



北京大學

本科生毕业论文

题目： 基于深度学习算法提取文本中有机反应

姓 名： 黄志贤

学 号： 1600011759

院 系： 化学与分子工程学院

专 业： 化 学

指导教师： 来鲁华 教授 裴剑锋 研究员

论文类型： 研究型

二〇二〇年 六 月

摘 要

利用计算机自动化提取专利文献中有机反应是化学信息学的重要课题,对于实现有机反应数据库高效、自动化的构建、更新与维护有着重要意义,同时也为人工智能技术应用于有机合成、药物合成提供数据上质与量的保障。本论文运用自然语言处理中命名体识别相关技术与深度学习算法,初步实现了一步识别段落文本中的有机反应。

论文工作首先确立了化学类文本的预处理方法。接着尝试了 Transformer (一种基于全多头注意力模块的算法)与 BiLSTM(双向长短期记忆, Bi-directional Long Short-Term Memory) + CRF (条件随机场, Conditional Random Field) 算法、不同标注模式与参数,最终搭建了双层 BiLSTM + CRF 的化学命名体识别 (CemNER) 模型, CemNER 在 CHEMDNER 与 CDR 数据集上 F1 score (准确率与召回率的调和平均值) 为 0.869。使用该模型处理了 USPTO 专利数据库的反应与文本数据,建立了有机反应命名体识别 (RxnNER) 数据集。目前 RxnNER 数据集包含了 37,787 条标注了反应物、产物与催化剂或溶剂三类反应命名体的段落文本。在 RxnNER 数据集上,我们尝试了不同算法与参数,最终搭建了加入预训练词向量的双层 BiLSTM + CRF 模型,实现了段落级文本中直接识别三类反应命名体,反应识别的 F1 score 为 0.862。最后用 OPSIN 与 Rdkit 后处理,将识别的反应命名体转化反应 SMILES 与相关信息。

在段落级文本识别有机反应的基础上,我们将文档级文本进行分段与简易解析,进而提取每个段落的有机反应,取得了初步成效。同时我们总结了文档级文本识别有机反应的困难,并提出相应的解决思路,为未来进一步的研究与开发作铺垫。

关键词: 文本挖掘 化学信息学 命名体识别 深度学习

ABSTRACT

With the continued growth of new publications, it is becoming increasingly difficult and costly to create and maintain up-to-date manually curated databases, and automated information extraction by machines is fast becoming a necessity. It is of great significance for the efficient, automated construction, update, and maintenance of organic reaction databases to develop a software system to automatically extract organic reaction. At the same time, it provides data of quantity and quality for developing artificial intelligence technology in the field of organic synthesis and drug synthesis. To solve this issue, some progress in this field has been made in the past 30 years. However, none of the software for organic reaction extraction is open source. So, it is essential for researchers in Cheminformatics and Organic Synthesis to develop a tool to extract organic reactions from texts in patents and literatures and then establish an organic reaction database of their own.

This project takes advantage of the related technology of Named Entity Recognition (NER) and Deep Learning algorithm in Natural Language Processing (NLP) to achieve the recognition of the organic reaction in a paragraph-level text through one core step, compared with those two-steps system developed before.

Beyond the core step of recognition, there are preprocessing step to deal with chemical text and post-processing step to convert chemical names to SMILES and assemble them into reactions.

Firstly, this project establishes a preprocessing method for chemical texts, including splitting sentences, tokenization, lowering, normalization and removing stop words, so as to remove meaningless words, function words and reduce vocabulary while maintain words and characters appearing in chemical entity mentions. After attempt for different algorithms like Transformer + CRF (Conditional Random Field) and BiLSTM (Bi-directional Long Short-Term Memory) + CRF, different tag modes and hyperparameters, a two-layer BiLSTM + CRF model for Chemical Entity Mentioned Recognition (CemNER) was built with precision of 0.869, recall of 0.869

and F1 score of 0.869 in chemical entities recognition task on CHEMDNER and CDR data sets. The CemNER model is built to process the patent data from USPTO with reactions and paragraph-level texts, recognizing the chemical entities in the texts and matching them with reactant, product and catalyst/solvent in reactions. After that, the patent paragraph texts were tagged with three types of reaction entities in the BIOES-style mode like (O, B-react, B-prod, B-catSolv, I-react, I-prod, I-catSolv) and were collected into the data set for Reaction Named Entity Recognition (RxnNER).

The RxnNER data set established in this project has 37,787 pieces of data and a vocabulary of 15,649 currently, which is focused on chemical entities mentioned with IUPAC nomenclature. On the RxnNER data set, we tried different algorithms and hyperparameters, and finally built a two-layer BiLSTM + CRF model with pre-trained word vectors to achieve the one-step recognition of three types of reaction entities in paragraph-level text. The RxnNER model we built achieved precision of 0.861, recall of 0.862 and F1 score of 0.862 in reaction entities recognition task on current RxnNER data set. In the future, RxnNER data set will be extended with more data to help improve the model by learning more features in organic reaction texts with low-frequency words and different sentence patterns.

After the recognition of reaction entities, OPSIN system and Rdkit module were used to post-process the prediction of RxnNER model. OPSIN system made help to convert chemical entities to SMILES and other chemical editable files. While Rdkit module assembled SMILES of reactants, products and catalyst/solvent into reaction SMILES and output the reaction information through pictures and textual records.

On the basis of RxnNER system of paragraph-level text we developed, we made attempt to tackle with RxnNER of document-level text through segmenting paragraphs and simple parsing the structure of chemical information, and achieved preliminary results which showed the potential of our model. At the same time, we have summarized the difficulties of RxnNER of document-level text, and put forward corresponding solutions to pave the way for further research and development in the future.

In a nut shell, we developed a RxnNER data set for reaction entity recognition and a system for extraction organic reactions in paragraph-level text. We also made attempt to extract organic reactions in document-level text and the demo had some promising result. Therefore, we will continue developing the document-level system in the future, and make use of the tool to establish organic reaction database of our own. Only by this way can we take full advantage of artificial intelligence technology to do further research in Cheminformatics and Organic Synthesis.

Keywords: Text Mining, Cheminformatics, Named Entity Recognition, Deep Learning

目录

1	背景及文献综述.....	1
1.1	化学分子信息识别.....	3
1.2	有机反应提取.....	5
1.3	命名体识别（NER）中的深度学习算法.....	6
2	数据和软件.....	8
2.1	有机反应数据集.....	8
2.2	美国专利数据库（USPTO）.....	8
2.3	CHEMDNER 与 CDR 数据集.....	9
2.4	开发环境与硬件.....	9
3	化学类文本预处理.....	10
4	有机反应命名体识别（RxnNER）数据集的建立.....	12
4.1	化学命名体识别（CemNER）数据集.....	12
4.2	CemNER 模型的搭建.....	13
4.3	CemNER 模型的优化.....	14
4.4	RxnNER 数据集的建立.....	15
5	有机反应命名体识别（RxnNER）模型的搭建、优化与应用.....	18
5.1	RxnNER 模型的搭建.....	18
5.2	RxnNER 模型的优化——加入预训练词向量.....	19
5.3	RxnNER 模型的优化——扩充数据集.....	21
5.4	RxnNER 模型误差定性分析.....	22
5.5	RxnNER 模型的应用.....	25
6	结论.....	29
7	参考文献.....	30
8	致谢.....	33

1 背景及文献综述

有机合成化学，作为现代制药与化工的基石，是化学领域最重要的分支之一，同时也是兼具规范性与创造性的一门学科。规范性，即有机反应的原理与规则；创造性，则是一位全合成化学家以个人的理解对于反应路线与反应本身进行创新改造。前者是需要通过学习大量已知反应来掌握规律，而后者则是在前者基础上进行灵感加工，偶然得到的成果。因此，全合成大师 K. C. Nicolaou 毫不夸张地称有机合成是一门“艺术”。

而在当今计算机技术飞速发展的时代，人工智能（Intelligence Artificial, AI）技术越来越多地被应用于各个学科领域。机器学习（Machine Learning）算法，其本质就是基于大数据的统计学模型，可以借助现代计算机的强大算力，学习到已知数据与已知结果之间的某种规律或关系，从而能够预测新数据的结果。而深度学习（Deep Learning）作为机器学习的一个重要分支，是一种模仿生物神经元连接模式，利用大量隐层单元与丰富的连接结构实现强大拟合能力的算法模型。它强大的特征提取能力，在自然语言处理（Natural Language Processing, NLP）与计算机视觉（Computer Vision, CV）两大领域发挥重要作用。

而在有机化学领域，人们尝试利用计算机来辅助有机合成，从最早 1969 年 Coery, E. J. 等人开发的基于人为制定反应规则与模板的逆合成辅助工具——Logic and Heuristics Applied to Synthetic Analysis (LHASA)¹ 开始，到后来出现基于机器学习与深度学习算法来学习反应数据中的特征，从而实现无模板的预测模型。相较人为制定反应规则与模板的模型，这种基于机器学习算法的模型泛化能力更强，对于未知模板的新反应有了更强的预测能力。例如 2020 年北京大学的鲁华教授课题组开发的 AutoSynRoute²，就是基于 Transformer 模型预测逆反

应,再辅以蒙特卡洛搜索算法对路线进行选择,从而实现自动化的化合物有机合成路线规划。其工作流程如图 1.1 所示。

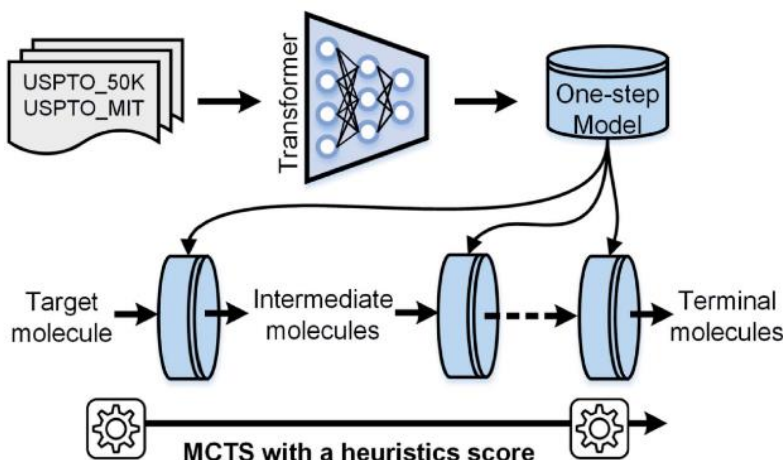


图 1.1 AutoSynRoute 工作流程示意图²

而对于基于机器学习与深度学习的模型,想要学习到有机反应的规律与特征来预测反应,就需要收集大量反应数据来训练模型。不仅如此,在有机化学不断进步发展的趋势下,新机理与新反应源源不断地被发现,因此有机反应的数据库在不断地更新,其中更新的反应对于预测模型来说是宝贵的“学习资料”,为模型提供新的特征进行学习。因此优质、丰富和及时更新的反应数据是将人工智能技术应用于化学信息学的基础和根本,也是化学领域搭上信息化、大数据化时代顺风车的必由之路。

目前已有的大型有机反应数据库,例如 Reaxys、Scifinder,它们的有机反应数据并不完全开源,只提供给有机化学家进行单个反应搜索,而不能大批量地爬取反应数据,因而无法为有机合成方向的机器学习提供充足的训练数据。

另一方面,目前反应数据库重要来源是相关期刊文献以及专利,传统方法是采用人工收集反应数据,需要耗费大量专业人力资源,并且在复杂化合物的识别上也容易出错。而随着新文献、专利的日益增长,人工更新和维护反应数据库变得越来越困难,因此通过计算机实现自动提取反应信息已经变得越来越有必要。而目前已经开发出来的自动提取有机反应的方法还远未完善,也没有开源。

