基于语素的化学术语分析和识别

摘　要：术语识别是领域文本，尤其是科技文本中自然语言处理中的基础环节，文本领域性也主要由术语体现。因而对术语的分析和识别具有巨大的语言学价值和实用价值。化学术语是种类多样、内部结构复杂、但规则性较强的一类术语，这对化学领域科技文本的处理造成很大困难，很少有在实际的发明专利上对化学术语进行识别的技术方案。本文分析了现有术语识别方法的优缺点，针对化学术语的组成特点和构词规律，从语素理论出发，构建了化学领域分类语素表，并用简单稳定的 HMM 模型对语素和语素类的构词关系建模，利用改进的前向算法计算构词概率实现术语识别，最终 F1 值达到了 91.58%。解决了传统统计方法遇到的化学术语词频低且构词过长的问题；同时也解决了规则方法中建立规律时的主观性和完备性问题。
关键词：术语识别；化学术语；语素；构词规律；语素类

Morpheme-based Chemical Term Analysis and Recognition

Abstract：Term recognition is the foundation of natural language processing in domain text, especially in scientific text, the domain of text is also mainly reflected in terms. Therefore, the analysis and recognition of terms have great linguistic and practical value. Chemical terms are a variety of terms with complex internal structures but strong regularity. It is very difficult to process the text in the field of chemistry, and there are few technical solutions for identifying chemical terms in actual invention patents. In this paper, the advantages and disadvantages of the existing recognition models are analyzed. According to the composition characteristics and word-formation rules of chemical terms, starting from the morpheme theory, a morpheme list in the field of chemistry is constructed, and an HMM model is established by taking chemical morphemes as observation sequences. Then, the improved forward algorithm is used to avoid the problem of too long terminology, and finally the F1 value reached 91.58%. The model solved the problems of low frequency and long word formation of chemical terms encountered in traditional statistical methods. At the same time, it also solved the problem of subjectivity and completeness in establishing rules.
Keywords：Term recognition; Chemical term; morphemes; Word-formation rules; Morpheme class

1　引言

术语识别，尤其是领域术语识别，是自然语言处理任务在实际使用场景中的重要组成部分。使用主流的边界词、上下文接续词等特征的识别方法，还不足以解决当前化学领域术语识别遇到的构词方式多样、词表庞大的问题。然而通过对大量科技术语，尤其是化学物质名称进行观察，发现其命名具有稳定的构成模式以及规模有限的化学用字，而这些化学用字与语素呈现出一一对应的关系。从语言学角度来看，语素是语言中最小的音义结合体，构词的主要方式是语素的组合。

故本文将语素观引入到化学术语识别，从化学语素入手，将语素构成术语的构词规律应用到术语识别上。具体为借助对化合物命名规范和领域语料的统计分析，总结整理出化学术语的构成语素及其语素类；再从领域词表中学习语素及其语素类的构词规律，得到反映构词规律的 Hidden Markov Model 模型，从而将化学术语的识别问题转化为计算化学语素有效构词概率的过程，解决了现有工作中的标注任务繁重、低频术语、过长术语的遗留问题。

2　相关工作

目前主流的命名实体和专业术语的识别方法是规则和统计相结合的方法。梁樑（2002）根据构词特点及词频分布识别商品文本中的药物名称等；宋丹等（2009）将化学特定词分成三类，提出基于规则的化学物质名称识别方法；李楠等（2010）使用规则方法和 CRF 模型将

术语识别转化为有效词类序列识别的过程；马建红（2018）根据化学资源文本的语言规律及特点，使用统计方法进行初步识别，再使用词典与规则的方法进行校正。

从统计学习的角度解决术语识别问题，需要搜集特定标注资源并制定对应算法，信息集成难度大；或者利用上下文进行识别，但仍然不能解决化学低频术语很多的长尾问题。从知识规则的角度看，化学领域的构词方式多样、词表庞大，没有严格的构词规律且无法穷尽，过长的低频术语、通用词类也带来干扰，难以构建完整的专名资料库。

