计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(6):271−277 [doi: 10.15888/j.cnki.csa.007985]

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

基于 DC-Value 的西班牙语文本词语提取方法①

于　娟, 颜煜铃, 简梓炜, 张　晨

(福州大学 经济与管理学院, 福州 350108)
通讯作者: 张　晨, E-mail: Zhangchenfzu@163.com

摘　要: 西班牙语 (以下简称西语) 是仅次于汉语的世界第二大母语语言, 是联合国 6 种官方语言之一. 西语复杂的词形变化和语法规则, 导致 C-value 等经典的词语提取方法的效果无法保证, 进而影响基于西语文本挖掘的效果. 因此, 本文研究西语文本词语提取方法, 为西语文本的结构化建模提供完备的词库. 给定待分析的西班牙语文本, 该方法分 3 步提取得到词语集合: 文本预处理、候选词语提取和 DC-value 成词度计算. 其中, 前两步所得的候选词语集合可直接用作文本挖掘的词库; 第三步所得的候选词语成词度可辅助判断候选词语成词的可能性, 减轻人工判断的工作量. 实验结果表明, 本文方法自动提取的西文词语集合的准确率达到 80%, 且召回率远高于经典方法, 能够为西语文本挖掘提供有效的词库.

关键词: 西语文本; 文本挖掘; 词语提取; DC-value

引用格式: 于娟,颜煜铃,简梓炜,张晨.基于 DC-Value 的西班牙语文本词语提取方法.计算机系统应用,2021,30(6):271–277. http://www.c-s-a.org.cn/1003-3254/7985.html

Extracting Terms from Spanish Corpora Based on DC-Value

YU Juan, YAN Yu-Ling, JIAN Zi-Wei, ZHANG Chen

(School of Economics and Management, Fuzhou University, Fuzhou 350108, China)

Abstract: As one of the six working languages of the United Nations and a major mother tongue second only to Chinese, Spanish has complex morphological changes and grammatical rules. These result in the inability of classic term extraction methods such as C-value and thus affect the effect of Spanish text analysis. This study proposes a Spanish term extraction method to automatically construct a complete lexicon for text modeling. Given a Spanish text or corpus, the method extracts terms in three steps: preprocessing the texts, extracting candidate terms, and calculating term-hood indexes of the candidate terms based on DC-value. The set of candidate terms obtained in the first two steps can be used directly as the lexicon for text mining. Meanwhile, the term-hood indexes obtained in the third step are essential for reducing the manual workload in determining whether the candidates are really terms. According to experiments, the proposed method has a high accuracy of 80% and a recall much higher than that of classic methods, providing the effective lexicon for Spanish text mining.

Key words: spanish text; text mining; term extraction; DC-value

1　引言

随着“一带一路”倡议的推进和全球化进程的加快, 国家之间的经济交流日益频繁与深入, 跨国组织的管理决策依据也不再限于单一语种的信息, 而是希望基于来自全球各语种数据的全局视图. 然而, 相比图像、视频等其它非结构化数据, 文本具有更为显著的语种

① 基金项目: 国家自然科学基金 (71771054)
Foundation item: National Natural Science Foundation of China (71771054)
收稿时间: 2020-09-28; 修改时间: 2020-10-28; 采用时间: 2020-12-12; csa 在线出版时间: 2021-06-01

差异——阅读不懂的语种的文本, 人们能从中接受到的信息几乎是零. 为此, 有必要研究多语种文本的融合分析方法, 以快速获取瞬息万变的国际情况信息, 支持跨国组织的国际化战略管理决策.

西班牙语 (以下简称西语) 是联合国 6 种官方语言之一, 是全球 19 个国家的官方语言, 有四亿多人作为母语使用, 是仅次于汉语的世界第二大母语语言[1]. 相关统计数据显示, 我国与西语国家的双边经贸关系发展迅速, 已成为包括西班牙、智利、秘鲁、墨西哥等大部分西语国家的主要贸易伙伴之一, 未来有着广阔的合作前景[2,3]. 另一方面, 我国尚缺乏西语专门人才, 因此, 如何对来自西语国家的海量文本进行高效的数据挖掘, 已成为我国相关组织的一个重要的管理方法问题.

