文章编号：1003-0077(2021)05-0017-10

基于分布式表示的汉字部件表义能力测量与应用

梁诗尘[1,2]，唐雪梅[1,2]，胡韧奋[1,2]，吴金闪[3]，刘智颖[1,2]

(1. 北京师范大学 中文信息处理研究所，北京 100875；
2. 神州泰岳-北京师范大学 人工智能联合实验室，北京 100875；
3. 北京师范大学 系统科学学院，北京 100875)

摘　要：汉字的表义性是其区别于表音文字的一大特点。部件作为构字单位，同汉字的意义之间有着很大的联系。然而，汉字部件的表义能力究竟如何是学界尚待讨论的课题。针对这一问题，该文从汉字部件入手，提出了融合部件的字词分布式表示模型。该模型在向量内部评测任务上性能获得了一定提升，在汉字理据性测量任务上也与人工打分结果显著相关。基于该模型，进一步提出了部件表义能力的计算方法，对汉字部件的表义能力做了整体评估，并结合部件的构字能力建立了现代汉字部件的等级体系。测量结果显示，现代汉字部件具有一定表义能力，但整体而言表义能力偏低。最后，将测量结果应用于对外汉语教学中，确立了适用于部件教学法的部件范围，并提出了对应的汉字教学顺序方案。

关键词：汉字部件；表义能力测量；分布式表示

中图分类号：TP391　　**文献标识码**：A

Measurement and Application of Chinese Component Semantic Ability Based on Distributed Representation

LIANG Shichen[1,2], TANG Xuemei[1,2], HU Renfen[1,2], WU Jinshan[3], LIU Zhiying[1,2]

(1. Institute of Chinese Information Processing, Beijing Normal University, Beijing 100875, China;
2. UltraPower-BNU Joint Laboratory for Artificial Intelligence, Beijing Normal University, Beijing 100875, China;
3. School of Systems Science, Beijing Normal University, Beijing 100875, China)

Abstract: The semantic representation of Chinese characters is one of the characteristics that distinguishes them from phonetic characters. As a unit of character construction, components are closely related to the meaning of Chinese characters. However, how to measure the meaning of Chinese character components is an issue remains to be discussed. In this paper, we focus on components in Chinese character and train a multi-granularity Chinese word embedding, which are proved positive in the internal evaluation task of word embedding and the motivation measurement of Chinese character. Based on this model, we further put forward a formula to calculate the semantic ability of components, revealing that components in Chinese characters have certain but limited semantic ability. Meanwhile, we further establish the grading system of components by taking the semantic ability of components into account. Finally, for the teaching of Chinese as a foreign language, We establish the scope of component teaching, and put forward a scheme of teaching sequence of Chinese characters.

Keywords: Chinese character component; semantic ability measurement; distributed representation

收稿日期：2019-09-19　**定稿日期**：2019-10-19

基金项目：国家语委科研项目(ZDI135-42)；国家社会科学基金(18CYY029)；教育部人文社会科学基金(18YJAZH112)

0 引言

汉字是世界上唯一未曾中断使用而延续至今的表义文字系统。[1]在学界和大众的普遍认知里,汉字都是形、音、义三者的结合体。与表音文字不同的是,汉语系统具有特殊的表义性。汉字的理据性就在于汉字的字形直接与汉字的音、义发生联系。然而,经过几千年语言和文字的发展,汉字系统的表义能力究竟如何?这仍然是一个没有定论的议题。

部件作为由笔画组成的具有组配汉字功能的构字单位[2],在一定程度上与汉字的音、义有关。对汉字系统表义能力的考察应该从汉字的字形入手,其中部件的表义能力自然是不能忽略的一环。值得注意的是,部件与部首、偏旁等含义不同。部首是汉语字典里属于同一偏旁的部目;偏旁是汉字以及部件中具有表音或表义功能的部分。而"所谓'部件',是按汉字的结构分解出来的,它们是构成汉字的常用零件,并不等同于部首或偏旁。"[3]其概念比偏旁和部首更广一些。同时部件也与意符(或称"形符")、声符等概念有所差别。意符、声符是形声字范畴下的概念,它们都能算作汉字部件,但部件概念适用范围并不局限于形声字。在本文中,我们认为不仅仅是部首或是意符才具有表义的功能,其他部件也可能与汉字意义产生关联。因而我们选择了"部件"这一内涵更广的概念,从部件的表义能力出发考察现代汉字的表义性。

相对而言,在语言系统中,语音的物理属性更为突出,也更容易进行测量,而语义信息则要更为复杂,表示和计算都存在一定的困难。在以往的研究中,有学者曾经利用计算机对汉字声符的表音能力进行测量[4]。而关于部件表义能力的测量也有学者使用人工统计的方法测量小部分部件的表义能力,但这种方法只适用于小规模部件,并且容易带有主观性。

近年来,随着分布式表示的发展,词语语义信息能够在向量空间得到很好的表示。受到这一方法的启发,我们提出了融合部件的字词分布式表示模型,将汉字部件与字词嵌入到同一向量空间中进行表示,这一方法在向量内部评测任务上取得了一定的提升,在判定汉字理据性的任务上也与人工打分结果显著相关。更进一步,我们提出了部件-字相似度和字间相似度两项衡量部件表义能力的指标,并结合部件的构字能力建立了现代汉字部件的等级体系。此外,我们还将部件表义能力测量的结果应用于对外汉语教学,用以确定适用部件教学法的部件范围以及汉字教学的顺序,并提出了具体的可行方案。

