差比句结构及其缺省现象的识别补全研究

周鹏飞[1]，**曲维光**[1,2,4,*]，**魏庭新**[3]，**周俊生**[1]，**李斌**[2]，**顾彦慧**[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，江苏南京210097;
3.南京师范大学国际文化教育学院，江苏南京210097;
4.南京师范大学中北学院，江苏丹阳212334;
*通讯作者，Email：wgqu_nj@163.com)

摘要

差比句是用来表达两个或多个事物之间的相似或不同之处的句子结构，常用句式为“X比Y+比较结果”。差比句存在多种结构变体且大量存在省略现象，造成汉语语法研究和自然语言处理任务困难，因此实现差比句结构的识别和对其缺省结构进行补全非常有意义。本文采用序列化标注方法构建了一个差比句语料库，提出了一个能够融合字与词信息的LatticeBERT-BILSTM-CRF模型来对差比句结构自动识别，并且能对缺省单位进行自动补全，实验结果验证了方法的有效性。

关键词： 差比句结构；神经网络；缺省补全；信息融合

A Study on Identification and Completion of Comparative Sentence Structures with Ellipsis Phenomenon

ZHOU Pengfei[1]，**QU Weiguang**[1,2,4,*]，**WEI Tingxin**[3]，**ZHOU Junsheng**[1]，**LI Bin**[2]，**GU Yanhui**[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;
4.Zhongbei College, Nanjing Nornal University, Danyang , Jiangsu 212334, China;
*Corresponding, Email: wgqu_nj@163.com)

Abstract

Comparative sentences are sentence structures used to express the similarities or differences between two or more things. The common pattern is ”X compare to Y+comparison result.” There are various structural variants and they often involve ellipsis, which poses challenges in Chinese grammar research and natural language processing tasks. Therefore, it is of great value to recognize the comparative structures and complete their ellipse. In this paper, first we constructed a comparative sentence corpus using a sequential tagging method. Then we proposed a LatticeBERT-BILSTM-CRF model that can integrate the information of Words and Phrases to identify the comparative, and complete the ellipsis units automatically. Experimental results confirm the effectiveness of our approach.

Keywords: comparative sentences structure , neural network , ellipse completion , information fusion

1 引言

差比句是汉语语法中的一种特殊句式，因其简洁明了的表达方式和丰富多样的语义功能，被广泛应用于各个领域的语言表达。差比句在比较句的研究中占据重要的地位。现在通常认为一个完整的差比句需要四个要素：比较主体（记作X）、比较基准（记作Y）、比较标记和比较结果（记作R）。并且按照现代汉语语序应该表现为“X+比较标记+Y+W”。吴颖菲(2020)提到差比句有以下几种类型:

例1：摩托车的车速比自行车的车速更快。

例2：小明的身高跟小李的身高比更高。

例3：当今的公务员考试相比几年前的已经有了很大进步。

例4：你吃的饭比我多。

其中例1-2就是典型的完整差比句，如例1其中“摩托车的车速”为比较主体（X）,“比”为比较标记，“自行车的车速”为比较基准（Y），“更快”为比较结果（R）。例2中“跟.....比”作为比较标记。例3-4虽然句子结构很完整，但是其结构内部存在省略，如例3中的“几年前”作为比较基准（Y）省略了“公务员考试”这一部分内容，这类句子称为省略差比句。能够正确识别出这类差比句中的省略部分并将其进行补全，对于自然语言处理任务如机器翻译、信息抽取等工作都有重要的意义。然而，差比句的多种结构给机器识别差比句结构及自动缺省补全带来很大的挑战。

目前，对于差比句的研究主要集中在语言学领域。而在自然语言处理领域，对于差比句结构识别的研究相对较少。这是因为差比句的结构相当特殊，传统的句法识别方法并不适用。此外，目前还缺乏专门针对差比句的语料库，而且差比句中存在大量省略现象，通用的解析器方法往往只能进行省略判别而无法进行补全，导致效果不尽如人意。虽然其他一些特殊结构缺省补全工作已经取得了不错的效果(侍冰清等, 2019; 施寒瑜等, 2020)，但是它们的规则模型不符合差比句中的省略特点，因此并不适用于差比句。

本文的贡献如下:

1. 设计并制定了一套语料库标注规范，构建了一个包含多种类型差比句的语料库。该语料库包括5800条句子，其中涵盖了具有完整成分的差比句以及比较主体和比较基准先行语省略的差比句。

2.提出一个LatticeBERT-BILSTM-CRF模型来实现判别差比句类型并对差比句成分进行识别，该模型在BERT模型基础上加入了词的信息，引入新的位置编码和预训练任务，将字和词的信息融合输入BERT。模型在对差比句类型判别和差比句结构识别任务上取得了很好的效果，差比句结构识别任务F值达到了91.4%。

