面向中文AMR标注体系的兼语语料库构建及识别研究

侯文惠[1], 曲维光[1,2], 魏庭新[2,3], 李斌[2], 顾彦慧[1], 周俊生[1]
(1.南京师范大学 计算机科学与技术学院，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)

摘要

兼语结构是汉语中常见的一种动词结构，由述宾短语与主谓短语共享兼语，结构复杂，给句法分析造成困难，因此兼语语料库构建及识别工作对于语义解析及下游任务都具有重要意义。但现存兼语语料库较少，面向中文AMR标注体系的兼语语料库构建仍处于空白阶段。针对这一现状，本文总结了一套兼语语料库标注规范，并构建了一定数量面向中文AMR标注体系的兼语语料库。基于构建的语料库，采用基于字符的神经网络模型识别兼语结构，并对识别结果以及未来的改进方向进行分析总结。

关键词： 中文AMR ； 兼语结构 ； 识别

Research on the Construction and Recognition of Concurrent corpus for Chinese AMR Annotation System

HOU Wenhui[1], QU Weiguang[1,2], WEI Tingxin[2,3], LI Bin[2], GU Yanhui[1], ZHOU Junsheng[1]
(1.School of Computer Science and Technology, 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)

Abstract

The concurrent structure is a common verb structure in Chinese. The predicate phrase and the subject-predicate phrase share the concurrent structure. The structure is complex and difficult to analyze. Therefore, the construction and recognition of the concurrent corpus is of great significance for semantic analysis and downstream tasks. However, there are few existing concurrent corpora, and the construction of concurrent corpora for the Chinese AMR labeling system is still in the blank stage. In response to this situation, this paper summarizes a set of concurrent corpus annotation specifications, and builds a number of concurrent corpora for Chinese AMR annotation systems. Based on the constructed corpus, this article uses an character-based neural network model to recognize the concurrent structure, and analyzes and summarizes the recognition results and future improvements.

Keywords: Chinese AMR , Concurrent structure , Recognition

收稿日期: **定稿日期**:
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

兼语结构是由述宾短语与主谓短语套接而成的一种动词结构，述宾短语的宾语同时做主谓短语的主语，其结构通常表示为NP1+V1+NP2+V2(周鸣, 2018)。如“老 师 让 大 家 补 选 一 名 劳 动 委 员 。”是一个典型的含有兼语结构的兼语句，该句中的“大家”既充当“让”的宾语又充当“补选”的主语。兼语结构与连动结构以及主谓短语做宾语结构相似，且存在共享省略成分，使得兼语句的识别与解析十分困难。据李斌(2017) 统计，兼语结构普遍存在于汉语语料中。因此，正确识别兼语结构，对句子的语义解析及其他下游任务具有重要意义。

语言学领域对兼语结构做了大量的研究，其工作集中在兼语句分类、语义研究、偏误分析等方面。然而在自然语言处理领域，针对兼语结构识别及相应的语料资源构建的研究较少。部分现有语料(周强, 2004; 郭丽娟, 2019) 中包含兼语结构的标注，但其不针对兼语构建，规模较小，规范不统一，无法用于兼语结构的识别及深入研究。现有的兼语结构识别工作依赖分词以及词性标注的效果，对未经人工校对的语料识别效果较差，对于低频兼语动词的识别能力有限。

抽象语义表示(abstract meaning representation, AMR)是一种新型的语义表示方法，它从语义角度出发，通过补充句子中的隐含或省略成分，更全面地描述句子的语义(曲维光 et al., 2017)，在语义解析任务中更具优势。AMR 在对语义标注时需要对兼语缺省的论元进行补充，因此，自动识别出兼语结构，将其转化为AMR图，可以辅助中文AMR语料的构建及解析，为语义解析及下游任务提供帮助。然而，中文AMR语料中兼语句较少，不足以用于训练，因此需要构建一定规模的兼语语料。

针对这一现状，我们构建了一个面向中文AMR标注体系的兼语语料库，并对语料库进行了统计分析。基于该语料，我们使用添加词典信息的字符神经网络模型识别兼语结构的边界，并对识别结果进行分析总结，讨论了未来可以改进的方向。

本文后续组织如下：第一节对以往的相关研究工作进行总结；第二节介绍面向中文AMR标注体系的兼语语料库构建工作，其中包括兼语结构界定、语料库构建规范以及统计分析；第三节主要介绍兼语结构识别的问题定义及模型；第四节介绍了兼语结构识别实验的结果及分析；最后一节总结全文并对未来的改进研究提出方向。