本文的主要贡献在于:①提出了部件的分布式表示方法和部件表义能力的自动测量方案,对汉语部件的表义能力有了整体的把握。②提出了将计算机自动测量部件表义能力应用到实际教学之中的可行方案,有助于减轻教师和学生的负担,提升教学的科学性。

1 相关工作

1.1 汉字部件表义能力测量

针对汉字部件及表义的研究在语言学领域目前已有比较丰富的学术成果,研究对象主要包括部首和意符、声符等,研究内容涵盖了表义状况分析[5]、表义机制探究[6]以及表义部件在教学中的应用[7]等方面,但涉及到部件表义能力测量的研究较少。施正宇[8]对现代汉语 3 500 个常用字和次常用字中的形声字进行统计,将形声字形符按表义功能分成了不表义、间接表义和直接表义三类,并发现间接表义的形符占绝大多数。李蕊[9]以形旁能表义的字数占总体部件构字的比例为表义度的衡量标准,从形声字的等级、形旁的表义、位置、是否成字以及构字数等多个角度分析了形声字形旁的表义状况。吕菲[10]利用义素分析法和核心义素与语境语素两个理论对古今形声字意符表义能力进行了考察,发现与古代形声字相比,现代形声字意符表义能力有所降低,但下降幅度较小;同时意符表义的方式也向精细、曲折的方向发展。陈爱华[11]根据意符直接表示意义或是表示意义范畴,以意符作为汉字主要义项的汉字比例为衡量标准来确定意符的表义度。他发现在不同汉字难度等级下,一些部件的构字能力和表义度发生了变化。这些对汉语部件表义能力的测量大部分局限在形容词的意符范围上,同时采用人工统计的方法,缺乏普适性。

1.2 引入汉字信息的词语分布式表示

分布式表示的方法为测量部件表义能力提供了新的思路。分布式表示最早出现于由 Mikolov 等[12]提出的 Word2Vec 模型。分布式表示模型基于 Firth[13]提出的分布式假说:出现在相同上下文

中的词往往具有相似的语义。因而在分布式表示模型得到的向量空间中，词义相近的词会有接近的向量表示，词与词的语义关系可以通过向量间的余弦进行计算。

近年来，在中文分布式表示上涌现了不少对词向量进行改进的研究，其中就包括利用汉字、字形和部件等信息进行的相关工作。Chen 等[14]提出了将字和词进行联合训练的 CWE 模型，由于大部分汉字词语和所组成的字之间存在语义关联，该模型比不考虑字信息的模型取得了更好的效果。Sun 等[15]及 Yin 等[16]在字词联合训练的基础上加入了目标词的部首信息，这些模型也在词类比和词相似任务上取得了更好的性能，同时可以更好地识别细粒度的词义。Tzu-Ray Su 等[17]在此基础上提出了利用繁体汉字图片来学习词语表示的新方法，该模型从汉字的图片中学习汉字的字形特征，也提升了词表示的效果。

以上研究证实了部件参与词表示的有效性，以及通过联合训练获得部件的向量化表示，计算部件与词语间的相似性的可能性。但值得注意的是，以往的训练方法或者只采用了部首信息，没有引入非部首的其他部件；或者引入汉字整体字形，但没有拆分部件。我们认为非部首之外的其他部件对于字义或词义也是有一定贡献的，因而我们的模型将汉字的所有部件都加入了训练。

2　融合部件的字词分布式表示

在以往的分布式表示模型基础上，我们加入了全部部件信息，并使用多粒度的向量模型，同时进行词向量与部件向量的训练。

2.1　数据与语料来源

本文所用汉字拆分数据来源于汉字结构网络与理解型学习系统①，共包括 3 993 个简体汉字。汉字部件拆分，以文献为参考，主要根据汉字的简体字形拆分，对于简体难以拆分的汉字，也参照汉字的古体字形。在汉字构字类型上，主要遵循传统的“六书”观点进行划分：形声字的数量占据绝大多数，共 2 896 个，此外有会意字 747 个，象形字 286 个，指事字 64 个。在汉字所含部件数上，数据的拆分颗粒度相对较粗，85％的汉字被拆分成两个部件，最多被拆分成 4 个部件。如图 1 所示，不同的构字类型在部件数量上也存在一定的差异：会意字和形声字以两部件汉字为主，象形字和指事字以单部件汉字为主。

