计算机科学
COMPUTER SCIENCE

基于GAN和中文词汇网的文本摘要技术

刘晓影，王淮，乌吉斯古愣

引用本文

刘晓影，王淮，乌吉斯古愣.基于GAN和中文词汇网的文本摘要技术[J]．计算机科学, 2022, 49(12): 301-304.

LIU Xiao-ying, WANG Huai, WU Jisiguleng. GAN and Chinese WordNet Based Text Summarization Technology [J]. Computer Science, 2022, 49(12): 301-304.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210600166

基于 GAN 和中文词汇网的文本摘要技术

刘晓影　王　淮　乌吉斯古愣
华北计算技术研究所网络安全工作组　北京 100083
(xiaoying81@126.com)

摘　要　随着神经网络技术的广泛应用,文本摘要技术吸引了越来越多科研人员的注意。由于生成式对抗网络(GANs)具有提取文本特征或学习整个样本的分布并以此产生相关样本点的能力,因此正逐步取代传统基于序列到序列(Seq2seq)的模型,被用于提取文本摘要。利用生成式对抗网络的特点,将其用于生成式的文本摘要任务。提出的生成式对抗模型由3部分组成:一个生成器,将输入的句子编码为更短的文本表示向量;一个可读性判别器,强制生成器生成高可读性的文本摘要;以及一个相似性判别器,作用于生成器,抑制其输出的文本摘要与输入的摘要之间的不相关性。此外,在相似性判别器中,引用中文的WordNet作为外部知识库来增强判别器的作用。生成器使用策略梯度算法进行优化,将问题转化为强化学习。实验结果表明,所提模型得到了较高的ROUGE评测分数。

关键词:文本摘要;生成式对抗网络;WordNet;强化学习;自然语言处理
中图法分类号　TP391

GAN and Chinese WordNet Based Text Summarization Technology

LIU Xiao-ying, WANG Huai and WU Jisiguleng
Network Security Group, North China Institute of Computing Technology, Beijing 100083, China

Abstract　Since the introduction of neural networks, text summarization techniques continue to attract the attention of researchers. Similarly, generative adversarial networks(GANs) can be used for text summarization because they can generate text features or learn the distribution of the entire sample and produce correlated sample points. In this paper, we exploit the features of generative adversarial networks(GANs) and use them for abstractive text summarization tasks. The proposed generative adversarial model has three components: a generator, which encodes the input sentences into shorter representations; a readability discriminator, which forces the generator to create comprehensible summaries; and a similarity discriminator, which acts on the generator to curb the discorrelation between the outputted text summarization and the inputted text summarization. In addition, Chinese WordNet is used as an external knowledge base in the similarity discriminator to enhance the discriminator. The generator is optimized using policy gradient algorithm, converting the problem into reinforcement learning. Experimental results show that the proposed model gets high ROUGE evaluation scores.

Keywords　Text summarization, Generative adversaial network, WordNet, Reinforcement learning, Natural language processing

1　引言

文本摘要是用较少的语义准确的词来概括给定的整篇文档的技术。通过这样的方式,关键的信息能够得以保留,而文档中不重要的冗余信息则被过滤掉,从而提高人们对信息的接受效率。文本摘要技术分为两种类型:抽取式和生成式。抽取式文本摘要需要从语料库中提取出已经存在的、包含正确含义的词或句子,来维持文档的总体意义。然后,将这些抽取出的词或短语进行组合以形成摘要。而在生成式的文本摘要中,需要在理解了文本的含义之后进行概括,生成新的句子或词,以形成语义相关的摘要。显然,相比抽取式文本摘要,生成式文本摘要的任务具有更吸引人的应用场景,因为它与人类总结摘要的方式相似。但是生成式文本摘要任务通常是一个难以完成的任务,比抽取式更难执行,因为它需要全面理解文章的内容才能生成一个好的摘要,这对于机器来说是难以实现的目标。进行自动文本摘要时存在的另一个问题是输入语句的长度不一致。如果输入的语料库和其中的句子很长,则很难捕捉到上下文语义关系并找到语义相关的关键词或短语。此外,受语义表示、推理和自然语言生成等技术的牵制,抽取式摘要的结果通常比生成式摘要的结果更好。这些问题比抽取式摘要这样的基于数据驱动的方法更难,因为数据驱动的方法一般只涉及定位和提取目标单词或短语,不

