中文变体词的识别与规范化综述

沙 瀛, 梁 棋, 王 斌

中国科学院信息工程研究所第二研究室 北京 中国 100093

摘要 网络书写具有随意性、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, 人们往往出于避免审查、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(如新闻等)具有巨大的差异。由此如何识别出这些变体词及其所对应的目标实体词对于下游的自然语言处理技术具有重要的意义。本文首先介绍了变体词的定义和特征, 变体词的生成规律, 总结了当前变体词的识别和规范化的主要技术进展和成果, 最后是此领域发展方向的展望。

关键词 社交网络; 变体词识别; 变体词规范化; 深度学习; 神经网络; 表示学习
中图法分类号 TP309.2 DOI 号 10.19363/j.cnki.cn10-1380/tn.2016.03.006

Chinese Morphs Identification and Normalization

SHA Ying, LIANG Qi, WANG Bin

China Institute of information engineering, CAS, Beijing 100093, China

Abstract Internet language is a casual informal language. Entity morph is an important feature of Internet Language. In some situation, Internet users are keen on creating kinds of morphs, special kinds of fake alternative names to achieve some goals, express strong sentiment or humor, and avoid censorship. Entity morphs and their target entities respectively appear on informal and formal text. And in some situation, entity morphs even appear on formal text. Although using entity morphs has some advantages, but morphs are big barriers for natural language processing (NLP). So it is very important to research on morph identification and normalization. First, we will introduce the definition of morphs and the features of morphs; second, we will show the rules of generating morphs; third, the current progress of morph identification and normalization will be demonstrated. Finally, it is the prospect of this field.

Key words social network; morph identification; morph normalization; deep learning; neural network; representation learning

1 引言

作为网络信息安全的重要组成部分, 网络内容安全一直受到工业界和学术界的广泛关注。特别是近年来, 针对威胁网络内容安全的行为研究越来越多, 已经成为内容安全新的研究热点。社交网络平台已经成为产生网络安全事件的重要源头, 大量的不良敏感信息通过社交网络平台进行传播并进一步引起特定事件的爆发, 为了躲避相关的审查和过滤, 变形词成为利用自然语言处理技术来传播秘密消息的一种重要而有效的手段(变体词就是将关键的不良敏感信息用另外不敏感的词来代替, 但是不影响理解)。因此研究变体词的识别和规范化对于网络内容安全是十分重要的。

目前自然语言处理技术较多关注正式的文本。但是随着互联网特别是社交网络的发展, 社交网络平台已经成为人们获得信息、发表观点看法、传播意见舆论的重要平台。根据社交网络的特点, 人们倾向于在社交网络上使用新式的、自创的语言、术语和习惯。这种网络语言及其使用习惯都对自然语言处理技术带来了冲击。

采用变体词是网络语言作为一种不规范语言的显著特色, 人们往往出于躲避审查、情感、娱乐等需求将相对严肃、规范的词用相对不规范、不敏感的

通讯作者: 沙瀛, 博士, 副研究员, Email: shaying@iie.ac.cn。
本课题得到国家科技支撑计划(编号: 2012BAH46B03), 中国科学院战略先导专项(编号: XDA06030200)资助。
收稿日期: 2016-04-01; 修改日期: 2016-06-16; 定稿日期: 2016-07-06

词来代替。用来代替原来词的新词就叫做**变体词(Morph)**。变体词和其对应的目标实体词(原来的词)会分别在非规范文本和规范文本中共存, 甚至变体词渗透到规范文本中。

变体词可以看作是一种特殊的有意隐藏背后的真实实体的假名[1,2]。变体词通常被认为是"**社会媒体用户为了某种目的需要隐藏真实的实体或事件, 代替这些实体或事件所采用的化名或假名等**"[3]。采用变体词的目的包括: 采用委婉的说法以避免当事人的过激反应; 表达对相关人或事的强烈的正面或负面情感; 表达讽刺或幽默; 使行文更加简练; 达到娱乐的效果让实体或事件描述得更加鲜明有趣, 让文本传播更广。也有恶意用户发布不良敏感信息的时候,为了避免被过滤会对不良敏感词汇进行变形处理。Zhang[4]分析了随机选取的 548 个变体词, 其中不良敏感信息的占 6.56%, 表达强烈情感的占 15.77%, 使描述更加幽默或生动活泼的占 25.91%, 上述 3 者都有的占 25.32%, 其他的占 23.44%。

变体词可以是一个具有新意思的常规词、现有词的重新排列组合或者完全是一个新造的词。例如, 现在各大 BBS、博客等讨论历史的版块, 经常可以看到用"常凯申①"来代替"蒋介石"。目前变体词在社交媒体中获得了广泛的使用, Chen 等人[3]分析新浪微博的数据发现, 提取的 37 个主题中有 11 个主题的推文中含有变体词, 甚至有的主题含有 5 个变体词。

实际上可以将变体词看作一种反语言(anti- language)。反语言是由著名语言学家 M.A.Halliday 提出, 是指与主流语言相背离的, 具有自身特有表达内涵的语言形式[5]。反语言具有如下的特征[5]:

1) 反语言是一种全新的构词语言, 在构建的过程中对词语进行重新编码, 重新编码的方式很多, 最直接的就是用新出现的词汇代替之前的词汇。反语言的语言规范是一种比喻性表达, 非直译表达, 不能根据表达的表面意思进行理解, 有一词多义的现象。

