文章编号：1003-0077（2017）00-0000-00

基于联合模型的藏文实体关系抽取方法研究

夏天赐[1,2] 孙媛[1,2]

（1.中央民族大学 信息工程学院，北京 100081；
2.中央民族大学 国家语言资源监测与研究中心少数民族语言分中心，北京 100081）

摘要：从无结构文本中抽取实体与实体之间的关系是自然语言处理领域的重要研究内容,同时也为构建知识图谱、问答系统等应用提供重要支撑。基于联合模型的实体关系抽取任务将实体识别和关系抽取同时进行，克服了传统实体关系抽取任务中先识别句子中的实体，然后在进行实体关系判断这两次任务中的错误累加。该文针对藏文语料匮乏，实体识别准确率不高等问题，提出了基于联合模型抽取藏文实体关系的方法。该文基于藏文实体关系抽取任务，提出以下方案：（1）针对藏文分词准确率不高的问题，对藏文进行字级和词级两种方式进行预处理，并给出对比实验，结果表明采用字级处理方式较词级处理方式效果有所提高。（2）藏语是一种语法规则比较强的语言，名词、格助词等能明确指示句子各组块之间的语法和语义结构关系，因此该文将藏文的词性标注特征加入到藏文的字词向量中，实验结果表明方法的有效性。（3）该文借鉴了联合模型处理的优势，提出基于联合模型处理方式，采用端到端的 BiLSTM 框架将藏文实体关系抽取任务转变为藏文序列标注的问题，实验结果表明，该文的方法较传统的基于藏文处理方式如 SVM 算法和 LR 算法，准确率提高了 30%-40%。

关键词：联合模型;藏文实体关系;词性标注

中图分类号：TP391 **文献标识码：**A

Tibetan Entity Relation Extraction Based on Joint Model

Tianci Xia[1,2], Yuan Sun[1,2]
(1.School of Information Engineering, Minzu University of China, Beijing 100081, China ;
2. Minority Languages Branch, National Language Resource and Monitoring Research Center,
Minzu University of China, Beijing 100081, China)

Abstract : Extracting the relationship between entities and entities from unstructured texts is an important research area in the field of natural language processing. It also provides important support for the construction of knowledge graph, question answering systems and other applications. The entity relation extraction task based on the joint model performs the entity identification and relation extraction at the same time, which overcomes the errors by first identifying the entity in the sentence, and then performing relationship judgment task. In lack of Tibetan corpus and accuracy of entity recognition is low, this paper proposes a method for extracting Tibetan entity relations based on joint model. This paper proposes the following: In order to solve the problem that the accuracy of Tibetan word segmentation is low, Tibetan language is preprocessed in both word-level and character-level, and comparative experiments are conducted. The results show that the use of character-level processing method is more effective than the word-level. (2) Tibetan is a language with strong grammatical rules. Verbs, lattices, and auxiliary particles can clearly indicate the grammatical and semantic structure relationships between the various chunks of the sentence. Therefore, this paper adds the Tibetan part of speech tagging feature to the Tibetan. In the character vector, experimental results show the effectiveness. (3) This paper draws on the advantages of the joint model processing, proposes a method based on the joint model, and adopts the end-to-end BiLSTM framework to transform the Tibetan entity relationship extraction task into a Tibetan sequence annotation. The experimental results show that the method is greated than tranditional methods such as SVM and LR, the accuracy rate is improved by 30%-40%.

Key words: Joint model; Tibetan entity relation; POS tagging

收稿收稿日期：2017-03-16；**定稿日期**：2017-04-26

基金项目：国家自然科学基金（No. 61501529, No. 61331013）；国家语委项目(No. YB125-139, ZDI125-36)

0 引言

实体关系抽取任务作为信息抽取领域的重要研究课题，其主要目的是抽取句子中已标记实体对之间的语义关系，即在实体识别的基础上确定无结构文本中实体对间的关系类别，并形成结构化的数据以便存取。例如，<e1>马云</e1>是<e2>阿里巴巴</e2>的创始人之一。实体关系抽取能自动识别实体“马云”和“阿里巴巴”是雇佣关系。

