doi:10.3772/j.issn.2095-915x.2020.06.004

基于依存句法分析的科技政策领域主题词表无监督构建

开放科学（资源服务）标识码（OSID）

邵卫 化柏林

北京大学信息管理系 北京 100871

摘要：为了解决科技政策领域词表构建的问题，本文提出一种基于依存句法分析的科技政策文本关键词抽取算法。在此基础上，提出文本主题词指数来构建文本主题词，利用同义词识别算法及百科知识发现和确定词与词的同义关系，采用字面匹配的方法判别上下位词，最终汇合四个部分形成科技政策领域主题词表。为了适应缺乏标记的实际情况，使得文章更具有实际应用价值，本文使用了无监督方法。结果表明，此方法产生的词表具有显著的领域特征，可以解决领域未登录词切分，主题词之间关系缺乏等问题，有效地支持分词及文本分析。

关键词：科技政策；无监督构建；依存句法分析；主题词表；文本挖掘

中图分类号：G35

Unsupervised Construction of Thesaurus in the Science and Technology Policy Based on Dependency Syntax Analysis

SHAO Wei HUA Bolin

Department of Information Management, Peking University, Beijing 100871, China

Abstract: In order to solve the problem of vocabulary construction in the field of science and technology policy, this paper proposes a keyword extraction algorithm for science and technology policy texts based on dependency syntax analysis. On this basis, the text topic index is proposed to construct the text topic words; using the synonym recognition algorithm and encyclopedia knowledge to discover and determine the synonymous relationship between words and words; utilizing the word matching method to discriminate the upper and lower words; converging four parts to form a thesaurus of science and technology policy. To adapt the real situation that labeled data is always lacked and improve the application value of

作者简介：邵卫（1999-），研究方向为文本挖掘；化柏林（1977-），助理教授，硕士生导师，研究方向为大数据情报分析，文本挖掘，E-mail：huabolin@pku.edu.cn。

doi:10.3772/j.issn.2095-915x.2020.06.004

this paper, all methods proposed by us belong to unsupervised methods. The results show that the vocabulary generated by this construction method has significant domain characteristics and can effectively support word segmentation and text analysis.

Keywords: Science and technology policy; unsupervised construction; dependency syntax analysis; thesaurus; text mining

引言

在政策文本分析过程中，通用的中文分词器对科技政策领域的文本适应性较差，可能导致后续分析的不准确。因为科技政策文本中经常性会出现一些特殊的名词术语，如“京津冀一体化”，“一带一路”，由于分词器对新词、特殊词等未登录词的敏感度很低[1]，所以很难将这些词完整的切分出来，这样就会造成关键信息的丢失。一种比较好的实践方式是借助科技政策有关的论文的关键词进行切分（将其导入分词器中），但是对于知名度较低的具有地方特色的术语，如安徽省的“合芜蚌示范区建设”，可能很少有论文会提及。而一个领域主题词表（又称叙词表）可以提供领域词汇信息，将其导入分词器中可以有效解决这个问题[2]。另一方面，利用词汇的类别信息、同义关系以及上下位关系会提升科技政策文本分析的准确性和内容丰富性。而主题词表包括词汇、词汇分类、词间关系，可以为科技政策文本的分析提供支持，有效地提升科技政策文本分析的效果，所以构造出主题词表支撑政策文本分析是很有必要的。

本文以省一级科技主管部门发布的科技政策文本为实验对象，提出了科技政策领域词表构建系列方法，结果表明该系列方法能够有效地构造出可用性较强的词表，并与其他相关方法进行了比较，说明本文方法的效果。

1 相关研究

领域主题词表的相关研究根据对象可分为领域主题词表的构建进展，构建方法，构建实践三个大类，其中构建方法又涉及到关键词、主题词提取以及上下位等级、相关等词间关系的判断。

1.1 领域主题词表的构建进展

近年来，面向细分领域的主题词表自动构建受到关注。安亚巍等人[3]提出基于共现特征矩阵的面向特定领域大规模语料的主题词表构建方法，其结果具有较高的准确率。张昱等人[4]利用了知识组织的方法对档案领域主题词表进行构建。孙立媛[5]等人面向高校应急决策领域提出了主题词表的构建流程。

当下，主题词表的自动构建也受到互联网发展的影响。一方面网络环境下主题词表的构建备受关注，另一方面，搜索引擎等网络工具的发展对主题词表的地位造成了挑战。Birger Hjørland[6,7]对传统主题词表在现代网络环境下信息检索中的地位进行了讨论，指出深入到学科术语领域的主题词表依然是非常重要的。但

