基于词信息嵌入的汉语构词结构识别研究

郑婳[1,2], 殷雅琦[1,2], 王悦[1,2], 代达劢[1,2], 刘扬[1,2]*
[1]北京大学计算语言学教育部重点实验室，北京100871
[2]北京大学计算机科学技术系，北京100871
{zhenghua,wyy209,daidamai,liuyang}@pku.edu.cn
yaqiYin@outlook.com

摘要

作为一种意合型语言，汉语中的构词结构刻画了构词成分之间的组合关系，是认知、理解词义的关键。在中文信息处理领域，此前的构词结构识别工作大多沿用句法层面的粗粒度标签，且主要基于上下文等词间信息建模，忽略了语素义、词义等词内信息对构词结构识别的作用。本文采用语言学视域下的构词结构标签体系，构建汉语构词结构及相关信息数据集，提出了一种基于Bi-LSTM和self-attention的模型，以此来探究词内、词间等多方面信息对构词结构识别的潜在影响和能达到的性能。实验取得了良好的预测效果，准确率77.87%，F1值78.36%；同时，对比测试揭示，词内的语素义信息对构词结构识别具有显著的贡献，而词间的上下文信息贡献较弱且带有较强的不稳定性。该预测方法与数据集，将为中文信息处理的多种任务，如语素和词结构分析、词义识别与生成、语言文字研究与词典编纂等提供新的观点和方案。

关键词： 汉语构词结构；词内信息；词间信息；语素；self-attention机制

Chinese Word-Formation Prediction based on Representations of Word-Related Features

ZHENG Hua[1,2], YIN Yaqi[1,2], WANG Yue[1,2], DAI Damai[1,2], LIU Yang[1,2]*
[1]Key Lab of Computational Linguistics (MOE), Peking University, Beijing 100871
[2]Department of Computer Science and Technology, Peking University, Beijing 100871
{zhenghua,wyy209,daidamai,liuyang}@pku.edu.cn
yaqiYin@outlook.com

Abstract

As a paratactic language, Chinese word-formations designate how the formation components combine to form words and become the key to understand semantics. In Chinese Natural Language Processing, most previous works on word-formation prediction follow the coarse-grained syntactic labels and use inter-word features, such as contexts, regardless of the inner-word features like morphemes and definitions. In this paper, we follow the word-formation labels defined from the linguistic perspective and first construct a formation-informed Chinese dataset. Then, we propose a Bi-LSTM-based model with self-attention to explore how the inner- and inter-word features influence the Chinese word-formation prediction task. Experimental results show that our method achieve high accuracy (77.87%) and F1 score (78.36%) on the word-formation task.

*通讯作者
基金项目：国家自然基金项目（62036001）、国家社科基金项目（16BYY137，18ZDA295）

Comparative analyses further show that morphemes (as an inner-word feature) greatly improve the prediction results, whereas the context (as an inter-word feature) performs the worst and shows strong instability. Our method and dataset would provide a new perspective for multiple Chinese information processing tasks, including linguistic analysis on morphemes and word-formations, sense prediction and generation, research on languages and lexicography, etc.

Keywords: Chinese word-formation , Inner-word features , Inter-word features , Morphemes , Self-attention

1 引言

汉语构词结构的研究由来已久，从《马氏文通》 (马建忠, 1898) 开始，涉及语法、词汇学的论著大都关注构词的话题，该问题对汉语语言学的重要性不言而喻。赵元任 (1980)、朱德熙 (1982) 等学者指出，词的结构是影响词义的一个重要因素。谭景春 (2000)、曹炜 (2001) 等深入分析了汉语词在结构组配过程中的意义贡献。苏宝荣 (2011) 进一步指出结构能够从句法、词法和新词义生成三个层面对语言产生影响。

面向中文信息处理的需求，杨梅 (2006) 给出了一套较为完善的构词结构标签，并证明了采用构词进行计算处理的可操作性和优越性。吉志薇和冯敏萱 (2015) 、田元贺和刘扬 (2016) 尝试利用语素信息和构词规则实现对未登录词的理解和语义预测。陈龙等 (2019) 则以语素概念和构词结构为基础，实现了对具有隐喻和转喻现象的汉语非字面义词的表示和理解。

认识到汉语构词结构在理论和应用上的重要性，信息处理领域的学者开始关注构词结构的自动识别，但是迄今为止开展的计算性工作依然较少：在已有的研究中，Li (2011) 以句法结构标签表示构词结构进行识别，Zhang等 (2013) 利用四种常见构词结构帮助识别复合词的主体部分，孙静等 (2014) 根据前缀与后缀结构构建计算模型。这类计算中大多沿用句法层面的粗粒度标签，缺乏相对明晰的语言学分类标准；此外，目前的构词结构识别主要利用词间信息 (Zheng et al., 2013; Gui et al., 2018; Wang et al., 2020)，忽略了语素义和词义等具有较强指示性的词内信息。

基于杨梅 (2006) 的构词研究成果和刘扬等 (2018) 的语言知识工程基础，我们构建汉语构词结构及相关信息数据集，首次采用语言学视域下的构词结构标签体系开展计算，提出了一种基于Bi-LSTM和self-attention的模型，以此来探究词内（词、字、词义、语素义）、词间（上下文）等多方面信息对构词结构识别的影响。该预测方法与数据集将为中文信息处理的多种任务，如语素和词结构分析、词义识别与生成、语言文字研究与词典编纂等提供新的观点和方案。