此外，从语素的角度分析构词特征，能够为术语识别带来一定帮助。于东等（2014）认为，具体学科的复杂术语构词可描述为学科基元语素、固有语素组、复杂术语。李楠（2003）等将化学物质命名中的语素进行分类；王倩倩等（2015）定义和分类化学词素，利用规则方法建立化学术语的知识量计算模型。

现有从语素角度进行的化学术语识别工作大都是从规则的角度进行的，有一定的准确率，但无法解决化学术语组合灵活多变、长词识别困难的问题，本文将语素作为化学术语识别过程中的基本单位，对化学语素进行分类，并通过模型学习和预测不同化学语素类间关系，将化学术语识别问题转化为化学语素构词概率计算问题，在一定程度上解决了现有工作中的遗留问题。

3 化学术语与语素

3.1 化学术语

广义的化学术语包括化学领域的基本概念名词、化学元素、化学式、化学仪器名称、化学反应原理名称以及与其它学科交叉形成的物理化学名词、生物化学名词等等，本文的化学术语指的是以化学式为代表的化学物质名词，这类词语往往长度大、数量多、扩展性强、不可能完全收录于化学词表中，是化学术语自动识别的难点。

同时，这些化学术语也具有一些规律性的特点。通过观察大量的化学术语词，发现化学术语的构词存在很强的规律性。以下是几条有机化合物的命名规则[6]：

(1) 天干 + 取代基/物质类属词；如：“甲基”“丁烷”“癸烷”；
(2) 化学元素名 + 物质类属词；如：“硼酸”“溴酸”；
(3) 中文数字 + 天干 + 取代基 + 天干 + 物质类属词；如：“三甲基丁烷”；
(4) 中文数字 + 化学元素名 + + 中文数字 + 化学元素名；如：“二氧化碳”、“三氧化二钒”；
(5) 取代位置用字 + 中文数字/天干 + 化学元素名；如：“聚四氟乙烯”、“对甲苯”、“间溴苯丙酮”；
(6) 阿拉伯数字+ ‘-’ + 中文数字 + 天干 + 化学元素/取代基 + 天干 + 化学元素；如：“2,7,8 -三甲基癸烷”

此外，从上述几条规则中我们可以发现，化学术语用字的范围是有限的，包括：化学元素表中的汉字、化学物质类属词（“酸”“酯”“醇”）、阿拉伯数字、中文数字、天干、化学结构用字（“对”“聚”“偏”“亚”）、连接符号等，具有可穷举性；而且，从语言角度来看，化学术语构成的最小单位与语素呈现出一一对应的关系，故本文将从化学语素角度入手，通过语言模型学习不同化学语素类间的组合关系，进而判断候选字符串序列成为化学术语的概率。

3.2 化学语素及分类

3.2.1 化学语素

在现代汉语语法中，术语“语素”是英语“morpheme”的汉译。这一概念是由美国结构主义语言学家布隆菲尔德提出的。在张斌（2017）《新编现代汉语》中把语素定义为：“语

素是语言中最小的音义结合体，是能够区别意义的最小的语言单位。语素的作用和职能主要是构词[19]。”从其表达形式、内容和意义的关系上来讲，它不能再切分，再分则必定会完全改变或破坏其原有语法、词汇的意义，汉语中一般一个汉字是一个语素。有些语素可以独立成词，语素与语素之间也可以通过相互组合成词。事实上，在现代汉语中，以单音语素与单音语素的相互组合确实产生了大量新词，甚至成为产生新词的最重要的方式之一。

化学术语中的语素与一般语素的基本语义功能一致，都指代基本意义单元，是最小的构词单位。化学领域中大多数语素都是表义明确的、具有专业性的单字语素，语素的含义具有更加明确的专业意义，例如“烷”“烃”“酸”等表示物质的种属。此外，化学术语中有较多的音译词，而这些词再切分为单字时，便不再具有实际意义了，我们将这些音译词整体当作一个语素处理，如“吡啶”“呋喃”等。因此，本文讨论的化学语素的概念是指具有化学术语特征或者辅助命名化学物质的最小构词单位，可以是一个数字、字母、汉字或者一个词。

