基于统一模型的藏文新闻摘要

闫晓东	解晓庆	邹煜	李维
中央民族大学	中央民族大学	中央民族大学	中央民族大学
信息工程学院	信息工程学院	信息工程学院	信息工程学院
国家语言资源监测	国家语言资源监测	国家语言资源监测	国家语言资源监测
与少数民族语言中心	与少数民族语言中心	与少数民族语言中心	与少数民族语言中心
yanxd3244@sina.com	xqplex@yeah.net	17820314536@163.com	1289773612@qq.com

摘要

Seq2seq神经网络模型在中英文文本摘要的研究中取得了良好的效果，但在低资源语言的文本摘要研究还处于探索阶段，尤其是在藏语中。此外，目前还没有大规模的标注语料库进行摘要提取。本文提出了一种生成藏文新闻摘要的统一模型。利用TextRank算法解决了藏语标注训练数据不足的问题。然后，采用两层双GRU神经网络提取代表原始新闻的句子，减少冗余信息。最后，使用基于注意力机制的Seq2Seq来生成理解式摘要。同时，本文加入了指针网络来处理未登录词的问题。实验结果表明，ROUGE-1评分比传统模型提高了2%。

关键词： 文本摘要；藏文；TextRank；指针网络；Bi-GRU

Abstractive Summarization of Tibetan News Based on Hybrid Model

Xiaodong Yan$_{1,2}$	Xiaoqing Xie$_{1,2}$	Yu Zou$_{1,2}$	Wei Li$_{1,2}$
yanxd3244@sina.com	xqplex@yeah.net	17820314536@163.com	1289773612@qq.com
Minzu University of China$_1$,		Minzu University of China$_1$,	
National language resource		National language resource	
monitoring &Research Center		monitoring &Research Center	
Minority Languages Branch$_2$		Minority Languages Branch$_2$	

Abstract

The sequence-to-sequence neural network model has achieved good results in the task of text rsummarization in Chinese and English, but the research of text summarization in low-resource languages is still in the exploratory stage, especially in Tibetan. What's more, there is no large-scale annotated corpus for summary extraction. In this paper, a hybrid model is proposed to generate Tibetan news summarization. We use the TextRank algorithm to solve the problem of lacking labeled training data in Tibetan. Then, we take two-layer Bi-GRU neural network to extract the sentences which represent the original news, and reduce redundant information. Finally, the Seq2Seq with attention model is used to generate the abstractive sumarization. Meanwhile, we add the pointer-network to deal with out-of-vocabulary words. The experimental results show that ROUGE-1 score increases by 2% than traditional model.

第十九届中国计算语言学大会论文集，第479页-第490页，海南，中国，2020年10月30日至11月1日。

Keywords: Text Summarization , Tibetan , TextRank , Pointer Network , Bi-GRU

1 引言

随着信息的爆炸式增长，人们难以高效、快速、准确地获取有价值的信息。为了解决这一问题，自动文本摘要技术应运而生，产生了对输入文本的简洁表示。文本自动摘要是自然语言处理领域的一个重要分支。它是一种利用计算机实现文本分析、内容归纳和自动文摘生成的信息压缩技术(Mani and Maybury, 1999)，帮助研究人员分析和总结冗长的文本，过滤掉多余的信息，从而提高浏览文本的速度。

文本摘要在信息检索中得到了广泛的应用，并取得了良好的效果。根据实现方法，文本摘要可以分为两类：抽取式摘要和理解式摘要。抽取式摘要是从原文中选择句子并将其组合起来生成摘要。而理解式摘要是对原文的重新解读而不是摘抄，对原文在语义上进行深层次理解，重新对文本进行表述，更加贴近人为表述方式。但这需要更先进的文本生成技术。由于抽取式摘要比理解式摘要更准确和可读，因此大多数研究都集中在抽取式摘要上(Gambhir and Gupta, 2017)。

随着深度学习技术的发展，基于注意力机制的seq2seq模型在文摘中取得了良好的效果(Rush et al., 2015)。与汉英相比，藏文文本摘要还处于探索阶段，面临着许多困难和挑战。首先，递归神经网络能够很好地对一个句子或一段文本进行编码，但不能很好地对整篇藏文文本进行编码。其次，缺乏大规模的文本摘要标注数据。最后，基于词的理解式摘要可能会出现未登录词的问题，从而影响摘要的可读性。

本文提出了一种将抽取式摘要和理解式摘要相结合的藏文摘要生成统一模型。首先，本文使用双向Bi-GRU神经网络从藏文新闻中提取句子。其次，将指针网络融入到基于注意力的seq2seq模型中，生成摘要。与其他模型相比，该模型能够有效地生成藏文摘要。

本文的主要贡献如下：

1) 提出了一个统一模型，它同时利用了抽取式和理解式的摘要方法。使用两层神经网络来提取能够表达原始语义的句子。采用基于注意力机制的seq2seq模型生成摘要，解决了藏文新闻篇幅过长的问题；