文本词语提取是文本挖掘的基础工作, 是指自动获取待分析文本中出现的词语, 包括单词和短语. 西班牙文 (以下简称西文) 的单词之间用空格分开, 易于实现自动提取; 但文本挖掘所用的特征词大多是面向文本内容的多词短语, 因此, 西文的短语提取是西语文本词语提取和文本挖掘的关键环节. 另一方面, 相比同属印欧语系的英语, 西语的词形变化规则更为复杂: 名词不仅有单复数变化, 还区分阴阳性; 动词、形容词和代词均需依据所修饰或指代的名词的阴阳性进行相应变化. 并且, 西语和英语的语序也有所不同: 西语中的形容词位置多样化, 即形容词或形容词短语可能位于其所修饰的名词或名词短语的前面、后面甚至其它位置[4]. 种种差异导致现有的比较完善的英语文本词语提取方法不能直接应用于西文词语提取.

为此, 本文研究西文词语提取方法, 结合西语语法规则和串频统计方法, 从西语文本中自动提取包含多词短语在内的词语集合, 以支持西语文本挖掘工作. 本文第 2 节介绍词语提取方法的研究现状; 第 3 节简述本文所提出的西文词语提取方法的框架流程; 第 4 节详细说明西文候选词语的提取过程和方法; 第 5 节介绍计算候选词语成词可能性的成词度算法; 第 6 节通过实验比较分析本文方法与传统的英文短语提取方法 C-value 和 NC-value; 第 7 节给出研究结论.

2 相关工作

词语是某一语言里的词 (也称原子词、单词等) 和固定短语 (也称合成词、词组、多词术语等) 的总称. 其中, 原子词是词语组成的基本单元; 合成词是由多个原子词依据一定规则组合成的短语, 具有其组成部分不能代表的特定含义. 词语提取是文本挖掘的基础工作, 为文本的结构化建模提供词库, 因此, 其召回率和准确率显著影响文本挖掘的效果. 现有的词语提取方法研究大多针对中文或英文文本词语提取[5,6]. 这些词语提取方法可分为主要的 3 类: 基于规则的方法、基于统计的方法和混合方法.

基于规则的方法首先根据语言的词法或句法特征总结构词规则, 然后从文本中提取符合规则的词串作为词语. 因各语言的词法和句法不同, 构词规则一般仅适用于某一特定语言的文本词语提取. 例如, 文献 [7,8] 总结了中文词语构词规则及中文词语提取方法; 文献 [9,10] 总结了英文网页或学术报告中的词语构词规则, 用于自动提取英文术语; 文献 [11–13] 等研究总结了西文的词语提取规则. 这些基于规则的词语提取方法受限于规则库的准确性和全面性. 由于总结构词规则耗时耗力, 且难以用少量规则覆盖多变的词法和句法, 因此这类词语提取方法的召回率大都不高.

基于统计的方法利用概率论和信息论, 从大规模语料中统计多个原子词的邻接共现概率, 提取得到原子词和固定短语. 文献 [14,15] 使用字符的共现率、字符串各部分的互信息、字符串的信息熵等指标从大规模语料库中提取词语. 文献 [16,17] 进一步将词语的上下文信息纳入考量指标, 提出了 C-value 方法及其改进方法 NC-value. 文献 [18] 研究了改进的 C-value/NC-value 方法, 用于提取西文词语. 这些基于统计的词语提取方法不受构词规则的限制, 也较少受到不同语言的影响, 仅在分词和词形规范化阶段因不同语言而异[19]. 这类方法一般都需要大规模语料的支持才能保证准确率, 不适用于可用语料较少的情况.

混合的词语提取方法, 结合使用前述两种方法, 以同时保证结果的高准确率和高召回率. 文献 [20] 结合词性分析与串频统计, 研究了一种提取中文词语的原子词步长法. 文献 [21] 提出一种结合 HITS 与 C-value 的 HC-value 方法, 用于提取英文短语. 文献 [22] 提出一种基于西文语义标注, 结合 TF-IDF 和 NC-Value 的术语提取方法. 此外, 还有一些基于机器学习的术语提取方法[23,24]. 文献 [25] 基于术语的词性特征和上下文等信息训练 SVM 模型, 以提取具有相似位置特征的命名实体. 这些混合方法既结合了基于规则和基于统计的词语提取方法的优点, 又能一定程度地克服两种方

法的不足, 因此优于非混合方法[26]. 混合方法是当前词语提取方法研究的主流.

