情报工程 ISSN: 2095-915X
第1卷 第4期 95-103 2015年8月
TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN: 2095-915X Vol.1 No.4 95-103 Aug 2015

doi:10.3772/j.issn.2095-915x.2015.04.011

面向专利的化合物和生物实体识别系统

赖鸿昌，朱礼军，徐硕

（中国科学技术信息研究所信息技术支持中心 北京 100038）

摘要：探索专利文献中的化合物和生物知识变得至关重要。为了识别化合物实体和生物实体，开发了面向专利的化合物和生物实体识别系统。系统基于开源的机器学习和自然语言工具进行开发。系统按照流水线模式进行，本文将详细阐述其三个主要过程：预处理（句子分割、词条化），识别（基于条件随机场的方法），后处理（基于规则的方法）。最后，利用系统在已标注的化合物专利语料库进行大量实验，进行十折交叉验证，得到了69.20%的F值。但是，从结果可以看到，在专利文献上的实验表现，要低于论文和新闻语料库中的表现。

关键词：条件随机场，化合物和生物实体，专利挖掘，交叉验证

中图分类号：G350，TP311

Chemical and Biological Entity Recognition System from Patent Documents

LAI Hongchang, ZHU Lijun, XU Shuo

(Center of Information Technical Support, Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: It is crucial to explore the chemical and biological space covered by patent documents. In order to recognize chemical and biological entities, a recognition system is developed on the basis of open-source machine learning and natural language processing (NLP) toolkits. The system processing pipeline consists of three major components: pre-processing (sentence detection, tokenization), recognition (conditional random field (CRF) based approach), and post-processing (rule-based approach). The paper introduces each part in detail.

基金项目：本研究得到国家自然科学基金项目“基于论文和专利资源的技术机会发现研究”（项目编号：71403255）、中国科学技术信息研究重点工作项目“大数据环境下融合多源信息的科技文献智能分析服务平台建设及应用示范”（编号：ZD2014-7-1）的资助。

作者简介：赖鸿昌（1992-），硕士研究生，研究方向：知识工程；朱礼军（1973-），博士，副研究员，研究方向：知识组织、语义检索等；徐硕（1979-），博士，副研究员，研究方向：数据挖掘、信息抽取等。通讯作者：徐硕，E-mail: xush@istic.ac.cn。

Finally, extensive experiments on annotated chemical patent corpus are conducted, and the balanced-F measure is 69.20% with 10-fold cross validation. The results indicates that the performance on patent documents is slightly lower than that of counterpart on paper and news corpus.

Keywords: Conditional Random Field (CRF), chemical and biological entity recognition, patent mining, cross validation

1 引言

探索专利文献中的化合物和生物知识变得至关重要。在早期的药用化学活动的研究中，能起到很好的加速作用 [1, 2]。专利文献中包含大量有价值的化合物和生物实体，例如一些化合物、基因、药物、药物靶点等。但是，面向专利文献的自动识别系统仍然非常有限。

然而，针对论文语料和新闻语料，已经提出了很多识别方法，也开发了较多识别系统。对于专利语料库和论文、新闻语料库的反差，本文认为可能存在两个原因：(1) 已经标注的专利文献对公众而言不容易获取；(2) 专利文献具有法律效应，内容复杂繁琐，难于理解。但是对专利挖掘兴趣的不断提升，使得专利文献命名实体识别的情况正在不断好转，面向专利挖掘的会议也越来越多，例如 BIOINFORMATICS [3], BioCreative[4], JNLPBA [5] 和 iPaMin[6]。Akhondi 等人发布了一个已标注的化合物专利语料 [8]，有助于进行化合物和生物实体识别系统的开发。多篇专利语料先进性自动预标注，再人工将化学名词按照不同的子分类、疾病、靶标、作用方式等进行分类标注。语料全集（Full Set）包含 198 篇专利文档共 400,125 个标注，统一集（Harmonized Set）包含 47 篇专利文档共 37,776 个标注。而且专利文档具有相当的复杂性 [7]，有些内容可能多达上百页。当前环境下，对于无结构文档，尤其是专利文档，自动识别其中的化合物和生物实体仍然是非常具有挑战性的任务。

