(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 107315734 B
(45)授权公告日 2019.11.26

(21)申请号 201710308896.3

(22)申请日 2017.05.04

(65)同一申请的已公布的文献号
申请公布号 CN 107315734 A

(43)申请公布日 2017.11.03

(73)专利权人 中国科学院信息工程研究所
地址 100093 北京市海淀区闵庄路甲89号

(72)发明人 沙瀛 施振辉 李锐 梁棋 邱咏钦 王斌

(74)专利代理机构 北京君尚知识产权代理有限公司 11200
代理人 司立彬

(51)Int.Cl.
G06F 17/27(2006.01)

(56)对比文件
CN 103699667 A,2014.04.02,
CN 105512334 A,2016.04.20,
CN 104765763 A,2015.07.08,
CN 104216875 A,2014.12.17,
CN 105608075 A,2016.05.25,
CN 104584003 A,2015.04.29,
EP 1952266 A4,2010.01.20,
US 7873654 B2,2011.01.18,
WO 2013118435 A1,2013.08.15,
沙瀛等.中文变体词的识别与规范化综述.《信息安全学报》.2016,第1卷(第3期),第77-87页.

审查员 黄长霞

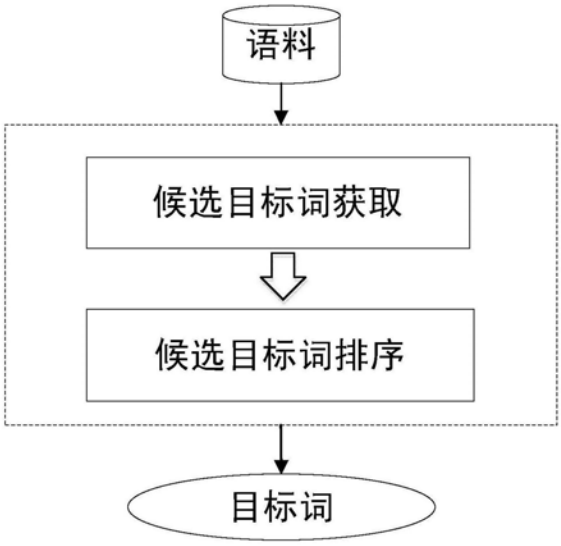
权利要求书2页 说明书9页 附图5页

(54)发明名称

一种基于时间窗口和语义的变体词规范化的方法和系统

(57)摘要

本发明公开了一种基于时间窗口和语义的变体词规范化的方法和系统。本方法为:1)根据给定变体词的出现时间,选取社交网络中该出现时间之前的设定时间段内的语料,作为候选语料库集合D1;2)将该候选语料库集合D1中和该变体词所在语料语义相似的语料加入到候选语料库集合D2;3)从该集合D2中提取出候选词,得到一候选词集合;4)根据每一候选词和变体词的字面相似度以及上下文特征相似度计算每对候选词和变体词的得分,根据计算结果确定该变体词对应的候选词,将确定出的候选词作为该变体词的规范词。本系统包括采集模块、过滤模块、获取模块和规范词获取模块。本发明使得社交网络的文本变的更加规范,便于舆情分析、热点时间追踪。



CN 107315734 B

1.一种基于时间窗口和语义的变体词规范化的方法,其步骤为:

1)根据给定变体词的出现时间,选取社交网络中该出现时间之前的设定时间段内的语料,作为候选语料库集合D1;

2)将该候选语料库集合D1中和该变体词所在语料语义相似的语料加入到候选语料库集合D2;

3)从该候选语料库集合D2中提取出候选词,得到一候选词集合;

4)根据每一候选词和变体词的字面相似度以及上下文特征相似度计算每对候选词和变体词的得分,根据计算结果确定该变体词对应的候选词,将确定出的候选词作为该变体词的规范词。

2.如权利要求1所述的方法,其特征在于,从该候选语料库集合D2中提取出候选词的方法为:利用分词工具、词性标注方法、名词词组检测方法、命名实体标注方法和事件提取方法分别从该候选语料库集合D2中提取出候选词;然后将提取出的候选词取并集得到所述候选词集合。

3.如权利要求1所述的方法,其特征在于,基于LDA文本相似性计算方法或基于Doc2Vec的文本相似性计算方法计算所述语义相似的语料。

4.如权利要求1或2或3所述的方法,其特征在于,所述步骤4)中,采用无监督的机器学习方法,利用神经网络在大规模语料库中自主学习词语的上下文语义表示计算每对候选词和变体词的得分。

5.如权利要求4所述的方法,其特征在于,分别提取变体词和候选词的词向量输入无监督的机器学习方法进行训练,其中在训练词向量的时候,将一个词语拆分成两部分:词语本身和组成该词语的汉字。

6.如权利要求5所述的方法,其特征在于,采用CWE模型训练词向量,加入组成词语的字的信息构成该词语的语义表示。

7.如权利要求1或2或3所述的方法,其特征在于,所述步骤4)中,采用有监督的机器学习方法,计算每对候选词和变体词的得分。

8.如权利要求6所述的方法,其特征在于,分别提取变体词和候选词的表面特征、语义特征、社交特征,输入有监督的机器学习方法进行训练,得到每一候选词和变体词的得分。

9.一种基于时间窗口和语义的变体词规范化的系统,其特征在于,包括采集模块、过滤模块、获取模块和规范词获取模块;其中,

采集模块,用于从社交网络中采集语料信息;

过滤模块,用于根据给定变体词的出现时间,从采集的语料信息中选取该出现时间之前的设定时间段内的语料,作为候选语料库集合D1;以及从该候选语料库集合D1中选取与该变体词所在语料语义相似的语料,加入到候选语料库集合D2;

获取模块,用于从该候选语料库集合D2中提取出候选词,得到一候选词集合;

规范词获取模块,用于根据每一候选词和变体词的字面相似度以及上下文特征相似度计算每对候选词和变体词的得分,根据计算结果确定该变体词对应的候选词,将确定出的候选词作为该变体词的规范词。