2) 引入文本秩算法对抽取的训练语料进行标记，作为神经网络模型的输入。它可以解决藏文标注语料库不足的问题；

3) 利用指针网络提高了藏文未登录词的处理精度，增加了摘要的可读性和新颖性。

2 相关工作

本文首先介绍抽取和理解式摘要的相关工作，然后介绍藏文文本摘要的相关工作。

IBM的Luhn首先提出了基于词频和分布的句子评分模型来提取“自动摘要”，这是机器生成的提取摘要的第一个例子(Luhn, 1958)。摘要抽取的目的是抽取句子来概括文章的中心思想。这些句子被称为关键句，它们是通过分析词频、标题、位置、句法结构、线索词等获得的。传统的提取算法大致分为四类：(1)基于统计的方法。句子的权重是根据词频、位置等信息计算出来的，然后按降序排列。权重值最高的句子被确定为摘要。这种方法的提取速度快，但不能提取

句子的内部信息，导致摘要质量差(Brandow, 1995)。(2)基于图的方法。将文章转化为拓扑图，通过递归和迭代运算使句子权重稳定。对句子进行排序和加权，选择权重最大的句子作为总结，例如TextRank和LexRank算法(Mihalcea and Tarau, 2004)。(3)基于文档主题的方法，利用主题模型提取隐藏信息，例如LDA算法(Sun, 2017)。(4)基于整数规划的方法。它通过将抽取的摘要转化为整数线性规划来寻找全局最优解(Xie, 2011)。目前，随着大数据、云计算等技术的发展，深度学习方法在NLP任务中取得了良好的效果，尤其是在文本摘要方面。SummaRuNNer是一个典型的文本过滤网络(Nallapati et al., 2017)，它将句子抽取问题转化为二分类问题。在英语语料库中，ROUGE-1的得分达到39.6%。Yin等人提出了一种新的基于CNN的网络语言模型（CNNNLM），将句子表示为一个密集向量进而计算句子冗余度，ROUGE-1评分达到42.3%(Yin and Pei, 2015)。Cheng等人使用基于注意力机制的LSTM对每个句子进行分类(Cheng and Lapata, 2016)，在长文本中，ROUGE-1得分达到33%。

随着语料库的不断扩展，机器学习方法被应用于抽象文摘中。传统的统计方法可分为三类：(1)朴素贝叶斯模型，它将朴素贝叶斯分类器与自动摘要结合起来(Chopra et al., 2016)。(2)隐马尔可夫模型，它将隐马尔可夫模型与自动摘要结合起来(Nallapati et al., 2016)。(3)将条件随机场与自动摘要结合起来的概率图模型(See et al., 2017)。同时，这种深度学习方法也取得了较好的效果。2015年，Rush等人使用序列到序列模型（Seq2Seq）和注意力机制生成文本摘要(Rush et al., 2015)。模型采用了编解码框架。编码器使用LSTM网络嵌入句子，解码器使用RNNLM生成摘要。在DUC-2004和Gagiword数据集中，ROUGE-1得分达到28.18%。但是嵌入层无法学习到深层的语义信息。为了解决这个问题，Sumit等人改进了模型，在编码器层，CNN被用来压缩字符作为GRU-RNN的输入。然后在解码层使用RNN(Chopra et al., 2016)。ROUGE-1得分达到32.75%。但是，摘要中的单词都来自词汇表，总是不断重复。2018年，谷歌推出了指针网络来解决词汇表外（OOV）问题(See et al., 2017)，它指向源文本并复制词汇表中没有出现的单词。此外，它还使用了覆盖机制来跟踪摘要的内容，从而减少重复。ROUGE-1得分达到39.53%。

