

一种基于词语抽象度的汉语隐喻识别方法^{*}

黄孝喜^{1,2} 张华¹ 陆蓓¹ 王荣波¹ 吴铤¹

¹(杭州电子科技大学 认知与智能计算研究所, 杭州 310018)

²(浙江大学 语言与认知研究中心, 杭州 310028)

摘要：【目的】设计一种自动计算汉语词语抽象度的方法，并将其用在自然语言理解中的隐喻识别任务。【方法】以统计学习理论中逻辑回归为计算模型，把神经网络语言模型获取的词语词向量作为特征，通过构建的抽象词库得到的特征权重向量来计算汉语词语抽象度。提出一种基于词语抽象度的汉语隐喻识别算法，验证该方法的应用效果。【结果】通过与已有的方法进行实验对比，结果表明本文设计的汉语词语抽象度计算方法更接近于人的认知常识；并且在隐喻识别任务中，也体现了更好的准确率。【局限】一方面词语词向量来表示词语抽象程度有一些缺陷，另一方面抽象词语库的规模影响了特征权重向量的学习。【结论】词语抽象度计算可以表现为人对概念的一种抽象分类能力，本文提出的汉语词语抽象度计算方法得到的结果能够较好地拟合人的认知，并且实验证明词语抽象度可有效提高隐喻识别的效果。

关键词：抽象度 神经网络语言模型 隐喻识别

分类号：TP391

An approach to Chinese metaphor identification based on word abstractness

Huang Xiaoxi^{1,2} Zhang Hua¹ Lu Bei¹ Wang Rongbo¹ Wu Ting¹

¹(Institution of Cognitive and Intelligent Computing, Hangzhou Dianzi University, Hangzhou 310018, China)

²(Center for the Study of Language and Cognition, Zhejiang University, Hangzhou 310028, China)

Abstract: [Objective] Design a method of computing the abstractness of Chinese word, and introduce word abstractness into metaphor recognition task in natural language understanding. [Methods] The word abstractness is computed by logistic regression model, where the features are the word vectors computed by neural network model and the weighted feature vectors come from the self-constructed dictionary. A metaphor recognition algorithm based on word abstractness is proposed to demonstrate the validity of this approach. [Results] By comparing with the existing methods of word abstractness computing, the method put forth in this paper is in better accordance with human's intuition and is an effective method of performing Chinese metaphor identification. [Limitations] One is that words vector in representing the term abstract level is defective. The other is that the size of the abstract words affects the learning of weighting vector. [Conclusions] Chinese words abstract computed by our method is better fitting the human cognition, and the experimental result shows that word abstractness can improve the effect of metaphor recognition.

Key words: Word abstractness; Metaphor identification; Neural network language model

^{*}本文系国家自然科学基金项目“引入涉身认知机制的汉语隐喻计算模型及其实现”(项目编号：61103101)、国家自然科学基金项目“基于马尔科夫树与 DRT 的汉语句群自动划分算法研究”(项目编号：61202281)、教育部人文社会科学研究项目“面向信息处理的汉语隐喻研究”(项目编号：10YJCZH052)和浙江省自然科学基金“全同态公钥密码体制的设计与应用研究”(项目编号：LY12F02006)的研究成果之一。

1 引言

对于每个词语来说,都有很多属性,比如词性、熟悉程度、抽象程度、词汇类别等。对于建立更好的语言理解理论和开发模拟人类认知系统来说,词语的相关属性是非常重要的,其中词语抽象程度这一属性是备受众多研究领域学者所关注。抽象词语存在于人们的意识之中,是人们对客观世界认识的过程或结果在大脑中抽象的反映,如“思想”、“感情”等;相反,具体词语指的是客观世界中所存在的,能够看得见摸得着的物体,如“苹果”、“骏马”等。词语抽象度就是指能够反映词语的抽象和具体程度的数值。

在文本衔接与语言的文本分析系统 Coh-Metrix^[1]中,词语抽象度是非常重要的指标。McCarthy 等^[2]人在识别句子的主题时也加入了词语抽象度。Feng 等^[3]人总结了词语抽象度在自然语言处理中相关应用,并利用词汇类型、潜在语义分析的维度、词语上下位理、语音领域和计算语言学都扮演着比较重要的角色。在 Graesser 等人开发的关系设计了一种自动化计算词语抽象度的方法。