本课题旨在借助深度学习算法来提取文本中的有机反应,为将来自动化搭建有机反应数据库,进行有机化学合成与反应研究的大数据和 AI 化打下基础。

1.1 化学分子信息识别

1.1.1 图像识别

想要从非结构化的期刊文献与专利文献中提取有机反应信息，首先要识别出一篇文献中的化学分子。在此有两种方式，一种是对文章中分子结构的图片进行图像识别，另一种是对文字内容进行文本识别。前者实现将文档内分子结构图像分割出来后，对分子结构图进行原子、化学键的识别，然后组装原子与化学键，得到分子的结构化文件，例如 OSRA³, ChemEx⁴, ChemInfty⁵, AsteriX⁶ 等软件都是基于不同图像识别算法实现分子结构图像识别。

1.1.2 文本识别

相比较图像识别，文本识别对于有机反应的提取更加重要，因为在绝大多数的期刊文献与专利中，有机反应信息是以文字形式记录，而且除了少数直接以反应式表述，更多的是以语句表述反应过程，或者在专利文本中常以反应实例进行反应操作流程的表述。本研究的重点是提取文本中反应的原始资料。

而想要提取有机反应，传统思路是先识别出文本中的化学命名体（Chemical Entity Mentions, CEM），再对识别出来的化学命名体进行解析分类，确定反应物、产物、溶剂、催化剂等。

识别化学命名体，其实就是自然语言处理中（Natural Language Processing, NLP）的命名体识别（Named Entity Recognition, NER）在化学中应用的问题，NER 指在一段文字中，识别出特定名称，例如地名、人名、机构名等。而识别命名体是通过对每个词进行标注，常见有 BIO 与 BIOES 两种标注模式⁷。其中 B 是指 Begin，即词头；I 指 Inside，即词中；O 是指 Other，即其他词；E 是指 End，即词尾；S 是指 Single，即单个词。两种标注模式差别在于是否标注词尾以及单个词，例如 PL 137,526 describes the hydrogenation of p-tert-butylphenol to form p-tert-butylcyclohexanol using a nickel catalyst. 其两种标注模式下的识别示例如下表 1.1 所示。

表 1.1 化学命名体识别实例

文本	BIO	BIOES
pl	O	O
137	O	O
,	O	O
526	O	O
describe	O	O

hydrogenation	O	O
of	O	O
p	B	B
-	I	I
tert	I	I
-	I	I
butylphenol	I	E
to	O	O
form	O	O
p	B	B
-	I	I
tert	I	I
-	I	I
butylcyclohexanol	I	E
use	O	O
a	O	O
nickel	B	S
catalyst	O	O
.	O	O

注：表格中文本已预处理（分词、小写化、标准化、删除停用词）

对文本中的每个词进行标注后，就可以确定化学命名体的边界，从而得到文本中化学命名体的内容与位置。而这一步命名体识别一般有两种算法来解决，一种是人工制定规则的方法，例如 *Lowe, D. M.*等人开发的 *LeadMine* 软件，就是通过人为制定目标命名体的匹配模式，同时建立目标词汇的词典来搜索与匹配⁸。另一种方法是用机器学习算法，即训练一个文本标注的模型，常见的算法有条件随机场算法（Conditional Random Field, CRF），朴素贝叶斯算法（Naive Bayesian, NB）等⁹⁻¹¹。

在 2015 年，*Krallinger, M.*等人为了推动化学医药自然语言处理的发展，创建了一系列命名体识别的任务与数据集，例如 *CHEMDNER* 是一个标注好化学命名体的文本数据集¹²。在该数据集上，基于规则的 *LeadMine* 方法得到 88.7% 的准确率、85.1% 的召回率与 86.9% 的 F1 score⁸；*Xu, S.*等人基于 CRF 算法的模型得到 88.8% 的准确率、79.1% 的召回率和 77.7% 的 F1 score⁹；*Leaman, R.*等人开发的基于支持向量机（SVM）、CRF、命名规则的 *tmChem*，在化学命名体识别任务中获得最高的 F1 score 87.4%¹³，具体信息列于表 1.3。

在识别化学命名体的基础上, Swain, M. C.等人开发的 ChemDataExtractor 软件实现了对化学文献专利中表征信息与表格信息的解析与提取,其工作流程示意图如下图 1.2 所示¹⁴。该软件基于 CRF 算法与词聚类实现化学命名体的识别, 再对文本句法进行解析, 将信息整合后, 得到文献中的有机化合物对应的理化、药化信息, 因此它对于自动化构建有机化合物的信息数据库有着重要意义。

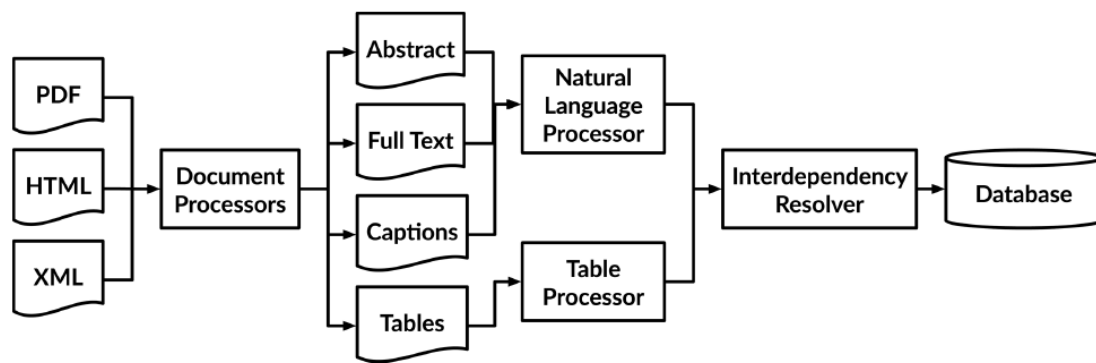


图 1.2 ChemDataExtractor 工作流程示意图

1.2 有机反应提取

1983 年至 1990 年, Blower, P. E.的课题组在自动化提取有机反应相关课题上取得进展, 他们通过识别文本中的化学命名体, 然后人工对特定动词设定相应框架, 从而匹配对应的化学命名体为反应物、产物、溶剂与反应条件¹⁵⁻¹⁸, 并最终在 50 篇文献中得到了 78% 的准确率。

2011 年 Jessop, D. M.等人开发了 PatentEye 用于提取专利文本中的反应¹⁹。这个软件分为两部分, 首先是用 ChemicalTagger²⁰ 实现化学命名体的识别, 接着用 OSCAR²¹ 基于特定词汇模式对文本进行句法解析与分类, 来判定每个化学命名体在反应中的角色, 最后得到 78% 的准确率与 64% 的召回率。

之后 2012 年, Lowe, D. M.在前者基础上, 优化了文本预处理流程, 例如文本段落分类、分句、分词, 同时提高化学命名体的识别率, 从而提升反应提取的准确率与召回率²²。

表 1.2 给出了目前已有的自动提取有机反应的软件概况。

表 1.2 目前已有的自动提取有机反应的软件概况

软件	时间	开发者	原理
/	1983-1990	Blower, P. E. et al. ¹⁸	化学命名体识别+设定动词框架
PatentEye	2011	Jessop, D. M. et al. ¹⁹	化学命名体识别+解析句法进行分类
LeadMine	2012	Lowe, D. M. ²²	化学命名体识别+解析句法进行分类

由以上内容可见, 过往在提取有机反应时, 都进行了两步识别, 先人为设定化学命名体的匹配式, 或者应用传统机器学习算法来识别化学命名体, 然后再基于语法框架解析语句成分, 从而找到每个化学名称在有机反应中的角色。以上报道的自动提取有机反应的软件的准确性还有待提高, 其元件也未开源无法获得并使用。

1.3 命名体识别 (NER) 中的深度学习算法

目前, 随着深度学习算法在自然语言处理各类任务中不断推陈出新, 其模型表现越来越好, 得到更多的青睐。在处理像命名体识别一类的自然语言处理问题时, 深度学习中的循环神经网络 (Recurrent Neural Network, RNN) 能够传递上一时刻的信息给下一时刻, 从而能够较好地处理语言类的序列问题。

长短期记忆网络 (Long Short-Term Memory, LSTM) 是在 RNN 的基础上, 在序列每个元素加入一个信息控制门, 控制哪些信息保留, 哪些不保留, 增加有用信息对于后续元素的影响, 减少长序列导致的信息丢失, 因而比 RNN 模型能更好地处理长序列问题²³。

而双向长短期记忆网络 (Bi-directional Long Short-Term Memory, BiLSTM) 是在 LSTM 的基础上, 再加上反向序列的 LSTM, 使得序列中每个元素能够得到前后双向的信息, 因此比 LSTM 在依赖上下文的任务中表现更好²⁴。

除了 RNN 及其衍生算法, 2017 年谷歌提出的 Transformer 模型在近几年表现越来越耀眼²⁵。它抛弃了 RNN 序列式的结构, 采用全多头注意力模块来学习序列特征。在提升了模型表现的同时, Transformer 的可并行性大大提高了训练速度。加之近几年 Transformer 通过增加对循环深度的控制、增加预训练, 衍生出许多变体, 一次次刷新自然语言处理各项任务的最高分²⁶⁻²⁷。

在命名体识别任务中, 只用 RNN、LSTM 或者 Transformer 等算法, 预测得到的标注概率只考虑了文本词汇上下文关系, 没有学习到标注的上下文的关系, 因此预测容易出现标注不合理, 例如在词头预测为 I、在词尾预测为 B。此时在这些深度学习算法的基础上辅助以 CRF 算法, 即预测时增加标注之间的转移概率, 从而避免出现不合理的标注。这种 LSTM、BiLSTM、Transformer 与 CRF 的结合算法, 在命名体识别任务中比单个算法表现更为出色。例如 Luo, L. 等人利用 BiLSTM-CRF 算法并且在其中加入预训练的词嵌入与注意力机制, 在 CHEMDNER 数据集上进行化学命名体的识别, 得到 91.14% 的 F1 score²⁸。下表 1.3 为 CHEMDNER 数据集上各算法表现。

表 1.3 各算法在 CHEMDNER 数据集上表现

算法	作者	F1 score
支持向量机 (SVM) + 条件随机场 (CRF) + 命名规则	Leaman, R. ¹³	0.874
命名规则 + 字典	Lowe, D. M. ⁸	0.869
最大熵 + 条件随机场 (CRF)	Xu, S. et al. ⁹	0.777
注意力机制的双向长短期记忆网络 (AttBiLSTM) + 条件随机场 (CRF)	Luo, L. et al. ²⁸	0.911

在 CHEMDNER 的化学命名体识别任务上，第一个算法结合了机器学习、规则，其表现比后两个基于规则、基于机器学习的算法表现更好。最后一个基于机器学习与深度学习的算法表现最好。

本课题的重点是应用自然语言处理中的深度学习算法，跳出传统两步识别的思路，一步实现有机反应命名体（反应物、产物、催化剂或溶剂）的识别（如下图 1.3 所示），从而提取文本中的有机反应，同时也为此类问题建立一个相应的有机反应命名体识别的数据集（RxnNER data set）。旨在探索不同深度学习算法与参数在化学命名体识别中的表现，从而搭建一个高效而泛化能力强的模型，实现从大量文献专利中提取有机反应数据，进而实现反应数据库的自动化搭建、更新与维护，为有机化学领域信息化、大数据化提供基础。