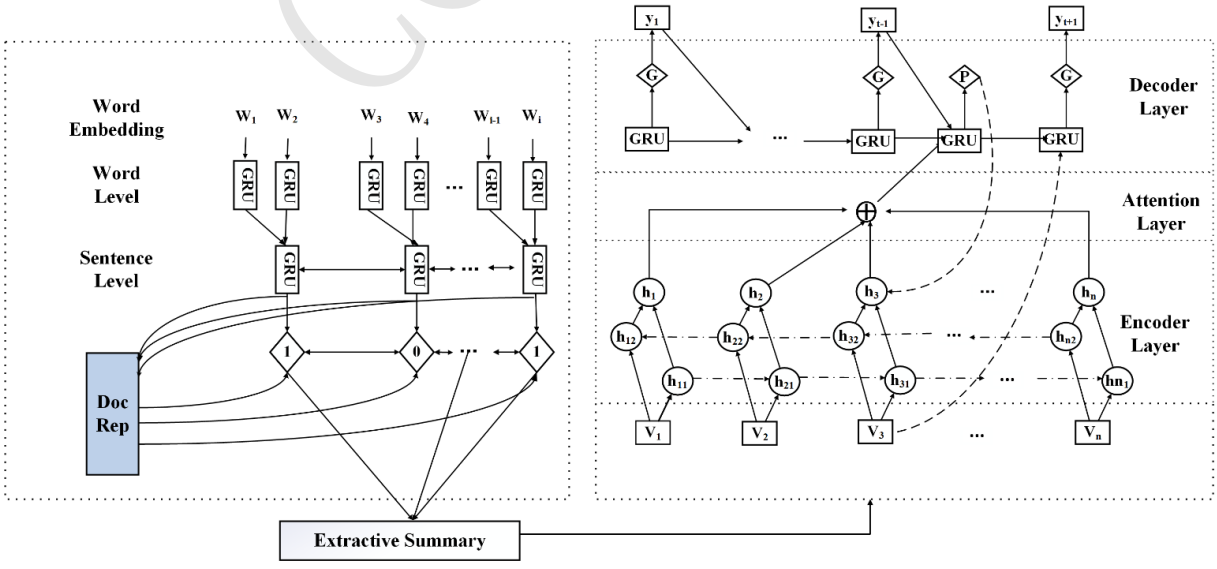

图 1: 统一模型示意图

由于缺乏大规模的训练语料库，目前对藏文文本摘要的研究较少。安见才让对藏文新闻进行了爬取过滤，提取了词频、标题、位置、句法结构、指示短语等五个特征作为权重(Anjian, 2010)。选取权重最大的句子作为新闻摘要的抽取。在此基础上，南奎娘若融合了分词、词性标注、缩略词、句子边界识别、停词选句等功能(Nankui and Anjian, 2016)。除此之外，在藏语中，没有提取摘要的基线，也没有关于提取摘要的研究。

3 模型架构

本文提出的模型架构如图1所示。该模型不仅采用TextRank算法构造了抽取式摘要的训练语料库，而且还训练了一个双层Bi-GRU网络来抽取藏文新闻中的句子。然后将提取出的句子输入seq2seq模型，根据注意力机制和指针机制生成摘要。模型主要由三部分组成:

1) 采用TextRank算法解决了低资源语言训练语料库不足的问题，并利用外部知识库对藏文新闻进行了标注。然后，在对TextRank算法进行迭代后，得到一个可用于训练抽取型网络的训练语料库。

2) 利用标注的语料库训练双层Bi-GRU抽取型网络，如图1左图所示。对于第一层，它用于获取字级信息。第二层则是从句子层面获取信息，获取藏文新闻中的文献信息。最后，根据新闻的文档表示、提取的摘要和隐藏层的状态来确定当前语句是否标记为1或0。

3) 以藏语句子1作为理解型模型的输入。如图1右图所示，理解型模型的总体架构采用seq2seq，编码端采用Bi-GRU，解码端采用RNN。为了获取关键信息，在解码端引入了注意力机制，结合指针网路解决了OOV问题。

4 模型描述

4.1 基于改进TextRank的摘要抽取

TextRank算法是PageRank算法的变体，PageRank算法是一种链接分析算法。谷歌用它来分类和估计网页的价值。它通常用于有向图中，并按指向前驱和后继的边数迭代。迭代运算如公式(1)所示。

$$S(V_i) = (1-d) + d \times \sum_{V_j \in In(V_i)} \frac{1}{|Out(V_j)|} \cdot S(V_j) \tag{1}$$

其中$S(V_i)$表示节点V_i的权重。$In(V_i)$是节点的入度，即指向此网页的URL数。$Out(V_j)$表示节点的出度数。d是阻尼系数，通常取0.85 。

本文将此思想应用于文本摘要的提取。根据句子的关联图确定句子的相对重要性。传统的算法大多忽略了文档的语义和语法信息。只把新闻看作是一个独立的词的集合，而没有考虑词与词之间的联系。将语料库等外部知识融入到文本摘要算法中，提高了算法的准确性。具体方法如下。首先，TextRank算法根据藏文新闻生成拓扑图，表示为$G = (V, E)$ 。 G表示无向图，其中V是顶点集，即新闻中的句子。E是一组边，表示句子之间的关系。本文使用TextRank算法迭代图模型直到收敛。然后每个顶点都有一个表示句子重要性的分数。分数最高的句子被提取出来作为摘要。该过程主要分为四个步骤:

1) 在对藏文新闻句子进行分割后，将每一个句子作为节点添加到图模型中;

2) 句子的矢量表示是同一维度上所有词矢量的平均值。边表示句子之间的相似度，如公式(2)所示。

$$WS(S_i, S_j) = \cos(S_{i1} \cdots S_{im}, S_{j1} \cdots S_{jm}) \tag{2}$$

其中S_i和S_j是句子向量，cos是句子S_i和S_j之间的余弦距离，n表示单词向量的维数。本文还比较了不需要用高维向量表示的共现矩阵计算的相似度。两个句子中同时出现的词的平均权重用作边的权重。