随着认知领域的发展,隐喻相关研究越来越热。隐喻识别的相关研究已经不仅仅关注与词语类别信息,而是通过更多的特征信息来获取隐喻的表示,比如说词语抽象度、词语成像性等。Gargett^[4]在隐喻识别中加入了词语抽象度、词语成像性这两个特征。Turney^[5]在英语隐喻过程识别中,提出了一种基于词语相似度自动计算词语抽象度的方法。因此,汉语词语抽象度的自动化计算工作已经越来越重要了。

目前针对汉语词语的抽象度研究工作都是集中在对它的语法特性进行深入描述上。比如,龙涛^[6]研究了抽象名词的语法意义,探讨了抽象名词的意义对它在语法形式和用法上的作用。官杨^[7]通过研究程度副词的修饰名词形式,发现“程度副词+抽象名词”是汉语中经常出现的一种结构,并且解析了这一结构与其他不同点。鲁晓雁^[8]对《小说月报》1998 年第 2 期中的抽象名词搭配结果进行了调查,其中抽象名词作宾语的句子占到了 72%。她把与抽象名词构成动宾搭配的动词分为存现动词、心理动词、表述性动词以及其他脱离具体行为的动词等。其中,存现动词中的“有”和“是”与抽象名词搭配的情况占到了 36%,所占比重较大。但是,对于汉语词语抽象度的量化工作目前还比较少见。在自然语言处理领域,赵红霞^[9]曾利用词语抽象度来进行隐喻识别。在她的硕士论文中,作者利用 Hownet 词库,利用人工进行分类,分别将词语分为抽象和具体,分别标记为 0 和 1,作为词语的抽象度。

鉴于目前汉语词语抽象度量方法研究的不足,本文提出一种自动计算汉语词语抽象度的方法,并把该词语抽象度用于汉语的隐喻识别中。本文剩余部分安排如下:第 2 节提出一种汉语词语抽象度的计算方法,并构造出相应的汉语抽象词库,通过设计实验来验证其有效性;第 3 节利用抽象词库,针对指称型模式的句子,给出了一种基于词语抽象度的隐喻识别方法,并与其他同类研究做比较;最后对论文做总结。

2 汉语词语抽象度计算

统计学习理论在自然语言处理中发挥着越来越重要的作用。在汉语词语抽象度的计算方面，我们可以利用词语在抽象度方面的相关特征，使用统计学习理论中的相关算法对词语抽象度进行量化计算。在统计学习理论里，判别式模型中的逻辑回归算法使用的逻辑函数是一个良好的阈值函数，具有连续、光滑、严格单调等特性，能够把样本的特征信息投影到(0,1)区间内，投影后的值可以用来估计某种事物的可能性。我们把计算词语抽象度的工作看成是分类问题，投影后的值也就是词语为抽象的可能性，因此这里可以利用逻辑回归算法进行词语抽象度计算。

在计算词语抽象度时，首先通过逻辑回归算法构建出计算词语抽象度的具体计算公式；其次，在互联网语料库下，利用神经网络语言模型得到语料库包含词语的词向量；再次，构建抽象词库，把词向量作为特征，在抽象词库的基础上计算抽象度计算公式的权重向量；最后通过词语的词向量和计算得到的权重向量对大规模的词语进行抽象度计算。

2.1 汉语词语抽象度的计算模型

逻辑回归是统计学习中一种十分常用的分类算法，常用于计算某种事件发生的条件概率。比如某用户购买某商品的概率，以及某广告被用户点击的概率等。根据人类的认知，判断一个词语为抽象的概率一般从几个方面入手：视、听、说、闻、感受、触感等。显然，有了这些特征，这里可以就利用此算法得到该词语为抽象的概率，即为该词语的抽象度。

逻辑回归的模型是用条件概率 $P(Y/X)$ 表示，形式为参数化的逻辑斯蒂分布。逻辑斯蒂分布的分布函数的图形是一条 S 性曲线，该曲线以点(0, 0.5)为中心对称，曲线在中心附近增长速度较快，在两端增长速度较慢。

条件概率 $P(Y/X)$ 如公式 1 所示：

$$P(Y=0|X)=\frac{1}{1+e^{wx}} \quad (1)$$

其中， $Y=0$ 表示该词语为抽象词语，随机变量 X 表示该词语所包含的代表词语抽象程度的特征信息， W 表示每一维特征在计算词语抽象度时的影响力。该式子表示该词语在已有特征向量 X 下为抽象词语的概率，也就是词语的抽象度。根据认知规律可以了解到，抽象词语如意识、思想，它们的抽象度应该接近 0，并且抽象度之间相差不大。同样的具体词语应该接近于 1。而根据逻辑斯蒂分布的特性可知，抽象的词语大部分会聚集在 $Y=0$ 端，而具体的词语大部分会聚集在 $Y=1$ 端，符合人类的认知规律，该模型能够较好的拟合抽象度这个概念。