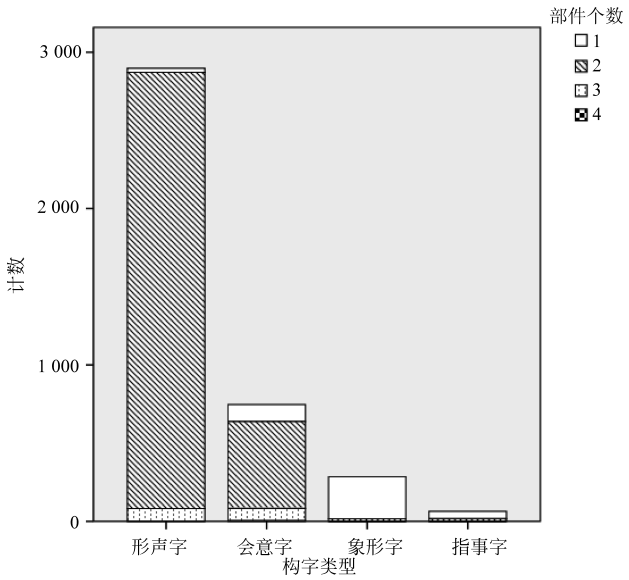

图 1　不同构字类型下汉字的部件个数统计

出于语料均衡性的考虑，本文所使用的现代汉语语料来源于网络爬取的百度百科语料，并进行了数据清洗，共计 4.1G。

2.2　分布式表示方法

为了将部件向量与字、词向量一起进行训练，我们在 Zhao 等[18]提出的 ngram2vec 工具包上进行了改进，这个工具包可以对词向量训练时的“目标词-上下文对”进行调整。与以往其他研究只加入部首信息的做法略有不同，本文将所有部件都作为词的上下文加入向量的训练。训练的主体模型为带有负采样的 Skip-Gram 模型(SGNS)。在 Skip-Gram 模型中，每个词语 w_i 由两个低维稠密的向量进行表示，分别是是中心词向量($\boldsymbol{w}'_i$)和上下文向量($\boldsymbol{c}'_i$)，Skip-Gram 通过梯度下降的方法取得，如式(1)所示。

$$\hat{p}(c_i \mid w_i) \propto \exp(w'_i \cdot c'_i) \tag{1}$$

其中，$\hat{p}(c_i|w_i)$表示上下文 c_i 在 w_i 附近给定的窗口内出现的概率。

SGNS 作为分布式表示模型同样基于分布式假说：出现在相同上下文中的词往往具有相似的语义。因而加入词语中字的部件信息后，具有相同部件的字词会有更接近的向量表示。

如图 2 所示，我们采用了两种略有不同的策略

① http://www.learningm.org

来获得分布式表示：①以词为基本单位，并融合部件信息。我们将每个词中包含的汉字拆解为部件，并将部件作为词语的上下文加入词向量训练中。这样出现在相同词语中的部件将能获得相近的部件表示，从而使得训练好的部件向量承载上语义信息。最终训练将得到词向量（其中单字词的向量即为字向量）与部件向量。②同样是以词为基本单位，但在融合部件信息的同时也加入了词中字及字的位置信息，也就是将词语中的字也加入词语的上下文，并且对字在词中的位置也进行了区分（词首字 B，词中字 E，词尾字 M，单字词 S）。最终的分布式表示包含词向量、字向量、带有位置标记的字向量。

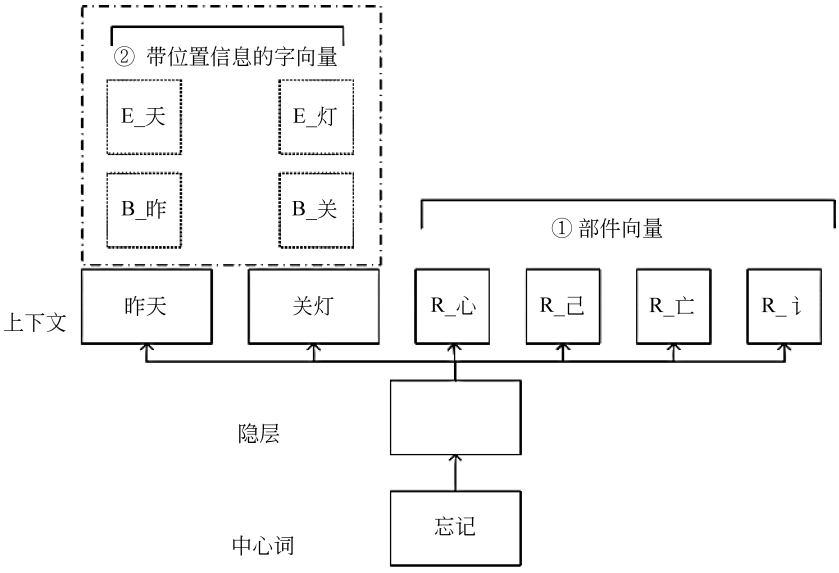

图 2 两种训练策略示意图

训练参数设置如下：向量维度为 300，窗口大小为 2，最小词频为 100，负采样为 5。

两个模型都将共同学习词、部件向量，此外，在文中第②种训练模型中还将学习带有位置信息的字向量。给定句子“不好意思，我昨天忘记关灯了。”，以“忘记”为中心词，模型需要预测窗口内的词语（“昨天”“关灯”）和“忘记”一词中含有的部件。同时，第②种训练模型还将预测窗口内字的位置。图中“R_”表示部件，“B_”“E_”及未在图中出现的“M_”“S_”为字的位置标记。