3.根据生成的句子标签，设计一套规则找到差比句中存在缺省的部分、识别出可补全的部分，并进行插入补全。

2 相关工作

2.1 差比句研究现状

差比句的语义功能丰富多样，包括了比较、强调、排比、修辞等多种功能。在研究差比句的语义方面，研究者通常会探讨差比句的语义特征、语义类别、语义关系等问题。例如，刘丹青(2004)从语言类型学的角度入手，结合Greenberg(1963)的研究结果，得出结论：汉语差比句的基本要素包括性质属性的主体、表示性质属性的形容词、基准和比较标记，分析在进行类型学调查时需要关注的差比句的要素。

差比句的语法结构由差异部分和比较部分组成，其表达形式多样。在研究差比句的语法结构方面，研究者通常关注其句子成分、成分的排列方式和缺省现象等问题。例如，任海波(1987)、邵敬敏，刘焱(2002)等人分别从句子成分的角度出发，探讨了差比句中的主语、谓语、宾语、定语和状语等成分在语法结构中的位置和作用，提出了差比句中成分排列的一些基本规律。在研究差比句的句法特征方面，研究者主要关注其语序、语法功能和句子结构等方面。例如吴颖菲(2020)通过对比分析汉语差比句的句子结构，发现汉语差比句中还存在一些仅仅含有差比语义但结构不固定的句子类型，引出了对多类非典型差比句的研究。

基金项目：国家社会科学基金重大项目(21&ZD288); 国家自然科学基金(62277031)。

这些研究者虽然对差比句句法语义做了大量研究，但均是从词典、公共语言数据集上做出的研究。未能构建专门差比句语料库进行研究，并且也未能与自然语言处理任务结合。

2.2 省略结构研究现状

省略结构是指在句子中省略某些成分而不影响句子的完整性和语义表达的现象，是语言学中一个重要的研究领域。在语法学层面，研究者通常会探讨省略结构的形式和语法特征。例如，赵元任(2002)等人研究了汉语省略结构的形式，提出了省略现象的三种类型，即语义省略、语用省略和结构省略，并探讨了各种省略现象的规律和特征，分析了不同省略形式的句法特征和语法功能。Marie Mikulová(2014)对英语和其他语言中省略结构的形式和语法特征进行分析，并提出了"链式理论"(Chain Theory)来解释省略结构的产生。

刘依欢(2020)在探讨汉语省略现象的抽象语义表示时，重点分析了省略句中省略内容和上下文之间的语义关系。她发现，大多数句子恢复省略成分依赖于先行语，先行语是指在句中的上下文中出现的与省略成分相同的词语。省略成分存在一个对应的先行语，省略现象与指代类似，其本质是回指。张伟男等(2009)采用决策树分类算法对中文对话中的省略现象进行判别，而杨国庆等(2011)则基于句子的基本结构，提出了省略三元规则进行省略识别。侍冰清(2019)研究了语义省略中的"的"字结构，发现其省略现象比较普遍，并对"的"字结构中的各个成分进行识别和补全，利用神经网络在构建的语料库上取得了87.1%的整体F值。施寒瑜(2020)同样采用神经网络的方法研究汉语数量名短语中的缺省成分，并通过联合学习方法，在构建的语料库上取得了85.07%的F值。

综上，差比句缺省形式多样、句法结构特殊，以往研究未能专门对差比句中的缺省现象进行系统性研究，并且目前尚未有对差比句内部缺省现象的补全工作。

3 差比句语料库构建

3.1 差比句语料库来源

本文选取1998年1至3月《人民日报》新闻语料以及北京语言大学语料库中心(BLCU Corpus Center)作为差比句标注的基础语料，此外还有一部分差比句来自CTB中文宾州树库，共计10300条，其中人工标注了5800条包含完整和省略的差比句供实验使用。

3.2 差比句语料库的构建

本文针对差比句的标注问题，使用了BIO 标注体系。该体系的优点在于能够很好地处理实体的边界问题。标注的基本组成结构为"比较主体(X) + 比较标记(C) + 比较基准(Y) + 比较结果(R)"。然而，如果仅将X、Y、C、R 作为标签进行标注，难以处理一些基本结构出现省略的情况。例如，"哥哥身高比弟弟更高"，这句话中的"弟弟"作为比较基准存在省略。若仅有四个标签，只能将其标注为Y，无法让机器学习到缺省的位置。另外，可以看到这类句子可以通过文中的先行语进行补全，而先行语作为补全的成分也需要一个标签。因此，本文对基本结构进行更加细粒度的拆分，引入标签X1、Y1 和K，分别标记缺省的比较主体、缺省的比较基准、先行语（句子中出现的与省略成分相同的词语）。这样做的好处是可以很好地处理省略的情况，提高标注的准确性和效率。因此最后的标签为X(比较主体) 、Y(比较基准)、C(比较标记)、R(比较结果)、X1(缺省比较主体)、Y1(缺省比较基准)、K(先行语)。

