小样本下材料领域的命名实体识别

北京大数据技能大赛

作者：徐早辉、胡俊宝、张熙熙、覃华清、李睿思

2022 年 7 月 6 日

目录

1 摘要

二氧化碳是一种主要的温室气体，容易导致全球气候急速变暖，但植物光合作用吸收二氧化碳效率较低，保护环境迫在眉睫，故应人工设计新型高效催化剂，将二氧化碳转化为许多下游碳基化合物。其中电化学转化是近年来较受关注的一个领域，在电化学法还原中，铜金属[1]做催化剂还原二氧化碳因其还原效果好、活性高[2]受到广泛关注。催化剂金属的种类即为研究问题的目标材料种类，调控因素为催化剂金属的形貌特征，产物种类是生成物的不同种类，在电化学催化反应中法拉第电磁感应效率是表征催化剂活性的重要标准。

本文基于初赛数据集的分析和广泛的文献调研，提出了包含上下文推理的标注规则与高效的标注方法，目前构建了约 150 份与主题相关的更加适合小样本下材料领域的实体命名数据集，结合迁移学习与负采样等技术设计了多种模型，包括 Word2Vec+LSTM+CRF 模型、Word2Vec+LSTM+CRF 模型以及 BERT 预训练模型 +CRF 模型。实验结果表明，目前 BERT+CRF+负采样模型表现最佳，整体 F1 值达到 0.8979，关键数据 M、R、P、F 标签平均 F1 值达到 0.3543。此外，针对小样本数据集的特点，本工作初步探索了半监督学习的可行性。预计在获得更多数据后，模型的性能将表现更佳，为基于数据驱动的机器学习在材料领域的应用奠定基石，以加速新材料的设计与发现。

关键词：　小样本　材料领域　BERT　CRF　负采样　半监督

开源项目地址为https://github.com/Xzaohui/chemical_ner

2 引言

2.1 问题描述

从材料领域科学文献中准确抽取相关命名实体是对该领域知识进行深层次分析的基础，对材料属性预测、催化方案生成以及新材料发现等方面具有重大意义。然而，材料领域中相关实体名称组成复杂、结构嵌套，且缺乏大规模人工标注语料，为抽取统一、完整、准确的领域实体带来了困难。

要求基于从材料领域科技文献数据库中抽取的专家标记的语料库，设计基于深度学习（包括但不限于 CRF、LSTM、BiLSTM、Transformer 等）及预训练模型的智能模型与方法，完成材料领域的催化原料、催化反应、催化生成物、法拉第效率属性等命名实体的识别任务，助力材料领域创新。

2.2 解决方案

从数据和模型两个方面切入，我们构建了高质量的数据集与多种神经网络模型，力求达到任务要求。

在数据方面，为解决数据集的标注规则不统一和标注密度过低的问题，我们基于初赛数据集的分析与整理，提出了更适合小样本下材料领域实体命名的标注方案，并由相关专业的参赛队员人工标注约 500 份的高质量数据集。

在模型方面，我们采用谷歌公司最新开发的与化学相关的 BERT 模型作为预训练模型加上条件随机场 CRF 作为标签序列预测，同时以 Word2Vec+LSTM+CRF 和 Word2Vec+LSTM+CRF 模型作为 baseline 和最新的模型

作对比。

实验结果表明，目前 BERT+CRF+ 负采样模型表现最佳，整体 F1 值达到 0.8979，关键数据 M、R、P、F 标签平均 F1 值达到 0.3543。从损失函数曲线分析可知，进一步扩大高质量的数据集规模，模型的表现效果可以进一步提升。这充分表明我们整体设计方案的有效性。

2.3 创新点

1. 基于初赛数据集的分析与整理，提出了新的更适合小样本领域材料实体命名的标注规则：建立统一、逻辑合理且包含上下文推理的标注规则

a. 设立了额外的大标签进行多任务目标实体预测，以提高模型对主任务的预测效果

b. 广泛采用字典方式进行自动标注，辅助和加快繁琐的人工标注过程。

2. 使用了当下由谷歌公司开发的最新的 BERT 预训练模型，采用迁移学习与 fine-tune 微调的方法对模型进行训练，对小样本数据集更为友好。

3. 针对专业领域的标注数据集较小、相对复杂的标注规则以及大量的非实体标签（O），我们设计了 Self-training/Co-training 半监督学习和负采样技术对训练进行了优化，大大增加了对于极少出现的标签识别的准确率。

3 初赛数据集分析

3.1 初赛数据集统计

1. 数据集的样本总数分别为 243 份，由化学领域的小组成员和大赛相关人员一致确认数据集的主题为“基于 Cu 催化 CO2 还原”。

2. 数据集的大标签一共有 4 种，包括：材料种类，调控因素，产物种类和法拉第效率，小标签共计 32 种，大小标签出现的频数统计如下所示：

(1) 大标签统计情况，包括材料种类（Materials types，M），调控因素 (Regulatory factors，R)，产品种类 (Product categories，P) 以及法拉第效率（Faradaic efficiency，F)：

