# 地质领域专有名词及固定短语提取

## 1.问题描述

地质领域有许多专有名词，如地名、地质历史时期、专有研究方法等。利用ICTCLAS分词后产生大量专有名词被错误切分，因此考虑将已有知识库中的地质领域叙词加入分词器自定义词典，确保部分专有名词能够正确切分。

本文选取中国地质大学地质图书馆整理的地质叙词表和中国分类主题词表作为叙词来源。其中地质叙词表中共有地质类叙词10235个；中国分类主题词表按照中国图书馆分类法中的天文学、地球科学大类选取地质类叙词共6408个；将两类叙词合并，并去重后共有14640个地质相关叙词。添加自定义词典后的分词效果如表1所示。

表1 添加自定义词典

|  |  |
| --- | --- |
| 添加自定义词典之前 | 添加自定义词典之后 |
| 泥 盆 纪  早 泥 盆 世  三 叠 纪  成 矿 作用  风暴 岩  高于 庄 组 | 泥盆纪  早泥盆世  三叠纪  成矿作用  风暴岩  高于庄组 |

由表1可以看出，添加地质叙词自定义词典后，能够将相应的地质领域叙词正确切分。但叙词是经过规范化处理的词语，除叙词外仍有许多非规范化的地质领域专有名词和固定短语被错误划分，见表2所举示例。

表2 未正确切分的地质领域固定短语举例

|  |
| --- |
| 频次大于10的未正确切分的固定短语举例 |
| 六 景； 大塘 坡 组； 淋滤 作用；  丹 池 盆地； 事件 沉积； 涠 洲 岛；  老 君山 组； 济源 盆地； 平行 层理；  洋 盆； 北 祁连； 灰 岩； |

由表2可知，添加自定义词典后依然有许多地质领域专有名词没有正确切分，如“涠 洲 岛”、“六 景”等地名；或者一些拥有特定含义的固定短语被切分，有可能造成部分主题缺失，如“老 君山 组”、“事件 沉积”等。因此，本文希望在分词基础上继续提取地质领域专有名词及固定短语，目的是为了能够提升主题聚类或主题抽取的效果，使主题由语义明确的词语表示，并且尽可能减少由于分词造成的主题词缺失。

## 2.汉语短语提取研究综述

短语，又称为语块、组块、语义块等。短语提取的基本方法分为基于规则的方法和基于统计的方法，或者是两者的结合。

赵军等[1]提出嵌套的基本名词短语规则，利用其对基本名词短语标注，根据与正确的标注结果对比，不断调整相应规则，是最早关于基本名词短语自动识别的文章；詹卫东等[2]对np、ap、vp、dj四种汉语短语结构规则进行了系统而具体的描述，制订了一个可以直接为计算机分析汉语短语结构提供支持的规则库；周雅倩等[3]根据上下文环境使用规则标注短语边界标记，并利用最大熵方法筛选规则，评估每条规则的重要程度；杨建国等[4]根据是否结合紧密、使用稳定，是否符合认知规律（短语长度一般为7±2），流通读是否达到一定的阈值这三个规则来判定短语是否为熟语，其中利用了频度、散步度、流通读等统计指标，并利用音节和短语结构信息对提取出的熟语进行分析；刘荣[5]利用动态语料库中的教育领域语料尝试大规模语料下的固定短语自动提取，从定量角度（互信息和左右熵）、句法角度（词性搭配规则）、语义角度（知网过滤）、时间角度抽取固定短语；罗盛芬和孙茂松[6]选取 9种常用统计量对字串内部结合紧密程度进行对比，通过对二字词的自动抽取实验表明，互信息的抽词能力最强；王素格等[7]在对互信息、Consine系数、x2 测试、似然比测试 4种词语相关性度量以及熵、方差、离散度3种词语结构分布度量分别进行比较分析后，提出基于互信息和熵融合的短语提取方法。

本文参考上述研究经验，选取互信息和左右熵作为统计信息标准，构建符合地质领域固定短语的词性搭配规则，经过3次迭代选取合适的固定短语，最后经过人工筛选确定地质领域固定短语。

## 3. 实验过程

1. 选取“杜远生”共87篇文章的题目、关键字、摘要作为原始语料，对语料分句、ICTCLAS分词、词性标注、词频统计。
2. 对分词后的词语穷举选取2-gram候选语块，如“南京 大学 学生”的候选语块有：“南京大学 大学学生”。
3. 计算互信息，互信息公式如下所示：

