基于 FP 序列树的法文词语提取方法研究

于 娟[1*]，吴晓鹏[1]，廖 晓[2]，刘建国[3]

(1. 福州大学经济与管理学院 福州 350108；2. 广东金融学院互联网金融与信息工程学院 广州 510521；
3. 上海财经大学会计与财务研究院 上海 杨浦区 200433)

【摘要】法语复杂的语法和词形变化规则导致 N-gram 等词语提取方法的效果无法保证，影响法语文本挖掘的准确性。该文提出一种高效的法文词语提取方法，从待分析的法语文本中自动获取包括单词和短语的词语集合，构建法语文本挖掘所需的词库。该方法把文本中的单词共现信息压缩为 FP 序列树结构，快速提取频繁词串并计算其成词度，得到法文词语集合。实验表明，该方法的准确率高达 90%，且具有比现有法文词语提取方法更高的召回率，能有效支持法语文本挖掘应用。

关 键 词 FP 序列树； 法语文本挖掘； 词语提取； 成词度； 文本压缩
中图分类号 TP182 **文献标志码** A doi:10.12178/1001-0548.2020273

Extracting Terms Form French Corpora with FP Sequence Tree

YU Juan[1*], WU Xiao-peng[1], LIAO Xiao[2], and LIU Jian-guo[3]

(1. School of Economics and Management, Fuzhou University Fuzhou 350108;
2. School of Internet Finance and Information Engineering, Guangdong University of Finance Guangzhou 510521;
3. Institute of Finance and Accounting, Shanghai University of Finance and Economics Yangpu Shanghai 200433)

Abstract French is one of the working languages of the United Nations. Its complex grammar and part-of-speech rules result in the inability of term extraction methods such as N-gram and thus affect the accuracy of French text mining. This paper proposes an effective and efficient French term extraction method, which can be used to extract words and phrases from the analyzing French text corpora and provide a complete lexicon for French text mining. Firstly, word co-occurrence information of the corpora being analyzed is compressed into an FP (Frequent Pattern) sequence tree for extracting frequent word sequences rapidly, and then the termhood of each frequent word sequence is calculated to obtain the term set. The FP sequence tree is a newly-designed data structure for reducing the time complexity of word co-occurrence statistics to linear time. Experiments show that the proposed method has a high accuracy of approximate 90% with a much higher than normal recall rate and thus has good potentials for French text mining applications.

Key words FP sequence tree; French text mining; term extraction; termhood; text compression

法语是联合国工作语言之一，是欧盟、北约、世贸等众多国际组织的官方语言及正式行政语言，是全球 29 个国家的官方语言，是除英语之外最多国家使用的官方语言，其影响力仅次于英语[1-2]。法语的使用范围主要集中于欧洲、非洲、北美洲的一些国家和地区。随着“一带一路”的建设和全球化进程的加快，我国与欧洲、非洲国家的经济文化交流越来越广泛和深入，相关的新闻、政策文件、社交媒体文件等文本数据成为跨国组织管理决策的重要依据。因此，我国亟需有效的法语文本挖掘方法技术来实现海量法语文本高效的自动分析和及时的信息提取。

但目前，国内外针对法语文本挖掘方法的研究成果较少[3]。其中，法文词语提取是法语文本挖掘的基础和关键步骤[4]，是指自动获取法语文本中出现的所有词语的集合，包括法文单词原形和由多单词组成的短语。由于文本的关键词或特征词大多是短语而非单词，所以短语的完整提取是法文词语提取方法的关键。尽管法文词语提取方法已应用于法语文本信息检索、命名实体识别、情感分析等法语文

收稿日期：2020－06－30；修回日期：2020－08－27
基金项目：国家自然科学基金 (71771054)
作者简介：于娟 (1981－)，女，博士，副教授，主要从事数据挖掘、信息与知识管理等方面的研究. E-mail：yujuan@fzu.edu.cn