到稿日期:2021-06-21　返修日期:2021-09-14
基金项目:国家重点研发计划(2018YFC0831200)
This work was supported by the National Key R&D Program of China(2018YFC0831200).
通信作者:王淮(498929906@qq.com)

涉及系统对文章语义的理解。然而，随着深度神经网络的广泛应用，自然语言处理模型对句子的语义表示掌握程度也得到了快速提升，生成式摘要也得到了很好的发展。

本文提出了一个基于生成对抗网络(Generative Adversarial Network，GAN)的文本摘要模型，并在中文新闻数据集上进行了生成式文本摘要的实验。该模型包含一个生成器和两个判别器，生成器负责将输入的句子编码成较短的长度；其中一个判别器负责创造人类可理解的词语，另一个判别器则负责提供反馈并约束生成器产生与输入词语相关的输出。此外，本文还将使用 WordNet 作为外部知识库来增强判别器的功能。在训练时，还使用了策略梯度算法将损失从判别器传回生成器。摘要的相关性用 ROUGE-N 评分[1]来评估。

2 相关研究

在过去 5 年里，由于神经网络和深度学习的发展，科研工作者针对文本摘要任务进行了许多的探索。Rush 等[2]提出了首个自动文本摘要模型，使用注意力神经机制对输入句子进行摘要生成。与几个强大的基线模型相比，他们的自动文本摘要模型获得了显著的性能提升。然而，这种只依赖于注意力机制的方法不能很好地捕捉长距离语义关系，因此只在短句上表现良好。后来，为了提升文本摘要系统的性能，研究人员做了许多改进，如利用编码器-解码器模型的不同变体进行实验。

Ramesh 等[3]使用了基于注意力的编码器-解码器和序列到序列的递归神经网络(RNNs)，对较长的文档的句子和语料库[4-5]进行了生成式的文本摘要。Ramesh 等还介绍了一个新的数据集，即 CNN/Daily Mail 数据集。该数据集的特点是其摘要由多个句子组成，因此需要模型有较好的长距离依赖捕捉能力，更具挑战性。

Abigai 等[6]用指针生成器网络进行文本摘要，ROUGE 得分相比当时最先进的模型提高了 20%。他们提出了一种新颖的架构，以抽取式和生成式相混合的模型，改进了标准的基于注意力的序列-序列模型。

Banafsheh 等[7]在 ROUGE-L 指标上改进了最先进的摘要模型，首次将 GAN 与策略梯度相结合用于文本摘要任务。他们将生成器建模为强化学习中以生成高质量摘要为目标的随机策略。判别器的作用则是计算和预测摘要来自训练数据集而不是生成器的概率，从而使生成器尽可能学习真实的数据分布。

Wang 等[8]介绍了一种新的基于内部注意力机制的深度强化模型。他们将监督学习与强化学习相结合，以消除监督学习造成的“暴露偏差”。通过将标准的词预测与强化学习的全局序列预测训练相结合，产生了可读性更高的摘要。

WordNet 是由普林斯顿大学的心理学家、语言学家和计算机工程师联合设计的一种基于认知语言学的英语词典。它不是简单地把单词以字母顺序排列，而是按照单词的意义组成一个“单词的网络”。Zhuang 等[9]使用 WordNet 来实现将提取到的句子语义特征进行抽取式的文本摘要。在他们的方法中，WordNet 被用来实现句子语义特征的提取。

3 基于生成式对抗网络的文本摘要方法

本文提出了一个基于生成式对抗网络的文本摘要模型，该模型能实现自动对输入文本提取关键信息，主要包括 3 个部分。

3.1 生成器

生成器以长句子作为输入，然后将其编码为长度较短的文本。生成器可以直接从文本或词汇表中预测组成摘要的单词的概率。生成器输入的第 i 个句子可以表示为 $x^i=\{x_1^i, x_2^i, x_3^i, \cdots, x_N^i\}$，其中 x_n^i 表示句子 i 的第 n 个单词。接受输入后，生成器将输出一个词汇表上的单词概率分布，然后对其进行抽样产生 $y^i=\{y_1^i, y_2^i, y_3^i, \cdots, y_T^i\}$。因此，生成器的损失函数可以定义为：

$$L_G=\frac{1}{N}\sum_{i=1}^{N}-\alpha C(x^i, y^i)-\beta D(y^i)$$

其中，$C(x^i, y^i)$是相似度判别器的输出，用于计算输入与输出概率分布之间的相似度；$D(y^i)$是可读性判别器的输出，用来衡量生成器的输出可读性的优劣；α 和 β 是两个模型自学习的超参数。