3) 迭代算法直到收敛，如公式(3)所示。

$$WS(V_i) = (1-d) + d \times \sum_{V_j \in In(V_i)} \frac{W_{ij}}{\sum_{V_k \in Out(V_j)} W_{kj}} \cdot WS(V_j) \tag{3}$$

其中W_{ij}表示节点V_i和V_j之间的边的权重，该权重由相似度表示。$In(V_i)$表示指向的V_i节点，$Out(V_j)$表示从V_j指向其他节点的节点。在TextRank算法中，所有节点的初始得分一般为1，当某个节点的误差小于0.0001时，迭代停止。

4) 根据收敛得分对节点进行排序。与标题相似的句子更有可能是摘要，因此在本文提出的改进模型中适当增加了这些句子的权重。在生成的藏文摘要中，几个权重较大的句子的相似度一般很大。本文引入惩罚系数以避免摘要中的句子冗余问题。摘要相似度高的句子乘以惩罚系数以降低权重。

4.2 基于Bi-GRU的文摘抽取

经过TextRank算法处理后，藏文新闻文章可以表示为一个由0和1组成的向量，这个向量的维数就是句子的个数。0表示句子未被选中，1表示摘要被选中。这样，句子抽取问题就可以概括为序列标记问题。递归神经网络（RNN）能很好地求解序列数据。但是，由于后面的节点对前面节点的感知度较低，本文使用了一种称为GRU的RNN变体。GRU由一个更新门和一个重置门组成。更新门决定将以前的内存保存到当前状态的程度，重置门决定如何将新信息与以前的信息融合。在时刻t，本文根据传输状态h_{t-1}和电流输入x_t得到两个门控状态。r_t是重置门，z_t是控制更新状态的门。更新和重置规则如式(4)-(5)所示。

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{4}$$

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{5}$$

其中σ是sigmoid函数，其目的是将输入数据转换为0-1的范围。W_r和W_z是训练好的参数。在GRU的隐藏层中，先前的状态h_{t-1}被重置并用x_t拼接。使用激活函数tanh得到的输出h_t。然后，使用更新门z执行遗忘和存储选择，如公式(6)-(7)所示。最后，得到新的隐藏状态h_t和输出y_t，如公式(8)所示。

$$h_t = \tanh(W_h \cdot [r_t \times h_{t-1}, x_t]) \tag{6}$$

$$h_t = (1 - z_t) \times x_t + z_t \times h_t \tag{7}$$

$$y_t = \sigma(W_y \cdot h_t) \tag{8}$$

其中y_t是值为0或1的句子的标签。单向GRU只能获得一个方向的信息，而双向GRU(Bi-GRU)可以连接隐藏层中的前向和后向传播状态。

本文采用两层Bi-GRU来更好地提取深层语义特征。第一层获得字级信息。执行最大池操作，将每个句子中单词的隐藏状态作为第二层句子单元的输入。第二层得到藏文新闻d的句子级信息和文档表示，如式(9)所示。

$$d = \tanh(W_d \frac{1}{N_d} \sum_{j=1}^{N^d} [h_j^f, h_j^b] + b) \tag{9}$$

其中h_i^f和h_i^b是句子的前向和后向隐藏层状态，N_d是文档中句子的个数，矩阵W_d和偏置b是可训练的参数。

在分类过程中，模型根据文档表示、隐藏层状态、位置信息、生成摘要四个方面共同决定句子是否被选中，如公式(10)所示。

$$P(y_j = 1|h_j, s_j, d) = \sigma(W_c h_j + h_j^T W_s d - h_j^T W_r \tanh(s_j) + W_{ap} p_j^a + W_{rp} p_j^r + b) \tag{10}$$

其中W_c，W_s，W_r，W_{ap}，W_{rp}和b是需要训练的参数，y_j表示是否选择此句子作为摘要，h_j是表示句子级网络隐藏状态的输出，d是经过非线性变换后的文本表示，s_j是位置$j^t h$的动态摘要表示，p_j^a是绝对的位置向量，p_j^r是相对位置向量。减法运算用于删除冗余信息。

4.3 基于指针网络的文摘

1) 基于注意力机制的seq2seq 序列到序列模型结合了两个递归神经网络。一个负责接收提取的句子；另一个负责根据前一个网络的隐藏状态生成藏文新闻摘要，分别称为编解码过程。编码过程实际上是利用RNN的记忆功能，根据上下文的顺序关系，将字向量按顺序输入网络，并保留最后的隐藏状态。它同样可以压缩整个句子并将其存储为上下文向量。