本挖掘任务[5-7]，但均为早期的 N-gram 词语提取方法[8-9] 或基于形容词与名词组合的方法[10-11]。这些方法受限于规则的不完备性，不能为文本建模提供完备的词库，影响法语文本挖掘的效果和效率。

另一方面，尽管中文和英文的短语提取方法研究已较为成熟[12-13]，但由于法文与中、英文在词法和语法方面有较大差异[14-15]，不能直接使用这些方法。例如，与中、英文相比，在词法方面，法文单词具有阴阳性的区别，动词、形容词、冠词需根据名词的阴阳性而变化；且不同语境的法文单词还有阴阳性的改变。在语法方面，法文中的定语需要根据具体语境搭配在名词前或名词后，搭配顺序不同则意思可能不同，如“un homme grand”意为“高大的人”，而“un grand homme”意为“伟大的人”。因此，法语文本的预处理和词语提取方法是法语所特有的，无法直接采用针对其他语言研发的方法。

上述原因导致法文词语提取成为当前制约法语文本挖掘准确性和高效性的瓶颈。因此，本文提出一种结合法语词法分析和单词共现统计规律的法文词语自动提取方法，并设计一种新的数据结构——FP 序列树，用于存储具有先后顺序的法文单词串，降低单词共现统计的时间复杂度。

1　法文词语提取方法框架

本文的法文词语提取方法综合考虑法语的词法分析和输入文本中单词共现的统计规律。该方法首先预处理输入的法语文本，接着将其压缩为 FP 序列树并基于该 FP 序列树提取无 N 元限制的频繁词串，依据不成词库筛选频繁词串，得到候选词语，然后计算候选词语的成词度，交由人工判别得到最终的法文词语集合。由于本文方法的高准确率和高召回率，未经人工判别的法文词语集合也可作为法语文本建模的词库。本文方法流程如图 1 所示：