是，他也通过数据统计的方式证明近些年领域主题词表的进展十分缓慢。

1.2 关键词识别与提取研究

对于关键词识别问题，仲云云等人[8]提出基于 N-gram 和 GL/GF 权重的识别方法，能够简单快速地识别出关键词，但是精度仍有较大提升空间。之后杜慧平等人[9]使用领域词典以及期刊文献的关键词进行筛选。张越等人[2]利用词性特征抽取出候选词，然后利用 TF-IDF、政策规范化距离、TextRank 三种指标混合进行关键词的筛选，具有较高的准确率。

对于主题词提取问题，曾文等人[10]提出基于 MI（互信息）和 TF-IDF 的识别模型，考虑了词语的互信息、位置权重、TF-IDF 等信息，能够处理不同的文本数据结构和内容。安亚巍等人[1]提出基于改进的 TF-IDF 选取候选词，然后利用词共现特征构建特征矩阵，并使用连通子图的方法将词聚成簇，从簇中选取语义贡献度最大的词作为中心主题词，具有较高的准确率和召回率。

1.3 词间关系识别研究

对于上下位关系识别问题，主要存在字面成族[8, 11]，相关度聚类[12]等方法。字面成族的方法利用具有相同词素的词之间经常存在意义上的联系的现象，根据后方一致性原则进行上下位类的判定，这种方法主要针对具有字面相似的词。相关度聚类的方法首先根据词频对不同的词划分等级，然后求上下等级词之间的关联度，并以此对相邻等级的词建立等级关系。张巍[13]等人利用概念词汇知识，上下位关系的统计和词性规则特征，提出一个多特征融合的概念关系识别模型，对同义关系和上下位关系进行识别。

对于同义关系识别问题，有同义词词典，模式匹配，字面相似度等方法[9, 10]，这些方法都是基于字符匹配的方式，其中同义词词典具有较高的准确度，但是许多领域并没有专门的同义词词典，所以使用范围会受到限制。模式匹配的方式则通过提前选定一些能够体现同义关系的连接词，如“俗称，亦称”等，然后选择被这些词连接的词作为同义词，这种方法准确度较高，但是召回率往往比较低。字面相似度的方法则是利用两个词的词素重叠程度表示两个词的相似度，然后确定一个阈值，选择相似度高于这个阈值的两个词作为同义词，这种方法可以提高召回率，但是精确度往往不足而且无法处理那些不满足字面相似度的同义词。

在机器学习特别是深度学习盛行的时代，词间关系识别被统一为标注问题，徐健[14]等人总结了深度学习兴起之前的实体关系识别的方法，将其概括为基于模式匹配的关系抽取、基于词典驱动的关系抽取、基于传统机器学习（如支持向量机 SVM 和条件随机场 CRF）的关系抽取、基于 Ontology 的关系抽取以及混合抽取方法，为进一步构建实体关系抽取系统提供良好借鉴。随着深度学习的兴起，实体关系识别也获得了新发展，鄂海红[15]等人围绕有监督和远程监督两个领域，描述了卷积神经网络 CNN，循环神经网络 RNN，深度强化学习，生成对抗网络 GAN 及其相应变体在实体关系识别上的应用，并与传统方法进行了比较。目前，最好的模型是基于双向 LSTM 融合 CRF 的方法。

doi:10.3772/j.issn.2095-915x.2020.06.004

1.4 领域主题词表构建实践

除了上述以词语为核心对象的研究，还有一些更加注重领域实践与应用，构建某些领域的主题词表，在构建的过程中综合使用现成资源以及多种方法。

电子政务主题词表方面，郑新燕等人[16]以国土资源和房地产领域为例，描述了专题电子政务主题词表的构建体系。仲云云等人[8]综合使用 N-gram，字面相似度，模式匹配，同义词字典等方法构建了电子政务主题词表。他们在电子政务主题词表构建的过程中除了提出了新的词语识别、抽取、筛选方法外，也注重改进已有相关方法使之适应当前领域。

汉语科技词方面，中信所建设集成了一套成熟的汉语科技词系统[17, 18]，并应用于新能源汽车等领域[19]，证明了领域科技词系统的可行性，并形成了知识架构处理、知识加工、内容审核、系统发布的科技词系统集成范式。同时，汉语科技词系统的初步成功也促进了新能源汽车、生物技术、材料等领域的研究。

