·人工智能及识别技术· 文章编号: 1000-3428(2016)02-0169-06 文献标识码: A 中图分类号: TP391

基于词向量的中文词汇蕴涵关系识别

张志昌,周慧霞,姚东任,鲁小勇

(西北师范大学计算机科学与工程学院,兰州 730070)

摘 要: 英文词汇蕴涵关系识别已有较多研究,并提出许多识别模型,但针对中文的词汇蕴涵关系获取则鲜有研究。为此,提出一种中文词汇蕴涵关系识别方法。利用词向量技术,在中文维基百科语料上进行训练,将词汇表示为词向量,设计各种基于词向量的分类特征,训练得到可用于名词词汇蕴涵关系分类的支持向量机分类模型。实验结果表明,与传统的余弦相似度方法相比,该方法以及设计的各种分类特征在词汇蕴涵关系识别方面具有明显优势。

关键词: 文本蕴涵; 词汇蕴涵; 词向量; 蕴涵特征; 支持向量机

中文引用格式: 张志昌,周慧霞,姚东任,等. 基于词向量的中文词汇蕴涵关系识别[J]. 计算机工程,2016,42(2):169-174.

英文引用格式: Zhang Zhichang, Zhou Huixia, Yao Dongren, et al. Recognition of Chinese Lexical Entailment Relation Based on Word Vector[J]. Computer Engineering, 2016, 42(2): 169-174.

Recognition of Chinese Lexical Entailment Relation Based on Word Vector

ZHANG Zhichang, ZHOU Huixia, YAO Dongren, LU Xiaoyong

(School of Computer Science and Engineering, Northwest Normal University, Lanzhou 730070, China)

【Abstract】 Automatic recognition of English lexical entailment relation has many researches, and many recognition models are presented. But study on Chines lexical entailment is not sufficient while there have many studies on English lexical entailment from different points of view. This paper proposes a recognition method of Chinese lexical entailment relation based on word vector, it uses word vector technology on Chinese Wikipedia corpora, and word is represented as word vector. Word vector based classification features are designed, and Support Vector Machine (SVM) model for Chinese noun lexical entailment classification is trained on manually created Chinese lexical entailment data set. Experimental results show that the method and designed classification features have good performance on lexical entailment relation recognition compared with traditional cosine similarity method.

【Key words】 textual entailment; lexical entailment; word vector; entailment feature; Support Vector Machine (SVM)

DOI: 10.3969/j.issn.1000-3428.2016.02.031

1 概述

文本蕴涵识别(Recognition Texutal Entailment, RTE)是自然语言处理领域的重要研究内容之一[1],任务是在给定 2 个句子或者段落(称其中一个为文本 T,另一个为假设 H)的条件下,判断 T 是否蕴涵 H,或者 H 的含义是否能从 T 中推导出来。文本蕴涵识别在信息检索、问答系统、机器翻译等方面都有重要应用。

已有研究表明,词汇蕴涵知识越丰富,对文本蕴涵识别的帮助会越大[1-3]。因此,借助大规模文本语料库,从中抽取大量的词汇蕴涵规则,则是提高文本蕴涵识别性能的关键。而在抽取词汇蕴涵规则时,经常需要判断给定的 2 个词之间是否存在蕴涵关系。

本文针对判断中文词汇是否存在蕴涵关系的问题,首先利用词向量技术在中文维基百科语料进行训练,将词汇表示成词向量,然后基于词向量构造各种有效的分类特征,通过训练 SVM 分类器对候选名词词汇蕴涵关系对进行分类判断。

基金项目: 国家自然科学基金资助项目(61163039,61163036,61363058);西北师范大学青年教师科研能力提升计划基金资助项目(NWNU-LKQN-10-2,NWNU-LKQN-12-23)。

作者简介: 张志昌(1976-),男,副教授、博士,主研方向为自然语义处理、Web 挖掘;周慧霞、姚东任,硕士研究生;鲁小勇,工程师。

收稿日期: 2015-08-17 修回日期: 2015-09-16 E-mail: zzc@nwnu.edu.cn

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2 相关研究

对已有的词汇蕴涵识别方法进行总结，可分为 4 类:

(1) 基于规则的方法。主要借助句子语法结构或百科知识库的特性来提取蕴涵规则。如文献[4]从维基百科中获取英文词汇蕴涵规则，利用了英文句子的系动词和表语的表达特点、以及维基百科中含义相同或者相近、但表达形式不同的词汇之间的链接、重定向关系等。