之后，生成器用策略梯度优化算法进行优化[10-12]。生成器由编码器和解码器两部分组成，它们均由两层的 LSTM 网络构成，隐状态向量维度分别为 128 和 256。生成器的输出为四维的分类张量，分别对应 4 个类别：相似、非完整、冗余和无相关性。所使用的损失函数是交叉熵损失函数，以数据中专家撰写的参考摘要作为真值。

3.2 相似性判别器

相似性判别器是一个 4-类别文本分类器，4 个类别与生成器一致：相似、非完整、冗余和无相关性。具体含义如下。

(1)相似：短文本与长文本非常相关。

(2)非完整：长文本中的某些特征不存在于短文本中。

(3)冗余：短文本中存在重复信息，从而导致生成了不恰当的摘要。

(4)无相关性：短文本与长文本不相关。

相似性判别器由两个编码器组成，它们都是基于 CNN 的模型，具有卷积、最大池化层和激活层。一个编码器用于短文，另一个则用于长文。在训练的最开始，CNN 的权重用 xavier 方法进行初始化。将两个编码器的输出 e_l 和 e_s 以及由 WordNet 提取的特征拼接在一起，进行 Hadamard 乘积、求差的绝对值后，接入一个全连接网络，并用 softmax 求得概率分布。相似度判别器的整体架构如图 1 所示。

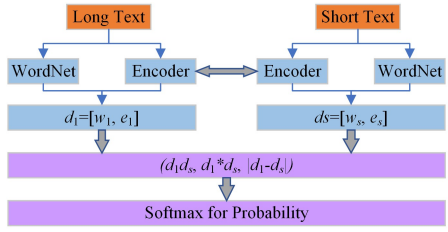

图 1 相似度判别器结构

Fig. 1 Architecture of similarity discriminator

相似度判别器中的编码器使用窗口大小分别为 3，5 和 7

的卷积过滤器，共 256 核，并按照文献[13]中的预训练方法进行了预训练。然后，将编码器输出的 256 个激活特征图连接到一个具有 512 个神经元的全连接层[10]。相似度判别器用 SGD 优化算法进行优化，学习率为 0.06，训练批次大小为 128。

3.3 WordNet 特征

TF-IDF 的计算基于两种统计方法，即术语频率(Term Frequency，TF)和逆文档频率(Inverse Document Frequency，IDF)。TF 指一个术语或单词在一个文档中出现的总次数，IDF 则表示出现某一特定术语的文档总数。IDF 的计算还依赖于语料库中存在的所有文档数。例如，在文献[12]中，TF-IDF 是 TF 和 IDF 的乘积。IDF 的计算方式是 N 的对数除以 DF。这里，N 是语料库中所有的文档数，DF 是一个词出现的文档数量。

WordNet 中的每一个词都构成了一个同义词网络。每个同义词组中的一个词可能对应于该词可能出现的不同的语境。模型在编码当前单词或句子的语义时，若能充分考虑到不同语境下单词所包含的不同语义，就能充分丰富语义。因此可以利用 WordNet 识别同义词组，以进行知识语义增强，并根据它们在文本中出现的频率来调整 TF-IDF 的结果。因此，我们可以通过 WordNet 中的同义词来改进 TF-IDF 的语义表示能力。最终，对短文本与长文本分别应用 WordNet，得到两个向量 $\boldsymbol{w}_s$ 与 $\boldsymbol{w}_l$。

3.4 可读性判别器

可读性判别器是另一个基于 CNN 的模型，用于判别生成的摘要是生成器生成的还是人类生成的。因此，它本质上是一个二元分类器，其结构包括一个卷积层、一个最大池化层，以及一个用于预测两个类别概率分布的 softmax 分类器。同样地，该判别器按照与相似性判别器相同的预训练方式进行预训练。

判别器将人类书写的参考摘要视为正类别，而由生成器生成的摘要则被视为负类别。可读性判别器的二元分类器中使用的损失函数是依照正负样本的输出概率而定义的，表达式如下：

$$L_D=\frac{1}{N}\sum_{i=1}^{N}-\log(1-D(x^i_-))-\log(D(x^i_+))$$

可读性判别器也是一个 CNN 模型，使用的过滤器窗口大小为 3，4，5。每个大小的过滤器有 128 个，因此能产生一个 128 维的激活特征图。在用生成器和相似性判别器对可读性判别器进行训练之前，需要对可读性判别器进行预训练。在预训练阶段，我们只用生成器来对可读性判别器进行预训练。生成器的输出被视为负样本，而参考文本摘要则是正样本。摘要生成模型的整体架构如图 2 所示。