3 本文方法框架

由于目前针对西语文本词语提取的方法研究较少, 因此, 为了支持西语文本挖掘, 本文借鉴前述中、英文文本词语提取方法, 提出一种结合语言学规则和统计学方法的西文词语提取方法, 利用词法规则和单词共现规律, 从西语文本中自动提取包含单词和短语在内的词语集合. 该方法分为 3 步: 文本预处理、候选词语提取和成词度计算. 方法流程如图 1 所示.

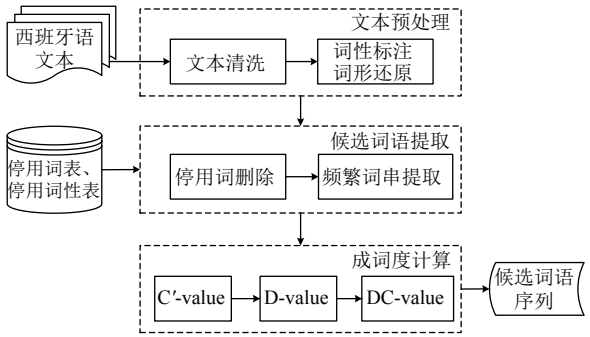

图 1 西语文本词语提取方法流程图

(1) 文本预处理模块, 首先清洗输入的西语文本, 删除其中与词语提取无关的图片、公式、标识符等, 输出统一格式的纯文本; 然后采用现成工具进行词性标注和词形还原, 输出带词性标注的标准化文本. 词性标注是指为每个单词标注其 POS 词性. 常用的西文词性标注工具有: Pattern.es[27]、NLTK[28]、Apache OpenNLP[29]、Stanford core NLP[30]、Polyglot[31] 等. 词形还原是指把名词复数、动词变位等变形的西语单词还原为单词原形. 常用的西文词形还原工具有 Pattern.es[27] 和 NLTK[28] 等.

(2) 候选词语提取模块. 本文依据西语语法特征总结构词率低的单词和词性, 总结形成停用词表和停用词性表. 该模块首先删除前一模块输出文本中的停用词, 得到一个单词串的序列; 然后计算每一单词串的子串及其出现频次, 超出频次阈值的子串为频繁词串; 删除出现频次与父串相同的频繁词串, 其余的作为候选词语输出. 后文第 4 节详细说明该模块的过程和方法.

(3) 成词度计算模块的输入为前一模块输出的候选词语集合, 输出为按成词度降序排列的候选词语序列. 该模块计算每一候选词语的 C′-value 和 D-value 值, 然后将加权和 DC-value 值作为候选词语的成词度. 把候选词语序列交由西语专业人士进行人工判断选择, 可得到最终的西文词语集合. 后文第 6 节详细介绍成词度计算方法.

4 候选词语提取

候选词语提取分为两个步骤: 停用词删除和频繁词串提取.

停用词删除的输入是带词性标注的西语纯文本、停用词表和停用词性表. 该子模块遍历输入文本, 删除其中出现在停用词表和停用词性表的单词, 仅保留位于句首的停用词, 输出一个西文单词串的序列. 停用词是那些参与造句但不参与构词的单词, 如 es(是)、y(和)等; 停用词性是一些鲜少参与构成短语的词性, 如代词、副词、从属连词等. 本文在文献 [20] 的基础之上, 依据经验总结了西文词语提取的停用词表和停用词性表.

频繁词串提取的输入是一个西语单词串的序列, 输出是候选词语集合. 该子模块以单词为步长, 对输入的每一词串以长度优先[32] 取子串, 并计算子串的出现频次, 出现频次大于阈值的作为频繁词串. 为了避免词语提取的截断问题, 出现频次与父串相同的频繁词串不列入候选词语. 例如, 若父辈词串“conjunto/NN de/IN dato/NNS”(数据集) 在文本中出现了 10 次, 且其子串“dato/NNS”(数据) 也出现了 10 次, 则“dato/NNS”不列入候选词语集合.