传统的实体关系抽取任务通常采用“流水线”方式。首先需要提取句子中相关实体，然后再识别实体之间的关系。这种方式的好处就是处理起来很方便，并且组合很灵活，但是它忽略了两个子任务之间的关联，并且会产生错误的叠加，比如，实体识别任务产生的错误会传递给关系识别的任务，导致整个模型错误率上升。

不同于上述的“流水线”方式，联合模型进行实体关系抽取时，能够从非结构或者半结构化的文本中提取出实体提及以及能够识别语句中的语义关系。通过这种方式，我们能根据语义信息，从预定义的关系表中匹配语句中出现的实体之间的关系。在提取实体和判别实体之间关系任务同时进行，大大降低了错误率的叠加，并且产生结果更加快速和高效。

联合模型的框架是将实体识别和关系识别任务用简单模型联合起来。有效的聚集了实体和关系的信息，并且在这个任务中得出一个比较好的结果。然而，目前存在的联合模型是基于特征的结构化系统，这个系统需要极其复杂的特征以及依靠很多的自然语言处理工具，在这种情况下，难免产生很多错误。为了降低人工处理的错误，目前，业界普遍采用端到端的神经网络模型，这种模型已经运用到各种序列标注任务：命名实体识别（NER）或者组合范畴语法（CCG）。而常用的神经网络模型是利用 BiLSTM 结构来获取句子表达或者句子信息来完成序列任务。

在本文中，我们将集中介绍联合模型抽取的任务，从一个生文本中抽取出包含两个（或以上）实体以及它们之间的关系，进而构成一个三元组（E1，E2，RE）[1]。因此，我们可以直接构建一个联合模型提取实体以及实体关系，基于这种想法，我们将实体关系转为一种序列标注问题，将句子切分成一个词或者字，并且给每一个词或字添加标签组（BIESO），同时，为了提高提取信息的准确率，我们也给每个词或者字进行词性标注。通过这种方法，我们仅仅通过神经网络就能构建相应的模型而不需要进行复杂的特征工程。

1 相关工作

实体关系抽取任务是构建知识库的重要一个环节，目前处理这个任务有两种方式：“流水线”方式和联合学习方式。

“流水线”方式处理这个任务分为两个步骤：命名实体识别和关系分类。

典型的命名实体识别模型是基于统计模型，比如 Passos[1]等提出从与实体相关的词典中学习一种新的词向量表达形式并且利用新的神经词向量作为单词语义表达。该方法在 CoNLL03 数据集上 F1 值达到 90.09%。Luo[2]等提出一种新的实体关系抽取模型——JERL（Joint Entity Recognition and Linking），该模型主要将实体识别和知识库中的实体进行联合来捕获实体和知识库中的依存关系，利用 CRF（Conditional Random Field）模型进行实体识别，然后利用知识库中已有的实体进行类别判断。该模型在 CoNLL03 数据集 F1 值达到 91.2%。目前，很多神经网络模型也运用到命名实体识别任务中，比如 Chiu[3]等利用 BiLSTM+CNN 联合模型进行字级和词级的特征提取。该模型首先利用 CNN 模型预处理的字级特征向量中提取出作为新的特征向量，然后将提取出的新的特征向量输入到 BiLSTM 中，进行词级的特征提取，最后输出该实体的类别概率值。该模型在 CoNLL03 数据集 F1 值达到 90.77%。Huang[4]等利用 BiLSTM+CRF 混合模型将命名实体识别任务转变为序列标注问题。该模型将分词后的词向量直接输入到 BiLSTM 中，提取出词级特征，在最后判断实体的类别时，利用 CRF 层将类别概率转变成序列概率值输出。该模型在 CoNLL2000 数据集上 F1 值为 94.40%。Lample[5]等提出利用 LSTM+CRF 模型提取词级特征同时基于过渡的方式构造标签片段。该方法的实验数据主要来源于有监督的字级语料库以及无监督的非标记的语料库。首先，对输出的句子利用依存句法的过渡方式进行处理，构造出有标记的单词，然后将预处理的单词输入到 LSTM 中，最后通过 CRF 输出序列概率值。该方法在 CoNLL2003(英文)上 F1 值达到 91.20%，在 CoNLL2003（德语）上 F1 值达到 78.76%，在 CoNLL2002（西班牙语）上 F1 值达到 85.75%。

