基于词向量的汉语复合词内部语义关系的量化研究

刘柱
清华大学人文学院中文系
liuzhu22@mails.tsinghua.edu.cn

摘要

复合构造是汉语词汇最常见的构词方式，它使得词汇内部的语义关系更为明显，也同时具有较强的可分析性。另一方面，计算语言学领域通过大规模语料学习词向量来作为词汇的语义表征。前人的研究大多关注于词向量如何对下游任务起作用，或者词汇之间的语义依赖，却较少关注词汇内部的语义关系，这对于以复合词构词为主、内部结构可分析的汉语来说，不失为一个重要缺失。本文探究了复合词向量可否如实反映两种语义关系：主导性和可组合性。前者表明复合词的哪一部分从语义上讲更加重要；后者体现了整体的词义多大程度上可以通过部分的意义推导出来。本文的研究发现通过词向量对于这两种关系的判断基本与语言学中的吻合。同时，通过对大规模词汇词向量的词义发掘，可以推断出主导性和可组合性受到多种因素影响，这些因素对于新词预测、语言教学、词典编撰等具有一定的实用参考价值。

关键词： 词向量；语义主导性；语义可组合性；汉语复合词

An Empirical Study on Semantic Relations within Chinese Compound Words based on Word Embeddings

Zhu Liu
School of Humanities, Tsinghua University
liuzhu22@mails.tsinghua.edu.cn

Abstract

Compounding construction is the most common word formation pattern in Chinese vocabulary. It makes the semantic relations within the vocabulary more apparent and also enhances its analyzability. On the other hand, in the field of computational linguistics, word vectors learned from large-scale corpora are used as semantic representations of vocabulary. Previous studies have mostly focused on how word vectors affect downstream tasks or semantic dependencies between words, but less attention has been given to the internal semantic relationships within individual words. This is more important especially for Chinese, which heavily relies on compound words with analyzable internal structures. This study investigates whether compound word vectors can accurately reflect two types of semantic relationships: lexeme meaning dominance and semantic compositionality. The former indicates which part of a compound word is semantically more important, while the latter reflects to what extent the overall meaning of a compound word can be recovered from the meanings of its parts. The findings of this study show that word vectors largely align with linguistic expectations in terms

of these two relationships. Additionally, by exploring the semantic properties of word vectors on a large-scale vocabulary, it is inferred that dominance and compositionality are influenced by various factors, which have practical implications for tasks such as new word prediction, language teaching, and dictionary compilation.

Keywords: word embeddings , dominance , compositionality , Chinese compound words

1 引言

通过两个或两个以上词根语素组成的复合构词方式[0]是汉语词汇最常见的构成路径，语言学界已经从方方面面对它的形成机制做了研究，包括中心理论（headness）(Packard, 2000)，构词和造词规则(葛本仪, 1985)，语素的自由度(董秀芳, 2004)等等。通过这种方式形成的复合词，其内部有一定程度的可分析性：词根之间、词根与整体词汇之间往往存在一定的结构和语义关系，例如表示“修饰-被修饰”的偏正结构。其中，语义主导性和组合性是两种突出的语义关系：语义主导性（lexeme meaning dominance）表示哪部分词根语素对于整体的语义更加重要，起到更加主导的作用。例如，“黑板”的右半部分语素“板”的语义更为重要。语义组合性（semantic compositionality）[1]则表明整体的语义多大程度上可以由部分的词根语素推断出来。例如，“道路”比“梦想”的组合性要更强。语义组合性较弱的词往往带有语义上的特异性（idiosyncrasy）,即无法从字面去预测出整体的意思，这类词汇可以放入到一个词库（lexical）中，也被称为词库词，另一种组合性高的词汇则不必单独列出，可通过词法规则地构造出来，也被称为词法词(董秀芳, 2004)。

不同于语言学领域通过内省的方法判断复合词的语义关系，计算语言学领域利用分布式语义假设(Harris, 1954)，通过大规模语料库训练出一个语言模型，给每个词汇学习一个词向量[2]。研究表明，这样训练出来的词向量可以反映一定的语义，例如，它可以用来计算词汇相似性(Mikolov et al., 2013)、寻找主题词(Xun et al., 2017)，或者作为各类下游任务的初始化向量(Kenton and Toutanova, 2019)。一些工作(Buijtelaar and Pezzelle, 2023)也探讨了合成词中的整体和部分的各种语义关系，包含可组合x性和主导性。这些工作主要以英语等印欧语系的语言作为研究对象，大多仅仅涉及到英语中的名词复合词，应用在中文的研究却较少。然而，印欧系语言的词汇词形变化都很丰富，它们的构词以曲折和派生两种方式为主，合成词数量较少，且常常用空格隔开（像sunlight这样合并在一个词单位的较少），这些与汉语的高分析性、派生构词的特点迥然不同，对于汉语的向量语义分析应该对合成词内部的语义关系予以高度重视。