以一段西语文本的处理为例说明本文的候选词语提取方法. 图 2 左侧为一段西语文本, 右侧为其对应的中文翻译. 图 2 中的文本不具有特殊性. 图 3 为图 2 文本经文本预处理的结果. 不失一般性, 本文采用 Pattern.es 进行西文词性标注和词形还原. 图 4 为图 3 文本删除停用词和停用词性词之后的结果. 为明晰起见, 采用“[”和“]”作为段落起始和终止符. 表 1 为图 4 文本提取频繁词串所得的候选词语, 其中的频次仅记录频繁词串独立出现的次数.

5 成词度计算

成词度是候选词语成词的可能性, 其主要指标是单元度 (unithood) 和领域度 (termhood). 其中, 单元度衡量候选词语内部语言结构的稳定性; 领域度衡量候选词语与某一领域相关的程度. 对每一个候选词语, 本文首先计算其单元度和领域度, 然后结合起来计算成词度.

[Los sistemas de información tratan el desarrollo, uso y administración de la infraestructura de la tecnología de la información en una organización. El mayor de los activos de una compañía hoy en día es su información, representada en su personal, conocimiento y innovaciones. Para poder competir, las organizaciones deben poseer una fuerte infraestructura de información, en cuyo corazón se sitúa la infraestructura de la tecnología de información. De tal manera que el sistema de información se centre en estudiar las formas para mejorar el uso de la tecnología de información que soporta el flujo de información dentro de la organización.]	信息系统涉及组织中信息技术基础设施的开发、使用和管理. 目前, 企业最重要的资产是以其人员、知识和创新为代表的信息. 为了竞争, 组织必须拥有强大的信息基础设施, 其核心是信息技术基础设施. 这样, 信息系统就可专注于研究信息技术的应用来改善组织内部信息的流动.

图 2 西语文本示例及其中文翻译

```
[el/DT sistema/NNS de/IN información/NN tratan/VB el/DT desarrollo/NN]
[uso/NN y/CC administración/NN de/IN la/DT infraestructura/NN de/IN la/DT tecnología/NN de/IN la/DT información/NN en/IN una/DT organización/NN]
[el/DT mayor/JJ de/IN el/DT activo/JJ de/IN una/DT compañía/NN hoy/RB en/IN día/NN es/VB su/PRP$ información/NN]
[representada/VBN en/IN su/PRP$ personal/JJ]
[conocimiento/NN y/CC innovación/NNS]
[para/IN poder/VB competir/VB]
[la/DT organización/NNS deben/VB poseer/VB una/DT fuerte/JJ infraestructura/NN de/IN información/NN]
[en/IN cuyo/WP$ corazón/NN se/PRP sitúa/VB la/DT infraestructura/NN de/IN la/DT tecnología/NN de/IN información/NN]
[de/IN tal/DT manera/NN que/WP$ el/DT sistema/NN de/IN información/NN se/PRP centre/VB en/IN estudiar/VB la/DT forma/NNS para/IN mejorar/VB el/DT uso/NN de/IN la/DT tecnología/NN de/IN información/NN que/WP$ soporta/VB el/DT flujo/NN de/IN información/NN dentro/IN de/IN la/DT organización/NN]
```

图 3 图 2 西语文本的文本预处理结果

```
[el sistema de información tratan el desarrollo]
[uso]
[administración de la infraestructura de la tecnología de la información]
[organización]
[el mayor de el activo]
[compañía]
[día]
[información]
[representada]
[personal]
[conocimiento]
[innovación]
[poder competir]
[la organización deben poseer]
[fuerte infraestructura de información]
[corazón]
[sitúa la infraestructura de la tecnología de información]
[manera]
[el sistema de información]
[centre]
[estudiar la forma]
[mejorar el uso de la tecnología de información]
[soporta el flujo de información dentro de la organización]
```

图 4 图 3 文本删除停用词后的结果

C-value 是常用的英文候选词语单元度计算方法, 但其仅考虑由两个及以上单词组成的词语[17,18], 不能用于计算仅由一个单词构成的词语的单元度. 为了全面比较包含单词和短语在内的候选词语的单元度, 本文对 C-value 计算公式进行改进, 将$C=\log_2|t|$改为$C=1+\log_2|t|$, 使其可以计算单词的单元度. 改进后的算式如式 (1) 所示:

$$C'(t)=\begin{cases} C\cdot tf(t), & t\text{未被嵌套} \\ C\cdot\left(tf(t)-\dfrac{1}{|T_t|}\sum_{b\in T_t}tf(b)\right), & t\text{被嵌套} \end{cases} \tag{1}$$

式 (1) 中, t 为候选词语, $C=1+\log_2|t|$, $|t|$表示 t 的长度; $tf(t)$是 t 在西语文本中出现的频次; T_t表示嵌套 t 的非 t 候选词语的集合, $|T_t|$表示T_t集合中元素的个数.