本文结构如下：在引言中，介绍了汉语构词结构识别的需求、现状和可能的发展；第2节对相关的理论问题、数据研发与计算方法作了梳理和评述；在第3节中，介绍了我们研发的汉语构词结构及相关信息数据集；第4节给出了一种基于多种词信息嵌入的汉语构词结构识别方法；在第5节中，我们阐述了实验结果，进行了详细的对比分析，并进一步探讨了模型的泛化能力；在结语部分，总结了本文工作以及未来可以深入展开的一些研究方向。

2 相关工作

2.1 汉语构词的研究与开发

对于汉语构词方式，语言学界目前有语法构词、语义构词等不同看法。语法构词的观点以偏正、主谓等语法结构对构词成分之间的关系进行分类。郭绍虞 (1979)、朱德熙 (1982) 等认为汉语句子的构造原则与词的构造原则基本一致。陆志伟 (1964)、赵元任 (1980)、王洪君 (2000) 等学者的研究，也支持复合词内部结构和句法结构类似这一观点。语义构词的观点则强调以主体、客体等语义标签分析构词成分 (张国宪, 1992; 朱彦, 2003)。刘叔新 (1990)、徐通锵 (1996) 等认为字与字之间是按语义关系构成字组。基于以上观点，考虑到计算的需求，傅爱平 (2003) 指出，虽然语义构词在表示词义时有天然优势，但其结构产生依据过于复杂，难以达成统一的标签集，因此不利于计算处理。而语法构词的结构体系简单，标准统一，且词法与句法结构有天然相似性，更适合计算处理。在语言知识工程方面，苑春法和黄昌宁 (1998) 利用语法结构标

签统计分析复合词的结构，构建语素知识库。刘扬等 (2018)、陈龙等 (2019) 依据这些前期研究，建立了以语素概念为基础语义单元、涵盖十余种构词结构的汉语概念词典。

除构词方式外，语言学界的另一个关注点是构词单位。学界普遍认为，语素是汉语中最小的音义结合体，也是构词的基本单位，能够对词相关信息的识别与研究起到关键作用 (尹斌庸, 1984)。徐枢 (1990) 对《现代汉语词典》中语素参与组词的数量进行了统计，结果表明语素在构词中非常活跃，处于重要的地位。苑春法和黄昌宁 (1998) 的统计结果显示，语素在构成名、动、形三类主要词汇后，语素义保持原本意义的比例均高于85.0%，说明了语素义研究对理解词义的必要性。另一方面，在信息处理中，语素对词的分析与表达提供了有效帮助。Qiu等 (2014) 利用语素嵌入增强词嵌入，为缺少上下文的新词提供表达，并在类比推理任务和词相似度任务中证明了语素嵌入的优势。Cao和Rei (2016) 将语素及其词内权重纳入词嵌入的生成过程，展现了语素信息对新词理解的优势。Lin和Liu (2019) 建立基于构词分析的语素嵌入，在语义相似度等内部任务中相比传统方法取得显著提升。

2.2 汉语构词信息的计算与应用

目前的中文信息处理以利用及分析词间信息为主 (Zheng et al., 2013; Gui et al., 2018; Wang et al., 2020)，对词内信息的关注相对较少。以往的词内信息研究大体上分为三类：

第一类研究将对词的分析细化为对字的分析，进行字符级的研究。Zhao (2009) 用基于字依赖的表示代替词向量。Dong等 (2010) 先从字进行分析，再由字组词来代替传统分词模式。Zhang等 (2013) 在设计字符级结构树标签时考虑了主谓、动宾、联合、偏正四种结构，将基于词的依赖树扩展为基于字的结构。Zhang等 (2014) 利用前文的标注结果，整合词间句法依赖和词内依赖。Li等 (2018) 捆绑了字、词的词性标签及其依赖标签，将字符作为神经网络学习的基础单元，提出了字符级依赖解析器。字符级的研究是词内结构研究的热门方向，但在语言学的视域下，构词的基本单位为语素，而非字符。因此，忽略了语素的字符级研究，存在语义理解与计算上的局限性；

第二类对于词内结构的研究，关注介于字和词之间的联系，即子词的概念。对于提取子词，Sennrich等 (2015) 给出了双字节BPE编码算法，Schuster和Nakajima (2012) 则提出了WordPiece词切分算法，以概率而非频率提取新的子词。Kudo (2018) 的一元语言模型以最大化句子分词结果概率为目标，同时输出分词结果与各词概率。Yang等 (2018) 利用BPE算法获得中文子词列表，再使用Lattice-LSTM模型将子词嵌入与字符嵌入结合。Zhang等 (2019) 结合词嵌入与子词嵌入，获得子词增强嵌入，从而增强文本理解任务的结果。Gong等 (2020) 建立字、子词、词的树状结构表示，组合成HiLSTM模型，应用于命名实体识别任务。子词的研究在近两年得到了研究者的关注，介于字与词之间的粒度让其应用更加灵活。但子词在语言学上没有确切的对应概念，这类方法更偏向统计学计算，而非基于语言本体的研究；