10.如权利要求9所述的系统,其特征在于,所述规范词获取模块包括分词模块和相似度计算模块;其中,分词模块,用于对输入的语料进行分词处理,将之作为词向量训练的输

入;相似度计算模块,用于根据变体词和候选目标词的词向量计算每对候选词和变体词的相似度得分。

一种基于时间窗口和语义的变体词规范化的方法和系统

技术领域

[0001] 本发明涉及社交网络数据分析领域，是一种基于时间窗口和语义的变体词规范化的方法，以实现更有针对性、准确性的社交网络中变体词的规范化的方法和系统。

背景技术

[0002] 随着社交网络的飞速发展，每天有亿级的信息发布在社交网络平台中，带来了信息的爆炸式增长。信息的形式多种多样，包括文本、图片、音频、视频等。其中社交网络中的文本具有随意性、非正规性等特点。变体词就是网络语言作为一种不规范语言的显著特色，人们往往处于避免审查、表达情感、讽刺、娱乐等需求将相对严肃、规范、敏感的词用相对不规范、不敏感的词来代替，用来代替原来词的新词就叫做变体词(Morph)。变体词和其对应的原来的词(目标实体词)会分别在非规范文本和规范文本中共存，甚至变体词会渗透到规范文本中。变体词使行文更加生动活泼，相关事件、消息也传播得更加广泛。但是因为变体词通常是某种隐喻，已不再是其表面字词的意义了，从而使网络上文体与正式文本(如新闻)具有巨大的差异。由此如何识别出这些变体词所对应的目标实体词，即规范化，对于下游的自然语言处理技术具有重要的意义。进一步，研究变体词的规范化对于舆情分析、热点事件追踪等都有重要意义。

[0003] 变体词的规范化

[0004] 变体词规范化指变体词的解析，即找到变体词所对应的目标实体词。

[0005] 形式化描述如下：已知文档集合$D=\{d_1,d_2,\cdots,d_{|D|}\}$，文档集合D中唯一词集合为$T=\{t_1,t_2,\cdots,t_{|T|}\}$，定义候选的变体词$m'_j$是T中的一个唯一词$t_j$。则定义一个候选变体词的提及(morph mention)$\mathbf{m}_j^p$为m_j在一个特定文档d_j里的第p次出现。

[0006] 这里需要注意如果一个提及的表面形式是与m_i相同的，但是如果其指向其原来的含义，那么就不认为是变体词的提及。

[0007] 例如：如果词语“小马哥”通过上下文获知其指向的是香港电影《英雄本色》里的一角色，则就不是一个变体词的提及；但是如果其指向的是一公司总裁马某某，则认为是一个变体词的提及。

[0008] 因此变体词规范化任务是针对每一个变体词提及$\mathbf{m}_j^p$解析出其目标实体词e_1。针对上例，则需要解析出变体词“小马哥”的目标实体词为“马某某”。

[0009] 最终目标是获得变体词对应的目标实体词。

[0010] 变体词的特点

[0011] 1)变体词可以看作是一种利用自然语言处理技术来传播秘密消息的一种手段。绝大多数的变体词可以看作是基于深层语义和背景知识的编码，而不是简单的字典式的替换，因此变体词更接近于行话、黑话、术语等。

[0012] 2)变体词与目标实体词之间的映射关系不是全射关系，也即不是标准的对应关系，多个变体词可以对应一个目标实体词，一个目标实体词也可以对应多个变体词。

[0013] 3)社交网络平台对变体词的产生和发展起着至关重要的推动作用。社交网络作为一种自媒体,更是激发了广大群众的创造欲望、集成了广大群众的集体智慧。众多流行的变体词都是互联网上普通用户自发创造获得广泛传播的。

[0014] 4)变体词随着时间的推移迅速演化。根据新的新闻热点、特殊事件,会不断地产生新的变体词,这是变体词的一大特点,也反应到了变体词的生成规律上。有些变体词会逐步消亡,而有些甚至进入了规范的文本中获得更广泛的认可。

[0015] 变体词规范化的研究现状

[0016] 明确的变体词概念出现在Huang的论文中(参考Huang,Hongzhao,et al."Resolving Entity Morphs in Censored Data."ACL(1).2013),但是变体词相关的概念和技术一直在不良文本过滤、社交媒体文本规范化等领域有所体现。下面主要从规范化技术角度详细阐述变体词规范化的发现现状。

[0017] 1)基于规则的规范化方法

[0018] 最早与变体词相关的研究主要有网络不良文本的过滤技术,前期主要使用精确匹配、分类器等方法。但是发现变体词的出现会严重影响到过滤的准确度。因此逐步引入了对变体词的处理,如Yoon将某特殊字符转化成形状相似的字母,然后再进行检测(参考Yoon T,Park S Y,Cho H G.A smart filtering system for newly coined profanities by using approximate string alignment[C]//Computer and Information Technology(CIT),2010IEEE 10th International Conference.IEEE,2010,643-650.)。例如:将特殊字符“!”转换成字母“i”,遇到“sh!t”词后,将这个词转换成“shit”来处理。

[0019] 陈儒等人提出了面向中文特定关键词变体的过滤技术(参考:陈儒,张宇,刘挺.面向中文特定信息变异的过滤技术研究[J].高技术通讯,2005,15(9):7-12.),针对中文网络的5种变体方法提出了变异规则:1)对关键词进行同音字替换或拼音替换;2)对关键词进行拆分;3)在关键词中插入无意义的非汉字符合;4)关键词的组合;5)上述4种方法的组合。

[0020] Sood在对不良文本及其变体信息进行检测的时候,利用了"众包"的思想,使用"众包"来对文本进行标记,采用机器学习的技术来对不良文本信息过滤,通过采用bigram、词干等作为特征值来对文本信息做分类分析,以检测不良信息(参考Sood S O,Antin J,Churchill E F.Using Crowdsourcing to Improve Profanity Detection[C]//AAAI Spring Symposium Series.2012:69-74.)。

[0021] Xia和Wong考虑中文聊天室等环境下动态非规范语言的规范化问题,以标准汉语语料库为基础建立了汉字的语音映射模型,对信源/信道模型进行扩展(eXtended Source Channel Model,XSCM),然后基于汉字语音之间的相似度进行替换,但需要手工确定相似度的权重(参考Yunqing Xia,Kam-Fai Wong,and Wenjie Li.2006.A phonetic-based approach to chinese chat text nor-malization.In Proceedings of COLING-ACL2006,pages 993-1000.;K.F.Wong and Y.Xia.2008.Normalization of Chinese Chat Language.Language Resources and Evaluation,pages 219-242)。

[0022] 2)基于统计和规则的规范化方法

[0023] Wang从非规范词的规范化角度(参考Aobo Wang,Min-Yen Kan,Daniel Andrade,Takashi Onishi,and Kai Ishikawa.2013.Chinese informal word normalization:an experimental study.In Proceedings of International Joint Conference on Natu-

ral Language Processing(IJCNLP2013)),首先提取统计特征和基于规则的特征,通过分类实现中文非规范词的规范化。通过语音建立了汉字-汉字之间的映射关系,通过缩写建立了汉字-词的映射关系,通过意译建立了字-词,词-词的映射关系。

