计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(5):144−151 [doi: 10.15888/j.cnki.csa.007418]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

结合关联置信度与结巴分词的新词发现算法①

曹　帅

(中国石油大学 (华东) 计算机科学与技术学院, 青岛 266580)
通讯作者: 曹　帅, E-mail: s17070785@s.upc.edu.cn

摘　要: 在中文自然语言处理领域中, 分词是非常重要的步骤之一, 它是关键词抽取、文本自动摘要、文本聚类的基础, 分词结果的好坏直接影响进一步文本处理的准确性. 近年来随着微博平台、直播平台、朋友圈等自由舆情平台的兴起, 大量不规范使用的舆情文本尤其是不断出现的新词给分词结果的准确性带来了巨大的挑战, 新词发现成为分词算法必须解决的问题. 为解决在新词发现过程中, 新词整体数据体量小、新词用法灵活以及过度合并词语易形成短语块等问题, 本文提出了结合关联置信度与结巴分词的新词发现算法, 该算法以结巴分词的初步分词结果为基础, 通过计算词语与其左右邻接词集中各个词语之间的关联置信度, 将被错误拆分的词语合并成候选新词, 并通过切分连接词以防止多个词语被连接成短语的情况出现. 以微博言论数据进行测试的实验表明, 相比于其它基于置信度的分词方法结果, 本文提出的算法可以大幅度提升发现新词尤其是命名实体、网络用语的准确率, 在确保新词语义完整的前提下降低新词长度, 并且在少量测试语料的情境下, 本文提出的算法对低频新词依然具有识别能力.

关键词: 自然语言处理; 分词; 置信度; 新词发现; 命名实体

引用格式: 曹帅.结合关联置信度与结巴分词的新词发现算法.计算机系统应用,2020,29(5):144–151. http://www.c-s-a.org.cn/1003-3254/7418.html

New Word Detection Algorithm Combining Correlation Confidence and Jieba Word Segmentation

CAO Shuai

(College of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China)

Abstract: Word segmentation is one of the most important steps in Chinese natural language processing, it is the basis for keyword extraction, automatic text summarization, and text clustering, the quality of the word segmentation directly affects the accuracy of further text processing. In recent years, with the rise of free public opinion platforms such as Microblog, live broadcast platform, and WeChat Moments, a large number of new words have brought great challenges to word segmentation methods. To solve the problem such as the small overall amount of new words, the flexible usage of new words, and excessive merging of words leads to the formation of phrase blocks in the process of new words discovering. This study proposed a new word detection algorithm combining correlation confidence and Jieba word segmentation. The algorithm is based on the preliminary word segmentation results by Jiaba library in Python, then calculates the correlation confidence between adjacent words to merge incorrectly split words into candidate new words, and by splitting the conjunctions to prevent multiple words from being connected into phrases. Compared with other confidence-based word segmentation methods, the proposed algorithm can greatly improve the accuracy of discovering new words, especially named entities and network terms, and reduce the length of new words while ensuring the integrity of new words. In the context of a small amount of test corpus, the proposed algorithm still has the ability to recognize low frequency new words.

Key words: natural language processing; word segmentation; confidence; new word detection; named entities

① 收稿时间: 2019-10-09; 修改时间: 2019-11-04; 采用时间: 2019-11-22; csa 在线出版时间: 2020-05-07

随着自然语言处理以及大数据分析技术的快速发展, 基于网络中的文本数据来获取网络舆情空间中的热点话题已经越来越快速、准确. 同时以微博、微信为代表的社交网络平台已经融入百姓的日常生活, 人们不仅可以基于社交网络平台获取新鲜资讯, 同时可以在平台上自由发表个人观点与日常生活状态, 这使得民众自由发表的言论成为社交网络平台中文本数据的主体. 然而, 社交媒体中的文本具有口语化程度高、来源广泛的特点[1], 这对自然语言处理技术的准确性带来了新的挑战.

在中文自然语言处理技术中, 文本分词是十分重要的过程, 分词过程是将一段中文文本进行切分, 从而识别一段文本中的各个词语. 分词之后的文本数据可以进行词向量训练、语义聚类等进一步的分析处理, 从而得到准确的情感分析、文本摘要等目标结果. 但多数分词算法如词典匹配、正逆向最大匹配法、双向最大匹配法以及 CRF 序列标注[2]等, 均从文本的语法、词性规律入手, 依赖于标注过词性、词频等信息的词典来对文本进行分词, 从而得到尽可能消除歧义的分词结果, 然而面对网络舆情空间中充斥着的大量不规范使用、不符合词性规律、口语化严重以及包含大量词典未登录词的文本语料时, 传统分词算法对命名实体以及网络用语等新词的识别能力较差.