第三类研究则将词结构分析作为独立的自然语言处理任务。方艳和周国栋 (2015) 定义了词结构分析任务，并提出了基于层叠CRF模型的词结构分析方法，即在传统分词方法后，利用层叠CRF识别词的内部结构。在后续的研究中，孙静等 (2014) 提出了基于词缀的词结构分析模型，考虑了前缀式与后缀式这两种构词结构。蒋万伟和刘娟 (2017) 在此基础上针对伪未登录词的特点，设计了一般化的特征集，试图识别构词层次结构。但这类研究并未提供语言学视域下的细粒度构词结构标签，而更多地关注词内切分的位置与层次。

3 汉语构词结构及相关信息数据集

在汉语构词结构识别中，我们把构词结构的影响因素分为两大类：词内信息与词间信息。

3.1 汉语的词内信息

汉语的词内信息包括词、构词结构、字、语素义与词义。其中，词指的是词型(word type)，字指的是构成词的字型，语素义指的是构成词的语素的释义，词义指的是词的释义。

考虑到词典的权威性，同时为了保证数据的覆盖度与细粒度，我们从《现代汉语词典（第五版）》（以下简称《现汉》）中收集数据。包括《现汉》中全部45,311个有释义和例句的汉语二字词（双音节词）词条，其中有8,684个多义词。我们把不同的义项视为不同的词条，并给了每个词条唯一的ID。以“题字$_1$”为例，其ID为“52061-01-01”，依次代表“该词的ID-该词在词典中的第几次条目出现-当前是该词的第几个义项”。

对于汉语构词结构的划分，从语言学的视角出发，杨梅 (2006) 给出了18种构词结构；在此基础上，为了中文信息处理的应用需求，刘扬等 (2018)、陈龙等 (2019) 提出并标注了16种构词结构。根据现有的前期工作，我们整理了一个包含构词结构及其相关信息的数据集，在辅助构词结构预测任务的同时，也为下游任务提供数据资源，具体的构词结构解释和使用实例如表1中所示，即：定中、联合、述宾、状中、单纯、连谓、后缀、述补、主谓、重叠、方位、介宾、名量、数量、前缀与复量。注意到，一些多义词的不同义项在构词结构上存在着差异，如表2中列举的“题字”一词，当表示“为留纪念而写上字”时，构词结构为述宾，而表示“为留纪念而写上的字”时，构词结构为定中。

构词结构	构词结构描述	用例	%
定中	后语素是体词性对象，前语素修饰后语素	长子	38.62
联合	前后语素地位平等，同义、反义或相互补充	长短	22.87
述宾	前语素是动作，后语素是前语素的支配对象	助长	16.44
状中	后语素是动作或性质、状态，前语素修饰后语素	疯长	8.45
单纯	词本身是独立的语素	沙发	3.51
连谓	前后语素地位平等，是连续发生的动作	生长	3.43
后缀	前语素是实词，后语素是词缀	鸭子	2.70
述补	前语素是动作行为，后语素是动作行为的结果或趋向	延长	1.28
主谓	前语素是动作主体或被说明对象，后语素是动作或说明	年轻	1.06
重叠	前语素与后语素完全一样	爸爸	0.59
方位	前语素是实词，后语素是方位词，形成参照-方位关系	眼前	0.37
介宾	前语素是介词，后语素是介词宾语	从前	0.31
名量	前语素是名词，后语素是与之匹配的量词	叶片	0.13
数量	前语素是数词，后语素是量词	一天	0.10
前缀	前语素是词缀，后语素是实词	老虎	0.10
复量	前后语素都是量词	人次	0.03

Table 1: 构词结构与用例（% 表示该类型所占的百分比）

词	词义	构词结构	例句
题字$_1$ (52061-01-01)	为留纪念而写上字	述宾	主人拿出纪念册请来宾～
题字$_2$ (52061-01-02)	为留纪念而写上的字	定中	书上有作者的亲笔～

Table 2: “题字”的两个义项及释义例句

为了区分字的不同使用及意义，即语素的情况，接下来需要对构词结构下的语素成分进行义项标注。我们从《现汉》中收集了10,527个汉字和20,855个语素释义，并赋予每个语素释义唯一的ID。表3展示了“长”字的不同语素义及其ID编码，其中“长$_1$”的释义为“两点之间的距离大”，其ID为“长1-06-01”，依次代表“该字在词典中的第几次条目出现-该条目共有几个语素义-当前是该条目的第几个语素义”。

语素	语素义
长$_1$ (长1-06-01)	两点之间的距离大
长$_2$ (长2-04-02)	排行最大
长$_3$ (长3-03-02)	生长

Table 3: “长”字的三个语素及定义示例

在此基础上，我们对每个词条的构词结构与语素义进行了标注。标注人员包括中文系两位教授与六名研究生，他们根据词条释义为每一个词条标注构词结构并绑定对应的语素义ID（如

表4所示）。每个词条由三位标注人员独立标注并交叉验证，每位标注人员在标注的同时也会给出一个置信度。如果三位标注人员的标注结果完全相同，则直接收入数据集，如果三位标注人员的标注结果不完全相同，则由另一位标注人员进行审阅，依据之前三位标注人员的标注与置信度决定最终标注并收入数据集。在全部45,311个词条中，81.92%的词条三位标注人员的标注完全相同，90.86%的词条至少两位标注人员的标注完全相同。

词	长子(50037-01-01)
构词结构	定中
前语素义	长$_2$ (长2-04-02)：排行最大
后语素义	子$_1$ (子1-13-01)：古代指儿女，现在专指儿子
词义	排行最大的儿子