因此，对于以理解汉语语义为目标的词向量而言，它能否反映语言学中关于词内部的两类关系是评估向量好坏的重要方面。本文首先利用一个已有的语言模型提取出来的词向量，通过计算部分向量和整体向量间的相似性，利用两个指标LMD和ST来分别量化语义主导性和语义组合性。之后通过线性回归的方式，找出影响LMD和ST的影响因素。最后，本文选择了词汇的结构信息（即联合、偏正、动宾、动补等结构）作为语义主导性的测试场景，选择语素的自由度以及是否进入词库这些特征来检验语义组合性，观察这些向量体现出来的语义关系，与语言学的研究是否吻合，从而更好地衡量语言模型学习到的词汇向量是否学习到了这些知识。

本文研究有以下的贡献和结论：

- 通过LMD和ST两个指标量化了汉语词汇内部的语义主导性关系和组合性关系。
- 找出了影响语义主导性和组合性的影响因素，包括词汇频率、词汇多义性、词汇及组成部分的语素的具体程度等。

[0]汉语中的复合方式还包括操作在词上的词复合，本文参考朱德熙先生的定义，仅考虑词根复合的方式。

[1]心理语言学也将这个性质称作透明性（transparency）(Buijtelaar and Pezzelle, 2023)。

[2]由于词向量相当于嵌入在高维语义空间上的一个点，它又被称为词嵌入（word embeddings）

- 测试了在不同词汇结构信息下的语义主导性和语义组合性，并发现它们的分布与语言学和人类的直觉较为吻合。
- 测试了语素的自由程度和能否成为词库词对于组合性的影响，与语言学上的相关结论也较为一致。

2 相关工作

2.1 复合词的两类语义关系

复合词是心理语言学中热衷讨论的话题，主要原因是它们具有很高的能产性，往往可以造出新的词汇，并且在语用过程中词汇化 (Gagné and Spalding, 2006)，发生语义特异性从而进入到词库中，也因此复合词又被视为“词库的后门”（backdoor into the lexicon）(Downing, 1977)。许多语言学的研究探讨复合词和它们的部分之间的语义关系，包含中心理论(Packard, 2000)、语义的可组合性(董秀芳, 2004)等，同时也发现很多影响因素，包括词汇频率、语义透明度等等(Gagné and Spalding, 2009; Ji et al., 2011; Marelli and Luzzatti, 2012; Pennington et al., 2014; Buijtelaar and Pezzelle, 2023)。

汉语的复合词众多，有许多研究复合词内部的结构和语义的工作。结构方面，一直以来就有两种说法(池昌海, 2019)，一派的观点(王力, 2015; 陆志韦, 1964; 朱德熙, 1982; 吕叔湘, 1979)认为词结构和短语结构是平行的，并归纳出偏正结构、动宾结构等句法也会采纳的分类，如吕叔湘、朱德熙 (2013)认为“双音词的构成跟短语相似”；另一派(刘叔新, 1990; 彭迎喜, 1995)则持有相反观点，本文采用前者的观点，并对词汇进行了分类。语义方面则包括复合词的认知机制(刘正光, 2004)、内部语素的词类鉴定和自由度(董秀芳, 2004)、词汇化(董秀芳, 2003)等。

2.2 词向量

计算语言学使用词向量来表征词汇的语义。Mikolov等人 (2013)在其论文中引入了Word2Vec模型，该模型通过神经网络训练来学习词向量表示。这项工作通过高效的训练过程（Skip-gram模型以及负采样）捕捉词语语义。另一个重要的贡献是Pennington等人 (2014)提出的GloVe模型，它结合了全局语料统计信息和局部上下文信息来生成词向量表示。GloVe模型在语义和语法任务上取得了显著的成果，并为词向量研究带来了新的思路。针对处理未登录词和具有丰富形态变化的词的挑战，Bojanowski等人 (2017)提出了FastText模型，该模型基于子词级别的向量表示。通过对单词进行子词分解和向量化，FastText模型能够更好地处理这些复杂情况。上述模型都是静态词向量，没有考虑的词汇的多义性以及上下文的重要性。随着对上下文的理解变得越来越重要，Peters等人 (2018)引入了ELMo模型，该模型能够提供上下文相关的词向量表示。ELMo通过使用双向语言模型来学习词语在不同上下文中的表示，从而捕捉到词语含义的多样性和上下文敏感性。近年来，基于Transformer模型(Vaswani et al., 2017)的预训练语言模型引起了广泛关注。Devlin等人 (2019)的BERT模型利用大规模文本数据的预训练，在自然语言处理任务上取得了革命性的突破。BERT通过学习上下文相关的词向量表示，极大地推动了自然语言处理领域的研究和应用。此外，Dai等人 (2019)提出的Transformer-XL模型对Transformer模型进行了改进，通过引入循环机制扩展了模型的上下文能力，从而在处理长文本和长距离依赖关系时表现出色。本文的词向量模型DSG (Song et al., 2018)基于Skip-gram，并增加了词汇位置信息，有效提高了模型性能和效率。