2 相关工作

语言学领域对于兼语句的理论及应用研究十分深入，也为自然语言处理领域的研究奠定了基础，但是兼语语料资源的匮乏限制了兼语结构识别的研究。

2.1 兼语语料库构建研究现状

兼语结构广泛存在于汉语中，语言学领域关于兼语结构的研究层出不穷。其研究工作主要集中在兼语句分类、语义研究、偏误分析等方面。胡裕树(1962/1979)将兼语句分为使令、促成类和有无类两类。邢福义、汪国胜(2010)则主张将兼语句分为使令式、爱恨式、有无式三类,并提到了连动兼语混用的情况。李婷玉(2017)从V1出发将兼语句式分为八个大类，并在语义分类和描述框架的基础上对八个大类进行细分。马德全、王利民(2010)对V1的二价动词和三价动词的应用进行了考察。司玉英(2010)对双宾兼语句各成分之间的语义关系进行了分析。

然而，针对兼语语料资源构建的工作较少，只有少数综合语料中包含兼语结构标注。周强(2004) 构建TCT(Tsinghua Chinese Treebank)的句法标记集采用功能分类的方法对汉语短语进行描述。其中，兼语结构是一类特殊的动词短语，使用“vp-JY”作为标识，兼语动词使用“vJY”标注。TCT对兼语结构这类特殊的动词短语有明确的边界标注。但该语料库对于兼语结构以及主谓短语做宾语结构的界定模糊，且未对兼语结构中的兼语以及V2进行标注。中文依存句法树库中也包含对兼语的标注，HIT-CDT(Harbin Institute of Technology Chinese Dependency Treebank)中使用“DBL”这个依存关系类型标注V1以及兼语中心词的依存关系，SU-CDT(SUDA Chinese Dependency Treebank)(郭丽娟, 2019) 在此标注系统的基础上增加一个“pred”依存关系类型，该关系类型用来标注兼语指向V2的关系，使得兼语与V2的语义关系更加紧密。李斌(2017)等构建的中文AMR语料将一个句子抽象为一个AMR 图，通过补充句子中的隐含或省略成分，完善句子的语义信息。兼语结构是典型的包含省略成分的结构，AMR会将共享的兼语进行补充。但以上的语料并不针对兼语构建，规模较小，且各语料对于兼语的定义不统一，无法直接用于兼语结构的识别工作。

第十九届中国计算语言学大会论文集，第528页-第538页，海南，中国，2020年10月30日至11月1日。

2.2 兼语结构识别研究现状

现有的兼语结构识别研究主要分为两类，一类是基于规则的识别方法，另一类是基于机器学习的识别方法。傅成宏(2007) 统计分析了1998年《人民日报》1月份语料，通过建立兼语动词词表识别V1，并在此基础上使用规则方法识别兼语边界，进而确定兼语结构的存在。但兼语动词词表的建立需要依赖语料，难以建立适合所有语料的兼语动词词表，无法识别新产生的兼语动词。且该方法只能识别符合语法规则的简单兼语结构，无法处理不符合语法规则的句子以及包含连动、复句等其他结构的复杂兼语结构，也无法对完整的兼语结构进行识别，难以达到应用层面。陈静(2012) 等将兼语结构边界识别问题转化为序列化标注问题，使用CRF模型识别兼语结构的边界。但是，该工作基于人工校对的语料进行，依赖分词以及词性标注的效果，对于大量未加工的语料识别效果较差。兼语中存在大量低频兼语动词，且其“使令”含义不强，CRF模型的识别效果有限。近年来，神经网络模型的出现有效提高了序列化标注任务的效果。词性标注以及命名实体识别通常被建模为序列化标注任务解决，神经网络因其具有更好的泛化性和不依赖手工选择特征等特点而被广泛应用于词性标注以及命名实体识别。Pinheiro和Collobert(2014)等首次将CNN模型与CRF模型结合，并用于命名实体识别任务。Chiu和Nichols(2016)结合CNN模型与LSTM模型，然后与CRF模型拼接，进一步提高了命名实体识别的效果。近年来，基于字符的神经网络模型被广泛应用于命名实体识别任务，Zhang(2018)等提出的Lattice LSTM模型在中文命名实体识别中取得了较好的结果。目前为止，还没有使用神经网络模型识别兼语结构的研究。

相关语料的缺乏限制了兼语结构边界识别任务的解决和提升。为了对兼语结构进行语义解析，需要构建更为细致的语料，因此本文构建了一个面向中文AMR标注体系的兼语语料库，并使用添加词典信息的字符神经网络模型识别兼语结构，缓解了分词系统造成的错误传播，有效提高了兼语结构的识别效果。

3 兼语语料库构建

本文首先对兼语结构进行界定，筛选出兼语句，然后根据标注规范构建语料库，最后对构建的语料库进行了统计分析。

3.1 兼语结构的界定

兼语结构是一种套接的动词结构，通常将结构表示为NP1+V1+NP2+V2(周鸣, 2018)，其中V1和V2的关系为递系式(张志公, 1957)，NP2为兼语，V1一般是具有“使令”含义的动词，V1和V2共享NP2，NP2分别做V1和V2的受事宾语和施事主语。AMR在标注兼语结构时，将NP2标注为V1的arg1，V2的arg0。具体示例如图1所示。

