基于句法分析与词向量的领域新词发现方法

赵志滨[1] 石玉鑫[1] 李斌阳[2]
(东北大学计算机科学与工程学院 沈阳 110819)[1] (国际关系学院信息科技学院 北京 100091)[2]

摘 要 很多已经存在的词汇和词组可能会被运用于它们之前从未被运用过的领域文本中,这样的词汇或词组被称为领域新词。领域新词的发现可以为该领域的研究人员提供最新的领域发展动态,帮助其分析该领域的最新舆情,因此具有非常重要的意义。针对领域新词发现这一问题,文中提出了一种基于依存句法分析与词向量的领域新词发现方法。首先,提出了句法词典的概念,并基于依存句法分析,结合TF-IDF值的计算,提出了构建领域句法词典的方法;然后,使用领域句法词典,结合词向量技术,完成了领域新词发现方法的设计;最后,使用来自于护肤品论坛的真实文本数据集对所提方法进行了正确性验证。实验结果表明,构建的句法词典的质量较高,所提方法在进行领域新词发现时具有良好的性能。

关键词 句法分析,词向量,领域新词发现,句法词典
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.06.003

Newly-emerging Domain Word Detection Method Based on Syntactic Analysis and Term Vector

ZHAO Zhi-bin[1] SHI Yu-xin[1] LI Bin-yang[2]
(School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China)[1]
(School of Information Science and Technology, University of International Relations, Beijing 100091, China)[2]

Abstract Many existing words and phrases may be used in a domain in which they have never appeared before. These words and phrases are called newly-emerging domain words. The researchers can get insight into the latest development tendency and public opinions of a domain through these newly-emerging words. Therefore, it is significant to detect newly-emerging domain words. Based on dependency syntactic analysis and term vector, this paper proposed a newly-emerging domain words detection method. Firstly, the concept of syntactic dictionary was proposed, and its constructing method was proposed for some specific domains based on the dependency syntax of sentences and TF-IDF values of training corpus. Next, domain syntactic dictionary and term vectors were used to detect newly-emerging domain words. The comprehensive experiments were conducted to evaluate the proposed method with comment data from a skin-care products forum. The experimental results show that the syntactic dictionary is effective and the proposed method has good performance in newly-emerging domain word detection.

Keywords Syntactic analysis, Term vector, Newly-emerging domain words, Syntactic dictionary

1 引言

在互联网时代的大背景下,网络每时每刻都产生着大量的文本数据。在这些文本中,很多已经存在的词汇或词组被运用于新的领域。例如,“美白”一词经常用于描述洗面奶、面霜等产品的功效,但是假设某天有人发明了一款具有美白功能的香水,那么网络中对于这款香水的描述、评论、评测文本中都可能会出现“美白”一词,“美白”这个词就被运用到了香水领域。同时,由两个或多个词汇组成的一个词组也可能被运用于它从未出现过的领域。例如,某款面霜的说明书中出现了“补充黑色素”的功能,虽然“补充”和“黑色素”两个词可能在与面霜相关的文本中并不是新词,但它们作为一个词组是第一次出现在该领域的,代表了面霜产品的一个新功能。本文将类似于上述两个例子中的词汇和词组,即在某个领域文本中最新出现的有意义的词汇或者词组定义为领域新词。

如果能够从海量的网络文本中挖掘出类似于上述例子中的有意义的领域新词,就可以为相关研究人员提供该领域的最新信息,以更好地了解该领域的最新动态和发展趋势。例如,市场营销人员可以根据网络中某类产品的商品评论、评测文章中出现的领域新词来发现当前市场上该品类产品的新功

到稿日期:2018-08-18 返修日期:2018-10-22 本文受国家重点研发计划项目(2018YFB1004700),国家自然科学基金项目(61472070),航天专业部新技术研究高校合作项目(SKX182010023)资助。

赵志滨(1975-),男,博士,副教授,CCF会员,主要研究方向为分布式计算、Web数据挖掘和大数据管理,E-mail:zhaozb@mail.neu.edu.cn(通信作者);**石玉鑫**(1994-),男,硕士生,主要研究方向为Web数据挖掘;**李斌阳**(1982-),男,博士,副教授,CCF会员,主要研究方向为自然语言处理、社会计算。

能、新成分等最新信息，从而了解市场营销的最新动态。而从微博等社交网站中挖掘的领域新词可以帮助相关部门更好地了解社会舆情，掌握社会事件的最新动态。因此，领域新词的发现具有非常重要的意义。

如果只需要单纯地提取出某个领域文本中未出现过的词汇或者词组，那么只需要对该文本进行分词和对比即可。但是，这样提取出的词汇或词组往往有很大一部分都是无意义的。因此，我们需要找到一种有效的方法来发现有意义的词汇或词组，从而发现领域新词。

针对领域新词的发现问题，本文提出了一种基于依存句法分析与词向量的领域新词发现方法。首先使用以依存句法分析为基础的方法构建句法词典，之后将句法词典与词向量技术相结合进行领域新词的发现。

