文章编号: 1004-4353(2019)03-0254-06

一种多语种生成式自动摘要方法的实现

易志伟, 赵亚慧*, 崔荣一
(延边大学 工学院, 吉林 延吉 133002)

摘要: 为实现多语种生成式自动摘要,基于序列到序列(Seq2Seq)模型提出了一种多语种生成式自动摘要方法.首先,按照传统的多语种自动摘要方法,将中、英、朝 3 个语种的语料分开训练,得到 3 个模型,并观察各模型在测试集上的表现;其次,按照本文提出的多语种自动摘要法,将中、英、朝 3 种语言的语料放在一起共同训练出一个模型,然后运用该模型分别运行中文、英文、朝文语料的测试集,并观察模型的表现;最后,用同一个测试集测试模型改进前后的摘要生成效果.实验结果表明,本文方法生成多语种自动摘要的效果与传统方法相近,但因本文方法只用一个模型即可实现多语种自动摘要,因此更具有适用性.

关键词: 生成式; 自动摘要; 多语种; 共同训练
中图分类号: TP391.1 **文献标识码**: A

Implementation of a multilingual abstractive automatic summarization method

YI Zhiwei, ZHAO Yahui*, CUI Rongyi
(*College of Engineering*, *Yanbian University*, *Yanji* 133002, *China*)

Abstract: In order to realize multilingual abstractive automatic summarization, a multilingual abstractive automatic summarization method is proposed based on the sequence-to-sequence (Seq2Seq) model. Firstly, according to the traditional multilingual automatic summarization method, the corpora of Chinese, English and Korean languages were trained separately, and three models were obtained to observe their performance on the test set. Secondly, according to the multilingual automatic summarization proposed in this paper, the method combined the Chinese, English and Korean corpora together to train a model, and then used this model to run the test set of Chinese, English and Korean corpus separately to observe the performance of the model. Finally, the same test set was used to test the effect of summarization before and after improving the model. The experimental results show that the effect of multilingual automatic summarization method proposed in this paper is similar to the traditional method, but can realize multilingual automatic summarization with only one model, so it is more applicable than traditional methods.

Keywords: abstractive; automatic summarization; multilingual; joint training

0 引言

文本摘要通常是指从一个文档中生成一段包含原始文档主要信息的文本.由于文本摘要的篇幅相比原始文档大幅减少,因此可为读者节省大量阅读时间,同时也可起到信息压缩的作用.1958 年,Luhn 首次提出了基于"簇"的自动摘要方法[1].相比人工撰写摘要,由于自动摘要技术可以大幅提高撰写摘要的效率,因此引起了学术界的广泛关注.目前,生成自动摘要的方法可分为抽取式摘要(extractive sum-

收稿日期: 2019-04-17

* **通信作者**: 赵亚慧(1974—),女,副教授,研究方向为自然语言文本处理.

marization)和生成式摘要(abstractive summarization). 抽取式摘要的特点是摘要中的句子是原文中的句子，又叫做“句子摘录”. 该方法通常使用 TextRank[2] 和 LexRank[3] 算法在文本中进行摘要句的抽取，但由于这两种算法都是基于 PageRank 算法[4]对拓扑图进行迭代计算，所以抽取式摘要所抽取的句子往往含有大量冗余信息，并且句子之间连贯性不强，可读性较差. 生成式摘要的特点是摘要中的句子是重新生成的句子，其使用的方法主要是基于序列到序列(sequence-to-sequence，Seq2Seq)模型的深度学习方法[5]. 由于该方法生成的摘要具有长度较短、冗余性较低、句子的概括性较强等优点，因此生成式摘要更加受到了学者的青睐[6-8]. 但目前为止，基于 Seq2Seq 模型的生成式自动摘要系统只能处理单一语种的文本，若要处理其他语种的文本，需要利用其他语种的语料重新训练新的模型. 本文将中、英、朝 3 种语种的训练数据一起训练，得到一个可以同时处理中文、英文、朝文 3 种文本的多语种自动摘要模型，并通过实验验证本文方法的有效性.