在解码过程中引入了注意力机制来分配序列的权重。通过加权变换提高了精度。如公式(11)-(12)所示生成摘要。

$$s_r = f(y_{t-1}, s_{t-1}, c) \tag{11}$$

$$Y_i = softmax(S_t) \tag{12}$$

其中Y_i表示生成的藏文摘要的$i^t h$单词，由三个状态确定：y_{i-1}，s_i，c_i。s_i表示时刻i的隐藏状态，该状态由c_i，s_{i-1}，y_{i-1}决定。c_i表示注意加权的内容向量，其内容向量c_i如式(13)-(15)所示。

$$c_i = \sum_{j=1}^{T_x} \alpha_{i,j} h_j \tag{13}$$

$$e_{i,j} = a(s_{i-1}, h_j) \tag{14}$$

$$\alpha_{i,j} = \frac{exp(e_{i,j})}{\sum_{k=1}^{T_x} exp(e_{i,k})} \tag{15}$$

其中$e_{i,j}$表示解码器隐藏层状态s_{i-1}和编码器隐藏层状态h_j的线性组合，$\alpha_{i,j}$表示通过注意机制学习的每个单词的权重。

2) 指针机构

本文在解码层采用指针网络来解决藏文OOV问题。指针机制是注意力机制的变体。它用于确定目标单词y_t是由词汇表中的RNN选择的，还是通过设置指针开关直接从输入文本中复制的。当选择P模式时，解码器从输入句子中复制单词。当选择G模式时，解码器从词汇表中选择单词。该模型使用注意分布矩阵来确定所选择的指针模式，如公式(16)所示。

$$p(C_i|C_1,\cdots,C_{i-1},X) = softmax(e^t) \tag{16}$$

其中$C_{j(1<j<i-1)}$是已生成的摘要，X是解码器层的当前状态，e^t是输入的注意权重。

5 实验

5.1 数据集

英语文本摘要有一些开源数据集，如DUC(Barrera and Verma, 2011)、gigaword(Napoles et al., 2012)和CNN/Daily数据集(Nallapati et al., 2016)。由于缺乏大规模的藏文文本摘要数据集，本文采用新闻标题作为参考摘要。语料库来源于中央民族大学自然语言处理实验室的舆情项目，共收录藏文新闻约50,000条。

5.2 数据预处理

藏语是一种以字符为基本单位，以“·”分隔的字母语言。一条竖线“ | ”表示短句的结尾。因此，本文首先将句子以“ | ”分隔，然后使用TIP-LAS工具对爬网语料库进行分段和词性标记(Li et al., 2018; Li et al., 2015)，然后，将数据作为Word2vec和Fastext模型的输入，分别生成词向量。最后，该模型还生成句子向量。

5.3 评测方法

评测方法是自动文摘研究的关键。评价方法可分为内部评价法和外部评价法。前者通过直接分析摘要的质量来评价。后者将其应用于特定的任务，如自动问答、文本分类等，并根据客观结果对其性能进行评估。相较于自动评价方法，手工评估方法成本高，主观上缺乏一定的公平性。目前，Lin等人参考机器翻译自动评测方法BLEU(Papineni et al., 2002)，提出了ROUGE(Recall-Oriented Understudy for Gisting Evaluation)评测方法(Lin, 2004)。它首先形成由多个专家生成的标准汇总集。然后，与模型生成的自动摘要进行比较。最后，对重叠的基本单元进行统计，评价摘要的质量。ROUGE已成为总结评价技术的通用标准之一。ROUGE系列评价指标包括ROUGE-N、ROUGE-L、ROUGE-S、ROUGE-W。最常见的评价指标是ROUGE-N。它基于n-gram共现统计。n的范围是从1到4。计算如公式(17)所示。

$$ROUGE-N = \frac{\sum_{S\in\{Refsummaries\}}\sum_{n-grams\in S} Count_{match}(n-gram)}{\sum_{S\in\{Refsummaries\}}\sum_{n-grams\in S} Count(n-gram)} \tag{17}$$

其中$Refsummaries$表示引用摘要，$Count(n-gram)$表示引用摘要中的个数，$Count_{match}(n-gram)$表示生成的摘要和引用摘要中的公用个数。

ROUGE-L是基于最长公共子串的统计，ROUGE-S基于词对的统计序列，ROUGE-W则被认为是基于ROUGE-S的字符串的连续匹配。不同的方法对不同类型的总结评价有不同的影响。

5.4 参数设置

本文设置了TextRank、抽取型网络和理解型网络模型的参数，如表1-3所示。

Parameter	Value
Gini coefficient	0.75
Iteration number	1000
Stop iteration value	0.001
Redundancy coefficient	0.5

表 1: TextRank主要参数

Parameter	Value
Hidden size	64
N_layers	2
Batch_size	20
Epoch	50
Learning_rate	0.01

表 2: 抽取型网络主要参数

Parameter	Value
Hidden size	64
Batch_size	20
Epoch	100
Learning_rate	0.01
Vocab_size	5000