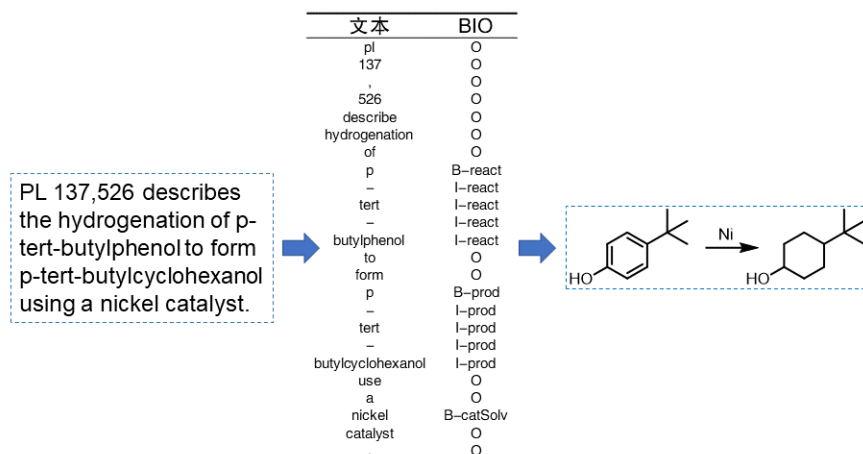


图 1.3 通过有机反应命名体识别一步实现文本反应提取流程示意图

2 数据和软件

2.1 有机反应数据集

本课题旨在从文本中提取有机反应，因此需要文本与对应的有机反应作为数据集让我们的模型学习其中语法规则。2017 年，LeadMine 的作者 *Lowe D. M.* 用该软件处理了美国专利数据库（USPTO）中化学相关专利的有机反应，以有机反应的简化分子线性输入规范（Simplified molecular input line entry specification, SMILES）进行表示²⁹。其主要信息如下表 2.1 所示。其中开源的 2001 年至 2016 年的有机反应约为 190 万条。

表 2.1 有机反应数据示例

Reaction SMILES	Patent ID	Paragraph Number
<chem>[C:1]([C:5]1[CH:10]=[CH:9][C:8]([OH:11])=[CH:7][CH:6]=1)([CH3:4])([CH3:3])[CH3:2]>[Ni]>[C:1]([CH:5]1[CH2:6][CH2:7][CH:8]([OH:11])[CH2:9][CH2:10]1)([CH3:4])([CH3:2])[CH3:3]</chem>	US200100 00035A1	0007

由上表可知，每一条反应会有对应的专利号与专利文本段落编号，因此为后续搭建有机反应命名体识别（RxnNER）数据集提供重要信息。

2.2 美国专利数据库（USPTO）

本课题选取的有机反应相关英文文本来自开源的美国专利数据库（USPTO）的专利 XML 文档。相比于其他有机化学期刊、有机化学会议的文献 PDF，专利文档的格式更为规范化、统一化，同时专利 XML 格式文档为树状信息组织结构，每一部分文本都有相应标签、属性、编号，便于大规模提取文本时进行索引，而 PDF 格式的文档在转化为线性文本时容易发生结构差错与文本错误。下表 2.2 为三种文本来源的优缺点比较。

表 2.2 各类文本来源的比较

	专利	科研期刊	其他（会议报告等）
格式统一度	高	中	低
内容获取成本	免费全文	需要购买	部分免费，部分收费
数据格式	PDF, XML	PDF	PDF

利用 2.1 中的有机反应数据与 USPTO 开源的专利 XML 文档，就可以得到有机反应与对应的文本段落，如下表 2.3 所示，这也是构建 RxnNER 数据集的原始资料。

表 2.3 有机反应与对应文本段落示例

Reaction SMILES	Paragraph
<chem>[C:1]([C:5]1[CH:10]=[CH:9][C:8]([OH:11])=[CH:7][CH:6]=1)([CH3:4])([CH3:3])[CH3:2]>[Ni]>[C:1]([CH:5]1[CH2:6][CH2:7][CH:8]([OH:11])[CH2:9][CH2:10]1)([CH3:4])([CH3:2])[CH3:3]</chem>	PL 137,526 describes the hydrogenation of p-tert-butylphenol to form p-tert-butylcyclohexanol using a nickel catalyst.

注：反应 SMILES 中的数字是 LeadMine 为了校对反应进行反应物、产物原子匹配时留下的信息

其中反应的 SMILES 是用 “>” 分隔反应物、催化剂或溶剂、产物。两个 “>” 左侧是反应物，右侧是产物，中间是催化剂或者溶剂。若同一侧有多个化合物，则用 “.” 来分隔。

2.3 CHEMDNER 与 CDR 数据集

CHEMDNER (BioCreative IV Chemical Compound and Drug Name Recognition)¹² 是 Martin Krallinger 等人在 2015 年公布的一个数据集。该数据集由 10,000 份 PubMed 上的文献摘要组成，其中涉及的 84,355 个化学命名体全部都是由化学领域专家人工标注出来的。该数据集主要用于进行化学命名体识别 (CEM) 与化学文档检索 (CDI) 两项任务的研究。

而 CDR (BioCreative V chemical-disease relation)³⁰ 则是 2016 年公开的数据集，其中包含 1,500 份 PubMed 上的文献摘要，标注了 4,409 个化学命名体、5818 个疾病命名体、3116 个化学-疾病相互作用命名体。因此也可以作为化学命名体识别任务的数据集。

2.4 开发环境与硬件

本课题的程序是在 Windows 10 的 PyCharm IDE 下进行开发，具体开发环境及所用的关键模块如下表 2.4 所示。

表 2.4 开发环境主要软件与模块

软件与模块	版本
anaconda	4.8.2
python	3.6.10
tensorflow-gpu	2.1.0
keras-gpu	2.3.1
keras-contrib	2.0.8
keras-transformer	0.1
gensim	3.8.0
nlTK	3.4.5
numpy	1.18.1

pandas	1.0.3
rdkit	2020.03.1.0
scikit-learn	0.22.1
chemdataextractor	1.3.0

3 化学类文本预处理

文本预处理方法根据文本内容、任务要求而有所区别。在本课题中，所使用的是英文文本，任务为命名体识别，同时由于目标命名体是化学类名词，因此在部分预处理细节上会与常规文本预处理有所不同。

分句:对于段落级文本,首先要做的是分句,使得段落文本可以划分为序列更短的句子,一方面便于化学命名体识别模型的训练,另一方面可以帮助我们删除部分与有机反应描述无关的文本(例如核磁质谱表征结果),减少对模型学习与预测的干扰。不同于中文文本的句号“。”,英文文本的句号“.”容易与缩写简称或者化学名词(螺烷)的“.”混淆,例如“U.S.”、“1,4-Dioxaspiro[4.5]decane”。因此分句采用的匹配方式是“.”前不是一到三个首字母为大写的字符串,而“.”后有空格且下一个字符是非小写,用正则表达式表示为:(?<![A-Z]\.)(?<![A-Z][a-z]\.)(?<![A-Z][a-z]{2}\.)(?<=\.)\s+(?=[A-Z0-9\(\ \{\}\/])(?<=?\?!\\:\.):\s+

的正则表达式为: $\backslash w+[\wedge\backslash w\backslash s]$ 。在这种介于单词级别与字符级别之间的分词模式下, 化学名词尤其是系统命名, 可以被分为多个官能团、数字、符号, 例如 4 - chloro - 1 - (4 - isopropyl - phenyl) - butan - 1 - one, 而我们的模型的词汇量只需要包含这些官能团、数字、符号, 即可组合出更多的化合物系统命名。

小写化: 对于英文文本, 句首的首字母大写也会给模型带来额外的词汇量, 因此需要在分词后进行文本小写化处理。这一步在极少数情况下, 可能会带来化学信息的丢失造成歧义, 例如 CO 与 Co 在小写化之后变成同一个单词。鉴于此类情况极少, 而小写化带来的词汇量减少的增益更可观, 因此依然采用小写化处理方法。

标准化: 英文文本还有一个特殊的地方就是名词会出现复数形式, 动词有各种时态, 这无形中也会增加模型不必要的词汇量, 因此我们考虑将名词动词复原, 采用的方法是利用 CHEMDNER 中已有的单词复原对应关系, 建立单词复原字典, 这样能解决大部分的常用单词复原问题。对于我们课题中采用的 XML 文档, 早期 2001~2004 年的文档中会出现无法直接编译的 HTML 命名体, 它们往往是一些标点符号和特殊符号、希腊字母等, 需要在文本中转换成 UTF-8 编码的字符串。

删除停用词: 最后一步是删除停用词, 也就是删除文本中部分介词、be、冠词等虚词以及标点符号, 保留实义词, 从一定程度上减少词汇量, 同时缩短序列长度。而对于化学类文本, 由于分词模式会导致文本中会出现 “,” “.” “-” “{” “[” 等符号, 以及 “s”, “r” 等单字符, 而这些事化学命名体的一部分, 需要保留。

有机反应段落文本预处理效果如下表 3.1 所示。

表 3.1 化学类文本预处理效果示例

原始文本	预处理后的文本
A round bottom flask was charged with 4-iodobenzoic acid (3.0 g, 12.1 mmol) in chloroform (100 mL). To the solution was added thionyl chloride (5.0 mL) in chloroform (10 mL) and 2-3 drops of dimethylformamide (DMF). The slurry was heated at reflux for 2 hours, while monitoring the reaction through an oil bubbler. A clear solution of 4-iodobenzoyl chloride was obtained. The volatiles were removed and a colorless oil was obtained which solidified upon cooling.	a round bottom flask charge 4 - iodobenzoic acid (3 . 0 g , 12 . 1 mmol) in chloroform (100 ml) . to solution add thionyl chloride (5 . 0 ml) in chloroform (10 ml) and 2 - 3 drop of dimethylformamide (dmf) . slurry heat at reflux 2 hour . monitor reaction oil bubbler . a clear solution of 4 - iodobenzoyl chloride obtain . volatile remove and a colorless oil obtain solidified upon cool .

4 有机反应命名体识别 (RxnNER) 数据集的建立

已有的 USPTO 提取的有机反应与对应文本段落是我们建立 RxnNER 数据集的原始资料。在此基础上,我们需要将文本段落中的化学命名体标注出来,并与有机反应中的反应物、产物、催化剂或溶剂相对应。确定每一部分都对应后,则对文本进行有机反应命名体的标注。在此我们主要考虑反应物与产物,二者都能对应上,则可以录入 RxnNER 数据集,对溶剂或催化剂不强制要求。

4.1 化学命名体识别 (CemNER) 数据集

建立 RxnNER 数据集,最重要的是需要先识别出段落文本中的化学命名体。在此我们使用 CHEMDNER (化学药物命名体识别) 数据集与 CDR (化学与疾病关系) 数据集来训练我们的化学命名体识别 (CemNER) 模型,实现句子级别的 CemNER 任务。

其中总共有 102,170 条语句,除去 (只出现一次的低频词) 词汇量为 33,260。我们将数据集按照训练集 (Train): 验证集 (Val): 测试集 (Test) = 8:1:1 进行划分,训练集用于让模型学习,验证集用于监测训练过程并作为模型调参的评价指标,而测试集用于最后衡量模型的泛化能力。

模型的表现评价标准通常有两个级别: tag 级与 cem 级。tag 级指的是模型预测 B、I、O 的表现; cem 级指的是模型预测化学命名体的表现,即将 B 与 I 拼接得到一个命名体,检验其是否为真实的化学命名体。而这两个级别都以准确率 (Precision)、召回率 (Recall) 以及 F1 分数 (F1 score, 前两者的调和平均数) 这三个指标来对模型表现进行评价。而 tag 级的指标数值高低不能准确反映模型是否预测出完整、准确的化学命名体,因而选择更具有实际意义 cem 级指标进行评价。同时本课题中 CemNER 模型需要兼顾准确性与召回率,因此最终选择以 cem 级的 F1 score 来衡量 CemNER 模型的表现 (如无特殊说明, CemNER 的模型表现默认为 cem 级)。计算准确率 (Precision)、召回率 (Recall) 以及 F1 score 的公式如下 (1~3) 所示。