1　生成式摘要模型

循环神经网络(recurrent neural network，RNN)[9]的主要用途是处理、预测序列数据和挖掘数据中的时序信息，常用于语音识别、语言模型以及机器翻译等领域. 循环神经网络的内部结构如图 1 所示.

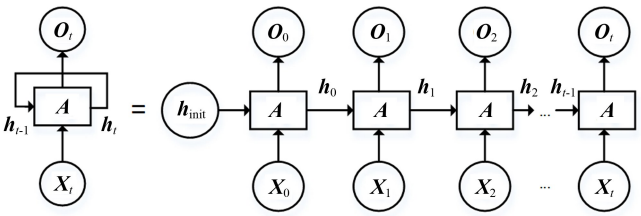

图 1　循环神经网络按时间展开后的结构图

在循环神经网络中，输入为一个序列，用 $\boldsymbol{X}=\{\boldsymbol{X}_0,\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_t\}$ 表示. 在每个时刻 t，RNN 的隐藏状态 $\boldsymbol{h}_t$ 由式(1) 更新.

$$\boldsymbol{h}_t=f(\boldsymbol{h}_{t-1},\boldsymbol{X}_t), \tag{1}$$

其中 f 代表一个非线性激活函数，可以是 sigmoid 函数或 tanh 函数. RNN 网络可以学习从开始到当前时刻的信息，并对下一个时刻的输出进行预测. 例如，在文本预测的任务中，对于当前时刻 t，输出 O_t 的概率分布为 $P(X_t\mid X_{t-1},X_{t-2},\cdots,X_1)$. 根据式(2)，对词袋中的每一个词 X_j 依次计算 $P(X_j)$，计算得到的概率即为词袋中每个词在下一个时刻出现的概率.

$$P(X_j)=\frac{\exp(\boldsymbol{W}_j\cdot\boldsymbol{O}_t)}{\sum_{j'=1}^{K}\exp(\boldsymbol{W}_{j'}\cdot\boldsymbol{O}_t)} \tag{2}$$

式(2) 中，$\boldsymbol{W}$ 为一个参数矩阵，$\boldsymbol{W}_j$ 是矩阵的第 j 行，$\boldsymbol{O}_t$ 是循环神经网络在 t 时刻的输出，K 是词袋的大小，$j\in[1,K]$. 由于标准的循环神经网络学习远距离信息的能力较弱，因此很多学者对标准的循环神经网络进行了改进. 例如：S.Hochreiter 等提出的长短时记忆网络(long short-term memory，LSTM)[10]，K.Cho 等提出的门控循环单元(gated recurrent unit，GRU)[11]，Tao 等提出的简单循环单元(simple recurrent unit，SRU)[12]. 由于长短时记忆网络能够记忆历史信息，学习远距离信息的能力较强，因此本文使用长短时记忆网络.

Seq2Seq 模型由 2 个循环神经网络组成：一个负责对输入序列进行编码，称为编码器(encoder)；一个负责对目标序列进行解码，称为解码器(decoder). Seq2Seq 模型的基本流程为：首先使用一个循环神经网络读取输入的句子，并将整个句子的信息压缩到一个固定维度的编码中；然后再使用另一个循环神经网络读取这个编码，将其“解压”为目标语言的一个句子. Seq2Seq 模型示意图如图 2 所示.