Table 4: 语义构词知识示例

3.2 汉语的词间信息

此外，影响汉语构词结构的词间信息主要是目标词的上下文。在前文中提到，不同义项的多义词可能会表现为不同的构词结构，这也有可能体现在上下文的差异中。

《现汉》中的例句和义项是彼此对应的，如表2所示，对于“题字”的两个义项，《现汉》中均给出了对应的释义与例句。我们收集了《现汉》中所有二字词的例句，作为数据集中的上下文信息。

综上所述，我们最终构建的汉语构词结构及相关信息数据集包含了词、构词结构、字、语素义、词义与上下文，如表5中呈现的例子所示：

词	平地$_1$ (18287-01-01)	平地$_2$ (18287-01-02)
构词结构	述宾	定中
字	平；地	平；地
语素义	平：使平；地：土地，田地	平：表面没有高低凹凸，不倾斜；地：陆地
词义	把土地整平	平坦的土地
上下文	播种前要翻地、～	找一块～修操场

Table 5: 构词相关信息示例

4 结合词内和词间信息的构词结构识别方法

4.1 任务描述

本文中的构词结构预测属于多分类任务，输入一个目标词w_*及其词内和词间信息，输出该目标词的构词结构类别。其目标函数如下：

$$p(m|w_*, Ch, Morph, Def, Con) = f(w_*, Ch, Morph, Def, Con), \quad (1)$$

$$\hat{m} = \arg\max_m p(m|w_*, Ch, Morph, Def, Con), \quad (2)$$

其中，m表示预测的构词结构，w_*为目标词，$Ch = \{ch_1, ch_2\}$为目标词中的字，$Morph = \{morph_1, morph_2\}$为目标词中的语素义，$Def$为目标词的词义，$Con$为目标词的上下文，$f(\cdot)$为构词结构识别的分类器。

4.2 基于Bi-LSTM的构词结构识别

为了探究词内和词间信息对汉语构词结构识别的影响，我们提出了用双向长短期记忆网络（Bidirectional Long-Short Term Memory, Bi-LSTM）(Graves et al., 2005) 网络模型来进行预测。模型架构包含四个部分：1）信息输入层；2）信息编码层，用来编码输入的词内和词间信

息；3）信息交互层，用来融合编码信息；4）输出层，根据编码的信息来进行分类，输出预测的构词结构。

4.2.1 信息输入和编码层

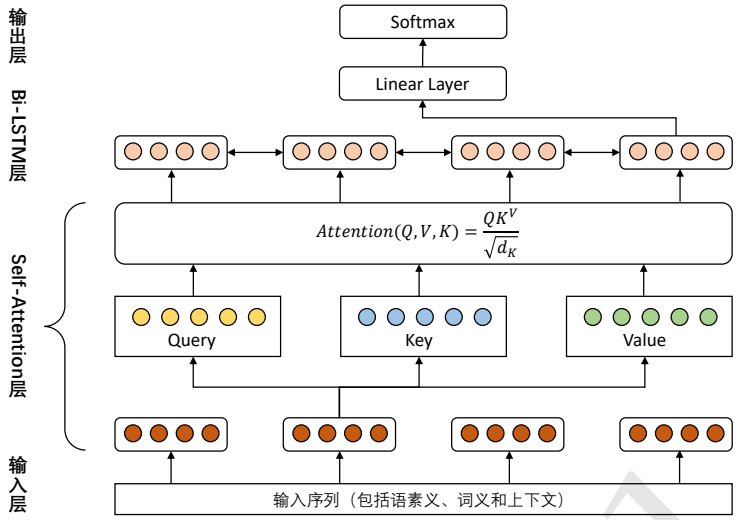

Figure 1: 模型结构图

信息编码层的架构如图1所示，在信息编码层，我们首先对五种输入的信息进行编码，分别是目标词、字、语素义、词义和上下文。对于目标词W_*和词中的字$Ch=\{ch_1,ch_2\}$，我们采用预训练的词和字向量来进行编码，其中，整体的字向量$\mathbf{ch}_*$由两个字向量$[\mathbf{ch}_1;\mathbf{ch}_2]$拼接得到，作为初始输入。

词内信息中的语素义$Morph=\{morph_1,morph_2\}$、词义$Def$和词间信息的上下文$Con$属于长序列输入。为了更加有效地捕捉到长距离信息，我们利用Bi-LSTM来分别对它们进行编码，以获得更丰富的语义信息。LSTM模型输入向量矩阵，利用遗忘门f_t、记忆门i_t和输出门o_t对隐层状态$hidden_t$和细胞状态$cell_t$进行更新，经过下列步骤来获得隐层向量的表示：

$$f_t=\sigma(W_f[h_{t-1},x_t]+b_f), \tag{3}$$
$$i_t=\sigma(W_i[h_{t-1},x_t]+b_i), \tag{4}$$
$$o_t=\sigma(W_o[h_{t-1},x_t]+b_o), \tag{5}$$
$$\hat{cell}_t=tanh(W_c[h_{t-1},x_t]+b_c), \tag{6}$$
$$cell_t=f_t\odot cell_{t-1}+i_t\odot cell_t, \tag{7}$$
$$hidden_t=o_t\odot tanh(cell_t), \tag{8}$$