3.2.1 完整差比句结构的标注

完整的差比句结构需要同时满足下列情况：句子中出现完整的差比句结构且比较主体和比较基准两个比较对象属于同一个范畴。下面以具体例子分析完整差比句标注方法:

例5：今年工厂的实际投资额$_X$比$_C$去年工厂的实际投资额$_Y$高$_R$。

在例5中"工厂的实际投资额"和"店铺的实际投资额"比较的是同一个范畴，比较的主题都是"实际投资额"。因此"工厂的实际投资额"作为完整的比较主体标注为X，"店铺的实际投资额"作为完整的比较基准标注为Y，另外这句话还有完整的比较标记和比较结果，因此例5为完整的差比句，完整的差比句不应该存在X1,Y1标签。因为该句子不存在标签X1,Y1，所以不需要进行补全，具体标注示例可见表1。

工	厂	的	实	际	投	资	额	比	店
B-X	I-X	I-X	B-X	I-X	I-X	I-X	I-X	B-C	B-Y
铺	的	实	际	投	资	额	高	。	
I-Y	I-Y	B-Y	I-Y	I-Y	I-Y	I-Y	B-R	O	

Table 1: 完整差比句标注示例

3.3 缺省差比句结构的标注

在差比句中，省略现象多数是由于比较主体与比较基准的比较范畴不对应所导致的。例如，在例句6中，“厦门气温”是完整的比较主体，因此将“厦门”标记为比较主体中的修饰成分X，而将“气温”标注为先行语K。而“南京”作为比较基准存在省略，因此将其标记为“Y1”，以此表示省略。另外，“比”标注为比较标记C，“更高”则标注为比较结果R。在例句7中，省略成分的先行语是比较基准中的一部分，即“我的成绩”的中心语为“成绩”。此时，“你”并非完整的比较主体，因此将其标记为X1；“我的成绩”作为比较基准是完整的，因此将“我的”标记为修饰，标记为Y，而将“成绩”标记为比较主题，标记为可供比较主体缺省补全的先行语K。具体的标注示例可以参见表2。

例6：厦门$_X$气温$_K$比$_C$南京$_{Y1}$更高$_R$。

例7：你$_{X1}$比$_C$我$_Y$的成绩$_K$优秀$_R$。

厦	门	气	候	比	南	京	温	和	。
B-X	I-X	B-K	I-K	B-C	B-Y1	I-Y1	B-R	I-R	O

你	比	我	的	成	绩	优	秀	。
B-X1	B-C	B-Y	I-Y	B-K	I-K	B-R	I-R	O

Table 2: 缺省差比句标注示例

3.4 差比句语料库统计

本文依据3.1，3.2 节中方法进行标注，数据的选择考虑均衡每种差比句类型比例，让模型能够充分学习到不同类差比句的特征，数据集划分类型如表3所示。为确保标注数据质量，先制定详细标注规范，提供相应培训确保标注人员掌握规范。试标阶段对两名标注人员进行一致性统计，以评估其标注准确性和一致性。统计结果显示标注人员一致性存在差异时，进行调整，以保证正式标注的准确性和一致性。此外，对标注数据反复校验和核对，确保标注数据质量和准确性。最终标注数据达到较高一致性和准确性水平，为后续实验提供可靠数据基础。

句子类型	标注数量	样本样例
完整差比句	3000	今年产量比去年产量高
缺省差比句	2300	现在的生活比以前更好
非差比句	2000	他是一位正直的人

Table 3: 语料库成分

此外，还对不同种类的比较标记类型做了统计如表4所示。

比较标记	比	跟	和	相对	于	跟...比	过
数量（句）	2006	64	13	63	22	116	16
占比（%）	87.22	2.78	0.57	2.74	0.957	5.04	0.66

Table 4: 不同类型比较标记分布

第二十二届中国计算语言学大会论文集，第388页-第398页，哈尔滨，中国，2023年8月3日至5日。

4 差比句结构识别与缺省补全模型设计

本文将差比句结构识别视为缺省补全的一个子任务，通过识别出的标签信息进行缺省补全。基于序列化标注任务建模，解决差比句结构识别。由于差比句结构类型特殊，每个字都对应一个结构标签，容易出现边界识别错误。由此使用Lattice结构将词信息加入到预训练模型BERT来解决此问题。模型如图1所示，由四个层级构成：1. Lattice 结构转换层，该层用于将文本数据转换为lattice形式，从而支持Lattice-based模型的输入；2. Lattice-BERT编码层，将经过Lattice转换层处理后的Lattice序列作为输入，进行编码，得到一系列的向量表示；3.BI-LSTM层，用于从Lattice-BERT的输出中提取特征，以便更准确地预测每个位置的标签；4.CRF层, 用于将经过BI-LSTM层提取的特征序列转换成标注序列。它通过考虑相邻标记之间的关系，并对相邻标记之间的转移施加约束，来解决标记之间的依赖关系和标记边界问题，从而确保结构识别的完整性。

