中文连动句语义关系识别研究

孙超[1] ， **曲维光**[1,2,*] ， **魏庭新**[2,3] ， **顾彦慧**[1] ， **李斌**[2] ， **周俊生**[1]
(1.南京师范大学 计算机与电子信息学院/人工智能学院，江苏省 南京市210023
2.南京师范大学 文学院，江苏省 南京市210097
3.南京师范大学 国际文化教育学院，江苏省 南京市210097
* 通信作者，Email：wgqu_nj@163.com)

摘要

连动句是形如“NP+VP1+VP2”的句子，句中含有两个或两个以上的动词（或动词结构）且动词的施事为同一对象。相同结构的连动句可以表示多种不同的语义关系。本文基于前人对连动句中VP1和VP2之间的语义关系分类，标注了连动句语义关系数据集，基于神经网络完成了对连动句语义关系的识别。该方法将连动句语义识别任务进行分解，基于BERT进行编码，利用BiLSTM-CRF先识别出连动句中连动词(VP)及其主语(NP)，再基于融合连动词信息的编码，利用BiLSTM-Attention对连动词进行关系判别，实验结果验证了所提方法的有效性。

关键词： 连动结构 ； 神经网络 ； 连动句语义关系识别

Research on Semantic Relation Recognition of Chinese Serial-verb Sentences

SUN Chao[1], **QU Weiguang**[1,2,*], **WEI Tingxin**[2,3], **GU Yanhui**[1], **LI Bin**[2], **ZHOU Junsheng**[1]
(1.School of Computer and Electronic Information/School of Artificial Intelligence, Nanjing Normal University,Nanjing, Jiangsu 210023,China;
2.School of Chinese Language and Literature,Nanjing Normal University,Nanjing,Jiangsu 210097,China;
3.International College for Chinese Studies, Nanjing Normal University,Nanjing,Jiangsu 210097,China
* Corresponding Author,Email:wgqu_nj@163.com)

Abstract

Serial-verb sentences are those in the form of ”NP+VP1+VP2” which contains two or more verbs (or verb structures) sharing the same agent. Sentences with the same serial-verb structures can express a variety of different semantic relation. Based on the previous works in classification of the semantic relation in consecutive verb sentences, we built a data set of serial verb sentences, and proposed a neural network model for recognition of semantic relation in serial verbs. First,the serial-verb sentences are encoded with Bert. Second, BiLSTM-CRF is used to identify the serial verbs and their subjects. Then, based on the encoded information of fused serial verbs, the relation recognition of serial verbs is implemented using BiLSTM-Attention. The experimental results prove the effectiveness of the proposed method.

Keywords: serial-verb structure , neural network , semantic relation recognition of serial-verb sentences

基金项目：国家自然科学基金“汉语抽象意义表示关键技术研究”(61772278)；江苏省高校哲学社会科学基金“面向机器学习的汉语复句语料建设研究”(2019JSA0220);国家社会科学基金“中文抽象语义库的构建及自动分析研究”（18BYY127）。

1 引言

连动结构是由两个或两个以上动词或动词短语构成的结构，且动词或动词短语共享同一个主语。连动句是包含连动结构的句子，是一种分布范围比较广泛、生命力比较旺盛的句式结构。连动句表示多个事件，这些事件相互依赖并呈现出语义上的方式、顺承、目的、因果等关系，相互依赖影响并产生相互关联的事件 (刘雯, 2017)。例如“我去图书馆看书”是一个连动句，在该句中包含由“我”完成的两个事件，分别是“我去图书馆”和“我看书”。对这两个事件而言，后一个事件是前一个事件发生的目的，需要先发生前一个动作，才能完成后一个动作。有效的连动句语义识别有助于自然语言处理中句子级别的语义分析任务和句法解析任务的研究。

抽象语义表示（Abstract Meaning Representation, AMR）是近年来一种新兴的句子级的语义表示方法，突破了传统的句法树结构的限制，将一个句子语义抽象为一个单根有向无环图。同时AMR会将缺省的论元进行补全，得到完整的语义表示，很好地解决了论元共享的问题 (曲维光 et al., 2017)。在连动句中存在着内部概念节点论元共享的现象 (戴茹冰 et al., 2020)，即在单句中多个谓词共享同一论元角色，AMR会将缺省主语的谓词进行补全，同时标注多个谓词间的语义关系。本文对连动句语义的研究借鉴了AMR 中对连动句语义的标注方法；同时本文对连动词的语义关系识别研究，也能够帮助中文AMR中连动句式的标注与解析工作。

2 相关工作

许多语言学家从不同的侧重点对连动结构的语义特征进行了不同的划分。赵元任 (1952)在《北京口语语法》中将连动式的语义关系分为了六类，(1)时间次序、(2)条件、(3)地点、(4)方法、(5) 目的、(6) 比较。随后他 (2002)又在《中国话的文法》中把连动结构的语义进行了进一步的划分，分为十类，除了以上(1)、(3)、(4)、(5)和(6)外还添加了“时间”、“恩惠和好处”、“手段”、“一般的联系”和“动作加动作”五个类别。丁声树 (1982)等人在《现代汉语语法讲话》中将连动结构的语义特征分为了五类，包括(a)拿东西次序分先后、(b)条件、(c)对象、(d)方式、(e) 时间和处所。朱德熙 (1982)在《语法讲义》中也介绍了几种连动结构的语义特征，如表伴随发生的动作、表目的、表假设、表因果等等。