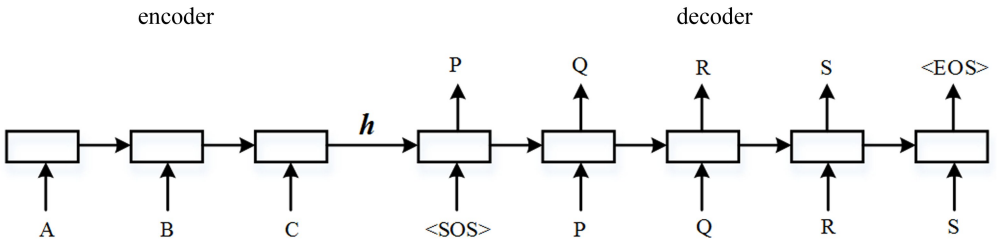

图 2 Seq2Seq 模型示意图

在图 2 中,A、B、C 是输入序列,〈SOS〉、P、Q、R、S、〈EOS〉是目标序列. Encoder 将来自输入序列的信息编码成一个中间的语义向量 $\boldsymbol{h}$,Decoder 接收语义向量 $\boldsymbol{h}$,并将其解码为输出序列. Seq2Seq 模型的目标函数是最大化对数似然函数 $P(Y \mid X)$:

$$P(Y \mid X) = \sum_{t=1}^{T} \log P(y_t \mid y_{1:t-1}, X), \tag{3}$$

其中 T 表示输出序列的时间序列大小,$y_{1:t-1}$ 表示输出序列的前 $t-1$ 个时间点对应的输出,X 为输入序列. 式(3)将自动摘要问题视为条件语言模型. 建立该模型的目的是为在已知输入文本的条件下,使生成整个目标摘要句的概率最大.

2 多语种自动摘要系统的设计

传统的基于 Seq2Seq 模型的自动摘要系统只能处理单一语种的文本. 在此以中文为例,描述其生成自动摘要的步骤.

Step1 对中文所有训练语料进行预处理,训练语料由文本和对应的摘要成对组成("文本-摘要"对). 首先分别获取训练语料的文本部分和摘要部分的词表,并在 2 个词表中都添加〈unk〉用以表示未登录词. 在输入端和输出端得到 2 个词表后,还需要在输出端的摘要词表中添加〈SOS〉和〈EOS〉,用以表示摘要的开始和结束.

Step2 对所有输入端的文本进行分词,并将文本的词项序列按照每个词项在词表中的 id 转换为数字序列,未登录词用词表中的〈unk〉来替代. 同时,将输出端所有摘要句的开头处加上〈SOS〉,在摘要句的末尾处加上〈EOS〉,并按照输出端的词表将摘要的词项序列转换为数字序列,未登录词用词表中的〈unk〉表示.

Step3 将所有的训练数据转换为数字序列后,用 Seq2Seq 模型对其进行训练.

若要使传统的自动摘要系统能够处理多语种的文本,则需要额外训练对应语种的模型,并将不同语种的模型结合起来使用. 以中、英、朝 3 种语种为例,传统的多语种自动摘要系统的工作方式如图 3 所示. 由图 3 可以看出,传统的多语种自动摘要系统需要先分别训练中文、英文和朝文的摘要模型,然后再将他们融合到一起. 输入文本后,系统首先进行语种识别,然后调用对应语种的摘要模型来生成摘要.

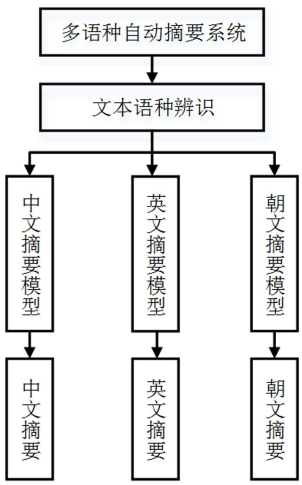

图 3 传统的多语种摘要模型

本文提出的多语种自动摘要系统的构建方式如下:

Step1 训练中文、英文、朝文 3 个语种语料的文本部分和摘要部分. 首先将 3 个语种的文本部分和摘要部

分分别放在一起，然后分别获取文本部分和摘要部分的词表，并在这 2 个词表中都添加〈unk〉用以表示未登录词. 在输出端得到摘要部分的词表后，在该词表中添加〈SOS〉和〈EOS〉，用以表示摘要的开始和结束.