1.5 研究述评

综上所述，为了无监督地构建各种类型的主题词表，现有研究主要是通过词频、共现等方式识别出关键词，通过共现聚类、字面相似匹配等方法识别等级关系。然而，现有研究抽取得到的关键词主要是来自分词器分词后的识别结果，领域里一些重要的未登录词难以得到准确的切分，分词结果的领域指示度仍然不足。同样词间关系也没有考虑词对领域的指示度。同时在科技政策领域的词表构建研究较少。为解决上述问题，考虑到依存句法分析可以得到词与词之间的句法关系，可以帮助组合细粒度的词形成领域指示度更高的短语，本文将利用依存句法分析来组合分词结果，得到具有较高领域指示度的关键词，并在词间关系识别中引入词长、词领域指示度等指标来改进原有识别方法，从而构建出科技政策领域词表。

2 数据搜集及研究方法

先利用爬虫从政府科技政策网站上爬取政策文本，经过预处理后利用依存句法分析方法提取关键词，导入百科知识库并进行上下位词、主题词、同义词的抽取，得到上下位词、主题词表与同义词表。研究设计流程框架如图 1 所示。

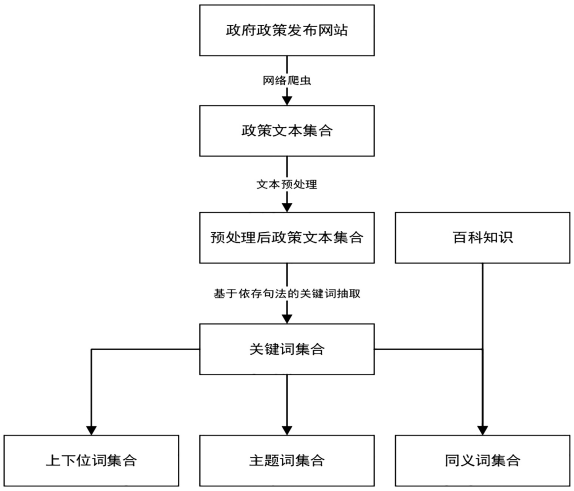

图 1 科技政策主题词表无监督构建流程

2.1 数据采集及实现工具

本文研究数据为北京、江苏、天津等 27 个省及直辖市的政策文本。获取方式是利用 python 爬虫程序，从各省直辖市科技厅（科委）官网下的科技政策栏目下采集了全部的政策文本数据（共 2620 条，爬取完成时间：2019 年 9 月）。

本文中使用的方法的具体实现基于 python3.6 及哈工大 ltp 的 python 接口 [20]。

2.2 关键词无监督抽取方法

经过调研发现，虽然现有的机器学习（如 CRF）或深度学习（如 BiLSTM+CRF）方法能够在测试集上取得很好的效果，但是这些方法都依赖于大量的训练数据，不适合本文面临的无监督抽取问题，即在没有训练语料的情况下从无标签语料中抽取关键词。目前已有的关键词无监督抽取算法如（TextRank，LDA）主要是对分词结果进行筛选的方法，得到的关键词对领域的指示程度不明显，也有通过对分词结果根据词性进行组合（如“名词”+“动词”）的方法来抽取较长的关键词，这种方式能够得到具有较高领域指示度的词，但是缺乏对词间关系的考虑，会产生较多无意义的组合，为了解决以上问题，本文采用基于依存句法分析的方法，通过词间的依存关系来抽取具有强领域指示度的关键词。

依存句法通过分析语言单位内成分之间的依存关系揭示其句法结构，可以发现句子中词之间的依存关系，如动宾、主谓、状语修饰等。关键词的词性一般以名词，动词及名词性动词为主。而一般主语、宾语的词性以名词为主，兼有名词性动词。而谓语以动词为主。故方法首先抽取出具有主谓关系（SBV）、动宾关系（VOB）的词对作为备选词，并将其中的谓语动词识别为关键词。同时为了增加关键词的领域指示力，对词对中的主语词、宾语词的所有依存关系进行遍历，寻找其修饰词（与主语词、宾语词的依存关系为定中关系（ATT）、状中关系（ADV）的词）。之后将其与修饰词按照修饰位置的顺序拼接起来，形成一个具有强领域指示能力的关键词。如：“非公有制”+“经济”=“非公有制经济”。具体算法流程如图 2 所示。