但是汉语句式复杂多变，仅凭结构难以识别，其中连动结构以及主谓短语做宾语结构与兼语结构尤为相似，需要结合结构和语义进行判定。具体界定过程一般分为两步。(1)筛选具有NP1+V1+NP2+V2结构，且NP2充当V1宾语，V2主语的句子。(2)判断V1宾语涉及的范围是NP2还是整个主谓结构构成的短语或从句，如果只涉及NP2则判定为兼语句，否则，判定为非兼语句。

TCT中对兼语结构和主谓短语做宾语的界定模糊。“建议纪委介入调查”是一个典型的主谓短语做宾语句，“建议”的内容是“纪委介入调查”，涉及的范围是其后的整个从句，但TCT将其标注为兼语结构。本文在构建语料库时，综合考虑了以上两个界定步骤，有效避免了两类结构界定模糊的问题。

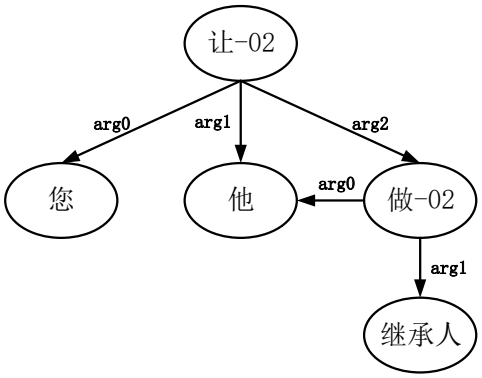
3.2 兼语语料库构建规范

本文构建的兼语语料库主要对兼语结构的边界、兼语的中心词以及V1、V2 进行标注。

3.2.1 兼语结构的前边界

本文语料库将兼语结构的前边界标注在V1前，并用“【【”标注。如果兼语结构的V1存在于连动结构中，则将兼语结构的前边界标注到连动结构的第一个动词前。

例1：**他 【【号_V1 召_V1 和 动 员 全 体 指_JY 战_JY 员_JY 节_V2 衣_V2 缩_V2 食_V2】】 。**



句子：您让他做继承人
x2/让
:arg0 x1/您
:arg1 x3/他
:arg2 x4/做
:arg0 x3/他
:arg1 x5/继承人

Figure 1: 兼语结构AMR图

3.2.2 兼语结构的后边界

本文语料库将兼语结构的后边界标注在V2所在的动词短语后，用“】】”标注，如果兼语结构的V2存在于连动结构中，则将兼语结构的后边界标注到连动结构中最后一个动词所在的动词短语之后。

例2：**它 能 【【帮_V1 助_V1 人_JY 类_JY 开_V2 拓_V2 未 知 的 领 域 和 获 得 新 的 知 识】】 。**

说明：例2中的“开拓”和“获得”的主语都是“人类”，且可以将其拆分成两个兼语结构，一个是“帮助人类开拓未知的领域”，一个是“帮助人类获得新的知识”。后边界标注到连动结构的最后一个动词所在的动词短语之后。

3.2.3 V1的标注

本文语料库使用“_V1”标注V1，如果兼语结构的V1存在于连动结构中，则只标注连动结构中的第一个兼语动词。具体情况如例1所示。

3.2.4 兼语标注

本文语料库使用“_JY”标注兼语。汉语中的兼语通常为一个名词、代词、名词短语或一个主谓宾结构，本文在标注兼语时只标注其中心词。针对各类复杂情况，本文对兼语标注规范进行以下细化规定。