Step2　对所有输入端的文本进行分词，并按照每个词项在词表中的 id 将文本的词项序列转换为数字序列，未登录词用词表中的〈unk〉来替代. 同时，将所有输出端的摘要句的开头加上〈SOS〉，在摘要句的最后加上〈EOS〉，按照输出端的词表将摘要的词项序列转换为数字序列，未登录词用词表中的〈unk〉表示. 这样，输入端和输出端的词表中都含有 3 种语言的单词. 训练数据由 3 部分组成：中文的“文本-摘要”对、英文的“文本-摘要”对和朝文的“文本-摘要”对.

Step3　将所有的训练数据转换为数字序列后，用 Seq2Seq 模型对其进行训练.

上述的多语种自动摘要模型如图 4 所示.

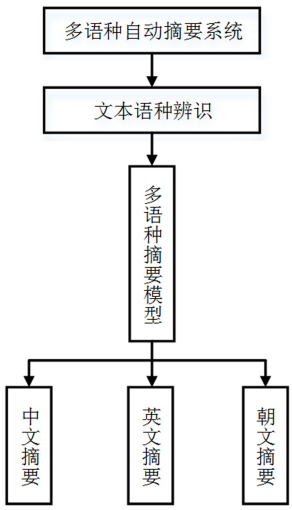

图 4　本文提出的多语种自动摘要模型

3　实验结果与分析

3.1　实验语料

本实验采用的数据均来自科技文献（包含生物、海洋和航空 3 个领域）的摘要和标题，其中将摘要作为模型的输入，将标题作为模型的输出. 实验的训练语料和测试语料的数量如表 1 所示.

表 1　训练和测试语料的数量

训练方式	训练语料数量	测试语料数量
单独训练中文	24 602	6 151
单独训练英文	24 562	6 141
单独训练朝文	24 602	6 151
中、英、朝一起训练	73 766	18 443

3.2　评价指标

评价指标采用 ROUGE（recall-oriented understudy for gisting evaluation）方法，该方法目前被广泛应用于摘要的自动评测任务中[13]. ROUGE 评价方法中包含 5 种评价指标，分别为 ROUGE-N、ROUGE-L、ROUGE-W、ROUGE-S、ROUGE-SU，其中每一种指标都要分别计算 P（precision，准确率）、R（recall，召回率）和 F（F-measure，F 值）. 目前，通常使用 ROUGE-1、ROUGE-2 和 ROUGE-L 指标中的 F 值作为自动摘要的评价指标. ROUGE-1 和 ROUGE-2 均属于 ROUGE-N 指标，其计算公式为：

$$ROUGE\text{-}N(R)=\frac{\sum_{S\in\{ref\text{-}summaries\}}\sum_{n\text{-}\mathrm{gram}\in S}Count_{\mathrm{match}}(n\text{-}\mathrm{gram})}{\sum_{S\in\{ref\text{-}summaries\}}\sum_{n\text{-}\mathrm{gram}\in S}Count(n\text{-}\mathrm{gram})},\tag{4}$$

$$ROUGE\text{-}N(P)=\frac{\sum_{S\in\{ref\text{-}summaries\}}\sum_{n\text{-}\mathrm{gram}\in S}Count_{\mathrm{match}}(n\text{-}\mathrm{gram})}{\sum_{S\in\{sys\text{-}summaries\}}\sum_{n\text{-}\mathrm{gram}\in S}Count(n\text{-}\mathrm{gram})},\tag{5}$$

$$ROUGE\text{-}N(F)=\frac{2PR}{P+R}.\tag{6}$$

其中：n-gram 表示 n 元词；$\{ref\text{-}summaries\}$ 表示参考摘要，即人工撰写的摘要；$\{sys\text{-}summaries\}$ 表示计算机生成的摘要；$Count(n\text{-}\mathrm{gram})$ 表示句子中的 n-gram 数量；$Count_{\mathrm{match}}(n\text{-}\mathrm{gram})$ 表示计算机生成的

