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，通过构建抽象词库得到特征权重向量，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，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，也体现出更好的准确率。【局限】词语词向量表示词语抽象程度有一些缺陷；抽象词语库的规模影响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: 词语抽象度 神经网络语言模型 隐喻识别
分类号: TP391

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备受众多研究领域学者关注。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程度的数值。

在 Graesser 等[1]开发的文本衔接与语言的文本分析系统 Coh-Metrix 中，词语抽象度是非常重要的指标。McCarthy 等[2]在识别句子的主题时也加入了词语抽象度。Feng 等[3]总结了词语抽象度在自然语言处理中的相关应用，并利用词汇类型、潜在语义分析的维度、词语上下位关系设计了自动计算词语抽象度的方法。由此可知，词语抽象度在计算语言学中扮演着比较重要的角色。

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词语类别信息，而是通过更多的特征信息获取隐喻的表示，比如词语抽象度、词语成像性等。Gargett等[4]在隐喻识别中加入词语抽象度、词语成像性这两个特征。Turney等[5]在英语隐喻过程识别中，提出一种基于词语相似度的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越来越重要。

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]研究抽象名词的语法意义，探讨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

通讯作者: 黄孝喜，ORCID: 0000-0003-4483-3664，E-mail: huangxx@hdu.edu.cn。

*本文系国家自然科学基金青年基金项目“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)、国家自然科学基金青年基金项目“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)和教育部人文社会科学研究青年基金项目“面向信息处理的汉语隐喻研究”(项目编号:10YJCZH052)的研究成果之一。

常出现的一种结构，并且解析了这一结构的不同点。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调查，其中抽象名词作宾语的句子占到 72%。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。其中，存现动词中的“有”和“是”与抽象名词搭配的情况占到 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，赵红艳[9]利用词语抽象度进行隐喻识别，利用 HowNet 词库，进行人工分类，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化方法研究的不足，本文以统计学习理论中逻辑回归为计算模型，通过构建抽象词库计算汉语词语抽象度；并提出一种基于词语抽象度的汉语隐喻识别算法，以验证该词语抽象度可以在隐喻识别中反映出词语的具体和抽象程度。

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。在汉语词语抽象度计算方面，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。在统计学习理论中，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。本文将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，通过逻辑回归算法构建计算词语抽象度的具体计算公式；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；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；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视、听、说、闻、感受、触感等方面入手。显然，有了这些特征，就可以利用此算法得到该词语为抽象词的概率，即为该词语的抽象度。

逻辑回归模型用条件概率 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图形是一条 S 型曲线，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:

$$P(Y=0\mid X)=\frac{1}{1+e^{WX}} \tag{1}$$

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。公式(1)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于 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。由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，该模型能够较好地拟合抽象度这个概念。

综上可知，在计算抽象度时，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量词和名词的组合。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 等[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},…, 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 的研究人员于 2013 年发布的 word2vec①。本文以搜狗实验室提供的 2012 年 6 月–7 月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

根据词语抽象度计算模型可知，在已经得到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，需要尽可能满足：对于训练数据中的具体词语，使由公式(1)计算的条件概率尽量小，而对于训练数据中的抽象词语，使之尽量大。用联合概率表示如公式(2)所示:

$$\max P(W) = \prod_{j=1}^{M} \frac{1}{1+e^{-y_j WX}} \tag{2}$$

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。

通过上述分析可知，构建合理的抽象词库对于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，已被标注过词语抽象度的词语一共有 8 004 个，标注区间为 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。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需要的特征权重向量 W。

本文构建了一个包含 2 091 个词语的汉语抽象词库，其中具体词语 1 057 个和抽象词语 1 034 个。在此抽象词库的基础上，通过迭代计算，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地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。

利用逻辑回归算法进行预测词语抽象与具体的实验，得到的结果如表 1 所示:

表 1 基于逻辑回归的词语抽象与具体分类的结果

次数	召回率(%)	准确度(%)	F 值(%)
1	88.02	91.30	89.63
2	80.79	89.94	85.12
3	83.25	95.90	89.13
4	81.71	95.89	88.23
5	82.95	97.87	89.79
6	84.40	98.22	90.79
7	90.80	80.61	85.41
8	81.79	98.98	89.57
9	88.83	91.62	90.20
10	80.58	99.70	89.12

其中，10 次实验的平均准确度约为 94%。平均召回率为 84.31%，平均 F 值为 88.70%。对实验结果进行分析可知，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地预测结果，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 计算词语抽象度。

据此本文已经得到计算词语抽象度的特征向量和特征权重向量，通过这些变量就可以计算词语抽象度。

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表示抽象的概念。因此，本文通过隐喻本体和喻体之间的抽象与具体程度识别隐喻。基于这一思想，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

①https://code.google.com/p/word2vec/.