2) 反语言与主流语言的语法大体一致。

3) 反语言中的某些词汇, 虽然与主流语言中的词汇一致, 但是其表达的内容与主流语言表达的内容很可能差异非常大。

4) 反语言就像一种密码, 只有其圈子内的人员才能了解明白反语言表达的内容内涵, 而外部人员一般是不会明白的。

由上述的内容可知, 变体词完全符合上述反语言的特征。因此可以借鉴反语言的研究成果加深对变体词的理解。

变体词通常进行了某种隐喻, 已不再是其表面字词的意义了, 因此变体词的存在使得现有的自然语言处理工具直接应用于社交媒体文本时效果不甚理想, 比如词性标注、依存分析、分词、命名实体识别等, 而很多基于关键词的算法和应用也经常达不到预期效果, 比如情感分析、事件发现等。因此, 研究变体词的识别和规范化对于深度自然语言理解任务是十分关键和重要的。

变体词的识别主要是指在文本中发现哪些是变体词, 变体词的规范化是指找出变体词所对应的被替换的目标实体词。

变体词的识别和规范化可用于自动化理解快速演化的社交媒体语言, 帮助人们理解新出现的词汇, 有利于信息提取、语义的深层理解等方面。为下游的自然语言处理任务提供强有力的支撑, 如命名实体识别、分词、消歧、隐喻识别、实体关联等。

此文是对当前变体词的识别和规范化研究进展的综述。主要包括: 变体词识别和规范化的形式化定义; 基于变体词的特点, 综合分析了变体词的生成规律以及变体词识别和规范化的主要难点; 当前主要研究成果及其代表性工作; 最后是变体词识别和规范化技术发展趋势的展望。

本文的结构如下:第 2 节为变体词识别和规范化的形式化定义; 第 3 节详述了变体词的特点和生成规律; 第 4 节为变体词的识别和规范化技术; 第 5 节是变体词识别和规范化的发展趋势; 最后是总结。

2 变体词的识别和规范化的形式化定义

变体词的识别和规范化指发现变体词的提及(morph mention)和变体词的解析(找到变体词所对应的目标实体词)。

已知文档集合 $D=\left\{d_1,d_2,\cdots,d_{|D|}\right\}$, 文档集合 D 中唯一词集合为 $T=\left\{t_1,t_2,\cdots,t_{|T|}\right\}$, 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 m_j^p 为 m_j 在一个特定文档 d_j 里的第 p 次出现。

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

① 常凯申, 为蒋介石之错译名。出自清华大学历史系副主任王奇所著《中俄国界东段学术史研究: 中国、俄国、西方学者视野中的中俄国界东段问题》一书中, 对 Chiang Kai-shek(即蒋的韦氏拼音写法及介石的粤语拼音)的翻译。

认为是变体词的提及。例如, 如果“小马哥”通过上下文获知其指向是电影《英雄本色》里周润发饰演的角色, 则不是一个变体词的提及; 但如果指向的是台湾地区领导人马英九, 则认为是一个变体词的提及。

因此变体词识别和规范化的首要任务是判断 m_j^p 在 d_j 上下文环境下是否是变体词的提及;

下一步针对每一个变体词提及 m_j^p 解析出其目标实体词 e_1。针对上例, 则需解析出变体词“小马哥”的目标实体词为“马英九”。

最终目标是获得变体词集合 $\mathrm{M}=\{m_1,m_2,\cdots\}$, $M \subseteq T$ 和对应的目标实体词集合 $\mathrm{E}=\{e_1,e_2,\cdots,e_{|E|}\}$。

3 变体词特征分析及生成规律

为了实现对变体词的识别和规范化, 首先需要分析变体词的特点, 其次分析目前人工产生的变体词都符合哪些生成规律。

3.1 变体词的特点

总的来说, 变体词具有下面的特点。

1) 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。众多流行的变体词都是通过社交网络自媒体产生并广泛传播的。

2) 绝大多数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替换,因此变体词更接近于行话、黑话、术语等。

3) 变体词与目标实体词之间映射关系不是全射关系, 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。

4) 变体词随着时间的推移会迅速演化, 根据新的新闻热点、特殊事件不断地产生新的变体词。有些变体词会逐步消亡, 而有些则可能进入规范文本。

3.2 变体词的生成规律

要实现对变体词的高效识别和规范化, 首先要了解变体词是如何生成的。

首先, 早期的变体词多采用同音异形异义词, 这也是生成变体词的一个重要手段。Li 等人[6]注意到同音异形异义词在中文中是十分普遍的, 中文字数虽然很多, 但是语音是有限的。据统计中文中 80%的单语音字是有歧义的, 而且其中有一半对应 5 个甚至更多的字。

其次, 充分利用中文的特点生成变体词。Chen 等人[7]发现中国互联网用户喜欢利用中文文字的拆分组合、翻译、昵称等手段来创建变体词。

当前变体词的生成方法逐步丰富, 主要利用深层语义信息、背景知识、特定事件等综合生成变体词。Zhang 等人[4] 基于 548 个随机选择的变体词, 从社会认知角度分析了人工产生的变体词的生成方法, 总结了人们创建变体词的意图和 8 个主要的生成方法。