3 研究方法

3.1 研究对象

本文主要的研究对象是双音节[3]汉语合成词，记做XY，其中X和Y分别表示它的左部分的语素[4]和右半部分语素。其中，合成词包含占数量优势的复合词，也包含附加和重叠构词，这是考虑到附加构词尽管没有很强的语义关系，它的缀部分的意义已经非常虚化了，但也可以作为一种特殊的语义主导性关系，即，词根部分占据主要意义。而重叠构词则可以看作是两部分意义主导性平均分配。这两种情况都可以作为一种平凡（trivial）情形，供研究参考。

[3]为了方便识别左右部分，本文仅考虑双音节。
[4]本文采用的语素的概念为最小的音义结合体，在汉语合成词中，一个字即为一个语素。

3.2 衡量指标

参考前人研究(Buijtelaar and Pezzelle, 2023)，针对词向量[5]，语义主导性和语义组合性可以利用相似性刻画。具体而言，以目标词w作为带有参数θ的模型f的输入，最终可输出一个d维的词向量e：

$$e = f(w; \theta). \tag{1}$$

针对任意两个词w_i，w_j对应的词向量e_i，e_j，可以通过一个相似度度量算子$\mathcal{S}$来衡量它们的相似程度，这里我们采用余弦相似度来表示：

$$\mathcal{S}(w_i, w_j) = \cos(e_i, e_j), \tag{2}$$

值越大表示越相似，1表示最相似，0表示最不相似。

语义主导性（记做LMD, **L**exical **M**eaning **D**ominance）旨在求得XY中X抑或Y的语义更加重要。本文基于如下的假设：与整体的语义更加相似的那部分，其语义更加重要。通过以下的方式求得：

$$\text{LMD} = 0.5 \times (\mathcal{S}(Y, XY) - \mathcal{S}(X, XY)) + 0.5 \tag{3}$$

当右边相比左边占据绝对主导性时，$\mathcal{S}(Y, XY) = 1$, $\mathcal{S}(X, XY) = 0$, 此时，LMD值为1，与之相反当LMD值为0时，左边更加重要。LMD的值越大，说明右半部分更加具有语义主导性。

语义组合性（记做ST, **S**emantic Compositionality/**T**ransparency）也是基于以下的假设：整体的语义与各个部分都更加相似的时候，语义的组合性越强。通过如下的方式求得：

$$\text{ST} = 0.5 \times (\mathcal{S}(Y, XY) + \mathcal{S}(X, XY)). \tag{4}$$

当ST值为1时候，$\mathcal{S}(Y, XY) = 1$, $\mathcal{S}(X, XY) = 1$，此时语义组合性最强，反之ST值为0时，语义组合性最差。即，ST值越大，语义组合性越强。

4 实验过程

4.1 知识库

中国知网（HowNet）：知网（HowNet）(Dong and Dong, 2003)最早是由董振东和董强先生在20世纪90年代设计和构建的一部更加适用于中文的语言知识库，它利用常见汉字构建出最小的语义单元（即义原）集合，并利用它们对十几万的中英文词条进行语义标注。知网作为一个大型知识库，体现在一方面，义原标注采用了较为结构化的方式来进行，即罗列属性和对应的属性值（或也称为特征）以及复杂的语义角色关系的方式。另一方面，2500多个义原概念之间也存在多种关系，例如上下位关系、同义关系、反义关系、对义关系等。

4.2 语言模型

DSG: 本文提取词向量的语言模型DSG参考的腾讯实验室的一篇工作(Song et al., 2018)，它通过引入一个方向向量，将词汇在上下文的位置信息融入到skip-gram模型中。它采用的语料库是维基百科的文章，大约包含20亿个词。最终提取出来的词汇有200万个，每个词汇对应一个200维的向量。