(2) 基于词典的方法。从特定的词典中抽取蕴涵规则，一般是利用词典中的各种语义关系，如同义关系、上下位关系、部分与整体关系等。如文献[5]从 WordNet 和 Word Similarity Database 抽取蕴涵规则。在 WordNet 中 dog 是 animal 的下位词，因此 dog 蕴涵 animal。

(3) 基于统计的方法。

1) 基于语义相似度的蕴涵规则判定。一般是计算词汇间的某种蕴涵度量，若值高于一定阈值，则判定为有蕴涵关系。如文献[6-7]认为，如果词汇 u 的上下文是 v 的上下文的子集，则 u 蕴涵 v，以此提出判断词汇蕴涵程度的度量 WeedsPrec。但 WeedsPrec 方法会抽取到大量不太常用的词汇蕴涵规则，于是文献[7]提出用 WeedsPrec 结合文献[8]相似度来对稀有词汇进行均衡，给出新的度量方法 balPrec。文献[9]利用词汇语义相似度进行类似蕴涵关系的本体匹配。文献[10]利用基于 WordNet 的动词语义相似度来学习模板蕴涵规则。文献[11]认为，不仅要看词汇 u 的上下文有多大比例包含在 v 的上下文中，也要看所包含的上下文的权重，从而提出 balAPinc 方法。另外，文献[12]将维基百科网页表示成词汇-文档矩阵，用 jLSI(java Latent Semantic Indexing)工具对矩阵进行潜在语义分析、构造词汇潜在语义向量，然后计算词汇的相似度值，若该值高于一定阈值，则该词对之间存在蕴涵关系。

2) 基于统计机器学习分类的方法。文献[13]根据“形容词修饰的名词短语蕴涵中心词(如 big cat ⇒cat)”这一特点，自动构建词汇蕴涵对训练语料，并通过词汇或短语在句子中的上下文，构建词汇点互信息矩阵，进而得到词汇语义空间向量，并以此为特征训练 SVM 分类器分类、识别新的蕴涵规则。文献[14]通过对“动词对”构建可判断是否蕴涵的特征，包括句子间的连接词、子句间的依存关系等特征，训练 SVM 分类器识别词汇蕴涵。文献[15]提出 SimDiffs 方法，使用 2 种矩阵: 领域矩阵和函数矩阵，来构造 4 种相似度差，将这些差值作为分类特征训练分类模型。

在以上各种方法中，规则方法虽然抽取到的词汇蕴涵规则准确率高，但对语言的覆盖灵活性差，且人工构造规则费时费力。基于词典的方法，受限于词典规模，抽取到的词汇蕴涵规则也往往有限。因此，基于统计机器学习的方法是研究的主流。

基于深度学习技术的词向量在各种应用中已经取得了不错的性能。因此本文针对中文的词汇蕴涵识别问题，基于中文维基百科语料训练词汇的词向量，然后构建基于词向量的各种分类特征，通过训练 SVM 分类模型，对候选词汇蕴涵关系进行分类判别。

本文方法有以下特点: (1) 基于词向量，构建适合词汇蕴涵关系分类的各种向量特征; (2) 通过训练 SVM 分类器，将多种不同类型的特征综合在一起; (3) 根据对已有相关研究工作的调研，本文针对中文词汇蕴涵关系识别进行研究工作，并标注了一定规模的词汇蕴涵关系分类识别的训练和测试语料。

本文所构造的词汇蕴涵关系识别训练和测试数据集，以及在中文维基百科语料上训练得到的100 维的词向量、分类特征构造 python 程序源代码，均可在 http://pan.baidu.com/s/1gdfIXuN 网址下载。

3 基于词向量的分类特征设计

3.1 词汇的词向量表示

判断 2 个词汇之间是否存在蕴涵关系时，仅根据词汇本身显然是无法进行的。因此，以往研究都是借助一个较大规模的语料库，将词汇表示一种分布向量的形式，再根据向量之间的关系来判别词汇之间的关系。如对一个词汇 w，常用的向量表示方法，一种是根据语料库中词汇 w 在特定上下文窗口中的所有上下文共现词 $c_i(i=1,2,\cdots,N)$，计算 w 和每个共现词 c_i 之间的共现次数值、点互信息 PMI 值等，得到 w 的向量表示 $<w_1,w_2,\cdots,w_N>$，其中的每个分量对应一个上下文词。该方法的缺点是上下文窗口大小难以确定，向量的维数太高，无法解决一词多义、多词一义的问题。还有一种是利用语料库构建词汇-文档矩阵 $\boldsymbol{M}$ 并进行奇异值 SVD 分解等，得到词汇的低维分布向量。该方法的缺点是对于大语料库，构建词汇-文档矩阵、进行奇异值分解降维，计算复杂性太高。