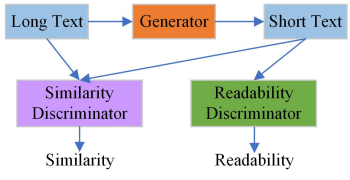

图 2 模型整体结构

Fig. 2 Overall architecture of the proposed model

3.5 策略梯度优化算法

由于无差异性问题，生成器的输出并非直接输入两个判别器中，而是对输出进行采样，将得到的离散序列输入两个判别器中。通过使用 SeqGAN[11-12] 中的技巧，文本摘要的任务被转换为一个强化学习的问题。当生成器产生语义相关的摘要时，相似性判别器就会给予生成器奖励，鼓励生成器继续生成好的语义相关的摘要。同样地，当生成器产生的摘要是人类可理解的摘要时，可读性判别器也会为生成器产生奖励。生成器的训练用策略梯度优化算法进行优化。策略梯度有助于以强化学习的方式训练生成式对抗网络，并消除监督训练时引入的“暴露偏差”(Exposure Bias)。

4 实验设置与结果分析

4.1 实验数据及评价标准

实验采用的数据集来自国内实际新闻网站的新闻数据。为了充分检验文本摘要模型的有效性，本文使用了多个不同规模大小的数据集。

(1)数据集 1。该数据集包含 350 个新闻网页，有 87 个主题。这些网页均发表于 2007 年 3 月和 4 月，内容都是关于搜索引擎公司的。它们都来自国内的新闻网站，包括新闻报道和评论。最大的主题有 20 篇报道，而最小的只有 1 篇。

(2)数据集 2。该数据集包含 953 个新闻网页，有 108 个主题。这些网页来自国内主要新闻门户网站之一的搜狐网体育频道的新闻。最大的主题有 151 篇新闻报道，而最小的只有 1 篇。

(3)数据集 3。该数据集包含了几十家国内新闻网站体育频道的 24782 个新闻页面以及指定网络日志和网络论坛列表中的 1339 篇文章，发表于 2007 年 9 月和 10 月。

实验采用 ROUGE-N 来评价摘要的效果。ROUGE(Recall-Oriented Understudy for Gisting Evaluation)是一套用于评估自然语言处理中自动文本摘要和机器翻译效果的指标。这些指标将自动产生的摘要或翻译与参考(或一组参考)摘要(如人类书写的摘要)进行比较，得出相应的分值，以衡量自动生成的摘要与参考摘要之间的相似度。ROUGE-N 指摘要系统生成的摘要与参考摘要之间 n-gram 的重叠程度。ROUGE-1 指一元语法(Unigram，或单个单词)的重叠程度，ROUGE-2 指二元语法(Bigram)的重叠程度。其定义如下：

$$ROUGE\text{-}N=\frac{\sum_{S\in\{\text{ReferenceSummaries}\}}\sum_{gram_n\in S}Count_{\text{match}}(gram_n)}{\sum_{S\in\{\text{ReferenceSummaries}\}}\sum_{gram_n\in S}Count(gram_n)}$$

其中，n 表示 n 元语法的长度，也即 $gram_n$，且 $Count_{\text{match}}(gram_n)$ 表示候选摘要和一组参考摘要中共有的最大数量。显然，ROUGE-N 是一个基于召回率的衡量指标，因为其分母是参考摘要的 n 元语法数量的总和。此外，为了便于计算，基于 ROUGE-N 还衍生出了 ROUGE-L，其是基于句子 X 与句子 Y 之间的最长公共子序列长度计算得来的。

4.2 实验结果分析

本实验基于 ROUGE-1，ROUGE-2 以及 ROUGE-L 这 3 个指标来对比本文的摘要生成模型与其他摘要模型之间效果

的优劣。模型在3个数据集上的*R*-1,*R*-2以及*R*-*L*平均值如表1所列。

表1 各个模型*R*-1,*R*-2和*R*-*L*的结果对比

Table 1 Comparison of results of each model in *R*-1, *R*-2 and *R*-*L*

方法	*R*-1	*R*-2	*R*-*L*
Pointer-generator	39.51	17.18	36.28
SummaRuNNer	39.57	16.1	35.3
GAN+policy gradient	37.83	15.7	37.33
Deep-RL	39.85	15.86	36.87
WGAN	35.51	9.36	23.97
GAN	39.91	**17.64**	36.69
Zhuang 等	40.1	16.22	37.16
本文模型	**40.65**	17.01	**37.53**