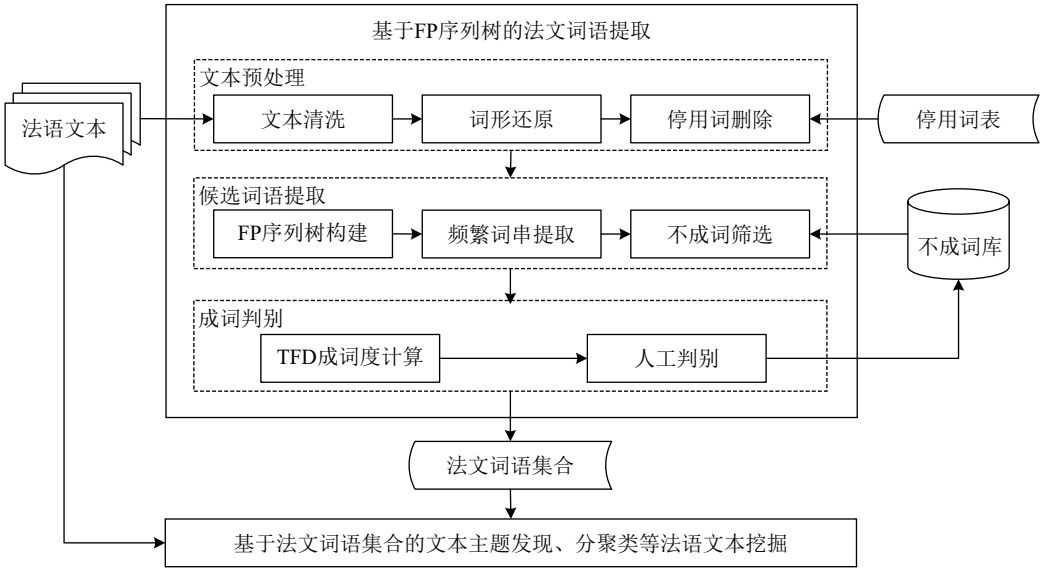

图 1　基于 FP 序列树的法文词语提取方法流程图

1) 文本预处理模块包含文本清洗、词形还原、停用词删除 3 个子模块。其中，文本清洗子模块删除输入文本中的图片、公式、标点符号和文本标记等，输出法语句子序列；词形还原子模块，采用现成的法文词形还原工具将法语句子序列中的每一个单词转换成单词的原形，常用工具有 Treetagger[16-17]、Spacy[18]、CST’s Lemmatiser[19] 等；停用词删除子模块，删除那些用来构成句子但不参与构词的单词，如系动词être(是)、代词 toi(你)、连词 si(如果) 等，输出一组法文单词串。

2) 候选词语提取模块包含 FP 序列树构建、频繁词串提取、不成词筛选 3 个子模块。其中，FP 序列树构建子模块将前一模块输出的一组法文单词串转存为树形结构，即 FP 序列树；频繁词串提取子模块，基于 FP 序列树，将频次超出阈值的词串输出为频繁词串；不成词筛选子模块，依据不成词库，删除不成词的频繁词串，输出候选词语集合。本文第 2 节详细介绍该模块的 FP 序列树构建与频繁词串提取方法。

3) 成词判别模块包含 TFD 成词度计算和人工判别两个子模块。其中，TFD 成词度计算子模块采用 TFD 算法逐一计算前一模块输出的候选词语

的成词度，将候选词语按成词度降序输出；人工判别子模块由法语专业人员对候选词语是否成词进行人工判别或轻微修改，得到最终输出的法文词语集合。经人工判别不成词的候选词语被加入到不成词库中，通过丰富不成词库，不断提高本文词语提取方法的准确率。本文第 3 节详述该模块的 TFD 成词度计算方法。

2 候选词语提取

在候选词语提取阶段，为了加速单词共现分析和频繁词串提取，本文设计并实现了一种新的数据结构——FP 序列树，在压缩文本数据集的同时不丢失单词在句子中出现的先后顺序信息。FP 序列树的设计受到了用于购物篮数据压缩的 FP 增长树[20]的启发。二者的主要区别在于：FP 序列树分支上的结点不是按其在数据集中出现总频次的大小排列，而是按其在句子中出现的先后顺序排列。

2.1 FP 序列树构建

构建 FP 序列树时，根结点不存放词语，除根结点以外的其他结点都存储一个单词及其所属单词串的出现频次。FP 序列树上的一个分支存储文本中出现的一个连续的法文单词串模式及其频次。重复出现的单词串及其子串不作为新分支，而是对已有分支上的每一结点的频次计数加 1。这样，仅需遍历一次待分析的法语文本，即可完成 FP 序列树的构建。

FP 序列树构建完成后即构建其相应的频次表。把频次计数相同的结点链接成一个链表，不同频次链表的头结点组成该 FP 序列树的频次表。

为了明晰起见，以一段法语文本为例，解释 FP 序列树及其频次表的构建过程。图 2 为一段法语文本及其中文翻译，不具有特殊性。图 3 为图 2 中的法语文本经过文本预处理的结果。本文采用目前法文词形还原效果最佳的 Treetagger 工具实现对图 2 文本的词形还原，然后采用基于大量实验总结出的停用词表删除其中的停用词。图 4 为压缩图 3 文本所构建的 FP 序列树及其频次表。

Le big data désigne des ensembles de données devenus si volumineux qu'ils dépassent l'intuition. Le volume des données stockées est en pleine expansion. Le volume des données stockées dans le monde fait plus que doubler tous les deux ans. La gestion des données et le traitement des données en entreprise sont des éléments clés de réussite. Le big data exige des méthodes automatiques de gestion des données et de traitements des données.	大数据是指庞大到超出人们经验和直觉的数据集。数据存储量不断增长，全球存储的数据量每两年翻一番以上。数据管理和数据处理是企业成功的关键因素。大数据需要自动的数据管理和数据处理方法。