（1）如果兼语为名词短语，则标注名词短语的中心词。

例3：**奏 鸣 曲 【【让_V1 专 修 音 乐 的 妹_JY 妹_JY 大_V2 吃_V2 一_V2 惊_V2】】 。**

说明：该例句中“专修音乐的妹妹”构成的名词短语充当兼语，“妹妹”为该名词短语的中心词，故只对“妹妹”进行标注。

（2）如果兼语是由多个名词或名词短语并列组成，则对其中的每一个名词或名词短语的中心词进行标注。

例4：**能 够 【【让_V1 灾 区 的 孩_JY 子_JY 、 学_JY 生_JY 得_V2 到_V2 相 应 的 关 怀】】 就 够 了 。**

说明：该例句中“灾区的孩子、学生”两个并列的名词短语充当兼语，“孩子”和“学生”分别为两个名词短语的中心词，故对这两个词进行标注。

（3）如果兼语由一个完整的主谓宾结构构成，则标注该结构的中心谓词。

例5：**【【使_V1 高 速 度 大 容 量 异 种 机 传_JY 输_JY 信 息 成_V2 为_V2 可 能】】 。**

说明：该例句中“高速度大容量异种机传输信息”为一个完整的主谓宾结构，其中的谓词“传输”为该结构的中心词，故对其进行标注。

3.2.5 V2的标注

本文语料库使用“_V2”标注V2，针对兼语句中包含连动、复句、以及其他修饰成分等复杂情况，对V2 的标注规范进行以下细化规定。

（1）如果兼语结构主谓词组的谓词存在于连动结构中，则将V2标注为连动结构中的第一个动词。该类型是包含连动的复杂兼语结构，具体例子如下。

例6：**把 读 书 当 成 【【使_V1 人_JY 信_V2 教_V2 修 行】】 的 一 种 手 段 。**

说明：例6中的“信教”和“修行”构成连动结构，我们将V2标注为连动结构的第一个动词“信教”。

（2）如果句中包含“去吃饭”、“来做客”这类连动结构，AMR会将“去”和“来”这类无实际含义的词省略，本文在标注V2时标注第一个动词。

例7：**他 们 【【邀_V1 请_V1 全 国 1 8 家 甲 级 城 市 规 划 设 计 院 的 专_JY 家_JY 来_V2 考 察 论 证】】 。**

（3）如果主谓词组为情态动词加动词的结构，则将V2 标注为情态动词。

例8：**要 重 视 理 论 队 的 建 设 ， 【【使_V1 确 有 成 就 的 青 年 理 论 人_JY 才_JY 能_V2 脱 颖 而 出】】 。**

（4）如果存在一个动词作为另一动词的“方式”的句子，则将V2标注为兼语之后的第一个动词。

例9：**【【让_V1 乡_JY 亲_JY 们_JY 集_V2 中_V2 到 一 个 碾 子 上 碾 米】】 。**

说明：AMR标注体系会将“碾”作为“乡亲们”的谓语，而将“集中到一个碾子上”作为“碾米”的方式，为了与前面的标注标准一致，故将V2 标注为兼语后的第一个动词“集中”。

（5）如果兼语结构中存在主谓词组后有补语的情况，则将V2标注为兼语后的第一个动词。

例10：**我 们 也 尽 可 能 【【让_V1 她_JY 过_V2 得 充 实 如 意】】 。**

（6）如果兼语结构中含有复句，对于并列以及递进等没有主次关系的复句，将V2标注为复句第一部分的谓词，对于其他带有主次关系的复句，将V2标注为主要子句中的谓词。

例11：**老 师 【【让_V1 她_JY 一 边 听_V2 语 音 一 边 记 笔 记】】 。**

例12：**干 吗 【【让_V1 人_JY 家_JY 一 进 门 就 赶_V2 上_V2 一 顿 熊】】 呢**？

3.3 兼语语料库的统计分析

本文选取了来自文学、新闻、微博等不同领域的67419 个句子作为语料构建的原始语料，从中筛选得到了4760 个兼语句以及5248个兼语结构，并按照本文设计的兼语结构标注规范完成了兼语语料库的构建。我们对兼语结构中V1出现的频率进行统计，其中出现频率最高的六个词如图2所示。根据图2可以看出兼语结构中的兼语动词多集中在“让”、“使”、“令”、“请”、“叫”、“要求”等词，这六个词构成的兼语结构占所有兼语结构的70.8%。

本文对低频兼语动词也进行了统计，其中出现频次低于5次的兼语动词数量如表1所示，根据表1可以发现兼语语料库中包含大量低频兼语动词，其中出现频率为1次的有128个，出现频率为2次的有51个。低频动词多为高频动词的近义词，使用规则以及统计的方法难以识别此类动词，低频兼语动词的大量存在使得兼语结构识别工作十分困难，因此有效处理低频兼语动词对兼语结构的识别具有重要意义。

4 兼语结构识别研究

基于构建的兼语语料库，我们使用神经网络模型自动识别兼语结构的边界，辅助中文AMR语料的构建及解析。由于兼语结构的语义关系复杂，句式变化丰富，因此兼语结构的识别任务具有一定的挑战性。