4.3 数据准备

本实验从上述单词中选择双音节的词汇XY，并且X和Y都出现在词汇库中的那些词，将这个原始版本记做$\mathcal{D}$。由于我们之后要利用知网中知识，我们从中找出那些出现在知网中的词汇，并记做$\mathcal{D}_H$。为了提供更加“极端”的例子，我们利用知网中的相似度算法 (Liu et al., 2013)，将$\mathcal{S}(X, XY) = 1$和$\mathcal{S}(Y, XY) = 1$的那部分词汇挑选出来，记做$\mathcal{D}_F$。这些例子表明这些词汇的相似度会偏向某一端。这三个数据库中词汇的数量可以参考表 1。其对应的所有实例已经上传到网上[6]。

为了评估LMD和ST的能否有效衡量相似性，本文通过它和知识库知网中反映的相似度做一个对比。将公式 3和 4中的余弦相似度换做知网中求相似度的算法，也可以计算出相应

[5]本文称作的词向量中的词并非语言学意义严格的词，也包含语素，以下表述中并没有严格区分这两者

[6]https://github.com/RyanLiut/Compound_Words

数据集	数量	LMD			ST	
		均值	相关性	准确性	均值	相关性
$\mathcal{D}$	325,300	0.49 ± 0.03	-	-	0.51 ± 0.02	-
$\mathcal{D}_H$	49,481	0.51 ± 0.01	22.78	51.14	0.50 ± 0.01	25.76
$\mathcal{D}_F$	11,495	0.50 ± 0.01	44.02	69.72	0.52 ± 0.01	8.60

Table 1: 不同规模的数据集下对应的LMD和ST的结果

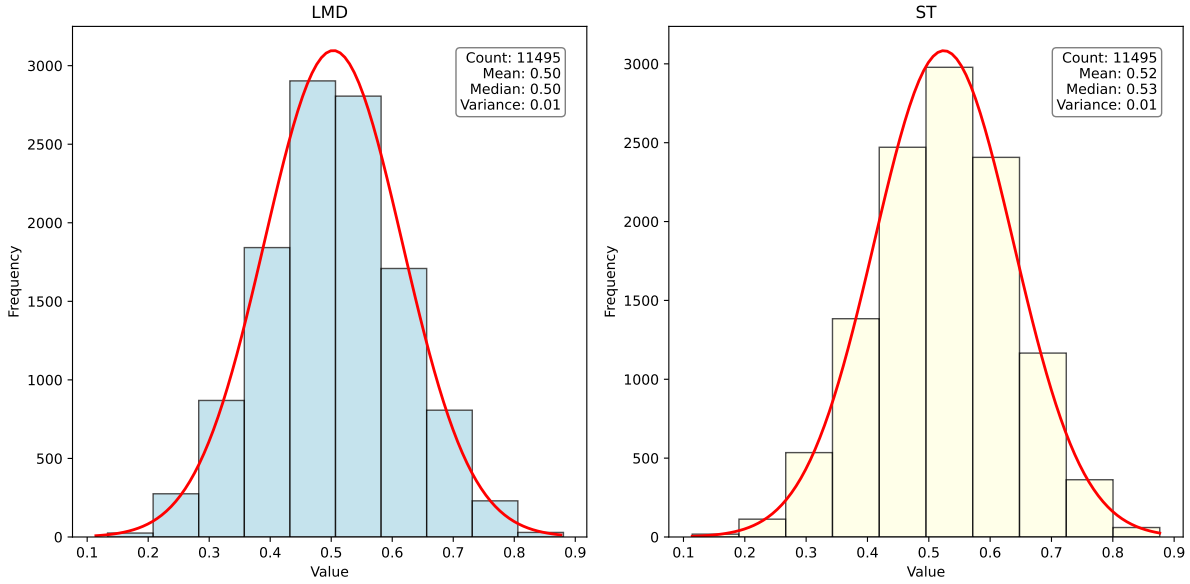

Figure 1: LMD和ST在$\mathcal{D}_F$上的分布

的LMD和ST值，将知网中计算出的这两个值和词向量余弦相似度计算出来的值求一个相关度CORR，可以反映二者的相关性。另外，我们也通过一个分类准确率ACC来计算它和知识库分类的差距：

$$\text{ACC} = \frac{\mathbb{1}\{(\text{LMD}_{\text{COS}} - 0.5) \times (\text{LMD}_{\text{HN}} - 0.5) > 0\}}{N} \tag{5}$$