表 3: 理解型网络主要参数

5.5 实验结果

1) 抽取式摘要

本文使用ROUGE作为评价，并进行以下实验。

TF-IDF：本文使用TF-IDF方法计算单词的权重。词的权重之和构成句子权重。提取的按权重排序的摘要用作基线。

TR+WF：本文使用TextRank算法提取句子，并用词频共生矩阵计算相似度。

TR+Fasttext：在TextRank迭代中，本文使用Fasttext模型生成句子向量并计算相似度。

TR+Word2vec：在TextRank迭代中，本文使用Word2vec模型生成句子向量并计算相似度。

Bi-GRU：本文使用双层Bi-GRU神经网络提取句子作为摘要。

为了提高TextRank算法的性能，引入外部知识库。Word2vec和Fastext模型生成的藏文文件大小见表4，实验结果见表5。

Word2vec		Fasttext	
Corpus	3.2GB	Corpus	3.2GB
Size	167MB	Size	157MB
Dimension	100	Dimension	300

表 4: 语料库以及Word2vec和FastText生成文件大小

Model	Rouge-1	ROUGE-2	ROUGE-L	Time(h)
TF-IDF	16.4	7.9	11.4	0.5
TR+WF	21.3	10.4	21.6	14
TR+Fasttext	26.6	11.1	22.6	30
TR+Word2vec	32.7	18.9	29	23
Bi-GRU	20.1	11.3	15.2	1

表 5: 抽取式摘要结果

根据表5可以发现，TextRank算法比其他方法取得了更好的性能。与传统的词共现矩阵和TF-IDF相比，在整合外部知识库时，ROUGE评分分别提高了5.3%、0.7%和1.0%。这意味着

Title:

དབྱི་རན་གྱི་ཧྲུང་ཧྲང་ལྷའུ་ཀ་ཉེས་ཡུག་རེན་གྱི་གནམ་གྲུའི་དོན་ཆུན་གྱི་རྒྱུ་ཆུན་ལ་ཞུར་བས་རྟོག་ཞིབ་བྱ་རྒྱའི་རེ་བ་བཏོན་པ།

(Iranian President Rouhani calls for further investigation into the cause of the Ukrainian plane incident.)

Source Text:

དབྱི་རན་གྱི་ཧྲུང་ཧྲང་ལྷའུ་ཀ་ཉེ་ཡིས་ཚེས་14ཉིན་དབྱི་རན་རྒྱལ་ཁབ་བརྒན་འཛིན་ཁང་གི་ཞང་གོང་ཁོད་དབྱི་རན་ཁྱིམས་འཛིན་རྩེ་ལག་གིས་མཐོ་རིམ་ཁྱིམས་ཁང་དང་ཚེད་མཁས་པ་བཅུ་ཡིས་གྲུབ་བའི་ ”དམེགས་བསལ་ཁྱིམས་ར་”འཛུགས་དགོས་པའི་བློ་ལང་ཐེབས། གོམ་གང་མའུན་སྤྱོས་ངང་ཤུའུ་ཁོ་ལན་གཅམ་གྲའི་དོན་ཆུན་གྱི་རྒྱུ་ཆུན་ལ་རྟོག་ཞིབ་དགོས་པ་བཤད། ཁོ་ང་གིས། ”འཕྲལ་རྒྱེལ་གནམ་གྲ་ས་ལ་ཤུང་བའི་དོན་ཆུན་ནི་རྒྱུན་ཧན་ཞིག་མིན། ”འཛམ་གླིང་ཡོངས་ཀྱིས་དབྱི་རན་གྱིས་ཇི་ཤ་ཐག་གཅོད་པར་ཤ་བཞིན་ཡོད། ”ཅེས་གསུངས་ཡོད། དུས་མཚུངས་སུ། ལྷའུ་ཀ་ཉེ་ཡིས་དབྱི་རན་དམག་སྡོགས་ཀྱིས་འདི་ཇོན་དུ་“འཛེམ་ཟོལ་མེད་པར་ནོར་འཁྲུལ་ཁས་བླངས་པའི་”རྣམ་འགྱུར་ལ་ཡི་རང་ཡོད་པ་དང་། རྟོག་ཞིབ་མཇུག་འཕྲས་ཀྱིས་མེ་ཚང་མར་ལན་ཁ་སྨོད་ཧྲུབ་པའི་རེ་བ་ཡོད་ཅེས་བཤད།

(Iranian President Rouhani said on Iran's national television live broadcast on the 14th that the Iranian judiciary called for the establishment of a "special court" consisting of ten high courts and experts. Further investigate the causes of the Ukrainian aircraft incident. He said,"The crash of the passenger plane is abnormal. The whole world is watching how Iran handles it." At the same time, Rouhani said that the Iranian military expressed its admiration for its "sincere admission". I hope the survey results will give you a response.)