本文采取了和 Xu S 等人类似的方法 [11]，开发了面向专利文献的化合物和生物实体识别系统。文章的第二部分对数据集进行概述，第三部分介绍了系统的构成和所使用的方法，第四部分是介绍了实验的相关信息，并介绍了实验所使用的数据集。

2 数据概述

Akhondi 发布了一个黄金标准的化合物专利语料，包含两个数据集：统一语料集（harmonized corpus）和全语料集（full corpus）。统一语料集由 47 篇专利文档组成，包含了 9,813 个唯一术语的 36,537 个标注，另外，还含有 1,239 个 OCR（Optical Character Recognition，光学字符识别）错误，包括 1189 个拼写错误和 50 个换行错误；全语料集由 198 篇专利文档组成，包含了 80,977 个唯一术语的 400,125 个标注，其中含有 5,096 个 OCR（Optical Character Recognition，光学字符识别）错误，包括 4,403 个拼写错误和 693 个换行错误。专利的全文和实体标注文档公开发布在 www.biosemantics.org. 另外，还包括了只含两篇文档的训练集。在实验中，主要对统一语料集进行处理和实验。

在对训练集和统一语料集进行分析之后，

在统一语料集中发现了一些化合物和生物实体具有两个或者多个标注，具有嵌套关系。由于在系统中，将 CRF++ 作为处理序列标注问题的具体实现，而 CRF++ 无法处理和识别嵌套的实体，因此我们对此进行了一些相对合理的处理，使每个实体都只具有一个标签。在统一语料集中，标注为 OCR 错误的实体（OCRERRORSPELL，OCRERRORLINE）中，有些还在同一篇文档中标注为其他类型的实体，为保证准确性，我们将其统一为 OCR 错误。最后，我们移除了 488 个嵌套实体，将标注数量从 37,776 减少到 37,288。最后，由 47 篇文档所构成的统一语料集，一共包含了 14 个类别，含有 9857 个唯一实体的 37288 个标注（见表格 1）。

如表格所示，在统一语料集中，标注为 M（IUPAC，International Union of Pure and Applied Chemistry，国际理论和应用化学联合会）和 G（Generic）的实体明显多余其他类别的实体。而另一方面，类别为 Y (InChI, International Chemical Identifier，国际化合物标识), C (CAS(Chemical Abstracts Service，化学文摘社）注册码）and I (SMILES, Simplified molecular input line entry specification, 简化分子线性输入规范）在化合物专利文档中出现极少。由于在数据预处理中，移除了一些嵌套实体的标签（含有多个 tag 的实体），因此术语数量和 Akhondi 在文献中提到的 [8] 会有不同。

3 系统描述和相关方法

在对 MUC−4(Mucin 4) 系统中使用的基本方法进行总结后，Hobbs 提出了一个包含十个基础模块的信息抽取的一般方法模型 [10]。这是基于大量实践所提出的理论模型，也是我们系统设计的理论基础之一。另一方面，也参考了 Xu S 等人发表的化合物实体识别系统 [11]。如图 1 所示，系统由三个主要模块组成，犹如构成一个序列化的流水线。首先，将会对专利文档原文进行句子分

表格 1 黄金标准语料集中的各类术语数量（移除嵌套实体后）

	描述	标注术语量	唯一术语量
M	IUPAC	13943	4592
I	SMILES	20	20
Y	InChi	0	0
D	Trademark	2355	897
B	Abbreviation	2087	146
C	CAS number	6	5
F	Formula	1115	160
R	Registry Number	140	95
G	Generic	8381	811
T	Target	3221	654
Disease	Disease	3765	1205
MOA	MOA	1016	197
OCRERRORSPELL	Spelling error	1189	1029
OCRERRORLINE	Spurious line break	50	46
	Total	37288	9857