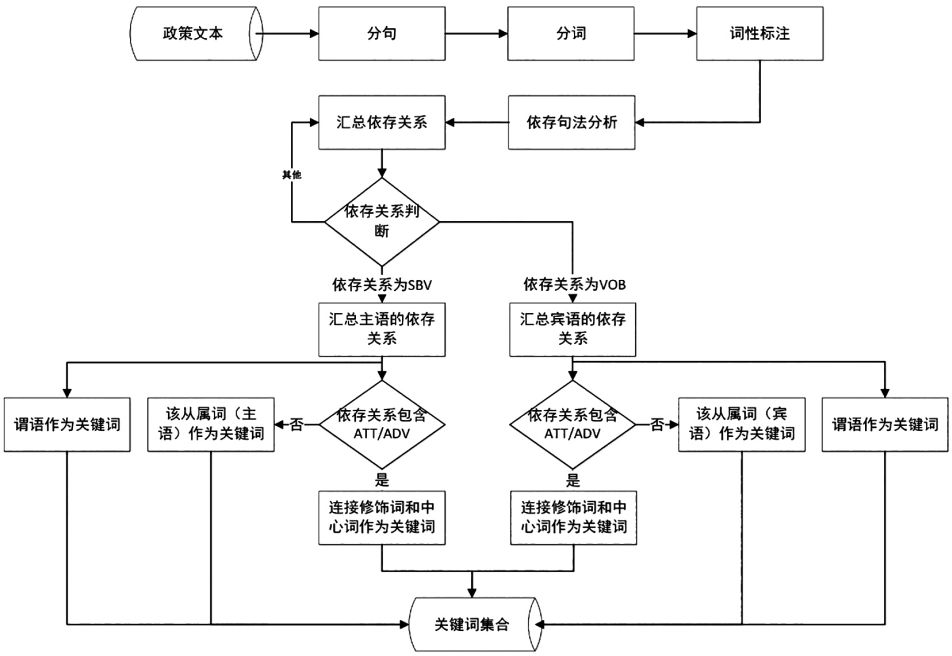

图 2　关键词提取算法流程图

doi:10.3772/j.issn.2095-915x.2020.06.004

首先对每个政策文本进行分句、分词、词性标注、依存句法分析，之后汇总依存关系并遍历，寻找是否存在 SBV 或 VOB。如果一个句子中不存在 SBV 或 VOB 的依存关系，则处理下一个句子；如果存在 SBV，则抽取谓语（V）作为关键词，找到 SBV 中的主语词（S）的依存关系字典，然后遍历，寻找是否存在 ATT 或 ADV，如果存在则连接修饰词和被修饰词组成一个长关键词，反之则抽取主语（S）作为关键词。VOB 处理过程与之类似，不再赘述。

2.3 文本主题词无监督抽取方法（连贯性）

文本主题词表是指所有政策文本的主题词形成的词表，用于确定单个政策文本的主题以及对抽取的关键词汇进行分类，需要与题目中的主题词表作区分。为了选择出具有领域指示度的、突显出一个政策文本主题的词，在词频的基础上考虑词长这个因素。一般越长的词，其领域指示度越高，这样就可以筛选出能表示较细分领域的主题词。具体做法是：每个政策文本的主题词的构造以从该文本中抽取出的关键词为基础，综合词的长度以及相对词频，设计出一个主题词指数。对于一个政策文本，选取主题词指数最大的若干个词作为主题词。主题词指数公式如下：

$$h=(f/t)/l \quad (1)$$

其中，h 表示当前词的主题词指数，f 表示当前词在该政策文本中的词频，t 表示与当前词长度相同的所有词的词频和，l 表示当前词的长度。

该主题指数考虑了不同长度的词的词频分布不同（较短的词的整体词频较高）以及不同长度词的领域指代粒度不同（越长的词对领域指代粒度越细）等因素，弥补了长词词频相对较小的缺陷，使得长词被选择的概率更大，增加了主题词的表示能力。具体算法流程如图 3 所示。

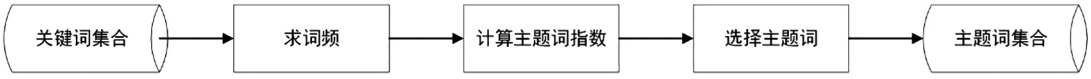

图 3 主题词筛选算法流程图

首先，对于一个政策文本的关键词集合，先求出各个词的词频，以及各个词长的总词频（即不同长度的词的词频），然后对该政策文本中的每一个关键词，根据公式求出其主题词指数，最后，根据主题词指数对关键词进行排序，选择出主题词指数最大的若干个词（本文选择的是 10）作为该政策文本的主题词。