而在近年来的研究中，根据文献[16]提出的词嵌入或者词向量思想，利用神经网络方法，通过在大规模语料库中进行训练，将每一个词映射成一个固定长度的短向量，则是词汇向量表示研究和应用的热点。依据这种思想开发的各种工具中，word2vec 是 Google 在 2013 年开源的一款利用神经网络方法将词表征为 k 维实数值向量的高效工具，采用一个三层的神经网络“输入层-隐藏层-输出层”，对每个词根据词频进行 Huffman 编码，所有词频相似的词

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

汇隐藏层激活的内容基本一致,而出现频率越高的词语激活的隐藏层数目越少,因此有效降低了计算的复杂度。由于 word2vec 在计算上的高效性,被广泛应用在自然语言处理的很多应用中。

3.2 分类特征

本文通过对各种词汇蕴涵关系进行分析,发现确定词汇间存在蕴涵关系的因素较多,而单独的某个因素无法覆盖所有的蕴涵规律。因此,通过训练 SVM 分类器,将多种分类特征有机综合在一起,并通过大量实验来检测这些特征的有效性。

对于 2 个词或者短语 u 和 v,首先利用 word2vec 将它们表示成维数相同的词向量 $\boldsymbol{U}=<u_1,u_2,\cdots,u_n>$ 和 $\boldsymbol{V}=<v_1,v_2,\cdots,v_n>$。根据词汇蕴涵关系的特点,所设计的分类特征有:

(1) 向量差特征 f_{diff}

如果 2 个词汇或者短语各自的向量中在相同维度分量度上相差很小的特征越多,说明 2 个词越相似,进而说明它们具有蕴涵关系的可能性越大。因此,将向量差特征定义如下:

$$f_{\text{diff}}=\boldsymbol{U}-\boldsymbol{V}=<u_1-v_1,u_2-v_2,\cdots,u_n-v_n> \tag{1}$$

(2) 向量乘特征 f_{mul}

具有蕴涵关系的 2 个词汇或者短语,它们在含义上一定有较强的相似之处,这体现在它们各自的向量上,就是在表达词汇含义的不同的维度上有很大的交集。如果某个特征经常出现在词汇的上下文中,则该特征在词汇的词向量中对应维度上的值就为正,否则为负。

因此,对于词汇 u 和 v,对它们的词向量中相同分量上分别进行乘积,若这些乘积结果中正的分量越多、越大,说明这 2 个词汇的上下文交集越大,进而说明具有蕴涵关系的可能性越大。基于此,将2 个词向量的各个分量分别相乘组成新的向量作为训练特征 f_{mul},定义为:

$$f_{\text{mul}}=<u_1\times v_1,u_2\times v_2,\cdots,u_n\times v_n> \tag{2}$$

(3) 向量和特征 f_{add}

如果体现词汇含义的某个特征经常出现在词汇的上下文中,则该特征在词向量的对应分量上为正。如果该特征是 2 个词汇 u 与 v 共同的重要特征,则对 u 与 v 的词向量相应分量求和,结果也应该为正。因此,构造向量和特征,将 2 个向量的各个分量分别相加组成新的向量,并作为训练特征,来反映 2 个词向量的共同上下文特点。向量和特征的定义如下:

$$f_{\text{add}}=\boldsymbol{U}+\boldsymbol{V}=<u_1+v_1,u_2+v_2,\cdots,u_n+v_n> \tag{3}$$

(4) 向量连接特征 f_{cat}

将 2 个词汇的词向量整体全部作为分类器的特征,让分类器判别和寻找它们的相关性。因此,构造向量连接特征,即将一个词的词向量连接到另一个词的词向量尾部组成新的向量作为训练特征。向量连接 f_{cat} 特征定义为:

$$f_{\text{cat}}=<u_1,u_2,\cdots,u_n,v_1,v_2,\cdots,v_n> \tag{4}$$

(5) *balAPinc* 特征 f_{bal}

文献 [6] 针对词汇蕴涵问题,提出了“分布一般性”的概念,即如果词汇 u 的上下文包含在词汇 v 的上下文中,则词汇 v 在含义分布上要比 u 更一般,也就是 u 蕴涵 v。因此,判断 u 是否蕴涵 v 的一个方法就是看 u 的上下文特征中有多大比例包含在 v 的上下文中。而文献 [8] 进一步提出,不仅要看 u 的上下文特征有多少包含在 v 的上下文中,也要看包含在 v 的上下文特征集合中的特征对于 u 来说有多重要,同时也要看 u 和 v 在语义上有多相似。据此,文献 [8] 提出了 *balAPinc* 这种词汇蕴涵关系度量。