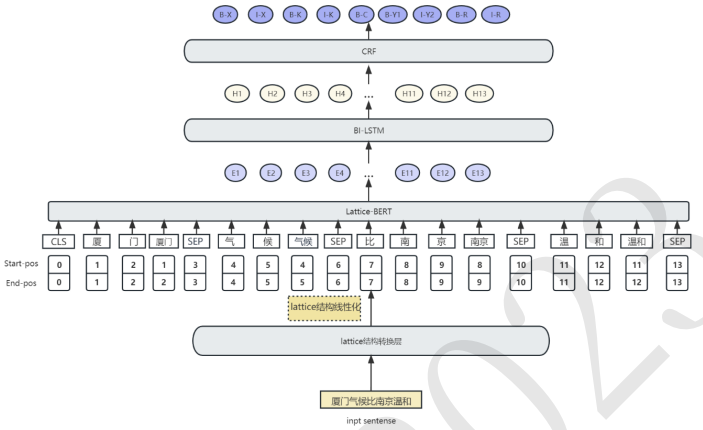

Figure 1: 模型结构组成图

4.1 Lattice结构转换

在预处理阶段，本文将中文文本转换为Lattice结构，转化过程通过以下步骤实现：

1.分词：首先使用字符词汇创建一个BERTTokenizer，再使用BERTTokenizer解析词典，获取每个单词的字符序列，然后测试字符范围是否与序列匹配。如果匹配，将其视为一个词。词汇的大小可以计算为字符词汇的大小加上单词词汇的大小。

2.构建词图：将分词得到的词语构建成一个有向无环图（DAG）。DAG的节点代表词语，边表示两个词语之间的关系。比如“厦门气候比南京温和”这句话中包含五个词语“厦门”，“比”，“南京”，“气候”“温和”则可以构建一个DAG，其中节点包括“厦”，“门”，“厦门”，“比”，“南”，“京”，“南京”，“气”，“候”，“气候”，“温”，“和”，“温和”，同时建立从“厦”指向“门”，厦门指向“夏”和“门”两个节点，同时又指向“气”节点，如此一直到最后，如图2 所示。

3.填充标签：将第三章设计的七个不同标签，采用BIO的标注方法对差比句各个结构进行标注。

4.线性化：将转换的lattice结构图线性化作为输入，如图2所示直接将词格图中各粒度的信息“拍平”，得到图1中输入的线性序列。

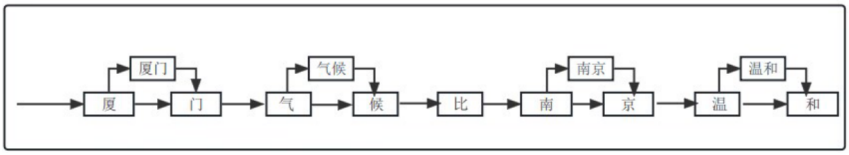

Figure 2: Lattice结构生成图

第二十二届中国计算语言学大会论文集，第388页-第398页，哈尔滨，中国，2023年8月3日至5日。

4.2 文本表示层

中文以词为语义基本单位，因此传统中文NLP都需要先进行分词处理，但是分词的难度很大，经常需要处理一些歧义和多义情况。在预训练模型中，中文模型通常采用字作为输入，这种输入会忽略中文的粗粒度信息，词语对于中文的语义理解来说至关重要。Lattice结构可以让BERT在预训练中学习到字和词的信息。

差比句结构识别任务对于普通序列化标注任务有如下特点：1、差比句结构标签分布很密集，几乎每个字都有一个对应的标签。这意味着对于每个字，其上下词信息对于正确预测它的标签非常重要；2、在差比句中，有时候会出现两个或多个实体的首尾相接或者互相嵌套的情况。因此词信息在差比句结构识别任务中尤为重要，lattice结构可以将输入融合词的信息(Matthias Sperber et al., 2019)，因此本文利用Lattice结构图输入到预训练模型BERT来得到一个同时包含字和词信息的文本表示。本文使用的方法是将Lattice结构图中各粒度的信息“拍平”，得到一个线性序列。然而，“拍平”词格的输入会导致重复和冗余的问题，进而影响位置编码的适应性。除此之外在“拍平”之后，原先二维的复杂图结构信息也会有所损失。为了解决这两个问题，模型设计了新的注意力机制。首先改进了BERT的绝对位置编码，式1里的P_i代表相对位置长度，由P^S 和P^E做差得到，P^S表示当前输入token的开始位置，P^E 表示结束的位置。式2将token的起始位置的绝对位置编码拼接，进行attention 操作，从而得到相对位置编码。模型除了保留原始BERT中的位置编码，还加入了词格输入。由于词格输入的每一项长度是不固定的，因此引入头尾位置，对应图1中的start-pos和end-pos。