上世纪的语言学家在连动结构的语义分类方面做了许多的研究工作，既有许多的相同之处，也存在一些有争议的地方。近年来也有一些研究人员从其他的角度去研究连动句的语义问题。徐情 (2012) 基于大规模真实文本语料库对现代汉语连动结构的语义结构类型和语义特征进行了系统的研究，采用了定性和定量相结合、形式和意义相结合的方法，在标注语料库的基础上，归纳得到了现代汉语连动结构五种特殊的语义结构，体现句法和语义的结合和统一。陈波等 (2013)针对现有的传统语义分析方法存在的不能很好反映汉语中各个词语或成分之间的语义关联，提出了基于特征结构的语义标注方法，探讨了连动句的语义标注模型，为面向汉语的自然语言处理提供了一种不同的语义分析方法。蒋梦娇 (2019)在语义语法学理论的指导下,基于标注的语料库,重点考察了连动句的词汇语义与句法语义的互动制约关系。在词汇语义和句法语义两方面对四类连动句V1进行词汇义征分析和句法范畴义征分析，最终建构了关于连动句的词汇与句法互动制约机制。

基于前人的研究 (邢欣, 1987; 杨月蓉, 1992)和中文AMR语料中连动句的标注规范，本文将连动词间的语义关系分为时序（temporal）、方式（manner）、并列（and）、目的（purpose）、因果（cause）、“来”类和“去”类，共7种语义关系，并用“other”来表示两个动词间不存在连动关系。同时在人工标注的数据集上采用神经网络的方法对连动句语义关系识别展开研究。

3 连动句语义关系语料标注

目前还没有一个公开的、完整的连动句语料库资源，故在前人研究理论的指导下，借鉴中文AMR的标注规范，本文首先构建了一个连动句的数据集，标注了连动句中的连动词及其主语，并且标注了连动词间的语义关系。

3.1 连动句中连动词及其主语的标注

本文使用基于AMR标注的小学一至六年级语文课本语料构建关于连动句的数据集。首先利用文献 (孙超 et al., 2020)中的方法挑选出小学语料中的连动句，再进一步标注出其中的NP、VP1以及VP2。利用序列化标注的思想进行标注，将主语NP标注为“n”，将连动

词VP标注为“v”，以“B”作为开始标签，“I”作为其他部分，单句中除去连动词和主语的其他部分全部标注为“O”。在对NP进行标注时，分为单主语和多主语两种，皆只标注主语的主体部分。若主语部分为“量词+名词”、“形容词+名词”、“名词+名词”皆只标记名词部分。若单句中不存在主语，则不做特殊标注。

例1：一个牧童挤进来喊着:

此句中将“牧童”标注为NP。

例2：英国发明家阿切尔到伦敦一家小酒馆喝酒。

此句中将“阿切尔”标注为NP。

例3：美国环球集团有限公司与满洲里国际贸易公司联合改组了满洲里服装厂。

此句中的主语分为两个部分，分别标注两个主语的中心词为NP，即“公司”、“公司”。

对于连动词部分只标注动词短语中心词；当某个连动词包含一个连动结构时，只标注连动结构的第一个连动词；同时对于类似“看一看”、“想一想”等一类的词语规定只标注第一个字作为连动词。综上，一个完整的连动句中连动词及其主语的标注示例如表 1所示。

例4：又挥锹填了几锹土。

此句中将“挥”和“填”标注为连动词。

例5：你跳出井口来看一看啊。

在此句中含有两个连动结构，第一个连动结构的连动词是“跳出”和“来看一看”，第二个的是“来”和“看一看”。在标注时，第一个连动结构将“跳出”和“来”标注为连动词，第二个将“来”和“看”标注为连动词。

你	跳	出	井	口	来	看	一	看
B-n	B-v	I-v	O	O	B-v	B-v	O	O

Table 1: 连动句中连动词与主语标注示例

3.2 连动句语义关系标注

连动句的语义关系一直是语言学家研究的重点，他们把连动句的语义关系做了细致的描述 (陈波, 2011)，对主语与构成谓语的若干动词短语之间的语义关系，学者们将其分为2-4类语义关系（范晓1980，陈昌来2000，吴启主1990，杨月蓉1992）；对构成谓语的若干动词短语之间的语义关系方面，学者们将其分为5-12类语义关系（范晓1980，陈昌来2000，黄伯荣1991，李临定1986）。本文研究的语义关系主要针对“构成谓语的若干动词短语之间的语义关系”，而对“主语与构成谓语的若干动词短语之间的语义关系”将不再做进一步的分类，只完成识别“主语”任务。基于前人的研究和中文AMR语料中连动句的标注规范，本文采用比较主流的分类方法，将连动词间的语义关系分为时序（temporal，VP1在时间上先于VP2发生）、方式（manner，VP1表VP2的方式）、并列（and，VP1和VP2表并列关系）、目的（purpose，VP2表VP1的目的）、因果（cause，VP1表VP2原因）、以及两类特殊的语义关系–“来”类(lai)和“去”类(qu)，共7种语义关系。以下是每种语义关系的例子。