其中，σ表示sigmoid函数，$\odot$表示哈达玛积。Bi-LSTM由前向LSTM和后向LSTM组合而成，并通过拼接前后隐层向量$h_t=[\overrightarrow{h_t};\overleftarrow{h_t}]$来更好地捕捉双向的语义依赖。同时，受前人工作的启发，我们同时引入self-attention机制 (Bahdanau et al., 2014) 来增强词表示。在self-attention中，首先对于每个单词分别创建Query向量（Q）、Key向量（K）和Value向量（V），然后对Q和K执行点积相乘并进行缩放操作，再对其执行softmax操作进行归一化，得到词之间的attention权重，最后利用attention权重对每一个V及进行加权求和，得到attention增强后的词表示，计算公式如下：

$$Attention(Q,V,K)=\text{softmax}(\frac{QK^T}{\sqrt{d_K}})V, \tag{9}$$

其中，softmax表示softmax函数，$\sqrt{d_K}$表示K的维度，用于缩放保持梯度稳定。

通过对语素义$Morph$、词义Def和上下文Con进行self-attention后得到语素义编码，利用Bi-LSTM进行编码得到输入，其公式如下：

$$\mathbf{mor}_i = \text{Bi-LSTM}(\text{Self-Attention}([morph_i])), \tag{10}$$
$$\mathbf{mor} = W_{mor}([mor_1; mor_2]) + b_{mor}, \tag{11}$$
$$\mathbf{con} = \text{Bi-LSTM}(\text{Self-Attention}(Con)), \tag{12}$$
$$\mathbf{def} = \text{Bi-LSTM}(\text{Self-Attention}(Def)), \tag{13}$$

其中的；表示向量拼接。最终得到目标词$\mathbf{w}_*$、字$\mathbf{ch}_*$、语素义$\mathbf{mor}$、上下文$\mathbf{con}$和词义$\mathbf{def}$，共五种编码后的词内词间信息，进入信息交互和输出层。

4.2.2 信息交互和输出层

在信息交互层，我们使用线性层来融合信息编码层中获得的特征，最后通过softmax层计算每种构词结构的概率分布，并输出识别概率最高的构词结构。计算公式如下：

$$\mathbf{k} = W_k[\mathbf{w}_*, \mathbf{ch}_*, \mathbf{mor}, \mathbf{con}, \mathbf{def}], \tag{14}$$
$$\alpha = \text{softmax}(\mathbf{k}), \tag{15}$$

其中，softmax表示softmax函数，$\mathbf{k}$表示五种词内和词间信息通过线性层信息融合的结果，α表示计算得到的构词结构概率。

5 实验结果与分析

5.1 实验设置

5.1.1 实验数据

我们采用第3节中的数据集，将其按照8:1:1的比例分为训练集、验证集与测试集，它们的统计信息如表6所示。对于多义词，我们视为不同的词条，保证每个多义词仅出现在一个集子里。

数据集	#词形	#词义	上下文句长	语素义$_1$句长	语素义$_2$句长	词义句长
训练集	29,169	36,248	7.22	7.69	7.29	12.02
验证集	3,673	4,531	7.32	7.45	7.30	11.91
测试集	3,666	4,532	7.26	7.51	7.01	12.03

Table 6: 数据集统计信息（语素义$_i$表示第i个语素的释义，长度按句子的平均汉字数计算）

5.1.2 评价指标

构词结构预测是一种多分类任务，本文使用准确率和F1值作为评价指标。其中，用TP表示预测正确的正例数，TN表示预测错误的正例数，FP表示预测正确的负例数，FN表示预测错误的负例数，准确率的计算公式为：

$$\text{准确率} = \frac{TP + TN}{TP + FP + TN + FN} \tag{16}$$

F1值的计算公式为：

$$\text{精确率P} = \frac{TP}{TP + FP} \tag{17}$$
$$\text{召回率R} = \frac{TP}{TP + FN} \tag{18}$$
$$F1 = \frac{2PR}{P + R} \tag{19}$$

5.1.3 参数设置

本文使用fastText (Bojanowski et al., 2017)在中文维基百科上预训练的词向量对词进行初始化，词向量维度为300，Bi-LSTM隐藏层的维度为300。超参的最优值通过验证集的结果获得，训练的批次大小为128。使用的优化器是Adam，学习率设置为10^{-3}。

5.2 实验结果与分析

注入特征	准确率(%)		F1 (%)	
	验证集	测试集	验证集	测试集
多数基准模型	5.61	5.27	7.55	8.32
随机基准模型	39.04	38.62	42.22	25.07
W	60.61	61.86	62.93	64.23
Ch	67.86	66.51	69.35	68.07
Def	60.91	60.13	63.90	63.91
Morph	76.51	75.19	76.98	76.56
Con	48.96	47.69	56.60	54.60
W+Ch+Def+Morph	**78.90**	**77.87**	**79.26**	**78.36**
W+Ch+Def+Morph+Con	76.51	75.19	76.98	76.56

Table 7: 构词结构预测结果（**粗体**表示最佳结果）

我们首次采用语言学视域下的构词结构标签体系进行预测，并重复进行三次实验取输出结果的平均值。在验证集和测试集上的指标如表7所示，测试集上的分类情况如图2的混淆矩阵所示。根据表中数据，我们观察得到如下结论：