3.2.2. 化学语素分类

根据化学语素构成化学术语时的位置及功能，本文将化学语素分为核心语素、限定语素、辅助语素、通用语素四类。

核心语素是指在化学术语中能明确表示物质的性质或类别，使其区别于其他类物质，起到最重要的标识和核心关键作用的语素。这类语素基本只在化学相关的领域文本中出现。核心语素又包括化学元素和化学专名两个小类，分别用字母 A 和 B 表示。其中化学元素是一个固定的集合，目前公布的化学元素共有 118 个。化学专名是指表示化学物质类属的语素和一些多音节的语素，如“酸”“脂”“盐”“吡啶”等。

限定语素是指在核心语素的基础上，对物质的具体性质、类别或结构起到限制作用的语素，它可以进一步限定缩小物质的种类或范围，从而有利于明确物质的概念，用字母 C 表示。它主要包括化学术语中经常出现在词头的限定语素，如“正、原、偏、焦、重、亚、次、高、连”等，表示化学物质生成方法的化学介词语素，如“化、合、聚、缩、代”等。

辅助语素是指在化学术语中，大写数字、罗马数字、阿拉伯数字、天干、英文字母、拉丁字母以及一些化学连接符号等，与核心语素及限定语素共同使用，以提高化学术语的专指度，用字母 D 表示。

通用语素是指在化学领域和其它领域的术语中均可出现的语素，其语素意义单一，跟领域没有相关性，对化学术语的专指度没有影响。它主要包括构成化学基础术语的语素，如“电、解、溶、液”等，用字母 E 表示；通用领域中语素，如“水、子、白”等，用字母 F 表示。

表 1 化学语素分类表

类别		化学语素
核心语素	化学元素（A）	氢、氦、锂、铍、硼、镍……
	化学专名（B）	酸、盐、碱、脂、醇、醚……
		吡啶、呋喃、嘧啶、吲哚……
限定语素	化学介词（C）	化、合、聚、缩、代……
	特定词头（C）	正、原、偏、焦、重、亚、次……
辅助语素（D）		大写数字、罗马数字、阿拉伯数字、天干、英文字母、拉丁字母、化学连接符号
通用语素	化学通用语素（E）	电、解、溶、液、器、质……

	一般通用语素（F）	水、子、白、光、蓝……

3.3 化学语素表构建

为训练化学术语识别模型，本文首先构建了化学术语词表，主要来自搜狗的化学词库[1]以及维基百科的有机化学[2]和无机化学[3]的术语，共计 22000 多个化学术语词。采取统计与人工相结合的方法构建了化学语素表。首先，对化学术语词表进行统计，得到了单音节语素 2045 个；随后，采取人工标注的方法，收集了词表中的多音节语素 105 个；之后，由于化学词表有限，不能覆盖所有的化学语素，而核心语素和辅助语素的集合相对封闭，故人工补充了核心语素和辅助语素；最后，结合语素出现的频次，对单音节语素进行了人工筛选，共得到 718 个语素。据统计可知，化学语素的数量远远小于化学术语的数量。可见，化学术语的语素组合灵活，构词能力强。利用化学语素及化学语素的组合规律进行化学术语识别简单可行。

表 2 化学领域频率最高的 20 个语素

语素	频次	语素	频次	语素	频次	语素	频次
酸	6707	基	3421	二	2721	化	2510
甲	2368	苯	2176	乙	1805	氯	1649
氧	1312	胺	1246	醇	1181	氨	1120
硫	1077	酰	1046	酯	1019	丙	921
三	913	盐	871	酮	866	烷	841

表 2 为化学领域频率最高的 20 个语素，其中有 16 个为核心语素，且语素“酸”出现的频率最高，总共出现 6707 次。图 1 为化学术语词表中所有语素频率分布表，也呈现出同样的趋势线。即分布靠前的基本为核心语素，其次为限定语素。