其中N表示样本总数。ACC仅仅用来反映LMD的好坏。

5 结果分析

图 1中展示了LMD和ST在$\mathcal{D}_F$上的数据分布，附录 7中显示出了$\mathcal{D}$和$\mathcal{D}_H$在LMD和ST上的数据分布，并且它们通过了正态假设性检验，可以视作是一个近似的正态分布，我们也将样本的分布模拟为一个正态分布。表格 1显示了词向量在不同的单词集合中的评估指标ACC和CORR，可以看到$\mathcal{D}_H$在LMD的相关性和准确性上表现得没有$\mathcal{D}_F$得好，这可能是由于原始的数据中X和Y的相似性较大（故，正态分布中的大多数样本都集中在中部），很多样本无法准确确定哪部分更加相似。而提前根据知网选取的语义极端的样本，可以去除掉这部分“困难样本”，使得基于相似性的词向量可以有效判断出具有语义主导地位的部分。

对于语义主导性而言，判断准确性接近70%，本文认为可以有效地反映词向量以及LMD计算指标的可靠性，因此在之后的分析实验中，针对影响因素（5.2）和语义主导性（5.3）采用$\mathcal{D}_F$。而针对语义组合性，$\mathcal{D}_F$上的相关性仅有8.60%，具有很低的相关性，侧面反映了向量的相似性计算过程和知识库（如知网）得到的相似性过程很不相同。

5.1 影响因素

语义主导性和语义组合性可能会受到哪些因素影响呢？本文从三个角度进行考虑：

影响因素	LMD		ST	
	系数	p值	系数	p值
nPD	0.1098	0.000	0.1324	0.000
nSememes	0.1858	0.000	0.2262	0.000
freq	-0.1726	0.000	0.0726	0.646
semefreq	0.1858	0.000	0.2262	0.000
conc	-0.1726	0.000	0.0726	0.646
conc_X	1.0780	0.000	0.6015	0.000
conc_Y	0.4197	0.000	1.6560	0.000
isDict	0.4502	0.000	0.4545	0.000

Table 2: 对于语义主导性和语义组合性的影响因素

- 多义性。多义性考虑词汇整体XY的多义程度。通过XY在知网中所包含的义项个数nPD，以及它所包含的所有义素的个数nSememes来量化表示。
- 频率。通过词汇在大规模语料中出现的频次[7]freq和词汇在知网中第一个义项对应的义素的频次和semefreq来表示。
- 具体性。刻画词汇的具体程度[8]，使用conc，conc_X, conc_Y分别来表示整体的、左部分和右部分的具体程度。
- 是否进入词库。该词汇是否进入到现代汉语词典中，对于进入到汉语词典的词，其词义往往更具有特异性，也有可能影响语义关系。记做isDict。

通过提取出这些量化的值与相应的LMD值和ST值做一个回归分析，可以得出那些变量对于目标值有显著影响，从而初步得出哪些因素更加重要。表 2展示了这些信息，其中p值表示该因素为0的显著性，其值越小，就越不为0，该因素也越明显。可以看出，对于LMD而言，这些因素都有影响，而对于ST而言，除了频率和具体程度外，其他因素也都对语义组合性有影响。

5.2 语义主导性

本文重点分析了词汇的结构对于语义主导性ST的影响。采纳之前工作(Zheng et al., 2021)的分类和标注，本文选取类型下样本量超过50的6种结构类型，包含后缀结构, 连谓结构, 状中结构, 述宾结构, 定中结构和联合结构。图 2(a)展示了它们的语义主导性分数的均值，以及对应数量，(b)展示了两两之间的显著性差异，其中差异显著的对应位置用星号做了强调。可以发现，同语言学知识类似，后缀情况下左边的语义要占据主导地位，而相比部分之间“势力均衡”的联合结构，“修饰-被修饰”的状中和定中的右边要占据语义主导地位。

5.3 语义组合性

语义组合性ST体现了部分的语义多大程度可以预测整体的语义，借助于前人 (董秀芳, 2004)关于半自由语素的研究:半自由语素介于自由语素和粘着语素之间，是不可以单用，但与一些成分结合之后就可以出现在一些典型的句法环境中，而与这些成分结合的可以是一个独立的词，因而这个语素似乎具有了独立的词的地位，例如“学校”中的“校”可以出现在“此校”中，成为一个典型的名词用法，而“此”后往往可以跟独立的名词，例如“此病”。因此“校”可以认为是一个半自由语素。