2.3 分布式表示评估

为了验证以上方法的有效性，我们采用了两类评测手段来对训练的向量表示进行评估：①词类比和词相似两个词向量内部评价手段；②训练向量计算所得的汉字理据性得分与人工打分数据的相关性。词类比和词相似是两个学界通用的词向量内部评价手段。在词相似任务上，我们使用了 wordsim-297（ws-297）[14]这一中文评测集；在词类比任务上，我们使用了 CA_translated[14]和 CA_8[19]两个中文评测集。评测结果如表 1 所示，其中对比了两种基础设置，分别是 SGNS 和 SGNS 加上不带位置信息的汉字信息（SGNS＋字）。结果表明我们的两个模型：加入部件信息的词向量（SGNS＋部件）以及加入部件和带位置的汉字信息（SGNS＋部件＋带位置的字）相比不加部件信息的模型在大部分评测指标上都取得了提升。除语义类比以外，其在语法类比和词相似上相较不加部件信息的向量而言都取得了更好的表现。其中，仅加入部件信息的词向量（SGNS＋部件）表现最佳。

此外，我们针对研究任务进行了汉字理据性的测量，并且同人工评分进行了相关性检验。汉字理据性指汉字的音、义与汉字字形之间的联系强度。“根据字符理论，现代汉字的字符分为三类，就是意符、音符和记号。意符、音符具有理据性，记号没有理据。[20]”因而，我们将汉字的理据性从声符表音度和非声符表义度两个方面来进行衡量，汉字理据性即汉字在这两项上得分的平均如式（2）所示。

$$\mathrm{moti}(C)=\frac{1}{2}[\mathrm{sim}(\boldsymbol{V}_C,\boldsymbol{V}_{Ry})+\mathrm{score}(C,Rs)] \quad (2)$$

其中，moti(C)表示汉字 C 的构字理据性，$\boldsymbol{V}_C$ 表示汉字向量，$\boldsymbol{V}_{Ry}$ 为汉字非声符部件的总向量。Rs 表示汉字声符，score(C,Rs)由下文所述形声字表音度数据直接获得。

表 1　词类比和词相似的评估结果*

模型	CA_8(语法)	CA_8(语义)	CA_translated	wordsim-297
SGNS	0.31	0.20	0.543	0.580
SGNS+字	0.42	**0.41**	0.565	0.581
SGNS+部件	**0.53**	0.38	**0.574**	**0.588**
SGNS+部件+带位置的字	0.44	0.36	0.557	0.571

* CA_8、CA_translated 为词类比评测集，wordsim-297 为词相似评测集。

形声字声符表音度相关数据来自胡韧奋等人[4] 2013 年的研究成果，包括 2 310 个形声字及其声符表音度。该数据综合考虑了声符现代发音的声母、韵母和音调信息，给出了每个形声字声符在 0-100 区间内的表音度。该表音度数值转换为[0-1]区间内的值，即我们汉字理据性公式中汉字的表音得分。由于表音度数据没有涵盖汉字拆分数据中的所有形声字，因而在计算时，我们将未涵盖的形声字及其他非形声字所含部件都视为非声符部件。非声符表义度由非声符部件向量和与汉字的余弦相似度来衡量，计算如式(3)所示。

$$\mathrm{sim}(\boldsymbol{V}_C,\boldsymbol{V}_{Ry})=\frac{\boldsymbol{V}_C\cdot\boldsymbol{V}_{Ry}}{\|\boldsymbol{V}_C\|\cdot\|\boldsymbol{V}_{Ry}\|} \quad (3)$$

其中，$\boldsymbol{V}_{Ry}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{V}_{Ryi}$，$\boldsymbol{V}_{Ryi}$ 表示汉字非声符向量，V_{Ry} 表示汉字非声符部件的总向量。

最后，我们将计算所得的汉字理据性得分情况与事先获得的人工为汉字理据性所打的分数进行相关性分析，其中对不同的字向量类型和不同部件数的字分别进行了相关性计算。

相关性计算结果如表 2 所示。整体而言，汉字理据性的计算结果与人工分值的整体相关性(部件数>0)达到 0.60，这从统计意义上说明计算机计算的汉字理据性和人工打分呈现显著正相关。从构字的部件数来看，部件数等于 2 的汉字相关系数更高。这部分的汉字占汉字总数的 85%，其中大部分是形声字。其他部件数下的汉字表现则相对较差。这些汉字大部分没有声符，可能说明汉字的声符对汉字理据性的贡献值比较大。其中从字向量的训练类型来看，单字词的表现要优于字向量，如果参考字在词中的位置信息，则是词中尾字的表现最好。综合词类比和词相似任务的表现，我们最终选择不引入位置信息，只加入部件信息的训练方法(SGNS+部件)进行之后的分析。

表 2　汉字理据性计算机计算结果与人工分值的斯皮尔曼相关系数

字向量的类型	部件数>0	部件数>1	部件数=1	部件数=2	部件数>2
字(SGNS+部件+字)	0.533	0.553	−0.232	0.566	0.045
单字词(SGNS +部件)	0.595	0.617	−0.130	0.615	0.190 0
词首字(SGNS+部件+带位置信息的字)	0.583	0.603	−0.223	0.615	0.034
词尾字(SGNS+部件+带位置信息的字)	**0.600**	**0.631**	**−0.230**	**0.638**	**0.204**
词中字(SGNS+部件+带位置信息的字)	0.580	0.602	−0.260	0.610	0.164

3　现代汉字部件表义能力测量及分级

本节我们将综合考虑汉字部件的构字能力和表义能力，形成对现代汉字部件的分级。若部件只能构成一个汉字，基本上汉字和部件是同形关系，这两个要素的意义将十分接近。因而我们只对拆分数据中能构成两个及以上汉字的 807 个部件进行分析。