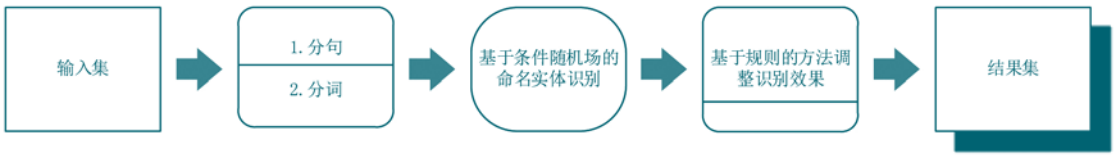

图 1 系统简化处理模块

割，每个句子间用制表符（“\t”）进行分割，并对每个句子进行分词，将句子分成一个个符合要求的词条。其次，利用条件随机场的方法（利用 CRF++ 作为具体实现）将化合物和生物实体从语料库中抽取出来并用十折交叉验证的方法对系统的效果进行评价。最后是对数据的后处理，使用基于规则的方法，提升实验效果并验证。

3.1 预处理：分句，分词

在统一语料集中，每个专利文献包含两种文档：被分割成若干部分的专利文档原文；各部分对应的标注实体后的文档。在语料库中，每篇专利都被分割成若干个部分，每个部分都包含了专利文档原文的不同部分，且没有重合部分。一般来说，每个子文档都是不规则的，每一行有可能是一个或几个句子，或者是几个元数据。例如，在训练集的 *US4659716_0001* 文档内，前几行是专利的元数据，包括标题、摘要、发明人等；而其余行则是专利的描述或者是正文，每行是一段话，包括一个或几个句子。

在系统中，利用了 openNLP 中的句子分割工具，并进行了一些调整。由于标点符号使用的模糊性，探测句子的边界具有一定程度的挑战性。在英文文档中，“.”并不一定代表一个句子的完结，有可能是单词的一个部分。为了使句子分割能够有更好的表现，在编写系统代码的同时，预先收集好了一些术语缩写集合，例如 *var., e.g., sp.* 等。特别是在标注文档中，许多实体词条包含半角英文句点，例如 *"EC 3.4.24.11"* 、*"MgCl2.6H2O"* 等。因此，在系统中构造了一些规则，例如：一个句子以缩写或者逗号结尾，当前分句和下一分句可以合成一个新的句子。而每一个元数据特征（标题、摘要、发明人等）也被视为一个句子，因为其和正常语句相比较短，包含更少的实体信息，因此这样处理并不会导致新的问题的出现。

最后，对所有子文档进行分句处理，并把所有子文档的分句结果合并成一个大文档。每一行是原来的一个子文档，以文档编号开头，其后接一个制表符，而每个句子间用空格进行分割。每一行的格式如下：

fileID sentence. sentence.

分词同样利用了 openNLP 中的分词工具，并进行了一些调整。调整之后可以将上述分句结果分割成所需的合理词条。而如果只使用原始的分词功能，将只会得到很差的结果，而且无法将分割后的词条应用在序列标注问题中。

根据语料库的特点，修改了分词的规则。包括但不仅限于以下几个方面：纠正拆开的计量单位；分离单个的数字；分离数字和单位；处理特殊符号；纠正特定前缀后缀的错误（主要针对盐类化合物）；一些复数形势；金属元素；其他。对于其中的一些规则，需要整理出相应的列表，以及一些常见的药物前缀等（如 anti, nano, dietary，"μ" 等）、计量单位（g,m,ml,mol），都对结果有影响。例如，盐类会整理出一个类似的列表：

```
private final static String[] aminoAcid3 = { "Ala",
"Arg", "Asn", "Asp",
"Cys", "Gln", "Glu", "Gly", "His", "Ile", "Leu",
"Lys", "Met",
"Phe", "Pro", "Ser", "Thr", "Trp", "Tyr", "Val" };
```