在中文自然语言处理工具中, Jieba(结巴分词) 是一个简单、高效、灵活的 Python 工具库, Jieba 工具库中的语句简洁凝练, 提供多种模式对中文文本进行分词, 并且可以自由修改词典文件, 深受广大自然语言处理初学者喜爱. 为解决当前分词算法对于新词识别的痛点, 本文提出使用关联置信度与连接词拆分的新词发现算法, 通过与结巴分词结果结合的方式, 对网络舆情文本中的新词进行发现, 该算法不依赖于完备的已标注词典以及庞大的训练语料, 采用计算一个词与其左右邻接词集中各个词之间的关联置信度的方法, 对被结巴分词错误拆分的词语进行合并, 并对包含连接词的较长新词进行拆分, 从而实现对新词尤其是命名实体、网络用语的发现功能, 并避免新词结果过长形成短语块的情况出现.

1 新词发现

在中文自然语言处理中, 新词指的是在分词词典中没有收录, 但又确实能称为词的词语[1], 如一些命名实体 (人名、地名) 以及网络用语, 此类词语灵活性大, 大多不符合语法、构词法规律, 但又能表示一定的语义, 是否能准确识别文本中的新词会直接影响后续算法对文本语义的判断, 新词识别的准确率已成为评价分词算法的重要标准.

目前国内外针对新词发现的算法分为两类, 包括基于频繁模式的新词发现算法以及基于序列标注的新词发现算法.

1.1 基于频繁模式的新词发现算法

基于频繁模式的新词发现算法通过统计方法计算词语在语料中的相关信息, 进而确定需要合并或重新切分的词语. 王雪瑞等[3]针对直播弹幕文本, 引入边界增强上下文熵的概念, 通过计算词语的上下文熵确定一个被拆分的词应该与其左右哪一侧的词语进行合并, 从而得到新词; 顾森[4]基于大量的文本语料, 以词语的内部凝聚程度以及自由运用程度为判别标准, 计算词语在文本语料中的熵值, 文献认为熵值低的词可能是被错误拆分的词, 应对这样的词进行合并处理从而得到新词; 王欣[5]利用多字互信息来判定一个词语的内部结合程度, 并通过邻接熵确定词语的左右边界, 该方法从内外两方面入手, 认为一个词语应该是内部紧密关联且与外部其它词语相对独立的, 通过计算两种标准提高了新词发现的准确性; Zhang 等[6]利用带有词性标注和词频标记的词典, 结合互信息和最大熵模型在分词过程中发现新词; 陈小红[7]等针对游戏领域文本, 通过从相关语料库中抽出部分游戏术语、简称与知识库进行结合, 对文本数据进行词性、领域标记, 该方法将分词算法转换为知识库实体链接问题, 在特定领域下提高了分词结果的准确性; 王珊珊等[8]通过对一个时间跨度的文本特征进行分析比对, 为每个词语添加时间跨度信息来判断该词是否为新词、热点词; 翟畅[9]提出了基于非结构化文本的目标领域未登录词识别策略和基于商业语料库融合的分词词典构建策略, 通过统计词语在文本语料中的频度信息, 结合领域术语知识库, 对特定领域的文本进行新词发现; 张婧等[10]通过词向量训练得到弱成词词串集合, 结合词频统计对候选新词进行了有效过滤, 在社交媒体语料分词实验中取得了较高的准确率; 袁华等[11]提出了基于最大置信度的中文复合新词发现方法, 该方法采用数据挖掘中 Apriori 算法的支持度与置信度, 将新词发现任务转化为频繁模式发现任务, 首先获取文本数据的频繁模式,

再利用剪枝操作对频繁模式结果进行精简, 最终得出复合新词结; 吴宏洲[12]与李亚松[13]等均通过从语料中提取候选词, 计算候选词的支持度与置信度, 并分别设定阈值对候选词进行筛选, 从而完成对文本语料的新词抽取工作.

基于频繁模式的新词发现算法首先要求候选词具有一定的频繁程度, 进而以词语的熵值、互信息、支持度与置信度的计算为主, 熵值包括信息熵、邻接熵, 熵值可以衡量词语在语料中的混乱程度, 一个词语的熵值越高, 则该词在语料中越灵活, 可以跟多个其它词语组成上下文, 这说明这种词语本身就是比较完备的整体, 可以算作一个词语, 而熵值越低说明能与该词语组成上下文的词语越少, 这说明该词可能是一个被错误切分的词, 需要与其左右邻接词组成一个完整的词语; 互信息用以衡量一个词内部各个文字之间的相互依赖程度, 互信息越大说明组成该词的各个字词单元不应再进行拆分; 支持度与置信度是数据挖掘中的常用指标, 分别衡量了事务项的频繁程度与关联程度. 基于频繁模式的新词发现算法通常是在初步分词后, 对分词结果进行再检验, 对错误拆分的词语进行合并, 对错误合并的词语进行拆分, 从而发现语料中的新词. 然而在包含新词的新事件发生初期, 总体数据量较少, 难以利用统计法发现频繁模式, 在以微博、微信朋友圈、直播弹幕为主的网络短文本中, 由于词语搭配的出现偶然性较大, 在少量语料中计算得到的熵值并不准确, 从而会导致多个词语被合并成短语块的情况出现, 从而使得分词结果的粒度变粗, 进而影响后续文本处理结果的准确性.