然而，光是绝对位置编码所提供的信息还不够充足，因为在理论上对绝对位置编码的限制只有一点，即不同位置的编码不同。这样会忽略了很多信息，比如:位置1和2的距离与位置5和6的距离应该一样，位置1和3的距离比位置4和10的距离要小等等。在绝对位置编码的设计上只能让BERT隐式地“学习”。因此后续还加入相对位置编码以及针对层叠问题加入层叠编码信息，来表示token 之间的相对距离。式3中的$\tilde{\alpha}_{ij}^{l}$第一项是字的表示得到的attention score，第二项是绝对位置编码，相对位置编码，层叠编码相加，相对位置编码为P^E，P^S之差，层叠编码根据这两个token 起始相对位置的不同，两个token 可以分成下列七种关系(Yuxuan Lai et al., 2021)T.1自身；T.2在左边，且无重叠；T.3在左边，且有重叠；T.4包含关系；T.5被包含关系；T.6在右边，且有重叠；T.7在右边，且无重叠。

$$P_i = \left| P^E - P^S \right| \tag{1}$$

$$\tilde{\epsilon} = \epsilon_i^{in,0} + P_i \tag{2}$$

$$\tilde{\alpha}_{ij}^{l} = \alpha_{ij}^{l} + f(i,j) \tag{3}$$

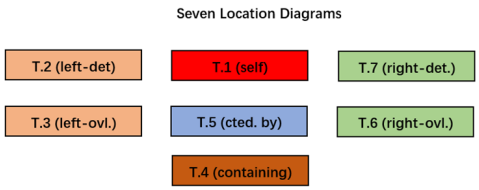

Figure 3: 七种位置关系

此外，Lattice-BERT引入了一些新的预训练任务：Masked Segment Prediction来取代原先的Masked Language Modeling（MLM）任务。MSP任务的训练数据是单个句子，因此每个训练样本都是单个句子的segment，其目标是预测输入句子中哪些部分是连续的。在MSP任务中，将输入句子切分为多个segment，并在其中随机mask掉一些segment，然后让模型预测哪些segment是连续的。这个任务的目的是让预训练模型学习到segment之间的连续性关系，进一步提升对于长文本的建模能力，以进一步提高中文自然语言处理的性能。

第二十二届中国计算语言学大会论文集，第388页-第398页，哈尔滨，中国，2023年8月3日至5日。

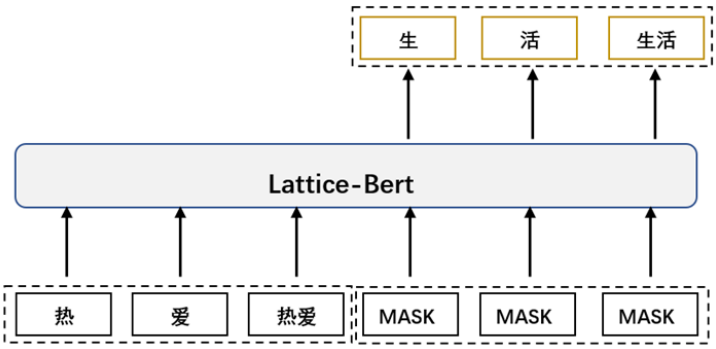

Figure 4: MSP任务mask示例

4.3 特征提取层

BI-LSTM（Bidirectional Long Short-Term Memory）是一种循环神经网络（RNN）的变体，它在自然语言处理领域广泛应用于序列标注任务，如命名实体识别（NER）和情感分析等。相比传统的单向LSTM，BI-LSTM不仅考虑了当前时刻之前的信息，还考虑了当前时刻之后的信息，从而更好地捕捉了序列中的上下文信息。BI-LSTM通过在正向和反向两个方向上分别处理输入序列，得到两个独立的隐藏状态向量，然后将这两个向量按位置相加，作为输出。这样可以使每个位置的输出同时考虑当前位置之前和之后的上下文信息。

相比于CNN，BI-LSTM具有更强的序列建模能力。Lattice-BERT的输入是嵌套的lattice结构，如果使用CNN对嵌套的结构进行建模，需要将lattice结构“拍平”成一维的序列，这样就可能损失掉一些与嵌套结构相关的信息。而BI-LSTM能够对序列中的上下文信息进行建模，更加适合处理嵌套结构。

相比于LSTM，BI-LSTM在提取特征时能够同时考虑上下文的信息。LSTM是一种单向模型，只能考虑当前时刻之前的信息，而BI-LSTM是一种双向模型，能够同时考虑当前时刻之前和之后的信息，因此能够更好地捕捉上下文信息。

因此，将BI-LSTM与Lattice-BERT结合能够充分利用Lattice-BERT的字和词级别信息，同时对序列中的上下文信息进行建模，提高序列化标注任务的准确性。