[0024] Choudhury针对SMS文本,提出了一种基于隐马尔可夫模型的文本规范化方法(参考M Choudhury,R Saraf,V Jain,et.al.Investigation and modeling of the structure of texting language[J].Internation Journal of Document Analysis and Recognition,2007,10:157-174.),该方法是一对一的规范化方法,通过构造常用缩写和非规范用法的词典,可以部分解决一对多的问题。Cook通过引入无监督的噪声信道模型对Choudhury提出的模型进行了扩展,模型对常用缩写形式和各种不同的拼写错误类型进行了概率建模。

[0025] 还有通过构建规范化词典用于文本规范化任务。例如,Han首先训练分类器用于识别非规范词候选,然后使用词音相似度得到规范化候选,最后利用字面相似度和上下文特征找出最佳的规范化候选(参考B Han,P Cook,T Baldwin.Automatically constructing a normalization dictionary for microblogs[C]//Proceedings of the 2012joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning,2012:421-432)。Han又提出基于上下文相似性和字面相似性构建规范化词典进行推特文本的规范化,使用词袋模型表示上下文分布,然后两两之间计算上下文分布相似度(参考B Han,T Baldwin.Lexical Normalization of Short Text Messages:Makn Sens a#Twitter[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:Human Language Technologies,2011,1:368-378.)。

[0026] Li提出了一个基于规则和数据驱动的对数线性模型从互联网语料中对规范与非规范中文短语的关系进行挖掘和建模(参考Zhifei Li and David Yarowsky.2008.Mining and modeling relations between formal and informal chinese phrases from web corpora.In Proceedings of Conference on Empirical Methods in Natural Language Processing(EMNLP2008),pages 1031-1040.)。他们主要针对同音异形异义词、缩略语、首字母缩写词、音译等。

[0027] 他们注意到一个现象,在非规范短语附近有时可以发现对应的规范短语,他们分为直接定义和间接定义。1)直接定义,如:“GF就是女朋友的意思”;2)间接的定义,如在聊天室中:A:“对不起,我先下线了。”B:“拜拜”。A:“88”。

[0028] Li提出的非规范词规范化的bootstrapping算法步骤如下:给定一个非规范词,利用搜索引擎搜索含有此非规范词的非规范文本(如博客、社交网络上的文本)。产生候选规范化词集,在含有非规范词的一定长度窗口内提取n-gram。基于正则化条件对数似然对候选集进行打分排序。规则驱动提取的特征包括:两者拼音之间的Levenshtein距离;两者拼音之间不同字符数;非规范词是否是规范词的拼音缩写;非规范词是否是规范词的汉字缩写。数据驱动提取的特征包括:两者共现频率;两者共现是否符合某一模式;搜索引擎搜索同时含有两者的网页数目。

[0029] Li主要是通过搜索引擎来发现非规范词－规范词对。此方法对于定义良好和高频的词效果比较好,而且严重依赖于搜索引擎返回的结果。

[0030] 3)基于语义表示的识别和规范化方法

[0031] 现有从语义角度入手变体词的识别与规范化的主要是基于分布假设和语义组合假设。1954年,Harris提出分布假说(distributional hypothesis),即“上下文相似的词,其语义也相似”(参考Zellig S Harris.Distributional structure.Word,1954.)。德国数学家弗雷格(GottlobFrege)在1892年提出:一段话的语义由其各组成部分的语义以及它们之间的组合方法所确定(参考Gottlob Frege.Über sinn und bedeutung.Funktion-Begriff-Bedeutung,1892)。为了得到句子和文档级别的语义表示,一般可以采用语义组合的方式。

[0032] 基于分布假设,给定一个变体词,如果另一个词与之上下文相似,则可以初步推断这个词很可能就是变体词的目标实体词。而上下文语义的获取则可以基于语义组合的方式。

[0033] Huang等人研究在给定变体词的情况下,挖掘跨数据源可比较语料的时空限制,找到对应的目标实体词。其基本框架如图1所示。给定一个变体词查询,获取多数据源的数据,进行对比分析,基于语义标注找到候选目标词集,然后根据:表面特征(surface features)、语义特征(semantic features)、社交特征(social features)等对候选目标词集进行打分,最终获得目标实体词。

[0034] 其中表面特征包括:字符串编辑距离,正则化字符串编辑距离,最长公共子串。语义特征指构建了信息网络(Information Network)。其中节点代表变体词(M);实体(E),包括候选的目标实体词;事件(EV);非实体名词(NP);边代表两者共现,边权重为其在所有推文中的共现频率。基于meta-path进行语义相似性比较。社会特征:对用户的行为建模,用社交行为的相关性来辅助语义相似性测量。因为观察发现变体词和对应目标实体词的用户往往具有相似兴趣和观点意见。

[0035] Huang的主要贡献在于:根据一定时间窗口内变体词和目标实体词是相关;根据社交媒体的动态特性提取变体词和目标实体词的时空分布;对多个数据源数据进行对比分析;对用户的行为建模,用社交行为的相关性来辅助语义相似性测量。其不足主要在于:此方法是在给定变体词的情况下,并且使用了大量的标注数据。此方法做到了语料级别,但是不是提及级别。此方法严重依赖于变体词的多个实例的聚合上下文和时空信息。

[0036] Zhang等人采用无监督的方法(参考Zhang,Boliang,et al.″Context-aware Entity Morph Decoding.″Proc.Annual Meeting of the Association for Computational Linguistics(ACL2015).2015),基于深度学习实现对变体词及其目标实体词的映射关系的发现。文章把变体词的识别和规范化分成如下步骤:

[0037] 1.先初筛出单个变体词提及(metion)的候选集。

[0038] a)潜在变体词的发现:基于4类特征(基本特征、特征字典、语音、语言模型)的分类问题来发现潜在的变体词。

[0039] b)潜在变体词的验证:基于2个假设:1)如果2个提及是共指的,则2者要么都是变体词的提及,要么都不是;2)高度相关的提及要么都是变体词的提及,要么就都不是。基于上述2个假设提出了一个半监督的学习方法利用小规模已标注数据集对大规模未标注数据集的变体词提及进行验证。

[0040] 2.变体词的规范化(发现其目标实体词)。主要通过深度学习技术来捕捉比较一个

变体词和它的候选目标实体词语义表示。

[0041] a)候选目标实体词的识别:主要是基于Huang的时空分布假设:变体词及其目标实体词应该有相似的时空分布。文章采用的标准:在变体词出现的7天之内应该可以找到变体词的目标实体词;