现将目前分析发现的变体词生成方法总结如表 1 [4,6-8]。

4 变体词的识别和规范化

4.1 变体词的识别和规范化的挑战

变体词的识别和规范化不同于传统的命名实体识别等技术, 有自身的难点:

1) 含有变体词的文档往往不规范,如社交网络上推文、BBS 上的帖子等。而且含有变体词的文本通常为短文本, 含有大量的噪声, 往往缺乏足够的上下文。

2) 面向海量的社交网络媒体数据, 变体词的比例并不大, 因此需要实现面对大规模语料的变体词快速识别技术。当前缺乏大量的标注数据, 此还需要关注可以减少标注代价的识别技术。

3) 由变体词的生成方式和目的所决定其含义通常都是暗示性质的, 因此也导致了变体词通常含有歧义。

4) 变体词与上下文环境缺乏其字面所感知意义的关联。在传统语言中“现代汉语动词的语义特征之间存在着内在的意义关联”[9]。但是由变体词产生的语境所决定的, 在网络语言中含有变体词的很多动词词组已经与其字面所感知的意义毫无相容之处。

5) 当前大量的变体词是根据人物映射、历史背景知识、特定事件等激发而产生的。(即表 1 中第 7、8、9 种生成方式)。单纯基于词汇上的特征是很难捕捉到的这 3 种方式所生成的变体词, 需要利用深层语义信息和上下文。

6) 变体词及其目标实体词通常具有不同的传播渠道和周期, 目标实体词多出现在规范文本且相对比较稳定, 而变体词多出现在非规范文本, 且随着时间迅速演化。

4.2 变体词的识别和规范化技术

目前变体词的识别与规范化的相关研究可以分成两个部分:

- 变体词的识别与规范化: 识别出相关文档中的变体词, 并且找到对应的目标实体词。
- 变体词的自动生成: 分析变体词的生成规律, 由计算机自动生成变体词, 与人工生成变体词进行比较, 分析其自动生成的可行性。

表 1　变体词生成规律总结

序号	类别		生成方法	示例
1	缩写	汉语拼音缩写	汉语拼音首字母	JS: 奸商 GCD: 共产党
		英文缩写	英文单词首字母	BF: 男朋友(Boy Friend) GF: 女朋友(Girl Friend) PLA: 中国人民解放军
		缩句词新成语	由一个句子或多个词语缩减而成的, 或者由原有成语中字面引申、改造等成为带有新意义的(即旧词新用)词语。以四字词居多, 也存在三字词及两字词	喜大普奔: 喜闻乐见、大快人心、普天同庆、奔走相告 不明觉厉: 虽然不明白是什么, 但是感觉好厉害啊
2	语音替换	汉字谐音及同首字母	将汉语转换成拼音, 然后根据容易混淆的、发音接近的、同音异形异义的拼音组合①来进行替换	萨科齐: 傻客气(”Sa Ke Qi”-> “Sha Ke Qi”)[6] 砖家: 专家 霉体: 媒体
		嗲与萌化(或方言谐音)	又称为娃娃音, 也就是故意把原本正确的发音取其相似音或者和音用汉字写出来	酱紫: “这样子”的连音。 稀饭: 喜欢。 肿么了: “怎么了”的谐音。
		数字谐音	用阿拉伯数字代替词语	1314: 一生一世。 3166: 日语“再见”(日语: さようなら, 音译“撒由娜拉”) 520/521: 我爱你
		复合	上述各种谐音的组合	3Q/QQQ: Thank you.(谢谢你。) 3X/3QS: Thanks.(=Thx) V5: “威武”的谐音
3	汉字拆分组合	拆字	某些字拆分后的词根依旧是表意的词, 则用拆分后的词来代替原词	古月: 胡
		合字	如果相邻的词合并起来的词也存在, 则用合并后的词来代替原来的多个词	壕: 土豪 煋: 火星
4	昵称		主要是重复一个实体名字的最后一个字	幂幂: 杨幂
5	语义解释		将实体名 e 分解成字的组合(c_1, c_2, …, c_k), 查找含有这些字(或谐音, 同音异形异义)的词组, 如果有偏负面的词组, 则用词组来代替原来的 e	空心菜: 蔡英文
6	翻译和音译	翻译	对应一个实体, 若其英语中一个成分是常见的英文词, 则用此英文词的汉语翻译来代替	拉里 鸟儿: 拉里 伯德(Larry Bird) 树丛: 布什(Bush) 河文档: 道格·里弗斯 (Doc Rivers)
		音译	从英语、日语、韩语等音译过来的词	纳尼: 什么(日语, なに,表示反问或惊讶) 欧巴: 哥哥(韩语, 오빠) 撸瑟/卢瑟: 失败者(英语, Loser 的谐音)
7	人物映射		用历史人物、小说中的人物来对应现实的人物	乔帮主: 乔布斯 (乔帮主, 乔峰, 武侠小说《天龙八部》里的人物) 小马哥: 马英九(小马哥, 香港电影《英雄本色》里的人物)
8	由历史知识、背景知识产生		由目标实体词的相关历史、相关背景而产生的	猴子: 丰臣秀吉(源自长相, 见司马辽太郎的《新史太阁记》) 乌龟: 德川家康(以善于隐忍着称) 帝都: 北京 魔都: 上海② 空一格③: 蒋介石
9	由特定事件促发		发生了与目标实体词相关的特定事件, 促发了变体词的产生	葫芦爹: 张艺谋④ 常凯申: 蒋介石 公孙永浩: 罗永浩⑤ 周带鱼: 周水平⑥