通过 word2vec 已经得到了每个词的词向量,而且每个词的词向量中的各个分量对应相同的特征。因此,在 word2vec 生成的词向量基础上,对 *balAPinc* 度量进行了一定的修改,并作为分类器的一项特征。对于词汇 u 和 v,将它们各自词向量的全部分量看作是 u 的上下文特征集合 F_u 和 v 的上下文特征集合 F_v,则 *balAPinc* 度量的定义如下:

$$f_{\text{bal}}=balAPinc(u,v)=\sqrt{APinc(u,v)\cdot LIN(u,v)} \tag{5}$$

LIN 相似度的计算公式如下:

$$LIN(u,v)=\frac{\sum_{f\in F_u\cap F_v}W_u(f)+W_v(f)}{\sum_{f\in F_u}W_u(f)+\sum_{f\in F_v}W_v(f)} \tag{6}$$

其中,$W_u(f)$ 表示特征 f 在 u 的词向量中对应分量的权值; $W_v(f)$ 同理。由于 2 个词的词向量每个维度对应相同的特征,因此规定,当 $W_u(f)>0$,记为 $f\in F_u$; 当 $W_u(f)<0$,记为 $f\notin F_u$; 对于 v 的词向量,同理。

对于 *APinc* 值的计算,本文的方法与 Kotlerman 所提出的计算有所不同,通过以下所述过程得到。

用 F_w 表示词 w 的上下文集合,也就是词向量的所有分量组成的集合,$|F_w|$ 是词向量中分量大于 0 的个数。将 F_w 中的特征按它们在词向量中的分量值降序排序,有如下公式:

$$rank(f_{wr},F_w)=r \tag{7}$$

其中,f_{wr} 是 F_w 中第 r 个特征。为了将 *rank* 值归一化到 0-1 区间,定义如下计算方法:

$$rel(f,F_w)=\begin{cases}1-\dfrac{rank(f,F_w)}{|F_w|+1} & \text{if } f\in F_w\\ 0 & \text{if } f\notin F_w\end{cases} \tag{8}$$

另外,引入 $inc(r,F_u,F_v)$ 来表示一个特征集合,表示 F_u 中的前 r 个特征也在 F_v 中,定义如下:

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$inc(r,F_u,F_v) = \{f \mid rank(f,F_u) \leqslant r \text{ and } f \in (F_u \cap F_v)\} \quad (9)$$

由于每个词的词向量中的各个分量对应相同的特征，在式(9)中，对于词 u 和 v，u 和 v 的对应特征如果都大于0，就将该特征计入 $inc(r,F_u,F_v)$ 中，否则不计入。对 inc 进行归一化：

$$p(r,F_u,F_v) = \frac{|inc(r,F_u,F_v)|}{r} \quad (10)$$

则将 $APinc$ 的计算定义如下：

$$APinc(u,v) = \frac{\sum_{r=1}^{|F_u|}[p(r,F_u,F_v) \cdot rel(f_{ur},F_v)]}{|F_u|} \quad (11)$$

4 实验与结果分析

本文基于中文维基百科语料，利用 Google 提供的 word2vec 工具来训练各个词的词向量；其次，需要构造词汇蕴涵关系训练和测试语料；然后，基于3.2节所述的特征，利用 libsvm 工具，训练词汇蕴涵关系分类的 SVM 分类器；最后，在测试语料上对训练得到的分类器性能进行综合评价，以检验本文方法以及各个分类特征的有效性。

4.1 数据集构建

首先从维基百科的语料库中下载中文维基百科 zhwiki-20150325-pages-articles. xml. bz2 数据集并用维基百科抽取器对其进行处理，得到 763 MB 语料。用 opencc 对语料进行繁体到简体的转换，去掉文本中的各种标签，用分词系统 ICTCLAS2015 对文本进行分词、词性标注。然后用 Google 提供的 word2vec 对全部语料进行训练，分别得到 100 维、200 维、300 维、400 维的词向量文件。