图 2 法语文本与其中文翻译示例

[big data désigner du ensemble de
donnée devenir]
[volumineux]
[dépasser]
[intuition]
[volume du donnée stocker]
[plein expansion]
[volume du donnée stocker]
[monde faire plus]
[doubler tout]
[deux an]
[gestion du donnée]
[traitement du donnée]
[entreprise]
[élément clé de réussite]
[big data exiger du méthode automatique
de gestion du donnée]
[traitement du donnée]

图 3 图 2 法语文本的文本预处理结果

2.2 频繁词串提取

基于 FP 序列树提取频繁词串时，根据预先设定的频次阈值以及频次表中所存储的头结点指针，从 FP 序列树的每一分支中深度最大 (最接近叶子结点) 且满足阈值的结点开始，获取从该结点到根结点的分支上的单词串，倒序输出即为频繁词串。若某一分支有两个以上结点，还需同时获取其子分支所形成的频繁词串。

为了避免词语提取被截断，本文的提词方法采用长度优先，即出现频次与母串相同的子串不列入候选词语。也即，若模式$a_i \cdots a_j$的出现频次大于阈值且等于模式$a_i \cdots a_{j-1}$的出现频次，则认为$a_i \cdots a_{j-1}$是不能单独成词的。例如，若“États Unis(联合国)”和“États”的出现频次均为 10，则后者不列入候选词语。具体实现方法为：若 FP 序列树某一分支满足频次阈值且有多个频次相同的结点，则对每一频次仅输出较长子分支所形成的频繁词串。