，

其中a，b为任意两个词语，P(a,b)表示两者共现的概率，即a、b共现的频次除以语料中所有二元搭配词组的频次，P(a)、P(b)表示a、b各自在语料中出现的频率。互信息降序排序见表1。

表1 互信息降序排序（前10）

|  |  |  |  |
| --- | --- | --- | --- |
| 二元短语 | 频次 | 词性 | 互信息 |
| 核心→部位 | 1 | n/n | 14.37051 |
| 锡→多金属矿床 | 1 | n/n | 14.37051 |
| 中保→存有 | 1 | n/v | 14.37051 |
| 压→实 | 1 | v/a | 14.37051 |
| 方→铈矿 | 1 | q/n | 14.37051 |
| 星体→坠入 | 1 | n/v | 14.37051 |
| 以南→广大 | 1 | f/b | 14.37051 |
| 埃→姆 | 1 | b/n | 14.37051 |
| 实→断层 | 1 | a/n | 14.37051 |
| 尚→需 | 1 | d/v | 14.37051 |

互信息值高说明词语搭配的内部结构稳定，但表1中互信息最高的前10个二元短语共现词频都为1，且组成候选短语的两个词语都是低频词。因此需要通过设定词频阈值过滤掉词频较少的词语组合。

从表2可以看出，当二元短语频次大于等于5时有候选短语个226个，满足阈值的短语的频次总数占全部候选短语的频次数的26.07%。因此本文选择频次个数5作为词频阈值。

表2 词频阈值

|  |  |  |  |
| --- | --- | --- | --- |
| 频次阈值 | 短语个数 | 频次数 | 占总频次的比例 |
| >=10 | 61 | 917 | 0.124170616 |
| >=8 | 85 | 1118 | 0.151387949 |
| >=5 | 226 | 1926 | 0.260798917 |
| >=1 | 4396 | 7385 | 1 |

表3 按照频次降序排序（前20）

|  |  |  |  |
| --- | --- | --- | --- |
| 二元短语 | 频次 | 词性 | 互信息 |
| 北→祁连 | 67 | f/ns | 7.461848 |
| 右江→盆地 | 32 | ns/n | 6.71739 |
| 2→O | 28 | m/nz | 7.625764 |
| 事件→沉积 | 28 | n/vn☆n/v | 6.30231 |
| 祁连→造山带 | 27 | ns/n | 6.828718 |
| 务→正道 | 26 | vg/n | 9.265175 |
| 灰→岩 | 26 | ag/ng | 8.210727 |
| 洋→盆 | 22 | ag/n | 7.460705 |
| 过程→中 | 21 | n/f☆n/b | 5.660095 |
| 正道→地区 | 21 | n/n | 7.072529 |
| 地球化学→特征 | 20 | n/n | 5.798196 |
| 山→带 | 20 | n/v | 6.403372 |
| 物→源 | 19 | ng/ng | 8.265175 |
| 碎屑→锆石 | 19 | n/n | 7.560697 |
| 裂→陷 | 19 | v/vi | 9.169148 |
| 铝土矿→形成 | 17 | n/v☆n/vn | 4.134959 |
| O→3 | 17 | nz/m | 7.223354 |
| 盆→山 | 17 | n/n☆q/n | 6.32091 |
| 地区→铝土矿 | 17 | n/n | 4.561224 |
| 叠加→模式 | 17 | vi/n☆vn/n | 8.976262 |

在表1中，互信息小于5的数据有4条，分别是“地区→铝土矿”、“铝土矿→形成”、“铝土矿→层”、“状→铝土矿”，说明这些词之间的内部结合不够稳定；互信息大于5的数据中，“2→O”、“山→带”、“务→正道”、“正道→地区”、“过程→中”、 “黔北→务”等是数据根据语感判断是语义不完全的搭配；因此，只通过频次的统计会产生较多噪声。