从语料库所包含的全部名词中，随机找出一些语义范围比较大的词(如“公司”、“酒”)共 280 个，然后利用 word2vec 训练得到的词向量，计算这 280 个词与其他所有词的余弦相似度，得到每个词最相似的另外 10 个词。通过人工检查发现，这些相似词对中，很多词对之间并不存在蕴涵关系，这说明相似与蕴涵并不是等价关系。从这些相似的词对之中，手工抽取、构建了 1 400 个存在蕴涵关系的词对和1 400 个不存在蕴涵关系的词对，共 2 800 个词对。

将 2 800 个词对分成两部分：1 400 个作为训练集，其中包含有 700 个为存在蕴涵关系的词对和 700 个为不存在蕴涵关系的词对；另外的 1 400 对作为测试集，其中的蕴涵关系词对和非蕴涵关系词对比例和训练集相同。

4.2 评价指标

为了方便定义准确率等评价指标，设一个 2×2 的混淆矩阵 $\boldsymbol{C} = (c_{ij})_{2 \times 2}$，其中，$c_{ij}$ 表示词对实际属于类 i 但分类器将词对判别为类 j 的总的词对数目($i,j \in \{0,1\}$)，其中类 1 表示词对之间存在蕴涵关系，类 0 表示不存在蕴涵关系。则分别定义准确率、召回率、F 值的如下：

$$Pre_0 = c_{00}/(c_{00} + c_{10}) \quad (12)$$

$$Pre_1 = c_{11}/(c_{11} + c_{01}) \quad (13)$$

$$Rec_0 = c_{00}/(c_{00} + c_{01}) \quad (14)$$

$$Rec_1 = c_{11}/(c_{11} + c_{10}) \quad (15)$$

$$F_0 = 2 \cdot Pre_0 \cdot Rec_0/(Pre_0 + Rec_0) \quad (16)$$

$$F_1 = 2 \cdot Pre_1 \cdot Rec_1/(Pre_1 + Rec_1) \quad (17)$$

其中，Pre_0，Rec_0，F_0 是针对不存在蕴涵关系的词对进行分类识别的准确率、召回率、F 值评价指标；Pre_1，Rec_1，F_1 则针对存在蕴涵关系的词对进行分类识别的评价指标。对蕴涵与不蕴涵 2 个类别的准确率、召回率、F 值通过权值进行综合，定义如下：

$$w_0 = (c_{00} + c_{01})/(c_{00} + c_{01} + c_{10} + c_{11}) \quad (18)$$

$$w_1 = (c_{11} + c_{10})/(c_{00} + c_{01} + c_{10} + c_{11}) \quad (19)$$

$$Pre = w_0 \cdot Pre_0 + w_1 \cdot Pre_1 \quad (20)$$

$$Rec = w_0 \cdot Rec_0 + w_1 \cdot Rec_1 \quad (21)$$

$$F = w_0 \cdot F_0 + w_1 \cdot F_1 \quad (22)$$

其中，Pre，Rec，F 是对所有词对进行蕴涵关系分类识别的综合准确率、召回率、F 值评价指标。

4.3 结果分析

基于 4.1 节所述数据集，利用 word2vec 训练出的词向量，然后依照 3.2 节所述的特征构造方法，在所构造的词汇蕴涵训练集上训练 SVM 分类器，最后在测试集上进行蕴涵关系分类测试与评价。选用 libsvm 作为训练 SVM 分类起的工具，训练时使用径向基核函数。通过在训练集上进行交叉验证，选择 libsvm 的参数 $c = 32.0$，$g = 0.007\ 812\ 5$。

利用 word2vec 在中文维基百科语料上进行训练时，可以得到不同维度的词向量表示。为了确定适合本文所提方法的词向量维度，用 3.2 节所述的不同向量特征，训练并测试 SVM 词汇蕴涵关系分类识别模型的性能。

表 1 给出了在 100 维、200 维、300 维、400 维的词向量表示条件下，基于向量差 f_{diff}、向量乘 f_{mul}、向量和 f_{add}、向量连接 f_{cat} 这 4 种不同分类特征时 SVM 分类模型的 F 值。