2.4 同义词无监督抽取方法

同义词的无监督构造方法一般基于词共现、字面相似度等方法。多数利用字符串匹配的方式没有考虑词长对匹配的影响，而本文得到的关键词词长对语义范围和内容影响很大，由于对具有特定依存关系的词进行连接，最终得到的词的词义可能会在原有词义的基础上发生变化。为此，本文提出了一个基于词长的阈值公式，帮助提升利用字符串匹配筛选同义词的效果。同时研究中发现，这些方法对于那些不具有字面相似特征的同义词并不适用，引入外部知识有助于同义词构造，本文借助了百度百科的知

识来帮助构造同义词。

在字符匹配方法中，如果短词和长词首尾相同，且长词包含短词，那么它们就有可能是同义词（如“中国科学院”与“中科院”）。但当长词和短词长度差距过大时，这种方式就存在问题。因为短词与长词的语义相似度会随着长词长度增加而减小，如“中国社会科学院”与“中科院”；另一方面由于长词的语义指代粒度很细，部分变动很可能会导致语义变化，如“中国科学院”与“中国社会科学院”。所以在判断两个词是否是同义词时，需要同时考虑这个两个因素。对此，本文提出的方法是：当短词被长词包含时，求出短词与长词的词长比，然后限定一个阈值，当词长比超过这个阈值时，认为这两个词是同义词。很明显，词的长度会影响这个阈值，在尝试构造阈值公式时，发现如下阈值公式效果较好：

$$\frac{1}{\sqrt{1+\sqrt{\frac{1}{L}}}} \qquad (2)$$

其中，L 为长词词长。

根据公式，长词长越长则要求短词覆盖的长度越长。而且长词和短词覆盖率不能为100%，分母上的 1 保证了这个限制。此外，根号保证了词长过长的长词产生阈值不会过长，避免了匹配率过低的问题。

前面提到的同义词抽取方法能够发现关键词之间的同义关系，但是该方法是基于字符匹配的方法，会在一定程度上丢失语义以及背景信息，如果想继续发掘已有关键词的同义词，如“国家科委”和“中华人民共和国国家科学家技术委员会”，就需要借助外部知识，本文借助了百度百科词条的同义表述发现关键词的同义词。

2.5 上下位类词无监督抽取方法

上下位类的无监督抽取则基于传统的字面相似度匹配的方式，这是在现有无监督条件及无辅助性语料库条件下较为有效便捷的方法。具体的措施为：对于一个词对，一个为长词，一个为短词，如果短词被长词包含，而且短词和长词末尾存在相同的字符串则认为短词是长词的上位类。这种方法的假设为长词的词义一般比短词的词义窄，而且大多数上下位词满足后方一致性原则。实验时采用末尾长度为 3 的字符串相同则认为两个词之间存在等级关系。如“示范点”——“服务示范点”，“所得税”——“个人所得税”，“企业所得税”。但是很明显，这种方法无法处理那些不具备后方一致性的上下位词，如“XXX 大学”和“XXX 大学 XXX 系”，这种上下位词可以通过正则匹配的方式识别。

遍历得到的关键词，对词对利用上面的方法进行判断，选择符合条件的词对，最终得到上下位类词表。

3 研究结果

3.1 词表体系说明

利用基于依存句法分析的构建方法，共得到关键词表、文本主题词表、同义词表、上下位词表四张词表。关键词表是所有政策文本中的关键词按先后顺序排列而成的词表。文本主题词表是所有政策文本的主题词按照流水编码而成的词表。同义词表按照“原始词 - 同义词”

doi:10.3772/j.issn.2095-915x.2020.06.004

形式进行流水编码而成。上下位词表是以上位词后跟下位词列表为一个记录的形式，将得到的结果按照流水编码的方式组织而成。

3.2 部分结果展示

最终从 2620 个科技政策文本中得到 56006 个关键词，3812 个文本主题词，利用字符匹配方式得到 3947 对同义词，基于百度百科的方法得到了 1275 对同义词，2784 个词的下位类。下面分别对结果进行分析说明。

3.2.1 关键词结果分析

对关键词抽取结果进行初步统计，共得到词 56006 个，其中 3 字词 8291 个，比例为 14.81%。4 字词 36830 个，比例为 65.76%。5 字词 8381 个，比例为 14.96%。6 字词 2031 个，比例为 3.63%。7 字词 362 个，比例为 0.65%。剩余的包括 8，9，10，11，12，20，23 字词，其总个数为 111，比例为 0.2%。