4.4 输出层

模型采用CRF作为序列标注的输出层，它可以通过整合上下文信息和标签约束来提高模型的性能。在Lattice-BERT模型中，CRF层用于将经过BI-LSTM层提取的特征序列转换成标注序列。它通过考虑相邻标记之间的关系，并对相邻标记之间的转移施加约束，来解决标记之间的依赖关系和标记边界问题，从而提高模型在序列标注任务中的性能。

4.5 缺省补全

缺省补全解析器基于句子中的先行语和缺省结构进行补全，即通过插入省略结构来实现缺省部分的补全。具体地，本文利用核心思想，即先找到先行语，然后在省略结构中插入先行语K，如果遇到K标签，X1或Y1则说明存在缺省内容，X1或Y1标签后面的位置就是先行语K需要插入的位置。为了实现这一目标，本文设计了一套补全规则流程图，如图5所示。该流程图具有一定的纠错功能，如果K标签和X1或Y1标签没有同时出现，则输出error并且需要进一步检查输出结果的准确性。

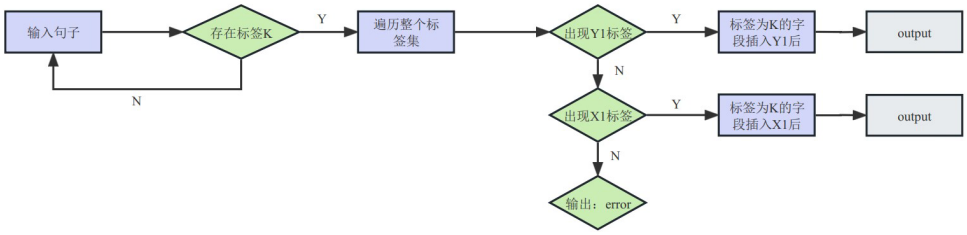

Figure 5: 缺省补全流程图

例如，当图6的第一个输入作为缺省补全解析器的输入时，解析器会首先分析句子的标签。在分析过程中，发现“气候”的标签为K，意味着“气候”是比较基准。随后，解析器会遍历整个标签集，发现“南京”的标签为Y1，说明比较基准部分存在缺省。为了补全缺省部分，先行语需要复制并补充在比较基准“南京”之后。最终，句子变成了“厦门气候比南京气候温和”，并被加入典型差比句语料库中。解析器具有标签错误检测功能，对于第二个输入，解析器检测到“成绩”为K标签，因此开始遍历标签集。然而，在遍历过程中，解析器发现句子中不存在X1和Y1标签，这意味着标签生成出现错误，句子无法被正确补全。因此，这个输入被视为含有错误生成标签，并输出到错误标签样本类进行人工处理。

通过这样的解析过程，缺省补全解析器可以对含有缺省部分的句子进行自动补全。对于能够正确补全的句子，可以将其加入典型差比句语料库中，以便后续的分析和应用。而对于无法正确补全的句子，则需要进行人工处理，以确保整个模型的准确性和可靠性。

Figure 6: 缺省补全解析器输入样例

5 实验

5.1 数据集划分和模型参数设置

实验将数据集由完整差比句、省略差比句、非差比句三类句型组成，按照8：1：1的比例划分为训练集、验证集和测试集，再将句子随机打乱。数据集如表5所示。其中训练集用于模型学习，拟合分类器的参数（普通参数、神经元的权重等）；测试集用于调整模型的参数，比如：确定隐藏单元数、确定神经网络结构和复杂程度的参数（超参数：如隐藏层数、每一层的神经元数等）；验证集用于测试模型的表现。

类别	训练集	测试集	验证集
完整差比句	2400	300	300
省略差比句	1840	230	230
非差比句	1000	500	500

Table 5: 数据集划分

本文采用Lattice-BERT预训练模型chinese_laBERT-tiny-std-512，超参数具体设置如表6。

超参数	参数值
Epochs(迭代次数)	12
batch size（单批次处理数量）	30
Dropout（丢弃率）	0.01
Learning Rate（学习率）	1e-5
Segment Size（每段最大token数量）	64
Embedding Size（输入最大token数量）	128

Table 6: 实验参数设置

5.2 评价方法

评价方法使用准确率P、召回率R和F1值对模型的性能进行评价，计算公式如下：

$$P = \frac{TP}{TP + FP} \tag{4}$$

第二十二届中国计算语言学大会论文集，第388页-第398页，哈尔滨，中国，2023年8月3日至5日。

$$R = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{6}$$

5.3 实验结果与分析

本文将编码层和特征提取层分别引入不同的模型与我们构建的模型作为比对。其中包括：BERT+BILSTM+CRF、LatticeBERT+CNN+CRF、LBERT+LSTM+CRF。整体实验结果如表7所示。实验结果显示，本文提出的模型在P、R和F1三个指标上都优于基线模型。可以看出，仅使用BILSTM+CRF时，F值仅为77.28，当加入BERT编码层时，性能得到提升可以达到88.19，而将编码层换成Lattcie-BERT可以进一步提升性。在编码层选择用Lattice-Bert的基础下，特征提取层选用BILSTM可以使得性能达到最优值91.57。这说明本文的模型能够更准确地识别标签，并且相较于基线模型具有更好的综合性能。