例6：小白兔连忙挎起篮子往家跑。

此句中VP1“挎起”和VP2“跑”具有时序关系。

例7：小白兔弯着腰在山坡上割草。

此句中VP1“弯”表示VP2“割草”的方式。

例8：燕子边飞边说:“要下雨了。”

此句中VP1“飞”和VP2“说”具有并列关系。

例9：他们没有砍树造房子。

此句中VP2“造”表示VP1“砍”的目的。

例10：许多人和动物都焦渴而死。

此句中VP1“焦渴”表示VP2“死”的原因。

例11：蜻蜓说：让我来送小蚂蚁吧！

此句中VP1“来”作为趋向动词与VP2“送”直接连动，作为特殊连动结构。

例12：妈妈知道他又去耍水了。

此句中VP1“去”表示施事者位移的运动趋向，与VP2“耍水”（施事者位移后进行的行为动作）直接连用，作为特殊的连动结构。

在中文连动句中可以在第一个动词VP1位置出现的单音节趋向动词仅限于“来”和“去”。在“来”类连动句中分为三种形式，分别是“来+NP+VP”、“来+VP”和“VP+来”。在“去”类的连动句中也分为以上三种形式。在第一类“来/去+NP+VP”中，“来”和“去”都充当了真正的动词，表示较强动作性的移位行为。例如，在句子“母亲的学生来我家学戏”中“来我家”的目的是为了“学戏”，所以在CAMR的标注会将“来”和“学戏”用“purpose”联系在一起，具体标注见图 1(a)。对于第二种“来+VP”是“来”作为趋向动词与VP直接连用，是一种特殊的连动，例如“皮埃尔·居里坚持让她来讲。”一句中，“她来讲”中的“来”表示的动作性就没有上例中的强烈，在这种情况下，“来”出现动词虚化现象，此时CAMR会让“来+VP”共享同一个概念结点，具体标注见图 1(b)。而在第三种情况“VP+来”中，此时“来”的意义已经被完全抽象化或完全虚化了，例如“乌鸦看见旁边有许多小石子，想出办法来了”此处的“来”已经完全没有动词意义，在CAMR中“来”将不在图中进行标注，具体标注见图 1(c)，此类情况也不作为本文研究的连动句。

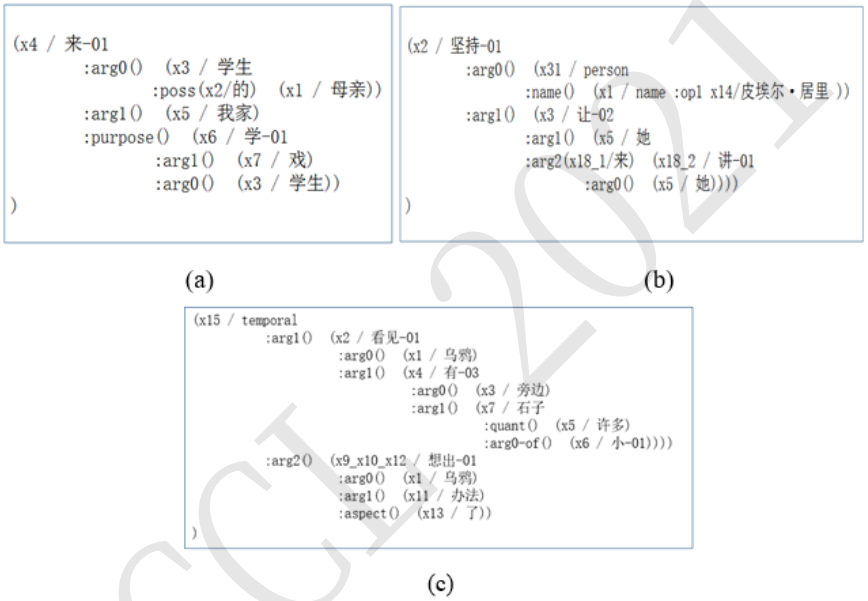

Figure 1: CAMR文本表示

“去”类的情况与上述“来”类情况相同，所以本文研究的“来”类和“去”类语义关系指的是第二类“来/去+VP”，对于第一类情况“来/去+NP+VP”将标注为语句表现出的更强烈的语义关系–“purpose”。

当某句中含有多组连动关系，为使连动词识别与连动词语义关系识别两个任务可以串联起来，同时避免当连动词识别错误时语义关系必然学习出错的情况，本文将连动句中所有的动词短语进行组合，罗列所有可能产生语义联系的组合，正确组合的样本根据语义关系标注语义标签，对于组合不正确的样本统一标注为“other”。

综上所述，对于每一个连动句，对其完整的语义关系标注，以“你跳出井口来看一看”一句为例，句中含有三个动词结构分别是“跳出”、“来”和“看”，两两进行组合，共有三种情况，分别进行语义关系的判断，标注如下，用<v_i></v_i>标注出句子中的连动词。

purpose(v_1, v_2)　　你<v_1>跳出</v_1>井口<v_2>来</v_2>看一看吧。

lai(v_1, v_2)　　你跳出井口<v_1>来</v_1><v_2>看</v_2>一看吧。

other　　你<v_1>跳出</v_1>井口来<v_2>看</v_2>一看吧。

4 连动词及其主语的识别

对于连动句s=(w_1,w_2,⋯,w_n)，首先需要识别出其中的连动词和其主语，再进行连动词间的语义关系识别。为此将连动词及其主语的识别看作一个命名实体识别的任务 (Greenberg N et al., 2018; Chen J et al., 2020; Kruengkrai C et al., 2020)，此时的"实体"包含两种，分别是连动句中的连动词v以及其主语n。采用的模型如图 2所示。