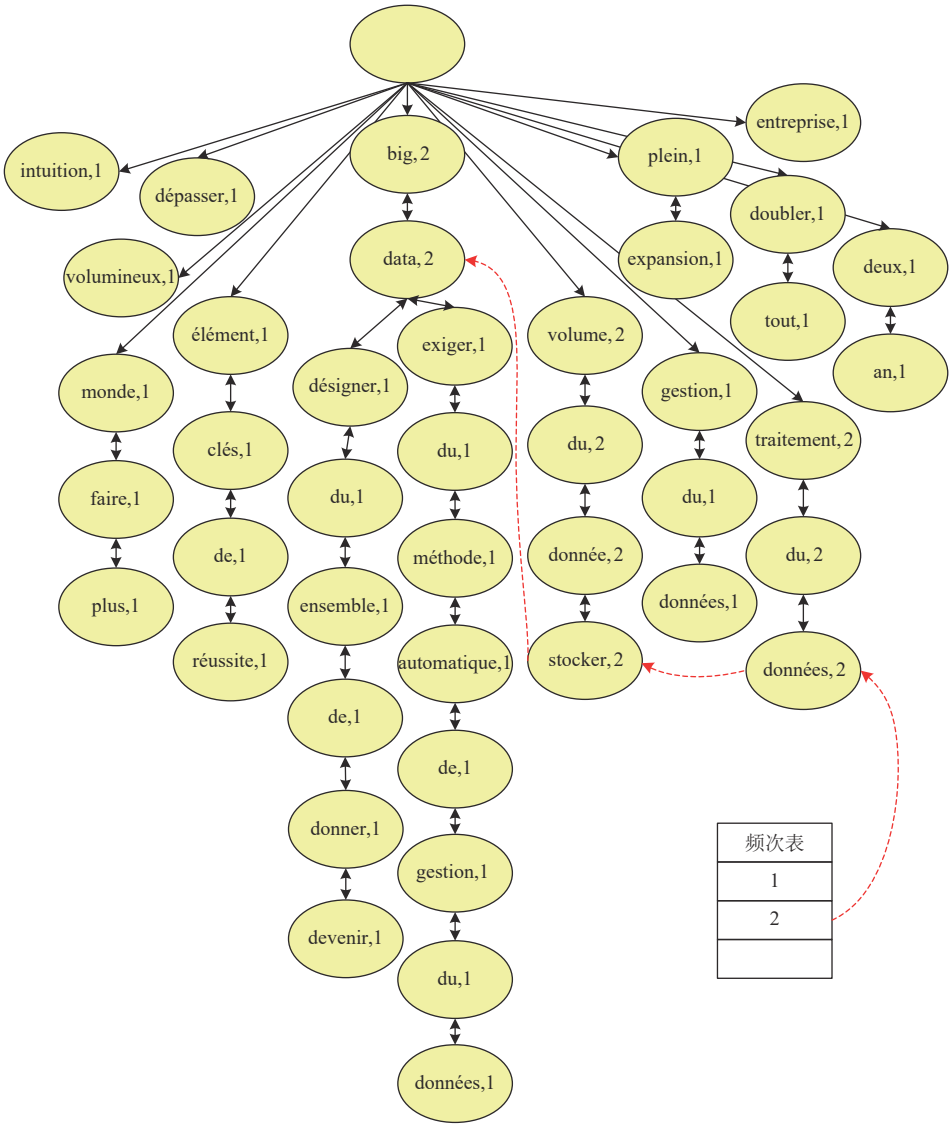

图 4　图 3 文本的 FP 序列树

表 1 是基于图 4 的 FP 序列树所提取得到的频繁词串和候选词语，此处设定频次阈值为 2。

表 1　图 4 的 FP 序列树的频繁词串提取结果

频繁词串	中文意思	频次
traitement du donnée	数据处理	2
volume du donnée stocker	数据存储量	2
big data	大数据	2

FP 序列树的时间复杂度是线性的。构建文本的 FP 序列树时，仅需遍历文本 1 次，即时间复杂度为 $O(n)$。基于 FP 序列树提取频繁词串时，根据频次阈值直接读取一个链表，时间复杂度为 $O(1)$。

3　TFD 成词度计算

文本数据集合中出现的候选词语 t 是否成词，与其出现频次 (term frequency, TF) 有关，也与 t 的分布有关[21]。并且，实际应用的文本数据集合中，词语在不同文本的出现频次往往变化较大。即，若 t 在整个文本集合中的分布 D(distribution) 不均匀，则 t 成词的可能性较高；反之，若 t 在整个文本集合中分布得较均匀，则成词的可能性较小。因此，受文献 [22-23] 的启发，本文依据词频和词分布两项因素计算候选词语的成词度为：

$$\mathrm{TFD} = \mathrm{TF}(t) \times D(t) = \sum \frac{\mathrm{TF}_i(t)}{\mathrm{DF}_i} \times \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{TF}_i(t) - \overline{\mathrm{TF}^*(t)}\right)^2} \tag{1}$$

式中，$\mathrm{TF}_i(t)$表示候选词语 t 在第 i 篇文本中出现的频次；DF_i表示第 i 篇文本中出现的候选词语的总频次；n 表示文本集合中有 t 出现的文本数；$\overline{\mathrm{TF}^*(t)}$

表示 t 在 n 篇文本中出现的平均频次。TFD 的值大，表示 t 在单篇文本内出现的频次高，且 t 在每个文本出现的频次波动大，则 t 成为词语的可能性大。TF(t) 越大，表示 t 出现的频次越高；$D(t)$ 越大，表示 t 在不同文本中出现频次的区别越大。

4 实验分析

目前还没有检验法文词语提取方法优劣的通用数据，也没有标准的评价指标。本文采用两组实验比较分析本文提出的词语提取方法与经典方法的性能。

4.1 数据介绍

采用两个题材不同的文本数据集进行实验分析：联合国平行语料库[24] 和 Europarl[25]，分别代表书面法语和口语法语。

联合国平行语料库是一个由联合国文件组成的平行文件档案库，本文采用其中 2014 年的 200 篇法语文件作为第一组实验数据，共 7.9 MB。

Europarl 是从欧洲议会议事录中收集的平行文本语料库，由欧洲议会讨论记录组成。本文选取其中法语文件的前 2 000 行作为第二组实验的数据，共 356 KB。