① http://en.wikipedia.org/wiki/Pinyin#Initials_and_finals 可以查到容易混淆的拼音组合。
② 出自旅居上海的日本作家村松梢风的小说《魔都》。
③ 以前台湾写到“总统”、“总裁”、“蒋总统”或“蒋中正”时, 都必须使用挪抬(在人名及称谓的前面空一格)以示尊敬。后来中国大陆部分网民在提到蒋的姓名、别名、别号以及绰号时, 也会“空一格”(如“千古完人　空一格”)以示调侃。
④ 源于 2013 年 5 月媒体曝光张艺谋严重超生, 当时传言其育有七个孩子, 因而被网友谑称“葫芦爹”。
⑤ 罗永浩曾在新浪微博承诺, 如果锤子 Smartisan T1 手机价格低于 2500, 就是孙子。发售不久该手机价格即降到 1980 元, 网民遂称之“公孙”, 意为“公共的孙子”。
⑥ 一名网络写手, 被称作“带鱼”是因为曾在一篇文章中声称浙江舟山有养殖带鱼, 后经调查发现目前尚无人工养殖带鱼技术。

下面分别介绍当前的主要进展和代表性成果。

4.2.1　变体词的识别与规范化

通用的变体词识别与规范化的架构如图 1 所示, 包括如下步骤:

1) 变体词的识别: 候选变体词的发现, 候选变体词的验证。

2) 变体词的规范化: 变体词的候选目标实体词的发现, 变体词的候选目标实体词的打分排序, 输出最优的目标实体词。

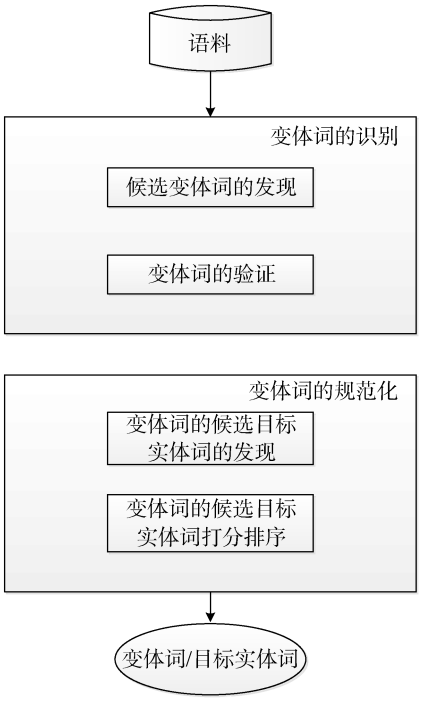

图 1　通用的变体词识别与规范化的架构图

明确的变体词概念出现在[3][10]以及同时期相关的论文中, 但是变体词的相关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。

综合上述研究成果, 下面主要从变体词的识别与规范化的技术角度来进行阐述。

变体词的识别与规范化基本上就是从 3.2 节介绍的变体词的生成规律入手, 由易到难。

- 早期主要是基于规则处理一些简单的相似的字符、数字之间转换的情况。
- 后续逐步注意到中文汉字的特点: 如同音异形异义字、缩写、语音的替换, 拆字组字等。
- 然后从中文英文的对比分析入手, 增加了翻译、音译等识别手段。
- 目前逐步增加了基于语义表示的分析和比较, 特别是随着深度学习的兴起, 研究人员开始利用神经网络获得变体词及其目标实体词的语义特征, 然后通过比较词向量的相似度来实现变体词的识别和规范化。

但是针对人物映射、特定事件、特定历史背景知识生产的变体词的识别和规范化目前还缺乏有效的技术手段。今后的方向应该更进一步深入到语义理解层面, 只有从深层语义层面把握变体词及其目标实体词之间的差异性和相似性, 才能进一步提高识别的准确度, 提高针对人物映射、特定事件、特定历史背景知识生产的变体词的识别和规范化的能力。

1) 基于规则的识别和规范化方法

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[11,12], 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, 具体包括: 首先通过观察变体词总结变体词的变体规则, 进一步提取变体词的 bigram、词干等特征基于分类的方法实现对变体词的识别, 或者根据汉语的语音特点建立语音映射模型, 基于语音的相似性度量实现对变体词的识别。

Yoon[13]总结发现某些变体词实际上是将某些字母转化成形状相似的特殊字符, 如“shit”转换成“sh!t”。陈儒等人[14]提出了针对中文网络的 5 种变体词变异规则: 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。李钝等人[15]根据 ASCII 码, 繁体 BIG 码, 简体中文 GB2312 码等不同编码的固定编码规则, 建立变体词变体规则识别出信息中夹杂的汉字拼音、简繁体混排、特殊符号等。

Sood [16]在对不良文本及其变体信息进行检测的时候, 采用机器学习的方法, 通过采用 bigram、词干等作为特征值来对文本信息做分类分析, 以检测出变体词。李少卿[17]针对拉丁语或英语, 从语音相似和字形相似等角度来计算不良文本变体的相似度, 基于相似度来对不良文本变体进行检测。