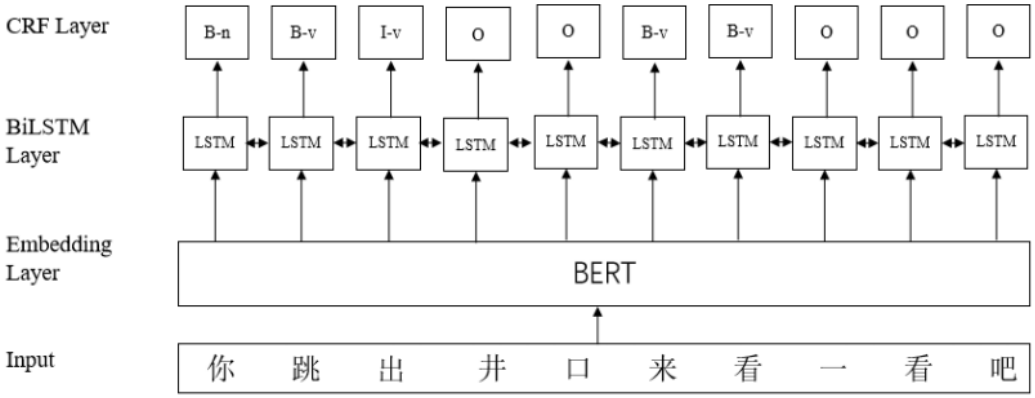

Figure 2: 连动词及其主语的识别模型图

(一) Word Representation Layer

使用预训练语言模型BERT-WWM-EXT (Cui Y et al., 2019),该模型是基于BERT模型 (Devlin J et al., 2019)并改变了BERT的MASK方式训练而来，在训练过程中使用了大量中文维基百科数据和通用数据，进一步提升了预训练语言模型的性能。

BERT中使用了Masked Language Model(MLM),即在模型进行训练时会随机从输入的序列中覆盖(MASK)一些单词，随后通过上下文来预测被MASK掉的单词。在对英文语料进行训练时MASK一个单词或单词的部分对句子本身的影响不太大，而对于中文语料的训练，MASK掉某些字，会导致有些中文词语被分割，从而可能会导致中文句子本身的语义发生了变化。而全词覆盖(Whole Word Masking,WWM)技术主要是更改了原本BERT在预训练阶段的样本生成策略，采用WWM时会覆盖整个词语，而非词语中的某个字。通过全词覆盖（WWM）的方法更有利于学习词语的搭配，有助于帮助连动词及其主语边界的识别。

(2) Contextual Representation Layer

BiLSTM层由前向LSTM和后向LSTM组成，前者用于学习前向的序列信息，后者用于学习后向的序列信息，二者考虑了句子前后的信息，充分结合上下文的特征。将输入的单句进行字级别的编码后将结果送入BiLSTM层，BiLSTM将利用字在句子中的前后顺序，同时还可以捕获双向的较长距离的语义依赖关系，从而更好地判断当动词间相隔较远情况下是否可以构成连动关系。

(3) CRF Layer

将连动句中连动词和主语的识别看作序列化标注问题，利用命名实体识别的思路，使用BIO（Beginning，Inside，Outside）编码方式。为句中每个字都分配一个标签，可以识别连动句中连动词（主语）的开始位置和结束位置，此处的实体类型分为两种(n,v)。在BiLSTM层之上，采用CRF层 (Kruengkrai C et al., 2016)来计算每个token的最可能标签，计算每个字的标签score:

$$s(h_i) = Vf(Uh_i + b) \tag{1}$$

其中f(·)表示激活函数，$V \in R^{p\times b}$，$U \in R^{p\times 2d}$,$b \in R^l$，d表示隐藏层单元数，p表示标签数，l表示BiLSTM层的宽度。假设输入连动句s=(w_1,w_2,⋯,w_n)，对应每个字标签的score为(s_1,s_2,⋯,s_n)，预测的标签序列为(y_1,y_2,⋯,y_n)，实际标签序列为($\tilde{y}_1$，$\tilde{y}_2$,⋯,$\tilde{y}_n$)。则线性链CRF的score为:

$$S(y_1, y_2, ..., y_n) = \sum_{i=1}^{n} s_{i,y_i} + \sum_{i=0}^{n} T_{y_i,y_{i+1}} \tag{2}$$

其中s_{i,y_i}是单词w_i被预测为y_i的得分，T是转移矩阵，$T_{y_i,y_{i+1}}$表示标签从y_i变为y_{i+1}的概率，用来表示相邻标签之间的依赖关系，此处添加了y_0和y_{n+1}表示句子的开始标志和结束标志，T_{y_0,y_1}可以限制句子开始的第一个文字w_1的标签不能标注为"I-"。产生标签序列y的概率表示为公式 3，通过最小化交叉熵损失L_ner，来达到优化网络参数和CRF 的目的。

$$Pr(y_1, y_2, ..., y_n|w) = \frac{e^{S(y_1,y_2,...,y_n)}}{\sum_{\tilde{y}_1,\tilde{y}_2,...,\tilde{y}_n} e^{S(\tilde{y}_1,\tilde{y}_2,...,\tilde{y}_n)}} \tag{3}$$