对于关系分类任务，主要有两种方式，一种是基于特征提取的人工处理方式，Rink[6]等采用 SVM 分类器进行语义关系类别识别，然后利用好语义关系类别进行关系分类。作者采用上下文、

语义角色索引、以及可能存在名词性关系等一系列特征进行分类。该模型在 SemEval-2010 Task 8 数据集上 F1 值达到 82.19%以及 Precision 达到 77.92%。Kambhatla[7]等利用最大熵模型组合不同地词汇、句法和语义等特征进行关系分类。该方法在添加了多种特征，包括实体类型，依存关系以及句法树等,F1 值达到了 52.5%以及 Precision 达到了 63.5%。另一种是基于神经网络的处理方式，Xu[8]等通过卷积神经网络（CNN）结合最短依存路径进行语义关系分类。首先将语句输入到 CNN 网络中，提取语句中的关系特征，最后通过依存特征进行类别判断。该方法在 SemEval-2010 Task 8 数据集上 F1 值达到了 85.6%。Zheng[9]等提出基于 CNN 的模型和基于 LSTM 的模型,为了学习关系模式信息和给定实体的语法特征。首先，利用 CNN 进行关系模式的提取，然后利用 LSTM 进行实体语义的特征提取，然后将两者结合进行最后的语义关系分类。该方法在 ACE05 数据集上 F1 值达到了 53.6%以及 Precision 到了 60.0%。

联合学习方式处理实体关系任务通常只需要一个模型。大部分联合模型是基于特征的结构，比如 Ren[10]等提出一种基于 Distant Supervision 和 Weakly Supervision 对文本中的实体和关系联合抽取的框架。该框架主要分为三个部分：1.候选集的生成；2.联合训练实体和向量空间；3.实体类型和关系类型的推理预测。该方法在三个公开集上做测试，分别是在 NYT 数据集上 F1 值为 46.3%以及 Precision 为 42.3%;在 Wiki-KBP 数据集上 F1 值为 36.9%以及 Precision 为 34.8%;在 BioInfer 数据集上 F1 值为 47.4%以及 Precision 为 53.6%。Yang[11]等利用联合推理模型进行观点类实体和观点类关系的抽取。在观点类识别任务中，采用 CRF 模型将识别任务转变成序列标注任务。在观点类关系抽取任务中，利用观点-参数模型识别观点类关系的识别。该模型 MPQA 数据集上 F1 值为 57.04%。Singh[12]等利用联合推理进行三个任务：实体标注，关系抽取以及共指。该模型利用联合图模式将三者结合在一起，相互作用，通过学习和推理的方式优化联合推理模型参数。该模型在 ACE2004 数据集上针对实体抽取任务 F1 值为 55.39%，针对实体标注任务达到了 82.9%的 Precision。Miwa 和 Bansal[13]等提出一种联合实体检测参数共享的关系抽取模型，模型中有两个双向的 LSTM-RNN，一个是基于 word sequence（bidirectional sequential LSTM-RNNs），主要用于实体检测；一个是基于 Tree Structures（bidirectional tree-structures LSTM-RNNs），主要用于关系抽取。后者堆在前者上，前者的输出和隐含层作为后者的输入的一部分。Zheng[14]等利用联合模型将实体关系抽取任务转变成序列标注任务，主要是采用 End-to-End 的模型直接抽取实体和关系。

藏文信息抽取处理技术相对落后，通常也是采用“流水线”方式进行实体关系抽取即藏文命名实体识别和藏文关系分类。

针对藏文命名实体识别，金明[15]等首次提出基于规则和 HMM 模型藏文命名实体的研究方案。罗智勇[16]等通过研究藏族人名汉译的方法，提出了利用藏族人名的字级特征以及命名规则，结合词典采用字频统计和频率对比策略，以及人名前后一个词为单位共现概率作为可信度度的藏文人名识别模型，需要先给出预先定义的域值。在新华网藏族频道文本和《人民日报》（2000-2001）上实验的召回率分别为 85.54%和 81.73%。华却才让[17]等提出基于音节的藏文命名实体识别方案，采用基于音节训练模型，准确识别藏文人名，地名和机构名，识别的 F1 值达到 86.03%。刘飞飞[18]等提出基于层次特征的藏文人名识别方法，将人名的内部和上下文特征作为 CRF 特征，然后将人名并列关系特征设计为规则进一步提高识别效果。识别的 F1 值达到了 95.02%。