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

4.2 CemNER 模型的搭建

参照 Luo, L. et al.²⁸ 的 AttBiLSTM + CRF 模型, 我们在此也采用 BiLSTM + CRF 模型, 而具体结构有所差别, 前者只用了单层 BiLSTM 与 CRF, 而在 BiLSTM 与 CRF 层之间加入注意力机制用以提高模型对于长序列的信息掌握能力; 而我们的 CemNER 模型采用双层 BiLSTM, 直接 BiLSTM 层得到的发射概率传递给 CRF 层, 进而计算标签之间的转移概率, 找到最优标注路径。我们的 CemNER 模型结构如下图 4.1 所示。

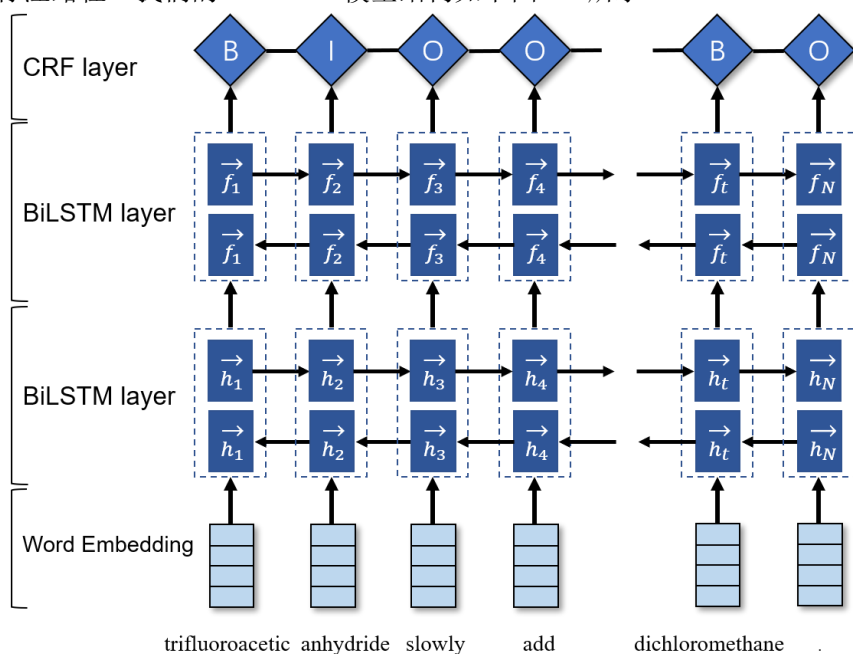


图 4.1 双层 BiLSTM + CRF 模型结构示意图

实验中, 我们对 BiLSTM + CRF 模型的参数进行调试, 得到兼顾训练速度与效果的超参数如下表 4.1 所示。其中最大序列长度为 200 可以包含数据集中 90% 的句子。

表 4.1 CemNER 模型主要参数

参数	数值
Tag Mode	BIO
Epoch	70
Word Embedding Dimension	256
Hidden Units	512
Dropout Rate	0.5
Optimization	Adam
Max Sequence Length	200

4.3 CemNER 模型的优化

实验中，我们尝试了不同的算法模型，包括双层 BiLSTM + CRF、单层 BiLSTM + CRF 与 Transformer + CRF；也尝试了两种标注模式 BIO 与 BIOES。以上模型在验证集 (Val) 上的表现如下表 4.2 所示。

表 4.2 不同算法、标注模式与参数的 cem 级模型表现

模型	每个 Epoch 耗时 (min)	标注模式	val F1 score
双层 BiLSTM + CRF	5.5	BIO	0.870
单层 BiLSTM + CRF	1.7	BIO	0.852
Transformer (2D) + CRF	2.8	BIO	0.829
Transformer (8D) + CRF	7.2	BIO	0.831
双层 BiLSTM + CRF	5.6	BIOES	0.865
Transformer (1D) + CRF	2.5	BIOES	0.854
Transformer (8D) + CRF	11.5	BIOES	0.860

注: D: Max Depth (Universal Transformer 循环深度)

Transformer 算法是一种摒弃传统 RNN 模型的序列结构，单纯使用注意力模块，来编码解码序列信息，从而实现训练速度的提升与模型表现的优化²⁵。而在此基础上 Universal Transformer 增加了对多头注意力模块的循环深度的控制，如下图 4.2 所示，从而进一步改善模型表现，在自然语言处理界有着举足轻重的地位²⁶。

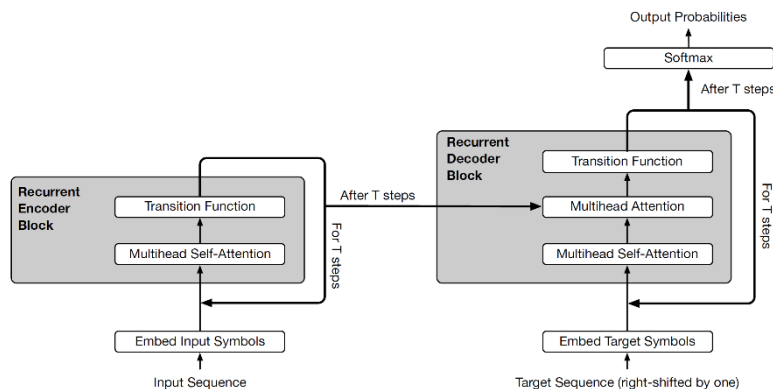


图 4.2 Universal Transformer 结构示意图²⁶

由表 4.2 结果可以看出，双层 BiLSTM 比单层 BiLSTM 表现更好，可能是因为双层结构能在两个维度提取文本特征，同时有更多的隐层单元对序列信息进行处理，可以掌握不同层次的信息。

单看 Transformer 模型，增加循环深度可以一定程度上提高模型表现，但是在实验过程中，增加循环深度后，训练时间明显增加，使得 Transformer 不再拥有训练速度快的优势。

在 BIO 标注模式下, Transformer 模型的表现不如 BiLSTM, 而在 BIOES 标注模式下, 二者差距缩小, 可能原因是 Transformer 结构更复杂, 目前 3 万级别的数据量无法满足 Transformer 模型所需训练量。另一方面, 在序列生成任务从三分类 BIO 模式变成五分类 BIOES 模式, 分类难度增加后, BiLSTM 表现稍有降低, 而 Transformer 表现有一定提升, 说明 Transformer 需要在更难的分类任务中才能体现优势。对于 Transformer 模型来说, 目前在机器翻译领域大显身手, 更适合分类体系复杂的序列生成任务。此外由于 Transformer 更加复杂的结构, 更多的参数量, 使得显卡在 8 G 的显存限制下每个 Batch 只能加入更少的数据, 使得实际训练速度比 BiLSTM 慢。因此综合考虑表现与训练效率, 在本实验的命名体识别任务中, 我们选用双层 BiLSTM + CRF 模型。

而最终我们的 CemNER 模型在测试集 (Test) 上的 cem 级准确率为 0.869, 召回率为 0.869, F1 score 为 0.869。该模型将用于有机反应专利文本的 CemNER。

4.4 RxnNER 数据集的建立

在有了自己的 CemNER 模型后, 我们可以将每一句文本中的化学命名体标注出来, 再用 Lowe D. M. 开发的 OPSIN³¹ 软件将这些化学名转化为化合物唯一的 InChIKey, 再与反应的中各部分化合物的 SMILES 转成的 InChIKey 进行匹配, 确定各个化合物在反应中是反应物、产物还是溶剂或催化剂。

InChIKey 是一个由 27 个字符组成的 InChI 压缩(Hash)版本, 用于互联网和数据库搜索与索引。这里选择 InChIKey 进行匹配而不直接用 SMILES 是因为每个化合物只有一个 InChIKey 编码, 而可以有很多种 SMILES 表示。因此选用具有唯一性的 InChIKey 来匹配识别的化学命名体与反应 SMILES。

如果某一条反应中的反应物或产物有部分化合物没有匹配到文本中识别的化学命名体, 则该条反应与文本无法进行完整的有机反应命名体标注, 因此不录入我们的 RxnNER 数据集。而这种问题主要发生于两种情况, 一个是文本中出现 Compound A、title compound 等化合物编号代替了原本的化学命名体, 这在我们段落级别的文本识别还无法实现, 需要在处理文档级别文本时, 建立相应的指代词字典进行索引。

另一个是 CemNER 模型预测的化学命名体不完整或有遗漏。例如在一些脂类的识别中会在 acid 处就结束, 而后续的 ester 识别为另一个词; 另一些遗漏的命名体是在训练集构成的词汇表以外 (Out of Vocabulary, OOV) 的单个词汇, 例如 oxalylchloride。但是一些出现

在系统命名中间的 OOV 则不会影响化学命名体识别, 例如 4 - nitrobenzenesulfonyl chloride, 由于模型学习到系统命名的命名模式与规律, 可以将 OOV 的 nitrobenzenesulfonyl 识别为化学命名体的词中 I。

本课题研究目前已经处理了 190 万反应文本数据中的前 80 万个, 得到 26,489 条准确标注了有机反应命名体的段落文本, 词汇量为 13,343, 示例如下表 4.3 所示, 召回率为 3%。而在没有进行系统的文本预处理召回率为 1%。预计处理完 190 万数据, 可以得到 6 万左右的 RxnNER 数据。

表 4.3 RxnNER 数据集标注示例

文本	BIO
pl	O
137	O
,	O
526	O
describe	O
hydrogenation	O
of	O
p	B-react
-	I-react
tert	I-react
-	I-react
butylphenol	I-react
to	O
form	O
p	B-prod
-	I-prod
tert	I-prod
-	I-prod
butylcyclohexanol	I-prod
use	O
a	O
nickel	B-catSolv
catalyst	O
.	O

注: react: 反应物; prod: 产物; catSolv: 催化剂或溶剂

经由以上步骤处理得到的 RxnNER 数据集, 其中文本主要有四大特点:

- 1、文本中所有反应物与产物的命名体都是系统命名或者明确唯一的常用名，都可以用 OPSIN 软件转化成 SMILES，而缩写（如 DMF）、化学式（如 EtOH）、官能团、化合物家族等有歧义的命名体不包含在内；
- 2、反应中的反应物与产物的命名体都在文本中至少出现一次，不存在被指带而没有出现的情况，例如 title compound、the product、compound 1 等（指代情况在原始数据中占据三分之一）；
- 3、文本中只标注了一个反应所对应的反应物、产物，存在一段文本包含多个反应的情况，但是在该情况下只有一个反应的有机物被标注，其余为 Other。
- 4、RxnNER 数据集中的文本都是我们 CemNER 模型能够准确识别化学命名体的文本，CemNER 识别失败的无法进入 RxnNER 数据集。

以上特点既解释了 RxnNER 数据集在原始数据中只召回了 3%，同时也确保每一条数据都是对于 RxnNER 模型最理想的学习资料。而在今后扩充数据集时，可以针对以上特点进行改进。

由于数据量庞大，且处理速度较慢，以下实验主要使用 2.6 万的 RxnNER 数据集（后续扩充到 3.7 万数据量）进行模型训练与评估，按照训练集 (Train)：验证集 (Val)：测试集 (Test) = 8 : 1 : 1 进行划分。