5 连动句中连动词语义关系识别

在连动句语义关系识别任务中，模型既需要依赖句子的信息，又对需要判别语义关系的两个连动词具有很强的依赖性 (Wei Z et al., 2020; Wu S and He Y, 2019)。为此本文首先提出了一个基于神经网络的方法，如图 3，它既利用了预训练的BERT 语言模型，又结合了来自连动词的信息，模型可以定位连动词并通过预训练结构传递信息，并添加连动词与连动句的交互信息，来判断连动词间的语义关系。

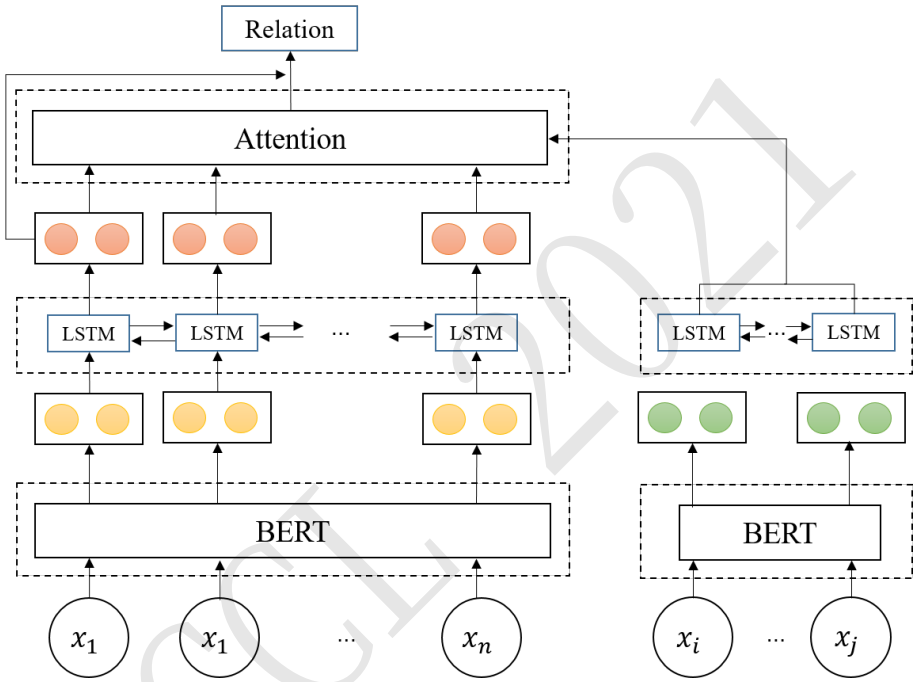

Figure 3: 连动句中连动词语义关系识别模型图

（一）Word Representation Layer

本任务中采用BERT作为模型的词向量表示层，由于需要判断的是连动句自身的语义情况，所以只需要输入单个句子，因此在句子输入时不需要在句尾添加[SEP]，但为每个连动词个体添加了相应的标签，对第i 个连动词以<v_i>开始，以</v_i>结束。

由于希望模型学习到两个连动词间以及和连动句之间的语义关系，对BERT模型的输出部分进行特殊处理方式，然后分别进行训练。将BERT的输出分为三部分，如图 4所示，第一部分是[CLS] 的隐含向量，它保存了整个连动句的信息；第二部分是第一个连动词的隐含向量；第三部分是第二个连动词的隐含向量，后两个部分分别保存了两个连动词的信息 (Liu W et al., 2020)。

对于BERT模型输出的三个部分，分别进行处理。H_0中含有整个句子的语义信息，i、j、k、m分别表示了第一个连动词的首字符的位置、第一个连动词的结束位置和第二个连动词的首尾字符的位置。将BERT 输出的三部分内容分别送入三个前馈神经网络进行学习，计算方式分别式 4、 5和 6。对于单音节的连动词直接使用BERT对其编码的输出内容H_i作为前馈神经网络的输入，对于多音节的连动词使用所有位置的字编码的平均值作为前馈神经网