[0042] b)候选目标实体词的打分排序:基于深度学习技术习得变体词及其目标实体词的语义表示,文章提出了2种算法,并且比较两者的效果。

[0043] 首先是基于多数据源的监督学习,如图2所示。但是效果不好,因为建立词向量的时候主要是采用wikipedia的数据进行训练,但是wikipedia和含有变体词的社交媒体文本有很大的不同。

[0044] 第2个模型采用的如图3所示的连续词袋模型。利用连续词袋模型训练推文,获得变体词和实体词的语义表示,比较两者的相似度。

[0045] 变体词规范化的评测标准

[0046] 一个社区发现算法的效果如何,需要在具体的网络上进行测试。当前,社区发现算法的测试网络主要有人工构造的网络和真实的网络。

[0047] 人工测试集的典型代表有Girvan与Newman提出的标准测试集和Lancichinetti等人提出的更为严格的测试集。标准测试集是对人工l-分割模型的一种实现,它规定网络中的结点组数l=4,每个组的节点数为32,则顶点总数为128,同时规定节点的平均度$\langle k\rangle=16$。通过调整z_{in}和z_{out}的值,可以调整网络社区结构的显著程度。其中z_{in}表示结点连接同一社区内点的平均边数,z_{out}表示连接不同社区的点的边数。显然有$z_{in}+z_{out}=\langle k\rangle$。标准测试集里隐含着这样的假设:网络里节点和社区是同质的。这与现实网络的特性是不符合的。因此,Lancichinetti等人提出了新的测试集,用于解决节点度和社团规模的异质性问题。在该测试集中,节点度和社团规模都服从幂律分布,混淆参数μ用于控制社区结构的显著程度。

[0048] 真实的网络测试集是那些根据人们的观察和知识而得到社区结构划分的现实网络。当前,用得比较广泛的有Zachary空手道俱乐部网络,Lusseau等人提供的宽吻海豚的社会关系网络以及根据美国大学橄榄球队之间的比赛关系构建的网络。真实的网络测试集存在的一个重要问题是:已知的社区结构是根据人们的观察和经验获得,而社区发现算法一般从拓扑结构入手,无法预知两者之间有多大的关联。

[0049] 有了测试集之后,还需要有相应的方法来度量社区发现算法获得的社区结构和网络已知社区结构之间的相似程度。最简单的方法是以节点正确划分的比例来衡量,而当前使用的最广泛的划分相似度度量方法是归一化互信息、信息变差等。

[0050] 对于扩展了主题的社区发现,可以采用社区的平均主题相似性作为衡量一个社区内部成员的紧密程度的标准。一个社区的平均主题相似值越大,说明社区中成员的共同兴趣越一致,该社区划分越合理。

发明内容

[0051] 本发明的目的是提供一种基于时间窗口和语义的变体词规范化的方法和系统。本发明基于时间窗口和语义来对社交网络上的变体词进行规范化操作,使得社交网络的文本变的更加规范,为接下来对于社交网络的舆情分析、热点时间追踪等分析操作做准备。

[0052] 当前变体词的规范化需要迫切解决的问题主要有：

[0053] 1)找到高质量的候选目标词集合。

[0054] 2)提高变体词规范化的准确度。

[0055] 以上2点其本质问题就是要加深对变体词的理解。这里以往都是强调变体词与目标实体词的相似性，实际上需要从相似性和差异性两个角度进行思考。即首先对变体词的生成规律的理解，需要从相似性和差异性两个方面来对变体词及其目标实体词进行对比分析：

[0056] 1)变体词和目标实体词的相同之处：只有识别出了变体词和目标实体词的相同之处，才可能找到变体词所对应的目标实体词。

[0057] a)首先变体词的语义和目标实体词的语义应该是一致的，这也是变体词能够产生的原因。变体词和目标实体词的语义相似性主要体现在文档级别、句子级别和字的级别。而词级别的应该主要是体现变体词和目标实体词之间的差异性。

[0058] b)变体词的字面组合(surface name)与目标实体词应该也具有一定相似性，其字面组合的意义也可以用来辅助对变体词的目标实体词的发现。既然出现了surface name，也就是说既然使用了变体词指向目标实体词，则说明surface name与目标实体词之间有一定相同\相似的特征，因此需要基于语义表示来研究变体词的surface name与目标实体词之间的共同特征以及在图上、词向量空间上如何展示的。

[0059] 2)变体词和目标实体词的不同之处：只有识别出变体词和目标实体词的不同之处，才可能在语料中找到变体词。

[0060] a)两者之间的差异性应该主要体现在语义表示上的词的级别。这种差异性主要体现在语义上，而上层文档、句子的语义相似性可以提供发现这种差异性的线索，而知识图谱、社交媒体的关系也可以提供辅助信息，加快这种搜索的过程。

[0061] 以往只强调了变体词和目标实体词的相似性，实际上应该是相似性和差异性的权衡，即“存大同，求小异”，这样才能体现变体词和目标实体词之间的微妙的关系。

[0062] 因此在充分研究两者相似性和差异性基础上，总结出变体词的特性和使用变体词的规律，然后才能提到识别的方法。因此需要对变体词及其目标实体词的特征进行分析，分析语义表示中各节点之间的相似性和差异性。在获得变体词和目标实体词之间的相似性和差异性之后，进一步依托句子、文档级的语义表示，研究变体词和目标实体词的使用环境的相似性和差异性。

[0063] 为了能够准确地解析出变体词的目标实体词，首先需要对变体词及其目标实体词准确地给出语义上的描述，能体现两者的深层语义联系(这样才能解析出其目标实体词)。因此首先要研究能够体现这种“求大同，存小异”的合适的语义描述，可以通过神经网络分别构建字/词级别、句子级别和文档级别的语义表示来体现这种“大同，小异”。

[0064] 因此需要在表达能力强的语义表示基础上，充分利用多源多维度的信息，充分利用社交媒体的关系信息，利用相关知识图谱的先验知识，以提高识别的准确度。

[0065] 基于此，为了实现对社交网络中变体词的有效规范化，本发明提出了一种基于时间窗口和语义的变体词规范化方法和系统。

[0066] 本发明主要包括两个方面：(1)提出了基于时空分布的候选词提取模型和基于语义相似度的候选词提取模型；(2)提出了一种基于时间窗口和语义的变体词规范化方法和