3.1　部件构字能力分级

汉字部件的构字数，也就是一个部件能作为多少个汉字的组成部分，反映了部件的构字能力。从

统计结果来看，汉字部件的构字能力差异很大，在分布上呈现长尾分布。80%的部件构字数在10个以下，其中167个部件构字数为2，但构字最多的部件构字数达到了236个。

为了控制部件构字能力差异对表义能力结果产生的影响，我们依照构字能力采用了分级策略。由于相同构字数下的部件数量众多，我们在构字数的基础上引入了部件频次作为分级参照，部件字频也就是部件构成汉字在语料中的字频总和。描写语料中词语分布的齐夫定律指出，在自然语言的语料库里，一个单词出现的频率与它在频率表里的排名成反比。我们通过实验发现，分别以构字数和部件频次为主要及次要排名依据，部件频次在语料中的分布也大体符合齐夫定律。

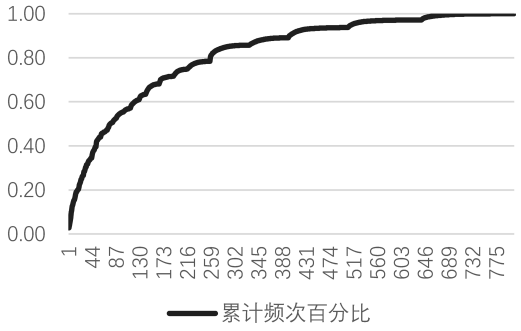

图3 汉字部件频次累计百分比统计图

注：横轴为部件序号(按构字数及部件频次从高到低对部件进行排序)。

图3是依照排名对汉字部件频次累计百分比的统计，我们发现少部分构字数多、频次高的部件的出现频次占据了语料中部件总频次的大多数。我们选择了累计占比60%及80%所对应的排名(126和428)为切分点，分别划定对应部件排序区间中的部件为构字能力强、中、弱的部件。因而排名1～126的部件为构字能力强的部件，其构字数在12至236之间，共计126个部件；排名127～428的部件为构字能力中等的部件，其构字数在4至12之间，共计302个部件；排名在429～806的部件为构字能力弱的部件，其构字数在2至4之间，共计378个部件。

3.2 部件表义能力计算方法

对于部件表义能力，我们从两个方面衡量：部件向量与字向量的平均相似度(以下简称“部件-字相似度”)以及部件构字集合中的字间平均相似度(以下简称“字间相似度”)。在基于分布式表示的词义计算研究中，研究者一般采用如夹角余弦值、欧氏距离等来衡量两词间语义相似度。本文也沿袭了这一方法，对于字与字之间、部件和字之间的语义相似度也以其向量间的余弦相似度来进行衡量。

我们认为部件的表义能力与部件与其所构成汉字的语义相似度正相关，如果部件与其所构成的汉字语义相似度越高，那么它的表义能力就越强，反之越弱。因而，在部件-字相似度计算中，我们将部件与其构成的汉字分别计算部件与字的相似度，并将这些相似度进行平均。部件向量与字向量的平均相似度计算如式(4)所示。

$$\mathrm{sim}(\boldsymbol{V}_R,\boldsymbol{V}_c)=\frac{1}{n}\sum_{i=1}^{n}\cos(\boldsymbol{V}_{ci},\boldsymbol{V}_R) \tag{4}$$

其中，$\cos(\boldsymbol{V}_{ci},\boldsymbol{V}_R)=\dfrac{V_{ci}\cdot V_R}{||V_{ci}||\cdot||V_R||}$，$\boldsymbol{V}_R$ 为部件向量，$\boldsymbol{V}_{ci}$ 为其组成字的向量。

另外，我们认为部件的表义能力也体现在其组成的汉字集合间的语义相似度上。如果一个部件所构成的汉字都具有相似的语义，那么我们认为它的表义能力较强。比如说{“松”“柏”“树”、……}是部件“木”所构成的汉字集合，假设这些字的意义都与树木相关，那么我们可以认为这个集合的语义凝聚力很强，从而推断出“木”的表义能力相对较强。据此，我们对两部件构字集合的汉字两两进行相似度计算，再将这些相似度进行平均，最后的字间相似度计算如式(3)所示。

$$\mathrm{sim}(V_{cs})=C_n^2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\cos(\boldsymbol{V}_{ci},\boldsymbol{V}_{cj}) \tag{5}$$

其中，$\cos(\boldsymbol{V}_{ci},\boldsymbol{V}_{cj})=\dfrac{V_{ci}\cdot V_{cj}}{\|V_{ci}\|\cdot\|V_{cj}\|}$，$\boldsymbol{V}_{ci}$、$\boldsymbol{V}_{cj}$ 为部件组成的字向量。

总相似度为以上两项相似度取平均的结果，最终计算如式(6)所示。

$$\begin{aligned}\mathrm{sim}&=\frac{1}{2}(\mathrm{sim}(\boldsymbol{V}_R,\boldsymbol{V}_c)+\mathrm{sim}(\boldsymbol{V}_{cs}))\\&=\frac{1}{2n}\sum_{i=1}^{n}\cos(\boldsymbol{V}_{ci},\boldsymbol{V}_R)+\frac{C_n^2}{2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\cos(\boldsymbol{V}_{ci},\boldsymbol{V}_{cj})\end{aligned} \tag{6}$$