运用这些规则后，系统能够得到相对较好的细粒度的分词结果。分词的结果中，除了分开单词，一般还把标点符号（包括 - ，（ ）｛ ｝[ ]. 等）、数字、希腊字母等都分开，例如，在文档编号为 *US5650521_0003* 的文档中，实体 "(S)-(-)-α ,α -diphenyl-2-pyrrolidinemethanol" 被标注为类别 M，意为“IUPAC”，其分词结果和标注结果如表格 2 所示：

表格 2 化合物实体分词示例

token	...	(	S	)	-	(	-
label	*O*	*-B*	*-I*	*-I*	*-I*	*-I*	*-I*
token	)	-	α	,	α	-	dipheny
label	*-I*	*-I*	*-I*	*-I*	*-I*	*-I*	*-I*
token	-	2	-	pyrrolidinemethanol			....
label	*-I*	*-I*	*-I*	*-E*			*O*

而且，每一个嵌套的实体，最后都统一成单个类别的实体，例如在文档编号为 *US5650521_0003* 的文档中：

T109 D 4726 4738 siruvastatin

T343 OCRERRORSPELL 4726 4738 siruvastatin

然而，有些“*OCRERRORSPELL*”实体只含有一个类别，这意味着只有一些是嵌套的实体，而其余不是。因此，如前文所述，将最后的标注文档统一为单类别的实体，避免出现嵌套实体；最后，移除了 488 个嵌套实体。

3.2 识别：基于条件随机场的方法

如前所述，我们将化合物和生物实体识别视作序列标注问题（如表格 2）。条件随机场作为建立概率模型进行分割和标签序列的数据一个框架，提供了隐马尔科夫模型和随机语法的优点。条件随机场也避免基于定向图形模型的 MEMMs（maximum entropy Markov models，最大熵马尔可夫模型）和其他判别式马尔可夫模型的基本限制。

条件随机场模型会将上下文考虑在内。例如，在自然语言处理任务中，对于一系列的输入集，线性链条件随机场能预测其对应的标签序列。对于给定的观测值 和随机变量 ，变量 随 的变化而变化，并服从条件概率 。由于拟牛顿方法能很好地求解多项式方程，CRF++ 采用 L-BFGS(Limited-memory BFGS (Broyden-Fletcher-Goldfarb-Shanno)) 处理大规模优化问题，解决了参数估计的无约束优化。另一方面，CRF++ 利用线性搜索的方法计算无约束优化的步长。在专利语料中，实体被分成 14 个类别中的一个：

C={M,I,Y,D,B,C,F,R,G,T,Disease,MOA,OCRERRORSPELL, OCRERRORLINE}

一般采用 B/I/E/O 的四标签法来对实体进行序列标注，分别表示：实体开始、实体内的词、实体结束、其他。如 3.1 节所述，由于 CRF++ 无法处理多标签的实体，嵌套的实体都被统一成一个类型，使实体和类别一一对应。

3.3 CRF 特征选取

系统利用了四个不同类型的特征。

一般语言学特征。包括了原始词条、进行词干还原后的词条（利用斯坦福大学 CoreNLP 的 Porter 词干器进行处理）。

字符特性。因为许多实体包含数字、希腊字母、罗马数 (Roman)、氨基酸 (Amino Acid)、化学元素 (element) 和特殊字符，系统为每个词条计算一些统计特性，包括其数字的数量（num of digital）、大写或小写字母的数量（Num of upper/lower case letter），所有字符数量和是否存在特定的字符或希腊字母、罗马数字、氨基酸或化学元素。