Xia[18] 和 Wong[19]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。

2) 基于统计和规则的识别和规范化方法

主要是将统计的方法与规则的方法相结合, 分别提取统计特征和基于规则的特征, 建立变体词与目标实体词之间的映射关系, 然后通过分类的方法基于上下文相似性和字面相似性实现对变体词的规

范化。

Wang[20]从规范化角度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。Choudhury[21]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, 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[22]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拼写错误类型进行了概率建模。

还有通过构建规范化词典用于文本规范化任务。例如, Han[23]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。Han[24]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。

Li[25]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, 主要针对同音异形异义词、缩略语、首字母缩写词、音译等。他们注意到一个现象, 有时可以在非规范短语附近发现对应的规范短语。Li[25]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, 但是严重依赖于搜索引擎返回的结果。

3) 基于语义表示的识别和规范化方法

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”[26]。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[27]。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。

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。

因此基于语义表示的方法主要是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。

Huang 等人[11]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 2 所示[11]。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:字面特征(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。

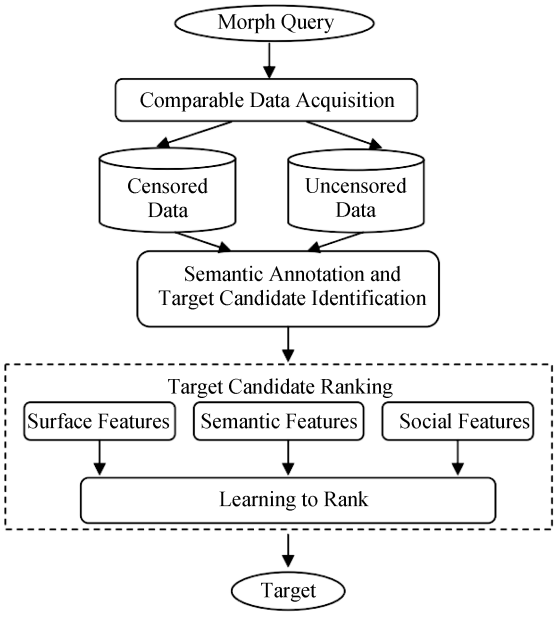

图 2 变体词的识别与解析流程图[11]

其中社会特征主要是对用户的行为建模, 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。

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。

Zhang 等人[28]提出了一个端到端的无监督的方法,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。文章基本上按照图 1 的步骤进行: 1)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; 2)采用半监督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; 3)在发现目标实体词阶段提出了 2 个算法: 基于多数据源的监督学习和连续词袋模型。

基于多数据源的监督学习如图 3 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。第 2 个算法采

用连续词袋模型(如图 4 所示)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。

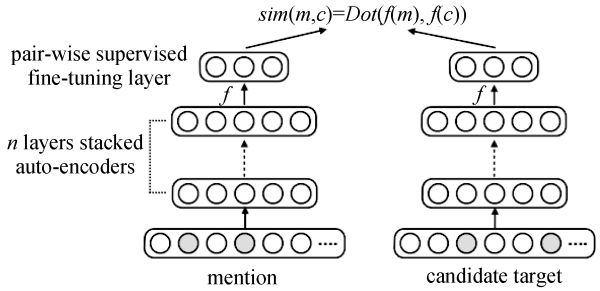

图 3　多数据源的监督学学习

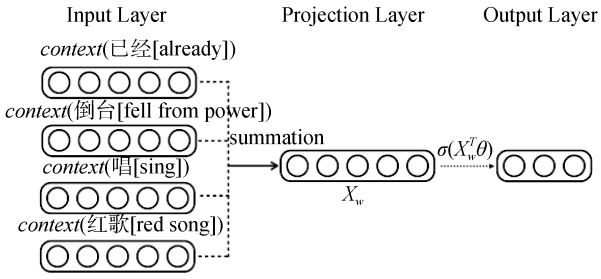

图 4　连续词袋模型

其端到端的变体词的识别与规范化的性能与 Huang[10]方法的比较如表 2 所示, 这也是目前比较好的结果, 从中可以看出对社会媒体的变体词识别与规范化还有很大的提升空间。

表 2　端到端变体词的识别与规范化的性能比较

方法	准确率	召回率	F1
Huang[11]	40.2	33.3	36.4
Zhang[12]	41.1	35.9	38.3

4) 与其他应用的结合

因为变体词的识别与规范化与下游的自然语言处理任务实际上是相互影响相互作用的, 例如分词。因此可以将变体词任务与下游的任务结合起来, 形成一个闭环相互反馈相互提高。

Wang [29]将中文微博变体词的发现与中文分词结合起来。这两个本身就是相互依赖的, Wang 提出了 2 层 FCRF(阶乘条件随机场)模型。在将两者结合起来后,两者的性能都有所提高。而且此文也注意到此方法发生错误的地方, 包括: 观察到的非规范词不完整的时候; 特别短的句子(非规范的词本身就构成了一个句子, 与上下文的句子语用是相关的, 但是词汇上的相关性很弱); 随心所欲创造的新的命名实体。