系统。

[0067] 该发明包括以下内容：

[0068] 1)社交网络中候选词的发现。在大规模语料库中提取出与给定变体词所匹配的可能的实体词。先是给语料分块。本发明借助变体词的时间分布以及变体词所在句子的语义，在大规模语料库中挑选出合适的语料，然后基于一些分词和词性标注等工具，提取合适的词语加入候选目标实体词集合中。

[0069] 2)社交网络中候选词的排序。通过计算候选词和变体词字面相似度和上下文语义特征相似度进行排序。对于有监督的机器学习方法，挑选合适的特征，对候选词在当前上下文背景下，计算变体词-候选词得分或相对排序。对于无监督的机器学习方法，利用神经网络在大规模语料库中自主学习词语的上下文语义表示计算得分或相对排序。

[0070] 3)基于时间窗口和语义的变体词规范化方法和系统。在第一阶段，拟采用基于时空分布并结合文档语义相似度，聚合语料，弥补候选词结合过大或过小的缺点；在第二阶段，拟采用机器学习的算法，挖掘词项上下文中的可用特征，结合词项或字的表面特征，构建候选词排序模型。采取神经网络语言模型，在大规模语料上训练词表示，然后计算相似度排序。

[0071] 与现有技术相比，本发明的积极效果为：

[0072] 1、充分利用变体词所在社交网络文本的时间和上下文语义，极大缩小了候选目标词的规模。

[0073] 2、分析了变体词和目标词之间的异同之处，结合变体词和候选目标词的上下文以及组成词语的字的信息，通过字词联合训练出词语的语义表示，对候选目标词进行排序。

附图说明

[0074] 图1为变体词的识别与解析流程图；

[0075] 图2为多数据源的监督学习；

[0076] 图3为连续词袋模型；

[0077] 图4为候选词集合提取模块架构图；

[0078] 图5为候选词排序模块架构图；

[0079] 图6为变体词规范化架构图；

[0080] 图7为候选目标词获取框架图；

[0081] 图8为候选目标词排序框架图。

具体实施方式

[0082] 本发明的变体词规范化架构如图6所示，具体步骤如下：

[0083] (一)社交网络候选词发现。具体可以分为两个步骤：

[0084] 候选词提取模块的模块架构如图4所示，该实验方案能够弥补前文分析的候选词集合过大或过小的缺点。

[0085] 实验步骤如下：

[0086] 1)语料库的划分

[0087] a)按时间划分，基于时空分布假设，在变体词出现的时间前7天内，根据语料库每

条微博的时间，划分出一个候选语料库集合D1。

[0088] b)按语义划分，基于语义相似假设，将候选语料库集合D1中和变体词出现的微博语义较相似的微博加入到候选语料库集合D2。计算相似度的方法是基于LDA(Latent Dirichlet Allocation)文本相似性计算方法和基于Doc2Vec的文本相似性计算方法。

[0089] 2)候选词的识别提取

[0090] 在候选语料库集合D2中，运用多种工具提取出候选词，如：分词工具、词性标注、名词词组检测、命名实体标注、事件提取等。然后综合上述工具得出的结果，本发明取结果集的并集作为本发明最后的候选词集合。

[0091] (二)社交网络候选词的排序

[0092] 候选词排序即对上述提取出的候选词集合中的所有词进行打分并排序，如图5所示：

[0093] 1)有监督的方式

[0094] 对候选词是否是变体词对应的目标词建立分类模型。现有的方法是根据如下4类特征：表面特征(surface features)、语义特征(semantic features)、社交特征(social features)等对候选目标词集进行打分，最终获得目标实体词。

[0095] 2)无监督的方式

[0096] 现有的方法是在大规模语料库上用word2vector模型学习词语的语义表示，然后计算变体词和候选词的语义相似度，从而根据相似度进行排序。一方面现有的方法没有考虑到词语中的字的表示，但是本发明考虑到大多数变体词和目标词在字的层面上会有共同点，所以在大规模语料库中训练词和字联合表示可能会有所提高。另一方面可以利用其它神经网络模型，比如记忆网络，在候选的语料库中自己学习到目标词。

[0097] 本发明采用无监督的方式对候选目标词进行排序，考虑到大多数变体词和目标词在字的层面上具有相同字的特点，所以在训练词向量的时候，将一个词语拆分成两部分：词语本身和组成这个词语的汉字。本发明采用CWE模型训练词向量，加入组成词语的字的信息构成这个词语的语义表示。

[0098] 在变体词候选目标词排序任务中，CWE模型有以下亮点优势：

[0099] (1)CWE模型输出的是融合了字向量信息的词向量。一些变体词会基于目标词中的某些字而形成，如变体词“吃省”，它的目标词是“广东省”，此时变体词和目标词有一个共同的“省”字，词向量表示中加入字向量后，在变体词的候选目标词排序中，CWE模型因为能更有效地计算变体词和目标词的相似度而使得排序结果更加准确。

[0100] (2)CWE模型单独输出了字向量。我们可以通过组合字的向量来合成未登录词的词向量，然后可以计算新的变体词和候选目标词之间的相似度，而不用重新训练词向量模型，减少了重新训练词向量所带来的时间开销成本。

[0101] (三)基于时间窗口和语义的社交网络变体词规范化方法和系统

[0102] 在时间属性和语义属性上，获取到变体词候选集，且对候选词打分排序的基础上，实现基于时间窗口和语义的社交网络变体词规范化方法和系统。

[0103] a)社交网络中变体词规范化方法：根据当前变体词规范化方法的相关研究现状，本方法采用先根据时间和语义属性来划分候选集的方式，来发现并提取候选词、排序候选词来实现变体词的规范化。

[0104] b)社交网络中变体词规范化系统:系统由目标候选词发现模块,目标候选词排序模块构成。

[0105] 由此实现了基于时间窗口和语义的变体词规范化方法和系统。

[0106] 社交网络变体词规范化方法和系统有两部分组成:1)候选目标词获取框架;2)候选目标词排序框架。

[0107] 候选目标词获取框架由3部分组成:采集模块,过滤模块和获取模块,如图7所示。各模块主要功能如下:

[0108] 采集模块:主要负责获取社交网络文本数据,如新浪微博消息数据,Twitter中文消息数据和Web新闻等。

[0109] 过滤模块:这是获取框架中的重点部分,分为根据时间窗口过滤和根据话题相似过滤。

[0110] 获取模块:主要负责对上述过滤后的语料进行分词和词性标注等,提取出需要的候选词。

[0111] 候选目标词排序框架由3部分组成:分词模块,词向量训练模块和相似度计算模块。

[0112] 如图8所示,各模块主要功能如下:

[0113] 分词模块:主要负责给输入的语料(如新浪微博等)进行分词处理,将之作为词向量训练的输入。

[0114] 词向量训练模块:这是排序框架中的重点部分,其中本文采用了两种字词联合训练方法:融合字信息的词向量法(CWE)模型和融合偏旁信息的词向量法(MGE)模型。

[0115] 相似度计算模块:主要负责对变体词和候选目标词的词向量进行余弦相似度计算,并对候选目标词进行排序操作。

[0116] 由此实现了基于时间窗口和语义的变体词规范化方法和系统。

[0117] 以上实施例仅用以说明本发明的技术方案而非对其进行限制,本领域的普通技术人员可以对本发明的技术方案进行修改或者同等替换,而不脱离本发明的精神和范围,本发明的保护范围应以权利要求所述为准。

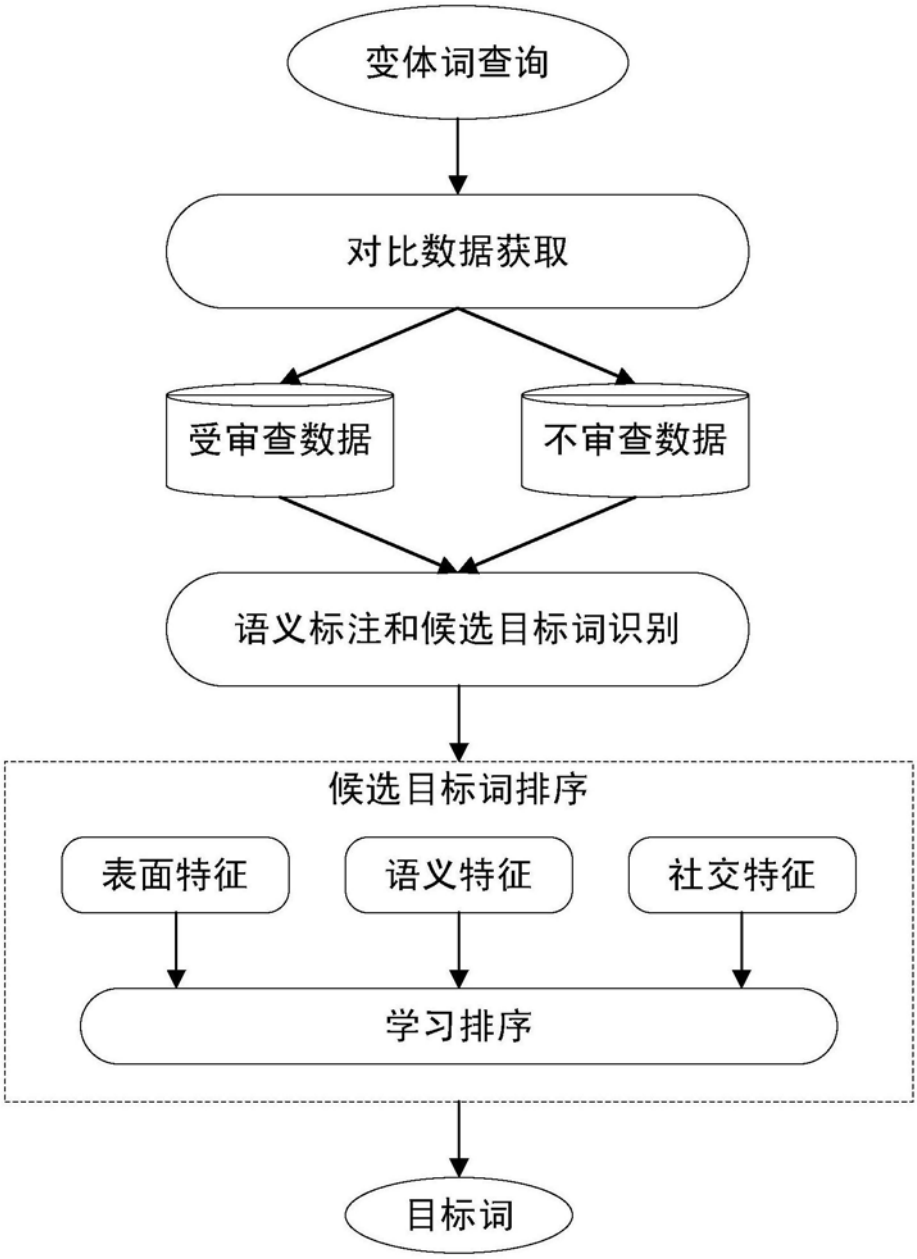

变体词查询
对比数据获取
受审查数据
不审查数据
语义标注和候选目标词识别
候选目标词排序
表面特征
语义特征
社交特征
学习排序
目标词

图1

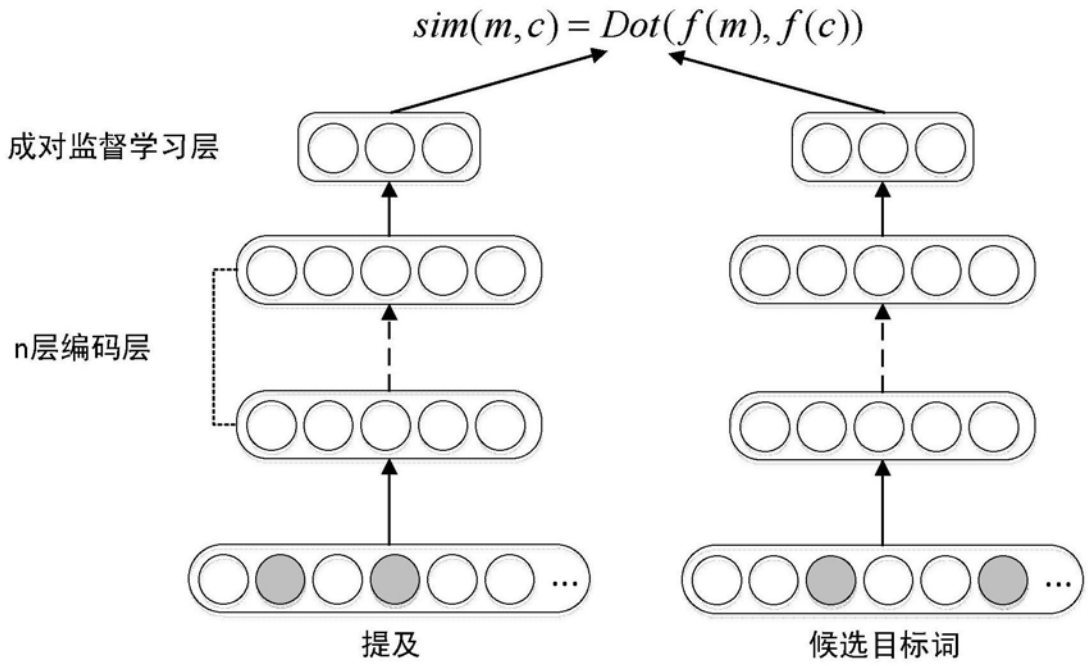


图2

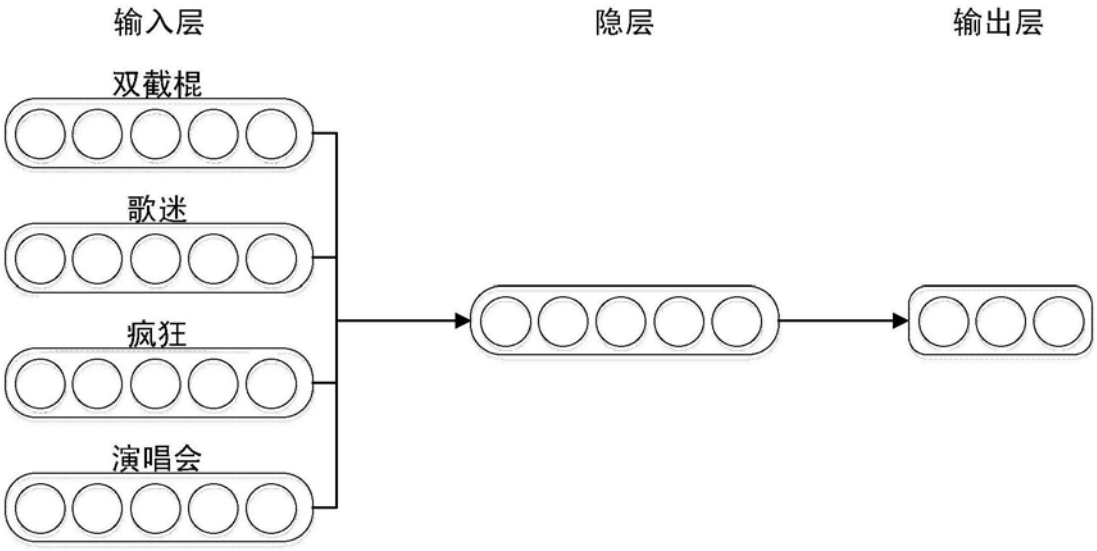


图3

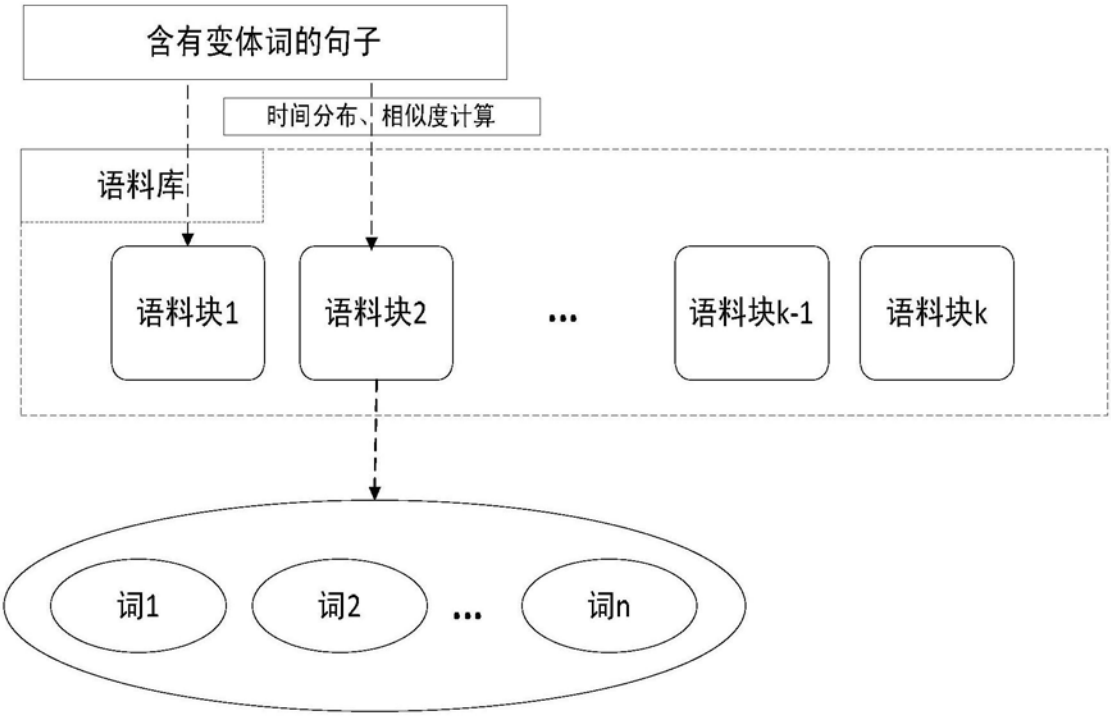

含有变体词的句子
时间分布、相似度计算
语料库
语料块1
语料块2
...
语料块k-1
语料块k
词1
词2
...