Extract Result:

དབྱི་རན་གྱི་ཧྲུང་ཧྲང་ལྷའུ་ཀ་ཉེ་ཡིས་ཚེས་14ཉིན་དབྱི་རན་རྒྱལ་ཁབ་བརྒན་འཛིན་ཁང་གི་ཞང་གོང་ཁོད

(Iranian President Rouhani said on Iran's national television live.)

གོམ་གང་མའུན་སྤྱོས་ངང་ཤུའུ་ཁོ་ལན་གཅམ་གྲའི་དོན་ཆུན་གྱི་རྒྱུ་ཆུན་ལ་རྟོག་ཞིབ་དགོས་པ་བཤད།

(Further investigation into the cause of the Ukrainian plane incident.)

འཕྲལ་རྒྱེལ་གནམ་གྲ་ས་ལ་ཤུང་བའི་དོན་ཆུན་ནི་རྒྱུན་ཧན་ཞིག་མིན།

(the airliner crash is unusual.)

ལྷའུ་ཀ་ཉེ་ཡིས་དབྱི་རན་དམག་སྡོགས་ཀྱིས་འདི་ཇོན་དུ”འཛེམ་ཟོལ་མེད་པར་ནོར་འཁྲུལ་ཁས་བླངས་པའི་”རྣམ་འགྱུར་ལ་ཡི་རང་ཡོད་པ་དང་།

(Rouhani said that the Iranian military expressed its admiration for its "sincere admission".)

Our Model:

དབྱི་རན་ ལྷག་ཀ་ཉེན་ སྤྱོན་པ་ སྤྱོན་པ་ གནམ་གྲ། རྒྱུན་ཞན་མེན་པ་ བརྟག་དཔྱད། དབྱི་རན་

(Iran Rouhani says says planes abnormal investigated Iran.)

表 6: 基于统一模型的摘要生成实例

外部知识库提高了TextRank算法的抽取性能。与Fastext模型相比，Word2vec模型的ROUGE评分提高了9%。证明了Word2vec模型的有效性。然而，TextRank算法的迭代时间太长，不适合大规模语料库。Bi-GRU神经网络的性能不如TextRank算法。但是，在相同的语料库规模下，所需的迭代时间小于1小时，更适合大规模语料库。

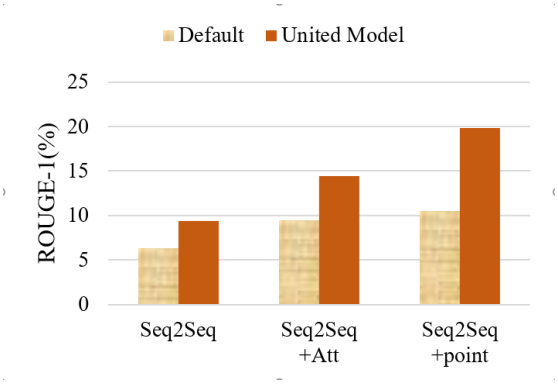

图 2: ROUGE-1实验评测结果

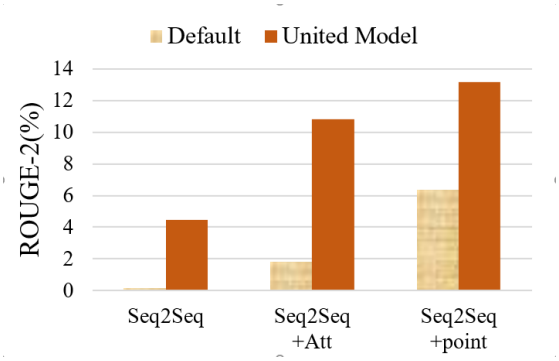

图 3: ROUGE-2实验评测结果

2) 理解式摘要

默认：本文分别使用带有注意机制的Seq2Seq模型和带有指针机制的Seq2Seq模型生成摘要。

统一模型：首先使用双层Bi-GRU提取最能表达原始新闻的句子，然后分别使用带有注意力机制的Seq2Seq模型和带有指针机制的Seq2Seq模型生成摘要。实验结果如图2和图3所示。

从图2，3中，本文可以观察到以下情况： 1) 默认模型显示的总体效果很差。主要由于藏文新闻文本太长，导致神经网络无法对新闻进行良好编码。在藏文新闻篇幅不缩短、摘要直接从原文中产生的条件下，“ROUGE-1”和“ROUGE-2”的得分趋于接近零，说明所产生摘要的可读性和连贯性都较差。与默认模型相比，统一模型得到的结果有了很大的改进，进一步证明了抽取式摘要和理解式摘要的结合不仅可以压缩文档，而且可以删除冗余信息，从而解决了藏文新闻长文本无法编码的问题。2) 新闻标题中的许多词汇都来源于新闻，而使用默认模型生成的摘要只包括藏文词表中的词汇。但藏语词表中不存在人名、地名等专有名词，评价结果较差。添加指针机制后，文本可以根据注意力从原始文本中复制出来，生成的摘要更接近新闻标题。那么Seq2Seq+attention+point模型得到的ROUGE-1分数比Seq2Seq+attention模型高5%。而且，得到的ROUGE-2评分提高了2%，证明了指针机制能够更好地摘要提高的质量。