还有采用基于图的方法。Hassan[30]提出了一种基于二部图随机游走的方法, 该方法首先通过随机游走得到全局优化的基于上下文相似性的规范化候选列表, 然后利用非规范词与规范词之间的字面相似度, 对规范化候选列表进行排序。Sönmez[31]提出了一种综合使用字面特征、上下文特征和语法特征的社交媒体规范化方法, 其中上下文特征和语法特征是从构建好的词关联图中得到。

综上所述, 目前基于语义的变体词的识别和规范化的基本思路如下:

1) 基于变体词及其目标实体词的特征使用分类的方法对变体词进行初筛, 采用的特征包括: 字面上的特征, 语音上的特征, 语言模型, 基于生成规律总结的规则。

2) 基于变体词及其目标实体词的时空分布假设进行验证, 包括: 两者时间上分布的相似性, 共现、共指的规律, 所属用户的兴趣、行为相似性。

3) 基于神经网络获得变体词及其目标实体词的语义表示, 通过语义相似度比较对候选目标实体词进行打分排序。

4.2.2　变体词的自动生成

变体词的自动生成即分析人工生成的变体词, 总结其生成规范, 基于规则或统计的方法实现变体词的自动生成, 使行文更加生动有趣, 使相关主题、事件传播的更加广泛。变体词的自动生成与变体词的识别实际上是相辅相成, 相互促进的, 变体词的自动生成技术也会促进其识别和规范化技术的进步。

变体词的自动生成

基于 3.2 节介绍的变体词的生成规则, 除了最后三条规则外, 其他的都可以由计算机自动生成。最后三条生成规则需要增加相关的人物、特定事件、历史和背景的知识。

首先出现的就是自动生成同音异形异义的变体词。中文是一个音调语言, 每个字的音是由根音(root sound)和它的音调决定的。有些字通过多个语音来代表不同的意思。从汉语的特征出发, 词由字组成, 由音调来决定一个字的意思, 字的意思组合构成了词的意义。虽然中文的书写只有一种标准, 但是存在着各种各样的方言。虽然音调的改变会改变一个字的意思, 但是人们通常会通过上下文来判断出一个不准确的音调背后的真实的意思。

Hiruncharoenvate[32]研究针对新浪微博自动生成非确定的同音异形异义的变体词, 并且不影响用户的理解。Hiruncharoenvate 从新浪微博的语料中计算字的出现频率, 共获得 12,166 个字, 419 个根音(忽略音调), 其中有 3365 个字含有多个根音。根据字的出现频率, 计算了每个根音中各个字所占的百分比。

对应字 c 及其对应的语音 r,计算 r 的百分比 p: c 对也发语音 r 的其他字符的相对频率。若一个词 W 由 $w^1w^2w^3\cdots w^n$ 组成, 则 W 的同音异形异义词 W_i 由每个字的同音异形异义词组合而成 $W_i^1W_i^2W_i^3\cdots W_i^n$ 通过下面的公式来计算一个同音异形异义词的频率分数:

$$\text{score}(W_i)=\sum^{n}p(w_i^k)$$

为了避免选择冷僻的词会对包含冷僻的词的组合进行惩罚。为了保证每个不会选择同一个同音异形异义词, 会随机从前 20 个中进行选择。

Zhang[4]根据表 1 变体词生成规律中: 语音替换、汉字的拆分、昵称、翻译和意译、语义解释的定义通过计算机实现了变体词的自动生成。针对人物映射, 尝试了基于历史人物映射的变体词的自动生成: 收集了 38 个著名的历史人物, 包括: 政治家、国王、诗人、将军、总理、学者等。

Zhang 还提出了一种叫做特征建模的变体词生成方式。首先收集尽可能多的语料, 然后基于上述语料使用谷歌(Google)的 Word2vec 计算出所有词向量。给定一个实体词,计算语料中的词与这个词的语义关系,然后根据余弦相似度、正面倾向性、负面倾向性、是否低频等综合指标进行排序, 把排序前面的词加上原来实体词的姓,形成一个新的变体词。文章中的例子是: 姚明=>姚奇才。

变体词的评测

目前变体词生成效果的评测主要是采用用户问卷调查的方式[4,32], 用户在看到含有变体词的媒体内容后回答问题。这些问题主要包括: 1)哪个是变体词, 指向的目标实体词是哪个? 是否合适? 2)理解内容是否有困难? 3)变体词是否让内容有趣?

调查结果[4,32]显示计算机自动生成的变体词可以达到 66%人工生成达到的效果。而且基于翻译与意译的方法, 计算机产生的结果要优于人工生成的, 可能的原因是计算机搜索的字典空间更大。一个有趣的现象是评测的人只能理解 76%的人工产生的变体词, 可能的原因是: 1)变体词新近产生的, 还不能很好地描述目标实体词的特征; 2)评测的人如果没有跟踪当前的热点, 或者不具备相应的背景知识, 则很难理解此变体词。在趣味性方面,人工产生的变体词要优于计算机自动生成的变体词。

Zhang[4]用 Huang[10]的变体词的识别和规范化方法来验证他们从新浪微博提取的人工产生的 151 个变体词和计算机自动生成的 247 变体词。结果发现, 计算机自动生成的变体词更不容易被发现, 毕竟此变体词的识别系统是基于人工产生的变体词进行训练的, 计算机自动生成的变体词的某些特征还没有被此识别系统所掌握。但是计算机自动生成的变体词的规范化准确度要高于人工生成的, 可能的原因是人工生成的变体词的含义更加隐蔽。