通过上面描述可知，在计算抽象度的时候，需要特征向量 X 及特征权重向量 W 。因此找到能够表示词语抽象度的特征向量 X 以及构建合理的训练数据就是词语抽象度计算模型的关键问题。

2.2 基于神经网络语言模型的特征词向量

一个词语抽象程度，在词语的视、听、说、闻、感受、触感等都可以得到反馈。而具体名词和抽象名词的区别除了意义上的界定之外，还体现在很多语法现象中。杨玉玲^[10]认为抽象名词和具体名词在与副词搭配时，会有明显的区别。比如可以说“他很有野心”，而“他很有书”则不通顺。具体名词和抽象名词的差别还体现在特定的量词和名词的组合时。绝大部分抽象名词不能与量词形成搭配，能够搭配的抽象名词可以搭配的量词数量也比较少。而且，多数情况下，与抽象名词搭配的量词之前仅能出现“一”这个数词。如：一种学说，一些理论、一个建议等。

通过上文对词语抽象与具体在语义上区别的讨论，可以获悉文本的上下文信息能够用来表示词语抽象度的特征信息 X 。

神经网络语言模型由 Bengio^[11]在 2003 年提出。该模型的目标是根据前 $n-1$ 个词语 $w_{t-n+1}, \dots, w_{t-2}, w_{t-1}$ 预测下一个词语 w_t 出现的概率，在优化该目标时，该模型能够充分利用上下文信息，从而得到能够客观反映词语的形态、句法、语义和语用等特点的词向量。因此这里把该词向量作为计算词语抽象度的特征信息 X 。

目前，使用较为广泛的神经网络语言模型工具包是由 Google 的研究人员与 2013 年发布的 word2vec¹。我们使用搜狗实验室提供 2012 年 6 月至 7 月期间搜狐新闻数据作为训练语料，共 1.65GB，使用 word2vec 将输入的词语转化为 200 维的词向量，作为计算抽象度的特征向量 X 。

2.3 特征权重向量的计算

根据词语抽象度的计算模型可知，在我们已经得到了词语所包含的特征向量 X 的情况下，还需要词语特征向量对应的特征权重向量 w 。

在逻辑回归模型中，寻找一个合适的特征权重向量 w ，就是要使得对于训练数据中为具体词语的计算所得条件概率(公式 1)尽量小；而对于训练数据中的抽象词语来说，就是让公式 1 得到的值尽量大。用联合概率来表示也就是公式 2 所示的形式。

$$\max P(w) = \prod_{j=1}^M \frac{1}{1 + e^{-y_j w x}} \quad (2)$$

寻找合适的特征权重向量 w 就可以转化为使公式 2 中目标函数 $P(w)$ 最小的 w ，这是一个无约束最优化问题，经过对数转化后使用梯度下降迭代算法解决这个问题。

通过上面分析可知，构建合理的抽象词库对于我们所要计算的特征权重向量具有十分重要的作用。心理语言学数据库 MRC Psycholinguistic Database^[12]中，标注过词语抽象度的词语一共有 8004 个，取值分别为 100-700 之间。但是对于汉语来说，至今没有针对词语抽象度这类隐含属性的词典。我们需要构建一个抽象词库，作为词语抽象度计算的训练语料库，用来计算词语抽象度时用到得到特征权重向量 w 。

¹ <https://code.google.com/p/word2vec/>

本文构建了一个共 2091 个词语的汉语抽象词库，其中包含具体词语 1057 个和抽象词语 1034 个。在此抽象词库的基础上，我们可以通过迭代计算，从而得到词语抽象度所需要的特征权重向量 w 。

得到的特征权重向量是否能够较好的表达词语抽象和具体的属性，需要使用根据神经网络语言模型得到的词向量为特征，在构建的抽象词库上进行测试验证，通过结果进行判断。在利用逻辑回归算法进行预测词语抽象与具体时，得到的平均准确度为 94%。召回率的均值为 84.31%，平均的 F 值则为 88.70%。从实验结果上进行分析可知，在此抽象词库的分类实验中，准确度比较高，在测试集上能够比较正确的预测结果，此模型在计算词语抽象度方面是可行的。因此，词向量能够较好地描述词语抽象程度，可以使用该抽象词库得到的特征权重向量 w 来计算词语抽象度。