对抽取结果进行人工检查发现，抽取得到的 3 字词中一部分不能单独成词（如“补措施”、“要处理”、“人以上”等），需要从抽取结果排除出去。经过观察发现，这些词绝大部分都可以切分成两部分，如“在展开”可以被切分为“在”和“展开”，“对承担”可以被切分为“对”，“承担”等。其中 1 字词大部分都可以被归为停用词。所以考虑对得到的 3 字词使用结巴分词器进行切分，然后检查是否有停用词，以此筛选出合适的 3 字词。对其他长度的词检查发现，这些词基本不会出现该问题。而 8 字以上的词将多个小词进行拼接，多个语义融合在一起，成词性大大降低，不宜将其作为词，故舍去。最终得到的关键词部分结果如表 1 所示（每一行为一个政策文本中的部分关键词）。

表 1 关键词部分列表

	Word1	Word2	Word3	Word4	Word5	Word6	Word7
Text1	股东权利	整体布局	国际评估	试点范围	主要文件	规范词	省级机关
Text2	能源供需	转化行动	众筹试点	银行信贷	备案数据库	市场主体	开源社区
Text3	岗位分红权	金额年增长率	国际化程度	标准化示范区	行政区域国民经济	非在岗人员	高等学校实验室
Text4	创新文化	孵化载体	云平台	运营服务商	信用体系	城乡一体化	企业光纤

3.2.2 文本主题词结果分析

从 2620 个政策文本中得到 3812 个文本主题词，根据算法阈值选择，每个文本产生 5 个主题词。经过人工判定，本文产生的结果对于文章主题的揭示性较为良好，选择出的主题词能够体现文章主题，但是存在一些合适的主题词被遗漏的情况，这个问题和固定每个文本的主题词个数有关，后续研究可以考虑针对文本的类型，特征，长度进行动态的主题词个数选取。部分结果如表 2 所示（每一行为一个政策文本的部分主题词）。

表 2　文本主题词部分列表

	Word1	Word2	Word3	Word4	Word5
Text1	天使投资人	企业家精神	大学生	简政放权力度	决定性作用
Text2	科学技术研究	企业化转制方向	高新技术企业	重大技术装备	个人所得税制改革
Text3	纪检监察部门	评审组织者	奖励委员会	监督委员会	科技奖
Text4	专家库登记表	社会经济效益	弄虚作假记录	高级专家库	信息化建设

3.2.3　同义词结果分析

从 2620 个政策文本中，利用字符匹配方式得到 3947 对同义词，基于百度百科的方法得到了 1275 对同义词。字符匹配的方式得到的同义词主要是字面近似的同义词，且多为较短词，全部来自文本语料中，准确率比较高。基于百度百科的方法得到的同义词准确率很高，集中在语义相近的词对，特别是长名词缩写。这种方法得到的同义词包括文本中未曾出现的词，且外来词主要是长词。部分结果如表 3 所示（每一行的两个词为同义词，前面四个为方法一构造出的同义词对，后面四个为方法二得到的同义词对）。

3.2.4　上下位词结果分析

从 2620 个政策文本中得到了 2784 个词的下位类词，每个词的下位类词个数从 1 到 70 个不等，以 1 到 10 个下位词居多。对得到的结果人工校验，其准确率比较高，但是同时也受三字词成词性低的影响，会有一些无意义的词。可以通过上面提到的对三字词的优化来缓解。另一方面，这些词大部分是字符相关的，而那些字符不相关的上下位词则很难被识别。后续可以考虑融合词汇语义进行上下位词的判定。部分结果如表 4 所示。

表 3　同义词部分列表

	Word1	Word2
1	国际人才	国际化人才
2	开放平台	开放式平台
3	实质进展	实质性进展
4	脱贫带头人	脱贫致富带头人
5	国家科委	中华人民共和国国家科学技术委员会
6	进出口交易会	中国进出口商品交易会
7	博士后流动站	博士后科研流动站
8	国家级高新区	中国高新技术产业开发区

表 4　上下位词部分列表

	上位词	下位词1	下位词2	下位词3	下位词4
1	一体化	种养一体化	生产一体化	城乡一体化	
2	出资人	基金出资人	境外出资人	社会出资人	其他出资人
3	领导人员	人财物领导人员	副职领导人员	党员领导人员	岗位领导人员
4	自主权	法人自主权	分配自主权	机构自主权	学术自主权
5	合格证	卫生合格证	防疫合格证	质量合格证	设施合格证
6	示范区	改革示范区	创新示范区	安全示范区	农业示范区

doi:10.3772/j.issn.2095-915x.2020.06.004

4 对比研究

为了体现提出的关键词无监督抽取方法的效果，使用 TF-IDF，TextRank 和 LDA 等无监督关键词抽取方法对相同的文本抽取关键词，与提出的方法进行对比。同时，为了探究依存句法分析结果对关键词抽取的影响，随机丢弃若干分析结果，如某些依存关系，然后利用经过此种处理的依存句法分析结果进行关键词抽取。此外，也比较了已有的同义词判断方法与提出的方法的结果。