本文模型在单元语法项的重叠程度(Unigram Overlapping)上,即*R*-1和最长公共子序列*R*-*L*上取得了最优的性能效果。然而,该模型的*R*-2得分,即参考摘要与模型生成的摘要之间的2元语法项的重叠程度上稍稍落后。造成这种情况的原因可能是模型失去了完整的2元语法项的上下文语义,这可以通过使用在大规模数据集上进行预训练的通用语言模型改进词嵌入表示来进一步改进2元语法项的上下文语境表示。

结束语 本文提出了一种基于生成式对抗网络和策略梯度优化算法的生成式文本摘要方法。实验结果表明,所提模型能够对给定长文章或语料生成高可读性的摘要。模型在3个测试数据集上的*R*-1和*R*-*L*得分均达到了最优的效果。所提模型主要包含3个部分:生成器、相似性判别器以及可读性判别器。相似性判器用来提高生成器生成高相关性的文本摘要的能力,而可读性判别器用来提高生成器生成高可读性文本摘要的能力。同时,为了增强相似性判别器的作用,本文提出了一种基于中文WordNet的加权TF-IDF改进方法,进而提升相似性判别器判断相似程度的能力。模型的3个部分均首先进行单独的预训练,然后再用强化学习的策略梯度优化算法进行联合训练。后续的工作可以通过使用文献[14]中介绍的BERT词嵌入来进行进一步改善。这些是预先训练好的词向量,可以通过迁移学习使用,也可以自定义训练。这些预训练的通用语言模型有助于进一步提高ROUGE指标。

参 考 文 献

[1] LIN C Y. Rouge: A package for automatic evaluation of summaries[C]// Proceedings of the Workshop on Text Summarization Branches Out(WAS 2004). 2004.

[2] RUSH A M, CHOPRA S, WESTON J. A neural attention model for abstractive sentence summarization[J]. arXiv: 1509.00685, 2015.

[3] RAMESH N, ZHAI F F, ZHOU B W. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2017.

[4] ILYA S, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[J]. arXiv: 1409.3215, 2014.

[5] RAMACHANDRAN P, LIU P J, LE Q V. Unsupervised pretraining for sequence to sequence learning[J]. arXiv: 1611.02683, 2016.

[6] SEE A, LIU P J, MANNING C D. Get to the point: Summarization with pointer-generator networks[J]. arXiv: 1704.04368, 2017.

[7] BANAFSHEH R, MOUSAS C, GUPTA B. Generative adversarial network with policy gradient for text summarization[C]// 2019 IEEE 13th International Conference on Semantic Computing(ICSC). IEEE, 2019.

[8] WANG Y S, LEE H Y. Learning to encode text as human-reaable summaries using generative adversarial networks[J]. arXiv: 1810.02851, 2018.

[9] ZHUANG H J, ZHANG W B. Generating semantically similar and human-readable summaries with generative adversarial networks[J]. IEEE Access, 2019, 7: 169426-169433.

[10] YU L T, ZHANG W L, WANG J, et al. Seqgan: Sequence generative adversarial nets with policy gradient[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2017.

[11] SUTTON R S, MCALLESTER D, SINGH S, et al. Policy gradient methods for reinforcement learning with function approximation[J]. Advances in Neural Information Processing Systems, 2000, 12: 1057-1063.

[12] ZHANG H Y, XU J J, WANG J. Pretraining-based natural language generation for text summarization[J]. arXiv: 1902.09243, 2019.

[13] SARKAR K. Bengali text summarization by sentence extraction[J]. arXiv: 1201.2240, 2012.

[14] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv: 1810.04805, 2018.

LIU Xiao-ying, born in 1981, Ph.D, senior engineer, is a member of China Computer Federation. Her main research interests include natural language processing, artificial intelligence and network security.

WANG Huai, born in 1996, master, engineer. His main research interests include network security, threat intelligence analysis and knowledge graph.

(责任编辑:杨雪敏)