1.2 基于序列标注的新词发现算法

基于序列标注的新词发现算法通过对文本中词语的词性、文字在词中的位置等信息进行标注, 通过预测序列变化的方式对文本进行分词, 在分词过程中发现新词. 曹菲[14]解决了基于 Hash 的正向回溯算法解决分词过程中的歧义问题, 并将 CRF 模型与正向最大匹配相结合, 从而提高对文本中命名实体的识别准确率; 李博涵等[15]针对 Prefixspan 算法在文本序列标注过程中容易出现的问题, 如序列模式不连续、序列模式项之间存在包含关系等, 融合了词语的频繁模式对算法进行了改进, 该方法基于词语的频繁程度、词性与语义对序列标注的结果进行过滤从而得到新词; 色差甲等[16]将最大熵模型嵌入到隐式马尔科夫模型 HMM 中, 进而对文本的 BEMS 序列进行标注, 该方法在藏文新词如人名、地名、机构名、事件名等方面具有良好的效果, 提高分词结果的正确率、召回率以及 F 值接近 2 个百分点; 周霜霜等[17]提出了一种融合规则和统计的微博新词发现算法, 该方法通过对微博文本词语的构词规则进行归纳总结, 同时结合 CRF 模型对文本的 BEMS 序列进行标注, 从而提高对微博文本分词和词性标注的精度.

基于序列标注的新词发现算法, 是在基于序列标注的分词算法基础上进行改进, 使得序列标注的准确性有所提高, 以 BEMS 序列为标注法的新词发现算法居多, 其中 B 表示词语开始的字, M 表示词语中间的字, E 表示词语结束的字, S 表示单字词语. 对文本的序列标注依赖于有标注的训练语料, 通过训练语料才能得出序列变化的概率分布模型, 而针对社交媒体中的短文本数据缺乏可靠的已标记语料, 且人为标注的方法费时费力.

针对上述两大类新词发现算法中的不足, 本文受数据挖掘中 Apriori 算法的启发, 提出结合关联置信度与结巴分词的新词发现算法, 该算法以结巴分词中精确模式的分词结果为基础, 通过计算词语与其左右邻接词集中各个词语之间的关联置信度, 找出被错误切分的词语, 将多个词语间关联置信度高的词语合并成候选新词, 之后通过识别候选新词中的连接词来防止多个词语被连接成短语块的情况出现, 从而实现在单条舆情文本而非整体数据集中的新词发现工作.

在第 2 部分中, 本文介绍了结巴分词的基本原理, 对本文提出的新词发现算法中的 6 个组成部分进行定义, 并阐述算法的具体流程; 第 3 部分中, 本文以微博文本为实验语料, 测试本文提出的新词发现算法的性能, 并与结巴分词、文献[12]中基于最大置信度的中文复合新词发现方法的结果进行对比实验, 并设置不同参数来验证参数对新词结果的影响; 在第 4 部分中, 将对本文的贡献进行总结, 并对后续工作进行展望.

2 结合关联置信度与结巴分词的新词发现算法

结巴分词是一种融合了 Trie 树词图扫描、动态规划以及 HMM 模型的分词算法, 结合了基于词典匹配的分词算法简单、准确以及基于序列标注的分词算法对词典未登录词具有区分能力的优点, 结巴分词提供

3 种分词模式, 分别是精确模式、全模式以及搜索引擎模式, 精确模式试图将句子最精确地切分, 全模式则是将句中所有可能成词的词语都找出, 搜索引擎模式是在精确模式的基础上, 对长词再次切分得到适合用于搜索引擎的分词结果. 由于全模式以及搜索引擎模式的分词结果会将一个词语多次拆分, 因此其结果会出现一个词包含另一个词的情况, 从而改变了原文本的顺序, 若在此结果上对错误拆分的词语进行合并会得到错误的结果. 本文采用以结巴分词提供的精确模式分词结果为基础, 该模式的分词结果粒度细, 极少出现词语被错误合并的情况, 但词语易被错误拆分, 因此本文提出的算法通过计算一个词语与其左右邻接词集中各个词语的关联置信度来对错误拆分的词语进行合并得到候选新词, 进而采用判断候选新词中的连接词左右平均关联置信度的情况对过度合并的候选新词进行拆分. 针对算法中的分词结果、左右邻接词集、关联置信度、关联置信度阈值、候选新词以及连接词有以下定义:

定义 1. 分词结果: 设一段文本为 T, 通过调用 Python 中的 Jieba 库中的 cut() 方法, 将文本 T 作为参数传入, 得到一个有序的列表 $Tc=[w_1, \cdots, w_i, \cdots, w_n]$ $(1 \le i \le n)$, 其中 w_i 表示一个字词单元, 它可能是一个单独的字, 也可能是一个词语, 也可能是标点符号, 下标 i 并不是字词单元在序列中出现的位置, 而是表示字词单元的唯一标识, 即一个 w_i 表示一个唯一的字词单元, 同一字词单元可以在 Tc 中多次出现, 称 Tc 为文本 T 的分词结果.

定义 2. 左右邻接词集: 在文本 T 的分词结果 Tc 中, 对于一个字词单元 w_i, 其左侧邻接词集 N_L 是由 w_i 在 Tc 中每个出现位置左侧的字词单元 (标点符号除外) 构成的集合; 其右侧邻接词集 N_R 是由 w_i 在 Tc 中每个出现位置右侧的字词单元 (标点符号除外) 构成的集合.

定义 3. 关联置信度: 对于两个字词单元 w_i 与 w_j, 若 w_i 与 w_j 分别存在于彼此的左侧邻接词集与右侧邻接词集, 或分别存在于彼此的右侧邻接词集与左侧邻接词集时, 则可以计算 w_i 与 w_j 的关联置信度, 关联置信度用以衡量在文本 T 中一个字词单元出现的情况下, 另一个字词单元与之相邻出现的概率, 计算公式如式 (1) 所示:

$$Conf\left(w_i \to w_j\right) = P\left(w_j|w_i\right) = \frac{P\left(w_i w_j\right)}{P(w_i)} \tag{1}$$

$Conf(w_i \to w_j)$ 表示字词单元 w_i 出现时, w_j 与其邻接出现的关联置信度, 计算方式为 w_i 与 w_j 在文本 T 中邻接出现的概率除以 w_i 在文本 T 中出现的概率, 基于大数定律, w_i 与 w_j 在文本 T 中邻接出现的概率可以用 w_i 与 w_j 在文本 T 中邻接出现的次数除以分词结果 Tc 的列表长度表示, 同理, w_i 在文本 T 中出现的概率可以用 w_i 在文本 T 中出现的次数除以分词结果 Tc 的列表长度表示.

定义 4. 关联置信度阈值: 在计算两个字词单元 w_i 与 w_j 的关联置信度时, 需要分别计算关联置信度 $Conf(w_i \to w_j)$ 与 $Conf(w_j \to w_i)$, 关联置信度阈值 Th 用以规定两个字词单元可以合并成候选新词时需要达到的最小关联置信度的值, 只有当 $Conf(w_i \to w_j)$ 与 $Conf(w_j \to w_i)$ 均大于 Th 时, 才称两个字词单元 w_i 与 w_j 满足关联置信度阈值.

定义 5. 候选新词: 当两个字词单元 w_i 与 w_j 满足关联置信度阈值 Th 时, 可对两个字词单元进行合并得到 (w_i+w_j), 称 (w_i+w_j) 为一个候选新词.

定义 6. 连接词: 若字词单元 w_i 与 w_j 可以合并为 (w_i+w_j), 且字词单元 w_j 与 w_k 可以合并为 (w_j+w_k), 则可合并为候选新词 $(w_i+w_j+w_k)$, 称字词单元 w_j 为候选新词 $(w_i+w_j+w_k)$ 的连接词; 若候选新词由多个字词单元组成, 则除了第一个和最后一个字词单元, 其余构成候选新词的字词单元均为该候选新词的连接词.

基于上述定义, 结合关联置信度与结巴分词的新词发现算法包括文本分词、关联置信度计算以及连接词拆分 3 个步骤.

2.1 文本分词

文本分词步骤采用 Jieba 工具库提供的精确分词模式, 对待进行新词发现的文本 T 进行拆分得到其分词结果 Tc. 本文所提出的新词发现算法, 是对少量网络舆情文本而非整体数据集进行新词发现, 由于网络舆情文本长度较短, 若一个词语在一条网络舆情文本中只出现一次, 则通过定义 3 计算得到的该词与其左右邻接词的关联置信度容易达到 100%, 因此该词与其左右邻接词极易满足关联置信度阈值从而成为一个候选新词.

为避免此类偶然情况的发生, 本文提出的新词发

现算法采用将多条网络舆情文本进行合并以加长待分词文本长度. 设网络舆情文本数据共有 n 条, 每 m 条文本数据进行合并得到合并文本段 $T'=T_1+T_2+\cdots+T_m$, 共得到$\lceil n\div m\rceil$个合并文本段, 对每个合并文本段进行分词, 得到多个分词结果 Tc', 对每个分词结果 Tc'进行新词发现.