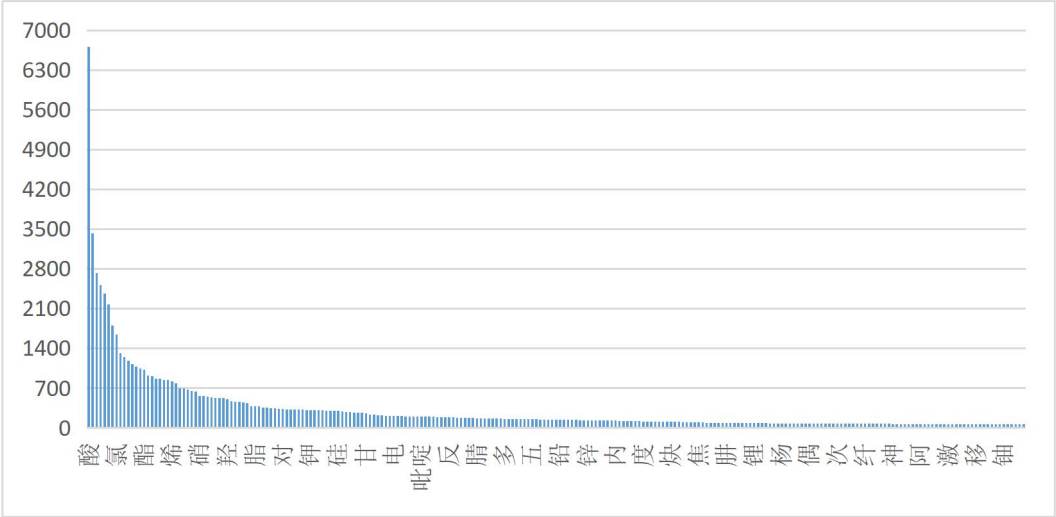

图 1 化学语素频率分布图（部分）

在构建了化学语素表之后，按照上文提到的化学语素分类说明，对化学语素进行了分类的人工标注，并统计了不同类别的化学语素所构成的化学术语数量。根据表 3 可知，化学专名（B）、限定语素（C）类的语素数量最多，分别占所有语素的 22.98%、20.75%，一般通用语素（F）语素数量最少，占所有语素的 10.58%，其它几类语素数量接近。而 A、B、C、

[1] 搜狗细胞词库-化工：https://pinyin.sogou.com/dict/cate/index/109
[2] https://zh.wikipedia.org/wiki/有机化合物列表
[3] https://zh.wikipedia.org/wiki/无机化合物列表

D 类均为相对封闭的集合，语素数量有限，因此，利用语素进行化学术语识别能够达到事半功倍的效果，在保证识别效果的同时，减少了人力投入。

表 3 化学语素分类情况统计

	化学元素（A）	化学专名（B）	限定语素（C）	辅助语素（D）	化学通用语素（E）	一般通用语素（F）
数量	109	165	149	101	118	76
百分比	15.18%	22.98%	20.75%	14.08%	16.43%	10.58%