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图 1 所示:

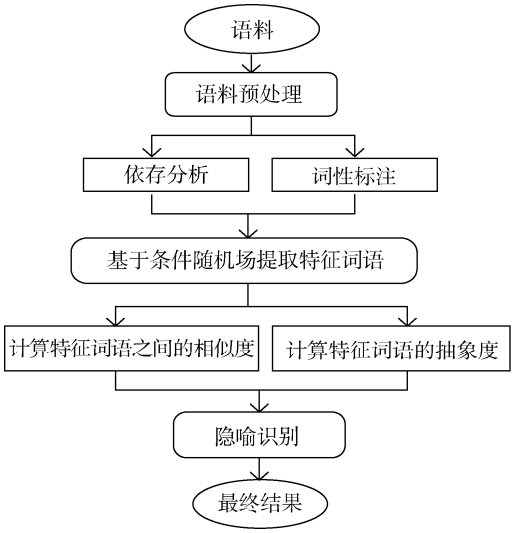

图 1　基于词语抽象度的汉语隐喻识别算法流程

(1) 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;

(2) 计算特征词语的抽象度和它们之间的相似度;

(3) 使用支持向量机模型对隐喻进行识别。

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 等提出的，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，本文主要考虑 4 个方面: 词语、词语词性标记、依存关系和窗口的大小。本文把 4 个方面特征进行组合实验，得到结果如表 2 所示。其中，W 表示词语，nW 表示窗口为[–n, +n]之间的词语; P 表示词性，nP 表示窗口为[–n, +n]之间词语的词性; R 表示依存关系，nR 表示窗口为[–n, +n]之间词语的依存关系。

分析表 2 可知，当窗口为 7，并包含词语、词性标记、依存关系的特征时，提取特征词语可以达到最出色的效果，准确度为 86.9%。例如“沉默是金”经过提取之后得到“沉默”和“金”，为之后的隐喻识别提供基础。

表 2　条件随机场提取特征词语的结果

特征模板	准确度
2W+2P	78%
3W+3P	77.4%
2W+2P+2R	83.6%
3W+3P+3R	85%
4W+4P+4R	85.8%
5W+5P+5R	86%
6W+6P+6R	86.4%
7W+7P+7R	86.9%
8W+8P+8R	86.3%

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(对特定训练样本的学习精度)和学习能力(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。

本文选取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 等[13]的概念隐喻理论，人们往往会在隐喻中用具体的事物表示抽象的概念。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0.000 133，而金的抽象度则为0.617 081，两者相差较大，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0.008 939，相似度较低，具备成为隐喻的基本条件。

4　实验及结果分析

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，本文在该抽象词库上同时按照 Turney 等[5]的算法思想，用相似度的方法计算词语的抽象度。具体算法如下: 计算该词语 w_i 与抽象词库中的抽象词语之间的相似度之和

a_{wi}，以及该词语与抽象词库中的具体词语之间的相似度之和 c_{wi}，把 a_{wi} 和 c_{wi} 相减之后经线性归一处理得到的值作为该词语的抽象度 v_{wi}，如公式(3)所示:

$$v_{wi}=\frac{(a_{wi}-c_{wi})-\min(a_{wi}-c_{wi})}{\max(a_{wi}-c_{wi})-\min(a_{wi}-c_{wi})} \quad (3)$$

由于抽象度是一个较为主观的概念，为使算法对比结果更为客观，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：邀请 5 位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。

(3) 问卷调查结果：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计算得到的抽象度更加合理的词语样本共有20个，比例为10.0%；而其他4.5%的词语样本被认为两种算法均不够合理，如表3所示:

表 3　词语抽象度算法对比的问卷调查结果

结果	所占比例
两者均不合理	4.5%
逻辑回归算法更合理	85.5%
词语相似度算法更合理	10.0%

通过对结果进行分析发现，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 0.447 684 和 0.457 421，“马克思”计算的结果分别为 0.275 683 和 0.214 427。而在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良好的效果。因此，根据问卷调查的结果，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