针对藏文关系分类，龙从军[19]等通过研究藏语名次语义关系，提出组织名次的基本单位是义类，联系名词和名词、名词与其他词之间的关系是语义关系。马宁和李亚超[20]等模板的方式从互联网中抓取纯藏文文本，然后对文本进行分词，词性标注和命名实体识别，并对关键字和实体进行过滤，抽取出候选模板。最后对抽取出的候选模板计算语义相似度，超过一定阈值就成为关系模板。

本文基于以上设计思路，同时考虑到藏文信息抽取任务的研究相对滞后，藏文的语料稀少，结构复杂，处理领域单一等缺点，考虑将联合模型运用于藏文实体关系抽取任务中，按照字级或者词级处理语料，然后利用词性标注特征进行补充，同时也将藏文关系抽取任务转变成藏文序列标注任务。

2 方法介绍

2.1 总体框架介绍

首先，我们对藏文语料分别按照词级或者字级进行序列标注处理（3.2 节介绍），然后利用自然语言工具，给每个词或者字进行词性标注（3.3 节介绍），在输入到神经网络编码层（3.3.2 节介

绍），经过编码层解析，然后通过解码层（3.3.3 节介绍），最后通过输出层输出结果（3.3.4 节介绍）。如下图 3-1 所示：

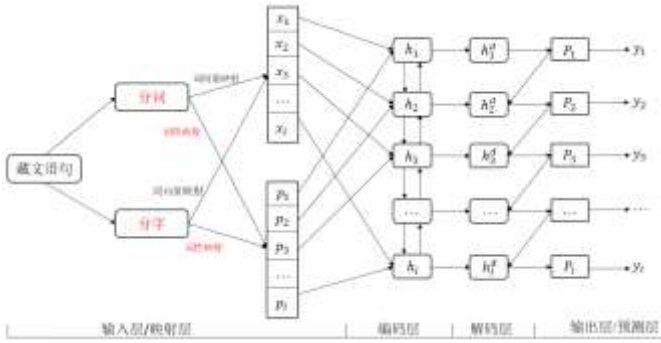

图 3-1 总体框架

其中，模型最终输出 [illegible] 代表输入藏文分词或者分字的序列标签。如图 3-2 所示（中文释义：扎西顿珠出生于迭部村庄），以分词为例，其中‘/’表示词与词之间分隔符，“BP”表示关系分类中“BirthPlace”类别。最后的输出与分词结果一一对应。

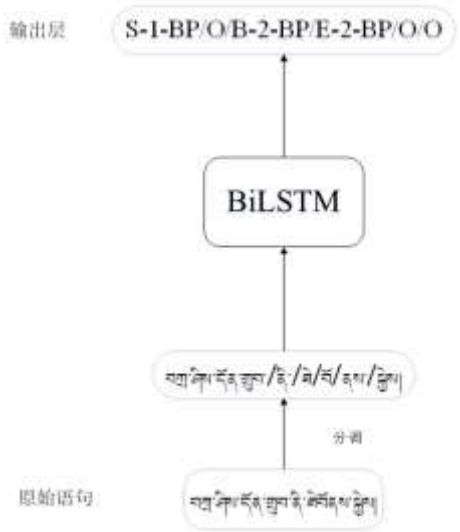

图 3-2 示例图

2.2 词级、字级处理

2.2.1 藏文词级处理

首先，本文利用 CRF++工具对藏文进行分词。然后，对分词的后的每个单词分配一个标签。标签“O”代表该单词与提及实体无关。除了标签“O”，其他单词标签分为三个部分：实体位置，关系类型以及关系角色。实体位置，本文使用“BIES”来表达，“B”代表实体起始位置，“I”代表实体中间位置，“E”代表实体结束位置，“S”代表单个实体。关系类型，从已知的关系集中查找。关系角色，根据上下文信息，确定关系角色，并同时设置“1”和“2”。