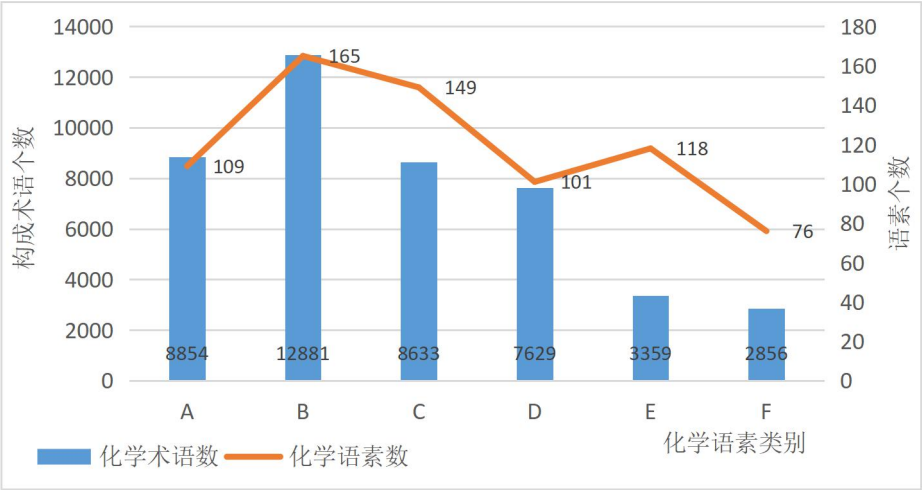

图 2 化学语素个数与所构成化学术语个数对比图

图 2 为每一类化学语素的数量与该类化学元素所构成的化学术语数量的对比图。经观察发现，C、E、F 类的化学语素数量较多，而其构成的化学术语数量却相对较少。分析化学术语词表得知，C 类为化学限定语素，一个化学术语中可能出现多个不同类型的化学限定语素，如“亚/C 铁/A 氰/B 化/C 钙/A ”；E、F 类为相对通用的语素，类型多样，且在化学术语出现只起辅助作用，出现情况无规律可循，故元素数量较多，而每个元素出现的频率较低，故构成的化学术语个数较少。

4 化学术语的识别方法

我们将化学领域语素集合固定、语素歧义少、语素组合有规律可寻、语素黏着程度高等特点作为术语识别的入手点，以化学语素作为基础，对化学术语词表中的语素的构词规律进行建模，之后计算句子中若干相连语素构成术语的概率，从而达到化学术语识别的目的。

具体来说，首先利用 HMM 拟合术语中语素的构词规律，之后由改进的前向算法计算语素构成术语的概率，再根据概率大小来判断是否合并语素，最后将合并结果作为识别结果。若干语素构成术语的概率使用阈值二值化，用于划分若干语素组合是否可以构成术语；在识别过程中，对于不同语素个数的候选术语用不同的构词概率阈值判断是否构成术语。

4.1 化学术语构词规律建模

本文采用 Hidden Markov Model 对化学术语的构词规律进行建模，隐马尔科夫模型能够用来描述未知参数并用这些参数进行问题分析与处理，且能够很好地描述观测状态与隐藏状态之间的隐含关系，本文将其应用于汉语化学术语识别。形式化的来说，模型 λ 可以表示为：

$$\lambda = (\Omega_x, \Omega_o, A, B, \pi)$$

其中，化学语素分类是模型的隐藏状态，即$\Omega_x = \{q_1, q_2, ..., q_6\}$，包含6个化学语素分类；化学语素是模型的观测状态，即$\Omega_o = \{y_1, y_2, ..., y_{718}\}$，包括718个化学语素；初始状态矩阵$\pi$是术语词首语素的分类频率估计得到的概率矩阵；状态转移矩阵A是由前一个语素分类转移到后一个语素分类的频率估计的概率矩阵；观测概率矩阵B是由语素分类转移到语素的频率估计得到的概率矩阵，也就是：

$$\frac{\#(语素y_j的频次)}{\#(y_j所在分类下所有语素的频次)}$$

在B矩阵中，会出现矩阵中的值等于0的情况，本文简单使用了加一平滑处理。

4.2 改进的前向算法

给定HMM模型λ，可以利用前向算法计算语素构成术语的概率，但是在迭代计算前向概率的过程中，随着观测序列长度的增加，导致递推概率值越来越小甚至趋于0。本文在改进的前向算法过程中，加入了对观测序列长度的处理，保证了最终概率不会算术下溢，可以用于计算语素构成术语的概率。改进的前向算法形式化如下：

给定模型参数 $\lambda = (A, B, \pi)$，计算观测序列$\Omega_o = \{y_1, y_2, ..., y_N\}$的概率$P(O|\lambda)$，其求解迭代过程为：

1) 计算时刻1的各个隐藏状态的前向概率：$\alpha_1(i) = \pi_i b_i(y_1), i = 1,2,...,6$

2) 计算时刻t的前向概率，$t \in [1, N]$：

$$\alpha_{t+1}(i) = [\sum_{j=1}^{6} \alpha_t(j) a_{ij}] b_i(y_{t+1}), i = 1,2,...,6$$