本文的主要贡献如下：

1)给出了“依存词对”和“句法模板”概念的定义，并且提出了一种新形式的词典——句法词典，为后续的新词发现提供了可靠的基础方法；

2)提出了句法词典的构造方法，即选择已标注的某一领域的文本对其进行依存句法分析，抽取出句法模板，并利用句法模板结合 TF-IDF 值的计算构建了句法词典；

3)利用贡献 2)中构建的句法词典，使用 word2vec 工具生成词向量，并利用词向量计算语义的相似度，进而发现领域新词。

本文第 2 节总结了近年来新词发现的最新研究成果，以及依存句法分析和词向量技术的最新应用和发展；第 3 节明确了本文要解决的问题，对问题进行了举例说明，并给出了问题的形式化定义；第 4 节介绍了本文工作的总体框架，并分别讨论了构建句法词典和发现领域新词这两方面的工作；第 5 节介绍了实验的设计及结果评价；最后总结全文，并分析了未来可继续改进的方向。

2 相关工作

新词发现通常是指对某一时刻之前未登录词汇的发现。新词发现始终是学术界的一个热点问题，相关研究众多。杨阳等[1]使用基于统计量的方法发现微博中的新词，然后利用 word2vec 工具判断新词与已有情感词的相关性，从而判断新词是否为情感词汇，实现了对情感新词的发现。Liang 等[2]提出了一个无监督的新词检测框架，用于检测推特上的中文新词，该框架不依赖于训练数据，而是用已有的词汇来标注新词。Yan 等[3]针对中文金融领域的新词发现问题，提出了一种动态的特征提取方法，该方法利用迭代模式来发现金融领域中文文本中的新词。Su 等[4]分析了经典统计量方法在微博文本中进行新词发现的性能，并针对统计量方法的缺陷和微博文本的特点，提出了一种基于分支熵的微博新词发现方法。Wang[5]提出了一种基于计算重复内容的广义后缀树来实现对候选新词的提取，并使用互信息和信息熵等统计量来筛选新词。Shen 等[6]利用未标记数据中高频率子串的特征来解决分词时的新词识别问题，提出了一种简单而有效的方法来提取特定类型的高频率子串，这些子串提供了对未知单词边界的良好估计，并应用后处理技术有效减少了所提取子串中的噪声。Xu 等[7]提出了一种基于支持向量机的新词发现方法，并通过实验证明了该方法可以提高新词识别的准确率和召回率。

依存句法分析是对句子结构的分析，即对句子中词汇之间依存关系的分析。依存句法分析在自然语言处理领域被广泛应用。He 等[8]提出了一种基于依存句法分析的评论观点挖掘方法，该方法可以有效地从评论中挖掘观点。Li 等[9]将 LDA 模型和依存句法分析结合为一个算法框架，提出了一种有效的微博舆情监测方法。在这个算法框架中，依存句法分析有效地提高了舆情监测的召回率。针对机器翻译、语音识别、信息检索中汉语词汇容易出现歧义的问题，史兆鹏等[10]提出了一种基于依存句法分析的词义消歧方法，其提高了消歧义的准确率。Guo 等[11]提出了一种改进的依存句法分析方法，并基于这种方法来分析微博情感倾向，取得了良好的效果。

词向量是使用算法对自然语言中的词汇进行向量化后得到的数学模型。关于词向量在自然语言处理中的应用，近年来已有了许多研究，其中不乏与依存句法分析结合使用的案例。Zhi 等[12]利用训练好的词向量来构建情感词典和方面词典，并结合依存句法分析来实现方面抽取和情感判定。Lin 等[13]提出了一种基于张量空间的词向量模型，将文本映射到张量特征空间，从而能够更好地解决社交网络中的年龄预测问题。Hayran 等[14]基于词向量和融合技术，提出了一种新的情感分析方法，对推特上的数据进行了情感分析。该方法降低了词向量的维度，避免了维数灾难，并提高了情感分析的准确率。Meng 等[15]提出了一种基于词向量对不同语言的概念进行匹配的方法，结合双语词典，其可以对不同语言的概念进行匹配。Kusner 等[16]定义了 Word Mover's Distance，用 word2vec 工具生成词向量，将求解句子之间相似度的问题转化为一个优化问题，从而可以使用词向量来计算句子之间的相似度。

3 问题描述

本文要解决的问题是领域新词的发现问题。领域新词的发现与传统的新词发现有所区别。传统的新词发现是指对某一时刻前词典中未登录词汇的发现，被发现的对象往往是单个的词汇，并且这些词汇在任何领域的文本中都从未出现过。

本文要发现的是某一领域中未出现过的词汇和词组，这与传统的新词发现问题主要有两点不同：1)本文要发现的领域新词只是在某个领域从未出现过，而不是在所有领域都未出现过；2)本文所指的领域新词，有可能是一个词汇，也可能是一个词组，即多个词汇的集合。通过发现领域新词可以挖掘出该领域最新的发展动态，例如对某类产品评论中的领域新词进行发现，可以帮助人们了解该类产品当前最新出现的功能、成分、包装等。本文所描述的领域新词的形式化定义如下。