表6给出了使用带指针机制的双层Bi-GRU网络从新闻中提取摘要的例子。用双层Bi-GRU神经网络选出四个句子。本文可以看到像“ཨང་གི་མ།”这样的多余句子和短语被删除了。理解式模型生成的摘要可以粗略地表达标题中包含的信息，证明了模型的有效性。单词“ལུའ་ཏ་ཞི་”不出现在词汇表中，但指针网络可以在原始文本中指向该单词并复制它以生成摘要。证明了指针机制可以解决OOV问题，增加了摘要的新颖性。此外，本文发现“སོོན་པ་”一词出现了两次。这说明生成的摘要存在重复性问题，这与传统的中英文文本摘要模式相似。

6 总结

本文提出了一个藏文新闻摘要生成的统一模型。在该模型中，结合了抽取式摘要和理解式摘要的优点，解决了神经网络无法对太长的藏文新闻进行编码的问题。在藏文摘要的生成过程

中，采用了指针机制和注意机制来解决与OOV相关的问题。然而，仍有许多困难有待解决。首先，作为参考摘要的标题不能包含原文的重要信息。其次，生成的摘要存在语义重复问题。今后，本文将使用K-Means聚类方法生成参考摘要，以提高原始信息覆盖的准确性。然后，本文将使用覆盖机制来解决语义重复问题。

参考文献

Anjian Cairang. 2010. Research on Automatic Summarization of web pages in Tibetan search engine system. *Microprocessors*, 31(5):77–80. *In Chinese.*

Araly Barrera and Rakesh Verma. 2011. Automated Extractive Single Document Summarization: Beating the Baselines with a New Approach. Proceedings of the 2011 ACM Symposium on Applied Computing, TaiChung, Taiwan, China, 268–269.

Ronald Brandow, Karl Mitze, and Lisa F. Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675–685.

Jianpeng Cheng and Mirella Lapata. 2016. Neural Summarization by Extracting Sentences and Words. *Association for Computational Linguistics.* Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 484–494.

Sumit Chopra, Michael Auli, Alexander M. Rush. 2016. Abstractive Sentence Summarization with Attentive Recurrent Neural Networks. *Association for Computational Linguistics.* Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, California, 93–98.

Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47:1–66.

Bohan Li, Huidan Liu, Congjun Long and Jian Wu. 2018. Tibetan word segmentation based on deep learning. *Computer Engineering and Design*, 39(01):194–198. *In Chinese.*

Yachao Li, Jing Jiang and Jiayangji. 2015. Tip-las: an open source tagging system for Tibetan word segmentation. *Journal of Chinese Information Processing*, 29(6):203-207. *In Chinese.*

Chin Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. *Association for Computational Linguistics.* Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 74–81.

H. P. Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Inderjeet Mani and Mark T. Maybury. 1999. Adavnces in Automatic Text Summarization. *Computational Linguistics*, 26(2):280–281.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text. *Association for Computational Linguistics.* Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 404–411.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. SummaRuNNer: a recurrent neural network based sequence model for extractive summarization of documents. *AAAI'17.* Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 3075–3081.

Ramesh Nallapati, Bing Xiang, and Bowen Zhou. 2016. Sequence-to-Sequence RNNs for Text Summarization. *ArXiv.* abs/1602.06023.

Nankui Niangruo and Anjian Cairang. 2016. Research on Extraction of Tibetan text Abstract Based on sensitive information. *Network Security Technology and Application*, 4:58–59. *In Chinese.*

Courtney Napoles, Matthew Gormley and Benjamin Van Durme. 2012. Annotated Gigaword. *Association for Computational Linguistics.* Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction, Montreal, Canada, 95–100.

Kishore Papineni, Salim Roukos, Todd Ward, Weijing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. *Association for Computational Linguistics*. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, 311–318.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. *Association for Computational Linguistics*. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 379–389.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. *ArXiv*. abs/1704.04368.

Guochao Sun. 2017. Research and implementation of Web Text Summarization System Based on LDA topic model. Shandong University of science and technology. *In Chinese*.

Yan Xie. 2011. Research on Automatic Summarization System Based on LSA and paragraph clustering. Liaoning University of science and technology . *In Chinese*.

Wenpeng Yin and Yulong Pei. 2015. Optimizing sentence modeling and selection for document summarization. *IJCAI'15*. Proceedings of the 24th International Conference on Artificial Intelligence, 1383–1389.