

文章编号:1001-9081(2016)11-3146-06

DOI:10.11772/j.issn.1001-9081.2016.11.3146

基于词向量和条件随机场的领域术语识别方法

冯艳红^{1,2}, 于红^{1,2*}, 孙庚^{1,2}, 赵禹锦³

(1. 大连海洋大学 信息工程学院, 辽宁 大连 116023; 2. 辽宁省海洋信息技术重点实验室(大连海洋大学), 辽宁 大连 116023;
3. 大连海洋大学 经济管理学院, 辽宁 大连 116023)
(*通信作者电子邮箱 yuhong@dlou.edu.cn)

摘要:针对基于统计特征的领域术语识别方法忽略了术语的语义和领域特性,从而影响识别结果这一问题,提出一种基于词向量和条件随机场(CRF)的领域术语识别方法。该方法利用词向量具有较强的语义表达能力、词语与领域术语之间的相似度具有较强的领域表达能力这一特点,在统计特征的基础上,增加了词语的词向量与领域术语的词向量之间的相似度特征,构成基于词向量的特征向量,并采用CRF方法综合这些特征实现了领域术语识别。最后在领域语料库和SogouCA语料库上进行实验,识别结果的准确率、召回率和F测度分别达到了0.9855、0.9439和0.9643,表明所提的领域术语识别方法取得了较好的效果。

关键词:词向量; 条件随机场; 术语识别; 相似度特征

中图分类号:TP391.4 **文献标志码:**A

Domain-specific term recognition method based on word embedding and conditional random field

FENG Yanhong^{1,2}, YU Hong^{1,2*}, SUN Geng^{1,2}, ZHAO Yujin³

(1. College of Information Engineering, Dalian Ocean University, Dalian Liaoning 116023, China;
2. Key Laboratory of Marine Information Technology of Liaoning Province (Dalian Ocean University), Dalian Liaoning 116023, China;
3. College of Economics Management, Dalian Ocean University, Dalian Liaoning 116023, China)

Abstract: Domain-specific term recognition methods based on statistical distribution characteristics neglect term semantics and domain feature, and the recognition result are unsatisfying. To resolve this problem, a domain-specific term recognition method based on word embedding and Conditional Random Field (CRF) was proposed. The strong semantic expression ability of word embedding and strong field expression ability of similarity between words and term were fully utilized. Based on statistical features, the similarity between word embedding of words and word embedding of term was increased to create the feature vector. term recognition was realized by CRF and a series of features. Finally, experiment was carried out on field text and SogouCA corpus, and the precision, recall and F measure of the recognition results reached 0.9855, 0.9439 and 0.9643, respectively. The results show that the proposed method is more effective than current methods.

Key words: word embedding; Conditional Random Fields (CRF); term recognition; similarity feature

0 引言

领域术语识别是自然语言处理领域的关键任务,对数据挖掘、信息检索、机器翻译等方面的研究和应用有重要的意义,引起了国内外学者们的关注^[1]。吴海燕^[2]利用互信息对旅游领域术语识别问题进行研究;李丽双等^[3]利用信息熵和词频变化对汽车领域的术语进行抽取。这类方法主要根据文本的互信息和信息熵等统计信息对术语进行识别,取得了较好的识别效果,但该类方法只考虑了文本的统计分布特性。近年来机器学习技术在自然语言处理领域得到广泛应用并取得丰硕成果,例如机器学习中的条件随机场(Conditional Random Field, CRF)算法^[4],利用文本的多种上下文特征完成对领域术语的识别。孙丽萍等^[5]用其预测企业简称,取得

了很好的效果;栗伟等^[6]将CRF算法用于医学领域术语识别;施水才等^[7]针对领域术语的特点,设计了词性、词长等多个统计特征,利用CRF算法对领域术语进行识别。这类方法将术语识别问题转为序列标注问题,利用机器学习中的CRF算法对术语识别问题进行研究。该类方法考虑了词语的多种特征,克服了使用单一特征的局限性,提高了术语的识别效果,但这些特征在本质上仍然属于词语的统计分布特性。然而,对于大部分特定领域的术语而言,都具有丰富的语义特性和领域特性,这也是领域术语区别于其他词语的重要方面。统计特征无法表达词语的语义和领域特性,影响了识别效果。所以本文研究如何将领域术语的语义特性和领域特性融入到基于CRF的领域术语识别模型中,克服统计特征的局限性,缓解高维特征向量的数据稀疏问题^[8],提高了术语识别的性

收稿日期:2016-04-22;修回日期:2016-06-20。

作者简介:冯艳红(1980—),女,黑龙江绥化人,讲师,硕士,CCF会员,主要研究方向:自然语言处理、信息检索; 于红(1968—),女,辽宁大连人,教授,博士,CCF会员,主要研究方向:数据挖掘、信息检索; 孙庚(1979—),男,黑龙江齐齐哈尔人,副教授,硕士,主要研究方向:嵌入式系统; 赵禹锦(1990—),女,辽宁营口人,硕士研究生,主要研究方向:数据挖掘、信息检索。