摘要和参考摘要中同时出现 n-gram 的数量；N 表示 n-gram 的长度；$ROUGE\text{-}N(F)$ 为 $ROUGE\text{-}N(P)$ 和 $ROUGE\text{-}N(R)$ 的调和平均数. ROUGE-L 指标的计算公式如下：

$$ROUGE\text{-}L(R)=\frac{LCS(X,Y)}{len(X)}, \tag{7}$$

$$ROUGE\text{-}L(P)=\frac{LCS(X,Y)}{len(Y)}, \tag{8}$$

$$ROUGE\text{-}L(F)=\frac{2PR}{P+R}. \tag{9}$$

其中：X 表示参考摘要；Y 表示计算机生成的摘要；$len(X)$ 表示 X 的长度(词项个数)；$len(Y)$ 表示 Y 的长度；$LCS(X,Y)$ 表示 X 和 Y 的最大公共子序列的长度；$ROUGE\text{-}L(F)$ 为 $ROUGE\text{-}L(P)$ 和 $ROUGE\text{-}L(R)$ 的调和平均数.

3.3 实验及结果分析

实验的训练参数设置如下：Seq2Seq 模型中的编码器和解码器都采用长短时记忆网络结构，隐层神经元个数设为 100，迭代次数设为 20 次. 训练单语种摘要模型时，先对输入端的文本进行预处理，然后按词频从大到小的顺序保留前 5 000 个词作为输入端词表；输出端生成词表的方式和输入端一样，但由于输出端的摘要比输入端的文本长度短，因此保留前 2 000 个词作为输出端词表. 训练中、英、朝 3 种语种摘要模型时，输入端将 3 种语料的输入端词表均放在一起并去重，由此得到一个新的词表；输出端将 3 种语料的输出端词表均放在一起并去重后，由此得到一个新的词表.

实验时首先分开训练中、英、朝 3 个语种的语料，然后用各自的模型分别运行它们的测试集，并记录 ROUGE 评分. 再将中、英、朝 3 个语种的语料放在一起训练，然后运用该模型分别运行中、英、朝 3 种语料的测试集，并记录 ROUGE 评分. 实验结果如表 2 所示.

表 2 不同训练方式下的实验结果

训练方式	语言	ROUGE-1	ROUGE-2	ROUGE-L
单独训练(3 个模型)	中文	23.88	4.71	20.34
	英文	25.90	5.54	21.91
	朝文	22.64	5.15	18.86
共同训练(1 个模型)	中文	23.46	4.70	19.78
	英文	24.96	5.32	21.13
	朝文	22.64	5.16	19.98

从表 2 可以看出，本文方法在中文和英文测试集上的 3 个 ROUGE 指标仅略低于单独训练模型所得的 3 个指标，可忽略不计；而在朝文测试集上，两种方法的 3 个 ROUGE 指标基本相同. 这表明，本文的训练方法有效.

表 3 为同一文本在两种模型下所得的摘要样例. 由表 3 可以看出，在中文样例和英文样例中，使用本文方法生成的摘要比使用传统方法生成的摘要更接近参考摘要的语义；而在朝文样例中，使用传统方法的效果相对更好. 综合来看，本文方法生成摘要的效果与传统方法接近，因此表明本文方法有效.

4 结论

实验结果表明，将中、英、朝 3 个语种的语料放在一个模型中训练，其效果与各语言单独训练的效果接近，说明本文提出的将 3 种语种共同训练的方法是有效可行的. 由于本文方法比传统方法简洁、高效，因此具有更好的潜在应用价值. 本文在实验中，使用的训练语料规模较小，模型的泛化能力较弱，因此在今后的研究中我们将扩大训练数据规模，提高模型的泛化能力，以此进一步验证和完善本文方法.

