结果超过目前抽取式摘要的最好结果。

关键词：神经网络；抽取式摘要；回归模型

Research on regression-based extractive summarization system

ZHAO Huaipeng, CHE Wanxiang, LIU Ting
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 Automatic text summarization is the process of generating a concise representation of original text while retaining the core information. Summarization algorithms can be broadly classified into two categories: extractive and abstractive. Extractive approaches aim to select salient words, phrases or sentences from the original text while the abstractive methods focus on rewriting the content without the constraint of reusing words or phrases from the original text. Automatic summarization can aid many downstream applications (e.g., news digests, social media). Recently, neural networks based data-driven approaches have become popular for modeling the extractive summarization task. A few recent approaches conceptualize extractive summarization as a sequence labeling task. Another problem is the discrepancy between training and testing, in which during the test time, it is treated as a ranking problem. Thus the paper presents a regression model to solve it. The proposed model learns to score sentences to fit ROUGE during the training. Experiment results show the proposed model outperforms than other extractive summarization systems.

【Key words】 neural networks; extractive summarization; regression model

0 引 言

随着互联网的迅猛发展，信息量正以指数级别在积累和增长。而摘要则能以精炼的文字帮助人们在海量数据中快速获取自己需要的信息。但鉴于目前信息量潮涌般的生成态势，故而亟需研发一套自动摘要系统来为文本自动总结重要信息，从而快速获取想要的信息。

摘要算法大致可以分为 2 个类别：抽取式摘要和生成式摘要。近年来随着深度学习的日趋成熟，尤其是随着 sequence to sequence[1] 的提出，生成式摘要方面涌现出数目可观的研究成果。而抽取式摘要却因其简单，低成本，能够生成逻辑连贯的摘要等优势，仍然具有重要的研究价值。本课题的目的即旨在设计构造一套抽取式摘要系统。

研究可知，传统的方法大多是利用无监督学习来得到文本的摘要。代表性的研究有：向量空间模型（the vector-space methods）[2-3]、基于图的模型（the graph-based methods）[4-5]、组合优化方法（the combinatorial optimization methods）[6-7]。这些方法依赖大量手工设计的特征来建模句子或篇章，例如位置信息，TF-IDF 等。

近些年，神经网络吸引了学界的高度关注，而 Hinton 等人[8] 发表了优化深层网络的方法后，随却就陆续见到了许多基于神经网络的抽取式摘要工作。这些工作均是将抽取式摘要任务看作序列标注任务。分类的类别有两类：0 代表不是摘要，1 代表是摘要。具体来说，Cheng 等人[9] 提出了基于 sequence to sequence 框架来进行句子分类。Singh 等人[10] 对篇章表示层进行了优化。同时，基于分类

作者简介：赵怀鹏（1993-），男，硕士研究生，主要研究方向：自然语言处理；车万翔（1980-），男，博士，教授，博士生导师，主要研究方向：自然语言处理；刘 挺（1972-），男，博士，教授，博士生导师，主要研究方向：自然语言处理、文本挖掘、文本检索等。
通讯作者：车万翔 car@ ir.hit.edu.cn

收稿日期：2018-12-06

的方法也呈现出一定的弊端与缺陷。Nallapati 等人[11]就提出了基于循环神经网络(Recurrent Neural Networks)的分类模型。首先,在训练过程中,将该任务当成序列标注来建模,但在测试的时候是根据分类概率大小来选择最优的几个句子。这就导致了训练和测试存在不一致性的问题。其次,标注为 1 的句子间也不能区分各自的重要程度。综合前文分析可知,本文则有针对性地研发提出了基于神经网络的回归模型来解决上述问题。

1　基于回归的抽取式摘要模型

1.1　分类模型存在问题及分析

最近几年展开了基于序列标注的神经网络来建模抽取式摘要的研究。这种利用交叉熵来优化与标准答案的最大似然方式并没有在训练过程中考虑排序句子。摘要任务的本质是对句子进行排序,然后选择排序靠前的几个句子。基于分类的模型在训练目标中却忽略了这一点。而且,摘要的分类数据集常常是利用人工摘要通过一定规则得到句子的分类标签。这样就会导致正例的个数过多,模型容易过拟合,而且仅是利用模型也无法区分相同标签的不同句子间的重要程度。