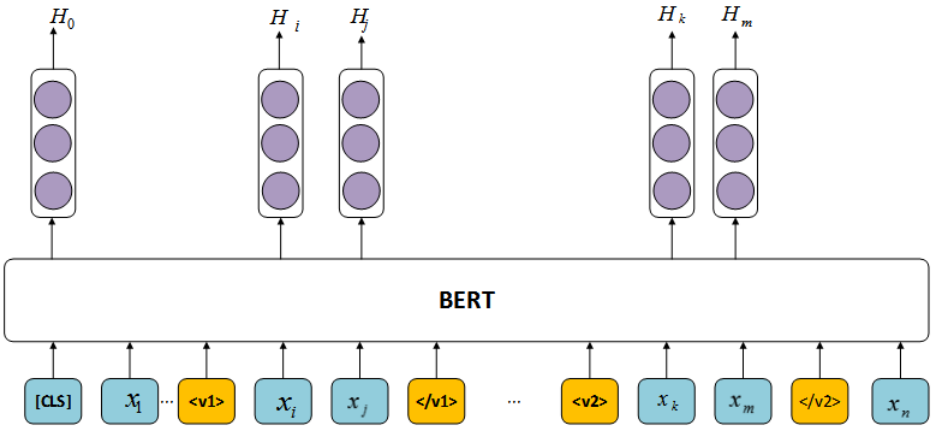

Figure 4: BERT模型输出

络的输入。

$$H_0^{'} = W_0(tanh(H_0)) + b_0 \tag{4}$$

$$H_1^{'} = W_1[tanh(\frac{1}{(j-i+1)}\sum_{t=i}^{j} H_t)] + b_1 \tag{5}$$

$$H_2^{'} = W_2[tanh(\frac{1}{(m-k+1)}\sum_{t=k}^{m} H_t)] + b_2 \tag{6}$$

（二）Contextual Representation Layer

将两个连动词的表示形式$H_1^{'}$、$H_2^{'}$进行拼接，随后利用双向LSTM分别获取连动词和连动句的上下文信息。对于连动句$H^l = h_0^l, h_1^l, \cdots, h_n^l$，编码使用BiLSTM 的隐藏层输出，见公式 7；对于连动词h^v，由于其长度比较短，使用最后一个隐藏层的输出，见公式 8。

$$h_i^t = [h_l^t; h_l^t] = BiLSTM(\overrightarrow{h_{(l+1)}^t}, \overleftarrow{h_{(l-1)}^t}, w_l^t) \tag{7}$$

$$h_v = [h_n^v; h_n^v] = BiLSTM(\overrightarrow{h_{(n)}^a}, \overleftarrow{h_{(n)}^a}, w_n^a) \tag{8}$$

（三）Attention Layer

使用Attention Layer计算连动词与连动句的交互信息，获取其中的语义联系。通过使用余弦函数计算连动词与连动句中每个词的相似性获取权重矩阵。

$$\alpha_i = cosine(h_i^t, h^v) \tag{9}$$

6 实验及结果分析

本文将连动句的语义识别任务分解成两个任务，利用pipeline模型解决该问题 (Zhong Z and Chen D, 2020),并分别进行实验验证了所提方法的有效性。在实验中使用精确率(precision,P)、召回率(recall,R)和F1-score (F1)和正确率（accuray,Acc）作为评价依据。

6.1 连动词及其主语识别

6.1.1 实验数据

实验采用本文2.1中所介绍的人工标注的连动句数据集，包含2071句连动句，其中包含4444个连动词和1417个主语。以句子为基准，按照6:2:2的比例划分训练集、开发集和测试集。

6.1.2 对比实验及结果分析

设置对比实验验证所提方法对连动句结构识别的有效性和模块各部分的必要性。表 2和 3展示了各模型对连动词及其主语的识别效果，表 4展示了句子整体的标注效果和连动结构的识别效果，其中BiLSTM+CRF 作为基线模型，使用基于字级别的向量表示；BERT+BiLSTM+CRF模型中用BERT代替了基线模型中的Character embeddings;设置与BERT相对应的BERT_WWM_EXT实验，验证WWM 机制对本任务的有效性。

模型	连动词		
	P	**R**	**F1**
BiLSTM+CRF	77.66%	76.06%	76.86%
BERT+CRF	84.72%	86.84%	85.77%
BERT+BiLSTM+CRF	87.88%	90.02%	88.94%
BERT_WWM_EXT+CRF	87.10%	89.80%	88.43%
BERT_WWM_EXT+BiLSTM+CRF	**90.56%**	89.36%	**89.96%**

Table 2: 连动词识别实验结果

模型	主语		
	P	**R**	**F1**
BiLSTM+CRF	74.41%	75.55%	74.97%
BERT+CRF	83.90%	84.69%	84.29%
BERT+BiLSTM+CRF	87.39%	86.61%	87.00%
BERT_WWM_EXT+CRF	87.05%	87.05%	87.05%
BERT_WWM_EXT+BiLSTM+CRF	**87.44%**	**87.05%**	**87.25%**

Table 3: 主语识别实验结果

模型	整句			
	P	**R**	**F1**	**Acc**
BiLSTM+CRF	76.84%	75.90%	76.37%	49.19%
BERT+CRF	84.53%	86.33%	85.42%	60.24%
BERT+BiLSTM+CRF	87.72%	88.89%	88.30%	70.79%
BERT_WWM_EXT+CRF	87.08%	88.90%	87.98%	71.95%
BERT_WWM_EXT+BiLSTM+CRF	**89.52**%	88.59%	**89.05**%	**73.17**%

Table 4: 整体识别实验结果

表 2至表 4分别展示了模型对连动词、主语和连动句整体的识别标注结果。表4中的Acc评测的是以每个连动句为单位的标注效果，不仅句中的连动部分要标注正确，其他部分也不能标注出错才判断为一个正例。通过对比可以发现，预训练语言模型BERT和BERT_WWM_EXT的效果远远好于基于word2vec训练得到的字符级别向量的模型。在对连动词、主语和两者总体的评价指标中，本章所提出的模型在P值和F1 值都是最优的，R值相对低一些，但相差不大，验证了本章中提出模型的有效性。对于每一个模型，会发现Acc的数值与F1值相比还是低很多，主要的原因是评测的单位不同所致的，模型对“连动结构”的学习效果还不够理想，会对连动结构外的成分进行误识别，导致对连动句整体的标注出错。

观察模型的预测结果可以发现在标注错误的句子中主要分为以下几种情况:

第一种情况是连动词的边界识别错误，例如在“就皱紧眉头硬咽下去。”一句中标注的两个连动词是“皱”和“咽”，而模型中识别出的两个连动词是“皱紧”和“硬咽”，此时模型对连动部分的识别是正确的，但在分词上发生了问题，将修饰动词的部分加入了动词本身。又如在“蔺相如见廉颇来负荆请罪”一句中，标注为“廉颇”、“来”、“负荆请罪”，而模型会将“负荆请罪”识别为两个词“负荆”和“请罪”，评测也会判别为该句识别错误。

第二种情况是模型可以正确识别原本标注的连动结构，但除此之外还会识别出其他无效的部分。例如在“老班长猛抬起头看见我目不转睛地看着他手里的搪瓷碗”一句中，模型可以正确识别出标注的gold部分“老班长”、“抬起”和“看见”，但除此之外，模型还将句子后半部分的“看着”一词也识别为该句中的一个连动词。这可能是因为句子的后半部分是“看见”的宾语，句式杂糅造成识别容易出错。

第三种情况是模型对连动词或主语的识别出错，与前两种情况不同，此时的模型不能识别出句中的连动词或识别出错，例如在“把狗尾草当做谷穗留着。”模型只能识别出一个动词“留”，此时对于该句，模型会认为该句中不含连动结构。再如“可用千树万树梨花开来形容”一句中模型识别出的动词结构是“开来”和“形容”，连动词识别出错。这些问题后续还需要再进行改进。

6.2 连动句语义关系识别

6.2.1 实验数据

实验采用本文2.2中所介绍的人工标注的连动句数据集，包含2071句连动句，各类语义关系的分布情况如表 5所示。按照6:2:2的比例划分训练集、开发集和测试集。

Relation	and	cause	lai	manner	purpose	qu	temporal	other
Num	432	26	155	576	537	230	411	176

Table 5: 数据集语义关系分布

6.2.2 对比实验及结果分析

表 6展示了模型对各类语义关系的识别效果，以及模型整体的Acc,其中support表示各类语义在测试集中的数量。

Relation	**P**	**R**	**F1**	**support**
and	0.82	0.80	0.81	123
cause	0.67	0.20	0.31	10
lai	0.97	0.97	0.97	35
manner	0.77	0.83	0.80	174
purpose	0.78	0.74	0.76	160
qu	0.94	0.96	0.95	69
temporal	0.79	0.79	0.79	126
other	0.82	0.87	0.84	53
Acc	**0.811**			

Table 6: 语义关系识别结果

通过表 6的实验结果可以看出，除了数量较少的“cause”类识别效果比较差外，其他类别的语义关系识别效果还不错。尤其对于“lai”类和“qu”类，虽然这两类的数量并不是最多的，但由于这两类语义的规则性比较强，都含有很明显的标志词“来”和“去”，模型可以比较容易获得其特征，所以识别效果比较好。但仍然无法保证百分之百的正确性。主要原因是对于“来/去+NP+VP”表目的的句子，模型很难与“来/去+VP”区分开来。例如在“< v_1 >去< /v_1 >洛阳< v_2 >拜< /v_2 >大思想家老子为师。”一句中“去洛阳”的目的是“拜师”，而模型会将其仍识别为“去”的动作性不强的“qu”类，此类情况在整个数据集中所占比例相对较少，故模型学习到的特征不够充分。对于“and”和“temporal”两种语义关系识别比较容易混淆，“temporal”和“and”的区别主要在与“temporal”侧重于时间上的先后，两个动作的发生要有先后顺序，当先后顺序不明显时，模型容易识别错误。同时“manner”与“and”两种语义关系也容易识别混淆，这两类中都可以表达VP1和VP2不分先后，例如“他边笑边说”和“他笑着说”，两句中都需要判断“笑”与“说”在句中的语义关系，此时模型对二者的区分度不够高。同时也存在一些因模型无法识别出语义关系而错误分为“other”类的情况。

同时也设置了消融实验，实验结果如表 7所示，证明各层设计的有效性。其中BERT_1模型只使用了H_0'，即只使用连动句的编码方式；而BERT_2模型使用了H_0'、H_1'和H_2' 三者进行拼接的编码方式，通过全连接层，最终送入分类层进行分类。为综合各种语义关系的识别效果，使用Macro_P、Macro_R和Macro_F1作为评价指标，具体计算方式如公式所示。其中P_i和R_i代表第i类语义关系识别的P值和R值，n表示语义关系的种类数。

$$Macro\_P = \frac{1}{n}\sum_{i=1}^{n} P_i \tag{10}$$

$$Macro\_R = \frac{1}{n}\sum_{i=1}^{n} R_i \tag{11}$$

$$Macro\_F1 = \frac{2 * Macro - P * Macro - R}{Macro - P + Macro - R} \tag{12}$$

Model	Macro_P	Macro_R	Macro_F1	Acc
BERT_1	0.70	0.71	0.71	0.73
BERT_2	0.76	0.74	0.75	0.75
BERT_2+BiLSTM	0.79	0.75	0.75	0.79
BERT_2+Attention	0.77	0.73	0.73	0.77
BERT_2+BiLSTM+Attention	**0.82**	**0.77**	**0.78**	**0.81**