3.3 部件表义能力测量与分析

如3.2节所述，我们通过部件-字相似度衡量部件与其所组成字之间的意义相似度，以部件构字集合的字间相似度衡量部件意义的凝聚性，并将两者的平均值作为总相似度衡量部件的表义能力。总相似度越高，表义能力越强，反之越低。我们对807个部件都进行了上述三个相似度的计算。

统计结果如图 4 所示，部件总相似度呈现较为明显的左偏分布，50％的部件总相似度在 0.4 以下，也就是说大部分部件的表义能力较弱。同时部件-字相似度和字间相似度的峰度差异明显：部件-字相似度具有负峰度，集中分布在 0.2～0.5，说明部件-字相似度在部件间的整体差异不是很大；而字间相似度具有较高的正峰度，在 0.1～1 之间都有分布，并且分布在每个区间的部件数量差距较大，说明不同部件间的字间相似度相差比较大。综上，部件-字相似度和字间相似度的分布差异反映了部件在字间相似度上的取值差异比在部件和字集合的相似度上更明显。这背后的原因可能是大部分部件和其构成的字的相似度都相对较低，但一些部件构成的字集合中字和字之间仍然保留着较高的相似度，同时另一些部件的字集合意义凝聚力则与部件-字相似度一致，保持在比较低的水平。

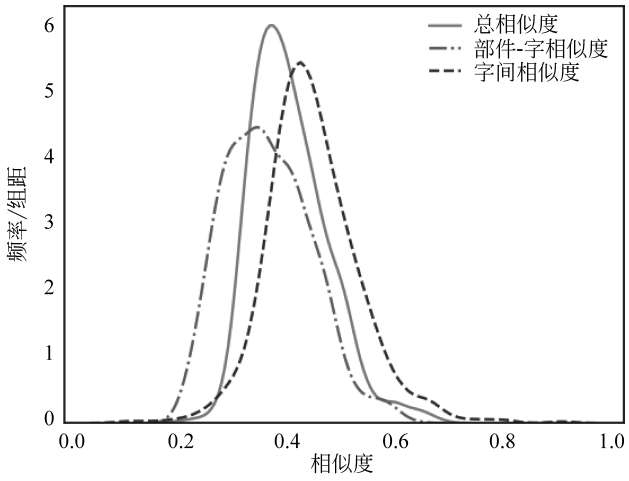

图 4　总相似度、部件-字相似度、字间相似度的整体分布情况

除了比较不同相似度的分布差异，我们也考察了现代汉字部件不同相似度之间以及相似度与部件构字数之间的相关性，结果如表 3 所示，可以看出三种相似度两两之间都呈现显著的正相关。部件-字相似度与字间相似度的正相关反映出：在整体趋势下，部件与构成汉字的意义越接近，其构成的字集合中的汉字之间的字义关联也更强。此外，相似度与构字数之间的相关性显现了很有意思的结果：在全体部件上，总相似度、部件-字相似度分别与构字数负相关，也就是说构字数少的部件其表义能力和部件-字相似度也就越小；而字间相似度则与构字数无显著相关关系。

表 3　总相似度、部件-字相似度、字间相似度与构字数的两两相关性

	总相似度	部件-字相似度	字间相似度	构字数
总相似度	1	**0.836****	0.860**	−0.179**
部件-字相似度	**0.836****	1	0.440**	−0.322**
字间相似度	**0.860****	**0.440****	1	0.007
构字数	**−0.179****	**−0.322****	**0.007**	1

注：**表示在 0.01 水平（双侧）上显著相关。

值得注意的是，当我们按 3.1 节中的方法将部件按构字能力分成三级时，发现构字能力强的部件和其他部件有不一样的表现：其总相似度、字间相似度与构字数正相关，部件-字相似度与构字数无显著相关性；构字能力强的部件相似度与构字数的相关性与总体一致（表 4）。

表 4　不同构字能力下构字数与总相似度、字间相似度、部件-字相似度的相关性

	总相似度	字间相似度	部件-字相似度
构字能力强	**0.182***	**0.216***	**0.102**
构字能力中	−0.152**	0.040	−0.352**
构字能力弱	−0.160**	−0.053	−0.287**
所有部件	−0.179**	0.007	−0.322**

注：① **表示在 0.01 水平（双侧）上显著相关。
② * 表示在 0.05 水平（双侧）上显著相关。

这也就反映了当部件的构字能力强时，其总体表义能力与构字数呈现正相关，尤其是字间相似度与构字数的相关系数较高；当部件的构字能力不强时，其总体表义能力与构字数呈现负相关。这背后的原因可能是，当部件只能构成很少数量的汉字时，字与字之间的构字数相差不大，因而构字数越多，其意义更可能分散，造成部件与字、字与字之间的相似度较低；但高构字能力的部件可能反映了古人造字时的倾向性：部件的构字数很多时，说明该部件在构字时更经常被使用，选择意义辨识度更高、更容易被理解的部件进行构字更符合人类的认知。

3.4　结合构字能力的部件表义能力分级

综合上述分析，由于汉字部件的构字能力相差悬殊，并且在不同的构字能力等级下汉字部件的表义能力和在语料中的频次分布差异性明显，因而我们认为，对部件表义能力进行分级引入汉字构字能力的区分是有必要的。