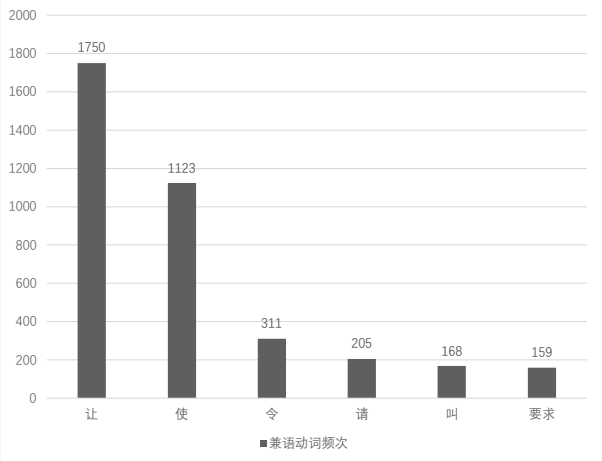


Figure 2: 兼语动词频次图

频次	数量	示例
1	128	胁迫、恳请、指派、催促
2	51	吁请、责令、诚邀
3	18	打发、煽动
4	8	任命、扶持
5	10	督促、选派、提请

Table 1: 低频兼语动词表

4.1 任务定义以及数据划分

我们将兼语结构的识别任务建模为序列化标注任务。给定输入的句子序列$X = \{x_1, x_2, ..., x_n\}$，模型需要预测出对应输入句子序列的标签序列$Y = \{y_1, y_2, ..., y_n\}$,其中$y_i \in \{B, M, E, O\}$。B标签对应兼语结构的起始字，E 标签对应兼语结构的结尾字，M标签对应兼语结构除以上成分的其他字，O标签对应句子的非兼语结构，句子对应标签序列示例如表2所示。我们将标注好的语料导出为序列化标注格式的文件，并随机打乱顺序，选取其中的10%作为测试集，然后从剩余的语料中选取90% 作为训练集，10% 作为开发集。

文本	林	老	师	让	大	家	补	选	一	名	劳	动	委	员	。
标签	O	O	O	B	M	M	M	M	M	M	M	M	M	E	O

Table 2: 兼语句标注示例表

4.2 模型

兼语结构中通常包含两个动词，因此使用规则或者机器学习的方法对兼语结构进行识别时，通常将词性作为重要的特征进行统计分析。然而现存的语料大部分没有分词以及词性标注，使用自动分词以及词性标注工具处理语料容易造成错误传播。单独使用字符信息对兼语结构进行识别容易丢失词语本身携带的信息，因此我们对字向量及其对应的词典信息进行拼接，获得句子完整的向量表示。将该向量传入表示层，获得包含上下文信息的句子表示。常用的表示层模型有BiLSTM、CNN、Transformer等，本文选用BiLSTM模型作为表示层获取句子的上下文信息。兼语结构的标签具有很强的依赖性，比如M标签必须在B之后，而不能在O之后，如果对每个标签进行独立决策，则无法考虑其间的依赖性。因此我们在BiLSTM模型之后拼接了CRF(李航, 2012) 模型，实现对含有约束关系的序列标签解码。最终构成的LexcionAugemented-BiLSTM-CRF(Peng et al., 2020)模型(LA-BiLSTM-CRF) 可以完成纯文本的兼语结构边界识别任务。

4.2.1 LA-BiLSTM-CRF模型

使用添加词典信息的字向量表示句子，既有效运用了文本中包含的词语信息，又避免了分词工具带来的错误传播。本文模型的输入为不包含任何分词以及词性信息的文本内容，然后使用公式(1)(2)(3)(4)获取每个字对应的向量表示。

$$x_i = [x_i^c; e^s(B, M, E, S)] \tag{1}$$

$$e^s(B, M, E, S) = [v^s(B) \bigoplus v^s(M) \bigoplus v^s(E) \bigoplus v^s(S)] \tag{2}$$

$$v^s(S) = \frac{1}{Z} \sum_{w \in S} Z(w + c) e^w(w) \tag{3}$$

$$Z = \sum_{w \in B \bigcup M \bigcup E \bigcup S} Z(w) + c \tag{4}$$

其中x_i表示当前句子中第i个字的向量表示，该向量由字向量与词语向量拼接构成，字向量通过查找字表获得对应的向量表示。$e^s(B, M, E, S)$表示包含当前字的所有词向量信息，$v^s(B)$表示以当前字为开始的所有词的向量表示，$v^s(M)$表示当前字在词的中间组成部分的所有词向量表示，$v^s(E)$表示以当前字为结尾的所有词的向量表示，$v^s(S)$表示当前字独立构成的词的向量表示。S表示某一种词集合，Z表示词语出现的频率，$e^w(w)$表示w的词向量。将词语的频率作为该词的权重，对集合内的所有词向量进行加权求和得到词集合的向量表示。

根据上述公式获得句子的向量表示$(x_1, x_2, \ldots, x_n)$，其中，n表示句子包含词的数量。将句子的向量表示传入BiLSTM模型(Yang et al., 2018)，获取包含上下文信息的句子表示$(h_1, h_2, \ldots, h_n)$。然后将其传入CRF 模型中，对于一个预测序列$y = (y_1, y_2, \ldots, y_n)$，将该序列的得分定义为公式(5)所示。

$$s(X, y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \tag{5}$$

其中，P定义为BiLSTM网络输出的分数矩阵，是一个规模为n×k的矩阵，其中k是标签的数量，P_{ij}对应的是在一个句子中第i个词语对应第j个标签的得分。A是一个转移得分矩阵，A_{ij}表示从第i个标签转移到第j个标签的得分。y_0和y_n是句子标签的开始和结束标志，我们将开始和结束标志也加入到候选标签集合，所以A是一个$(k + 2) \times (k + 2)$维的矩阵。