2.2 关联置信度计算

对于一个分词结果 Tc', 通过遍历 Tc'中的各个字词单元得到每个字词单元 w_i 在 Tc'中出现的次数以及 w_i 与其左右邻接词集中的各个字词单元在 Tc'中邻接出现的次数, 基于式 (1) 计算每个字词单元与其左右邻接词集中各个字词单元之间的关联置信度, 将所有满足关联置信度阈值 Th 的字词单元对进行合并, 得到候选新词集 W'.

候选新词集 W'是将所有满足关联置信度阈值的字词单元对合并得到的集合, W'中的候选新词可能是将被结巴分词过度拆分的词语合并得到的正确新词结果, 也可能是被过度合并形成的短语块, 形成短语块的原因可能是多个字词单元构成的短语在文本中多次、单调出现, 单调出现指的是构成该短语块的字词单元在合并文本段 T'中只与构成该短语的其它字词单元邻接出现, 导致这些字词单元之间均满足关联置信度阈值而被合并, 有些短语可能是较长的命名实体, 而有些则是包含动词结构的实际短语, 例如“我/去/上学”, 短语不属于新词发现的范畴, 因此需要对候选新词集 W'中的候选新词进行过滤筛选, 从而得到最终的新词结果.

2.3 连接词拆分

在本文提出的新词发现算法中, 定义了构成短语的字词单元中连接词的存在, 在连接词拆分步骤中, 遍历候选新词集 W'中的候选新词, 若一个候选新词由 3 个及以上的字词单元构成 (即候选新词中有存在连接词), 则对其进行以下操作:

(1) 找出候选新词中的连接词, 判断每个连接词与其左右邻接词的平均关联置信度的大小情况, 平均关联置信度为 $Conf(w_i\rightarrow w_j)$ 与 $Conf(w_j\rightarrow w_i)$ 的平均值;

(2) 若一个连接词与其左侧、右侧字词单元的平均关联置信度值不同, 则将该候选新词进行拆分, 拆分点为连接词与其左、右侧字词单元平均关联置信度较小的两字词单元之间;

(3) 若一个连接词与其左侧、右侧字词单元的平均关联置信度值相同, 则保持两个字词单元合并的状态, 继续判断候选新词中的下一个连接词;

(4) 通过将带有连接词的候选新词进行拆分, 去除由单个字词单元组成的结果, 得到最终新词结果.

通过对候选新词中的连接词进行拆分, 可以把由多个字词单元合并成的短语块进行拆解, 降低了最终新词结果的粒度, 使得拆分出来的新词更为独立, 从而防止新词淹没在短语中.

3 实验分析

为验证本文提出的新词发现算法的有效性, 笔者使用网络爬虫在新浪微博中爬取了 2019 年 9 月至 10 月内涉及体育赛事以及国庆档电影总共 1 GB 的纯文本数据, 包括 2644 494 条微博言论, 512 948 156 个字符. 实验通过设置不同合并文本数 m 与关联置信度阈值 Th 进行了多组实验. 由于实验所使用的数据时效性较新, 网络中缺乏相应的正确词语标注内容, 因此本文无法使用正确率、召回率等性能指标对新词发现结果进行评价, 本文将结合具体情况对结果进行分析.

专利[11]与文献[12,13]均采用了置信度作为新词发现的有效工具, 对初步分词结果进行处理以得出新词结果. 然而在计算置信度之前, 需要计算词语的支持度, 只有满足最小支持度也就是在文本语料中出现次数高于一定阈值的字词单元, 才有资格与其它字词单元计算置信度, 并且在得出候选词后, 没有对候选词进行切分操作. 实验中选用文献[12]中的新词抽取方法, 同样对结巴分词精确模式的初步分词结果进行处理, 与本文所提出的新词发现算法的结果进行了对比.

实验所用操作系统采用 Windows 10 专业版操作系统, 处理器为 Inter(R) Core(TM) i7-8700K CPU, 3. 70 GHz, 16 GB 内存, 实验代码使用 Python 3.5 编写, 其中 Jieba 工具库的版本号为 0.39.