能。

1 特征选择

特征选择是术语识别的关键,不同类型的特征会产生不同的识别效果。特征包括统计特征和语义特征。统计特征以词语的频率为核心,采用统计学的方法给出特征值,表达能力单一,无法表达出词语的语义信息,从而影响术语识别的效果。对于特定领域的术语而言,有两个很重要的特点:第一,这类术语具有丰富的语义含义,可表达词语的内涵;第二,这类术语具有很强的领域性,即同一领域的术语具有很强的相关性。所以本文深入分析了这两个特点,给出相似度特征的计算方法,并将相似度作为术语识别的重要特征。

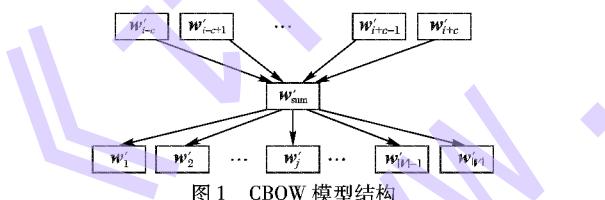
用词向量表达文本中的词语是将深度学习算法引入自然语言处理领域的一个核心技术。词向量是通过训练神经网络语言模型得到的一种分布表示特征^[9-10],即用一个连续的实数向量表达文本中的一个词语,该向量能表达词语的语义含义;语义上相似的词语在空间中的向量也相似。所以,本文采用词向量表达领域术语的语义含义。每个领域都有其核心词汇,一般以领域词典的形式存在。假定领域词典中的词语都是领域术语,如果某个词语与词典中的词语在语义上相似,那么,该词语被识别为领域术语的概率就会很大,所以本文采用词语与领域术语的词向量之间的相似度表达领域性。

1.1 相似度特征

为了将领域术语的语义和领域性融入术语识别模型中,首先要将词语的语义和领域性以适当的数据结构或形式表达出来,具体表达形式如下。

1.1.1 词向量

词向量可由 Mikolov 发布的开源 word2vec^[11-12]训练得到。Mikolov 提出了两种用于训练词向量的模型:连续词袋(Continuous Bag Of Words, CBOW)模型和 Skip-gram 模型。CBOW 在训练效率上高于 Skip-gram,所以本文使用 CBOW 模型,模型结构如图 1 所示。



给定文本的词语序列 $w_1, w_2, \dots, w_i, \dots, w_n, w_i \in V, V$ 是有限大小的词表,词表大小为 $|V|$ 。词语 w_i 对应的词向量为 w'_i 。模型的输入为 $w'_{i-c}, w'_{i-c+1}, \dots, w'_{i+c-1}, w'_{i+c}$ 表示词语 w_i 的上下文词语对应的词向量,上下文窗口大小为 c ;将输入累加后得到 w'_{sum} ,作为神经网络的隐藏层;将通过上下文预测的目标词语作为输出层。模型的训练目标为:由该上下文得到的正确的目标词语的概率最大,目标函数为式(1)所示的对数似然函数:

$$\frac{1}{n} \sum_{i=c}^{n-c} \log p(w'_i | w'_{i-c}, w'_{i-c+1}, \dots, w'_{i+c-1}, w'_{i+c}) \quad (1)$$

其中: $p(w'_i | w'_{i-c}, w'_{i-c+1}, \dots, w'_{i+c-1}, w'_{i+c})$ 是给定上下文 $w'_{i-c}, w'_{i-c+1}, \dots, w'_{i+c-1}, w'_{i+c}$ 条件下 w'_i 的概率。模型中输出层的神经元数目为 $|V|$,该值通常很大,导致训练时间长,可采用基于层次结构的哈夫曼树或负采样的方法提高训练效

率,本文采用基于层次结构的哈夫曼树的方法进行训练。训练结束后,得到 w_i 的 m 维的词向量 w'_i 为 (v_1, v_2, \dots, v_m) ,语义上相似的词语的词向量在向量空间上也相似,所以本文用词向量表达词语的语义特性。

1.1.2 相似度特征计算

本文利用渔业领域的词典——水产辞典^[13]作为领域的核心词汇集,采用待识别词语与领域词典中词语的词向量之间的语义相似度来表达领域性,相似度特征 $Simi$ 的计算方法如式(2)所示:

$$Simi(w_t, w_d) = \max_d(f(w_t, w_d)) \quad (2)$$

$w_d \in D, D$ 是领域词典。函数 $f(w_t, w_d)$ 为计算 w_t 和 w_d 的相似程度,本文取 w_t 和 w_d 的词向量的夹角的余弦值作为相似度;相似度特征 $Simi$ 取 $f(w_t, w_d)$ 中最大值作为词语的相似度。相似度特征 $Simi \in [0, 1]$,若 $w_t \in D$,则相似度特征 $Simi$ 的值为 1。