据此我们已经得到计算词语抽象度的特征向量和特征权重向量，通过这些变量我们就可以计算出词语抽象度。

3 基于词语抽象度的汉语隐喻识别方法

概念隐喻理论^[13]认为“隐喻通常以具体的事物来表示抽象的概念”，因此我们可以通过隐喻本体和喻体之间的抽象与具体程度来识别隐喻。基于这一思想，本节针对指称型模式的句子给出一种基于词语抽象度的隐喻识别方法。指称型模式的句子包含源域和目标域可能出现的部分——主语和宾语（本文统称为“特征词语”），根据它们的抽象度可以对句子是否为隐喻进行判断。隐喻识别的具体流程如下图 1 所示：首先，对语料进行词性标注和依存分析等预处理，使用条件随机场模型对特征词语进行提取；其次，计算特征词语的抽象度和它们之间的相似度；最后，使用支持向量机模型对隐喻进行识别。下面结合“沉默是金”的例子具体阐述该算法。

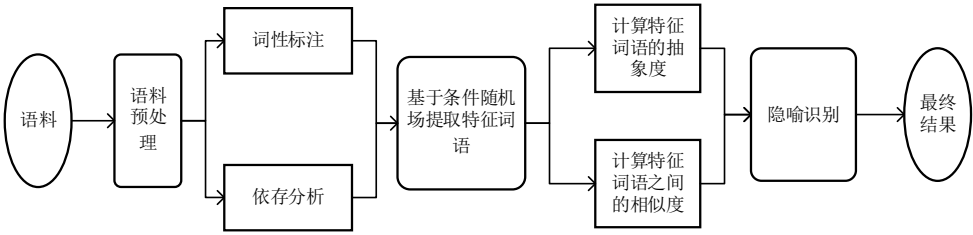


图 1 基于词语抽象度的汉语隐喻识别算法流程

3.1 基于条件随机场的特征词语提取

条件随机场^[14]是由 Lafferty 第一次提出的，它是一个在给定输入节点（观察值）条件下，计算输出节点的条件概率的模型，常用来标注和分析序列资料。特征词语的提取可转化为序列标注的问题，即把源域和目标域分别标注出来，因此可以使用条件随机场模型进行提取。

特征选取是条件随机场模型的关键问题。根据对句法结构和句意信息的分析，我们主要考虑四个方面：(1)词语。(2)词语词性标记。(3)依存关系。(4)窗口

的大小。

在经过多次试验发现，采用窗口为 7，包含词语、词性标记、依存关系的特征，提取特征词语时可以达到较出色的效果，准确度达到了 86.9%。因此，“沉默是金”经过提取之后得到“沉默”和“金”，为下面的隐喻识别提供了基础。

3.2 基于词语抽象度的隐喻识别方法

隐喻识别问题可看作一个分类问题。支持向量机模型^[15]能根据有限的样本信息在模型的复杂性（即对特定训练样本的学习精度）和学习能力（即无错误地识别任意样本的能力）之间寻求最佳折中，以获得最好的推广能力。隐喻识别过程中使用的特征比较少，适合使用非线性支持向量机的方法进行隐喻识别。

我们选取了以下两个度量作为隐喻识别的特征：

(1)特征词语的抽象度。根据 Lakoff 的概念隐喻理论，人们往往会在隐喻中用具体的事物来表示抽象的概念。抽象度可以给出特征词语的抽象与具体的度量，用来表示特征词语更倾向具体事物还是抽象概念，例如沉默的抽象度为 0.000133，而金的抽象度则为 0.617081，两者相差较大，可以作为隐喻识别的依据。

(2)特征词语之间的语义相似度。隐喻一个重要特征就是源域和目标域之间存在明显差异，而完全相同的事物是不可能构成隐喻的。通过余弦相似计算，沉默和金的相似度为 0.008939，相似度较低，具备成为隐喻的基本条件。

4 实验设计与实验结果分析

4.1 词语抽象度的评测

为了对抽象度的合理性进行对比验证，我们在该抽象词库上同时按照 Turney 的算法思想，用相似度的方法来计算词语的抽象度。具体算法如下：计算该词语与抽象词库中的抽象词语之间的相似度之和 s_a ，以及该词语与抽象词库中的具体词语之间的相似度之和 s_c ，把 s_a 和 s_c 相减之后经线性归一处理得到的值作为该词语的抽象度。