而 RxnNER 数据集的评价标准分为三个等级：tag 级、rpc 级与 rxn 级，其用途总结在表 4.4 中。

表 4.4 RxnNER 数据集的三个级别评价标准

评价级别	评价内容	用途
tag 级	O、B-react、B-prod、B-catSolv、I-react、I-prod、I-catSolv	运算快，监控模型训练进程
rpc 级	reactant、product、catalyst/solvent	反映模型对反应角色分类情况
rxn 级	反应物与产物组合而成的反应	反映模型提取有机反应表现

tag 级是评价标签 (O、B-react、B-prod、B-catSolv、I-react、I-prod、I-catSolv) 预测表现，该级别评价运算速度快，因此作为训练 RxnNER 模型时的监控指标；

rpc 级是评价反应物 (reactant)、产物 (product)、催化剂或溶剂 (catalyst/solvent) 预测表现，即需要将词汇按照标签进行组合成命名体进行判断，该级别评价可以输出混淆矩阵，让我们模型了解对于三类命名体的预测水平；

rxn 级是评价反应预测表现, 需要将反应物、产物组合成反应进行判断 (在此我们评价 rxn 级指标不考虑催化剂或溶剂, 允许模型对催化剂或溶剂的预测错误), 该级别评价是可以直接反映模型提取文本中反应的表现。

在这三个级别的评估中, 以准确率 (Precision)、召回率 (Recall) 以及 F1 分数这三个指标来对模型表现进行评价。考虑模型要兼顾准确率与召回率, 因此选择 F1 score 作为最终评价标准。

而 tag 级和 rpc 级属于多分类问题的评价, 其准确率、召回率和 F1 score 有 macro 和 micro 两种平均值算法。macro 是计算各分类的指标的算术平均值, 而 micro 是计算整体分类情况直接得到的准确率与召回率, 再得到整体 F1 score, 相当于计算了样本数量的加权平均值, 因此 macro 指标更能体现样本不平衡时, 小样本对于模型表现的贡献。在我们的 RxnNER 数据集中, tag 级的 O 类样本比其他 B、I 类样本多了一到两个数量级, 所以在 tag 级与 rpc 级评价时, 我们选择 macro F1 score 来评估模型表现。

5 有机反应命名体识别 (RxnNER) 模型的搭建、优化与应用

通过之前在 CemNER 模型的搭建、训练、优化过程中总结的经验, 我们选用双层 BiLSTM + CRF 模型作为主要调试、优化的对象。在训练模型的过程中, 我们不设定固定的训练 Epoch, 而是选用 tag 级 val F1 score 作为监控指标, 在 5 个 Epoch 内不再提升时, 减小学习率; 在 10 个 Epoch 内不再提升时, 停止训练并回溯到之前最优模型。在这种训练模式下, 模型一般在 40 ~ 50 个 Epoch 的训练后停止。

5.1 RxnNER 模型的搭建

RxnNER 选用双层 BiLSTM + CRF 模型的结构如之前图 4.1 所示。除了该模型, 我们还尝试搭建其他类似的基于 RNN 结构的模型, 例如单层 BiLSTM + CRF 模型、双层 BiLSTM 模型、双层 LSTM 模型进行对比。这些模型的表现如下表 5.1 所示, 所有评价都是在验证集上进行的。

表 5.1 不同参数、算法下的模型表现 (标注模式为 BIO)

模型	tag 级 F1 score	rpc 级 F1 score	rxn 级 F1 score
双层 BiLSTM (1024 Units) + CRF	0.93	0.915	0.818
双层 BiLSTM (512 Units) + CRF	0.92	0.901	0.790
单层 BiLSTM (512 Units) + CRF	0.92	0.891	0.778
双层 BiLSTM (512 Units)	0.50	<0.1	<0.1
双层 LSTM (512 Units) + CRF	0.91	0.888	0.744

由上表 5.1 可以看出, 更多的隐层单元数给模型带来更高的精度, 同时双层 BiLSTM 比单层 BiLSTM 表现更好, 说明在一定范围内增加模型参数与复杂度可以提高模型准确性。

而双层 BiLSTM 模型如果不加上 CRF 层, 模型在 tag 级 F1 score 只有 0.50, 而 rpc 级与 rxn 级的 F1 score 小于 0.1, 证明在 NER 任务中, CRF 算法是非常必要的, 在没有 CRF 层时, 模型在 rpc 级预测结果显著低于 tag 级, 说明有许多不合理标注 (例如词头标注 I, 词尾标注 B), 而 CRF 层可以让模型学习到标签的概率分布, 从而保证标注的合理性。

对于 LSTM 与 BiLSTM 的比较, 通过上表 5.1 的数据我们可以看出只考虑上文的 LSTM + CRF 模型表现不如 BiLSTM + CRF 模型, 验证了前文所述二者在原理上的优劣。

经过不同算法模型的尝试以及参数调试后, 得到 RxnNER 模型在验证集上 rxn 级 F1 score 为 0.818。

5.2 RxnNER 模型的优化——加入预训练词向量

5.1 得到的模型还可以继续优化, 例如用预训练的词向量代替原生的词嵌入 (Word Embedding) 向量。词嵌入也被称为分布式词表示, 是一种可以捕获来自未标记的大型语料库中词的语义和句法信息的词编码形式。

一个词汇量为 N 的词汇表中的第 n 个词会使用 onehot 编码, 即用一个 $N \times 1$ 的向量代表这个词, 这个向量除了第 n 位是 1, 其余为 0。而我们可以通过算法学习这个词在文本中上下文词法、句法的规律, 将其压缩至 $W \times 1$ 的向量 ($W < N$), 而此时的词向量在空间上有了新的取向, 而一般词义、语境接近的词向量取向更相近。因此训练好的词向量是包含词的词法、句法信息, 可以辅助其他自然语言处理模型获得更好的表现。

而在我们的双层 BiLSTM + CRF 模型中, 词嵌入层 (Word Embedding Layer) 就是用一个矩阵将高维度 onehot 编码的输入向量, 转化为较低维度的词向量, 进一步再输入到 BiLSTM 层。而词嵌入层可以使用原生随机初始化的矩阵, 让矩阵中的权重随着模型训练而

不断优化；也可以使用预训练的矩阵，即矩阵中的权重是通过算法已经学习了外部语料库中的词法、句法信息。而在过往的研究中表明，通过学习更庞大数据量的外部语料库得到的词向量，比需要在训练中学习数据集文本特征的原生词向量表现更好³²。

而想要得到优质的预训练词向量，一方面需要足够大的语料库，另一方面需要语料库与我们任务中的数据集文本内容尽可能接近。因此我们选用与有机反应相关的专利作为语料库，得到 2 千万条语句，2.5 GB 的数据量。除去文本中出现不超过 2 次的低频词，得到 917,569 的词汇量。由于语料库数据量较大，选用 Gensim 模块中的连续词袋 (CBOW) 算法来训练，最终经过 5 个 Epoch 得到 917,569 个词的 256 维词向量。利用 TensorFlow 开源的 projector 网站对其中出现频率最高的 10,000 个词的词向量进行 PCA 降维至三维空间，并进行可视化。我们可以搜索感兴趣的词，以及与其最相似的 10 个词，例如 glycine (甘氨酸)、obtain (得到)、nsaid (Nonsteroidal Antiinflammatory Drugs 非甾体抗炎药)，其可视化结果如图 5.1 所示。

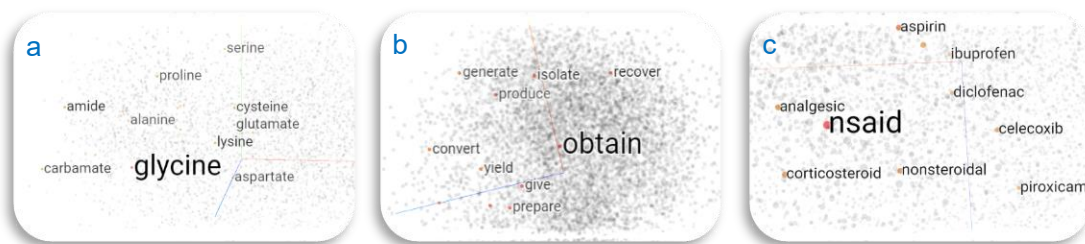


图 5.1 词向量与语义相近词的可视化结果

由上图可以看出，通过有机反应语料库训练出来的词向量能够很好地提取词汇的语义特征。图 5.1(a)中 glycine (甘氨酸)与一系列天然氨基酸以及衍生物 (carbamate 氨基甲酸酯) 有着相似的词向量；图 5.1(b)中 nsaid (非甾体抗炎药)与一系列特定的非甾体抗炎药 (aspirin 阿司匹林、ibuprofen 布洛芬等) 以及 analgesic (止痛药)、corticosteroid (皮质类固醇)、nonsteroidal (非甾体) 等含义相关的词有着相似的词向量。

而图 5.1(b)中 obtain 更加值得关注，因为在有机反应文本中 obtain 往往与产物相关联，因此我们发现训练好的词向量中，与 obtain 语义相近的词 yield、produce、generate、give、isolate、prepare 等也都是与产物直接相关联的动词。因此在我们的 RxnNER 模型可能就是学习到了文本中类似语义规律，在这一类动词附近标注的产物命名体。

我们在预训练词向量中找出 RxnNER 数据集词汇表 13,345 个词对应的词向量，组成一个 13345×256 维的词嵌入矩阵，用于替换原生的词嵌入矩阵，而其中有 168 个词无法在预训练词向量中找到。

那么新的词嵌入层可以选择固定权重 (Fixed)，在训练过程中不调整权重；也可以选择固定权重，在训练中微调权重 (Fine-tune)。这两种情况我们都进行了实验，二者与原生词嵌入层的结果比较如下图 5.2 所示（这些结果都是在测试集上的表现）。

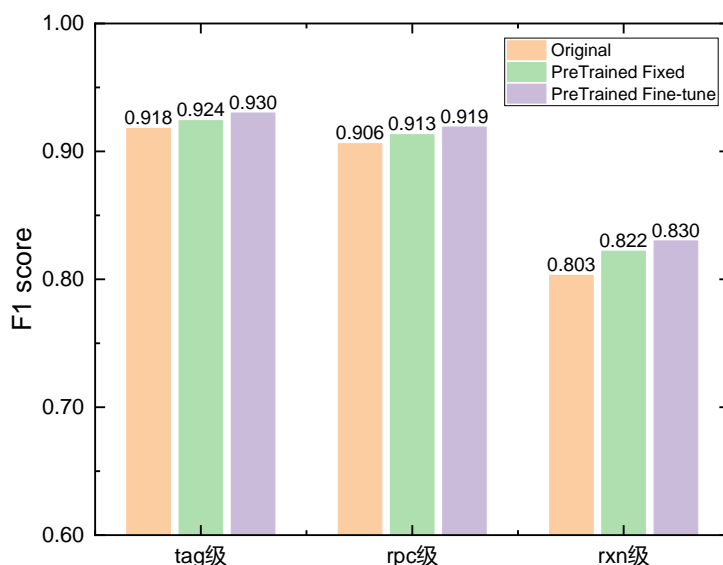


图 5.2 不同词嵌入层对模型表现的影响

由上图 5.2 可知加入预训练词向量可以改善模型表现，而允许预训练的词向量随着训练进行微调可以进一步提高模型表现。可能原因是 RxnNER 数据集中有 168 个词不在预训练词向量中，如果 Fixed 词向量的权重，那么这些词的词向量将都是初始设置的零向量，而无法学习到语义特征。因此 Fine-tune 模式下使得这些词向量可以跟随训练进行微调，生成新的词向量。

5.3 RxnNER 模型的优化——扩充数据集

以上实验都是在 26,489 条反应文本的 RxnNER 数据集上进行的，得到最好结果为测试集上 rxn 级 F1 score 0.830。而后续我们增加 RxnNER 数据集的数据量至 37,787 条，增加了 42% 的数据量，词汇量由 13,343 增加至 15,649。在此基础上训练加入预训练词向量的双层 BiLSTM+CRF 模型，得到最好结果为测试集上 rxn 级 F1 score 0.862。在不同数据量的数据集上模型表现如下表 5.2 所示。