从实验结果可以看出，相较于常用的BERT编码器Lattice-BERT实验结果上有明显的优势。因为一般我们常用差比句中几乎每个词都会打上一个标签，因此容易错误地将距离近的标签作为错误生成，尤其体现在先行语和与其构成完成比较主体（基准）的其他内容上。BERT模型没有词的信息很容易扩大或缩小生成先行语的标签，从而造成缺省补全的困难，引入词的信息明显可以帮助模型理解差比句的各个结构，使词与词之间位置信息更加明确。

本文的模型采用了多层Lattice-BERT、BILSTM和CRF等技术，能够有效地解决结构识别任务中的嵌套标签问题，同时还具有更好的语义表示能力。这些技术的组合使得本文的模型在处理结构识别任务时能够更加全面地考虑句子中字和词的信息，从而提高了识别的准确率和召回率。

实验模型	P(%)	R(%)	F1(%)
BILSTM+CRF	78.34	76.25	77.28
BERT+BILSTM+CRF	87.63	88.76	88.19
LatticeBERT+CNN+CRF	90.16	89.24	89.70
LatticeBERT+LSTM+CRF	91.31	90.34	90.82
本文模型	**92.39**	**90.76**	**91.57**

Table 7: 差比句结构识别结果

表8展示了用于缺省补全的先行语识别结果。由于先行词通常存在于比较基准和比较主体之间，因此会造成先行语和缺省比较主体（基准）标签之间的边界信息和位置信息识别困难。从而经常造成生成的标签之间相互越界。观察表8可以发现，相对于其他模型，各模型对于先行词的识别整体水平普遍下降。然而，在本文模型中，通过在编码过程中引入词的信息和多种位置编码的信息，使得本文模型受到的影响最小。这说明本文模型的识别更加稳定，同时也显著提高了后续缺省补全的准确率。

实验模型	P(%)	R(%)	F1(%)
BILSTM+CRF	73.25	71.51	72.37
LatticeBERT+CNN+CRF	86.63	87.16	86.89
LatticeBERT+LSTM+CRF	88.47	87.56	88.01
BERT+BILSTM+CRF	88.65	88.37	88.51
本文模型	**91.23**	**90.41**	**90.82**

Table 8: 省略差比句中先行词识别结果

5.4 消融实验

本文通过使用Softmax层代替CRF层；减去BILSTM层分别做了两次实验，如表9所示。

实验模型	P(%)	R(%)	F1(%)
Ours	**92.39**	**90.76**	**91.57**
Ours-BILSTM	88.70	88.24	88.47
Ours-CRF	91.85	90.10	90.97

Table 9: 差比句结构识别消融实验结果

实验发现，两个消融实验在P、R和F1三个指标上相较于本文模型均有所下降。原因是虽然Lattice-BERT编码已经可以通过引入词信息，很好地表示结构特征。但由于输入内容是一张“拍平”后的图，输入过于冗长因此会丢失一些长距离信息。而BILSTM层的加入能够更好地学习长距离依赖关系且BILSTM很适合处理嵌套关系信息，由此加入BILSTM作为特征提取层可以很好缓解由于输入过长而导致信息丢失的问题。此外加入CRF是由于其可以自动学习序列标注的规律且能够更好地处理复杂的标注模式，包括多个标签之间的依赖关系和标签转换模式，从而使系统更加准确和稳定。基于将这些模块结合，能够更好地处理差比句识别和补全任务，实验结果表明，我们所选取的另外两层模块每一层对于整体模型都有一定提升。

5.5 缺省补全实验

本实验采用了两种基线模型，分别是BILSTM+CRF和BERT+BILSTM+CRF，这两种模型生成的标签被送入缺省补全解析器中，用于对比本文设计的解析器的性能。实验结果见表10。由于LatticeBERT+BILSTM+CRF解决了先行语标签生成容易和存在缺省标签互相“越界“问题，从而能够精准识别出补全的关键标签—先行语和缺省比较主体（基准）。实验结果表明LatticeBERT+BILSTM+CRF模型与本文设计的解析器适应度最高，P、R、F1值均达到了最高水平。

Model	P（%）	R（%）	F1（%）
BILSTM+CRF+缺省补全解析器	68.77	75.28	71.88
BERT+BILSTM+CRF+缺省补全解析器	80.34	81.50	80.92
LatticeBERT+BILSTM+CRF+缺省补全解析器	**86.49**	**88.21**	**87.34**