通过softmax函数对所有可能的标记序列进行概率计算，使用公式(6) 计算序列y的概率。在训练过程中，使用交叉熵损失函数来更正标签序列的预测，具体如公式(7)所示。

$$P(y|X) = \frac{e^{s(X,y)}}{\sum_{\tilde{y} \in Y_X} e^{s(X,\tilde{y})}} \tag{6}$$

$$\log(P(y|X)) = s(X, y) - \log(\sum_{\tilde{y} \in Y_X} e^{s(X,\tilde{y})}) \tag{7}$$

其中，Y_X代表对应于句子X的所有可能的标签序列集合。使用Viterbi 算法解码，得到最优输出序列(Graves and Schmidhuber, 2005)。解码过程中，预测得到的输出序列的最大得分由公式(8)计算得到。整体的模型结构图如图3所示。

$$y^* = \arg\max_{\tilde{y} \in Y_X} s(X, \tilde{y}) \tag{8}$$

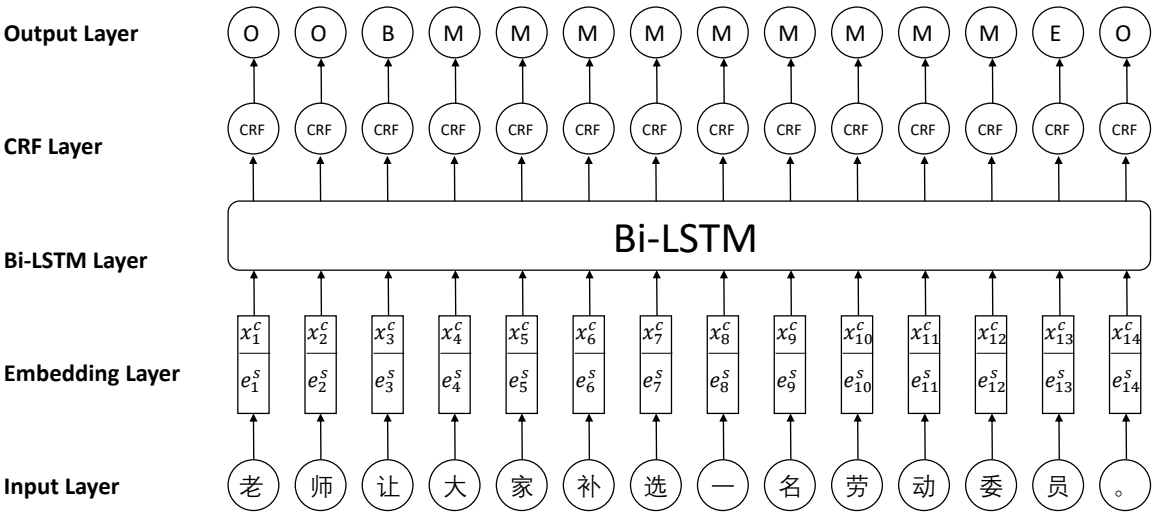


Figure 3: LA-BiLSTM-CRF模型结构图

5 兼语结构识别研究

实验中使用pytorch编写LA-BiLSTM-CRF模型。实验使用的语料是本文构建的面向中文AMR标注体系的兼语结构标注语料。

5.1 实验参数设置

LA-BiLSTM-CRF模型的参数设置如表3所示。本文的词向量使用预训练的CTB6.0(Xue et al., 2005)50 维词向量，字向量以及使用word2vec(Mikolov et al., 2013)训练的Giga-Word 50维字向量，通过BiLSTM模型获得隐藏层为300维的含有上下文信息的句子向量表示。然后将此向量表示传入CRF模型，获得对应输入序列的输出标签预测序列。使用Adam(Kingma et al., 2014)优化函数训练模型，学习率为0.0015，学习衰减率为0.05。对所有语料循环训练30 次，保留在开发集上预测结果最佳的模型，使用该模型对测试集数据进行预测。

超参数名称	参数值
epoch	30
Learning rate	0.0015
Learning rate decay	0.005
Word embedding size	50
Char embedding size	50
Hidden size	300

Table 3: 超参数设置表

5.2 实验结果及分析

由于没有使用神经网络模型识别兼语结构边界的相关研究工作，因此本文对神经网络模型的构建进行了探索。我们分别使用了CNN、Transformer以及BiLSTM(Lample et al., 2016)这三个基础模型作为表示层提取句子特征，其识别结果如表4 所示。

根据表4的实验结果可以发现，BiLSTM模型在兼语结构边界识别任务中表现最好，其F1值达到了86.06%。CNN模型的识别效果最差，其F1值为67.54%，CNN模型对于文本局部特征的捕捉能力较强，但是难以捕捉长兼语结构的特征，因此其识别效果较差。Transformer模型采用注意力机制提取文本特征，解决了文本的长距离依赖问题，其识别效果优于CNN模型。但Transformer模型难以捕捉句子中字词的位置方向信息，对于兼语结构中包含的连动以及宾