定义 1(领域新词) 在某个时刻 t_0，在一段描述领域 B 的文本集合 T_B 中出现了一组在 t_0 时刻前从未在文本集合 T_B 中出现过并且有实际含义的词汇集合 $W_i=\{w_1,w_2,\cdots,w_n\}(n\geqslant 1)$，那么这个词汇集合 W_i 就叫做领域 B 的领域新词。

表 1 列出了从某日化用品论坛抓取的文本中挑选出的 3

个例句。例句 r_1 是某品牌洗发水产品的广告，其中“维他命 B5”一词第一次出现在该商品的相关文本中，因此可以认为该品牌产品第一次添加“维他命 B5”这一成分，即“维他命 B5”在该品牌的洗发水这个领域范围内是一个领域新词。例句 r_2 中出现了“补充黑色素”这一词组。显然，在化妆品领域的文本中，“补充黑色素”是一个领域新词。例句 r_3 是一款护手霜的广告，“蕴含玉米须萃取物”同样是一个领域新词。可以看到，例句 r_2 和例句 r_3 的领域新词都是词组（“补充＋黑色素”和“蕴含＋玉米须＋萃取物”），而不是单个的词。

综上所述，本文的目标研究问题是：针对领域 B，设计领域新词发现函数 $F_B^{t_0}$，挖掘文本集合 T_B 在某一时刻 t_0 之后出现的领域新词集合 $W=\{W_1,W_2,\cdots,W_m\}$。这个问题可以形式化地表示为：$F_B^{t_0}:T_B\rightarrow W$。

表 1　领域新词的举例

Table 1　Examples of newly-emerging domain words

编号	例句	领域新词
r_1	三大黄金营养——脂醇、维他命 B5 和氨基酸帮助发丝锁住营养成分，为发丝注入能量并防止发丝氧化	维他命 B5
r_2	曾经有补充黑色素的化妆品，后因致癌而被禁用	补充黑色素
r_3	这 4 款护手霜都蕴含玉米须萃取物及天然精油成分，富含维生素 E 及 B5，能有效抗氧化	蕴含玉米须萃取物

4　领域新词发现方法的设计

4.1　总体框架

领域新词发现方法的总体框架分为两部分：句法词典的构建及领域新词的发现。领域新词发现的流程如图 1 所示。

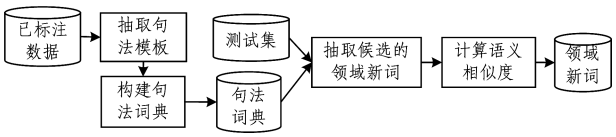

图 1　总体框架

Fig. 1　Overall Framework

1）句法词典的构建。在进行领域新词的发现之前，需要确定新词的边界，因此需要用已有的语料构建一个非新词词典。本文提出了一种新形式的词典——句法词典。首先，对领域 B 的文本集合 T_B 进行标注，标注的内容是我们想要获得的属性或领域信息。例如，若想知道面霜这个品类的最新功能和最新成分，则可以抓取护肤品论坛中有关面霜的文本后对其进行标注，并以短语的形式标注出文本中描述了哪些功能和成分。然后，利用依存句法分析技术对标注出的短语进行一定规则的匹配，挖掘出现频率较高的句法模板，并结合 TF-IDF 值的计算提取出符合这些模板的词汇或词组，构建初始句法词典。所构建的领域句法词典是下一步领域新词发现工作的基础。

2）领域新词的发现。运用已经构建的句法词典，结合依存句法分析和词向量技术，对某个领域 B 的最新文本集合 T_B' 进行领域新词的发现。首先，依据第一步中抽取出的句法模板，从文本集合 T_B' 中抽取出符合这些模板的词汇或词组。然后，通过词向量技术以及一定规则对每一个词汇或词组与其对应模板的句法词典中的词汇或词组进行相似性计算，相似性高于一定阈值的词汇或词组即可被认为是领域新词。

4.2　构建领域句法词典

在构建词典之前，首先需要收集领域 B 的文本集合 $T_B=\{c_1,c_2,\cdots,c_n\}$，其中 $c_1,c_2,\cdots,c_n$ 是文本集合 T_B 中的 n 条文本，一条文本可能是一个句子，也可能是一段由多个句子构成的语义连贯的长文本。然后，对文本集合 T_B 进行标注。可以将领域 B 中有价值的属性分为 $A_1,A_2,\cdots,A_m$ 等 m 个属性，对于文本 $c_i\in T_B$，需要将描述这些属性的文本分别标注出来。标注完成后，文本 c_i 将对应一个标签集合 $L_i=\{a_1,a_2,\cdots,a_m\}$，其中 $a_1,a_2,\cdots,a_m$ 是文本 c_i 中分别对应属性 $A_1,A_2,\cdots,A_m$ 的标签集合。在属性 A_i 的标签集合 $a_j=\{l_1,l_2,\cdots,l_p\}(a_j\in L_i)$ 中，$l_1,l_2,\cdots,l_p$ 均为文本 c_i 原文的一部分，被称为标签。表 2 列出了人工标注的一个示例，该示例文本来自某护肤用品论坛。从表中可以看出，人工标注了功能、成分、质地 & 泡沫等 3 个属性的标签。其中，功能属性标注了标签“防晒”，成分属性标注了标签“温和”，质地 & 泡沫属性标注了“清爽”“不油腻”两个标签。