Table 10: 缺省补全各模式实验结果

解析器生成的错误可以分为三类，如表11所示。第一种类型是模型仅生成了K标签，而没有生成X1或者Y1。当该标签被送到解析器时，解析器无法找到合适的位置插入先行语，从而导致错误。第二种类型的错误是模型生成的K标签覆盖了比较主体或比较基准的过多部分，在补全后形成的句子语义混乱。第三种类型的错误是模型生成的比较主体和比较基准标签存在中断，同样会在补全后形成语义混乱的句子。这三种类型的错误都与模型标签生成的准确性直接相关。对于这样的错误案例分析错误的原因如下:1.词典构建不完善，在分词时出现错误进而导致生成错误的Lattice结构。2.数据集规模仍相对较小，对于一些句子类别语料库内较少，因此模型学习时有一定困难。后续会继续在这两个问题上做出相应改进。

原句子生成补全标签	结果类型	输出句子
若长此以往，以后$_X$的情况$_K$会比现在$_{Y1}$更差。	正确输出	若长此以往，以后的情况会比现在的情况更差。
两者斟酌，前者$_X$严重程度$_K$固然比后者$_Y$轻。	错误1	Error
太平洋面积$_X$是其余三大洋的总和$_K$，比北冰洋$_{Y1}$大十四倍。	错误2	太平洋面积是其余三大洋的总和，比**北冰洋面积是其余三大洋的总和**大十四倍。
1993年$_X$新钢利润$_K$比上$_{Y1}$年下降88.9%	错误3	1993年新钢利润**比上新钢利润年**下降88.9%

Table 11: 补全转换示例

6 结语

本文通过神经网络对差比句结构及其缺省情况进行了研究，利用BIO方式标注的数据集进行模型训练，采用Lattice-BERT与BILSTM结合的复合模型成功识别差比句结构并补全缺省部分。未来，本文将继续研究不规则的差比句并进一步优化模型以提高其性能。具体来说，我们将尝试通过篇章级别的学习，让模型脱离先行语的限制对差比句的所有省略结构进行识别。

参考文献

侍冰清, 戴茹冰, 曲维光, 顾彦慧, 周俊生, 李斌, 徐戈, 史胜旺. 2019. 基于组合神经网络的语义省略“的”字结构识别. 北京大学学报(自然科学版), 2019, 55(01): 75-83.

施寒瑜, 曲维光, 魏庭新, 周俊生, 顾彦慧. 2022. 基于组合深度模型的现代汉语数量名短语识别. 南京师大学报(自然科学版), 2022, 45(01): 127-135.

刘丹青. 2004. 差比句的调查框架与研究思路. 现代语言学理论与中国少数民族语言研究. 北京: 民族出版社, 2004: 1-22.

邵敬敏, 刘焱. 2002. 比字句强制性语义要求的句法表现. 汉语学习, 2002（5）: 3-7.

任海波. 1987. 现代汉语”比”字句结论项的类型. 语言教学与研究, 1987(4):91-103.

吴颖菲. 2020. 汉语非典型差比句的研究与教学. 华东师范大学.

赵元任. 2002. 赵元任语言学论文集. 北京：商务印书馆，2002：61-72.

刘依欢. 2020. 基于抽象语义表示的省略现象研究. 南京师范大学.

张伟男, 张宇, 刘挺. 2009. 基于决策树的中文对话省略句判别.中国中文信息学会会议论文集, 2009:315-322.

杨国庆, 孔芳. 2011. 基于规则的中文缺省识别研究. 计算机科学, 2011(12): 255-258.

邓依依, 邬昌兴, 魏永丰, 等. 2021. 基于深度学习的命名实体识别综述. 中文信息学报, 2021, 35(9): 30-45.

郑远汉. 1998. 省略句的性质及其规范问题. 语言文字应用, 1998(02): 12-19+3.

李艳惠. 2005. 省略与成分缺失. 语言科学, 2005(02): 3-19.

Yuxuan Lai, Yijia Liu, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2021.

Lattice-BERT: Leveraging Multi-Granularity Representations in Chinese Pre-trained Language Models. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1716–1731.

Greenberg J H. 1963. Some universals of grammar with particular reference to the order of meaningful elements. Universals of language, 1963: 73-113.

Marie Mikulová. 2014. Semantic Representation of Ellipsis in the Prague Dependency Treebanks. In Proceedings of the 26th Conference on Computational Linguistics and Speech Processing (ROCLING 2014), 125–138.

Matthias Sperber, Graham Neubig, Ngoc-Quan Pham, and Alex Waibel. 2019. Self-Attentional Models for Lattice Inputs. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 1185–1197.

Devlin J, Chang M W, Lee K, and Toutanova K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT, 2019: 4171-4186.

Jason P. C. Chiu and Eric Nichols. 2016. Named Entity Recognition with Bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics, 357-370.

Yue Zhang and Jie Yang.2018. Chinese NER Using Lattice LSTM. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics,1554-1564.

Song K, Tan X, Qin T, et al. 2019. MASS: Masked Sequence to Sequence Pre-training for Language Generation. International Conference on Machine Learning, 5926-5936.