3.1 合并文本数 m 对比

本文所提出的新词发现算法, 采用了合并多条文本数据的方式以扩大单条文本长度, 从而减少偶然出现的词语搭配的关联置信度达到 100% 的情况出现, 因此合并文本数 m 的设置会影响最终新词发现结果的情况. 在本组实验中首先设置 m=100, 即每 100 条微博言论数据融合为 1 条文本数据, 融合处理后共有 26 445 条融合后的文本数据, 使用 Jieba 工具库的精确模式对每条文本数据进行分词, 标出标点符号, 之后针对每条融合文本数据的分词结果, 统计每个字词单元

在当前融合文本中出现的次数并计算字词单元与其左右关联词集中各字词单元的关联置信度, 设置关联置信度阈值 Th=0.9, 根据平均关联置信度的大小对合并出的带有连接词的候选新词进行切分, 共得到新词结果 1874 个. 新词是从多个合并文本段中得出, 最终新词结果是由每个合并文本段中得到的新词合并而来, 新词的出现次数则是在全部合并文本段中出现的总数, 部分新词结果如表 1 所示; 令 m=10, Th=0.9, 使用相同语料共挖掘出新词 567 个, 部分新词结果如表 2 所示, 在表 2 中去除了表 1 中出现过的新词. 其中在文献[12]的新词抽取方法中同样对融合文本数据进行处理, 并设支持度为 10%, 置信度为 90%.

表 1 合并文本数 m=100 部分新词结果

本文算法新词结果	出现次数	新词抽取方法结果	结巴分词结果	注释
郭艾伦	3536	郭艾伦	郭艾/伦	中国男篮运动员
快闪	1412	快/闪	快/闪	一种行为艺术
易烊千玺	1193	易烊/千玺	易/烊/千玺	我国男歌手
张常宁	1003	张常宁	张常/宁	中国女排运动员
正能量	835	正/能量	正/能量	一种积极的情感
文牧野	581	文/牧/野	文/牧/野	中国导演
虎扑	501	虎/扑	虎/扑	国内体育网站
拉皮诺埃	480	拉皮/诺/埃	拉皮/诺/埃	美国女足运动员
黄景瑜	480	黄景/瑜	黄景/瑜	我国男演员
国庆档	345	国庆档期间	国庆/档	指国庆期间的电影
卡拉库尔特	336	卡拉/库尔特	卡拉/库尔特	土耳其女排运动员
秒拍	302	秒/拍	秒/拍	一种短视频
四川航空	293	四川航空	四川/航空	我国航空公司
流浪地球	253	流浪地球	流浪/地球	我国科幻电影
世界波	213	世界/波	世界/波	形容精彩的进球

表 2 合并文本数 m=10 部分新词结果

本文算法新词结果	出现次数	新词抽取方法结果	结巴分词结果	注释
微博视频	61 320	微博视频	微博/视频	微博短视频名称
中国机长	21 889	国庆档电影中国机长	中国/机长	微信体育公众号
推广曲	1236	推广/曲	推广/曲	宣传歌曲
北京时间	856	北京时间	北京/时间	北京时区的标准时
段奥娟	586	段奥/娟	段/奥/娟	我国女歌手
韩东君	583	韩东/君	韩东/君	我国男演员
X 玖少年团肖战	553	X/玖/少年团/肖战	X/玖/少年/团肖战	偶像团体成员
五犯离场	362	五犯离场	五犯/离场	体育用语
博斯科维奇	332	博斯/科维奇	博斯/科维奇	土耳其女排运动员
郭子瑄	326	郭子/瑄	郭子/瑄	中国女排运动员
红海行动	318	红海/行动	红海/行动	电影名
三巨头	286	三/巨头	三/巨头	形容 3 名主力运动员
黄金联赛	263	黄金/联赛	黄金/联赛	赛事名
不敌	261	不/敌	不/敌	形容输给某方
曲尼次仁	173	曲尼/次/仁	曲尼/次/仁	我国女演员

由上述实验可见, 本文新词发现算法在对命名实体的识别上具有出色表现, 可合并两组及以上被结巴分词过度拆分的词语, 准确发现国内外人名、影视剧名及网络用语等; 参数 m 的设置对新词发现结果具有一定影响, m 值越小则单条语料中字符数越少, 因此一组词语搭配在单条语料中出现的频次更少, 词语搭配的多个词语之间更容易达到关联置信度阈值从而被本文算法认定为新词. 如“微博/视频”、“北京/时间”等词语, 组成新词的两个词语均具有明确的意义, 组合后则成为具有不同含义的命名实体词; 如“X/玖/少年/团肖

战”、“曲尼/次/仁”等复杂命名实体的词语, 在结巴分词完全错误的情况下可清晰准确对过度拆分的词语进行合并. 组成此类复杂命名实体新词中的连接词如“少年”、“次”在总体语料中独立出现的次数非常多, 通过减少参数 *m* 的设置可减少此类连接词在单条语料中出现的次数, 从而使其更易与其它词语满足关联置信度阈值以得出准确的新词结果.

文献[12]中的新词抽取方法在字词单元较为频繁时, 但由于缺少对候选词的拆分过程, 因此容易将字词单元连接成较长的短语, 如“国庆档电影中国机长”; 而当字词单元的频繁程度低于最小支持度时, 新词抽取方法则不认为该字词单元可能新词的组成部分, 因此无法识别出现次数较少的新词.