目前还缺乏对自动生成的变体词的评价标准和机制, 主要还是采用人工判断的方式。这里的自动化评价标准主要是指如何判断自动生成的变体词是否符合网络用户使用语言的习惯, 是否达到人工生成的变体词的水准, 富有生动活泼的特性, 易于被人接受和传播等等。

4.3 总结

综上所述, 虽然变体词的识别和规范化技术获得了长足的进步, 但是还有很多空白的领域有待研究。

1) 识别和规范化的准确度还有待提高, 目前最好的结果: F1 值为 38.3;

2) 目前还缺乏对人物映射、特定事件、历史和背景知识(即第 7,8,9 种变体词生成规则)产生的变体词的有效的识别和规范化手段;

3) 缺乏对变体词的演化规律的研究; 变体词也是在不断地发展变化, 同一个目标实体词在不同的时期会有不同的变体词, 其中有无规律可循, 这些变体词的共同点和差异点。研究变体词的演化规律也就是研究网络语言的演化规律。

4) 变体词的自动生成及其相应的评价标准和手段方面还缺乏足够的研究成果。

5 变体词的识别和规范化的发展趋势和展望

目前变体词的识别与规范化需要迫切解决的问题主要有:

1) 提高变体词识别与规范化的准确度。

2) 找到基于人物映射、特定事件、历史和背景知识产生的变体词的识别与规范化方法。

3) 变体词的演化规律及其对网络语言的影响。

4) 变体词的自动生成技术及其评价标准。

以上 4 点实际上是相通的, 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。

5.1 变体词及其目标实体词之间的相似性与差异性

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。

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。

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。

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。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。

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, 我们才可能在语料中找到变体词。

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。

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关系。

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。

为了能够高效地识别变体词, 并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, 即能体现两者的差异性(这些才能判断某提及是否是变体词), 又能展示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"。

5.2 变体词及其目标实体词的语义表示

基于人物映射、历史与背景知识、特定事件下产生的变体词的识别和规范化实际上与研究变体词的演化规律是相通的。上述 3 种变体词的识别和规范化需要在一个长的时间窗口内及时捕捉到相关事件的发生、发展、消亡, 获取到更丰富长期的背景知识和人物关系映射, 研究变体词及其目标实体词的动态演化规律也会辅助提高变体词的识别的准确度, 因此需要能够体现这种动态演化的语义表示。

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。

自 2006 年 Geoffrey Hinton 等人发表了关于深度学习的文章[33], 深度学习逐渐受到了来自不同领域的研究者们广泛的关注。近年来, 深度学习技术也越来越多的被应用到自然语言处理当中, 其中一种重要的应用方式就是通过深度学习技术学习到一种重要的词汇表达方式, 即词向量 [34,35](又叫 word embedding 或 word representation), 是指用一个 N 维的向量来表示词汇, 其中的每一维都是相应词语的隐含特征。一般来说, 词向量包含了有用的句法、语义信息, 具有领域独立性。仅使用词级别的语义表示不足以完全地展示变体词及其目标实体词的深层语义关联。因此还需要通过模型, 得到句子和文档级别的语义表示, 具有一定的记忆功能的神经网络如 Memory Network[36]等在变体词的识别和规范化方面应该会有用武之地。

因此变体词的识别和规范化的关键在于找到: 能够展示变体词的动态演化、能够体现这种"**求大同, 存小异**"的变体词及其目标实体词特殊属性的字/词、句子(段落)、文档不同层面的语义表示。

6 总结

变体词的出现降低了自然语言处理技术面对社交媒体等非规范文本的效果, 因此变体词的识别以及目标实体词的发现对于自然语言处理技术是十分重要的。本文是对当前变体词的识别和规范化技术的回顾和总结, 包括变体词的定义和特征, 变体词的生成规律, 当前变体词的识别和规范化的主要技术进展和成果, 最后指出"求大同、存小异"是变体词及其目标实体词的特殊属性, 变体词的识别和规范化关键在于如何找到其准确恰当的语义表示。

参考文献

[1] Paul Hsiung, Andrew Moore, Daniel Neill, and Jeff Schneider. Alias detection in link data sets. *In Proceedings of the Interna-*

tional Conference on Intelligence Analysis, May.2005.

[2] Patrick Pantel. 2006. Alias detection in malicious environments. *In AAAI Fall Symposium on Capturing and Using Patterns for Evidence Detection*, pp. 14–20.

[3] Le Chen, Chi Zhang, and Christo Wilson. 2013. Tweeting under pressure: analyzing trending topics and evolving word choice on sina weibo. *In Proceedings of the first ACM conference on Online social networks*, pp. 89–100.

[4] Boliang Zhang, Hongzhao Huang, Xiaoman Pan, Heng Ji, Kevin Knight, Zhen Wen, Yizhou Sun, Jiawei Han and Bulent Yener, Be Appropriate and Funny: Automatic Entity Morph Encoding ,*Proc. the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2014.

[5] 丁建新.作为社会符号的“反语言”——“边缘话语与社会”系列研究之一[J].*外语学刊*.2010(02).