本文通过国家语言文字工作委员会语料库[①]以及北京大学中国语言学研究中心的语料库检索系统[②]搜索指称型模式的句子，从中选取 1 000 句作为本次实验的语料，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的有效性，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，即无法进行隐喻识别，因此本文先在测试语料中进行隐喻识别，再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尔滨工业大学的语言云系统(LTP)[16]、条件随机场模型工具包 CRF++[③]和支持向量机模型工具包 LibSVM[④]。把隐喻识别实验语料分 10 份，每次选 9 份作为训练语料，其余 1 份作为测试语料，经过 10 次实验得到的结果如表 4 和表 5 所示:

表 4　隐喻识别结果对比

实验	F 值	召回率	准确率
1	0.555	0.524	0.600
2	0.521	0.484	0.578

表 5　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	0.630	0.601	0.671
2	0.594	0.554	0.646

①http://www.cncorpus.org/CCindex.aspx.

②http://ccl.pku.edu.cn:8080/ccl_corpus/index.jsp?dir=xiandai.

③http://crfpp.googlecode.com/svn/trunk/doc/index.html.

④http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html.

(3) 实验分析

从表 4 和表 5 的实验结果可知，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。这在一定程度上反映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。

另外，从表 4 和表 5 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进一步提升，词语词向量的训练语料还需扩大，以提高算法的准确度。

5 结　语

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。本文提出一种自动计算汉语词语抽象度的方法，并用它对指称型句子进行隐喻识别，根据实验结果得到如下结论:

(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) 考虑其他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, & 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 24th International Florida Artifical Intelligence Research Society Conference, Palm Beach, Florida, USA. 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, Neuman Y, Assaf D, et al. Literal and Metaphorical Sense Identification Through Concrete and Abstract Context [C]. In: Proceedings of the 2011 Conference on the Empirical Methods in Natural Language Processing, Edinburgh, UK. 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): 307-308. (Guan Yang. Analysis of the Nouns Modified by Adverbs [J]. Anhui Literature, 2008(12): 307-308.)

[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. An Investigation into the Semantic Collocation of Abstract Nouns [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(5): 31-34. (Yang Yuling. Cognitive Salience and the Contained “Have” Sentence Format [J]. Rhetoric Learning, 2007(5): 31-34.)

[11] Bengio Y, Ducharme R, Vincent P, et al. A Neural Probabilistic Language Mode [J]. Journal of Machine Learning Research, 2003, 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, Johnsen M. Metaphors We Live by [M]. The University of Chicago Press, 1980.

[14] Lafferty J, McCallum A, Pereira F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data [C]. In: Proceedings of the 18th 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 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, 25(6): 53-62. (Liu Ting, Che Wanxiang, Li Zhenghua. Language Technology Platform [J]. Journal of Chinese Information Processing, 2011, 25(6): 53-62.)

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：进行实验；
张华，黄孝喜，吴铤：采集、清洗和分析数据；
黄孝喜，张华，陆蓓：论文起草；
黄孝喜：论文最终版本修订。

收稿日期: 2014-10-28
收修改稿日期: 2014-12-22

An Approach to Chinese Metaphor Identification Based on Word Abstractness

Huang Xiaoxi[1,2] Zhang Hua[1] Lu Bei[1] Wang Rongbo[1] Wu Ting[1]
[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 to automatically compute Chinese word abstractness, and introduce it into metaphor identification task in natural language understanding. [**Methods**] The word abstractness is computed by logistic regression model. The features are the word vectors computed by neural network model and the feature weight vectors come from a hand coded abstractness dictionary. A metaphor identification algorithm based on word abstractness is proposed to demonstrate the validity of this method. [**Results**] By comparing with the existing methods of word abstractness computing, this method has better accordance with human cognition and is an effective method in metaphor identification task. [**Limitations**] The utilization of word vectors for word abstractness is defective. The scale of the abstract words affects the learning of feature weight vectors. [**Conclusions**] Word abstractness computing reflects the ability to concept classification, Chinese word abstractness computed by this method is better fitting the human cognition, and the experimental results show that word abstractness can improve the effect of metaphor identification.

Keywords: Word abstractness Neural network language model Metaphor identification