表 3　同一文本在两种模型下所得的摘要样例

原文	参考摘要	本文方法生成的摘要	传统方法生成的摘要
从缢蛏(Sinonovacula constricta)cDNA 文库中筛选出一条铁蛋白同源序列，直接扩增质粒获得全长 cDNA 序列，共1 106 bp，包括 128 bp 的 5 非翻译区和 309 bp 的 3 非翻译区，以及 669 bp 的开放阅读框. 阅读框共编码 222 个氨基酸，N 端含有 17 个氨基酸的信号肽序列，推算的分子量约为 25.47 ku，理论等电点为 5.48……	缢蛏铁蛋白基因的分子特性及其表达分析	基因的克隆与表达分析	铁蛋白基因的表达分析
high frequency motion is always ignored in dynamic positioning operation and relevant paper pays attention to the effect of first order response on dynamic positioning accuracy via a deep water semi submersible drilling numerical model is calculated in both time domain simulations considering low frequency part only and considering that combined with first order motion then the statistic motion results with and without first order response are analyzed...	influence of the motion of deepwater platform on dynamic positioning precision (深水平台运动对动态定位精度的影响)	effect of platform on dynamic positioning (平台对动态定位的影响)	study on the effect of dynamic (动态效应的研究)
아미노산 분석 방법 은 유리 아미노산 폴리펩티드 또는 단백질 을 함유 한 샘플 중 의 아미노산 함유량 을 측정 하는 방법 이다 지난 여 년 동안 아미노산 분석 에 주로 전통 적 인 후컬럼 유도체 화 이온교환 크로마토그래피 와 전 컬럼 유도체 화.....	질량 분석 기술 이 아미노산 분석 방법 에서의 이용 (质谱法在氨基酸分析方法中的应用)	아미노산 의 아미노산 및 질량 분석 (氨基酸和氨基酸的质量分析)	아미노산 의 품질 분석 (氨基酸质量分析)

参考文献：

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

[2] MIHALCEA R, TARAU P. TextRank: bringing order into texts[C]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA: ACL, 2004:404-411.

[3] 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.

[4] PAGE L, BRIN S, MOTWANI R, et al. The pagerank citation ranking: bringing order to the web[J]. Stanford Digital Libraries Working Paper, 1998,9(1):1-14.

[5] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[EB/OL]. [2019-3-17]. https://arxiv.org/pdf/1409.3215.pdf.

[6] CHOPRA S, AULI M, RUSH A M. Abstractive sentence summarization with attentive recurrent neural networks [C]//Proceedings of NAACL-HLT. San Diego: NAACL, 2016:93-98.

[7] ZHOU Qingyu, YANG Nan, WEI Furu, et al. Selective encoding for abstractive sentence summarization[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2017:1095-1104.

[8] MA Shuming, SUN Xu, LIN Junyang, et al. A hierarchical end-to-end model for jointly improving text summarization and sentiment classification[EB/OL]. [2019-3-17]. https://arxiv.org/pdf/1805.01089.pdf.

[9] SATHASIVAM S, ABDULLAH W A T W. Logic learning in hopfield networks[EB/OL]. [2019-3-17]. https://arxiv.org/ftp/arxiv/papers/0804/0804.4075.pdf.

[10] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997,9(8):1735-1780.

[11] CHO K, GULCEHRE C, BAHDANAU D, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[EB/OL]. [2019-3-17]. https://arxiv.org/pdf/1406.1078.pdf.

[12] TAO Lei, ZHANG Yu, WANG Sida, et al. Simple recurrent units for highly parallelizable recurrence[EB/OL]. [2019-3-17]. https://arxiv.org/pdf/1709.02755.pdf.

[13] LIN C Y. ROUGE: A package for automatic evaluation of summaries[C]//In Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL. Stroudsburg, PA: ACL, 2004.