1.2 统计特征

本文的统计特征首先选择词语本身、词性、词长和是否在词典中 4 个统计特征,根据领域术语的特殊性,加入了词的特定偏旁部首数目特征。以渔业领域为例,5 个统计特征提取和分析如下。

特征 1 词语本身 Word。利用分词软件,对文本切分后生成的词语。词语是构成术语的基本符号,例如渔业领域中,镜鲤、乌鳢、苗种、亲虾、养殖等词语或者为渔业领域术语、或者为术语的后缀、或者为术语的前缀,若一个词语中包含这类词语,那么该词语通常为领域术语,所以选取词语本身作为术语识别的重要特征。

特征 2 词性 POS (Part Of Speech)。根据汉语词性对照表(北大标准/中科院标准)标注的词语的词性。术语的词性一般为名词、名词短语、动词或动词短语,而几乎不会是介词、连词等词性。术语和词性有一定的相关性,所以选取词性作为术语识别的特征。

特征 3 词长 WordLen。按照 UNICODE 编码,词语包含的字符的数目。对术语识别研究的文献中,该特征表现出较好的识别效果,所以本文研究了该特征对识别效果的影响。

特征 4 词语是否在词典中 InDic。根据预先建立好的渔业领域词典,判定该词语是否在词典中。由于假定词典中存储的词语是领域术语,如果该词语在词典中,则该词语是术语,否则不一定。该词语是否在领域词典中与该词语是否被识别为术语直接相关,所以,选取该特征为渔业领域术语识别重要特征。

特征 5 词的偏旁部首数目 Num。由于构成渔业领域的词语的字符一般包含特定的偏旁部首,根据这个特点,计算出词语中包含指定偏旁的字符的数目。渔业领域中,描述鱼类的词语的字符中一般包括偏旁部首“鱼”,描述水产类的词语的字符中一般包括偏旁部首“氵”,所以本文将该特征为术语识别的重要特征。

2 领域术语识别方法

CRF 是一种基于概率图的统计模型,可以添加多种特征,利用更丰富的上下文信息;具有表达长距离依赖的能力,克服了隐马尔可夫模型严格的条件独立性假设带来的弊端;并且该模型实现了全局归一化,可求得全局最优解,有效解决了标

记偏置问题^[4]。所以本文采用 CRF 作为术语识别模型。

2.1 CRF 模型

CRF 是一种可用于解决自然语言处理中的序列标注问题的模型,所以本文将领域术语识别问题建模为序列标注问题。将文本中的词语序列 $w_1, w_2, \dots, w_i, \dots, w_n$ 定义为序列标注问题中的观察序列 X ;标注集采用 {B,I,O} 符号集,B 和 O 分别表示术语语的开始部分、中间部分和非术语部分,对词语序列用 B、I 和 O 标注后形成标注序列,将该标注序列定义为序列标注问题中的状态序列 Y 。根据观察序列 X 和状态序列 Y ,领域术语识别问题定义为:已知观察序列 X 的条件下,求解状态序列 Y 的概率 $p(Y/X)$ 最大时的状态序列,该状态序列即为领域术语识别问题的解,计算方法如式(3)所示:

$$p(Y/X) = \frac{1}{Z(X)} \exp\left(\sum_i \sum_k \lambda_k f_k(y_{i-1}, y_i, x, i)\right) \quad (3)$$

其中: f_k 为二值特征函数,由 CRF 特征模板生成。本文根据前文选择的 5 个统计特征和 1 个相似度特征,采用特征递进的方式定义特征模板。 λ_k 为模型的参数; $Z(X)$ 为全局归一化因子,如式(4)所示:

$$Z(x) = \sum_y \exp\left(\sum_i \sum_k \lambda_k f_k(y_{i-1}, y_i, x, i)\right) \quad (4)$$

该模型通过 L-BFGS 算法对模型的参数 λ 进行估计,通过 Viterbi 算法对状态序列解码,求得条件概率 $p(Y/X)$ 最大时的状态序列 Y ,即文本的标注序列,从而得到术语识别的结果。

2.2 识别方法

在 CRF 模型的基础上,结合前文中得到的统计特征和相似度特征,给出本文的基于词向量和 CRF 的术语识别方法,用于解决领域术语识别问题,识别方法如图 2 所示。

该方法的输入为未经过处理的原始文本语料。首先,对文本语料进行分词、去除停用词等预处理;再利用统计特征分析器抽取 Word、POS、WordLen、InDic 和 Num 5 个特征,得到统计特征向量 (Word, POS, WordLen, InDic, Num)。其次,CBOW 词向量学习器以文本语料为训练数据,得到词向量 (v_1, v_2, \dots, v_m);结合领域词典,根据前文所阐述的相似度计算方法得出相似度,该相似度为连续的实数。由于 CRF 模型的输入