表 1 图 4 候选词语提取结果

序号	候选西文词语	中文释义	出现频次
1	la infraestructura de la tecnología	技术基础设施	2
2	la tecnología de información	信息技术	2
3	el sistema de información	信息系统	2
4	la organización	组织	2
5	uso	使用	2
6	información	信息	4

C'-value 值越大, 说明候选词语出现的频次越高且被嵌套的情况越少, 则其单独成词的可能性就越大. 在出现频次相等的情况下, 较长的候选词语成词的可能性更大.

在计算候选词语的领域度时, 本文借鉴文献 [33,34], 采用式 (2) 计算领域度:

$$D(t)=\frac{tf(t)}{df(t)}\cdot\sqrt{\frac{1}{N}\sum_{i=1}^{N+1}\left(tf_i(t)-\overline{tf^*(t)}\right)^2} \tag{2}$$

式 (2) 中, t 为候选词语, $tf(t)$表示 t 在西语文本中出现的总频率, $df(t)$表示 t 出现的文本频率; $tf_i(t)$表示 t 在第 i 个文本中出现的次数; N 为 t 出现的文本数. 考虑到部分候选词语仅在 1 个文档中出现, 所以引入第 $N+1$ 个文本对候选词语的分布进行修正, 且$tf_{N+1}(t)$等

于 t 在修正后语料中出现的平均频次.

D-value 值越大, 则候选词语在输入语料中的不同文本之间的分布越不均匀, 说明其越有可能与特定领域相关, 则其是领域专有短语的可能性越大, 因而成词的可能性也就越大.

结合单元度 C′-value 和领域度 D-value 这两个指标计算候选词语的成词度, 算式如式 (3) 所示:

$$DC(t) = \alpha \cdot C'(t) + (1-\alpha) \cdot D(t) \tag{3}$$

式 (3) 中, α 为 0 到 1 之间的一个实数, 用于将 C′-value 和 D-value 融合进一个线性模型. 多次实验的结果表明, α 取值 0.2 时, 成词度的计算结果最准确.

6 实验分析

目前还没有检验西文词语提取方法优劣的通用数据集, 也没有标准的评价指标. 本文采用两组实验, 比较分析本文所提出的西文词语提取方法与传统的 C-value[17] 和 NC-value[18] 方法的性能.

6.1 实验数据

本文采用两个题材不同的西语语料比较分析: 联合国平行语料库[35] 和路透社文本分类语料库[36]. 从联合国平行语料库中随机选取 246 篇西语会议记录作为实验数据一, 共 12.8 MB; 从路透社语料中选取 200 篇新闻报道作为实验数据二, 共 162 KB.

6.2 评价指标

常用的文本挖掘方法检验指标是召回率和准确率. 召回率主要受所采用的候选词语提取方法的影响. 西文词语提取常用的 C-value 方法和 NC-value 方法在提取候选词语时, 只考虑名词和形容词, 基于词性规则, 仅能提取得到符合设定规则的名词和形容词组合. 本文在提取候选词语时, 全面考察各种词性, 仅删除不参与构词的代词、副词、从属连词等, 然后以单词为步长提取由各种词性单词组合而成的频繁词串, 删除其中仅作为子串出现的频繁词串之后得到候选词语集合. 因此, 本文的西文词语提取方法能够提取得到的候选词语数目大幅提高, 约为 C-value 方法和 NC-value 方法的 2.2 倍; 并且, 由于本文方法的准确率较高, 所以召回率也远高于 C-value 方法和 NC-value 方法. 因此, 本文不再比较 3 种西文词语提取方法的召回率, 仅重点评价三者的准确率.

6.3 实验结果与分析

首先对每组实验数据进行文本预处理, 接着以 2 为出现频次的阈值提取频繁词串, 删除频次与父串相同的频繁词串, 得到候选词语集合. 然后计算候选词语的成词度, 即 C-value、NC-value 和 DC-value 值, 并按成词度降序排列输出给西语专业人士判断哪些候选成词. 实验数据一和数据二的候选词语集合分别包含 17 058 条和 1983 条西文候选词语.