1.2　回归模型概述

给定一篇文章 D,其中包含句子序列 $\{x_1,x_2,\cdots,x_n\}$。抽取式摘要系统的目的就是要从 D 中选择 m 个句子组成摘要 S(其中 $m < n$)。对于每个句子 $s_i \in D$,研究对其预测一个分数 $score_i$。在训练时通过回归损失函数来优化网络。在测试时,对于每个句子 s_i 都会预测一个分数,即:

$$score_i = g(s_i, D, \theta). \tag{1}$$

此后,将选出 $score_i$ 最大的 m 个句子作为摘要。

基于回归的抽取式摘要模型的过程结构设计如图 1 所示。基于回归的抽取式摘要模型一般通过一定的规则来给每个句子打分。例如 Ren 等人[12]就利用当前句子与人工摘要的 ROUGE 值以及句子间的 ROUGE 值来为每个句子打分。在训练的过程中,该模型通过计算当前句子与篇章表示的相关程度和句子间的相关程度来为每个句子评判打分,通过网络训练让模型分数接近正确的分数。测试时,会给每个句子进行评分,然后选择分数最大的作为最终求得的摘要。基于回归模型的优势是分数能够更加精确刻画句子的重要程度,并以此作为依据来进行句子间的排序。另外,在构造分数的时候就考虑到了最终的评价指标 ROUGE[13],因此会更加合理。

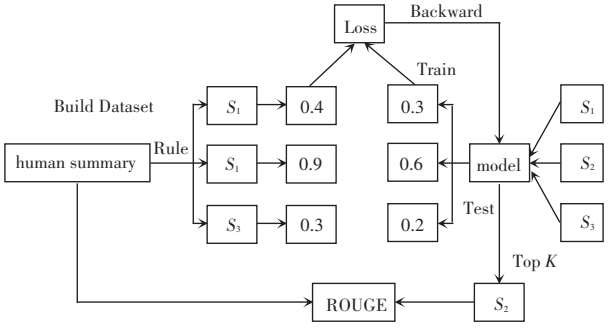

图 1　回归模型结构图

Fig. 1　The structure of regression model

1.3　基于神经网络的抽取式摘要模型

本文中的句子和篇章的表示层利用了 Yang 等人[14]提出的 Hierarchical attention networks。如图 2 所示,该结构分为 3 层:输入层、句子表示层和篇章表示层。该模型的设计初衷是用于篇章分类(document classification),而本次研究则将其用于抽取式摘要系统的表示层。

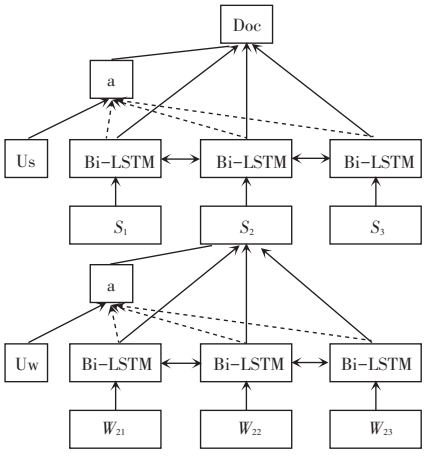

图 2　层次化注意力网络

Fig. 2　Hierarchical attention networks

本次研究的输入层采用了 100 维的词向量,而选择了训练词向量的工具是 word2vec[15],过程中训练词向量用到的训练数据是 CNN/DailyMail[16]数据集里面所有的文本。继而,文中设置的最小词频阈值为 8,这样就可得到 154 K 的词汇。Skip 窗口大小设置为 5,hierarchical softmax 的层数也是 1。

同时,对于句子表示层和篇章表示层,研究采用了 Bi-LSTM。LSTM 中包含 3 个门:输入门(input gate)、输出门(output gate)和遗忘门(forget gate),如图 3 所示。

在得到 LSTM 的隐层输出之后,研究利用 Attention[17]机制得到每个词或者句子的权重。设计

时,计算 Attention 的向量是随机初始化,并通过网络学习进行更新。以篇章表示层为例,假设 h_t 为第 t 个句子的表示,U_s 是计算 Attention 的向量。那么两者分数计算方式可表述如下:

$$score(U_s, h_t) = U_s^{\mathrm{T}} \cdot h_t, \tag{2}$$

$$a_t = \frac{\exp(score(U_s, h_t))}{\sum_t \exp(score(U_s, h_t))}, \tag{3}$$

$$d = \sum_t a_t h_t, \tag{4}$$

其中,d 就是研究中最终的篇章表示,h_t 就是求得的句子表示。

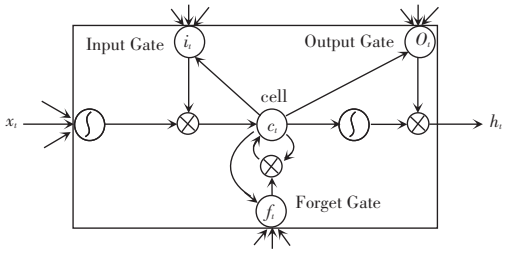

图 3 LSTM Cell 结构图

Fig. 3 The structure of LSTM Cell

在此基础上,研究推得的最终回归模型的打分函数将可写作如下形式:

$$score_i = \sigma(h_i^{\mathrm{T}} W d + b), \tag{5}$$

接下来,通过计算当前句子 s_i 与人工摘要 S^{ref} 的 ROUGE-2 $F1$ 值就可得到标准分数,其数学公式可表示为:

$$score_i^g = ROUGE(s_i, S^{ref}), \tag{6}$$

在得到了篇章表示后,就可以定义损失函数如式(7)所示:

$$L = -score_i^g \cdot \log(score_i) - (1 - score_i^g) \cdot \log(1 - score_i). \tag{7}$$

2 实验结果与分析

2.1 基本设置

词向量维度为 100 维,句子表示层和篇章表示层 Bi-LSTM 的维度为 200 维。训练采用的优化器为 Adam,初始学习率为 0.001。Batch size 为 20,随机种子设为 1,训练迭代了 10 轮。

研究对每篇文章进行了预处理,去除了文章日期,作者信息等。同时对所有单词做了小写化处理。为了降低时间和计算资源开销,同时还设置每篇文章最多 100 个句子,每个句子最多 50 个词,如果超过就进行截断。而在研究句子级别表示层时,选取一个 batch 中所有篇章词数最多的句子(超过 50 的按照 50 计算)作为 padding 的基准,词数未达此标准的句子增补若干个 100 维的 0 向量。在篇章表示层中,选取一个 batch 中篇章句子数最多的篇章(超过 100 的按照 100 计算)作为 padding 基准,句子数不够的予以补 0 向量处理。

2.2 数据集

实验用到的数据集是 CNN/Daily Mail 数据集。数据的内容是 CNN 和 Daily Mail 发布的新闻数据,每篇文章包含标题名称、正文和人工摘要三个部分,样本示例见表 1。该数据集最初是由 Hermann 用于完成阅读理解任务。后来 Cheng 等人[9]将其作为抽取式摘要的数据集。由于数据集的规模较大,在近段时间内已被广泛应用到文本摘要任务中。数据集的规模统计参见表 2。

表 1 数据集样本示例

Tab. 1 Sample of the dataset

类别	内容
标题	AFL star blames vomiting cat for speeding
正文	Adelaide Crows defender Daniel Talia has kept his license, telling a court he was speeding 36km over the limit because he was distracted by his sick cat. The 22-year-old AFL star, who drove 96km/h in a 60km/h road works zone on the South Eastern expressway in February, said he didn't see the reduced speed sign because he was so distracted by his cat vomiting violently in the back seat of his car. In the Adelaide magistrate court on Wednesday, Magistrate Bob Harrap fined Talia $824 for exceeding the speed limit by more than 30 km/h. He lost four demerit points, instead of seven, because of his significant training commitments.
摘要	Adelaide Crows defender Daniel Talia admits to speeding but says he didn't see road signs because his cat was vomiting in his car. 22-year-old Talia was fined $824 and four demerit points, instead of seven, because of his significant training commitments.

表 2 数据集规模统计

Tab. 2 The statistics of dataset

数据集	CNN/篇	DailyMail/篇
训练集	90 266	196 961
开发集	1 220	12 148
测试集	1 093	10 397

实验中,重点选用了 Daily Mail 数据集,因为近年来的大部分工作都在 Daily Mail 数据集上提交了结果,因而有利于后续的实验结果对比。Daily Mail 数据集中每篇文章的平均句子数为 25.6,人工摘要的平均长度在 3~4 句的范围内。