由于抽象度是一个较为主观的概念，为了使算法的对比结果更为客观，本文采用问卷调查的形式进行测评。

(1)问卷调查内容：从同义词词林（扩展版）中随机选出 200 个词语样本作为问卷调查的主体。每条样本的样式形如(w , value1 , value2 , classification)，其中 w 表示待判定的词语；value1 和 value2 分别代表两种不同方法得到的词语 w 的抽象度，其中 value1 表示逻辑回归算法得到的抽象度，value2 表示相似度算法得到的抽象度；classification 为调查对象对抽象度 value1 和 value2 合理性的判定结果。

(2)问卷调查对象：我们邀请五位有计算语言学背景的研究生，在不被告知 value1 和 value2 所对应的具体算法的情况下对样本进行评判，选出他们认为更加合理的词语抽象度。

调查结果：在 200 个词语样本中，认为逻辑回归算法所计算得到的抽象度更

加合理的词语样本共有 171 个，占 85.5%，认为基于词语相似度的方法更加合理的词语样本共有 20 个，比例为 10.0%，而其它 4.5%的词语样本被认为两种算法均不够合理。

通过对这 200 个词语样本的抽象度结果进行分析，我们发现一些特定名词的抽象度在使用这两种方法时都不能得到较好的结果，比如“金字塔”计算的结果分别为 0.447684 和 0.457421，“马克思”计算的结果分别为 0.275683 和 0.214427。而对于大部分词语的抽象度计算中，基于逻辑回归的算法都展现出了良好的效果。因此，基于问卷调查的结果，基于逻辑回归的词语抽象度算法在汉语词语抽象度计算方面与已有的词语抽象度算法相比展现出了巨大的优势。

4.2 隐喻识别实验设计与分析

(1) 实验语料

我们通过国家语委语料库²以及北京大学中国语言学研究中心的语料库检索系统³搜索指称型模式的句子，从中选取 1000 句作为本次实验的语料库，其中包含非隐喻句子 564 句，隐喻句子 436 句。为了得到更加可靠的结果，采用十折交叉验证。

(2) 隐喻识别实验结果

为了验证本文提出的基于抽象度的隐喻识别算法有效性，我们设计了两个对比实验。

实验 1：在实验语料中对提取到的特征词语使用本文提出的抽象度计算方法获取其抽象度，在此基础上使用本节提出的隐喻识别方法进行隐喻识别。

实验 2：在实验语料中对提取到的特征词语使用词语相似度的抽象度计算方法获取其抽象度，同样使用上述的隐喻识别方法进行隐喻识别。

由于提取的结果中存在无法得到词向量的词语，这些词语无法计算词语抽象度，也即无法进行隐喻识别，因此在我们首先在测试语料中进行隐喻识别，其次在过滤掉没有正确提取到特征词语的测试语料下进行识别。这里应用的软件工作主要有哈工大的语言云系统(LTP)^[16]、条件随机场模型 CRF++⁴和支持向量机模型工具包 libSVM⁵。把隐喻识别实验语料分为 10 份，每次把其中 9 份作为训练语料，其余 1 份作为测试语料，经过十次实验得到的结果如下表 1 和表 2：

表 1 隐喻识别结果对比

实验	F 值	召回率	准确率
1	0.5553922	0.5238288	0.600
2	0.52101886	0.48352012	0.578

表 2 隐喻识别结果对比（过滤没有得到特征的句子）

实验	F 值	召回率	准确率
1	0.629814	0.6007932	0.671
2	0.593536	0.5539257	0.646

² <http://www.cncorpus.org/CCindex.aspx>

³ http://ccl.pku.edu.cn:8080/ccl_corpus/index.jsp?dir=xiandai

⁴ <http://crfpp.googlecode.com/svn/trunk/doc/index.html>

⁵ <http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html>

(4) 实验分析

从表 1 和表 2 的实验数据分析可知,相比实验 2,本文设计的算法在准确率、召回率和 F 值上都有提升。这在一定程度上反应了逻辑回归算法得到的词语抽象度相比词语相似度算法得到的词语抽象度在隐喻识别中能够起到更好的分类效果,再一次在隐喻识别的角度上验证了使用逻辑回归算法的抽象度计算方法能够更加合理地衡量词语的抽象程度。

另一方面,从表中数据可以看出,经过过滤的实验结果优于没有过滤的结果。因此,在提取特征词语的准确度方面还需要提升,词语词向量的训练语料还需要扩大,以提高算法的准确度。