词n

图4

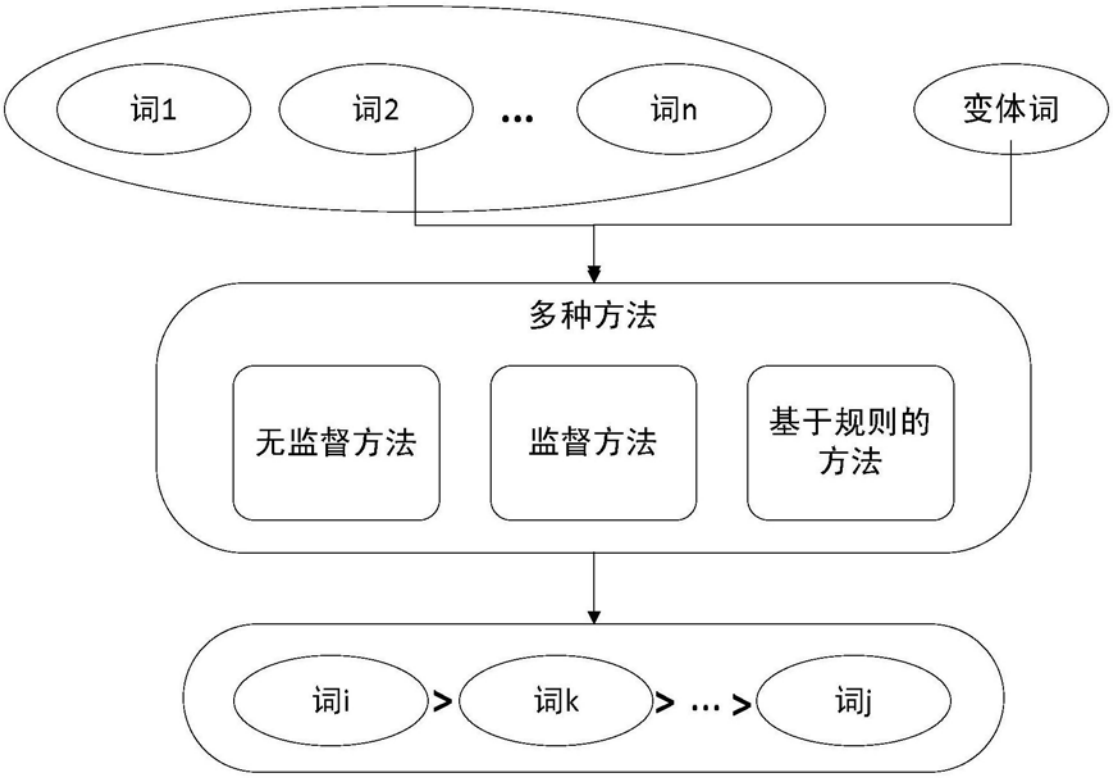

词1
词2
...
词n
变体词
多种方法
无监督方法
监督方法
基于规则的方法
词i
>
词k
> ... >
词j

图5

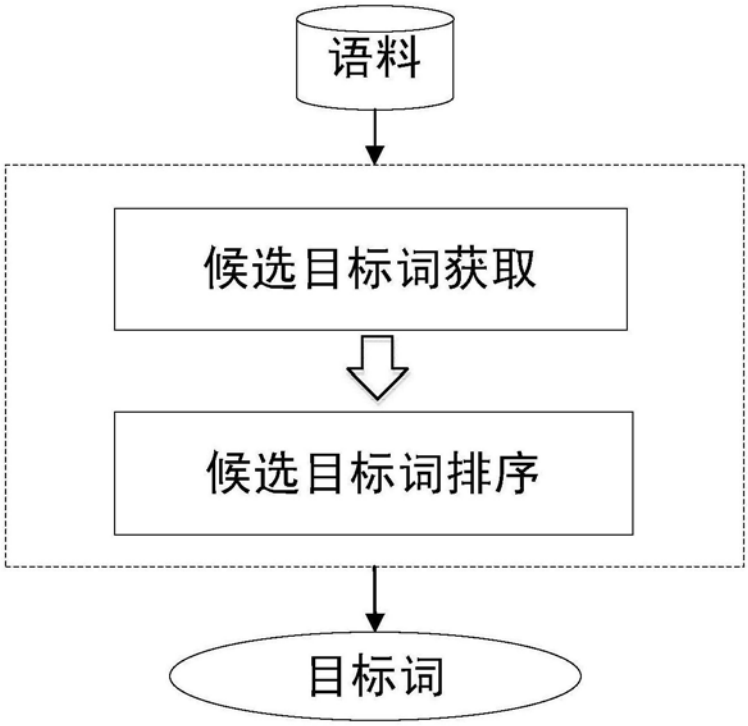

语料
候选目标词获取
候选目标词排序
目标词

图6

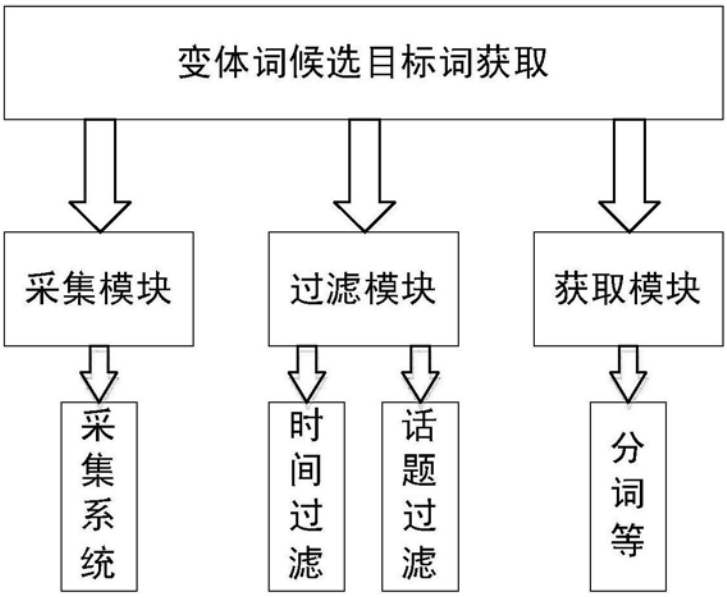

变体词候选目标词获取
采集模块
过滤模块
获取模块
采集系统
时间过滤
话题过滤
分词等

图7

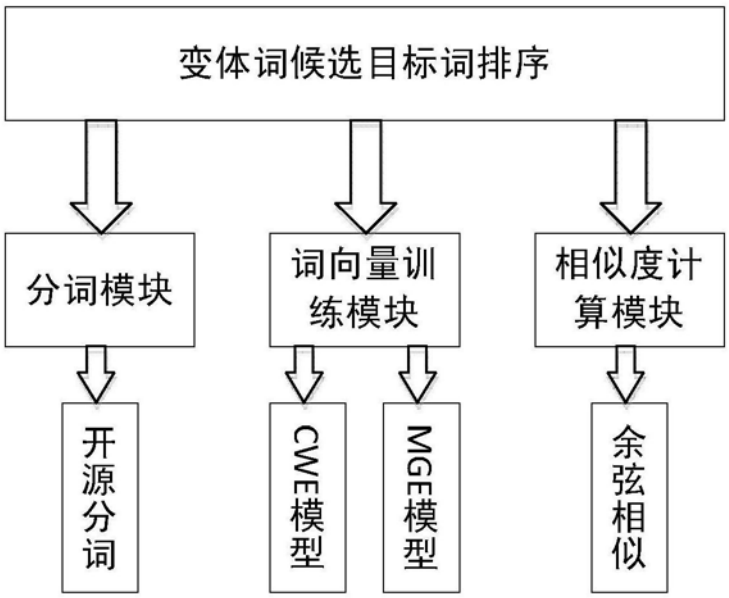


图8