表 2　人工标注的举例

Table 2　Manually annotated example

示例文本	功能	成分	质地 & 泡沫
这款防晒霜最赞的就是全家可以一起用哦。3 岁以上的小朋友用它防晒也完全没问题的，温和的成分不会对小朋友嫩嫩的肌肤有损伤。妈妈的敏感肌也可以放心去用了，涂上薄薄一层，很清爽，不油腻，超清薄。	防晒	温和	清爽 不油腻

本文基于依存句法分析和词向量进行领域新词的发现。依存句法分析的目标是通过分析句子的依存句法来构建该句子的依存句法树，从而描述出句子中词汇之间的依存关系。以句子“这种美白霜的美白效果很好。”为例，利用哈工大 LTP[17] 工具得到的依存句法树如图 2 所示。

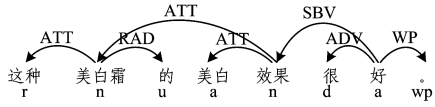

图 2　依存句法分析的实例

Fig. 2　Example of dependency syntactic analysis

图 2 中的有向弧被称为依存弧，表示两个词之间存在从属关系。每个依存弧上都有一个标注，表示两个词之间的依存关系类型，每个词汇下方标注了它的词性。例如，“很”与“好”之间存在依存关系 ADV（状中结构），“很”是程度副词，修饰形容词“好”。“好”是这对关系中的核心词，也叫支配词；“很”是用来修饰支配词的词语，也叫从属词。类似于“很”和“好”这样的词对，本文将其称为“依存词对”，其形式化定义如下。

定义 2（依存词对）　依存词对指存在依存关系的两个词语，本文定义为一个依存词对，可形式化表示为：$WordPair(w_i,w_j)=(id_i,w_i,pos_i,id_j,w_j,pos_j,relation)$。其中，$id_i$ 是从属词 w_i 的词号，即该从属词在句子中的位置；pos_i 是 w_i 的词性；而 id_j 和 pos_j 分别是支配词 w_j 的词号和词性；$relation$ 是词汇 w_i 和 w_j 的依存关系类型。

例如，依存词对“很”和“好”可以形式化表示为：

$WordPair$(很,好)=(6,很,d,7,好,a,ADV)

在某个领域的文本集合中,某种词性组合的依存词对可能较为频繁地出现。以某护肤品论坛中的帖子为例,帖子中出现了"收缩毛孔""焕发光泽""滋养肌肤"等关于护肤品功能的描述,它们均为"动词+名词"形式的依存词对。同时,多个依存词对的组合可能也会频繁出现,例如"去除多余杂质""恢复正常酸碱值""提供充足水分"等描述,这些均为"动词+形容词+名词"的形式,其中也包含"动词+名词"形式的依存词对和"形容词+名词"形式的依存词对。对于某一领域的文本中类似于"动词+名词""动词+形容词+名词"等包含一个或多个依存词对的频繁词汇集合,本文称为"句法模板",其形式化定义如下。

定义 3(句法模板) 在领域 B 的文本集合 T_B 中,存在文本 c_i,包含词性为 $\{pos_1,pos_2,\cdots,pos_n\}$ 的词汇集合 $W_{c_i}=\{w_1,w_2,\cdots,w_n\}(n\geqslant 1)$,且对于集合 W_{c_i} 中的任意词汇 w_j,至少存在一个词汇 $w_k\in W_{c_i}$ 与其存在依存关系,构成依存词对 $WordPair(w_j,w_k)$ 或 $WordPair(w_k,w_j)$。假设与 c_i 具有上述相同性质(即包含词性序列 $\{pos_1,pos_2,\cdots,pos_n\}$ 并且该序列中每个词都至少属于一个依存词对)的文本集合为 T_e,T_e 中文本数量占 T_B 中文本数量的比例大于一个给定的阈值 θ,则称元组 $e=(pos_1,pos_2,\cdots,pos_n)$ 为领域 B 的一个句法模板,每个符合该句法模板的词汇组合都是句法模板 e 的一个实例。

前文提到,领域 B 中我们想要获取的信息可以分为 A_1,A_2,…,A_m 等 m 个属性。本文将构建一个基于句法模板并包含这 m 个属性信息的词典,用来保存文本集合 T_B 中所包含的不同属性的信息。句法词典的形式化定义如下。