3) 计算改进的前向算法概率值：$P'(Y|\lambda) = \log_{1/t} \sum_{i=1}^{6} \alpha_T(i)$

由此可见，在给定模型 λ 的条件下，前向算法概率值$P'(O|\lambda)$就是化学语素观测序列O能够构成化学术语的概率。

5 实验

5.1 实验数据

实验数据包括化学语素表、化学词表和识别测试集。化学语素表和化学词表为自主构建和筛选，语素表包含 718 个语素，化学词表包含 22356 个词，词表平均词长是 4.3 个语素，详情见 2.2 节。化学专利数据是术语识别的重要场景之一，本文将《国际专利分类表》（IPC 分类）中 C01（无机化学）和 C07（有机化学）下的中文专利文献按整句随机抽取来构建测试集，并人工标注出句子中的化学术语。测试集共包含 531 句，平均术语长度 6.998，最大术语长度 48，平均句子长度 24.537，最大句子长度 172。

为方便交流学习，本文的实验数据和代码均已公开可下载[4]。

5.2 实验流程和评价指标

实验流程分为准备模型和识别测试两部分。在准备模型阶段，关键步骤是计算 HMM 模型参数、确定构词概率阈值；在识别测试阶段，关键步骤是根据构词规律模型的概率值合并化学语素，从而完成术语识别任务，伪代码流程如下：

[4] https://github.com/xiabo0816/ChemNer

```
// 定义返回值和临时语素数组
results = []
term    = []
// 统计估计模型参数，确定概率阈值
Pi, A, B, threshold = TrainHMM (morpheme, wordlist)
// 遍历待识别句中的语素
for morpheme in sentence:
      if prob(term + morpheme) > threshold:
            //前后语素可以构成术语
            term += morpheme
      else:
            //前后语素不可以构成术语
            results.append(term)
            term.clear()
return results
```

语素个数为 2 时使用的阈值是 15，语素个数为 3 时使用的阈值是 15.8，语素个数为 4 时阈值是 18.7。评价指标采用准确率（Precision）、召回率（Recall）和 F1 值。

5.3 结果和分析

利用 3.2、3.3 提到的识别模型和改进的前向算法，可以计算得出句子中若干相连语素构成术语的概率。

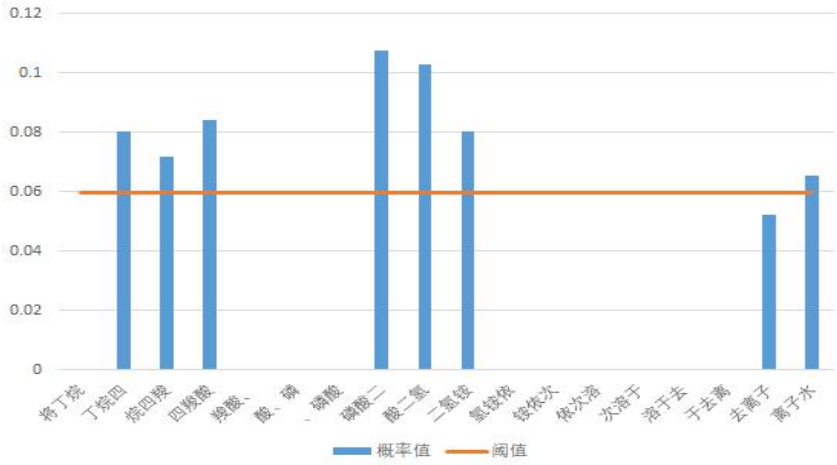

图 3 句子 1 中语素构词阈值与术语识别结果

由图 3、图 4 可以看出，基于语素的术语识别方法可以出色的完成对多术语、长术语、复杂术语的识别任务。比如，在句子 1“将丁烷四羧酸、磷酸二氢铵依次溶于去离子水中”中，根据图 3 可知，大于阈值的相邻语素是 “丁烷四”“烷四羧”“四羧酸”等，所以可以合并相连语素，将合并后的字符串序列作为识别结果，即“丁烷四羧酸”“磷酸二氢铵”。

再如，在句子 2“取正辛酰氯、乙烯基三乙氧基硅烷、丙烯酸-2-乙基己酯、羟基甲基膦酸二乙酯、羟基乙叉二膦酸钾于三口瓶中”中，通过图 4 可以看出，根据语素构词的阈值，该句中有五组可以合并的相邻语素，故最后识别出“正辛酰氯”“乙烯基三乙氧基硅烷”“丙烯酸-2-乙基己酯”“羟基甲基膦酸二乙酯”“羟基乙叉二膦酸钾”五个术语。