例如图 3-1 所示：（中文释义：扎西顿珠出生于迭部村庄）

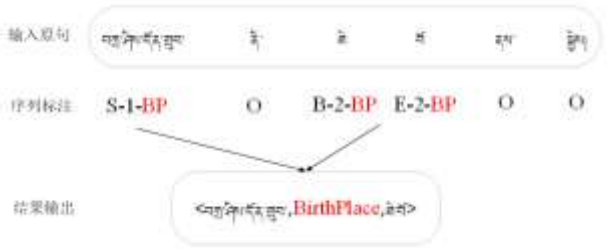

图 3-1 藏文词级处理

2.2.2 藏文字级处理

首先，本文按照藏文拼写特征，利用藏文音节点进行字级处理。然后对分字后的音节分配标签。与词级对应，标签“0”代表该音节与提及实体无关。其他的音节标签同样也分为三个部分：实体位置，关系类型以及关系角色。定义与词级一致。如图 3-2 所示：（中文释义：扎西顿珠出生于迭部村庄）

图 3-2 藏文字级处理

2.3 词性标注

由于藏文进行序列标注过后的信息较少，在与实体无关的单词或者音节上都默认标签为“0”，对结果的提取存在较大偏差。本文针对这种情况，在序列标注过后的藏文词或者字进行词性标注，对所有的词或者字分配词性标签，降低最后提取的错误率。如图 3-3 所示：（中文释义：扎西出生于迭部村庄）

图 3-3 词性标注

这里需要注意，我们字进行标注时，根据词的词性来定义，如下图 3-4 所示：（中文释义：泽旺拉姆）

图 3-4 字性的定义

不难发现，很多藏文特有的词性，例如格助词，属格助词等对帮助判断两个实体的关系有辅助的作用，同时，本文也借鉴了这种藏文特有的词性规则，比如利用属格助词来强哥表达“包含”，“属于”之类的关系，以此来强化和提高藏文实体抽取的准确率。

2.4 端到端模型

目前，基于神经网络的端到端模型在序列标注任务中起到良好的效果。本文也采用端到端的模型进行实体关系抽取任务。模型主要包括预处理阶段，BiLSTM 编码层，LSTM 解码层以及一个 Softmax 输出层。

2.4.1 预处理阶段

给定一句长度为 l 的藏文语句 W=$\{x_1,x_2,x_3,...,x_l\}$，先通过 word2vec[2] 生成词向量 T=$\{t_1,t_2,t_3,....,t_l\}$，然后经过 CRF 工具获取每个词的词性 P=$\{p_1,p_2,p_3,...,p_l\}$，并且通过 Word-POS[21]的方法将词向量和该词词性向量进行拼接，组成新的向量表达 TP=$\{(t_1,p_1),(t_2,p_2),(t_3,p_3),...,(t_l,p_l)\}$。流程如下图 3-5 所示：（中文释义：扎西出生于迭部村庄）

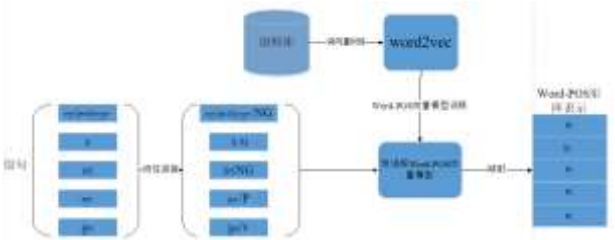

图 3-5 预处理流程

2.4.2 BiLSTM 编码层

将预处理阶段生成的向量表达 TP 输入到 BiLSTM 中。BiLSTM 能够捕获到句子中的语义信息。它主要包括前向 LSTM 层，后向 LSTM 层以及一个连接层。通过预处理得到的藏文语句向量表达，输入到 BiLSTM 中，这个结构包含一系列的循环连接单元，称谓记忆区块。每个当前的记忆区块能够根据前一层的隐向量 h_{t-1}，前一层的单元向量 c_{t-1} 以及当前的输入向量 TP_{t-1}，捕获当前的隐向量 h_t。具体公式定义如下：