表 1 不同词向量维度、不同特征下的分类模型 F 值

维度	向量和 f_{add}	向量差 f_{diff}	向量连接 f_{cat}	向量乘 f_{mul}
100	0.603	0.568	0.586	0.507
200	0.597	0.567	0.542	0.496
300	0.661	0.555	0.520	0.496
400	0.575	0.544	0.556	0.466

为了能更直观地观察不同向量维度对分类性能的影响，将表 1 中的数据以折线图的形式进行展示，如图 1 所示。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

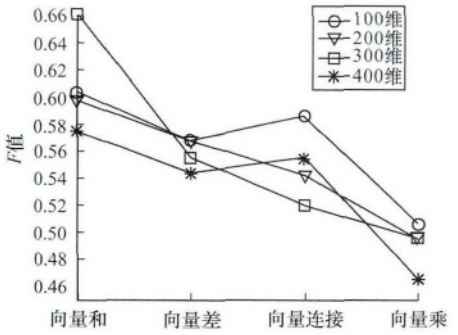

图 1　不同词向量维度在不同特征下分类模型 F 值对比

从表 1 和图 1 可以看出，不同的词向量特征构造对词汇蕴涵关系的分类识别有一定的性能差异。同时，当训练所得的词向量维度不同时，分类性能也会不同。总体来看，当词向量维度为 100 维时，基于向量差 f_{diff}、向量连接 f_{cat}、向量乘 f_{mul}3 种特征中的任何一个特征进行词汇蕴涵关系分类，都能有相对其他维度词向量较好的性能表现。而且，当词向量为 100 维时，计算复杂度相对更小。因此，在最终的实验中，选择100维的词向量，通过对基于词向量的不同特征进行组合，评测不同特征对 SVM 分类器性能的影响。在评测时，综合比较拥有蕴涵关系词对分类的准确率 Pre_1、召回率 Rec_1、F 值 F_1；对不具有蕴涵关系词对进行分类的准确率 Pre_0、召回率 Rec_0、F 值 F_0，以及 F_0 和 F_1 的综合 F 值。

另外，如果 2 个词汇之间存在蕴涵关系，则它们之间一定有较高的相似度，因此选择向量余弦相似度方法为性能评价基准(baseline) 。基于 100 维的词向量，对于测试集中的每一个词对(u，v) ，计算词 u 和 v 之间的余弦相似度，若该相似度超过阈值(本文设阈值为 0. 7) ，则认为它们之间存在蕴涵关系。对测试集中所有词对依此分类后，得到余弦相似度方法时的各项性能评价指标。

表 2 给出了在使用和组合不同的分类特征时，SVM 分类器在词汇蕴涵测试集上的性能情况。从结果可以看出，除了 f_{mul} 特征，选择其他任何特征或者特征组合，训练 SVM 分类器进行词汇蕴涵关系的有指导分类识别，性能都要比余弦相似度方法好。

表 2　不同特征组合时的词汇蕴涵分类性能

特征	Pre_0	Rec_0	F_0	Pre_1	Rec_1	F_1	Pre	Rec	F
f_{add}	0. 608	0. 583	0. 595	0. 599	0. 624	0. 611	0. 604	0. 604	0. 603
f_{cat}	0. 622	0. 469	0. 535	0. 574	0. 716	0. 637	0. 598	0. 592	0. 586
f_{diff}	0. 563	0. 611	0. 586	0. 575	0. 526	0. 549	0. 569	0. 568	0. 568
f_{mul}	0. 507	0. 529	0. 517	0. 507	0. 486	0. 496	0. 507	0. 507	0. 507
f_{bal}	0. 527	0. 573	0. 549	0. 532	0. 486	0. 508	0. 530	0. 529	0. 528
$f_{\text{add}}+f_{\text{diff}}$	0. 610	0. 564	0. 586	0. 594	0. 639	0. 616	0. 602	0. 601	0. 601
$f_{\text{add}}+f_{\text{cat}}$	0. 637	0. 393	0. 486	0. 561	0. 776	0. 651	0. 599	0. 584	0. 568
$f_{\text{add}}+f_{\text{mul}}$	0. 605	0. 589	0. 597	0. 599	0. 616	0. 607	0. 602	0. 602	0. 602
$f_{\text{add}}+f_{\text{bal}}$	0. 605	0. 636	0. 620	0. 617	0. 586	0. 601	0. 611	0. 611	0. 610
$f_{\text{add}}+f_{\text{bal}}+f_{\text{diff}}$	0. 605	0. 627	0. 616	0. 613	0. 590	0. 601	0. 609	0. 609	0. 608
$f_{\text{add}}+f_{\text{bal}}+f_{\text{cat}}$	0. 638	0. 441	0. 522	0. 573	0. 750	0. 650	0. 606	0. 596	0. 586
$f_{\text{add}}+f_{\text{bal}}+f_{\text{mul}}$	0. 602	0. 600	0. 601	0. 602	0. 604	0. 603	0. 602	0. 602	0. 602
$f_{\text{add}}+f_{\text{bal}}+f_{\text{diff}}+f_{\text{mul}}$	0. 596	0. 579	0. 587	0. 591	0. 609	0. 600	0. 594	0. 594	0. 593
余弦相似度	0. 545	0. 913	0. 683	0. 732	0. 239	0. 360	0. 639	0. 576	0. 521