4.1 关键词对比

针对同一文本，不同方法的得到的关键词结果可以见表 5，可以看出，这四种方法反映的主题大致相同，但是在具体的关键词上存在差异，TF-IDF 很难选择出较长的词作为关键词，而 TextRank、LDA 则可以抽取出部分较长的词，使得关键词的信息稍微丰富一些。而本文提出的方法抽取得到的关键词皆为含义比较具体的长词，能够体现出更多的可用于分析的信息。

表 5 关键词对比

方法	关键词结果						
本文方法	党纪律	民主权利	教育党员	遵纪守法	党中央权威	党内法规	处分工作
TF-IDF	处分	给予	规定	警告	党员	情节	组织
TextRank	处分	给予	警告	留党察看	职务	党内	开除党籍
LDA	章程	贯彻执行	纪检	权力	失察	国家机关	审批

4.2 依存句法分析结果对抽取结果影响的比较

针对同一文本去除不同依存关系后再进行关键词抽取，得到的结果见表 6，“/”表示原始的词从该位置截断的结果即为这种依存分析结果的对于该词的抽取结果。从表中可以发现，如果依存句法分析的结果无法捕捉到动宾关系（VOB），类似于“实施就业政策”的词组会被截断为“实施”、“就业政策”。如果无法捕捉到主谓关系（SBV），类似于“我省提出”的词会被截断为“我省”、“提出”。去除其他依存关系的结果见表 6。此外，根据提出的算法，如果是不涉及这些依存关系的词，则不受影响。

表 6 依存句法分析结果对抽取结果影响的比较

处理方式	关键词结果				
原始	实施就业政策	健全评价机制	带动重大项目	国务院	我省提出
去除”VOB”	实施/就业政策	健全/评价机制	带动/重大项目	国务院	我省提出
去除”SBV”	实施就业政策	健全评价机制	带动重大项目	国务院	我省/提出
去除”ATT”	实施/就业/政策	健全/评价/机制	带动/重大/项目	国务院	我省提出
去除”ADV”	实施就业政策	健全评价机制	带动重大项目	国务院	我省提出

4.3 同义词比较

我们选择传统字面匹配法与本文方法进行比较。针对同一文本，得到的部分结果如表 7 所示。本文方法与传统字面匹配方法的区别在于，在使用首尾匹配的同时，考虑短词对长词的覆盖以及长词的部分组成词变化导致语义变动等因素，这样可以有效防止由于词长过长中间语义变化过大引发的词义变更过大问题。我们可以看出，针对“中科院”一词的同义词，本文的方法选择出的同义词是“中国科学院”，但是传统字面匹配方式选择出了“中国社会科学院”，没有注意到“社会”对整个词义的影响。同样的，还有“分布式架构”与“分布式存储架构”，这两个词的词义也差别较大。

表 7 同义词比较结果

方法	同义词对						
本文方法	行政组织 , 行政部门组织	实质内容，实质性内容	国家实验区 , 国家级实验区	中科院 , 中国科学院	操作技术 , 操作系统技术	新农民 , 新型农民	分布式架构 , 分布式云架构
传统字面匹配	行政组织，行政人员及各组织	实质内容，实质性内容	国家试验区，国家自贸区	中科院，中国社会科学院	操作技术，操作系统技术	新农民，新型农民	分布式架构，分布式存储架构

5 结束语

本文基于依存句法分析，利用关键词的词性特征以及词间的修饰关系对细小词汇进行拼接，从而得到细粒度的关键词，这种关键词更能反映具体领域。在通过这种无监督方式抽取出关键词的基础上，提出主题词指数筛选法、同义词判断算法、百科知识引入等方法进行主题词表的无监督方式的构造，并与已有的相关方法进行了对比研究。