输入门：

$$i_t=\sigma(W_{wi}tp_t+W_{hi}h_{t-1}+W_{ci}c_{t-1}+b_i)，①$$

主要决定当前输入的文本信息中重要的词或者字进行更新。

遗忘门：

$$f_t=\sigma(W_{wf}tp_t+W_{hf}h_{t-1}+W_{cf}c_{t-1}+b_f)，②$$

这一步决定以前的文本信息中丢弃无表达无关的词或者字

输出门：

$$z_t=\tanh(W_{wc}tp_t+W_{hc}h_{t-1}+b_c)，③$$

$$c_t=f_tc_{t-1}+i_tz_t，④$$

$$o_t=\sigma(W_{wo}tp_t+W_{ho}h_{t-1}+W_{co}c_t+b_o)，⑤$$

最终输出当前时刻的文本信息状态以及最后的特征输出向量：

$$h_t=o_t\tanh(c_t)，⑥$$

①②③④⑤⑥式中，b 是偏置项，c 是记忆单元，$W_{(.)}$ 是随机设置的参数。对于每个向量 tp，前向 LSTM 层经过上下文信息进行编码，设为 $\overrightarrow{h_t}$。使用同样的方式，后向 LSTM 层也对向量 tp 进行编码，设为 $\overleftarrow{h_t}$。最后经过连接层将 $\overrightarrow{h_t}$ 和 $\overleftarrow{h_t}$ 联合输出表达向量 tp，记为 $h_t=[\overrightarrow{h_t},\overleftarrow{h_t}]$

2.4.3 LSTM 解码层

本文采用 LSTM 结构进行标签序列的预测，对于给定向量 tp，解码层的输入为：BiLSTM 层的输出 h_t，前者预测的输出 P_{t-1}，前者的单元值 c_{t-1} 以及前者隐层的输出值 h_{t-1}^d。具体公式定义如下：

输入门：

$$i_t^d=\sigma(W_{wi}^dh_t+W_{hi}^dh_{t-1}+W_{ti}P_{t-1}+b_i^d)$$

遗忘门：

$$f_t^d=\sigma(W_{wf}^dh_t+W_{hf}^dh_{t-1}^d+W_{tf}P_{t-1}+b_f^d)$$

输出门：

$$z_t^d=\tanh(W_{wc}^dh_t+W_{hc}^dh_{t-1}^d+W_{tc}P_{t-1}+b_c^d)$$

$$c_t^d=f_t^dc_{t-1}^d+i_t^dz_t^d$$

$$o_t^d=\sigma(W_{wo}^dh_t+W_{ho}^dh_{t-1}^d+W_{co}^dc_t^d+b_o^d)$$

2.4.4 Softmax 层

针对最后的 Softmax 层，基于输出的向量 P_t，来预测实体的概率标签：

$$y_t=W_yP_t+b_y$$

$$p_t^i=\frac{\exp(y_t^i)}{\sum_{j=1}^{Nt}\exp(y_t^i)}$$

其中 W_y 是输入 Softmax 矩阵，Nt 是整个标签的数量，b_y 是偏置项。

3 实验

3.1 数据集

数据集采用中央民族大学自然语言处理实验室处理的藏文数据集，数据格式同 NYT 数据集。该藏文数据集共包括了 2400 了三元组及其原句，并且关系分类有 11 种常见的关系，在实验中，我们采用训练集有 2000 句，测试集有 400 句。

3.2 评估

主要采用准确率和召回率以及 F1 分数作为评估指标，不同于传统的机器学习方法，我们没有使用标签类型来训练模型，因此在评估过程中不需要考虑实体类型。同时我们会在训练集中随机选出 10%的数据作为验证集来优化我们模型的参数。

3.3 参数设置

我们使用 word2vec 工具来生成词向量，对于词向量维度可选[20,30,50,80]。本文基于实验效果最好的维度 50 维，即 d=50。神经网络隐层的数量依据启发式规则，将 LSTM 编码层单元数量设置成 300 层，LSTM 解码层单元数量设置成 600 层，学习率初始值设为 0.002。具体参数如下 4-1 表所示：