模型	P	R	F
LA-CNN-CRF	65.04	70.25	67.54
LA-Transformer-CRF	73.48	66.14	69.62
LA-BiLSTM-CRF	**86.25**	**85.91**	**86.06**

Table 4: 神经网络模型对比实验结果表

语从句这种与位置方向有关的结构学习能力较差。该模型通常只捕捉兼语结构中的一个V2及该V2对应的宾语，对于包含连动以及宾语从句的兼语结构的后边界识别效果较差。BiLSTM 模型既可以捕捉句子中较长的上下文信息，又不会丢失句子中字词的位置方向信息，其对于长兼语结构以及包含连动或宾语从句的兼语结构识别效果优于以上两个模型，在兼语结构边界识别任务中表现突出，其精确率、召回率以及F1值分别为86.25%、85.91% 和86.06%。实验结果证明，BiLSTM模型更适合兼语结构边界识别任务。

为了证明基于字符的神经网络模型以及词典信息的有效性，我们做了相关的消融实验，其实验结果如表5所示，其中BiLSTM-CRF_W是基于词和词性信息的神经网络模型，BiLSTM-CRF_C 是基于字符的神经网络模型，LA-BiLSTM-CRF是本文的添加词典信息的字符神经网络模型。

模型	P	R	F
BiLSTM-CRF_W	71.72	75.87	73.73
BiLSTM-CRF_C	85.52	84.34	84.93
LA-BiLSTM-CRF	**86.25**	**85.91**	**86.06**

Table 5: 消融实验结果表

根据表5可以发现BiLSTM-CRF_C模型的精确率、召回率以及F1值比BiLSTM-CRF_W模型分别高13.80%、8.47%和11.20%，这证明基于字符的神经网络模型缓解了分词以及词性标注的错误传播问题，有效提高了兼语结构边界识别任务的效果。但该模型丢失了句子中包含的词语信息，本文在此模型的基础上添加了词典信息，使用LA-BiLSTM-CRF 模型识别兼语结构边界，使得识别结果的精确率、召回率以及F1值分别提高了0.73%、1.57% 和1.13%，实验结果证明添加词典信息可以有效提高基于字符的神经网络模型对兼语结构边界识别的效果。

目前为止，兼语结构边界识别的研究工作较少，只有陈静(2012)采用基于特征模板的条件随机场模型对兼语结构边界进行了识别研究，因此我们使用陈静(2012)的模型以及特征模板对本文构建的语料进行识别，并将其结果与本文模型的结果进行对比，实验结果如表6所示。我们还对两个模型的所有标签识别效果进行了研究，具体结果如表7所示。

模型	P	R	F
CRF	**87.12**	82.24	84.61
LA-LSTM-CRF	86.25	**85.91**	**86.06**

Table 6: 对比实验结果表

根据表6的实验结果可以发现，LA-BiLSTM-CRF模型识别兼语结构边界的F1值比CRF模型提高了1.45%。CRF模型识别的精确率为87.12%，略高于LA-BiLSTM-CRF模型，而LA-BiLSTM-CRF模型识别的召回率为85.91%，比CRF模型的召回率高3.67%。就表7的各标签识别效果而言，两个模型对兼语结构前边界的识别效果最好，其F1值分别为93.55%和94.51%，后边界识别效果最差，其F1值分别为85.00% 和86.67%。CRF模型对前边界与后边界识别的精确率为96.32%和87.53%，分别高于本文模型2.00%和1.03%，但本文模型对前边界与后边界识别的召回率较高，比CRF模型分别高3.77%和4.21%，且F1值比CRF模型分别高0.96%和1.67%。实

标签	CRF			LA-BiLSTM-CRF		
	P	**R**	**F**	**P**	**R**	**F**
B	96.32	90.93	93.55	94.32	94.70	**94.51**
M	88.58	91.43	89.98	94.03	89.15	**91.52**
E	87.53	82.63	85.00	86.50	86.84	**86.67**

Table 7: 各标签识别结果表

验结果证明，CRF模型基于特征模板进行训练，识别结果较为精确，但其难以识别包含低频兼语动词以及兼语动词存在分词错误的兼语结构。LA-BiLSTM-CRF使用向量对句子进行表示，有效提高了兼语结构前边界识别的召回率以及F1值。兼语结构本身较为复杂，其内部常包含许多修饰成分，且前边界识别的错误直接影响后边界的识别效果，因此，后边界的识别效果较差。总体而言，本文模型对三个标签的识别效果都有不同程度的提升，并且有效提高了兼语结构边界识别任务的效果。