特征为离散特征,所以将计算出的相似度离散化,得到相似度特征 Simi。然后,将统计特征向量与相似度特征 Simi 拼接为一个特征向量 (Word, POS, WordLen, InDic, Num, Simi),作为 CRF 模型的输入特征。最后,用标注好的训练数据训练 CRF,得到领域术语识别模型;用该术语识别模型在测试数据上进行术语的标注任务,得到领域术语识别的结果。

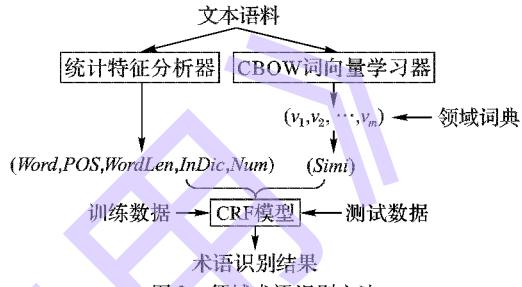


图 2 领域术语识别方法

3 实验

3.1 实验语料

实验所用语料分为两部分:渔业领域语料和通用语料。

第一部分语料来源于课题组已有的渔业领域语料,分别为《水产辞典》^[13]和海洋渔业领域的国家和地方标准文档。《水产辞典》分为渔业资源、水产捕捞、水产增养殖等 10 个类别,去掉文中特殊符号和插图,得到约 55 万字的文本;标准文档包括海水养殖类的文档和水产品类的文档,选取 167 篇,经过预处理后得到约 37 万字的文本。所以渔业领域语料共有约 92 万字的文本。

第二部分语料来源于公开的中文文本语料。本文研究的领域术语识别问题中的词语属于严谨的科学术语,中文语料库中新闻语料的词语用词相对严谨、规范,并且涉及的领域范围广,所以通用语料选择来自搜狗实验室发布的新闻语料库 SogouCA^[14],该语料库包括 2012 年 6 月到 7 月期间国内、国际、社会等 18 个频道的新闻数据,共计约 67 万张新闻网页,压缩后大小约为 436 MB。每条数据由 4 个数据项构成,具体文档格式如表 1 所示。

表 1 SogouCA 的数据格式

数据项	说明	示例
docno	Sogou 为每个文档分配的唯一 ID	< docno > 35d7097793e7bc78-49f37189a1acd500 </ docno >
url	抓取该文档的 URL	< url > http://biz.cn.yahoo.com/ypen/20120615/1115731.html </ url >
contenttitle	文档的标题	< contenttitle > 三部委决定在上海试行启运港退税政策 </ contenttitle >
content	文档的正文内容	< content > 关于在上海试行启运港退税政策的通知..... </ content >

3.2 语料标注及统计特征提取

语料标注和统计特征提取都需要使用文本语料中的词语,所以首先进行分词。采用中国科学院计算技术研究所的分词软件 ICTCLAS2016 对所有的语料进行分词并去除停用词。

由于渔业领域没有公开的标注好的数据,所以本实验数据的标注由领域专家通过人工方式采用 {B,I,O} 符号集进行标注。若完全标注工作量较大,所以选取渔业领域语料中的 10 万个词语进行标注,并将这部分语料作为术语识别模型的训练和测试语料。标注后对该 10 万个词进行统计特征提取。提取统计特征所用的领域词典为《水产辞典》^[13],收录了约 5200 个领域术语。5 个统计特征提取由统计特征分析器自动完成:分词软件可直接获得词本身 Word 和词性 POS 两个特

征;根据 UNICODE 编码方式,得到词的长度特征 WordLen;根据 UNICODE 字母表中汉字的编排顺序,得到汉字的偏旁部首,建立汉字的偏旁部首词典,计算词语包含特定偏旁部首的字符的数目,得到特征 Num;为判定词语是否在领域词典中,构建前缀词典树,若词语在词典中,特征值为 1,否则特征值为 0,得到特征 InDic。

3.3 词向量训练和相似度特征计算

3.3.1 词向量训练

词向量由实现了 CBOW 模型的开源 word2vec^[15]训练得到。不同的训练语料会得到不同的词向量,而词向量会间接影响术语的识别效果,所以本文分别用通用语料、渔业领域语料和混合语料(两种语料的合集)作为词向量的训练数据,训练得到

50 和 200 维的词向量。实验中的上下文窗口参数设为 5; 在通用语料和混合语料上训练时,词的采样频率设为 0.001, 初始学习率设为 0.025, 在渔业领域语料上训练时,词的采样频率设为 1.0, 初始学习率设为 0.0015; 学习率随着训练进行自适应调整。训练得到的词表和模型数据如表 2 所示。

表 2 词向量的训练结果

语料	词表大小	向量维数	模型大小/MB
通用语料	752 528	50	156
		200	587
渔业领域语料	19 749	50	3.91
		200	15.20
混合语料	759 259	50	158
		200	592