本文认为一个半自由语素和一个独立成分构成的词汇，要比它作为一个粘着语素构成的词汇，语义组合性更强，因为它的各个部分的活动性都更强，更像是一个短语，例如“此校”的组合性要比“学校”的组合性更强，我们针对前人研究中的半自由语素（主要参考动词性和名词性半自由语素）X，根据与它结合的另一部分Y的词类$\mathcal{C}$，找到语料库中更多这样的词汇（称为半

[7]词频是从150亿规模的语料中提取出来的，参考https://www.plecoforums.com/threads/word-frequency-list-based-on-a-15-billion-character-corpus-bcc-blcu-chinese-corpus.5859/

[8]词汇具体程度通过前人工作得知(Xu and Li, 2020),用1-5之间的数值表示，越小表示越具体

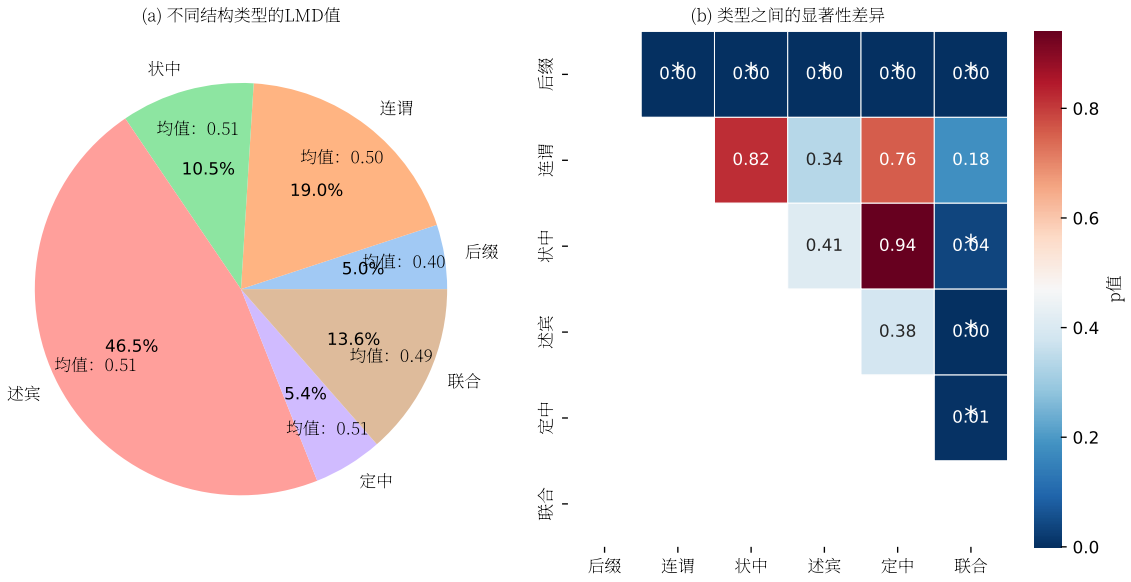

Figure 2: 不同结构类型下的LMD分布及其差异显著性

自由词），同时找出不在$\mathcal{C}$中的词类的Y，作为X实现为粘着语素的另一部分，称为粘着词。通过对于半自由词和粘着词内的平均ST值，来分析词向量可否反映出前述的假设。

表 3列出了以半自由语素为词根，对应的半自由词以及粘着词，由于这些半自由语素都处于词的右半部分，为节约空间，词汇举例时候只列出了它的左半部分。同时，我们也输出了相应词汇集合的ST均值，用*号表示差异显著。从表中可以观察到，除了“择”和“擅”两个语素对应的两类词，半自由的组合性小于粘着的，其余的都是半自由词的大于粘着词的，并且在很多实例中都是显著差异。这反映了词向量捕捉到了半自由语素的粘着程度更低，语义更加规则的特点。

6 结论

本文探讨了大规模语料中学习到的词向量是否可以捕捉到汉语复合词中的两类语义关系：语义主导性和语义可组合性。通过不同词汇结构的差别和半自由语素构成了两类词的差异，本研究说明了模型得出的词向量可以反映语言学上得到的这两类语义关系。通过线性回归，我们发现了影响这两类语义关系的影响因素，这可以启发后续对于复合词更深入的理解。

6.1 未来工作

本研究仍有许多工作要做，包括探讨基于相似度的LMD和ST是否可以全面反映语义主导性和语义组合性，影响两类语义关系的影响因素对计算语言学相关的启发，以及复合词的这些特点如何有效地帮助我们处理中文的语言数据等等。

参考文献

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Lars Buijtelaar and Sandro Pezzelle. 2023. A psycholinguistic analysis of bert's representations of compounds. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2222–2233.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.