Table 7: 消融实验结果

由表7可知，融合连动词信息的编码对判断两个连动词间的语义关系起到很大的作用，这主要是因为很多时候连动句的语义关系和连动词间的语义关系是不同的，且有时同一个连动句中可能含有多个连动词(2个以上)，表达多种语义关系，故对连动词的有效编码可以提升实验的效果。同时连动词所处的上下文语境对语义关系的识别也非常重要，同一组连动词处在不同的上下文语境中所表达的语义关系可能是不同的。例如在“燕子妈妈笑笑说”和“妈妈笑着说”两句中都含有（笑，说）这一对连动词，而在前者中更多地表达的是“temporal”这种语义关系，先发生了“笑”后发生了“说”；而后者表达的是“manner”这种语义关系，“笑”是“说”时的状态。所以模型使用BiLSTM更好地融合了上下文信息，同时利用Attention获得连动词对和连动句的交互信息，更好地识别连动句中连动词的语义关系。

7 总结展望

本文利用神经网络对连动句语义关系进行识别研究。首先对连动句中连动词及其主语进行识别，然后对连动词再进行组合和语义关系的识别，并在人工标注的语料上进行实验探究。构建了基于字符级别的神经网络模型(BiLSTM-CRF) 用以识别连动词及其主语；构建了融合连动词信息的神经网络模型（BiLSTM-Attention）用以识别连动词间的语义关系。两个任务基于不同输入特征独立训练模型更能捕获到每个子任务需要的特征，从而达到较好的语义识别效果。目前，构建的连动句数据集的规模相对较小，各类语义间存在着数据分布不均衡的问题，同时对连动词边界识别还存在一些错误，也会影响语义关系的识别。接下来将针对存在的这些问题对模型进行改进,以达到更好的识别效果。

参考文献

陈波,姬东鸿,吕晨. 2013. 基于特征结构的汉语连动句语义标注研究[J]. 中文信息学报, 27(05):60-66+74.

戴茹冰,侍冰清,李斌,等. 2020. 基于AMR语料库的汉语省略与论元共享现象考察 [J]. 外语研究, 37(02):16-23.

丁声树. 1982. 现代汉语语法讲话[M]. 北京: 商务印书社.

蒋梦娇. 2019. 基于词汇语义与句法语义互动制约的连动句研究[D]. 南京师范大学.

刘雯. 2017. 基于汉语连动句的常识获取方法研究[D]. 江苏科技大学.

陈波. 2011. 特征结构及其汉语语义资源建设[D]. 武汉大学.

曲维光, 周俊生, 吴晓东, 等. 2017. 自然语言句子抽象语义表示AMR研究综述 [J]. 数据采集与处理, 32(01):26-36.

孙超,曲维光,魏庭新,等. 2020. 基于神经网络的连动句识别[C]// In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*,pages 172–182.

邢欣. 1987. 简述连动式的结构特点及分析[J]. 新疆大学学报(哲学社会科学版), (01): 116-122.

徐情. 2012. 基于语料库的汉语连动结构语义研究[D]. 华中科技大学.

杨月蓉. 1992. 连动句和兼语句中的语义关系—兼论连动式与兼语式的区别[J]. 西南师范大学学报(人文社会科学版), (04): 96-100.

赵元任. 1952. 北京口语语法[M]. 北京: 开明书店.

赵元任. 2002. 中国话的文法[M]. 香港:香港中文大学出版社.

朱德熙. 1982. 语法讲义[M]. 北京: 商务印书社.

Chen J,Huang Y,Yang F,et al. 2020. A Novel Named Entity Recognition Approach of Judicial Case Texts Based on BiLSTM-CRF[C]// In *Proceedings of the 12th International Conference on Advanced Computational Intelligence*,pages 263–268.

Cui Y,Che W,Liu T,et al. 2019. Pre-Training with Whole Word Masking for Chinese BERT [J]. In *arXiv.preprint arXiv*,1906.08101.

Devlin J, Chang M-W, Lee K, et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]// In *The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,pages 4171-4186.

Greenberg N,Bansal T,Verga P,et al. 2018. Marginal Likelihood Training of BiLSTM-CRF for Biomedical Named Entity Recognition from Disjoint Label Sets[C]// In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*,pages 2824-2829.

Kruengkrai C,Nguyen T H,Mahani S A, et al . 2020. Improving Low-Resource Named Entity Recognition using Joint Sentence and Token Labeling[C]// In *Proceedings of the the 58th Annual Meeting of the Association for Computational Linguistics*,pages 5898–5905.

Lample G, Ballesteros M, Subramanian S, et al. 2016. Neural Architectures for Named Entity Recognition[C]// In *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,pages 260–270.

Liu W, Zhou P, Zhao Z, et al. 2020. K-BERT: Enabling Language Representation with Knowledge Graph// In *The Thirty-Fourth Conference on Artificial Intelligence.*,pages 2901-2908.

Wei Z, Su J, Wang Y, et al. 2020. A Novel Cascade Binary Tagging Framework for Relational Triple Extraction[C]// In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*,pages 1476-1488.

Wu S and He Y. 2019. Enriching Pre-trained Language Model with Entity Information for Relation Classification[C]// In *Proceedings of the 28th International Conference on Information and Knowledge Management*,pages 2361-2364.

Zhong Z and Chen D. 2020. A Frustratingly Easy Approach for Joint Entity and Relation Extraction [J]. In *arXiv.preprint arXiv*,2010.12812.