这两组实验数据均为随机选取，不具有特殊性。

4.2 评价指标

准确率和召回率是文本挖掘方法常用的评价指标。词语提取方法的准确率是候选词语中经人工判定成词的比率；召回率是指经人工判定成词的候选词语占文本中出现的全部词语的比率。但是，由于目前尚没有经过人工精确标注的法语文本词语提取语料库，无法确定语料中出现的全部词语数量，因此，本文采用正确提取词语的数目来代替召回率评价指标。

4.3 实验结果与分析

对两组实验文本数据分别进行文本预处理，使用 Treetagger 实现词形还原，删除停用词；接着将文本压缩为一棵 FP 序列树，提取频次超过 2 的词串作为频繁词串；然后直接将这些词串作为候选词语，不进行不成词筛选。由两名法语专业人员判别这些候选词语是否成词并互相检验。

本文的法文词语提取方法，从第一组实验数据提取得到 19 245 个候选词语，从第二组实验数据提取得到 1 713 个候选词语。法语专业人员人工判别的结果显示：第一组实验的自动提词结果准确率为 90.8%；第二组实验的自动提词结果准确率为 89.1%。

在对候选词语进行成词度计算时，分别采用本文的 TFD 方法与经典的 TF-IDF 方法计算频繁词串的成词度，将频繁词串按成词度降序输出。比较结果如图 5 和表 2 所示。

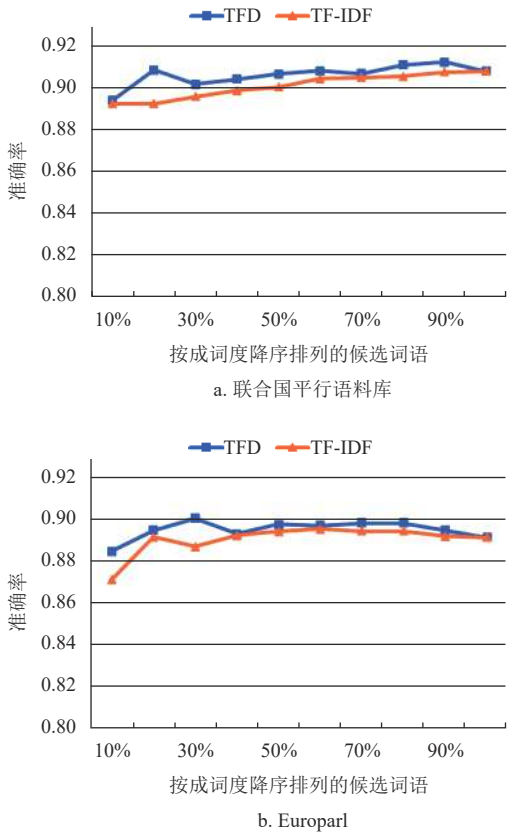

图 5 法文词语提取方法的准确率比较

表 2 法文词语提取方法正确提取的词语数目比较

词语数目	本文方法	二元词组法	词性规则法
联合国平行语料库	17 478	12 711	11 327
Europarl	1 526	1 443	1 028

图 5 展示 TFD 和 TF-IDF 两种方法在自动判别候选词语是否成词方面的准确率。其中，“按成词度降序排列的候选词语”的 m%(m=10, 20, …, 100) 是指成词度前 m% 的候选词语；准确率是指这些候选词语中经人工判定成词的比率。

表 2 为本文方法与法语文本挖掘中词语提取常用方法正确提取词语的数目，包括 N-gram 二元词组法和基于形容词与名词组合的词性规则法。

由实验可知：

1) 从图 5 可以看到，对书面法语和口语法语两种题材不同的语料进行词语提取时，在候选词语的成词度计算方面，TFD 的准确率均稳定优于 TF-IDF。这是由于 TFD 方法增加了词串在不同文本中分布