表 4-1 参数表

参数名	参数值
learning_rate	0.002
clip	10
dropout_rate	0.5
sequence_length	200
batch_size	64
biLSTM-num_units	300
LSTMd-num_units	600
embedding_shape	50

3.4 基线方法

我们比较了各种算法在藏文实体关系抽取上的结果，包括传统的 SVM 和 LR 方法，同时也比较了单一的 GRU 方法在任务上的结果，而我们的方法取得了最好的结果。

同时本文比较每个词性对于实体关系的抽取的影响，经过分析，选择词性 NG（名词），词性 P（格助词），词性 V（动词），词性 A（动词）作为特征控制变量输入。即本文只选取其中一种词性作为词性特征输入，并且将其他的词性设置为空，进行二次实验。

4 结果与分析

4.1 方法比较

在不同方法上的的测试结果，如下表所示 5-1 所示：

表 5-1 方法结果比较

方法	准确率	召回率	F1
SVM	0.34	0.24	0.28
LR	0.27	0.19	0.22
LSTM/GRU	0.60	0.32	0.41
LSTM+CRF	0.67	0.41	0.50
BiLSTM+藏文词分割	0.47	0.27	0.34
BiLSTM+藏文字分割	0.53	0.37	0.43
BiLSTM+藏文词分割+词性标注	0.63	0.36	0.45
BiLSTM+藏文字分割+词性标注	0.72	0.40	0.56

从上表可以看出，针对藏文的分割粒度以及词性标注的影响，我们的方法较传统的机器学习方法提升了很高的准确率，同时，在神经网络的方法中，综合比较了 LSTM 在藏文实体关系抽取任务上的不同处理，尤其是藏文语料的处理，我们采用了不同粒度对藏文进行处理，将藏文进行按照词分割和按照字分割，并在神经网络学习过程中添加词性标注进行优化，我们的方法较纯粹的神经网络模型也有很高的提升。

4.2 词性比较

这里，本文仅在藏文字级处理上进行进一步的词性标注的比较，结果如下表 5-2 所示：

表 5-2 词性结果比较

词性	准确率
+所有词性	0.72
+NG	0.69
+P	0.59
+V	0.54
+A	0.56

不难发现，词性 NG 的影响比较大，经过分析，我们发现，藏文中词性 NG 占所有词性中比例最大，约为 85%。同时缺少词性 NG 的情况下，对最后提取的准确率下降了至少 10%，可见词性 NG 对于藏文实体抽取的重要性很高。而词性 V 在所有

词性比重中占比最小，约为 2%。同时，我们也发现，词性 P 以及词性 A 对于结果的影响偏差很接近，藏文中的格助词以及形容词在一定程度上能帮助提高藏文实体抽取的准确度。

由于藏文的语料稀少，处理过程需要有专业人士进行校正，上述的切分过程都是先使用机器程序化处理，然后经过人工校正，处理周期较长，并且结果也需要有专业的人士来进行修正，帮助优化神经网络参数。

经过专业人士修正，我们发现实验中也存在以下不足：（1）在处理藏文词或者藏文字过程中，藏文的长度过长，往往几百行后才能找到相应的实体和关系；（2）藏文语句中表达的意思冲突，藏文中一个实体往往会表达多个意思，也就是说，藏文一句话中，除了标注实体以外，其他词或者字中也表达相同的意思，给神经网络模型造成误判的现象；（3）本文的方法中，在同一个句子中的两个实体，往往也会出现在其他句子中，但是关系表达不一致，导致关系紊乱，也造成了错误率提高。

5 总结

本文主要介绍一种针对藏文语料稀少的情况，提出一种将实体关系抽取任务转变成一个词性标注任务。同时，对藏文语料的处理也是本文的一大亮点，我们的实验相对于传统的机器学习以及普通的神经网路模型，取得最好的准确率。但是我们的方法在藏文的处理上也存在一些缺点，同时针对神经网络的优化也没有做对比试验。同时我们在针对藏文特有的语法规则以及性质上面，本文没有进行深入的研究。

在未来的工作中，我们会逐步优化藏文的处理，尽量减少人工的参与，同时不断优化模型，添加藏文的特有规则，继续添加藏文特有的词性规则，使模型更适应于藏文的实体关系抽取，以及为后续的藏文自然语言处理的深入研究提供基础。

参考文献

[1]Alexandre Passos, Vineet Kumar, and Andrew McCallum. 2014. Lexicon infused phrase embeddings for named entity resolution. In International Conference on Computational Linguistics. pages 78–86.