根据文献[16],主要有两种方法对词向量的效果进行评价。第一种方法从语言学的角度进行评价,也就是评价词向量本身的语义含义,如表 3 所示。第二种方法用词向量对自然语言处理任务的效果的提升来评价,这也是本文研究的术语识别任务的最重要的评价指标,在 3.4 节的实验中详细阐述。

表 3 中共选取了三个样例词语分别代表三类词语,其中前两个选取的是通用领域的词语:第一个为没有语义含义的词语“北京”,第二个为具有一定语义含义的词语“报道”;第三个选取的是渔业领域的词语“养殖”。同时给出与样例词语在向量空间上最相似的前五个词语,以及样例词语与每个词语的词向量的相似度。根据表 3 中的数据,从语义角度,对不同类别的词语、在不同语料和不同维度时的词向量的质量进行了分析,可以得出以下两点结论。

表 3 训练得到的词向量的效果评价

样例词语	语料	向量维度	与样例词语在向量空间上相似的词语/相似度
北京	通用语料	50	历届/0.6869, 临近/0.6762, 书法展/0.6485, 倒计时牌/0.6441, 雅典/0.6065
		200	姚仲三/0.6308, 梅冬/0.5797, 陈云朋/0.5795, 奥运会/0.5731, 奥运/0.5366
	渔业语料	50	无缝/0.5542, 夹带/0.5066, 藏书/0.5028, 多余/0.4991, 毁灭性/0.4937
		200	续/0.2601, 印染/0.2529, 万/0.2473, 船尾/0.2467, 直属/0.2379
	混合语料	50	书法展/0.6669, 倒计时牌/0.6463, 会旗/0.6342, 喜迎/0.6325, 迎/0.6261
		200	清运/0.4988, 前段/0.4925, 历届/0.4786, 雅典/0.4778, 悉尼/0.4708
	通用语料	50	介绍/0.6874, 引述/0.6769, 透露/0.6570, 援引/0.6381, 消息/0.6346
		200	引述/0.6657, 路透社/0.6532, 法新社/0.6491, 刊发/0.6422, 援引/0.6384
报道	渔业语料	50	竹竿/0.5155, 委内瑞拉/0.4715, 况/0.4541, 前额/0.4513, 物件/0.4423
		200	主干道/0.2751, 伸缩/0.2704, 竹叶/0.2692, 拟/0.2565, 传真/0.2478
	混合语料	50	东森/0.6584, 引述/0.6522, 熊伊眉/0.6433, 发言人/0.6415, 路透社/0.6342
		200	消息/0.5524, 介绍/0.4604, 了解/0.4447, 莫尼塔/0.4264, 华文/0.4258
养殖	通用语料	50	加工/0.7252, 奶牛/0.7156, 加工厂/0.7030, 种植业/0.6981, 饲料/0.6706
		200	奶牛/0.7650, 养殖业/0.7621, 生猪/0.7313, 畜禽/0.7235, 畜牧业/0.7115
	渔业语料	50	鱼/0.7548, 资源/0.7261, 鱼类/0.7233, 中/0.7155, 渔业/0.7135
		200	鱼/0.8356, 渔业/0.8206, 网/0.8093, 水产/0.7712, 水/0.7641
	混合语料	50	畜牧业/0.7632, 马铃薯/0.7417, 种植业/0.7326, 加工/0.7153, 畜/0.7082
		200	饲养/0.5638, 奶牛/0.5616, 养牛/0.5122, 养殖业/0.4965, 经济林/0.4863

第一 对于通用领域词语,通用语料和混合语料上训练得到的词向量的质量高于渔业领域语料上训练得到的词向量的质量;向量维度方面,对于没有语义含义的通用领域词语(北京),词向量的维度对词向量的质量没有显著的影响,而对于有语义含义的通用词语(报道),高维度的词向量的质量高于低维度的词向量的质量。

第二 对于特定领域词语,从特定的语义角度看,渔业领域语料训练得到的词向量的质量高于通用语料和混合语料上训练得到的词向量的质量,而从通用的语义角度看,三种语料训练得到的词向量的质量没有明显差别;向量维度方面,高维度词向量的质量高于低维度词向量的质量。

通过以上两点得出:首先,验证了词向量能表达丰富的语义含义,语义上相似的词语的词向量也相似,所以在 3.4.1 节的实验中加入相似度特征与不加入该特征进行对比;其次,通用语料和领域语料上训练得到的词向量的质量不同,所以在 3.4.2 节中,对不同语料训练得到的词向量进行验证,说明词语的领域性对识别结果的影响;最后,向量的维度对于有丰富语义含义的词语的质量有一定的影响,所以在 3.4.2 节中,对 50 维和 200 维的词向量分别实验。