表 5.2 不同数据量下模型的表现

数据量	词汇量	tag 级 macro F1 score	rpc 级 macro F1 score	rxn 级 F1 score
26,489	13,343	0.930	0.919	0.830
37,787	15,649	0.952	0.935	0.862

由上表 5.2 的比较可知,在目前 2~3 万的数据量级别,增加数据量可以有效提升模型表现。扩充数据集一方面增加有机名词、基团名词等词汇量,使模型掌握更多词汇特征,可以减少模型遇到的新词。对于新词,模型会识别其为<UNK> (unknown),而在训练时<UNK> 对应出现频率仅为 1 次的词,而所有<UNK>都共享一个词向量,因而新词在模型预测时自身特征无法体现在词向量中,只能通过 BiLSTM 与 CRF 对周围语境、词汇的概率分析来识别新词,这样会降低命名体识别的准确性与召回率。因此,在排除了不同时态、单复数等冗余词汇后,增加有效词汇量对提升模型表现是有明显作用。

另一方面,增加数据量可以丰富文本的风格,因为不同专利文本描述有机反应的模型所用词汇句式有所区别,学习更多风格的文本对于文本把握句式、词汇用法有所帮助,因此可以提升模型的表现。

虽然增加数据量的边际效益会随数据增加而递减,但是在目前 2~3 万量级下,有必要继续扩增 RxnNER 数据集。

目前得到 RxnNER 模型预测 Reactants、Products、Catalyst/Solvent 的 micro 准确率为 0.939,召回率为 0.938, F1 score 为 0.934;对比 Lowe D. M.报道的预测试剂 Reagent (对应我们的 Reactants、Catalyst/Solvent) 与 Products 的 micro 准确率为 0.889,召回率为 0.964, F1 score 为 0.925²²。由于后者评价所用测试集为 2008 年至 2011 年的 USPTO 专利文档,而我们的测试集为段落文本,比前者更为理想,所以比较结果仅供参考,无太大意义。

5.4 RxnNER 模型误差定性分析

除了以上分析不同算法、参数、数据量对于模型表现的影响,我们以目前最优模型(测试集上准确率为 0.861,召回率为 0.862, F1 score 为 0.862)为例,借助多分类问题中的混淆矩阵来定性分析其误差以及改进方案。

测试集上的 tag 级分类混淆矩阵如下图 5.3 所示。

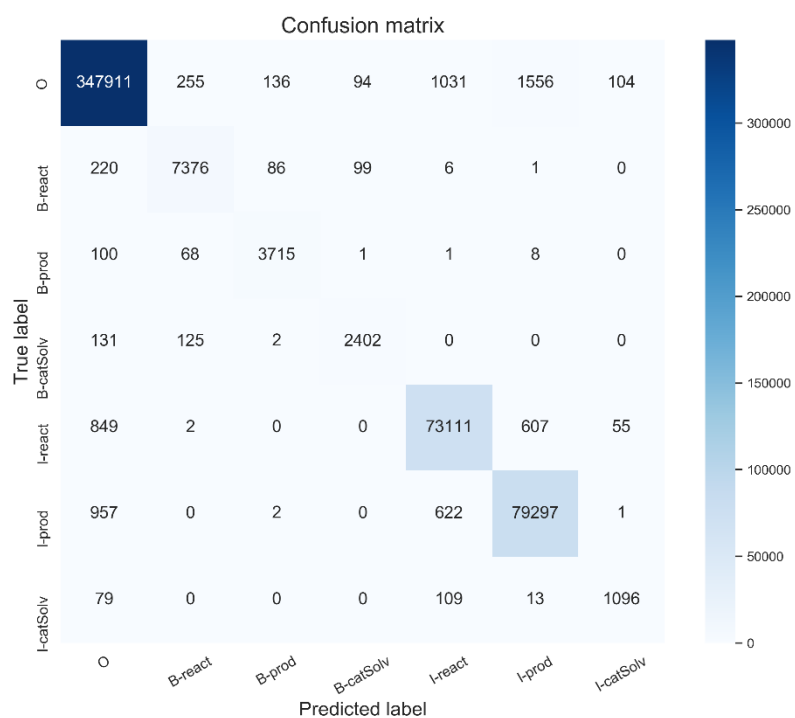


图 5.3 tag 级分类混淆矩阵

测试集上的 rpc 级分类混淆矩阵如下图 5.4 所示。

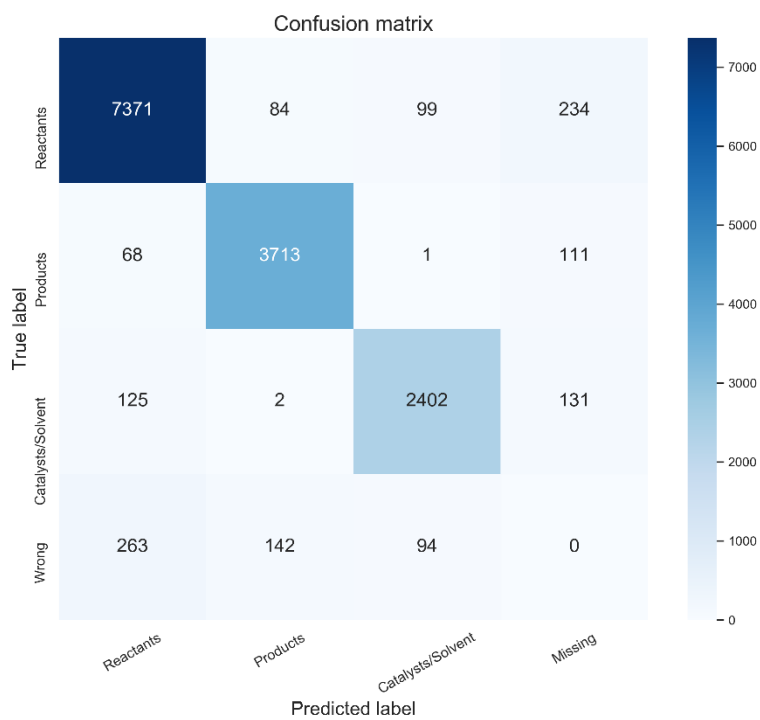


图 5.4 rpc 级分类混淆矩阵

在混淆矩阵中，左上到右下的对角线是 True Positive (预测为正且真实为正, TP) 对应的数量，其余是不同分类下的预测错误情况。在，分类误差主要出现在 O 被预测为反应命名体中的一类，这点会导致预测 rpc 时容易发生冗余预测 (False Positive)，在 rpc 级混淆矩阵的 Wrong 行可以得到这个结论；同时 tag 级测试下误差也出现在其他反应命名体被预测为 O，造成预测 rpc 时容易发生丢失 (True Negative)，在 rpc 级混淆矩阵的 Missing 列可以得到这个结论。这方面主要存在的问题与 CemNER 任务时遇到的类似，需要增加数据量，丰富词汇量与文本风格，同时也可以考虑进一步优化模型，尝试更复杂的算法。

而除了 O 有关的预测错误外，出现最多错误的是催化剂或溶剂被预测为反应物，这点在 rpc 级的混淆矩阵中更能直观体现，催化剂或溶剂被预测为反应物的数量为 125 个，比 rpc 之间预测错误都要多。这方面存在的问题其实是在情理之中，因为在英文有机反应文本中，反应物与催化剂或溶剂都是与“加入”这类动词相关联的，而产物多与“得到”这类词相关联，因此反应物与催化剂或溶剂在语境上容易混淆；另一方面，在有机实验中，反应物与催化剂、溶剂大多数情况下，共同构成反应的起始状态，有些溶剂本身也是反应物（或者说有的反应物就是溶剂，比如傅克酰基化中的芳香烃既是反应物也是溶剂），而催化剂在有机化学中的定义较为模糊，尤其是涉及到酸碱催化反应，催化剂明确参与质子转移过程，甚至有的吸收反应生成质子的碱性试剂，在反应过程中是不断被消耗的，我们也会称之为催化剂。因此，想从命名体识别的角度来优化这个问题是比较困难的，我们考虑可以找出主反应物与主产物，根据原子匹配与守恒定律来重新校对催化剂、溶剂。

5.5 RxnNER 模型的应用

在用以上得到的 RxnNER 模型对文本进行有机反应命名体识别后,一步得到了反应物、产物、催化剂或溶剂的命名体。之后进行后处理输出反应信息: 利用 OPSIN 软件将其转化为 SMILES, 并用 Rdkit 组装成反应 SMILES, 输出反应的图片与文字, 即实现了段落级文本识别出有机反应的流程, 其示例如下图 5.5 所示。

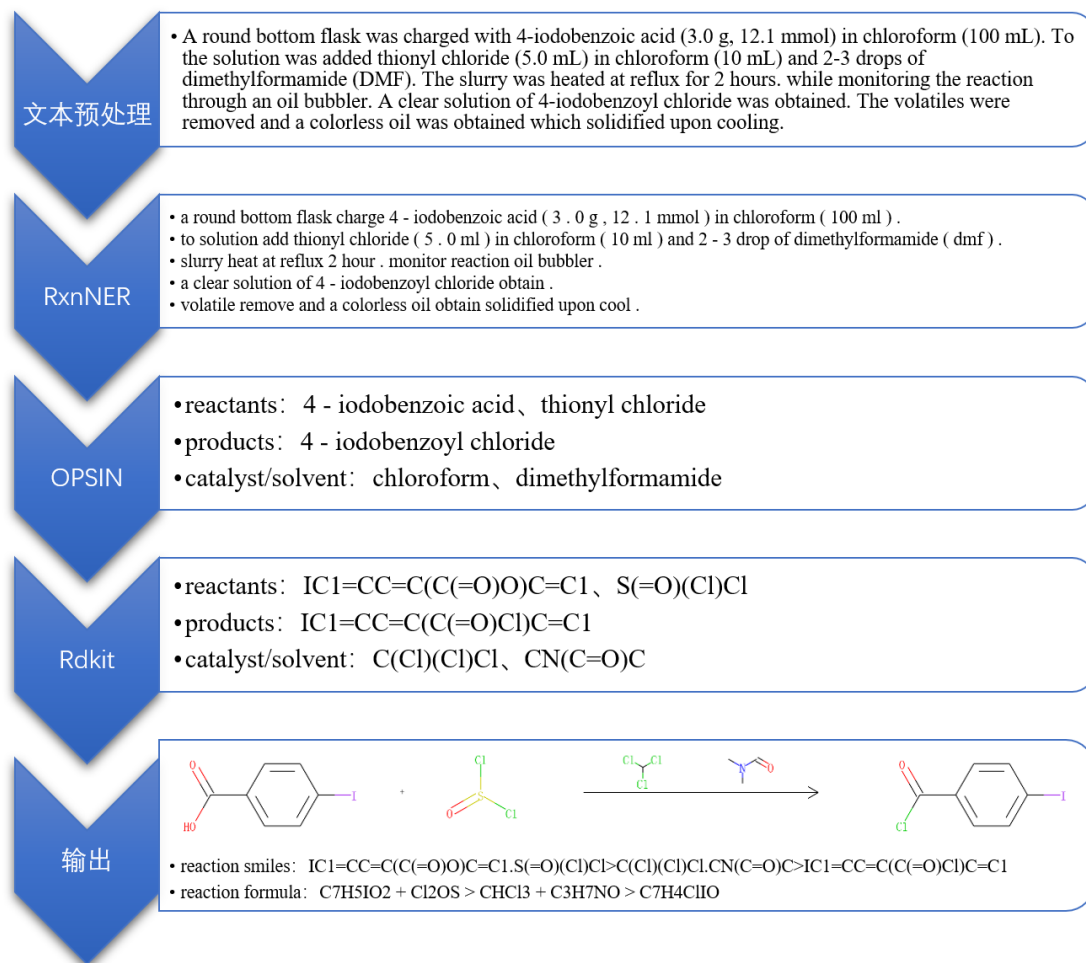


图 5.5 段落级文本中识别有机反应的流程示意图

在我们的 RxnNER 数据集上, 模型有着 F1 score 0.862 的不错表现。但是这仅是在我们建立的理想数据集上进行的测试 (一段文本只对应一个反应, 并且没有指代词), 而且目前只实现段落级文本的识别, 实际生产中应用意义更大的是文档级别的识别与提取。因此, 我们接下来尝试把段落级识别模型应用于文档级反应提取, 对于文档级文本反应提取进行初步的试验。