1. 五种词信息（包括词内、词间信息）都能在一定程度上捕捉构词结构知识，其准确率和F1值远超随机基准模型。最佳模型（W+Ch+Def+Morph）取得了良好的构词结构识别效果，准确率77.87%，F1值78.36%，证明了自动构词结构识别任务的可行性；
2. 在词内和词间信息中，对构词结构识别效果提升最为明显的是语素信息（Morph），其次是字（Ch）信息，表现最弱的是上下文信息（Con）。其中，相较于字信息，语素信息在准确率和F1指标上分别有13.05%和12.47%的提升，证明了语素信息能最有效地捕捉到词内部的构词结构知识。我们认为上下文信息表现最弱的原因在于其主要包含了词与词之间的组合关系，而相对难体现词内部状况，因此不容易准确预测构词结构；
3. 把使用全部词内信息（W+Ch+Def+Morph）、使用全部词间信息（Con）和使用所有词信息（W+Ch+Def+Morph+Con）的三种模型作比较，结果显示，仅用词内信息（W+Ch+Def+Morph）就能达到构词结构预测的最佳效果。和使用所有词信息（W+Ch+Def+Morph+Con）相比，使用词内信息（W+Ch+Def+Morph）在准确率和F1指标上分别有3.56%和2.35%的效果提升。这不仅证明了第2点结论，即上下文信息难以准确识别构词结构，而且表明了上下文会带来额外噪声。

我们根据测试集上的最佳结果制作混淆矩阵，颜色越深代表分类的准确率越高，如图2所示。由于不同构词结构下的词条的数量差异较大，我们对结果进行归一。根据图中趋势可得：

1. 对于定中、述宾、联合、述补、状中、介宾、后缀、主谓和方位这九类构词结构，模型的预测准确率较高。“名量”结构的预测准确率最低，可能是由于该结构下的词条数量最少，在训练时难以有效捕捉到该构词结构的特点，因此预测效果较差。“单纯”结构的预测准确率次低，可能是该构词结构代表“词是独立的语素”（如表1所述），因此模型同样无法有效地捕捉到词的内部结构；
2. 我们注意到，“连谓”和“重叠”结构经常被错误预测为“联合”结构，这可能是因为“连谓”、“重叠”和“联合”这三种构词结构在语言学上有很强的关联和相似性，都隐含有“前后语素地位平等”的意思，而其中“联合”结构的词条在训练数据中占比最高，因此“连

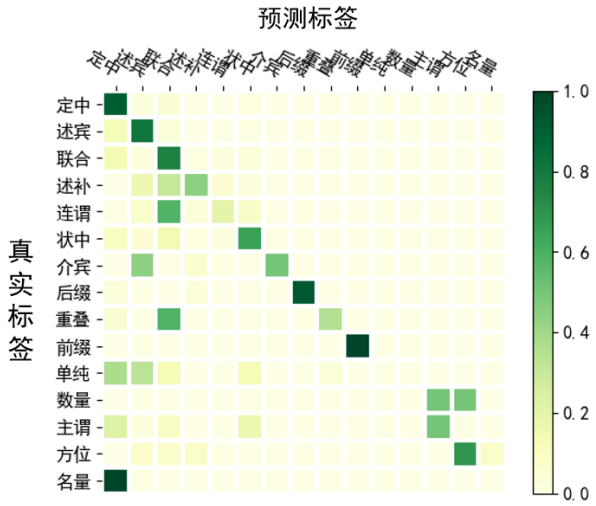

Figure 2: 构词结构预测结果混淆矩阵

谓”和“重叠”结构容易被错误预测为“联合”结构。这一现象符合语言学预期，也从侧面表明我们的方法能有效捕捉到构词结构的隐含特点。

根据第2节前人工作的经验，以上下文为代表的词间信息能有效辅助词义消歧、词义生成、词义识别等常见语义任务。然而，对于语言学视域下的构词结构识别任务，上述的实验结论表明上下文的贡献较小。这种情况说明，语义构词识别任务和其它常见语义任务在性质和特征体现方面有不同的状况和趋向。

为了进一步探究上下文对于构词结构识别的有效性，我们额外进行了针对上下文的稳定性实验。在实际下游任务应用中，可能存在上下文的信息量有限、质量难以保障的情况，因此我们设计了上下文替换模板，将训练集中的上下文替换成低信息量、低质量的句子。我们使用`jieba`库对上下文中的目标词标注词性，库中包含名词、形容词、动词、数词、方位词等28种词性，并针对每种词性设计了不同的替换模板。以部分词性为例的上下文替换模板如表8所示。

词性	替换模板	举例（替换前→替换后）
名词	这是$[w_*]$	树上的苹果成熟了→ 这是苹果
动词	我们$[w_*]$	夏天，他最喜欢去海边游泳 → 我们游泳
形容词	这很$[w_*]$	山河是多么庄严且美丽 → 这很美丽
数词	这有$[w_*]$个	这个节目吸引了数百万电视观众→ 这有百万个
方位词	这在$[w_*]$	旭日染红了东边的天空→ 这在东边