例模式特性。分别用“A”，“a”，“0”替代原文中的大写字母、小写字母、数字 (0 - 9)。此外，系统也合并连续字母和数字，生成额外的字母“a”和数字“0”。

上下文的特性。对于每一个词条，系统包括了两个相邻词条的语言特征，包括布朗聚类（brown clustering）。

例如，对于一个词条，其特征如下（见表 3）：

表格 3 实体特征示例

stemmer	Amino Acid	Element	Symbol
lymphocyt	true	true	false
Roman	Num Of Digitals	Num Of Upper Case Letters	Num Of Lower Case Letters
false	0	0	11
length	case Pattern	brown	label tag
11	aaaaaaaaaa	111011101100	-I

3.4 后处理：基于规则的方法

在实验中，发现仅用条件随机场的方法会获得一些假阳性（FP，False Positive）的结果。为了获得更好的效果，还对识别的词条进行了基于规则的调整。在专利文本中，难免会出现一些不常见的术语缩写和符号错误（如括号缺失），因此需要利用一些必要的规则来提升效果，包括对错误结果的移除和对实体偏移量的调整。但仍然存在一些错误：

在文档 *EP1481667_0004* 中，实体 "dopamine receptor" 出现了两次，但只标注了一次。在文档 EP1481667_0004 中，实体 "ACE inhibitors" 被分开标注为两个实体；在文档 *WO2004000294_0004* 中则视为单个实体。本文认为这违反了 Akhondi 在文献中“标注指南”的第一条规则 [8]：如果一个实体是嵌套或者和其他实体重叠，应该标注为特定的、包含更多信息的实体。而一些实体，诸如 "AMcAMP", "IcAMP" (Abbreviation), "amino acids", "agonist", "methane sulfonic acid"，在一些文档中未被标明实体，本文认为也会对实验结果造成一定影响。

3.5 详细流程

系统的详细流程如图 3 所示。

4 实验

专利语料包含两个数据集：统一语料集（harmonized corpus）和全语料集（full corpus）。此外还提供一个只含两篇专利文档的训练集（training set）。在对训练集和统一语料集进行分析之后，在统一语料集中发现了一些化合物和生物实体具有两个或者多个标注，具有嵌套关系。经过一些相对合理的处理，使每个实体都只具有一个标签，视作序列标注问题，使得 CRF++ 能够进行处理。之后，我们将预处理过的专利语料的原文和标注文档都存入 MySQL 数据库中进行实验处理。每一篇子文档都作为表中的一条记录处理，句子间用空格分割。被标注的实体存储在另一张表中，每个实体有一个唯一的 ID，保存了每个实体的分类信息、文档中的偏移量、所属文档等。在十折交叉验证中，数据集被分为十份，其中一份作为测试数据其余九份作为训练数据。每一轮的训练集大概含有 12,000 个句子和 500,000 个特征。

CRF++ 有四个主要参数："−a", "−c", "−f" , "−p"，用于控制训练条件。参数“−f ”设置特征的最低阈值。CRF++ 使用训练数据中至少 NUM 次出现的特征。默认值为 1。参数“−a”是规范化算法选择。默认是 CRF−L2。一般来说 L2 算法效果要比 L1 算法稍微好一点，虽然 L1 算法中非