定义 4(句法词典) 在领域 B 的文本集合 T_B 中,有句法模板集合 $e_1,e_2,\cdots,e_n$,其中任意一个句法模板 e_j 均存在实例集合 $W_{e_j}=\{W_1,W_2,\cdots,W_p\}$。假设我们想获取的领域 B 的信息分为 $A_1,A_2,\cdots,A_m$ 等 m 个属性,记集合 W_{ej} 中描述属性 A_k 的实例的集合为 $W(e_j,A_k)$,则这些集合可以构成一个新的集合 $D_B=\{W(e_j,A_k)\mid 1\leqslant j\leqslant n,1\leqslant k\leqslant m\}$,集合 D_B 就是领域 B 的句法词典。

为了将文本集合 T_B 中所包含的不同属性的信息提取出来,并且尽量提高词典的质量,在构建句法词典之前需要预先进行前文中提到的人工标注工作,然后对文本进行分词处理。本文使用哈工大的 LTP 工具对文本进行分词。分词后的词语对于领域 B 的重要程度,用该词语的 TF-IDF 值衡量。

TF-IDF 用来评估一个词汇对于一个文件的重要程度。TF 指的是某一个给定的词语在该文件中出现的频率;IDF 是逆向文件频率,是一个词语普遍重要性的度量。

将 T_B 看作一个文件,从微博上抓取一定数量的文本 $WE_1,WE_2,\cdots,WE_n$,将每条文本看作一个文件,与 T_B 组成文本集合 $WE=\{T_B,WE_1,WE_2,\cdots,WE_n\}$。对于词汇 $w_i\in T_B$,它对于 T_B 的 TF 值和 IDF 值的计算方式分别如式(1)和式(2)所示:

$$tf_{i,B}=\frac{n_{i,B}}{\sum_k n_{k,B}} \tag{1}$$

$$idf_i=\frac{|WE|}{|\{WE_j\mid w_i\in WE_j,1\leqslant j\leqslant n\}|+1} \tag{2}$$

其中,$n_{i,B}$ 是词汇 w_i 在本文集合 T_B 中出现的次数,$\{WE_j\mid w_i\in WE_j,1\leqslant j\leqslant n\}$ 是包含词汇 w_i 的微博文本集合。词汇 w_i 对于文本 T_B 的 TF-IDF 值的计算方法如式(3)所示:

$$tfidf_{i,B}=tf_{i,B}\cdot idf_i \tag{3}$$

根据词汇的 TF-IDF 值,可以构建一个重要词汇词典 $D_{\text{imp}}=\{w_i\mid tfidf_{i,B}>\theta'\}$,其中 θ' 是一个阈值,TF-IDF 值大于 θ' 的词汇均可看作领域 B 的重要词汇。

根据句法模板 e 的概念,从 T_B 中抽取出句法模板集合 $e=\{e_1,e_2,\cdots,e_n\}$。文本 c_i 标注的标签集合为 $L_i=\{a_1,a_2,\cdots,a_m\}$,其中 $a_1,a_2,\cdots,a_m$ 是文本 c_i 中分别对应属性 A_1,A_2,…,A_m 的标签集合,属性 A_k 的标签集合为 $a_k=\{l_1,l_2,\cdots,l_p\}(a_k\in L_i)$。对于 c_i 中符合句法模板 $e_j(e_j\in e)$ 的实例 $W_{cij}=\{w_1,w_2,\cdots,w_n\}(n\geqslant 1)$,若其满足以下两个条件之一,即可加入集合 $W(e_j,A_k)$。

1)存在标签 $l_s\in a_k$,对于 $W_{c_{ij}}$ 中任意一个词汇 w_q,均有 $w_q\in l_s$。

2)存在词汇 $w_q\in W_{c_{ij}}$,有 $w_q\in D_{\text{imp}}$,且存在标签 $l_s\in a_k$,使得 $w_q\in l_s$。

对 T_B 中所有句法模板的所有实例重复上述步骤,即可得到领域 B 的一个句法词典 $D_B=\{W(e_j,A_k)\mid 1\leqslant j\leqslant n,1\leqslant k\leqslant m\}$。

下面利用句法词典 D_B 进行下一步的领域新词发现方法的设计。

4.3 领域新词发现

4.2 节中使用依存句法分析的方法得到了句法词典 D_B,本节将借助句法词典 D_B,使用依存句法分析和词向量技术相结合的方法进行领域新词发现方法的设计。

设领域 B 在时刻 t_0 之后出现了一个新的文本集合 T'_B,且 $T_B\cap T'_B=\varnothing$。通过依存句法分析,我们可以从文本集合 T'_B 中提取出符合句法模板集合 $e=\{e_1,e_2,\cdots,e_n\}$ 的所有实例,并构成集合 W'_e,记为 $W'_e=\{W'_1,W'_2,\cdots,W'_n\}$,集合中的元素均为句法模板所对应的实例集合。对于模板 e_i 的实例集合 $W'_i=\{W'_{i1},W'_{i2},\cdots,W'_{im}\}(W'_i\in W'_e)$,若其中一个实例 $W'_{ij}\notin W(e_i,A_k)$,则可以通过计算 W'_{ij} 与 $W(e_i,A_k)$ 中已存在实例的相似度来确定 W'_{ij} 是否描述了领域 B 中与属性 A_k 有关的内容。如果可以确定 W'_{ij} 描述了领域 B 中与属性 A_k 有关的内容,那么 W'_{ij} 就是领域 B 中属性 A_k 的领域新词。因此,需要设计合适的方法来评价词组之间的相似度。