Table 8: 以部分词性为例的上下文替换模板（其中$[w_*]$和举例中下划线的部分表示目标词）

实验结果显示，利用模板替换后，仅用上下文的汉语构词结构识别在测试集上的准确率为43.62%，F1值为51.38%，相较替换之前分别降低了4.07%和3.22%；用所有词内和词间信息的汉语构词结构识别在测试集上的准确率为71.39%，F1值为73.20%，相较替换之前分别降低了3.80%和3.36%。上述结果表明，虽然上下文能够提供一定的句法、词义信息并辅助汉语构词结构识别，但是其有效性严重依赖于上下文的信息量和质量，而这些在实际下游任务应用中无法保障。因此，对于构词结构识别任务，上下文具有较强的不稳定性，且容易带来额外噪音。

5.3 关于模型泛化能力的讨论

为验证本方法的泛化能力，我们进一步在新词上展开实验。

新词的特殊性在于其催生出了新的词型或义项，也可能衍生出了新的语素义，这些给构词结构识别带来了挑战。为了评估本文方法在新词构词结构识别上的效果，我们构建了一个小规

模的新词数据集。其中，新词及词义来源于中文维基百科[0]。我们筛选了维基百科标签或释义中带有“新词”或“流行语”且未收入《现汉》的词条，最后选取了覆盖不同领域的100个词条。此外，考虑这里面缺少了“名量”等结构的样例，为了保证数据在构词结构上的分布一致，我们从王钧熙 (2011) 的《汉语新词词典：2005-2010》中挑选了特定结构的部分词条，也加入到数据集中去，共计得到108个新词。新词的上下文提取自微博[1]，并经过人工筛选以保证新词在上下文中的语义与释义一致。同时，我们对每个新词的构词结构进行了人工标注。

最终，数据集中的每个词条包含：1）新词，2）构词结构，3）新词释义，4）语素义，5）上下文。这些新词的来源覆盖了科技、经济、政治、生活、艺术、体育等多个领域。在表9中，给出了一个新词的示例，其中“菜”的语素义标注为“（空）”，这是因为目前的《现汉》中缺乏针对此类新衍生出的语素义的定义。

词	菜鸟
构词结构	定中
字	菜；鸟
语素义	菜：（空）；鸟：脊椎动物的一大类，体温恒定，卵生
词义	泛指对某个领域缺乏基本知识的人
上下文	无论你是职场～还是江湖老手，都来跟打工界的最强王者们取取经吧！

Table 9: 新词及构词相关信息示例

实验结果显示，使用词、字、语素义、词义和上下文信息的方法（W+Ch+Def+Morph+Con）在新词测试集上的准确率为68.89%，F1值为67.93%。考虑到上下文信息可能带来噪音，去除上下文后，在新词测试集上的准确率上升到69.92%，F1值上升到68.78%。这两个实验结果，远高于随机基准模型的效果，且符合主实验中以往汉语词汇的表现趋势，这说明本文方法可以进一步衍生到新词的构词结构识别中去。

对比主实验中以往汉语词汇上的最佳结果（见表7），新词数据集上的结果分别降低了10.21%（准确率）和12.23%（F1值）。我们猜想，导致这一现象的原因主要有两方面：1）大部分新词存在隐喻、转喻等非字面义 (陈龙等, 2019)，例如，”社畜”表示“社会底层上班族”而非“社会的牲畜”，“巨婴”表示“心理不成熟的成年人”而非“巨大的婴儿”。这些非字面义削弱了词和词义之间的直接联系，从而减低了算法中词义信息表达的有效性；2）此外，受限于新词中语素义的新的衍生与发展，部分语素无法在《现汉》中找到对应的语素义。例如表9中的“菜”，表示“弱；差”的概念，“卖萌”中的“萌”，表示“可爱”的概念，但在目前的《现汉》中均没有对应的语素义。

这种情况表明，现有语素的语义空间划分存在缺憾，无法覆盖新词中可能衍生出的语素义。在构词结构识别之后，通过计算性手段，有可能推测出新衍生出的语素义，为汉语语言文字研究和词典编纂提供帮助。

6 结语

本文旨在探究基于词信息嵌入的汉语构词结构识别，我们采用语言学视域下的构词结构标签体系，构建汉语构词结构及相关信息数据集，提出了一种基于Bi-LSTM和self-attention的模型，以此来探究词内和词间等多种信息对构词结构识别的影响，其中，词内信息包括词、构词结构、字、语素义和词义，词间信息为上下文。实验取得了良好的预测效果，对比测试揭示，词内的语素义信息对构词结构识别具有显著的贡献，而词间的上下文信息贡献较弱且带有较强的不稳定性。同时，为了证明模型的泛化能力，我们进一步将模型推广到新词的构词结构识别，并取得了良好的效果。

在未来，该预测方法与数据集，将为中文信息处理的多种任务，如语素和词结构分析、词义识别与生成、语言文字研究与词典编纂等提供新的观点和方案。在后续的工作中，我们计划将构词结构识别融入中文信息处理的下游任务，以进一步提升应用系统的性能。

[0] https://dumps.wikimedia.org/zhwiki

[1] https://weibo.com

参考文献

Dzmitry Bahdanau, Kyunghyun Cho and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *Proceedings of TGE International Conference on Learning Representations*.

Kris Cao and Marek Rei. 2016. A joint model for word embedding and word morphology. *arXiv preprint arXiv*: 1606.02601.

Zhendong Dong, Qiang Dong and Changling Hao. 2010. Word segmentation needs change - From a linguist's view. *Proceedings of CIPS-SIGHAN Joint Conference onChinese Language Processing*.