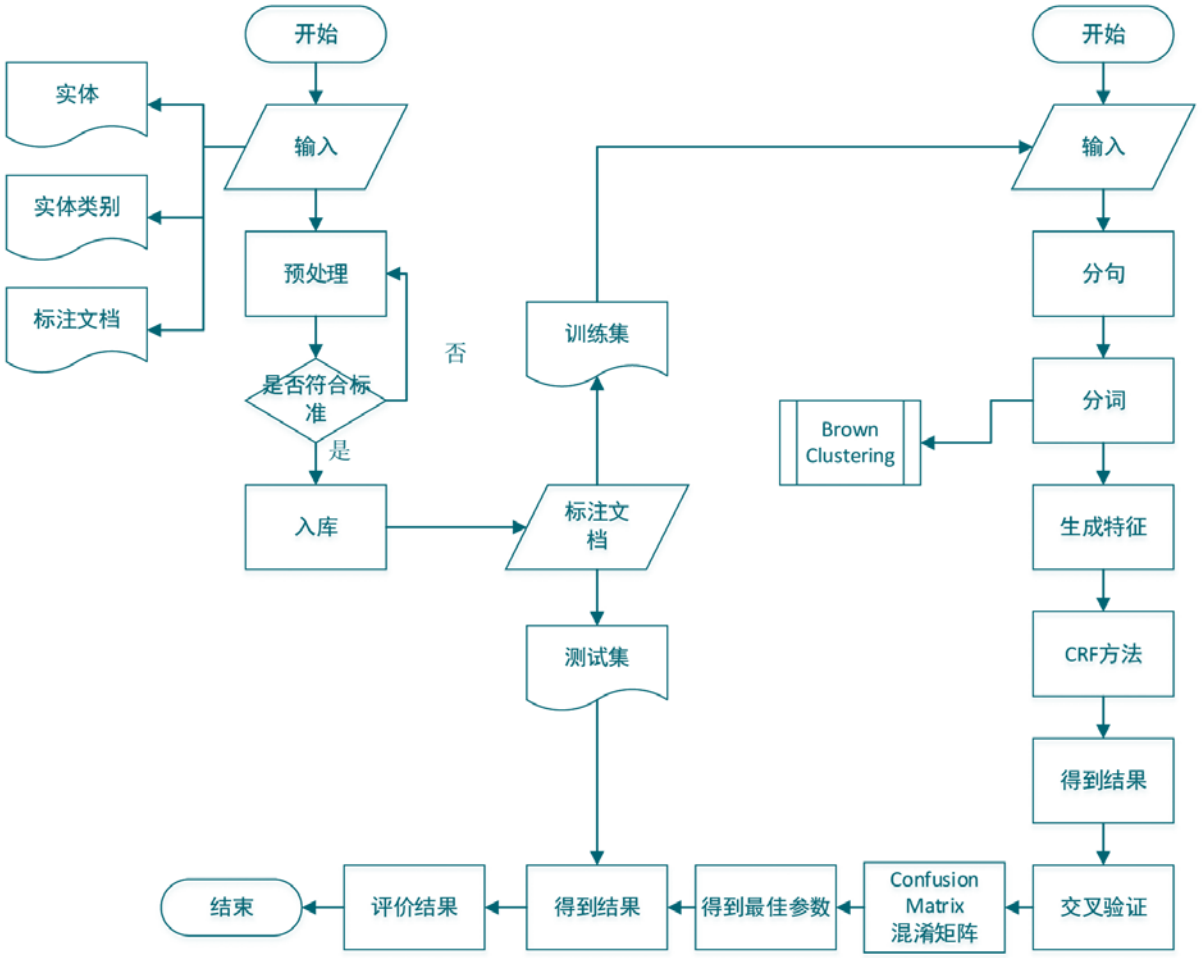

图 3 系统详细流程图

零特征的数值要比 L2 中大幅度的小。“−c”设置 CRF 的惩罚参数。c 的数值越大，CRF 拟合训练数据的程度越高。这个参数可以调整过度拟合和不拟合之间的平衡度。这个参数可以通过交叉验证等方法寻找较优的参数，且对实验结果有显著影响。“−p” 设置的是线程数量。由于实验时间限制，将“−c”的参数值设置为：$\{2^{-2}, 2^{-1}, 2^{0}, 2^{1}, 2^{2}\}$，根据“−c”的不同进行了五轮十折交叉验证。

系统使用布朗聚类 [14]（Brown Clustering）提高识别效果。布朗聚类是一种自底向上的聚类模式。通过基于分类的语言模型，利用基于文本对数概率的合并标准将词语按照二叉树进行组织分类。最后，按照是否有布朗聚类以及聚类数的不同生成了五个 CRF++ 的模板文件（分别为：无布朗聚类，500、1000、1500、2000 个布朗聚类），在此基础上进行实验。