半自由语素	半自由词		粘着词	
	举例	均值	举例	均值
感	不互倍偶共即多已很忽极此甚略百稍自虽颇	0.4686	业交代传伤体冰冷凉剧力动反口听味	0.4532
知	不为仅共即只哪如定岂已必怎才既更略皆真至至莫要颇	0.5109 *	乐亲人从会侦元先公出初告善国均	0.4601
予	不再多请首	0.5430∗	交付传借免准售均嫁宰寄小授施	0.4278
返	一三不再即已	0.6058∗	上会全劝南回复外带引归往忘快折抵	0.4751
珧	一以至	0.6814∗	古大小文日智	0.4658
愈	不则已老而自	0.6484	会伤全听得心末正治渐疗病痊看	0.5354
恐	不或极犹诚	0.6903∗	反唯密思惊惟慌暴涉社防震	0.4828
择	不再另张自	0.5467	中先决可后天慧抉拣选采	0.5711
畏	不首	0.6474	可怖敬生相	0.5317
受	另莫可无难	0.6047	寻杨林欲觅追锦	0.5537
擅	不独	0.7471	专精	0.776

Table 3: 共同词根对应的半自由词和粘着词举例以及ST均值

Zhendong Dong and Qiang Dong. 2003. Hownet-a hybrid language and knowledge resource. In *International conference on natural language processing and knowledge engineering, 2003. Proceedings. 2003*, pages 820–824. IEEE.

Pamela Downing. 1977. On the creation and use of english compound nouns. *Language*, pages 810–842.

Christina L Gagné and Thomas L Spalding. 2006. Conceptual combination: Implications for the mental lexicon. *The representation and processing of compound words*, pages 145–168.

Christina L Gagné and Thomas L Spalding. 2009. Constituent integration during the processing of compound words: Does it involve the use of relational structures? *Journal of memory and language*, 60(1):20–35.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Hongbo Ji, Christina L Gagné, and Thomas L Spalding. 2011. Benefits and costs of lexical decomposition and semantic integration during the processing of transparent and opaque english compounds. *Journal of Memory and Language*, 65(4):406–430.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Jiangming Liu, Jinan Xu, and Yujie Zhang. 2013. An approach of hybrid hierarchical structure for word similarity computing by hownet. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 927–931.

Marco Marelli and Claudio Luzzatti. 2012. Frequency effects in the processing of italian nominal compounds: Modulation of headedness and semantic transparency. *Journal of Memory and Language*, 66(4):644–664.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Word2vec: Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jerome L Packard. 2000. *The morphology of Chinese: A linguistic and cognitive approach*. Cambridge University Press.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 175–180.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xu Xu and Jiayin Li. 2020. Concreteness/abstractness ratings for two-character chinese words in meldsch. *PloS one*, 15(6):e0232133.

Guangxu Xun, Yaliang Li, Wayne Xin Zhao, Jing Gao, and Aidong Zhang. 2017. A correlated topic model using word embeddings. In *IJCAI*, volume 17, pages 4207–4213.

Hua Zheng, Lei Li, Damai Dai, Deli Chen, Tianyu Liu, Xu Sun, and Yang Liu. 2021. Leveraging word-formation knowledge for chinese word sense disambiguation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 918–923.

刘叔新. 1990. 汉语描写词汇学. 商务印书馆.

刘润清刘正光. 2004. N+ n 概念合成名词的认知发生机制. 外国语, (1):26–32.

吕叔湘. 1979. 汉语语法分析问题. 商务印书馆.

朱德熙吕叔湘. 2013. 语法修辞讲话. 商务印书馆.

彭迎喜. 1995. 几种新拟设立的汉语复合词结构类型. 清华大学学报: 哲学社会科学版, (2):34–36.

朱德熙. 1982. 语法讲义. 商务印书馆.

林志永池昌海. 2019. “汉语复合词的结构与句法结构平行” 说新议. 浙江大学学报(人文社会科学版), 5.

王力. 2015. 龙虫并雕斋文集: 二. 中华书局.

葛本仪. 1985. 汉语的造词与构词. 文史哲, 4:28–33.

董秀芳. 2003. ” x 说” 的词汇化. 语言科学, 2(2):46–57.

董秀芳. 2004. 汉语的词库与词法. 北京大学出版社.

陆志韦. 1964. 汉语的构词法. 科学出版社.