结合3.1节构字能力的分析结果，我们在不同的构字能力等级下分别以按2∶3∶5划分部件表义能力强、中、弱三个等级，比如说构字能力强的126个部件，按表义能力前20%的25个汉字为强表义能力部件，排名20%～50%的38个部件为中表义部件，后50%的部件为弱表义部件。最后，如表5所示，我们将汉字部件按表义能力划分成了3个大类，结合部件构字能力的等级划分形成了9个小类。

表5 汉字部件表义能力分级结果

	构字能力	部件个数	示例
表义能力强	强	25	病，鸟，鱼
	中	38	舟，革，豕
	弱	63	买，男，糕
	总计	126	/
表义能力中	强	60	石，月，日
	中	91	生，亡，束
	弱	151	三，竟，杀
	总计	301	/
表义能力弱	强	76	刂，八，页
	中	113	乔，化，夹
	弱	189	丐，总，或
	总计	378	/

4 部件表义能力在汉语教学中的应用

汉字教学是对外汉语教学的重要组成部分。部件教学法也就是利用汉字形体结构理据进行汉字教学的方法，是对外汉语教学的教学法之一。然而，部件教学法虽然受到学界的大力倡导，但多数研究成果只在理论层面取得了发展，却难以在实际应用中直接进行转换[21]。汉字部件教学体系设计、范围界定、顺序安排仍是部件教学法面临的难题。其中，部件的表义性是帮助汉字理解的关键突破口。表义部件也是最早应用于汉字教学的，是部件教学法重要的组成部分。

针对部件教学法面临的问题，我们认为若引入上述部件表义测量和分级情况，也许能推动部件教学法的实际应用。首先在部件教学的体系设计上，并非所有部件都适用于部件教学法，因而在教学时需要确定适合采用该方法的部件范围。同时，针对不同表义能力的部件，应当采取不同的教学策略，以免学习者误用部件表义性产生错误的理解。此外，我们认为表义部件是对字义的提示，因而将近义关系引入部件教学法中，通过将部件下意义相近的字一起进行教学，可以加深学生对部件和汉字的认识。

就教学范围而言，汉字部件在构字能力和表义能力两个维度上的分级能够帮助我们确立部件教学的范围。对于构字能力不强的部件，由于一个部件构成的字数太少，使用部件教学法的必要性不大，反而会增加老师和学生的负担。对于表义能力太弱的部件，使用部件教学法来进行汉字教学对意义理解帮助也不大。在实施部件教学法时，教师可以主要聚焦于高构字能力、高表义能力的部件，同时适当关注高构字能力中表义能力的部件。

对于高表义能力的部件，我们可以利用其表义能力和汉字的近义关系来安排汉字教学顺序。目前的汉字教学的一大依据是字频，在部件表义能力的指导下，我们认为具有高表义能力部件的汉字可以与同部件的其他汉字一起进行教学，比如说在教“河”一字时，可以联系“江”“湖”“海”等同部件的字。值得注意的是，对于高表义能力的部件，其构成的汉字集合大部分与部件意义相关；而表义能力越弱，构字集合中与部件意义关联少的汉字比例就越高，并非该部件在所有构字集合中的汉字中都有很强的意义指示性。比如说“演”“派”等字与“氵”意义关联度便不强。针对构字集合中部件和汉字意义关联度的差异，区分表义部件是否在汉字中有意义指示作用便很有必要。因而我们引入了词义相似度来决定：同部件的汉字中，哪些字可以一起教学，哪些字不适合一起教学。

对于汉字教学顺序的编排，我们主要参照了国家语委发布的现代汉语语料库字频表①，将字频更高的常用字教学顺序安排靠前，非常用字靠后。同时当汉字中含有高表义部件时，我们将计算该目标字与同部件其他汉字的语义相似度，若同部件汉字与其相似度大于阈值，则将同部件汉字与目标字一起教学，相当于将同部件汉字的教学顺序予以提前；如果字集合中没有字与目标字达到相似度的阈值，说明目标字中部件的表义性不强，则不对该字采用部件教学法。我们对现代汉语语料库字频表中字频排名前3 500个汉字实施了上述操作，最终为759个汉字找到了可以共同教学的同部件字。表6中是

① http://corpus.zhonghuayuwen.org/Resources.aspx

部分汉字与它们可以一起教学的同部件汉字示例。不难发现，表中可以一起教学的同部件字和目标汉字的意义比较接近，并且部件在这些字中的表义功能基本一致，因而这些汉字的共同教学将对汉字习得起到一定的促进作用。

表 6　汉字与可一起教学的同部件汉字示例

汉字	可一起教学的同部件字
心	意，爱，慧，悲，忍，惑，忿，愿，虑……
山	岭，峰，涯，岩，岗，峡，屿，峦，巅，……
光	荧，灿，炬，焰，焕，炫，灼，煌，……
神	祈，祀，祖，祷，禄，禅，礼，……
痒	痛，疼，疮，疹，癣，痘，痰，病，……

综上，部件教学法是对外汉语教学中的重要教学方法之一，但以往的教学缺乏定量的数据分析。部件表义度的测量能为确定适用部件教学法的部件范围以及汉字教学顺序提供一定的参考，促进部件教学法的良好发展。