然而实验结果没有预想中那么好（见表格 4）。在类似的实验中，例如 BioCreative IV 的 CHEMDNER 实体识别任务，官方的平均准确率为 89.21%，召回率为 66.41%，F1 值为 76.11%。由于系统自身的原因，也可以有一些因素影响了实验结果。利用论文语料进行实验的研究，往往采用论文的标题、摘要、关键字作为研究对象，噪声数据更少。而在这次实验中，我们采用了专利的全文。一般来说，专利会注重知识产权的保护，而论文更多是知识的传播和分享。而为了保护知识产权和创新性，专利文献会用特殊的方式进行书写；相反，论文作者会选择读者更容易理解的

表格 4 在专利数据集上的实验结果表现

	Run 1	Run 2	Run 3	Run 4	Run 5
best cost	21	21	20	21	21
TP	28981	29655	29473	29502	29451
TN	10517	15262	15626	15568	15668
FP	16607	11131	10790	11027	10875
FN	0	0	0	0	0
Precision (%)	63.57	72.71	73.20	72.79	73.03
Recall (%)	73.37	66.02	65.35	65.46	65.27
F1 score (%)	68.12	69.20	69.05	68.93	68.94

方式。这在一定程度上也增加了对专利文献进行解读的难度。

对于基因、蛋白质、疾病三类实体，如果在训练集中不存在或者说未被观测到，则很难确定实体类型，而且必须依靠上下文进行判断。化学实体具有很强的上下文特征。而化学实体的边界错误问题也是导致 FP 结果过多的重要原因，在训练集中实体 "ACE inhibitors" 被分开标注为两个实体；在文档 WO2004000294_0004 中则视为单个实体，这些都属于实体的边界重叠问题。而在边界错误中，影响最大的则是“改性剂”，例如，实体多环芳烃（polycyclic hydrocarbon）会返回“多环的”（polycyclic），因为在训练集中“polycyclic”标注为实体的情况多于其他多环化合物实体。另外，诸如化学式或者 IUPAC 等系统名称，则可能是化学名词的各种组合，较难判断边界。而通俗名称（TRIVIAL）相对较短，具有明显界限，因此识别准确率较高。另一方面，化学式的注解和化学式往往是通指的，虽然不会影响实体识别，但对它们之间的关系识别会有一定的影响，这种注解式的写法，往往会使得两个在现实中指代同一事物的共现关系变得很突出。而化学式及其所属族也会造成一定影响。例如，化学式 "2-acetamido-3-mercapto-3-methyl-N-aryl-butanamides"，标注为 Formula，且注解为 phthalate，事实上，两个化学式为上下位的关系。词典法利用编制词典，将化学式和通俗名称进行一一对应，这与基于规则的方法类似。但基于规则的方法在识别 IUPAC、SMILE 或者 CAS 号时能进行直接识别，但需要较大空间存储规则库。无监督的机器学习方法也能在一定程度上降低边界模糊的影响 [5]。

另外一种造成 FP 结果的是蛋白质的多肽组成。在化学专利中，出现蛋白质往往会描述其氨基酸序列。由于测试数据集为 CHEMDNER 提供的数据集，数据集从组成结构的角度来定义化合物，对于含肽链长度大于 15 的化合物将不会被标注，这使得对于一些蛋白质对应的肽链如果被标注，则有可能是 FP 的结果。比如毒素“kaliotoxin”，其氨基酸序列为 N - Gly - Val - Glu - Ile - Asn - Val - Lys - Cys - Ser - Gly - Ser - Pro - Gln - Cys - Leu - Lys - Pro - Cys - Lys - Asp - Ala - Gly - Met - Arg - Phe - Gly - Lys - Cys - Met - Asn - Arg - Lys - Cys - His - Cys - Thr - Pro - Lys - OH，肽链长度多达 40，这种短名称的蛋白质对应长氨基酸序列的现象在蛋白质中是很常见的。