Chen Gong, Zhenghua Li, Qingrong Xia, Wenliang Chen and Min Zhang. 2020. Hierarchical LSTM with char-subword-word tree-structure representation for Chinese named entity recognition. *Science China Information Sciences*, 63(10), 1-15.

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

Tao Gui, Qi Zhang, Jingjing Gong, Minlong Peng, Di Liang, Keyu Ding and Xuanjing Huang. 2018. Transferring from formal newswire domain with hypernet for twitter POS tagging. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv*: 1804.10959.

Haonan Li, Zhisong Zhang, Yuqi Ju and Hai Zhao. 2018. Neural character-level dependency parsing for Chinese. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Zhongguo Li. 2011. Parsing the internal structure of words: A new paradigm for chinese word segmentation. *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human language technologies*: 1405-1414.

Zi Lin and Yang Liu. 2019. Implanting Rational Knowledge into Distributed Representation at Morpheme Level. *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. No. 01. 2019.

Bojanowski Piotr, Grave Edouard, Joulin Armand and Mikolov Tomas. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135-146.

Siyu Qiu, Qing Cui, Jiang Bian, Bin Gao and Tieyan Liu. 2014. Co-learning of word representations and morpheme representations. *In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*: 141-150.

Rico Sennrich, Barry Haddow and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv*: 1508.07909.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Yinglin Wang, Ming Wang and Hamido Fujita. 2020. Word sense disambiguation: A comprehensive knowledge exploitation framework. *Knowledge-Based Systems* 190:105030.

Jie Yang, Yue Zhang and Shuailong Liang. 2018. Subword encoding in lattice lstm for chinese word segmentation. *arXiv preprint arXiv*: 1810.12594.

Meishan Zhang, Yue Zhang, Wangxiang Che and Ting Liu. 2013. Chinese parsing exploiting characters. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*: 125-134.

Meishan Zhang, Yue Zhang, Wangxiang Che and Ting Liu. 2014. Character-level chinese dependency parsing. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*: 1326-1336.

Zhuosheng Zhang, Hai Zhao, Kangwei Ling, Jiangtong Li, Zuchao Li, Shexia He and Guohong Fu. 2019. Effective subword segmentation for text comprehension. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*: 27(11), 1664-1674.

Hai Zhao. 2009. Character-level dependencies in chinese: Usefulness and learning. *roceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*: 879-887.

Xiaoqing Zheng, Hanyang Chen and Tianyu Xu. 2013. Deep learning for Chinese word segmentation and POS tagging. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

曹炜. 2001. 现代汉语词义学. 学林出版社, 上海.

陈龙, 饶琪, 刘扬. 2019. 汉语词的非字面义的表示与应用. 中国科学:信息科学: 49:1005-1018.

方艳, 周国栋. 2015. 基于层叠CRF模型的词结构分析. 中文信息学报, 29(04):1-7+24.

傅爱平. 2003. 汉语信息处理中单字的构词方式与合成词的识别和理解. 语言文字应用, (04):25-33.

郭绍虞. 1979. 汉语语法修辞新探. 商务印书馆, 北京.

蒋万伟, 刘娟. 2017. 基于条件随机场的词结构分析方法. 武汉大学学报(理学版), 63(03):251-258.

吉志薇, 冯敏萱. 2015. 面向普通未登录词理解的二字词语义构词研究. 中文信息学报, 29(05):63-68+83.

刘叔新. 1990. 汉语描写词汇学. 商务印书馆, 北京.

刘扬, 林子, 康司辰. 2018. 汉语的语素概念提取与语义构词分析. 中文信息学报, 32(02):12-21.

陆志韦. 1964. 汉语的构词法. 科学出版社, 北京.

马建忠. 1898. 马氏文通. 商务印书馆, 北京.

潘文国, 叶步青, 韩洋. 2004. 汉语的构词法研究. 华东师范大学出版社, 上海.

苏宝荣. 2011. 词（语素）义与结构义. 语文研究, 01:1-5.

孙静, 方艳, 丁彬, 周国栋. 2014. 利用扩展标记集的词结构分析. 中文信息学报, 28(05):39-45+82.

谭景春. 2000. 词的意义、结构的意义与词典释义. 中国语文, (01):69-78+94.

田元贺, 刘扬. 2016. 汉语未登录词的词义知识表示及语义预测. 中文信息学报, 30(6):26-34.

王洪君. 2000. 汉语语法的基本单位与研究策略. 语言教学与研究, 02:10-18.

王钧熙. 2011. 汉语新词词典：2005-2010. 学林出版社, 上海.

徐枢. 1990. 语素. 人民教育出版社, 北京.

徐通锵. 1997. 核心字和汉语的语义构辞法研究. 语文研究, 03:2-16.

杨梅. 2006. 现代汉语合成词构词研究. 南京大学博士论文, 南京.

尹斌庸. 1984. 汉语语素的定量研究. 中国语文, (05).

苑春法, 黄昌宁. 1998. 基于语素数据库的汉语语素及构词研究. 世界汉语教学, (02):8-13.

张国宪. 1992. 并列式合成词的语义构词原则与中国传统文化. 汉语学习, (05):28-31.

赵元任. 1980. 中国话的文法. 香港中文大学出版社, 香港.

朱德熙. 1982. 语法讲义. 商务印书馆, 北京.

朱彦. 1982. 汉语复合词语义构词法研究. 华东师范大学博士论文, 上海.