3.2 关联置信度阈值 *Th* 对比

关联置信度阈值 *Th* 是衡量两个字词单元能否合并为候选新词的界限, *Th* 的值越低则两个字词单元更容易被合并. 在本组对比实验中, 设置 *Th*=0.5, *m*=100, 得到的部分新词发现结果如表 3 所示, 在表 3 中词去除了在表 1 中出现过的新词. 其中在文献[12]的新词抽取方法中设支持度为 10%, 置信度为 50%.

表 3 关联置信度阈值 *Th*=0.5 部分新词结果

本文算法新词结果	出现次数	新词抽取方法结果	结巴分词结果	注释
中国机长	21 889	国庆档电影中国机长	中国/机长	电影名
我和我的祖国	20 651	电影我和我的祖国	我/和/我的/祖国	电影名
蔡徐坤	1476	蔡徐坤	蔡/徐坤	我国男艺人
开学第一课	1312	开学第一课	开学/第一课	节目名
龚翔宇	1253	龚翔/宇	龚翔/宇	中国女排运动员
女篮亚洲杯	1113	女篮/亚洲杯	女篮/亚洲/杯	赛事名
陈飞宇	803	陈飞/宇	陈飞/宇	我国男演员
东京奥运会	760	东京/奥运会	东京/奥运会	赛事名
欧阳娜娜	596	欧阳/娜娜	欧阳/娜/娜	我国女艺人
国际排联	556	国际/排联	国际/排/联	国际排球联合会
更多	543	更/多	更/多	形容还有很多
字母哥	532	字母/哥	字母/哥	男篮球员外号
搭起	496	搭/起	搭/起	动词
热搜	153	热/搜	热/搜	形容热点事件
C 位出道	142	C/位/出道	C/位/出道	形容团队核心人物

对比表 3 与表 1 的结果可以发现, 通过降低关联置信度阈值 *Th* 的取值, 可以提升对由“中国”、“我的”等高频词汇组成的新词的发现能力, 由于实验所用语料临近新中国成立 70 周年, 因此“中国”、“祖国”等词语出现的次数较多, 其它词语很难与“中国”、“祖国”达到较高关联置信度, 但在降低关联置信度阈值后, 本文提出的算法成功挖掘出新词“中国机长”、“我和我的祖国”, 这两个词语为 2019 年国庆档热映的电影名, 是不同于其组成部分词语意义的命名实体词; 且可以看出其在语料中出现极为频繁; 同时由于降低了关联置信度阈值, 如同“开学/第一课”、“女篮/亚洲/杯”、“东京/奥运会”等新词被挖掘, 这些新词的每个组成部分都具有实际意义, 合并在一起时可以让命名实体包含更多信息, 并且更加符合文本想要表达的实体.

在降低置信度取值后, 文献[12]中的新词抽取方法会结合出更多的短语结果, 同时对低频度的新词结果的识别能力较差, 如要增加该方法对低频新词的识别能力, 需要进一步降低支持度阈值, 然而降低支持度阈值后会导致计算时间增加以及错误结果的出现.

关联置信度阈值 *Th* 的取值需要结合实际的语料情况, 若设置过低则会挖掘出过多不正确的新词, 不正确的新词结果中容易出现包含“的”、“我”、“和”等高频字. 对于不正确的新词结果, 可以通过高频字词典对新词结果进行过滤, 去除包含高频字的新词结果.

通过实验验证, 本文所提出的结合关联置信度与结巴分词结果的网络舆情新词发现算法在新词尤其是命名实体的发现工作中具有出色的表现, 结巴分词在处理人名时, 虽可以准确识别姓氏, 但结巴分词难以准确分出包含 3 个及以上字数的人名, 结巴分词通常将姓氏字与跟在姓氏字后面的第一个字分为一词, 跟在姓氏字后面的第二个字则视为单独的字, 从而造成错误识别人名的情况, 本文提出的算法通过计算组成人

名的字词之间的关联性, 在面对 3 个及以上字数的人名时同样可以准确识别此类命名实体; 针对网络舆情中的网络用语以及网络媒体命名实体, 结巴分词通常会将此类词语过度拆分为单个的字, 本文提出的算法可以将其进行合并, 从而得出词典中未登录且具有实际意义的网络用语词.

本文所提出的新词发现算法是在少量文本语料中进行新词发现, 虽然对网络舆情文本尤其是以微博、直播弹幕为主的超短文本进行了合并以加长单条文本数据的长度, 但令文本合并数 m=10 时, 单条合并文本字段符数不超过 1000 个, 本文所提出的新词发现算法依然可以准确发现新词. 因此本文提出的新词发现算法不依赖于庞大的数据量, 在新的网络舆情事件发生的初期, 缺乏数据量的情况下依然可以及时发现新事件文本数据中包含的新词.