5 结论

我们开发了化合物和生物实体识别系统，并用已标注好的化合物相关专利文档进行了实验。

本文将实体识别视作序列标注问题进行处理，而不是一次抽出整个实体。此外，系统还利用了一些开源的自然语言处理工具包，例如 OpenNLP、斯坦福大学的 CoreNLP。 为了使之符合专利语料的要求，系统在代码上做了一些修改，增加相应的规则。在序列标注问题的处理上，系统利用 CRF++ 作为具体实现，对实体进行序列标注。同时运用十折交叉验证的方法进行训练、测试和评价。如表格 4 所示，得到太多的假阳性（FP）的结果。某些实体在一些文档中被标注出来而另一些没有，这可能影响了系统的实验结果。在今后的实验中，将会设计更多适应语料库的规则来提升系统的效果。

参考文献

[1]Muresan S, Petrov P, Southan C, Kjellberg MJ, Kogej T, et al. (2011) Making every SAR point count: the development of Chemistry Connect for the large-scale integration of structure and bioactivity data. Drug Discov Today 16: 1019-1030.

[2]Southan C, Boppana K, Jagarlapudi SA, Muresan S (2011) Analysis of in vitrobioactivity data extracted from drug discovery literature and patents: Ranking 1654 human protein targets by assayed compounds and molecular scaffolds. J Cheminform 3: 14.

[3]De Ridder, D. et al. 2013. Pattern recognition in bioinformatics. Briefings in Bioinformatics. 14, 5 (Sep. 2013), 633-647.

[4]Grego, T. et al. 2009. Identification of Chemical Entities in Patent Documents. Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living, Pt Ii, Proceedings. S. Omatu et al., eds. Springer-Verlag Berlin. 942-949.

[5]Campos, D. et al. 2013. Gimli: open source and high-performance biomedical name recognition. Bmc Bioinformatics. 14, (Feb. 2013), 54.

[6]Han, H. et al. 2014. Mining technical topic networks from Chinese patents. 1st International Workshop on Patent Mining and Its Applications, IPaMin 2014, Co-located with Konvens 2014, October 6, 2014 - October 7, 2014 (2014).

[7]Roman Klinger, Corinna Kolarik, Juliane Fluck, Martin Hofmann-Apitius, and Christoph M. Friedrich, 2008. Detection of IUPAC and IUPAC-Like Chemical Names. Bioinformatics, Vol. 24, No. 13, pp. i268-i276.

[8]Akhondi, S.A. et al. 2014. Annotated Chemical Patent Corpus: A Gold Standard for Text Mining. PLoS ONE. 9, 9 (2014), e107477. DOI: 10.1371/journal.pone.0107477

[9]Zimmermann, M. et al. 2005. Information Extraction in the Life Sciences: Perspectives for Medicinal Chemistry, Pharmacology and Toxicology. Current Topics in Medicinal Chemistry. 5, 8 (Aug. 2005), 785-796.

[10]Hobbs J R. The generic information extraction system[C]//MUC. 1993: 87-91.

[11]Xu S, An X, Zhu L, et al. A CRF-based system for recognizing chemical entity mentions (CEMs) in biomedical literature[J]. Journal of Cheminformatics, 2015 (Suppl 1): S11.

[12]Wei, C.H., Harris, B.R., Kao, H.Y., Lu, Z.: tmVar: A text mining approach for extracting sequence variants in biomedical literature. Bioinformatics 129(11) (2013) 1433-1439

[13]Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: ICML'01. (2001) 282-289

[14]Turian, J., Ratinov, L., & Bengio, Y. (2010, July). Word representations: a simple and general method for semi-supervised learning. In Proceedings of the 48th annual meeting of the association for computational linguistics (pp. 384-394). Association for Computational Linguistics.

[15]Zhang S, Elhadad N. Unsupervised biomedical named entity recognition: experiments with clinical and biological texts. J Biomed Inform. 2013; 46(6):1088-1098.