尽管 C-value 和 NC-value 方法在候选词语提取阶段的召回率远低于本文方法, 但为了公平比较 3 种方法的准确率, 在成词度计算时为 3 种方法提供了相同的候选词语集合, 均为由本文方法所得到的候选词语集合. 基于人工判断的结果, 从正确率和错误率两个方面分析词语自动提取的准确率, 如表 2、表 3 和图 5、图 6 所示.

表 2 联合国平行语料库词语提取正确率 (%)

数据集	C-value	NC-value	DC-value
Top100	95.0	96.0	96.0
Top800	96.0	93.5	91.5
Top8000	87.5	86.0	85.4
Top13000	88.7	87.6	86.9
Top16000	87.9	87.7	87.0

表 3 路透社语料词语提取结果正确率 (%)

数据集	C-value	NC-value	DC-value
Top100	87.0	87.0	87.0
Top500	82.2	80.8	82.6
Top1000	78.2	75.8	78.3
Top1500	78.2	77.5	77.4
Top1800	80.6	80.5	80.6

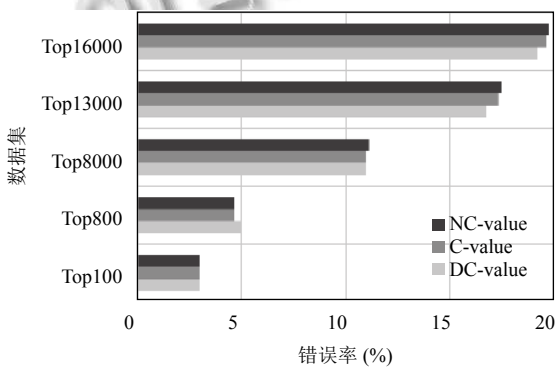

图 5 联合国平行语料库词语提取错误率

正确率和错误率的计算公式分别如式 (4)、式 (5) 所示:

$$\text{正确率} = \frac{\text{成词数}}{\text{候选词语数}} \times 100\% \tag{4}$$

$$\text{错误率} = \frac{\text{不成词数} + \text{误删词数}}{\text{频繁词串数}} \times 100\% \tag{5}$$

其中, 候选词语数是候选词语集合中的词语个数; 成词数和不成词数分别指由西语专业人士判断成词和不成词的候选词语的个数; 频繁词串数是指阈值大于 2 的词串个数, 其中包含那些出现频次与父串相同的子串; 误删词数是那些本应成词但因仅作为子串出现而未列入候选词语的频繁词串个数. 从表 2、表 3 和图 5、图 6 可以看到:

(1) 候选词语集合相同的情况下, 3 种成词度计算方法对候选词语的排序不同, 但准确率持平. 可见, 本文提出的西文词语提取方法在大幅提高召回率的同时, 与经典的 C-value 和 NC-value 方法的准确率持平.

(2) 语料的规模影响着本文方法的准确率. 实验数据二的词语提取准确率略低于数据一, 原因是其语料规模较小, 没能完备地展现串频共现的统计特征. 图 2 的西语文本较短, 其中的频繁词串数量较少, 且常因仅作为子串出现而被误删, 如“tecnología(技术)”、“sistema (系统)”等. 可见, 本文方法更适用于语料规模较大的西语文本的词语提取.

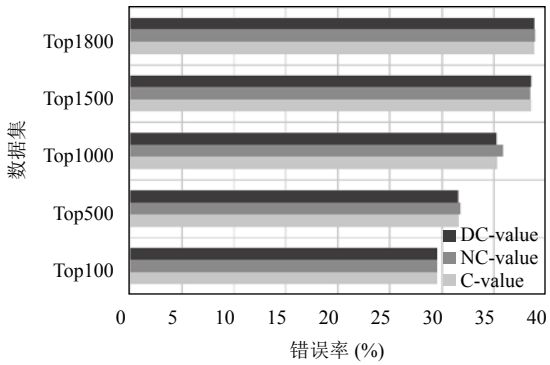

图 6 路透社语料词语提取结果错误率

总之, 在西语文本词语提取方面, 本文方法的召回率显著高于经典的 C-value 和 NC-value 方法; 在成词度计算方面, 3 种方法的准确率区别不大, 都较为令人满意. 并且, 语料规模越大, 本文方法的准确率越高.