表4 词频5以上按照互信息降序排序（前20）

|  |  |  |  |
| --- | --- | --- | --- |
| 二元短语 | 频次 | 词性 | 互信息 |
| 充→填 | 10 | v/v | 10.64369 |
| 钦→防 | 11 | vg/v☆ng/v | 10.50618 |
| 正安→道真 | 11 | nr2/ns | 10.50618 |
| 务川→正安 | 11 | ns/nr2 | 10.50618 |
| 六→景 | 10 | m/ng | 10.50618 |
| 滨→岸 | 8 | ng/n | 10.38065 |
| 热→液 | 9 | a/ng☆n/ng☆v/ng | 10.35418 |
| 丹→池 | 8 | b/ng | 10.33626 |
| 老→君山 | 10 | a/ns | 10.26517 |
| 大塘→坡 | 10 | ns/n | 10.26517 |
| 豆→鲕 | 9 | n/w | 10.26517 |
| Al/→Al+Fe+Mn | 11 | xu/n | 10.15826 |
| 配→分 | 10 | v/qt | 9.878151 |
| 矿→岩系 | 8 | n/n | 9.556223 |
| 砾→屑 | 11 | ng/n☆ng/ng | 9.530228 |
| 防→海槽 | 8 | v/n | 9.506183 |
| 同步→叠加 | 12 | vd/vi☆vi/vn | 9.473761 |
| 主→量 | 12 | ag/n☆b/n | 9.463114 |
| 铝土→矿层 | 9 | n/n | 9.380652 |
| 生态→系 | 11 | n/v | 9.316521 |

与表2相比，大部分数字组合都未在表4中被选出；与表1相比，“2、3、中、特征、形成、地区、状”等高频词都未在表4中出现。通过设置频次的阈值设置为8，互信息阈值设置为4，把互信息高但频次低的候选短语过滤掉。

在以上满足阈值共102条，选前50条数据通过人工检查发现错误的提取结果有条， “矿→岩系”、“防→海槽”、“主→量”、“务→正道”、“屑→灰”、“西北→陆”、“硅→质”、“南海→西北”、“分→模式”、“复合→盆”、“秦岭造→山”、“黔北→务”。

对应的正确情况如下：

“含→矿→岩系”、“钦→防→海槽”、“主→量→元素”、 “务→正道→地区”、 “屑→灰→岩”、 “西北→陆→缘”、 “硅→质→岩”、“南海→西北→陆→缘”、“配→分→模式”、 “复合→盆→山→体系”、 “秦岭造→山→带”、“黔北→务→正道”。

这些错误是由于候选短语的左邻边界和右邻边界没有确定，使得生成信息不全、语义不完整的候选词语。因此下一步加入左右熵测量候选词语的左右边界。

熵的计算公式：

右熵的计算公式：

左熵的计算公式：

例如“生态→系”的右邻字集合及其频次为：{'生态→系→演变’：1，'生态→系→特征' ：1，'生态→系→突变'：2，'生态→系→菌'，1}。

右信息熵为：-(1/5)log2(1/5)- (1/5)log2(1/5)- (2/5)log2(2/5)- (1/5)log2(1/5)= 1.9219280948873623

表5 加入左右上的统计结果，按互信息降序排序（前20）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 二元短语 | 频次 | 词性 | 互信息 | 右熵 | 左熵 |
| 充→填 | 10 | v/v | 10.64369 | 2.846439 | 2.947703 |
| 六→景 | 10 | m/ng | 10.50618 | 2.321928 | 1 |
| 钦→防 | 11 | vg/v☆ng/v | 10.50618 | 1.095795 | 1.918296 |
| 滨→岸 | 8 | ng/n | 10.38065 | 1.548795 | 2.75 |
| 热→液 | 9 | a/ng☆n/ng☆v/ng | 10.35418 | 1.435521 | 2.235926 |
| 大塘→坡 | 10 | ns/n | 10.26517 | 1.485475 | 2.5 |
| 铝土→矿层 | 9 | n/n | 9.380652 | 2.947703 | 2.5 |
| 生态→系““ | 11 | n/v | 9.316521 | 1.921928 | 2.419382 |
| La→N | 11 | u/o | 9.265175 | 1.348588 | 2.155639 |
| 裂→陷 | 19 | v/vi | 9.169148 | 2.739572 | 3.721612 |
| 黏土→矿物 | 10 | n/n | 9.158259 | 2.75 | 2.281036 |
| 地球→生物学 | 10 | n/n | 8.92122 | 2.75 | 2.584963 |
| 陆→缘 | 15 | ng/ng☆nr1/n☆nr1/ng | 8.889511 | 2.931209 | 2.004936 |
| 负→异常 | 10 | v/n | 8.85709 | 2.251629 | 1.521928 |
| 冷泉→碳酸盐岩 | 13 | n/n | 8.759163 | 3.169925 | 3.121928 |
| 孤立→台地 | 11 | v/n☆a/n | 8.573297 | 3.459432 | 3.095795 |
| 生境→型 | 10 | n/k☆n/ng | 8.473761 | 2.321928 | 2.807355 |
| 物→源 | 19 | ng/ng | 8.265175 | 3.664498 | 4.084963 |
| 氧→相 | 10 | n/d | 8.243148 | 2.641604 | 1.672933 |
| 灰→岩 | 26 | ag/ng | 8.210727 | 3.75 | 2.213092 |