图 2 以条形图的形式展示不同特征进行组合时，分类器对非蕴涵关系词对的分类 F 值 F_0、蕴涵关系词对分类的 F 值 F_1 以及 F_0 和 F_1 的综合 F 值。

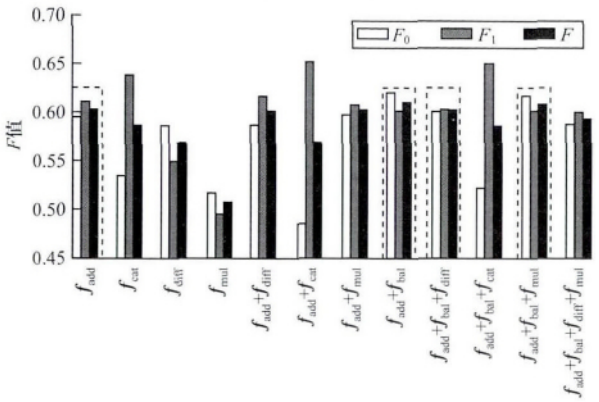

图 2　不同特征组合时词汇蕴涵分类性能 F 值对比

实验结果表明，在使用单个特征进行分类时，f_{add}特征的综合性能最好，达到 0. 603 的 F 值。因此，在 f_{add}特征基础上再组合其他特征，发现 f_{add}和 $balAPinc$ 特征组合表现出最好的综合性能，F 值达到 0. 610。进一步地，在 f_{add}和 f_{bal}的基础上组合其他特征 ，发现性能反而均有所下降。但相对来说，组合向量差特征 f_{diff}后，分类器综合 F 值为 0. 608，下降幅度较小。f_{add}，$f_{\text{add}}+f_{\text{bal}}$，$f_{\text{add}}+f_{\text{bal}}+f_{\text{diff}}$3 种特征选择条件下相对其他特征组合情况的分类性能对比如图 2所示。

为了更加深入地检验组合特征对于词汇蕴涵性能的影响，将测试集中存在蕴涵关系但并非同义词的所有词对及其类别标签{($leftw_i$，$rightw_i$，1) ，$i=1$，2，…，M } 复制出来，并将这些词对的方向进行交换，

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

形成新的不存在蕴涵关系的词对及类别标签 $\{(rightw_i, leftw_i, -1), i=1,2,\cdots,M\}$，再将它们加入到测试集中，形成扩展测试集。然后，利用 f_{add}，$f_{add}+f_{bal}$，$f_{add}+f_{bal}+f_{diff}$ 3 种特征组合分别训练分类器并在扩展测试集上进行评测，在 F_0，F_1，F 3 种评价指标下的性能情况如表 3 所示。

表 3 4 种特征组合时分类器在扩展测试集上的 F 值

特征	F_0	F_1	F
f_{add}	0.574	0.468	0.539
$f_{add}+f_{bal}$	0.593	0.464	0.550
$f_{add}+f_{bal}+f_{diff}$	0.605	0.482	0.564
$f_{add}+f_{bal}+f_{diff}+f_{mul}$	0.591	0.477	0.553

从原始测试集和扩展测试集上的实验结果可以看出，组合了向量和特征 f_{add} 和 $balAPinc$ 特征 f_{bal} 之后，要比单独用 f_{add} 特征好；在扩展测试集上的实验结果表明，f_{add}，f_{bal}，f_{diff} 3 种特征组合的性能要高于 f_{add}，f_{bal} 2 种特征组合的性能，说明向量差特征 f_{diff} 对具有蕴涵关系的词对之间蕴涵方向的识别有极大帮助，这和本文的直觉认识是一致的。当一个词 u 蕴涵 v 但 v 不蕴涵 u 时，u 的词向量 $\boldsymbol{U}$ 与 v 的词向量 $\boldsymbol{V}$ 的差 $\boldsymbol{U}-\boldsymbol{V}$，以及 $\boldsymbol{V}-\boldsymbol{U}$ 应该有一定的规律性差异，这可以通过 f_{diff} 特征来体现这种差异。