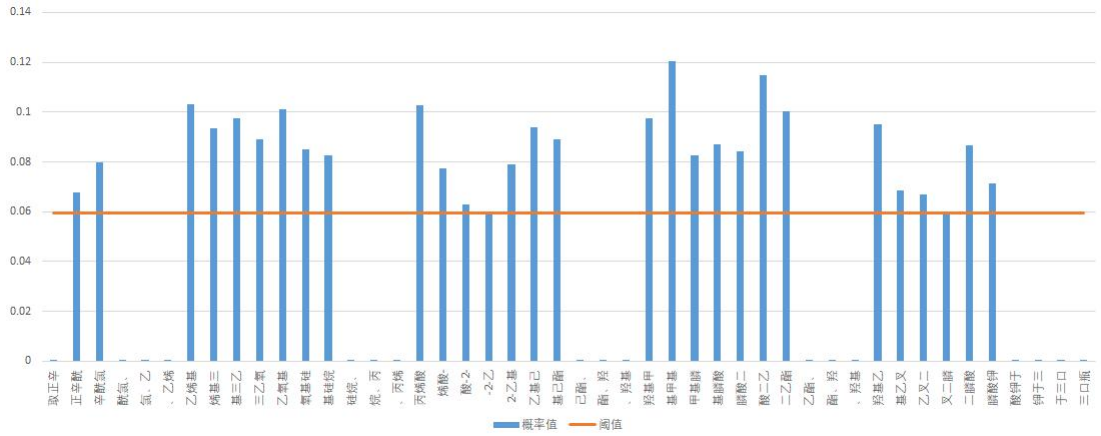

图 4 句子 2 中语素构词阈值与术语识别结果

在实际场景中，术语的命名复杂性、长尾效应体现的更加明显。表 4 为化学专利文献测试集的测试结果，根据表 4 可知：

1．实际情况下，作为专利核心发明的化学术语中，短术语很少，术语平均长度较大，术语复杂度高且复杂度呈非线性增长；

2．无论术语长度如何变化，召回率稳定在 98%以上，充分利用了化学术语虽然命名变化多、但化学语素稳定的特点；

3．术语长度在≤10 时，平均准确率达到 89.13%以上，充分利用了化学术语的黏着程度高、语素歧义性低的特点；

表 4 化学专利文献测试集的测试结果

语素数	总数	召回数	正确数	P 值	R 值	F1 值
≤3	6	6	6	100.00%	100.00%	100.00%
4	10	10	9	90.00%	100.00%	94.74%
5	288	286	275	**96.15%**	99.31%	98.39%
6	54	53	47	88.68%	98.15%	93.17%
8	30	30	28	93.33%	100.00%	96.55%
10	12	12	8	66.67%	100.00%	79.80%
≥12	131	130	75	57.69%	99.24%	72.97%
all	531	527	448	**85.01%**	99.25%	**91.58%**

在运行效率上，该测试集在实验机[5]上单线程耗时 3500ms，平均每句 7ms。粗略估算在主流服务器配置上，至少可以支持响应时间 300ms 以内的 30+并发请求。

当前模型识别效果较前人工作有很大提升，但仍有不少问题。经观察测试集结果，发现识别效果较差的化学术语主要有以下几种类型：

从化学语素类型上看，主要是含有多个“-”、数字的术语以及含有括号的化学术语，如“2-氯-6-三氯甲基吡啶”“四溴邻苯二甲酸双（2-乙基己基）酯”，主要是因为数字和括号具有通用性，领域性较弱，训练数据中该类型的术语数量少，模型学习不充分，且在进行计算时，没有考虑语素的位置信息；从化学术语长度上看，本文的模型在一定程度上解决了识别过程中术语过长的问题，对 10 字以内的化学术语识别效果较好，大于 12 字的术语效果也达到了 57.69%，相对短术语识别效果较差，可增加长术语的训练数据规模，针对长术语进一步优化模型。