词频阈值为8，互信息阈值为4，经过试验观察发现左右熵的阈值为1时候选词语抽取效果较好。满足三个阈值的数据共56条，与表4对比，在表5中“滨岸、热液、大塘坡”等独立成词的候选词语被明显选出，另外“地球生物学、陆缘、冷泉碳酸盐岩、孤立台地”等地质领域专有名词也被明显选出。

### 句法分析（定性）

句法分析主要根据词性序列的组合筛选符合句法规法的词语。

赵军根据汉语的特点将汉语基本名词短语定义为4种情况：

* baseNP →baseNP + baseNP
* baseNP →baseNP + 名词 | 名动词
* baseNP →限定性定语 + baseNP
* baseNP →限定性定语 + 名词 | 名动词

限定性定语 →形容词 | 区别词 | 动词 | 名词 | 处所词 | 西文字串 | (数词+ 量词)

baseNP是有一些存在一定规律性的词性序列的词语序列组成，如名词加名词可以构成baseNP，而副词加动词不会构成baseNP。

刘荣将最小词性组合的名词短语定义为：

* NP→形容词 + 名词
* NP→名词 + 名词
* NP→动词 + 名词

最小词性组合的动词短语定义：

* VP→动词 + 动词

根据地质领域的词性搭配特征，为了抽取名词短语，在刘荣的工作基础上，定义名词短语的最小词性组合为：

* NP→形容词 + 名词
* NP→名词 + 名词
* NP→动词 + 名词
* NP→数词/方位词 + 名词
* NP→名词 + 动词

最小词性组合的动词短语定义：

* VP→动词 + 动词

在定量分析后有56条符合阈值条件的数据，再根据各个词性组合对候选词语进行统计。其中候选词语序列有多个词性序列的情况，人工进行处理保留一个，如“热→液”，词性有“a/ng”、“n/ng”、“v/ng”，保留“a/ng”。

“形容词+名词”的候选词语有4个：热→液、孤立→台地、灰→岩、洋→盆，全为正确的抽取结果。

“名词+名词”的候选词语有26个：滨→岸、大塘→坡、铝土→矿层、黏土→矿物、地球→生物学、陆→缘、冷泉→碳酸盐岩、生境→型、物→源、淋滤→作用、量→元素、碎屑→锆石、土→状、锆石→年龄、阴沟→组、长安→组、源→区、济源→盆地、右江→盆地、盆→山、相→区、状→矿石、化学→特征、地球化学→特征、地区→铝土矿、状→铝土矿，其中“量→元素、源→区、状→矿石、地区→铝土矿、状→铝土矿”不是合法的固定短语。

“动词+名词”的候选词语有2个：负→异常、造→山，为正确的抽取结果。

“数词/方位词 + 名词”的候选词语有3个：六→景、北→祁连、2→O，其中“2→O”为错误的抽取结果。

“名词 + 动词”的候选词语有6个：钦→防、生态→系、山→带、事件→沉积、硅质岩→具有、铝土矿→形成，其中“山→带、硅质岩→具有、铝土矿→形成”是不合法的抽取结果。

“动词+动词”的候选词语有：充→填、裂→陷，都为合法的匹配结果。

剩余不符合最小词性序列的词语都是14个：La→N、F→F、岛→洋、SiO→2、鲕→状、东→段、O→3、值→平均、造山带→东、3→个、具有→明显、过程→中、铝土矿→层、型→铝土矿，其中“鲕→状、铝土矿→层”