3.3.2 相似度特征计算

获取词向量之后,根据式(2),计算 3.2 节中标注好的词语的相似度。由于要为 CRF 模型的语料中的每个词语计算相似度,并且需要分别计算渔业领域语料和混合语料上的 50 维和 200 维的相似度,为提高计算效率,将《水产辞典》^[13]中的约 5200 个术语的特征向量从词表 V 的词向量集中提取出来存入领域辞典向量表,同时为其建立检索树,进一步提高计算效率。由于 CRF 需要离散特征,而根据式(2)计算的相似度为 0~1 的连续特征,所以将区间 [0,1] 均匀划分为 10 个子区间,将相似度离散化为 0~9 的特征值。

3.4 实验结果及分析

根据 3.3.1 节的分析,本文设计两组实验。

第一组为验证实验:验证词语的语义和领域性对领域术语识别效果的影响。首先只用传统的统计特征作为 CRF 的输入特征;然后加入用于表达语义和领域性的相似度特征,与没加入该特征的实验结果进行对比,验证相似度特征对识别效果的影响,识别效果采用准确率 P、召回率 R 和 F 测度三个评价指标进行评价,P、R 和 F 计算如下:

$$P = \text{识别出的正确术语数目} / \text{识别出的术语总数}$$

$R = \text{识别出的正确术语数目} / \text{样本中的术语总数}$

$$F = P * R * 2 / (P + R)$$

其中: P 、 R 和 F 的值在 0 和 1 之间。 P 和 R 的值越接近 1, 准确率或召回率越高。 F 测度为 P 和 R 的调和平均值, 表达了术语识别的综合效果。

第二组为对比实验: 在第一组实验基础上, 加入通用语料和混合语料训练得到的不同维度的词向量, 对比通用语料、向量维度对术语的识别效果的影响, 评价方法同第一组实验。

实验中的 CRF 利用开源 CRF++^[17] 实现。CRF 模型提供了两种类型的模板: Unigram 和 Bigram, 分别表示一元特征模板和二元特征模板, 本文采用前者。将前文得到的 CRF 模型的语料数据按照 4:1 划分为两部分: 前一部分作为训练数据, 后一部分作为测试数据。实验中, 模型对训练数据的拟合程度的参数 c 设置为 1; 表示特征出现的次数的参数 f 设置为 5, 模板中每个特征的窗口大小设为 5, 正则化算子选择 L2。

3.4.1 验证实验及分析

采用特征递增的方式定义 CRF 特征模板, 选择渔业领域语料训练得到的词向量计算相似度特征, 通过实验验证相似度特征对识别结果的影响, 其 P 、 R 和 F 值如表 4 所示。

表 4 增加相似度特征的识别结果

指标	特征					
	1	1, 2	1, 2, 3	1, 2, 4	1, 2, 4, 5	1, 2, 4, 5, 6 (200 维)
P	0.9793	0.9802	0.9795	0.9797	0.9809	0.9855
R	0.9367	0.9357	0.9358	0.9362	0.9375	0.9439
F	0.9575	0.9574	0.9572	0.9575	0.9587	0.9643

实验中, 每次增加一个特征。1) 只有特征 1“Word”时, 准确率、召回率和 F 测度分别为 0.9793、0.9367 和 0.9575。2) 加入特征 2“POS”后, 其准确率略有提高, 但召回率和 F 测度略有下降, 说明词性对术语识别的效果影响不大, 在领域术语识别中可将该特征作为参考。3) 加入特征 3“WordLen”后, 准确率较上一步实验下降, 召回率较第一步实验略有下降, F 测度较上一步实验下降, 所以该特征未能提高术语识别效果, 说明词语的长度不是识别术语的有效特征, 在领域术语识别中不建议选取该特征; 接下来的实验中去掉该特征。4) 加入特征 4“Num”后, 三个指标较上一步实验均有提高, 所以该特征对于提高识别效果有积极的影响, 说明词语的字符结构是识别术语的有效特征, 在术语识别中, 可将该特征作为领域术语识别的重要特征。5) 加入特征 5“InDic”后, 3 个指标较上一步实验均有提高, 所以该特征对于提高识别效果有重要的正面影响, 说明词典对领域术语识别具有重要作用, 在术语识别中, 建议将该特征作为领域术语识别的重要特征。6) 加入特征 6“Simi”, 这里只考察相似度特征对识别效果的影响, 并不关注向量的维数, 所以只选择 200 维的词向量得到的相似度, 三个指标分别为 0.9855、0.9439 和 0.9643, 识别效果较之前有显著的提高, 验证了本文将词语的语义和领域性融入术语识别模型中的积极作用。

3.4.2 对比实验及分析

在验证实验的基础上, 将词向量分别替换为通用语料和混合语料训练得到的不同维度的词向量, 计算相应的相似度, 进行不同语料及不同维度的向量的相似度的对比实验, 实验结果如表 5 所示。