5 结束语

词汇蕴涵关系在自然语言处理领域有着非常重要的应用价值。本文提出利用词向量技术，设计基于词向量的各种词汇蕴涵关系分类特征，进行名词词对之间的蕴涵关系分类识别。实验结果表明，本文提出的方法以及设计的各种分类特征，在词汇蕴涵关系识别方面相对于传统的余弦相似度方法具有明显的优势；而“向量和”特征、“向量 $balAPinc$ 特征”、“向量差”特征 3 种的组合，在名词词汇蕴涵关系识别方面有较好的性能。同时，各种特征及其组合在测试集上的 F 值性能大部分都在 0.6 以下，说明词汇蕴涵关系识别存在较大难度的研究问题。在下一步工作中，需要设计和寻找更好的词汇蕴涵分类特征，并对词汇蕴涵关系进行更细的类别划分。

参考文献

[1] Androutsopoulos I, Malakasiotis P. A Survey of Paraphrasing and Textual Entailment Methods [J]. Journal of Artificial Intelligence Research, 2010, 38(1): 135-187.

[2] 袁毓林，王明华. 文本蕴涵的推理模型与识别模型 [J]. 中文信息学报，2010，24(2)：3-13.

[3] 盛雅琦，张 晗，吕 晨，等. 基于混合主题模型的文本蕴涵识别 [J]. 计算机工程，2015，41(5)：180-184.

[4] Shnarch E, Dagan I. Lexical Entailment and Its Extraction from Wikipedia [D]. Israel, Jaffa: Bar-Ilan University, 2008.

[5] Kouylekov M, Magnini B. Building a Large-scale Repository of Textual Entailment Rules [C]//Pro-ceedings of the 5th International Conference on Language Resources and Evaluation. Genoa, Italy: [s. n.], 2006: 2437-2440.

[6] Weeds J, Weir D. A General Framework for Distributional Similarity [C]//Proceedings of EMNLP'03. Sapporo, Japan: [s. n.], 2003: 81-88.

[7] Weeds J, Weir D, McCarthy D. Characterizing Measures of Lexical Distributional Similarity [C]//Proceedings of the 20th International Conference on Computational Linguistics). Geneva, Switzerland: [s. n.], 2004: 1015-1021.

[8] Lin Dekang. Automatic Retrieval and Clustering of Similar Words [C]//Proceedings of COLING-ACL'98. Montreal, Canada: [s. n.], 1998: 768-774.

[9] 何 娟，高志强，陆青健，等. 基于词汇相似度的元素级本体匹配 [J]. 计算机工程，2006，32(16)：191-193.

[10] Szpektor I, Dagan I. Learning Entailment Rules for Unary Templates [C]//Proceedings of the 22nd Inter-national Conference on Computational Linguistics. Manchester, UK: [s. n.], 2008: 849-856.

[11] Kotlerman L, Dagan I, Szpektor I, et al. Directional Distributional Similarity for Lexical Inference [J]. Natural Language Engineering, 2010, 16(4): 359-389.

[12] Kouylekov M, Mehdad Y, Negri M. Mining Wikipedia for Large-scale Repositories of Context-sensitive Entail-ment Rules [C]//Proceedings of the 7th Conference on International Language Resources and Evaluation. Washington D. C., USA: IEEE Press, 2010: 3550-3553.

[13] Baroni M, Bernardi R. Entailment Above the Word Level in Distributional Semantics [C]//Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon, France: [s. n.], 2012: 23-32.

[14] Weisman H, Berant J. Learning Verb Inference Rules from Linguistically Motivated Evidence [C]//Proceed-ings of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, Korea: [s. n.], 2012: 194-204.

[15] Turney P D, Mohammad S M. Experiments with Three Approaches to Recognizing Lexical Entailment [J]. Natural Language Engineering, 2015, 21(3): 437-476.

[16] Hinton G E. Learning Distributed Representations of Concepts [C]//Proceedings of the 8th Annual Conference of the Cognitive Science Society. Hillsdale, USA: [s. n.], 1986: 1-12.

编辑 索书志

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net