5　结论

本文基于分布式表示模型，采用将部件与字词共同进行向量表示的多粒度训练方法，通过部件-字相似度和字间相似度两项指标计算现代汉字部件的表义能力，并结合部件的构字能力进行了部件等级的划分。同时，本文提出将部件表义能力测量的结果应用于对外汉语教学，以确定部件教学法应涉及的部件范围以及结合语义相似度调整汉字学习顺序。在对现代汉字部件表义能力的测量中，我们发现现代汉字部件具有一定的表义能力，但汉字整体表义水平不高；同时部件表义能力与构字能力关系密切，当汉字的构字能力强时，汉字的表义能力与构字能力正相关，而汉字的构字能力不强时，两者则呈负相关关系。

参考文献

[1] 王宁.系统论与汉字构形学的创建[J].暨南学报(哲学社会科学)，2000，64(02)：15-21.

[2] 苏培成.现代汉字学纲要(增订本)[M].北京：北京大学出版社，2001：74.

[3] 林柏松(Patrick Lin)，周健.外国人汉字速成[M].北京：华语教学出版社，1996：2.

[4] 胡韧奋，曹冰，杜健一.现代汉字形声字声符在普通话中的表音度测查[J].中文信息学报，2013，27(03)：41-47.

[5] 李丽.古文字意符演变研究[D].重庆：西南大学硕士学位论文，2012.

[6] 张莹莹.会意字意符的认知功能分析[J].东南学术，2017，29(01)：238-245.

[7] 崔永华.关于汉字教学的一种思路[J].北京大学学报(哲学社会科学版)，1998，43(03)：113-117.

[8] 施正宇.现代形声字形符表义功能分析[J].语言文字应用，1992，1(04)：76-83.

[9] 李蕊.对外汉语教学中的形声字表义状况分析[J].语言文字应用，2005，2(02)：104-110.

[10] 吕菲. 现代形声字意符表义研究[D].北京：中央民族大学硕士学位论文，2012.

[11] 陈爱华. 汉语国际教育等级汉字意符表义状况及教学研究[D].合肥：安徽大学硕士学位论文，2017.

[12] Mikolov T，Sutskever I，Chen K，et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 26th Internatial Conference on Neural Information Processing Systems，2013：3111-3119.

[13] Firth J R. A synopsis of linguistic theory 1930-1955.[J]. Studies in Linguistic Analysis，1957，41(4)：1-32.

[14] Chen X，Xu L，Liu Z，et al. Joint learning of character and word embeddings[C]//Proceedings of the 24th International Conference on Artificial Intelligence. AAAI Press，2015：1236-1242.

[15] Sun Y，Lin L，Tang D，et al. Radical-enhanced Chinese character embedding[C]//Proceedings of the 2014 International Conference on Neural Information Processing. Springer International Publishing，2014：279-286.

[16] Yin R，Wang Q，Li P，et al. Multi-granularity Chinese word embedding[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing，2016：981-986.

[17] Tzu-Ray Su，Hung-Yi Lee. Learning Chinese word representations from glyphs of characters[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Springer International Publishing，2017：264-273.

[18] Zhao Z，Liu T，Li S，et al. Ngram2vec：Learning improved word representations from ngram co-occurrence statistics[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing，2017：244-253.

[19] Shen Li，Zhe Zhao，Renfen Hu，et al. Analogical reasoning on Chinese morphological and semantic relations[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics，2018：138-143.

[20] 李宝贵.汉字理据性与对外汉字教学[J].汉字文化,2005,16(01):41-43.

[21] 李明.常用汉字部件分析与对外汉字教学研究[D].北京:北京语言大学硕士学位论文,2006.

梁诗尘(1996—),硕士研究生,主要研究领域为计算语言学。
E-mail:shichen@mail.bnu.edu.cn

胡韧奋(1988—),博士,讲师,主要研究领域为计算语言学。
E-mail:irishu@mail.bnu.edu.cn

唐雪梅(1995—),硕士研究生,主要研究领域为计算语言学。
E-mail:tangxuemei@mail.bnu.edu.cn

(上接第 16 页)

[4] 吴丹.留学生汉语褒贬义词习得研究[D].上海:上海师范大学硕士学位论文,2015.

[5] 莫丹.美国学习者汉语写作产出性词汇知识的发展路径考察[J].海外华文教育,2017(5):579-605.

[6] 徐琳宏,林鸿飞,潘宇,等.情感词汇本体的构造[J].情报学报,2008(2):180-185.

[7] 郑毅.基于情感词典的中文微博情感分析研究[D].广州:中山大学硕士学位论文,2014.

[8] 刘海涛.计量语言学导论[M].北京:商务印书馆,2017.

[9] 莫丹.华裔与非华裔汉语学习者产出性词汇知识差异及其对写作质量的影响[J].云南师范大学学报(对外汉语教学与研究版),2015,(5):33-41.

张易扬(1994—),硕士研究生,主要研究领域为汉语国际教育。
E-mail:bournezyy@163.com

吴迪(1994—),硕士研究生,主要研究领域为汉语国际教育。
E-mail:wudi098@qq.com

王治敏(1972—),通信作者,教授,博士生导师,主要研究领域为计算语言学、汉语国际教育。
E-mail:wangzm000@qq.com