表 5 不同语料、不同维度的词向量的识别结果

指标	语料(维度)					
	渔业(50)	渔业(200)	通用(50)	通用(200)	混合(50)	混合(200)
P	0.9842	0.9855	0.9807	0.9812	0.9815	0.9824
R	0.9423	0.9439	0.9403	0.9419	0.9421	0.9428
F	0.9628	0.9643	0.9601	0.9612	0.9614	0.9622

通过对对比实验的结果分析得出以下 3 点结论:

1) 渔业领域语料训练得到的 50 和 200 维词向量计算得到的相似度特征, 前者的识别效果略差于后者。

2) 通用语料训练得到的 50 和 200 维词向量计算得到的相似度特征, 总体结果差于渔业领域语料训练得到的词向量, 说明词向量的训练语料来源于特定领域对识别效果有重要的积极作用。当然, 由于本文研究的背景是识别渔业领域文本中的术语, 所以 CRF 模型的测试语料来自该领域, 这也是得到这一结论的原因之一。

3) 混合语料训练得到的 50 和 200 维词向量计算得到的相似度特征, 总体结果差于渔业领域语料训练得到的词向量, 但好于通用语料训练得到的词向量。

综上 3 点, 若术语识别的应用背景是特定领域文本中的术语识别, 则选用该领域语料训练的词向量的效果较好; 同时, 高维度的词向量得到的相似度特征要好于低维度的词向量得到的相似度特征。

4 结语

本文利用词向量和 CRF 模型解决自然语言处理领域的术语识别问题。充分考虑领域词语的语义特性和领域特性, 将其融入基于 CRF 的术语识别模型中, 较传统的仅利用统计分布特性的方法进一步提高了领域术语识别的精度。分别在渔业领域语料、通用语料和混合语料上进行实验, 分析本文方法对术语识别效果的准确率、召回率和 F 测度的影响, 验证了将相似度特征加入 CRF 后, 对识别效果有显著的提升, 表明利用领域术语具有丰富的语义含义和很强的领域特性这两个特点对识别结果的积极作用; 并且在实验的结论中给出了解决领域术语识别问题的建议。本文方法可以推广到其他领域中, 但由于特定领域语料通常相对较少, 而词向量的质量与语料规模直接相关, 那么, 针对领域数据量少的普遍情况, 能否在小样本下进行更准确的术语识别, 这是进一步需要解决的问题; 另外, 本文以人工方式选取词语的多种特征进行术语识别, 能否减少或取消对人工选取特征的依赖, 即采用弱监督或无监督的机器学习方法, 直接利用原始文本进行术语识别, 也是下一步要研究的问题。

参考文献:

- [1] 祝清松, 冷伏海. 自动术语识别存在的问题及发展趋势综述[J]. 图书情报工作, 2012, 56(18): 104–109. (ZHU Q S, LENG F H. Existing problems and developing trends of automatic term recognition[J]. Library and Information Service, 2012, 56(18): 104–109.)
- [2] 吴海燕. 基于互信息与词语共现的领域术语自动抽取方法研究[J]. 重庆邮电大学学报(自然科学版), 2013, 25(5): 690–693. (WU H Y. Automatic domain term extraction based on word co-occurrence and mutual information[J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2013, 25(5): 690–693.)
- [3] 李丽双, 王意文, 黄德根. 基于信息熵和词频分布变化的术语抽

- 取研究[J]. 中文信息学报, 2015, 29(1): 82–87. (LI L S, WANG Y W, HUANG D G. Term extraction based on information entropy and word frequency distribution variety[J]. Journal of Chinese Information Processing, 2015, 29(1): 82–87.)
- [4] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional random fields: probabilistic models for segmenting and labeling sequence data[EB/OL]. [2016-05-51]. <http://www.seas.upenn.edu/~streln/bib/PDF/crf.pdf>.
- [5] 孙丽萍, 过弋, 唐文武, 等. 基于构成模式和条件随机场的企业简称预测[J]. 计算机应用, 2016, 36(2): 449–454. (SUN L P, GUO G, TANG W W, et al. Enterprise abbreviation prediction based on constitution pattern and conditional random field[J]. Journal of Computer Applications, 2016, 36(2): 449–454.)
- [6] 粟伟, 赵大哲, 李博, 等. CRF 与规则相结合的医学病历实体识别[J]. 计算机应用研究, 2015, 32(4): 1082–1086. (LI W, ZHAO D Z, LI B, et al. Combining CRF and rule based medical named entity recognition[J]. Application Research of Computers, 2015, 32(4): 1082–1086.)
- [7] 施水才, 王锴, 韩艳铧, 等. 基于条件随机场的领域术语识别研究[J]. 计算机工程与应用, 2013, 49(10): 147–149. (SHI S C, WANG K, HAN Y H, et al. Terminology recognition based on conditional random fields[J]. Computer Engineering and Applications, 2013, 49(10): 147–149.)
- [8] 刘海霞, 黄德根. 语义信息与 CRF 结合的汉语功能块自动识别[J]. 中文信息学报, 2011, 25(5): 53–59. (LIU H X, HUANG D G. Chinese functional chunk parsing employing CRF and semantic information[J]. Journal of Chinese Information Processing, 2011, 25(5): 53–59.)
- [9] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003, 3: 1137–1155.
- [10] MNIH A, HINTON G E. A scalable hierarchical distributed lan-