本文将两个词组 W_p 和 W_q 之间的相似度形式化表示为 $GroupSim(W_p,W_q)$,也称其为 W_p 和 W_q 的词组相似度。本文采用词向量技术来计算词组的相似度。

本文采用 Google 开源词向量工具 word2vec 进行词向量的训练,利用 Skip-Gram 模型,设定词向量的维度为 100 维。在得到词向量模型后,进一步进行词组的相似度的计算。本文使用两种不同的策略来计算词组的相似度。

第一种策略:将词组转化为一个向量,本文称这样的向量为词组向量。对于词组 $W_p=\{w_{p1},w_{p2},\cdots,w_{pn}\}$,其各个分词对应的词向量集合为 $\{\boldsymbol{v}_{p1},\boldsymbol{v}_{p2},\cdots,\boldsymbol{v}_{pn}\}$,记词组 W_p 的词组向量为 $\boldsymbol{v}_p$。进一步地,设计词向量转化函数 F_v,将词向量集合 $\{\boldsymbol{v}_{p1},\boldsymbol{v}_{p2},\cdots,\boldsymbol{v}_{pn}\}$ 转化为词组向量 $\boldsymbol{v}_p$,形式化表示为:F_v:

$\{\boldsymbol{v}_{p1},\boldsymbol{v}_{p2},\cdots,\boldsymbol{v}_{pn}\}\rightarrow\boldsymbol{v}_p$。本文所采用的转化方法有两种。

第一种转化方法（本文简称为 group-max 方法）将集合 $\{\boldsymbol{v}_{p1},\boldsymbol{v}_{p2},\cdots,\boldsymbol{v}_{pn}\}$ 中每个向量的第 i 维 $s^i_{p1},s^i_{p2},\cdots,s^i_{pn}$ 进行比较，并取最大值作为向量 $\boldsymbol{v}_p$ 的第 i 维 s^i_p。因此，向量 $\boldsymbol{v}_p=\{s^1_p,s^2_p,\cdots,s^{100}_p\}$ 中的任一元素 s^i_p 都可以由式(4)得到：

$$s^i_p=\max(s^i_{p1},s^i_{p2},\cdots,s^i_{pn})(1\leqslant i\leqslant 100) \tag{4}$$

第二种转化方法（本文简称为 group-avg 方法）将集合 $\{\boldsymbol{v}_{p1},\boldsymbol{v}_{p2},\cdots,\boldsymbol{v}_{pn}\}$ 中每个向量的第 i 维的平均值作为向量 $\boldsymbol{v}_p$ 的第 i 维 s^i_p。因此，向量 $\boldsymbol{v}_p$ 中任一元素 s^i_p 的计算公式如式(5)所示：

$$s^i_p=\mathrm{avg}(s^i_{p1},s^i_{p2},\cdots,s^i_{pn})(1\leqslant i\leqslant 100) \tag{5}$$

得到词组 W_p 和 W_q 的词组向量 $\boldsymbol{v}_p$ 和 $\boldsymbol{v}_q$ 后，采用余弦相似度计算方法，可以得到 W_p 和 W_q 之间的词组相似度 $GroupSim(W_p,W_q)$，如式(6)所示：

$$GroupSim(W_p,W_q)=\cos(\boldsymbol{v}_p,\boldsymbol{v}_q)=\frac{\boldsymbol{v}_p\cdot\boldsymbol{v}_q}{\|\boldsymbol{v}_p\|\|\boldsymbol{v}_q\|} \tag{6}$$

第二种策略：通过分别计算每个词汇之间的相似度来得到词组之间的相似度，计算方法有两种。

第一种方法（本文简称为 words-max 方法）分别计算每个词组对应位置的词汇之间的余弦相似度，并取这组相似度的最大值作为词组的相似度。计算方法如式(7)所示：

$$GroupSim(W_p,W_q)=\max(\{\cos(\boldsymbol{v}_{pi},\boldsymbol{v}_{qi})\mid 1\leqslant i\leqslant n\}) \tag{7}$$

第二种方法（本文简称为 words-avg 方法）分别计算每个词组对应位置的词汇之间的余弦相似度，并取这组相似度中的平均值作为词组的相似度。计算方法如式(8)所示：

$$GroupSim(W_p,W_q)=\frac{\sum_n\cos(\boldsymbol{v}_{pi},\boldsymbol{v}_{qi})}{n} \tag{8}$$

在句法词典 D_B 中，对于文本集合 T'_B 中符合模板 e_i 的未登录实例 W'_i，若存在已登录词组 $W_i\in W(e_i,A_k)\in D_B$，且两个词汇或词组的相似度大于一定的阈值 θ_{sim}，即 $GroupSim(W'_i,W_i)>\theta_{\mathrm{sim}}$，则可以认为 W'_i 是领域 B 中属性 A_k 的领域新词。