此外，我们还对模型的识别结果进行了错误分析。兼语结构前边界的识别错误主要发生在低频兼语动词中，大部分为语料中只出现一次的兼语动词。尽管使用字向量表示句子，缓解了低频兼语动词难以识别的问题，但是对于本身出现频率较低、“使令”义不强的兼语动词识别效果较差。比如“留她吃饭”中的兼语动词“留”在语料中只出现过一次，且其“使令”义较弱，因此模型未识别出该兼语结构。此外，模型会将部分高频兼语动词构成的非兼语结构错判为兼语结构，比如“使了个瞒天过海之计”中的“使”是高频兼语动词，模型将其误判为兼语结构。兼语动词的识别错误会导致错误传播，直接影响兼语后边界的识别效果。此外，兼语结构的后边界识别错误主要出现在包含定中结构或做定语的兼语结构。比如“让儿子买本养花的书参照执行”是包含定中结构的兼语结构，模型将兼语结构的后边界识别为“养花”。而在“战争是迫使敌人服从我们意志的一种暴力行为。”这一句子中，兼语结构作为定语修饰“暴力行为”，模型将该兼语结构的后边界识别为“行为”。由此可见，模型对于这两类兼语结构的后边界判别能力较差。

6 总结与展望

本文根据中文AMR标注体系的特点，制定了一套面向中文AMR标注体系的兼语结构标注规范，并利用此规范对收集的语料进行了兼语结构标注，得到4760句兼语句，5248个兼语结构，缓解了面向中文AMR标注体系的兼语语料库缺乏的问题。基于该兼语语料库，本文使用添加词典信息的字符神经网络模型识别兼语结构，避免了分词以及词性标注系统造成的错误传播，有效提高了兼语结构的识别效果，以期对今后的中文AMR语料构建及解析任务提供帮助，从而为语义解析及其下游任务奠定基础。

基于字符的神经网络模型缓解了低频兼语动词难以识别的问题，但低频兼语动词的存在仍然影响兼语结构前边界的识别效果。且模型对于包含定中结构或做定语的兼语结构识别效果较差。因此解决低频兼语动词的识别以及定中结构的边界判定是今后提高兼语结构识别的重点。此外，我们仍需要不断标注新的语料，使得模型学习到更多复杂的句子形式，提高模型处理复杂句子的能力。

参考文献

陈静, 王东波, 谢靖, 郑建明. 2012. 基于条件随机场的兼语结构自动识别. 情报科学,30(03): 439-443.

傅成宏. 2007. 现代汉语兼语结构的自动识别. 南京师范大学.

郭丽娟. 2019. 汉语依存句法分析树库构建与应用研究. 苏州大学.

胡裕树. 1962/1979. 现代汉语. 上海：上海教育出版社.

李斌, 闻媛, 宋丽, 卜丽君, 曲维光, 薛念文. 2017. 融合概念对齐信息的中文AMR语料库的构建. 中文信息学报,31(6): 93–102.

李航. 2012. 统计学习方法. 北京：清华大学出版社,pages 191-210.

李婷玉, 王亚, 曹聪. 2017. 兼语语义类的分类研究. 计算机应用研究,34(01):15-20.

马德全, 王利民. 2010. 兼语句的语义分析. 内蒙古民族大学学报（社会科学版）,36(04): 30-32.

曲维光, 周俊生, 吴晓东, 戴茹冰, 顾敏, 顾彦慧. 2017. 自然语言句子抽象语义表示AMR研究综述. 数据采集与处理,32(1): 26–36.

司玉英. 2010. 双宾兼语句的语法、语义和语用特征. 内蒙古大学学报（哲学社会科学版）,42(01): 148-152.

邢福义, 汪国胜. 2010. 现代汉语. 北京：高等教育出版社.

张志公. 1957. 修辞概要. 上海：上海新知识出版社.

周鸣. 2018(24). 浅谈兼语式定义问题. 汉字文化, pages 90–92.

周强. 2004(04). 汉语句法树库标注体系. 中文信息学报, pages 1-8.

Alex Graves and Jurgen Schmidhuber.2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*,18(5):602-610.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Design Challenges and Misconceptions in Neural Sequence Labeling.2016. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California : Association for Computational Linguistics*,pages 260-270.

Jason P. C. Chiu and Eric Nichols.2016(4). Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*,pages 357-370.

Jie Yang, Shuailong Liang and Yue Zhan.2018. Design Challenges and Misconceptions in Neural Sequence Labeling. *Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA: Association for Computational Linguistics*,pages 3879-3889.

Kingma, Diederik P and Ba, Jimmy.2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

Minlong Peng, Ruotian Ma, Qi Zhang, and Xuanjing Huang.2020. Simplify the Usage of Lexicon in Chinese NER. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistic*, pages 5951–5960.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer.2005. The Penn Chinese treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*,11(02): 207-238.

Pinherio, Ronan Collobert Pedro HO, and H. Pedro.2014. Recurrent Convolutional Neural Networks for Scene Parsing. *International Conference of Machine Learning.*

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean.2013. Distributed Representations of Words and Phrases and their Compositionality. *Proceedings of Advances in Neural Information Processing Systems 26*,pages 3111–3119.

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*,pages 1554-1564.