由于本文提出的新词发现算法不进行字词单元频繁程度的判定, 因此相比于其它应用置信度对新词进行抽取方法, 可以对低频度的新词进行识别, 同时在通过关联置信度连接字词单元后, 通过对连接词进行拆分, 确保了新词结果不会过长导致出现短语块结果.

4 结论与展望

本文提出的结合关联置信度与结巴分词的新词发现算法, 是在 Jieba 工具库精确模式分词结果的基础上, 通过计算字词单元与其左右邻接词集中各个字词单元之间的关联置信度, 将满足关联置信度阈值的字词单元进行合并得到候选新词, 之后根据候选新词中连接词与其左右邻接词的平均关联置信度大小关系对候选新词进行拆分, 从而弥补了结巴分词容易将词语过度拆分导致分词结果不正确的错误, 同时避免了字词单元在合并过程中过度合并导致形成短语块的问题.

实验表明本文提出的新词发现算法可以准确识别新词, 尤其是命名实体以及网络用语, 并且该算法可以在语料数较少的数据集中准确发现新词, 适应了在新事件发生初期, 包含新词的数据量较少的实际应用情况, 使得文本分析工作可以在新事件爆发前率先发现新词从而快速提取新事件信息.

本文提出的新词发现算法是对结巴分词的初步分词结果进行修正, 在今后的工作中将对本文提出的算法与其它较为成熟的分词工具进行融合, 使新词发现的准确率有进一步的提升.

参考文献

1 张华平, 商建云. 面向社会媒体的开放领域新词发现. 中文信息学报, 2017, 31(3): 55–61.

2 Peng FC, Feng FF, McCallum A. Chinese segmentation and new word detection using conditional random fields. Proceedings of the 20th International Conference on Computational Linguistics. Geneva, Switzerland. 2004. 562–568.

3 王雪瑞, 刘渊. 基于边界增强的中文直播弹幕新词发现. 传感器与微系统, 2018, 317(7): 142–146, 150.

4 顾森. 基于大规模语料的新词发现算法. 程序员, 2012, (7): 54–57.

5 王欣. 一种基于多字互信息与邻接熵的改进新词合成算法. 现代计算机 (专业版), 2018, (11): 7–11.

6 Zhang LY, Qin M, Zhang XM, *et al*. A Chinese word segmentation algorithm based on maximum entropy. Proceedings of 2010 International Conference on Machine Learning and Cybernetics. Qingdao, China. 2010. 1264–1267.

7 陈小红, 陈环环, 方之家, 等. 基于领域本体的游戏攻略文本标注算法研究与实现. 计算机应用与软件, 2017, 34(2): 80–86. [doi: 10.3969/j.issn.1000-386x.2017.02.014]

8 王珊珊, 冯利鑫. 基于新词识别的大数据聊天文本舆情热点挖掘. 电子商务, 2018, (1): 60–61.

9 翟畅. 面向非结构化中文文本的本体构建[硕士学位论文]. 武汉: 武汉工程大学, 2017.

10 张婧, 黄锴宇, 梁晨, 等. 面向中文社交媒体语料的无监督新词识别研究. 中文信息学报, 2018, 32(3): 17–25, 33. [doi: 10.3969/j.issn.1003-0077.2018.03.003]

11 袁华, 钱宇, 徐华林. 基于最大置信度的中文复合新词发现方法: 中国, CN201610779163.3. 2017-01-18.

12 吴宏洲. 分词技术的研究与应用——一种抽取新词的简便方法. 软件工程, 2015, 18(12): 64–68. [doi: 10.3969/j.issn.1008-0775.2015.12.025]

13 李亚松, 王玉龙. 一种新词自动提取方法. 电信工程技术与标准化, 2014, (12): 83–86. [doi: 10.3969/j.issn.1008-5599.2014.12.025]

14 曹菲. 基于 Hash 和 CRF 的中文分词算法研究[硕士学位论文]. 镇江: 江苏大学, 2017.

15 李博涵, 蔡永香, 邓舒颖, 等. 基于改进的 Prefixspan 算法的中文文本新词提取方法研究. 电脑知识与技术, 2018, 14(8): 160–163.

16 色差甲, 贡保才让, 才让加. 基于最大熵和 HMM 的藏文新词识别对比研究. 青海师范大学学报 (自然科学版), 2018, 34(1): 12–16.

17 周霜霜, 徐金安, 陈钰枫, 等. 融合规则与统计的微博新词发现方法. 计算机应用, 2017, 37(4): 1044–1050. [doi: 10.11772/j.issn.1001-9081.2017.04.1044]