7 结论

作为联合国和众多国际组织的工作语言, 西班牙语在全球具有广泛的使用范围, 是仅次于汉语的世界第二大语言, 但目前西语文本挖掘研究尚不成熟, 尤其是针对西语文本词语提取的方法研究. 我国与西语国家的双边经贸关系发展迅速, 为了支持基于西语信息的管理决策, 本文提出一种西语文本词语提取方法, 以支持针对西语文本的文本挖掘和自动分析.

给定待分析的西语文本或语料库, 本文分 3 步自动提取词语集合: (1) 对输入文本进行预处理, 包括: 文本清洗、词性标注和词形还原; (2) 根据西语语法特征总结停用词表和停用词性表, 删除文本中的停用词, 然后基于串频统计提取得到候选词语集合; (3) 计算候选词语的成词度, 以成词度降序输出给人工判断选择, 得到最终的词语集合. 实验表明, 本文方法的召回率显著高于 C-value 和 NC-value 等经典的西语文本词语提取方法, 且准确率与这些经典方法持平.

本文方法适用于大规模西语文本语料的词语提取. 在面向西语文本分类、聚类等文本挖掘任务时, 采用本文方法的前两个步骤 (文本预处理和候选词语提取) 即可无监督地提取待分析文本中的词语集合, 供文本建模从中选取特征词. 在面向西语文本命名实体识别、本体构建、机器翻译等需要准确词语的任务时, 除了前两个步骤, 还需采用本文方法的第 3 步 (成词度计算) 计算候选词语的成词度. 候选词语按成词度降序排列交由西语专业人士进行人工判断确定最终的词语集合, 能够降低人工选词的工作量.

本文的西词文本词语提取方法的准确率受到停用词表和停用词性表的影响, 因此, 未来将在应用研究中继续完善停用词表和停用词性表, 进一步提高西班牙语文本词语提取方法的准确率.

参考文献

1 维基百科. 西班牙语. https://en.wikipedia.org/wiki/Spanish_language. [2020-04-19].
2 商务部国际贸易经济合作研究院. 2018 年西班牙货物贸易及中西双边贸易概况. https://countryreport.mofcom.gov.cn/record/view110209.asp?news_id=63741. [2020-04-19].
3 商务部国际贸易经济合作研究院. 2018 年墨西哥货物贸易及中墨双边贸易概况. https://countryreport.mofcom.gov.cn/record/view110209.asp?news_id=63321. [2020-04-19].
4 王仲. 英语和西班牙语间的差异对英汉和西汉同传策略的影响 [硕士学位论文]. 北京: 北京外国语大学, 2016.
5 NLPIR 大数据语义智能分析平台. http://www.nlpir.org/wordpress/. [2020-04-19].
6 Hasan KS, Ng V. Automatic keyphrase extraction: A survey of the state of the art. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, ML, USA. 2014. 1262−1273.
7 朱波, 侯敏. 基于特征过滤的新词语提取. 北华大学学报