均匀程度的因素。

2) 从表 2 可见，本文的词语提取方法正确提取的词语数目明显高于常用的法语文本提词方法——二元词组法和词性规则法。这是因为：二元词组法限定了所提取词语的长度；词性规则法限定了所提取词语中每个单词的词性，只能提取得到形容词与名词的搭配。这些过滤规则能提高词语提取的准确率，但同时大幅降低了提取的词语数目。而本文方法通过总结停用词表来精准过滤频繁词串，在保证准确率的同时，显著提升了词语提取的召回率。

3) 语料的规模影响着本文的法文词语提取方法的召回率。图 2 的法语文本较短，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“gestion du données (词形还原前为 gestion des données，数据管理)”。因此，语料规模较小时，本文方法的召回率会降低。语料规模越大，本文的词语提取方法越具有优越性。

4) 由实验耗时可知，采用 FP 序列树存储待分析文本中的频繁词串及其出现频次，能够快速提取不同频次的频繁词串。尽管构建 FP 序列树需要花费一定时间，但能降低后续的频繁词串提取的时间复杂度，进而缩短词语提取的耗时。

此外，每进行一次法文词语提取，都应把人工判别不成词的频繁词串加入到不成词库。随着不成词库的丰富，本文方法的准确率会持续提升。这样就可以逐渐降低人工判别的工作量，提高词语提取的自动化程度和效果。

5　结束语

目前，关于法语文本挖掘的研究还在起步阶段。由于法语特殊的词法和语法规则与中、英文存在巨大差异，导致当前较为成熟的中、英文文本挖掘方法无法直接应用于法语文本挖掘。

为了支持基于法语信息的管理决策，本文提出了一种基于 FP 序列树的法文词语提取方法。该方法能够高效准确地从待分析的法语文本中自动获取包含法文单词原形和由多单词组成的法文短语的法文词语集合，为法语文本主题发现、分/聚类等文本挖掘任务提供词库。采用本文设计的 FP 序列树的数据结构压缩文本，能够快速提取文本中不同频次的频繁词串，将词语提取的时间复杂度降低到线性时间，从而提高文本自动分析的效率。同时，本文的法文词语提取方法在文本预处理阶段所使用的词形还原工具影响着最终结果的准确性。

参　考　文　献

[1] WIKIPEDIA. French language[EB/OL]. [2020-05-10]. http://en.wikipedia.org/wiki/French.

[2] UNITED NATIONS. Official languages[EB/OL]. [2020-06-27]. https://www.un.org/en/sections/about-un/official-languages/index.html.

[3] MARTIN L, MULLER B, SUAREZ P J O, et al. Camembert: A tasty French language model[EB/OL]. [2020-05-21]. https://arxiv.org/abs/1911.03894v1.

[4] HAN Jia-wei, WANG Chi, EL-KISHKY A. Bringing structure to text: Mining phrases, entities, topics, and hierarchies[C]//The 20th ACM Conference on Knowledge Discovery and Data Mining (KDD '14). New York, NY, USA: ACM, 2014: 1968.

[5] PRINCE V, LABADIE A. Text segmentation based on document understanding for information retrieval[C]// Proceedings of the 12th International Conference on Applications of Natural Language to Information Systems (NLDB'07). Berlin, Heidelberg: Springer-Verlag, 2007: 295-304.

[6] PAIS S, DIAS G, WEGRZYN-WOLSKA K, et al. Textual entailment by generality[J]. Procedia-Social and Behavioral Sciences, 2011, 27: 258-266.

[7] ABDAOUI A, NZALI M D T, AZE J, et al. ADVANSE: Sentiment, opinion and emotion analysis in French Tweets[C]//22nd Conference on Natural Language Processing. Caen, France: [s.n.], 2015: 78-87.