[2] Gang Luo, Xiaojiang Huang, Chin-Yew Lin, and Zaiqing Nie. 2015. Joint entity recognition and disambiguation. In Conference on Empirical Methods in Natural Language Processing. pages 879–888.

[3]Jason PC Chiu and Eric Nichols. 2015. Named entity recognition with bidirectional lstm-cnns. In Processings of Transactions of the Association for Computational Linguistics.

[4]Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991 .

[5]Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016.Neural architectures for named entity recognition.In Proceedings of the NAACL international conference.

[6]Bryan et al. Rink. 2010. Utd: Classifying semantic relations by combining lexical and semantic resources.In Proceedings of the 5th International Workshop on Semantic Evaluation. pages 256–259.

[7]Nanda Kambhatla. 2004. Combining lexical, syntactic,and semantic features with maximum entropy models for extracting relations. In Proceedings of the 43th ACL international conference. page 22.

[8]Kun et al. Xu. 2015a. Semantic relation classificationvia convolutional neural networks with simple negative sampling. In Proceedings of the EMNLP.

[9]Suncong Zheng, Jiaming Xu, Peng Zhou, Hongyun Bao, Zhenyu Qi, and Bo Xu. 2016. A neural network framework for relation extraction: Learning entity semantic and relation pattern. KnowledgeBased Systems 114:12–23.

[10]Xiang Ren, Zeqiu Wu, Wenqi He, Meng Qu, Clare R Voss, Heng Ji, Tarek F Abdelzaher, and Jiawei Han.2017. Cotype: Joint extraction of typed entities and relations with knowledge bases. In Proceedings of the 26th WWW international conference.

[11]Bishan Yang and Claire Cardie. 2013. Joint inference for fine-grained opinion extraction. In Proceedings of the 51rd Annual Meeting of the Association for Computational Linguistics. pages 1640–1649.

[12]Sameer Singh, Sebastian Riedel, Brian Martin, Jiaping Zheng, and Andrew McCallum. 2013. Joint inference of entities, relations, and coreference. In Proceedings of the 2013 workshop on Automated knowledge base construction. ACM, pages 1–6.

[13]Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. In Proceedings of the 54rd Annual Meeting of the Association for Computational Linguistics.

[14]Suncong Zheng,Feng Wang,Hongyun Bao,Yuexing Hao,Peng Zhou,Bo Xu. Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme[J]. Computational Intelligence and Neuroscience,2017,2017.

[15]金明,杨欢欢,单广荣.藏语命名实体识别研究[J].西北民族大学学报:自然科学版，2010(3):49-52.

[16]罗智勇,宋柔,朱小杰.藏族人名汉译名识别研究[J].情报学报，2009(3):475-480.

[17]华却才让,姜文斌,赵海兴,刘群.基于感知机模型藏文命名实体识别[J].计算机工程与应用,2014,50(15):172-176.

[18]刘飞飞,王志娟.基于层次特征的藏文人名识别研究[J/OL].计算机应用研究,2018(09):1-7[2018-03-22].

[19]龙从军,周学文.藏语名词语义关系研究[J].全国少数

民族青年自然语言处理学术研讨会,2008
[20]马宁,李亚超,于槐,加羊吉.面向互联网的藏文实体关系模板获取技术研究[J].中央民族大学学报(自然科学版),2015,24(01):35-39.
[21]何鸿业,郑瑾,张祖平.基于词性结合的卷积神经网络文本情感分析[J/OL].计算机工程:1-7[2018-03-13].

夏天赐（1993—），硕士研究生，主要研究领域为自然语言处理，信息检索，问答系统
E-mail：muctianciking@163.com

孙媛（1979—），通信作者，副教授，主要研究领域为自然语言处理，信息抽取。
E-mail：tracy.yuan.sun@gmail.com