(社会科学版), 2012, 13(5): 18–22.
8 化柏林. 针对中文学术文献的情报方法术语抽取. 现代图书情报技术, 2013, (6): 68–75.
9 Sánchez D. A methodology to learn ontological attributes from the web. Data & Knowledge Engineering, 2010, 69(6): 573–597.
10 Spasić I, Sarafraz F, Keane JA, *et al*. Medication information extraction with linguistic pattern matching and semantic rules. Journal of the American Medical Informatics Association, 2010, 17(5): 532–535. [doi: 10.1136/jamia.2010.003657]
11 García-Sánchez F, Fernández-Breis JT, Valencia-García R, *et al*. Combining semantic web technologies with multi-agent systems for integrated access to biological resources. Journal of Biomedical Informatics, 2008, 41(5): 848–859.
12 Ochoa JL, Almela Á, Valencia-García R. Identifying patterns for unsupervised learning of multiword terms. Educational Research and Reviews, 2011, 6(9): 645–656.
13 Gurrutxaga A, Saralegi X, Ugartetxea S, *et al*. ElexBi, a basic tool for bilingual term extraction from spanish-basque parallel corpora. IXA-group, University of the Basque Country. 2006. 159–165.
14 王璐, 张仰森. 基于典型句型的词语搭配定量分析及提取算法. 计算机科学, 2012, 39(S1): 232–234, 270.
15 刘剑, 唐慧丰, 刘伍颖. 一种基于统计技术的中文术语抽取方法. 中国科技术语, 2014, 16(5): 10–14. [doi: 10.3969/j.issn.1673-8578.2014.05.002]
16 Frantzi K, Ananiadou S. Extracting nested collocations. Proceedings of the 16th Conference on Computational Linguistics. Copenhagen, Denmark. 1996. 41–46.
17 Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms:. The C-value/NC-value method. International Journal on Digital Libraries, 2000, 3(2): 115–130. [doi: 10.1007/s007999900023]
18 Barrón-Cedeo A, Sierra G, Drouin P, *et al*. An improved automatic term recognition method for spanish. Proceedings of the 10th International Conference on Intelligent Text Processing and Computational Linguistics. Mexico City, Mexico. 2009. 125–136.
19 于娟. 基于文本挖掘的本体学习方法研究. 北京: 科学出版社, 2019. 51–74.
20 于娟, 党延忠. 结合词性分析与串频统计的词语提取方法. 系统工程理论与实践, 2010, 30(1): 105–111. [doi: 10.12011/1000-6788(2010)1-105]
21 徐会芳. 可比语料中双语多词术语互译对抽取方法研究 [硕士学位论文]. 大连: 大连理工大学, 2013.
22 Ochoa JL, Valencia-García R, Perez-Soltero A, *et al*. A semantic role labelling-based framework for learning ontologies from spanish documents. Expert Systems with Applications, 2013, 40(6): 2058–2068. [doi: 10.1016/j.eswa.2012.10.017]
23 Lossio-Ventura JA, Jonquet C, Roche M, *et al*. Biomedical term extraction: Overview and a new methodology. Information Retrieval Journal, 2016, 19(1): 59–99.
24 da Silva Conrado M, Pardo TAS, Rezende SO. A machine learning approach to automatic term extraction using a rich feature set. Proceedings of 2013 NAACL HLT Student Research Workshop. Atlanta, GA, USA. 2013. 16−23.
25 Kazama J, Makino T, Ohta Y, *et al*. Tuning support vector machines for biomedical named entity recognition. Proceedings of ACL-02 Workshop on Natural Language Processing in the Biomedical Domain. Philadelphia, PA, USA. 2002. 1−8.
26 Zhang ZQ, Iria J, Brewster C, *et al*. A comparative evaluation of term recognition algorithms. Proceedings of the 6th International Conference on Language Resources and Evaluation. Marrakech, Morocco. 2008.
27 Patten.es. https://www.clips.uantwerpen.be/pages/pattern-es/. [2020-04-19].
28 NLTK 3.5 documentation. http://www.nltk.org/. [2020-04-19].
29 Apache OpenNLP. http://opennlp.apache.org/. [2020-04-19].
30 Stanford CoreNLP-natural language software. https://stanfordnlp.github.io/CoreNLP/. [2020-04-19].
31 Polyglot. https://polyglot.readthedocs.io/en/latest/. [2020-04-19].
32 姜韶华, 党延忠. 基于长度递减与串频统计的文本切分算法. 情报学报, 2006, 25(1): 74–79. [doi: 10.3969/j.issn.1000-0135.2006.01.013]
33 周浪, 张亮, 冯冲, 等. 基于词频分布变化统计的术语抽取方法. 计算机科学, 2009, 36(5): 177–180. [doi: 10.3969/j.issn.1002-137X.2009.05.045]
34 于娟, 党延忠. 领域特征词的提取方法研究. 情报学报, 2009, 28(3): 368–373. [doi: 10.3772/j.issn.1000-0135.2009.03.007]
35 Ziemski M, Junczys-Dowmunt M, Pouliquen B. The united nations parallel corpus v1.0. Proceedings of the Tenth International Conference on Language Resources and Evaluation. Portorož, Slovenia. 2016.
36 Lewis DD, Yang YM, Rose TG, *et al*. RCV1: A new benchmark collection for text categorization research. The Journal of Machine Learning Research, 2004, 5: 361–397.