5　实验

5.1　实验数据集和预处理

本文的实验数据集由爬虫程序从某护肤品论坛抓取到的 500 个帖子和从新浪微博中抓取到的 1000 万条微博构成。将从护肤品论坛得到的语料库作为领域新词发现所需要的基础语料库，将从新浪微博中抓取到的数据用于训练 word2vec 模型。两者合并作为一个语料库，用于计算护肤品语料库中出现的词语的 TF-IDF 值。以上文本中，需要去除护肤品论坛中的无效帖子，并且将有效帖中的各种无意义符号去除。

在进行领域新词发现工作之前，首先对护肤品论坛中的帖子进行了 4.2 节中所提到的标注。由于人工标注难免有疏漏，因此对标注结果进行了细致的检查，并对 10% 的数据进行了重复标注。标注完成后，使用 LTP 对每个帖子进行了分句、分词、词性标注和句法分析。之后，用 LTP 对微博文本进行了分词，并去除停用词，将分词后的语料用于训练 word2vec 模型和计算 TF-IDF 值。

本文实验采用 Python 3.5 语言，数据库采用 MongoDB。为了提高处理效率，在两台物理机上构建了 6 个虚拟计算节点，平均分配数据，以实现均衡的并行处理。

5.2　领域新词发现实验的设计及结果评价

本文选择了已标注的 500 个帖子中的 400 条作为构建句法词典 D_B 的原始语料库 T_B。针对 T_B 中的每一条文本，预先进行了分词，之后去除停用词，并对其进行依存句法分析，抽取出句法模板。然后，结合 1000 万条微博语料，计算出语料库 T_B 分词后所得的每一个词汇的 TF-IDF 值，并结合句法模板构建句法词典 D_B。

在实验过程中选择了两种词典构建策略：策略 A 只将满足本文 4.2 节末尾条件 1）的词汇或词组加入词典，即只用字符串匹配的方法；策略 B 将同时满足本文 4.2 节末尾的条件 1）和条件 2）的词汇或词组都加入词典，即采用字符串匹配和计算 TF-IDF 值相结合的方法。其中，阈值 θ' 分别选择 0.001，0.01 和 0.05 3 个值来进行实验。本文采用查准率（Precision）、召回率（Recall）以及 F1-score 等 3 个指标来对实验结果进行评估，构建句法词典的实验结果如表 3 所列。

表 3　构建词典的实验结果

Table 3　Experimental results of constructing dictionary

策略	*Precision*	*Recall*	*F1-score*
策略 A	0.423	0.815	0.557
策略 B($\theta'=0.001$)	0.445	0.793	0.570
策略 B($\theta'=0.01$)	0.652	0.743	0.695
策略 B($\theta'=0.05$)	0.786	0.577	0.665

通过表 3 的实验结果可以看出，策略 A 构建的词典召回率较高，但是查准率过低；策略 B 可以有效地提高词典的查准率，但会降低召回率；随着阈值 θ' 的增大，查准率不断提高，而召回率不断降低。当采用策略 B，并且阈值 θ' 为 0.01 时，F1 值最高。实验结果表明，结合计算 TF-IDF 值的方法可以有效提高词典的查准率，但也会漏掉一些本该加入词典的词汇，降低召回率。其原因在于：TF-IDF 值是衡量词汇重要性程度的指标，因此可以筛选掉很多无意义的词汇或词组，但是也会过滤掉一些出现频率不高但是也很重要的词汇和词组。因此，选择合适的阈值 θ' 是决定词典质量的关键因素之一。本文将采用阈值 θ' 为 0.01 时构建的句法词典来进行后续的新词发现实验。

词典构造完成后，将分词后的 1000 万条微博用 Google 开源词向量工具 word2vec 进行词向量训练，生成 word2vec 模型，词向量的维数为 100。利用依存句法分析和词向量模型对余下的已标注的 100 条帖子进行领域新词的发现实验。

4.3 节共提出了 4 种方法来计算词组相似度：group-max 方法、group-avg 方法、words-max 方法和 words-avg 方法。本文将分别采用这 4 种不同的方法，结合不同的相似度阈值 θ_{sim} 来进行实验。

领域新词发现的实验结果如表 4 所列。从表 4 可以看出，使用 words-avg 方法和 words-max 方法计算词组相似度时的查准率、召回率均明显高于使用 group-max 方法和 group-avg 方法时的查准率、召回率，这说明 words-avg 方法和 words-max 方法要优于 group-max 方法和 group-avg 方法，而 words-avg 方法又要略优于 words-max 方法。以上说

明:在计算词组相似度时,如果我们将多个向量合并为一个向量来构建词组向量,则很可能会覆盖单个词汇的特征,从而影响最终的结果。同时,实验表明,阈值 θ_{sim} 对最终的结果有很大的影响,随着阈值的增大,4 种方法的准确率不断提高,但是召回率却不断降低;当 $\theta_{sim}=0.7$ 时,4 种方法的召回率有断崖式的下降。因此,需要选择合理的阈值 θ_{sim}。