7 附录

7.1 LMD和ST在另外两个数据集上的分布

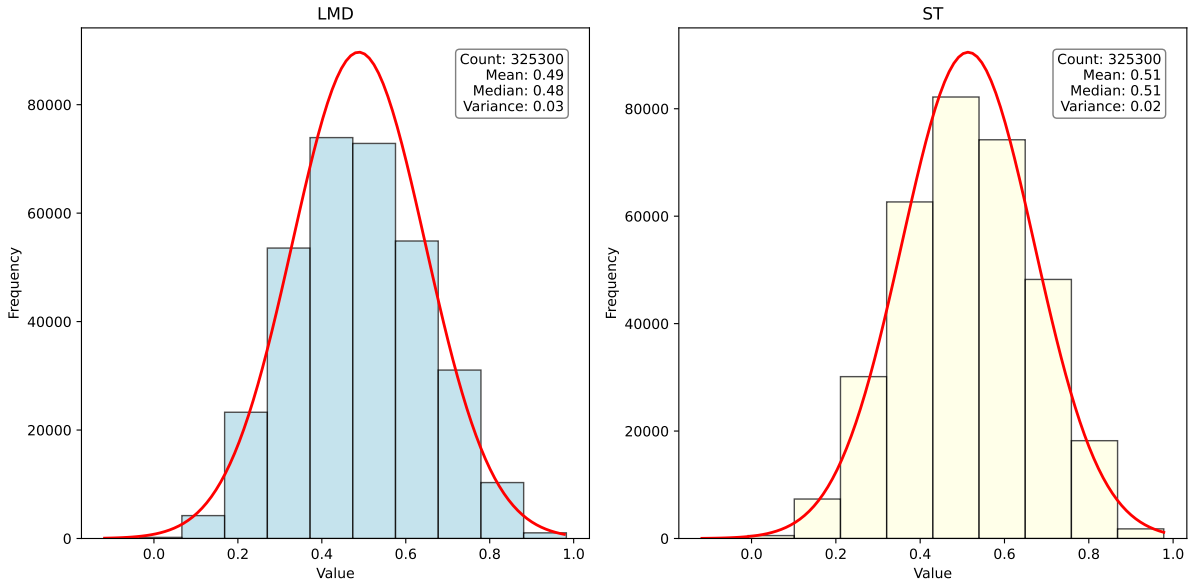

Figure 3: LMD和ST在$\mathcal{D}$上的分布

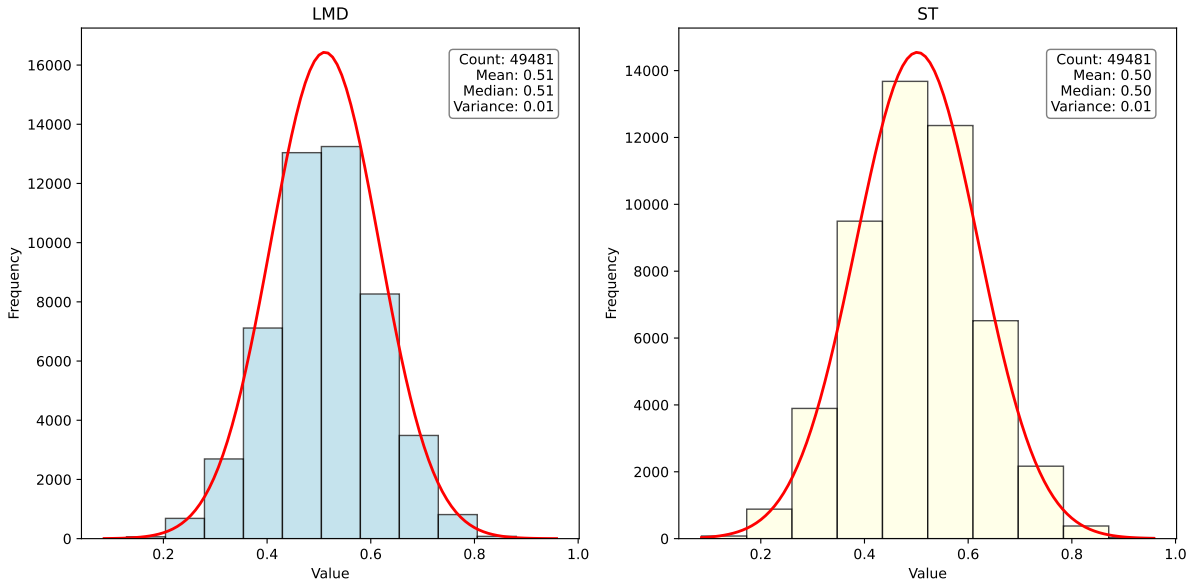

Figure 4: LMD和ST在$\mathcal{D}_H$上的分布