[5] 实验机配置：CPU i7-8568U，内存 16.0GB

6 结论

本文利用术语和语素本身的特点以及术语中语素和语素类间关系，对术语的语素构词规律建模，并运用于化学领域科技本文，很大程度上解决了化学术语的整体识别中的长尾效应明显、命名规则复杂、识别难度大的问题。综合来看，从语素角度入手进行化学术语识别具备了较高的准确性和高效性，克服了化学术语识别复杂度高的困难。

本文的化学术语识别模型仍有待进一步改进。从模型和算法的角度看，HMM 的假设只能捕捉相邻语素和相邻语素类的信息，需要用更全局的建模方式（如 CRF）对构词规律建模或使用深度学习将术语结构向量化；从术语和语素的角度看，术语的位置和边界信息目前还未加入到识别模型中进行预测；另外，不同领域的术语语素的通用性不同，除了科技术语以外，新闻、体育术语是否具有明显的语素稳定的特点有待研究。

参 考 文 献

[1] Frederick Jelinek. Statistical methods for speech recognition. The MIT Press, 1997.

[2] 术语在线, http://www.termonline.cn/index.htm

[3] Ma, Wei-Yun & Chen, Keh-jiann. A Bottom-up Merging Algorithm for Chinese Unknown Word Extraction. 2003

[4] Zhai, Zenan & Nguyen, Dat & Akhondi, Saber & Thorne, Camilo & Druckenbrodt, Christian & Cohn, Trevor & Gregory, Michelle & Verspoor, Karin. Improving Chemical Named Entity Recognition in Patents with Contextualized Word Embeddings. 2019.

[5] Xia, Bo & Xun, Endong. Distributed Representation of Chinese Collocation. Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation. 2018.

[6] 中国化学会. 有机化合物命名原则 2017. 2017

[7] 张静. 自动标引技术的回顾与展望[J]. 现代情报, 2009,29(04):221-225.

[8] 李质轩. 融合上下文信息的汉语分词方法研究[D]. 北京交通大学, 2018.

[9] 黄昌宁. 对自动分词的反思[C]. 哈尔滨工业大学计算机科学与技术学院、清华大学智能技术与系统国家重点实验室. 语言计算与基于内容的文本处理——全国第七届计算语言学联合学术会议论文集. 哈尔滨工业大学计算机科学与技术学院、清华大学智能技术与系统国家重点实验室:中国中文信息学会, 2003:36-48.

[10] 曾浩, 詹恩奇, 郑建彬, 汪阳. 基于扩展规则与统计特征的未登录词识别[J]. 计算机应用研究, 2019,36(09):2704-2707+2711.

[11] 岳金媛. 面向专利文献的汉语分词技术研究[D]. 北京交通大学, 2013.

[12] 马建红, 王立芹, 姚爽. 面向化学资源文本的命名实体识别[J]. 郑州大学学报(理学版), 2018,50(04):14-20.

[13] 梁樑, 李禕. 商品文本中药物名称和化学名称识别的研究[J]. 烟台大学学报(自然科学与工程版), 2002(04):280-285.

[14] 宋丹, 孙济庆, SongDan, et al. 基于规则的化学特征词自动标引研究[J]. 情报学报, 2009, 28(5):689-692.

[15] 李楠, 郑荣廷, 吉久明, 滕青青. 基于启发式规则的中文化学物质命名识别研究[J]. 现代图书情报技术,2010(05):13-17.

[16] 于东, 饶高琦, 唐共波, 荀恩东. 复杂科技术语构词中的语素化[J]. 中国科技术语, 2015,17(02):15-20.

[17] 李楠, 孙济庆, 吉久明. 汉语词素语义与知识发现研究初探[J]. 图书情报工作, 2013,57(17):109-113.

[18] 王倩倩, 陈荣, 李楠, 孙济庆. 面向化学名称的术语知识量计算模型研究[J]. 图书馆杂志, 2015,34(10):59-62+98.

[19] 张斌. 现编现代汉语[D]. 复旦大学出版社, 2017.