表 4 领域新词发现的实验结果

Table 4 Experimental results of newly-emerging domain words

	$\theta_{sim}=0.5$			$\theta_{sim}=0.6$			$\theta_{sim}=0.7$		
	Precision	*Recall*	*F1-score*	*Precision*	*Recall*	*F1-score*	*Precision*	*Recall*	*F1-score*
group-max	0.375	0.412	0.393	0.423	0.384	0.403	0.608	0.193	0.293
group-avg	0.434	0.482	0.457	0.510	0.432	0.468	0.596	0.176	0.272
words-max	0.615	0.604	0.609	0.686	0.536	0.602	0.735	0.437	0.548
words-avg	0.643	0.680	0.661	0.703	0.653	0.677	0.792	0.423	0.551

结束语 本文提出了一种基于依存句法分析与词向量的领域新词发现方法,并在人工预标注一定量的数据的基础上进行了有关护肤品的领域新词发现实验。首先,使用依存句法分析的方法,结合 TF-IDF 值,构建了领域句法词典;然后,使用词向量技术,计算候选新词与词典中已登录词的相似度,从而完成领域新词的判定。实验证明,本文提出的方法在领域新词发现方面表现良好。

本文工作需要大量的标注,属于有监督学习,需要耗费大量的人力和物力,因此接下来将会考虑采用基于无监督学习的方法进行领域新词的发现。另外,可考虑引入迁移学习技术来提高本文方法对于新领域的适应性。

参 考 文 献

[1] YANG Y,LIU L F,WEI X H,et al. New methods for extracting emotional word based on distributed representation of words [J]. Journal of Shandong University(Natural Science),2014,49(11):51-58. (in Chinese)
杨阳,刘龙飞,魏现辉,等. 基于词向量的情感新词发现方法[J]. 山东大学学报(理学版),2014,49(11):51-58.

[2] LIANG Y,YIN P,YIU S M. New Word Detection and Tagging on Chinese Twitter Stream[C]// International Conference on Big Data Analytics and Knowledge Discovery. Cham:Springer,2015:310-321.

[3] YAN L,BAI B,CHEN W,et al. New Word Extraction From Chinese Financial Documents[J]. IEEE Signal Processing Letters,2017,24(6):770-773.

[4] SU Q L,LIU B Q. Chinese new word extraction from MicroBlog data[C]// International Conference on Machine Learning and Cybernetics. IEEE,2014:1874-1879.

[5] WANG F. Research on New Chinese Words Detection in Microblog[J]. Computer Engineering & Software,2015,36(11):6-8.

[6] SHEN M,KAWAHARA D,KUROHASHI S. Chinese Word Segmentation and Unknown Word Extraction by Mining Maximized Substring[J]. Journal of Natural Language Processing,2016,23(3):235-266.

[7] XU Y,GU H. New Word Recognition Based on Support Vector Machines and Constraints[C]// International Conference on Information Science and Control Engineering. IEEE,2015:341-344.

[8] HE T,HAO R,QI H,et al. Mining Feature-Opinion from Reviews Based on Dependency Parsing[J]. International Journal of Software Engineering & Knowledge Engineering,2017,26(9n10):1581-1591.

[9] LI Y,ZHOU X,SUN Y,et al. Design and Implementation of Weibo Sentiment Analysis Based on LDA and Dependency Parsing[J]. China Communications,2016,13(11):91-105.

[10] SHI Z P,ZOU X X,XIANG R Z,et al. Multi-feature Word Sense Disambiguation Based on Dependency Parsing Analysis [J]. Computer Engineering,2017,43(9):210-213. (in Chinese)
史兆鹏,邹徐熹,向润昭,等. 基于依存句法分析的多特征词义消歧[J]. 计算机工程,2017,43(9):210-213.

[11] GUO F,ZHOU G. Research on micro-blog sentiment orientation analysis based on improved dependency parsing[C]// International Conference on Consumer Electronics. IEEE,2014.

[12] ZHI S,LI X,ZHANG J,et al. Aspects Opinion Mining Based on Word Embedding and Dependency Parsing[C]// International Conference on Advances in Image Processing. ACM,2017:210-215.

[13] LIN Z,WANG Y. Age Prediction in Social Networks Based on Word Embedding and Tensor Learning[C]// International Conference on Communication and Electronic Information Engineering. Paris:Atlantis Press,2017.

[14] HAYRAN A,SERT M. Sentiment analysis on microblog data based on word embedding and fusion techniques[C]// Signal Processing and Communications Applications Conference. IEEE,2017.

[15] MENG F,LU W,XUE R. Mapping senses in BabelNet to Chinese based on word embedding[C]// International Congress on Image and Signal Processing,Biomedical Engineering and Informatics. IEEE,2018.

[16] KUSNER M J,SUN Y,KOLKIN N I,et al. From word embeddings to document distances[C]// International Conference on International Conference on Machine Learning. JMLR. org,2015:957-966.

[17] CHE W,LI Z,LIU T. LTP:a Chinese Language Technology Platform[C]// International Conference on Computational Linguistics:Demonstrations. Association for Computational Linguistics,2010:13-16.