[6] Li, P., and Yip, M. C. 1996. Lexical ambiguity and context effects in spoken word recognition: Evidence from Chinese. *In Proceedings of the 18th Annual Conference of the Cognitive Science Society*, pp.228–232.

[7] Chen, L.; Zhang, C.; and Wilson, C. 2013. Tweeting under pressure: Analyzing trending topics and evolving word choice on sina weibo. *In Proc. COSN* ’13.

[8] 中国大陆网络语言列表, https://zh.wikipedia.org/wiki/中国大陆网络语言列表, 2015.12.

[9] 武文杰, 徐艳, 现代汉语视觉动词语义相容度认知分析[J]. 河北大学学报: 哲学社会科学版, 2013(6): 90-92.

[10] Hongzhao Huang, Zhen Wen, Dian Yu, Heng Ji, Yizhou Sun, Jiawei Han and He Li, Resolving Entity Morphs in Censored Data, *Proc. the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2013.

[11] Dinakar K, Reichart R, Lieberman H. Modeling the detection of textual cyberbullying[C], *Intemational Conference on Weblog and Social Media-Social Mobile Web Workshop*. 2011: 11-16.

[12] Yin D, Xue Hong L,et al. Detection of harassment on web 2.0[J]. *Proceedings of the Content Analysis in the WEB*, 2009, 2.

[13] 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), *2010 IEEE 10th International Conference. IEEE*, 2010, 643-650.

[14] 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.

[15] 李钝, 曹元大, 万月亮. 信息安全中的变形关键词的识别[J]. *计算机工程*, 2007, 33(21): 155-156, 159.

[16] Sood S 0, Antin J, Churchill E F. Using Crowdsourcing to Improve Profanity Detection[C]//AAAI Spring Symposium Series. 2012: 69-74.

[17] 李少卿, 不良文本及其变体信息的检测过滤技术研究, 硕士学位论文, *复旦大学*, 2014.4.

[18] Yunqing Xia, Kam-Fai Wong, and Wenjie Li. 2006. A phonetic-based approach to chinese chat text nor- malization. *In Proceedings of COLING-ACL2006*, pp. 993–1000.

[19] K.F. Wong and Y. Xia. 2008. Normalization of Chinese Chat Language. *Language Resources and Evaluation*, pp. 219–242.

[20] 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)*.

[21] 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.

[22] P Cook, S Stevenson. An unsupervised model for text message normalization[C]//*Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, 2009:71-78.

[23] 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.

[24] B Han, P Cook, T Baldwin. Automatically constructing a normalization dictionary for microblogs[C]//Proceedings of the 2012 joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012:421-432.

[25] 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)*, pp. 1031–1040.

[26] Zellig S Harris. Distributional structure. Word, 1954.

[27] Gottlob Frege. Über sinn und bedeutung. *Funktion - Begriff - Bedeutung*, 1892.

[28] Boliang Zhang, Hongzhao Huang, Xiaoman Pan, Sujian Li, Chin-Yew Lin, Heng Ji, Kevin Knight, Zhen Wen, Yizhou Sun, Jiawei Han and Bulent Yener, Context-aware Entity Morph Decoding,*the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2015.

[29] Aobo Wang and Min-Yen Kan. 2013. Mining informal language from chinese microtext: Joint word recog-nition and segmentation. *In Proceedings of the Association for Computational Linguistics (ACL2013)*.

[30] H Hassan, A Menezes. Social Text Normalization Using Contextual Graph Random Walks[C]//*Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 2013: 1577-1586.

[31] C Sönmez, A Ozgür. A Graph-based Approach for contextual Text Normalization[C]//*Porceedings of Conference on Empirical Methods in Natural Language Processing(EMNLP)*.2014:313-324.

[32] Hiruncharoenvate, C., Lin, Z. & Gilbert, E. (2015). Algorithmically Bypassing Censorship on Sina Weibo with Nondeterministic Homophone Substitutions.. *In M. Cha, C. Mascolo & C. Sandvig (eds.), ICWSM* (p./pp. 150-158), : *AAAI Press*. ISBN: 978-1-57735-733-9.

[33] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786): 504–507, 2006.

[34] T Mikolov, I Sutskever, K Chen, et al. Distributed representations of words and phrases and their compositionality [J]. *Advances in Neural Information Processing Systems*. 2013, 3: 3111-3119.

[35] Q Le, T Mikolov. Distributed Representations of Sentences and Documents[C]//*Proceedings of the 31st Internation Conference on Machine Learning(ICML-14)*. 2014:1188-1196.

[36] J. Weston, S. Chopra, and A. Bordes. Memory networks. In International Conference on Learning Representations (ICLR), 2015.

沙瀛　于 2002 年在中国科学院计算技术研究所计算机软件与理论专业获得博士学位。现任中国科学院信息工程研究所副研究员。研究领域为自然语言处理。研究兴趣包括: 社会计算、网络舆情等。Email: shaying@iie.ac.cn

梁棋　于 2014 年在电子科技大学信息安全专业获得硕士学位。现任中国科学院信息工程研究所研究实习员。研究领域为信息检索、舆情计算。研究兴趣包括: 社交网络数据采集与分析。Email: liangqi@iie.ac.cn