表6 词性序列抽取效果

|  |  |  |
| --- | --- | --- |
|  | 实际为短语 | 实际不为短语 |
| 抽取为短语 | 34 | 9 |
| 抽取不为短语 | 2 | 12 |

因此，词性抽取的Precision值为：79.06%，Recall值为：94.44%。由此可见，通过词性序列抽取，使得固定短语提取效果进一步提升。

### 迭代

在之前定量和定性分析后，加上人工筛选，选择合法的名词短语共32个，动词短语2个，将抽取出的固定短语替换之前的分词文本，迭代前两步操作。在迭代过程中，对于符合词性序列的词语搭配，人工筛选，尽可能保存在下一轮迭代中潜在的成词序列。比如“量→元素”符合“名词+名词”序列，虽然在此轮迭代中不是合法的词语序列，但是依然保留；对于“硅质岩→具有”虽然符合“名词 + 动词”词性序列，但无论在本轮还是下轮迭代中都不具备合法的成词可能，因此剔除。

1. 第一次抽取的词语集合

|  |  |
| --- | --- |
| 阈值 | 词频>=8；互信息>=4；左右熵>=1 |
| 名词短语 | 热液、孤立台地、灰岩、洋盆、滨岸、大塘坡、铝土矿层、黏土矿物、地球生物学、陆缘、冷泉碳酸盐岩、生境型、物源、淋滤作用、碎屑锆石、土状、锆石年龄、阴沟组、长安组、济源盆地、右江盆地、盆山、相区、化学特征、地球化学特征、负异常、六景、北祁连、造山、钦防、生态系、事件沉积、鲕状、量元素、源区、状矿石 |
| 动词短语 | 充填、裂陷 |

1. 第二次迭代抽取的词语集合

|  |  |
| --- | --- |
| 阈值 | 词频>=8；互信息>=4；左右熵>=1 |
| 名词短语 | 主量元素、钦防海槽、地球化学特征、屑灰岩、山带、北祁连造山带、北祁连岛弧 |
| 动词短语 |  |

|  |  |
| --- | --- |
| 阈值 | 词频>=5；互信息>=4；左右熵>=1 |
| 名词短语 | 旋回、平行层理、角砾、主量元素、滨岸湿地、钦防海槽、地球化学特征、南华海、体系域、石灰沟、被动大陆边缘、风暴涡流、屑灰岩、肃南一带、早中二叠世、泥盆、大地构造背景、杂质元素、台地边缘、黄龙组灰岩、鲕状矿石、古海洋环境、柱状、南海北部、风暴浪、质岩石、空间分布、山带、东坡组、竹园组、北祁连造山带、古浪、池盆地、岩性、风暴作用、平均含量、大塘坡组、铝土矿系、砂岩碎屑、北祁连岛弧、裂陷盆地、元素含量、状构造、沉积过程、分布特征、沉积型、型铝土矿 |
| 动词短语 |  |

1. 第三次迭代抽取的词语集合

|  |  |
| --- | --- |
| 阈值 | 词频>=5；互信息>=4；左右熵>=1 |
| 名词短语 | 砾屑灰岩、秦岭造山带、岛弧火山岩、地震事件沉积、组碎屑岩、晚古生代硅质岩 |
| 动词短语 |  |

### 参考文献

[1] 赵军, 黄昌宁. 结合句法组成模板识别汉语基本名词短语的概率模型[J]. 计算机研究与发展, 1999, 36(11): 1384-1390.

[2] 詹卫东. 面向中文信息处理的现代汉语短语结构规则研究[D]. 北京: 北京大学博士学位论文, 1999.

[3] 周雅倩, 郭以昆, 黄萱菁, 等. 基于最大熵方法的中英文基本名词短语识别[J]. 计算机研究与发展, 2003, 40(3): 440-446.

[4] 杨建国. 基于动态流通语料库的汉语熟语单位研究[M]. 北京语言大学出版社, 2009.

[5] 刘荣. 面向教育领域的固定短语提取方法研究[D]. 北京语言大学, 2009.

[6] 罗盛芬, 孙茂松. 基于字串内部结合紧密度的汉语自动抽词实验研究[J]. 中文信息学报, 2003, 17(3): 9-14.

[7] 王素格, 杨军玲, 张武. 自动获取汉语词语搭配[J]. 中文信息学报, 2006, 20(6): 33-39.