[8] BOUGOUIN A, BOUDIN F, DAILLE B. The impact of domains for keyphrase extraction[C]//21st Conference on Natural Language Processing. Marseille, France: [s.n.], 2014: 13-24.

[9] PANCKHURST R, LOPEZ C, ROCHE M. A French text-message corpus: 88milSMS. synthesis and usage[EB/OL]. [2020-05-10]. http://journals.openedition.org/corpus/4852.

[10] ALI C B, WANG R, HADDAD H. A two-level keyphrase extraction approach[C]//International Conference on Intelligent Text Processing & Computational Linguistics. [S.l.]: Springer, 2015: 390-401.

[11] LOSSIO-VENTURA J A, JONQUET C, ROCHE M, et al. Biomedical term extraction: Overview and a new methodology[J]. Information Retrieval, 2016, 19(1-2): 59-99.

[12] 于娟, 党延忠. 结合词性分析与串频统计的词语提取方法[J]. 系统工程理论与实践, 2010, 30(1): 105-111.
YU Juan, DANG Yan-zhong. Chinese term extraction based on POS analysis & string frequency[J]. Systems Engineering—Theory & Practice, 2010, 30(1): 105-111.

[13] HASAN K S, NG V. Automatic keyphrase extraction: A survey of the state of the art[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014). Baltimore, Maryland, USA: [s.n.], 2014: 1262-1273.

[14] 祁依虹, 茅于杭. 汉法机器翻译的难点分析[J]. 计算机工程, 2002, 28(9): 235-237.
QI Yi-hong, MAO Yu-hang. Analysis on difficulities of Chinese-French machine translation[J]. Computer Engineering, 2002, 28(9): 235-237.

[15] BROWN P F, PIETRA S D, PIETRA V J D, et al. The mathematics of statistical machine translation: Parameter estimation[J]. Computational Linguistics, 1993, 19(2): 263-311.

[16] SCHMID H. Treetagger[EB/OL]. [2020-05-10]. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/#Linux.

[17] SCHMID H. Probabilistic part-of-speech tagging using decision trees[C]//Proceedings of International Conference on New Methods in Language Processing. Manchester, UK: [s.n.], 1994, 12: 44-49.

[18] EXPLOSION. Spacy[EB/OL]. [2020-05-10]. https://spacy.io/models/.

[19] JONGEJAN B, DALIANIS H. Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike[C]//Proceedings of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Suntec, Singapore: [s.n.], 2009: 145-153.

[20] HAN Jia-wei, PEI Jian, YIN Yi-wen. Mining frequent patterns without candidate generation[C]//Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (SIGMOD'00). Dallas, Texas, USA: ACM, 2000: 1-12.

[21] FRANTZI K, ANANIADOU S, MIMA H. Automatic recognition of multi-word terms: The C-value/NC-value method[J]. International Journal on Digital Libraries, 2000, 3(2): 115-130.

[22] 于娟, 党延忠. 领域特征词的提取方法研究[J]. 情报学报, 2009, 28(3): 368-373.
YU Juan, DANG Yan-zhong. Domain feature and its extracting approach[J]. Journal of the China Society for Scientific and Technical Information, 2009, 28(3): 368-373.

[23] 周浪, 张亮, 冯冲, 等. 基于词频分布变化统计的术语抽取方法[J]. 计算机科学, 2009, 36(5): 177-180.
ZHOU Lang, ZHANG Liang, FENG Chong, et al. Terminology extraction based on statistical word frequency distribution variety[J]. Computer Science, 2009, 36(5): 177-180.

[24] ZIEMSKI M, JUNCZYS-DOWMUNT M, POULIQUEN B. The united nations parallel corpus[C]//Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Portorož, Slovenia: [s.n.], 2016: 3530-3534.

[25] KOEHN P. Europarl: A parallel corpus for statistical machine translation[EB/OL]. [2020-06-25]. http://www.statmt.org/europarl/.

编 辑 税 红