经过人工测评，使用本文方法得到的词表能够帮助解决新领域分词的问题，满足分析的需要，但同时这些方法的结果也存在着一些可以改进的地方，如抽取算法受依存句法分析效果的影响，不同的依存句法解析器可能会产生不同的结果，而且政策文本中的指代关系比较难以捕捉以及单句长度过长，都会使得依存句法分析的效果降低，使得无法发现潜在的语义关系，最终影响关键词的抽取。另一方面抽取得到词中依然有部分词不具有成词性，如“次担保服务”，“创新为原则”，“在影响力”等，后续需要提出新的方法加以筛选或改进抽取算法。同时同义词和上下位词的选择方法依然有很大的提升空间，后续可以考虑通过融合语义信息的方式选择出无法依靠字符特征来选择的同义词和上下位词。

由于缺乏标准的测试集，所以本文的测评主要采用人工判断，这种判断方式带有一定的主观性。但是由于文本直接面向大规模语料，使用无监督的方式去获得关键词，缺乏用于测评的标注数据，同时拼接的方式使得词语语义发生重组，难以实现对关键词抽取乃至最终的主题词表进行测评，故只能采用人工审核的方式进行主观判断，后续工作会对该类基于语义组合的主题词表构建方法的测评方式进行研究。

doi:10.3772/j.issn.2095-915x.2020.06.004

参考文献

[1] 唐琳 , 郭崇慧 , 陈静锋 . 中文分词技术研究综述 [J]. 数据分析与知识发现 , 2020, 4(Z1):1-17.

[2] 张越 , 刘琦岩 , 张玄玄 , 等 . 科技成果转化政策文本中的领域关键词汇提取研究 [J]. 中国科技资源导刊 , 2018, 50(3):68-75.

[3] 安亚巍 , 操晓春 , 罗顺 . 面向语料的领域主题词表构建算法 [J]. 计算机科学 , 2018, 45(S1):396-397+410.

[4] 张昱 , 于薇 . 档案领域词表自动化辅助构建及知识组织应用探析 [J]. 数字图书馆论坛 , 2018(6):67-72.

[5] 孙立媛 , 苏新宁 . 面向高校应急决策的领域主题词表构建研究 [J]. 情报科学 , 2019, 37(4):137-143.

[6] Birger Hjørland. Annual Progress in Knowledge Organization (KO)? Annual Progress in Thesaurus Research?[J]. Knowledge Organization, 2019, 46(3):238-239.

[7] Birger Hjørland. Does the Traditional Thesaurus Have a Place in Modern Information Retrieval?[J]. Knowledge Organization, 2015, 43(3):145-159.

[8] 仲云云 , 侯汉清 , 杜慧平 . 电子政务主题词表自动构建研究 [J]. 中国图书馆学报 ,2008(3):97-102.

[9] 杜慧平 , 侯汉清 . 网络环境中汉语叙词表的自动构建研究 [J]. 情报学报 , 2008, 27(6):863-869.

[10] 曾文 . 网络化数字化时代主题词表自动构建技术的探索与实践 [J]. 国家图书馆学刊 , 2012, 21(4):78-82.

[11] 张琪玉 . 字面相似聚类法辅助构造词族表、分面类表和自动标引 [J]. 图书馆论坛 , 2002, 22(5):95-96.

[12] 杜慧平 , 何琳 , 侯汉清 . 基于聚类分析的自然语言叙词表的自动构建 [J]. 国家图书馆学刊 , 2007(3):44-49.

[13] 张巍 , 于洋 , 游宏梁 . 面向词汇知识库自动构建的概念术语关系识别 [J]. 现代图书情报技术 , 2009(11):10-16.

[14] 徐健 , 张智雄 , 吴振新 . 实体关系抽取的技术方法综述 [J]. 现代图书情报技术 , 2008(8):18-23.

[15] 鄂海红 , 张文静 , 肖思琪 , 等 . 深度学习实体关系抽取研究综述 [J]. 软件学报 ,2019, 30(6):1793-1818.

[16] 郑新燕 , 李霖 . 专题电子政务主题词表体系 [J]. 科技创新导报 , 2007(33):143+145.

[17] 史新 , 乔晓东 , 张志平 , 等 . 汉语科技词系统的 Web 服务研究与实现 [J]. 现代图书情报技术 , 2008(12):37-42.

[18] 乔晓东 , 张运良 , 朱礼军 . 汉语科技词系统建设与应用进展 [J]. 情报学报 , 2010, 29(6):978-986.

[19] 朱礼军 , 乔晓东 , 张运良 . 汉语科技词系统建设实践——以新能源汽车领域为例 [J]. 情报学报 , 2010, 29(4):723-731.

[20] Che W X, Li Z H, Liu T. LTP: A Chinese Language Technology Platform. In Proceedings of the Coling 2010: Demonstrations, Beijing, China. 2010:13-16.