而在此我们选择文本格式较为规范、文本分段清晰且解析较为高效的 USPTO 专利 XML 文档进行文档级反应提取的尝试。

我们目前文档级反应提取的思路如下图 5.6 所示。

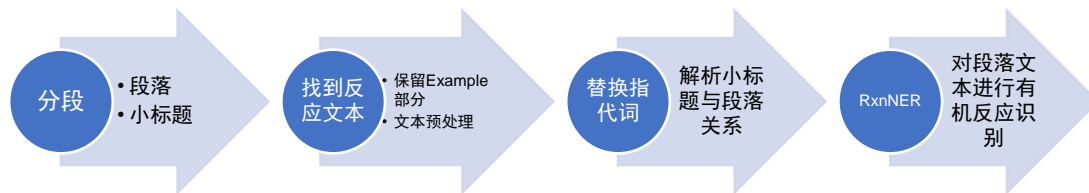


图 5.6 文档级有机反应提取的流程示意图

首先利用 chemdataextractor 模块解析 XML 文本，得到分好的段落以及小标题。由于几乎所有的有机反应、有机合成描述都集中在专利的 Example 部分，因此我们只保留 Example 段落文本，并进行文本预处理。同时由于反应文本中出现的核磁、质谱表征数据会干扰模型预测，因此利用分句及相关字符特征匹配，将表征部分删除。

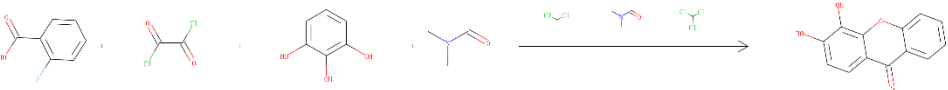
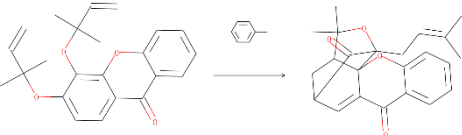
然后由于许多专利在描述有机反应时，会将主产物放在小标题中，而在反应描述中用 title compound、the product 等词进行指代，因此需要解析小标题与段落文本的对应关系，并把其中指代词替换为系统命名的化学命名体，便于我们的模型正确识别。

最终在分段、保留关键段落、替换指代词三个步骤结束后，我们用之前训练好的 RxnNER 模型进行反应识别。由于本课题文档级软件尚未完全开发，暂时无法得到大量测试结果，在此以一篇文档为例进行反应提取，提取结果与 Lowe D. M. 提供的结果进行比较。

以专利 US20050004026A1 为例，该专利介绍了治疗诱导凋亡的疾病的方法和筛选测定，涉及一些荧光分子的有机合成和细胞实验，反应文本中既有符合 RxnNER 数据集特征的完整描述，也有涉及 title compound 等指代词以及罕见化学命名体（例如天然产物藤黄酸 gambogic acid）和系统命名比较复杂的有机分子（例如 1,3,7-Triallyl-8-benzoyl-4-Oxa-tricyclo[4.3.1.0^{3,7}]]dec-8-en-2-one）。

经过人工核对可以提取的有机反应为 32 条，Lowe D. M. 提取的有机反应为 13 条，其中一条不完整（缺少关键反应物），还有两条反应对应的文本为细胞实验，并无可提取反应；而我们提取了 15 条反应，其中 6 条反应由于 OPSIN 无法正确将化学命名体转化为 SMILES 而无法输出反应 SMILES 与图片，但是识别了正确的反应物与产物；有 1 条反应对应的文本不包含有机反应；其余反应主要反应物与产物都预测正确，部分反应物、溶剂、催化剂遗漏。我们的部分具有代表性的预测结果如下表 5.5 所示（反应文本中的标注为人工标注：黄色为反应物，绿色为催化剂或溶剂，蓝色为产物；反应文本的第一、二行是小标题）。

表 5.5 专利 US20050004026A1 中提取有机反应的部分结果

反应文本 (1)	<p>Preparation of N-(2-Gambogylaminoethyl)biotinamide</p> <p>A mixture of gambogic acid (85.6 mg, 0.136 mmol), DMAF (19.9 mg, 0.164 mmol), EDC (31.3 mg, 0.164 mmol) and N-(2-aminoethyl)biotinamide (Molecular Probes, 50 mg, 0.14 mmol) in DMF (5 mL) was stirred at room temperature for 72 h. The solution was poured into water (50 mL) and was extracted with ethyl acetate (3×10 mL). The combined organic layer was dried and concentrated to give crude product, which was purified by chromatography (SiO₂, EtOAc/MeOH 4:1) to give the title compound (28 mg, 23%). ¹H NMR (CDCl₃): 12.92(s, 1H), 7.58 (d, J=6.9 Hz, 1H), 7.05-6.90 (m, 2H), 6.68(d, J=9.9 Hz, 1H), 6.15 (bs, 1H), 5.50 (d, J=10.5 Hz, 1H), 5.28 (m, 2H), 5.05 (m, 2H), 4.49 (m, 1H), 4.32 (m, 1H), 3.58-2.00 (m, 14H), 1.77 (bs, 3H), 1.73 (bs, 3H), 1.69 (bs, 6H), 1.65 (bs, 6H), 1.45 (bs, 3H), 1.29 (bs, 3H). MS: 919 (M+Na⁺), 897 (M+H⁺), 895 (M-H⁺).</p>
预测结果	<p>reactants: gambogic acid、ethyl acetate</p> <p>products: N-(2-Gambogylaminoethyl)biotinamide</p> <p>catalyst/Solvent: None</p>
说明	<p>无法识别文本中缩写的溶剂、催化剂；</p> <p>将后处理试剂 ethyl acetate 预测为溶剂或催化剂；</p> <p>由于 OPSIN 无法解析 gambogic acid 与 N-(2-Gambogylaminoethyl)biotinamide，无法得到反应的 SMILES 与图片</p>
反应文本 (2)	<p>Example 15</p> <p>3,4-dihydroxyxanthen-9-one</p> <p>To a stirring solution of 2-fluorobenzoic acid (5.09 g, 36.3 mmol) and dichloromethane (110 mL) in an ice bath under argon was added dropwise a solution of oxalyl chloride (2.0 M in dichloromethane, 21 mL, 42 mmol), followed by dimethylformamide (6 drops). The ice bath was removed and the solution was stirred at room temperature for 1.5 h. The solution was then concentrated by rotary evaporation. The product was dissolved in hexane (3×50 mL) and the mixture was filtered. The filtrate was rotary evaporated to yield 5.42 g of colorless oil. The oil was added dropwise to a mixture of pyrogallol (6.48 g, 51.3 mmol), aluminum chloride (14.6 g, 110 mmol), chloroform (250 mL) and dichloromethane (700 mL), and the solution was stirred for 17 h at room temperature. The solution was then refluxed for 3 h and cooled to room temperature. The solution was washed with 1 N HCl (3×500 mL). The organic layer was filtered, dried over sodium sulfate, and evaporated to yield an oil. The oil was added to dimethylformamide (120 mL) with sodium carbonate (8.11 g, 76.5 mmol) and it was refluxed for 3.5 h. The solution was concentrated by rotary evaporation with heating, and the residue was purified by column chromatography (95:5 chloroform/methanol) to give a solid. The solid was washed with hexane (2×35 mL), filtered and dried to yield 2.10 g (25 %) of the title compound as an off-white solid. ¹H NMR (DMSO-d₆, 300 MHz): δ 8.16 (d, J=7.42 Hz, 1H), 7.84 (t, J=7.69 Hz, 1H), 7.64 (d, J=8.52 Hz, 1H), 7.57 (d, J=8.79 Hz, 1H), 7.44 (t, J=7.41 Hz, 1H), 6.94 (d, J=8.52 Hz, 1H).</p>
预测结果	
说明	<p>遗漏关键催化剂 AlCl₃；草酰氯与 DMF 应该是催化剂与溶剂而非反应物</p>
反应文本 (3)	<p>Example 22</p> <p>1-(3-Methyl-2-butenyl)-3,3-dimethyl-1,3,3a,4,5,12a-hexahydro-7,13-dioxo-1,5-methano-furo[3,4-d]xanthene</p> <p>A solution of 3,4-bis-(1,1-dimethyl-allyloxy)-xanthen-9-one (229 mg, 0.587 mmol) in toluene (10 mL) was refluxed under argon for 2 h. The solvent was evaporated and the residue was purified by column chromatography (SiO₂, EtOAc:hexanes/10-30%) to give the title compound as white solids (145 mg, 63%): ¹H NMR (CDCl₃, 300 MHz) δ 7.95 (dd, J=1.5, 7.8 Hz, 1H), 7.53 (ddd, J=1.5, 7.2, 8.1 Hz, 1H), 7.44 (dd, J=0.6, 7.2 Hz, 1H), 7.07 (m, 2H), 4.42 (m, 1H), 3.50 (dd, J=4.5, 6.9 Hz, 1H), 2.63 (m, 2H), 2.46 (d, J=9.3 Hz, 1H), 2.35 (dd, J=4.2, 12.6 Hz, 1H), 1.73 (s, 3H), 1.31 (m, 1H), 1.31 (s, 6H), 0.92 (s, 3H); ¹³C NMR (CDCl₃, 75 MHz) δ 202.8, 176.3, 159.5, 136.1, 134.8, 134.7, 133.6, 126.8, 121.8, 119.0, 118.9, 118.0, 90.3, 84.6, 83.5, 48.8, 46.8, 30.4, 29.2, 25.4, 25.2, 16.8.</p>
预测结果	
说明	<p>准确提取了一个[3, 3]σ 迁移反应 (Claisen 重排)</p>
反应文本 (4)	<p>Example 25</p> <p>(2,5-Diallyl-3,4-dihydroxy)-benzophenone</p> <p>A solution of 3,4-bis-allyloxy-benzophenone (1.510 g, 5.1 mmol) in diphenyl ether (3 mL) was stirred at 200 °C. for 2 h. The reaction mixture was cooled and the mixture was purified by column chromatography (SiO₂, 30% EtOAc in hexanes) to give a light yellow oil (0.840 g, 56%): ¹H NMR (CDCl₃, 300 MHz) δ 7.79 (m, 2H), 7.57 (m, 1H), 7.44 (t, J=7.8 Hz, 2H), 6.78 (s, 1H), 6.16-5.86 (m, 3H), 5.60 (brs, 1H), 5.12 (m, 4H), 3.51 (d, J=6.0 Hz, 2H), 3.39 (d, J=6.6 Hz, 2H).</p>
预测结果	<p>reactants: 3,4-bis-allyloxy-benzophenone</p>

	products: (2,5-Diallyl-3,4-dihydroxy)-benzophenone catalyst/Solvent: diphenyl ether
说明	文本中没有提及产物或 title compound 等指代词，默认小标题中的化合物为产物； 由于 OPSIN 无法解析(2,5-Diallyl-3,4-dihydroxy)-benzophenone，因此没有得到反应 SMILES 与图片。