2.3 评价指标

早期，传统的摘要评价方式一般都包含人工的评分函数，包括语法、可读性、内容、一致性等。这些简单的人工评价规则能够较好反映摘要的质量，但是需要消耗大量的人力去进行评估。Lin[13] 提出 *ROUGE*（Recall - Oriented Understudy for Gisting Evaluation）用来评价摘要的质量，并和人工评价有着很强的一致性，目前即将其作为一种常用的摘要评价指标。分析可知，常用的评价指标有 *ROUGE* - 1、*ROUGE* - 2 和 *ROUGE* - *L*。前两者分别计算了 uni-gram 和 bi-gram 的覆盖度，表示了涵盖的信息量，后者计算了最长公共子序列（longest common subsequence）的覆盖度，描述了生成摘要的流畅程度。*ROUGE* - *N* 和 *ROUGE* - *L* 可由如下公式计算得出：

$$ROUGE - N = \frac{\sum_{s \in S^{ref}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{s \in S^{ref}} \sum_{gram_n \in S} Count(gram_n)}, \quad (8)$$

$$ROUGE - L = \frac{LCS(S, S^{ref})}{m}. \quad (9)$$

2.4 实验结果

本次研究中的 baseline 模型是 Lead-3，且只取文章中前 3 句话作为摘要。另外，研究中还对比了文献[9]和文献[11]中的仿真结果。这里，即研究给出了不同长度限制下的实验结果详见表 3、表 4。

表 3　DailyMail 测试集 75 bytes 下 *ROUGE Recall*

Tab. 3　75 bytes *ROUGE Recall* of DailyMail test set

	ROUGE - 1	*ROUGE* - 2	*ROUGE* - 3
LEAD-3(the proposed)	23.2	8.4	11.9
Cheng et al´16	22.7	8.5	12.5
SummmaRuNNer	26.2	10.8	14.4
The proposed regression	**29.2**	**15.2**	**15.8**

表 4　DailyMail 测试集 275 bytes 下 *ROUGE Recall*

Tab. 4　275 bytes *ROUGE Recall* of DailyMail test set

	ROUGE - 1	*ROUGE* - 2	*ROUGE* - 3
LEAD-3(the proposed)	40.4	15.6	32.1
Cheng et al´16	**42.2**	17.3	**34.8**
SummmaRuNNer	42.0	16.9	34.1
The proposed regression	42.1	**19.0**	34.0

由表 3、表 4 的实验结果来看，本文的模型在生成短摘要时，效果上要明显优于其它的抽取式摘要模型。在生成长摘要时，效果也能和 SOTA 相当。

3　结束语

本文分析了利用分类来做抽取式摘要的问题，并设计提出了一个基于神经网络的回归模型。结果表明，本文研发的模型不依赖任何手工设计的特征。而且，在 DailyMail 数据集上，研究提出的模型在不同长度限制下都取得了不错的效果。

参考文献

[1] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks [C]//Advances in neural information processing systems. Montreál, Canada: dblp, 2014: 3104-3112.

[2] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.

[3] HOFMANN T. Probabilistic latent semantic analysis[C]//UAI'99 Proceedings of the Fifteenth conference on Uncertainty in Artificial Intelligence. Stockholm, Sweden: ACM, 1999: 289-296.

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

[5] MIHALCEA R, TARAU P. TextRank: Bringing order into text [C]//EMNLP. Barcelona, Spain: ACL, 2004, 4: 404-411.

[6] GILLICK D, FAVRE B, HAKKANI T D. The ICSI summarization system at TAC 2008[C]//Proceedings of the First Text Analysis Conference. Maryland, USA: dblp, 2008: 1-8.

[7] MCDONALD R. A study of global inference algorithms in multi-document summarization [C]//ECIR´07 Proceedings of the 29th European conference on IR research. Rome, Italy: ACM, 2007: 557-564.

[8] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural computation, 2006, 18(7): 1527-1554.

[9] CHENG Jianpeng, LAPATA M. Neural summarization by extracting sentences and words[J]. arXiv preprint arXiv: 1603.07252, 2016.

[10] SINGH A K, GUPTA M, VARMA V. Hybrid memNet for extractive summarization[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. Singapore: ACM, 2017: 2303-2306.

[11] NALLAPATI R, ZHAI Feifei, ZHOU Bowen. SummaRuNNer: A Recurrent Neural Network based sequence model for extractive summarization of documents [C]//The Thirty-First AAAI Conference on Artificial Intelligence (AAAI-2017). San Francisco, California, USA: AAAI, 2017: 3075-3081.

[12] REN Pengjie, WEI Furu, CHEN Zhumin, et al. A redundancy-aware sentence regression framework for extractive summarization [C]//Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan: ACL, 2016: 33-43.

[13] LIN C Y. Rouge: A package for automatic evaluation of summaries[J]. Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004. Barcelona, Spain: ACL, 2004: 1-10.

[14] YANG Zichao, YANG Diyi, DYER C, et al. Hierarchical attention networks for document classification[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California: ACL, 2016: 1480-1489.

（下转第 207 页）