(上接第 3117 页)

- [8] ZHENG K, FUNG P C, ZHOU X. *K*-nearest neighbor search for fuzzy objects[C]// Proceedings of the 2010 ACM International Conference on Special Interest Group on Management of Data. New York: ACM, 2010: 699–710.
- [9] WANG L, CHEN H, ZHAO L, et al. Efficiently mining co-location rules on interval data[C]// Proceedings of the 6th International Conference on Advanced Data Mining and Applications. Berlin: Springer-Verlag, 2010: 477–488.
- [10] 欧阳志平, 王丽珍, 周丽华. 实例位置模糊的空间 co-location 模式挖掘研究[J]. 计算机科学与探索, 2012, 6(12): 1144–1152. (OUYANG Z P, WANG L Z, ZHOU L H. Mining spatial co-location patterns for fuzzy location of instances[J]. Journal of Frontiers of Computer Science and Technology, 2012, 6(12): 1144–1152.)
- [11] QIAN F, CHIEW K, HE Q, et al. Mining regional co-location patterns with *k*NN[G]. Journal of Intelligent Information Systems, 2013, 42(3): 485–505.
- [12] 温佛生, 肖清, 王丽珍, 等. 一种模糊对象的极大 co-location 模式挖掘算法[J]. 计算机科学, 2014, 41(1): 138–145. (WEN F S, XIAO Q, WANG L Z, et al. Algorithm of mining maximal co-location patterns for fuzzy objects[J]. Computer Science, 2014, 41(1): 138–145.)
- [13] FENWICK K D, MORRONGIELLO B A. Spatial co-location and infants' learning of auditory-visual associations[J]. Infant Behavior & Development, 1998, 21(4): 745–759.

guage model[C]// NIPS2008: Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2008: 1081–1088.

- [11] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J/OL]. [2015-08-16]. <http://arxiv.org/pdf/1301.3781v3.pdf>.
- [12] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]// NIPS2013: Advances in Neural Information Processing Systems 26. Cambridge, MA: MIT Press, 2013: 3111–3119.
- [13] 潘迎捷. 水产辞典[M]. 上海: 上海辞书出版社, 2007: 1–353. (PAN Y J. Dictionary of Fisheries[M]. Shanghai: Shanghai Lexicographical Publishing House, 2007: 1–353.)
- [14] 搜狗全网新闻数据[EB/OL]. [2015-09-08]. <http://www.sogou.com/labs/dl/ca.html>. (SogouCA[EB/OL]. [2015-09-08]. <http://www.sogou.com/labs/dl/ca.html>.)
- [15] word2vec [EB/OL]. [2015-11-02]. https://github.com/NLPchina/Word2VEC_java
- [16] LAI S, LIU K, XU L, et al. How to generate a good word embedding? [J]. IEEE Intelligent Systems, 2015, III(2): 1.
- [17] CRF ++[EB/OL]. [2015-11-25]. <http://sourceforge.net/projects/crfpp/files>.

Background

FENG Yanhong, born in 1980, M. S., lecturer. Her research interests include natural language processing, information retrieval.

YU Hong, born in 1968, Ph. D., professor. Her research interests include data mining, information retrieval.

SUN Geng, born in 1979, M. S., associate professor. His research interests include embedded system.

ZHAO Yujin, born in 1990, M. S. candidate. Her research interests include data mining, information retrieval.

- [14] 王新洲. 论空间数据处理与空间数据挖掘[J]. 武汉大学学报: 信息科学版, 2006, 31(1): 1–4, 8. (WANG X Z. Spatial data processing and spatial data mining[J]. Geomatics and Information Science of Wuhan University, 2006, 31(1): 1–4, 8.)

- [15] DUAN M, XIE X. Co-location visual pattern mining for near-duplicate image retrieval: US 8073818[P]. 2011-12-06.

- [16] LIU X, PEDRYCZ W. Axiomatic Fuzzy Set Theory and Its Applications[M]. Heidelberg: Springer-Verlag, 2009: 244.

Background

This work is partially supported by the National Natural Science Foundation of China (61370050, 61572036), the Natural Science Foundation of Anhui Province (1508085QF134), the Innovation Foundation of Anhui Normal University (2016XJJ074).

YU Qingying, born in 1980, Ph. D. candidate, lecturer. Her research interests include spatial data processing, information security.

LUO Yonglong, born in 1972, Ph. D., professor. His research interests include information security, spatial data processing.

WU Qian, born in 1994. Her research interests include data mining.

CHEN Chuanming, born in 1981, Ph. D. candidate, associate professor. His research interests include data mining, intelligent computing.