该专利原有 32 条反应，其中提取了 15 条，剩下 17 条文本无法提取的主要原因为以下 4 点：

- 1、无法准确识别反应命名体，包括无机物、低频词(天然产物)、特别长的系统命名等；
- 2、缩写（如 DMF）、化学式（如 EtOH）、常用名（如 Lindlar catalyst）等特殊命名体干扰模型预测与 OPSIN 解析转化过程
- 3、受限于 OPSIN 无法解析缩写、常用名、复杂系统命名、错别词等化学命名体；
- 4、多样化的指代词与指代内容，例如 “The title compound was prepared in an analogous manner as Example (198)”；

以上 4 点是在初步尝试提取文档级文本中的有机反应时所发现需要解决的问题。

对此我们也提出相应改进的办法：

- 1、增加 RxnNER 模型数据量，提升词汇量；进一步优化模型，提高其泛化能力；
- 2、建立缩写、化学式、常用名的字典库，在文本预处理时将这些特殊命名体替换为系统命名。
- 3、ChemOffice 提供的 ChemScript 模块可以处理更复杂的系统命名，也可以对拼写错误的命名体进行自动纠正，因此考虑之后代替 OPSIN 来解析化学命名体。
- 4、指代词问题是文档级识别中的最主要问题，也是最复杂的问题。目前考虑建立全文档的字典对指代词进行索引与替换。而指代词本身形式多样，也有丰富的表达方式，考虑可以借助词向量来捕获更多的指代词。

此外需要一个算法来校对提取出来的反应，这点可以考虑借鉴 *Lowe D. M.*所使用的 atom-atom Mapping (AAM) 算法²²，即将反应物与产物所有原子进行匹配，从而可以验证反应合理性，以及将预测错位置的反应物、催化剂、溶剂进行归位。而在我们进行实验比对的过程中发现 *Lowe D. M.*所使用的 AAM 算法仍有瑕疵，无法有效处理不同比例多反应物的情况，同时也存在“无中生有”情况，原本是生物细胞实验，而将添加的试剂当作反应物，并

由反应物编造出仅仅原子守恒的不合理产物。因此在这方面还有很大的提升空间, 值得我们未来继续深入研究与开发。

6 结论

本课题运用自然语言处理的技术与算法来解决文本中提取有机反应的问题。首先建立化学类文本的预处理方法。接着搭建了双层 BiLSTM + CRF 的化学命名体识别 (CemNER) 模型, 识别 cem 的准确率为 0.869, 召回率为 0.869, F1 score 为 0.869, 并用该模型处理 USPTO 专利数据库的反应与文本数据, 建立了有机反应命名体识别 (RxnNER) 数据集, 该数据集中的化学命名体。目前 RxnNER 数据集有 37,787 条标注了反应命名体的段落文本, 除去只出现一次的低频词, 词汇量为 15,651。

在该数据集上, 我们搭建了加入预训练词向量的双层 BiLSTM + CRF 模型, 实现段落级文本中直接识别反应物 (reactant)、产物 (product) 和催化剂或溶剂 (catalyst/solvent), 并用 OPSIN 将识别的反应命名体转化为 SMILES, 并用 Rdkit 组装反应。我们的模型在 RxnNER 数据集上识别有机反应的准确率为 0.861, 召回率为 0.862, F1 score 为 0.862。

在实现段落级文本识别提取有机反应后, 我们初步尝试进行文档级文本的有机反应提取, 在解决一部分指代词 (例如 title compound) 的指代问题后, 能够提取一部分专利文档中的有机反应。我们总结了文档级文本识别反应的困难, 并提出相应的解决思路。

文档级相比段落级, 对于大量提取专利、文献中有机反应更具实践生产意义。而处理文档级文本时会面临更丰富的化学命名体类型, 以及更加多样的文本描述风格, 因此我们需要对文档结构与信息进行深入解析、改造, 才能使基于以系统命名为主的 RxnNER 数据集搭建的模型能够有更出色的表现。而未来我们将会在这方面继续研究与开发。

7 参考文献

1. Corey, E. J.; Wipke, W. T., Computer-Assisted Design of Complex Organic Syntheses. *Science* **1969**, *166* (3902), 178-192.
2. Lin, K.; Xu, Y.; Pei, J.; Lai, L., Automatic retrosynthetic route planning using template-free models. *Chemical Science* **2020**, *11* (12), 3355-3364.
3. Filippov, I. V.; Nicklaus, M. C., Optical Structure Recognition Software To Recover Chemical Information: OSRA, An Open Source Solution. *J. Chem. Inf. Model.* **2009**, *49* (3), 740-743.
4. Tharatipyakul, A.; Numnark, S.; Wichadakul, D.; Ingsriswang, S., ChemEx: information extraction system for chemical data curation. *BMC Bioinformatics* **2012**, *13* (17), S9.
5. Fujiyoshi, A.; Nakagawa, K.; Suzuki, M., Robust Method of Segmentation and Recognition of Chemical Structure Images in ChemInfty. In *Pre-Proceedings of the 9th IAPR International Workshop on Graphics Recognition (GREC2011)*, 2011.
6. Lounnas, V.; Vriend, G., AsteriX: A Web Server To Automatically Extract Ligand Coordinates from Figures in PDF Articles. *J. Chem. Inf. Model.* **2012**, *52*, 568-76.
7. Dai, H.-J.; Lai, P.-T.; Chang, Y.-C.; Tsai, R. T.-H., Enhancing of chemical compound and drug name recognition using representative tag scheme and fine-grained tokenization. *J. Cheminform.* **2015**, *7* (1), S14.
8. Lowe, D. M.; Sayle, R. A., LeadMine: a grammar and dictionary driven approach to entity recognition. *J. Cheminform.* **2015**, *7* (1), S5.
9. Xu, S.; An, X.; Zhu, L.; Zhang, Y.; Zhang, H., A CRF-based system for recognizing chemical entity mentions (CEMs) in biomedical literature. *J. Cheminform.* **2015**, *7* (1), S11.
10. Tang, B.; Feng, Y.; Wang, X.; Wu, Y.; Zhang, Y.; Jiang, M.; Wang, J.; Xu, H., A comparison of conditional random fields and structured support vector machines for chemical entity recognition in biomedical literature. *J. Cheminform.* **2015**, *7* (1), S8.
11. Usié, A.; Cruz, J.; Comas, J.; Solsona, F.; Alves, R., CheNER: a tool for the identification of chemical entities and their classes in biomedical literature. *J. Cheminform.* **2015**, *7* (1), S15.
12. Krallinger, M.; Rabal, O.; Leitner, F.; Vazquez, M.; Salgado, D.; Lu, Z.; Leaman, R.; Lu, Y.; Ji, D.; Lowe, D.; Sayle, R.; Batista-Navarro, R.; Rak, R.; Huber, T.; Rocktäschel, T.; Matos, S.; Campos, D.; Tang, B.; Xu, H.; Valencia, A., The CHEMDNER corpus of chemicals and drugs and its annotation principles. *J. Cheminform.* **2015**, *7*, S2.

13. Leaman, R.; Wei, C.-H.; Lu, Z., tmChem: a high performance approach for chemical named entity recognition and normalization. *J. Cheminform.* **2015**, 7 (1), S3.
14. Swain, M. C.; Cole, J. M., ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *J. Chem. Inf. Model.* **2016**, 56 (10), 1894-1904.
15. Reeker, L. H.; Zamora, E. M.; Blower, P. E., Specialized information extraction: automatic chemical reaction coding from English descriptions. In *Proceedings of the first conference on Applied natural language processing*, Association for Computational Linguistics: Santa Monica, California, 1983; pp 109–116.
16. Zamora, E. M.; Blower, P. E., Extraction of chemical reaction information from primary journal text using computational linguistics techniques. 1. Lexical and syntactic phases. *J. Chem. Inf. Comput. Sci.* **1984**, 24 (3), 176-181.
17. Zamora, E. M.; Blower, P. E., Extraction of chemical reaction information from primary journal text using computational linguistics techniques. 2. Semantic phase. *J. Chem. Inf. Comput. Sci.* **1984**, 24 (3), 181-188.
18. Ai, C. S.; Blower, P. E.; Ledwith, R. H., Extraction of chemical reaction information from primary journal text. *J. Chem. Inf. Comput. Sci.* **1990**, 30 (2), 163-169.
19. Jessop, D. M.; Adams, S. E.; Murray-Rust, P., Mining chemical information from open patents. *J. Cheminform.* **2011**, 3 (1), 40.
20. Hawizy, L.; Jessop, D. M.; Adams, N.; Murray-Rust, P., ChemicalTagger: A tool for semantic text-mining in chemistry. *J. Cheminform.* **2011**, 3 (1), 17.
21. Jessop, D. M.; Adams, S. E.; Willighagen, E. L.; Hawizy, L.; Murray-Rust, P., OSCAR4: a flexible architecture for chemical text-mining. *J. Cheminform.* **2011**, 3 (1), 41.
22. Lowe, D. M. Extraction of chemical structures and reactions from the literature. Ph.D. Thesis, University of Cambridge, 2012.
23. Hochreiter, S.; Schmidhuber, J., Long Short-Term Memory. *Neural Comput.* **1997**, 9 (8), 1735-1780.
24. Graves, A.; Schmidhuber, J., Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* **2005**, 18 (5), 602-610.
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. In *Attention is all you need*, Advances in neural information processing systems, 2017; pp 5998-6008.
26. Dehghani, M.; Gouws, S.; Vinyals, O.; Uszkoreit, J.; Kaiser, Ł., Universal transformers. *arXiv preprint arXiv:1807.03819* **2018**.

-
27. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K., Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **2018**.
28. Luo, L.; Yang, Z.; Yang, P.; Zhang, Y.; Wang, L.; Lin, H.; Wang, J., An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics* **2017**, *34* (8), 1381-1388.
29. Daniel, L., *Chemical reactions from US patents (1976-Sep2016)*. 2017.
30. Li, J.; Sun, Y.; Johnson, R. J.; Sciaky, D.; Wei, C. H.; Leaman, R.; Davis, A. P.; Mattingly, C. J.; Wiegiers, T. C.; Lu, Z., BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database (Oxford)* **2016**, 2016.
31. Lowe, D. M.; Corbett, P. T.; Murray-Rust, P.; Glen, R. C., Chemical Name to Structure: OPSIN, an Open Source Solution. *J. Chem. Inf. Model.* **2011**, *51* (3), 739-753.
32. Lai, S.; Liu, K.; He, S.; Zhao, J., How to generate a good word embedding. *IEEE Intelligent Systems* **2016**, *31* (6), 5-14.

8 致谢

首先感谢导师来鲁华教授与裴剑锋老师在课题上给予的建议以及过去两年在组里学习过程中给予我科研学习的指导建议。

其次感谢师兄徐优俊博士后，在我构思课题时给予我很大的启发，同时热心帮助我解决一些程序编写上的问题；同时也感谢林康杰师兄，与我分享有机反应相关数据以及处理有机反应的相关经验；感谢同级的王聪同学一直与我分享科研学习的心得，一同在深度学习、化学信息学领域中进步；感谢 MDL 课题组的师兄师姐们在我本科两年科研学习中给予我的帮助，在一次次组会交流中给予我科研上的启发。

特别感谢我的父母、奶奶以及无论远近的亲人，在家做课题写论文的过程离不开你们对我生活上的照顾与关心；特别感谢我的同学与朋友，无论在这段期间是否见面，但是你们是我生活中一抹亮丽的色彩，让我能共享喜怒哀乐、体味不同的人生。

最后，特别感谢与缅怀我亲爱的爷爷。自从一月份回家后，我在家看着你的身体每况愈下，一直到三月十四日离我而去。在这不忍回首的两个月时间里，我逐渐承担起家里的重任，才体会到您以前为了照顾我上学时的辛劳，才一点点真正成长起来。您在最后的时间里还总是怕照顾您而耽误我学习，一直叮嘱我要好好做课题做研究。虽然这次课题还有许多不足的地方，但是我会继续努力钻研下去，做出更好的成果献给您！