5 结论

词语抽象度作为一种基础性的研究工作,是认知科学与计算语言学发展的基础。本文提出了一种自动计算的汉语词语抽象度,并用它对指称型句子进行隐喻识别,根据实验结果得到了如下结论:(1)基于逻辑回归算法得到的汉语词语抽象度相比已有算法更加符合人类的认知常识。(2)在隐喻识别中,使用基于逻辑回归算法的词语抽象度能够对汉语句子的隐喻识别起到一定的作用。

但后续仍需在以下方面进行改进:(1)扩大抽象词库,对词语的抽象度进行更好的学习,以达到更好的效果。(2)考虑其它更多句子类型的隐喻识别以及认知科学的其他应用领域,验证汉语词语抽象度的效果。

参考文献

- [1] Graesser A C, McNamara D S, Louwerse M M, et al. Coh-Metrix: Analysis of text on cohesion and language[J]. Behavior Research Methods, Instruments, and Computers, 2004,36(2):193-202.
- [2] McCarthy P M, Renner A M, Duncan M G, et al. Identifying topic sentencehood[J]. Behavior Research Methods, 2008,40(3):647-664.
- [3] Feng S, Cai Z, Crossley S A, et al. Simulating Human Ratings on Word Concreteness[C]. In: Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference, 2011:245-250.
- [4] Gargett A, Ruppenhofer J, Barnden J. Dimensions of Metaphorical Meaning[C]. In: Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon, 2014:166-173.
- [5] Turney P D, Neuman Y, Assaf D, et al. Literal and metaphorical sense identification through concrete and abstract context[C]. Proceedings of the 2011 Conference on the Empirical Methods in Natural Language Processing, 2011:680-690.
- [6] 龙涛. 抽象名词的隐喻性 “有界” 空间范畴义[J]. 武汉大学学报: 人文科学版, 2011,64(4):112-117. (Long Tao. Metaphorical "Bounded" space category meaning of abstract nouns [J]. Wuhan University Journal (Humanity Sciences), 2011, 64(4):112-117.)
- [7] 官杨. 程度副词修饰名词浅析[J]. 安徽文学 (下半月), 2008,12:199. (Guan Yang. Analysis of the nouns modified by adverbs[J]. Anhui Literature (The second half of month), 2008,12:199.)
- [8] 鲁晓雁. 抽象名词语义搭配情况调查 (之一)[J]. 学术交流, 2002(2):109-112. (Lu Xiaoyan.

- Abstract semantic collocation situations investigated (part one) [J]. Academic Exchange, 2002(2):109-112.)
- [9] 赵红艳. 基于语义知识的动词隐喻识别与应用[D]. 南京: 南京师范大学, 2012. (Zhao Hongyan. Chinese Verb Metaphor Recognition and Application Based on Semantic Knowledge[D]. Nanjing: Nanjing Normal University, 2012.)
- [10] 杨玉玲. 认知凸显性和带“有”的相关格式[J]. 修辞学习, 2007(05):31-34. (Yang Yuling. Cognitive salience and the contained "have" sentence format[J]. Contemporary Rhetoric, 2007(05):31-34.)
- [11] Yoshua B, Rejean D, Pascal V, et al. A Neural Probabilistic Language Model[J]. Journal of Machine Learning Research, 2003,0(3):1137-1155.
- [12] Coltheart M. The MRC psycholinguistic database[J]. The Quarterly Journal of Experimental Psychology, 1981,33(4):497-505.
- [13] Lakoff G, Johnson M. Metaphors We Live By[M]. Chicago: The University of Chicago Press, 1980.
- [14] Lafferty J, McCallum A, Pereira F C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]. In: Proceedings of the Eighteenth International Conference on Machine Learning, 2001:282-289
- [15] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods[M]. Cambridge: Cambridge university press, 2000.
- [16] 刘挺, 车万翔, 李正华. 语言技术平台[J]. 中文信息学报, 2011(06):53-62. (Liu Ting, Che Wanxiang, Li Zhenghua. Language Technology Platform[J]. Journal of Chinese Information Processing, 2011(06):53-62.)

(通讯作者:黄孝喜 Email: huangxx@hdu.edu.cn)

作者贡献声明

黄孝喜, 张华, 王荣波: 提出研究思路, 设计研究方案;
张华, 黄孝喜: 负责进行实验;
张华, 黄孝喜, 吴铤: 负责采集、清洗和分析数据;
黄孝喜, 张华, 陆蓓: 负责论文起草;
黄孝喜: 负责最终版本修订。