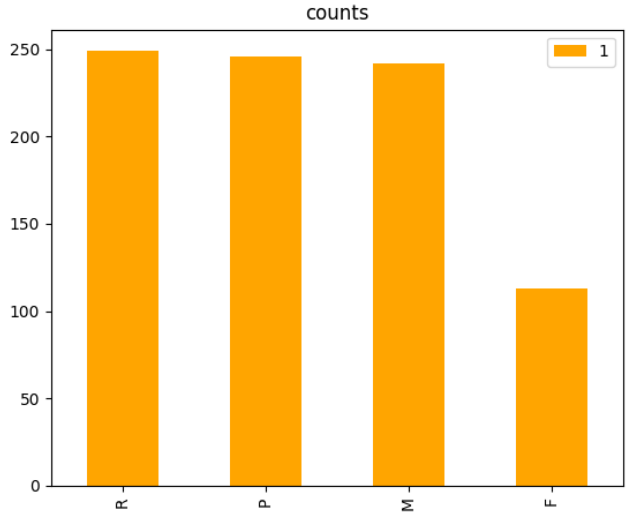

图 1: 大标签统计

(2) 小标签统计情况如下：

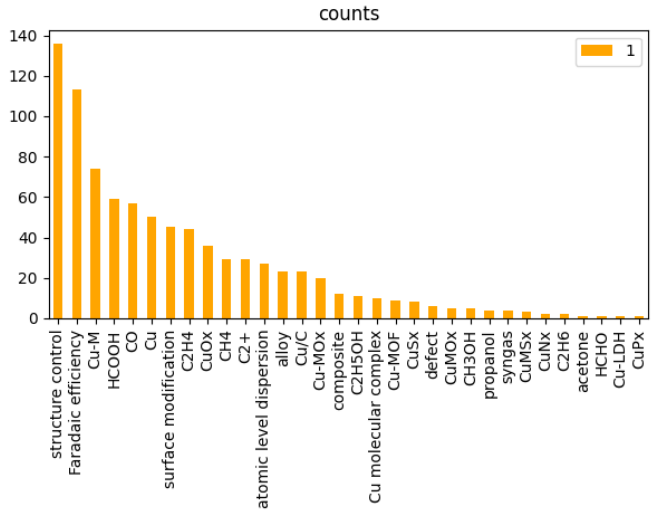

图 2: 小标签统计

3. 数据集的 data 内容中，一共出现 6161 种单词，出现的高频词（展示前 20 个）统计结果如下：

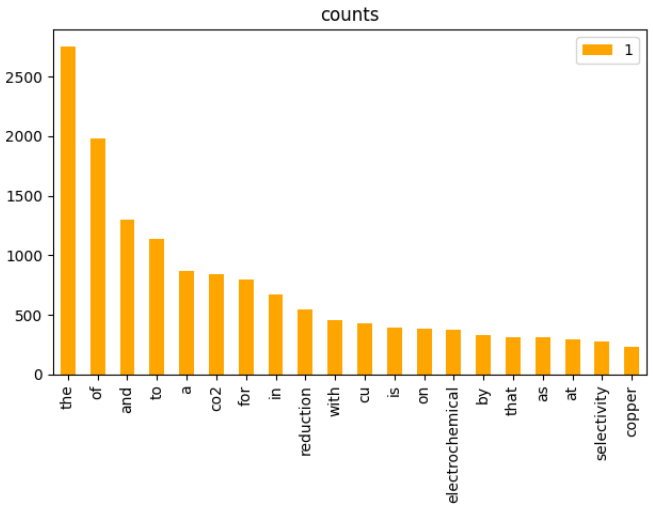

图 3: 高频词统计

4. 数据集中每个样本的标注次数平均为 3.5 次，而数据集的 data 标签内的字符总数分布图如下所示：

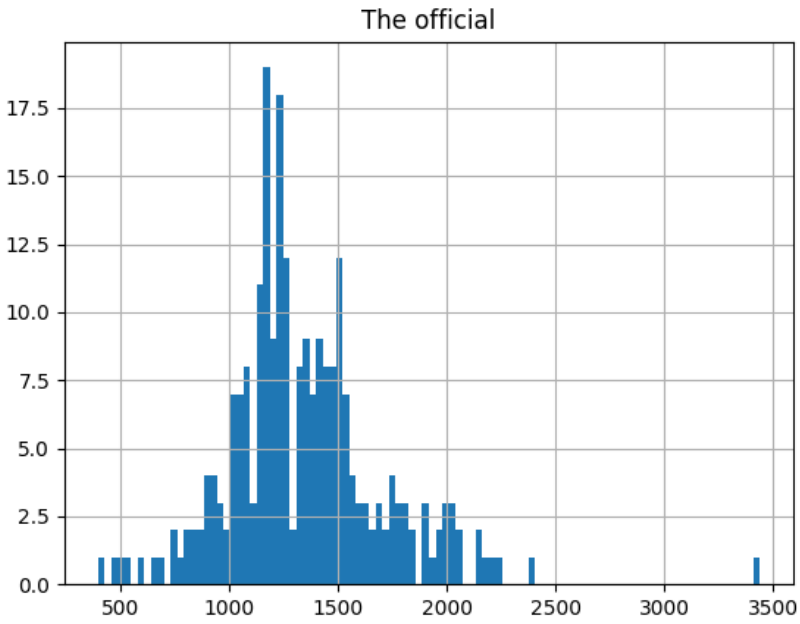

图 4: 字符总数统计

3.2 初赛数据集分析

1. 由图 4 可得，每个样本的字符总数分布主要从 500 到 2500 之间，而每个样本的平均标注次数为 3.5 次。表明标注的密度过低，数据中大部分单词的标注为非实体标签 O ，不利于模型训练。

2. 组委会标注逻辑混乱，标签标注存在规则前后不统一的情况，以 Faradaic efficiency 为例，有样本标记原文的长达 103 个字符（49.7% and a high current density of 28.5 ma cm(-2) at -1.19 v vs. a reversible hydrogen electrode

(rhe))，而有的样本则几个字符（90.24 % 或 17.9)。

3. 存在较多误标的情况，如 82.3% 应该标记为 Faradaic efficiency，但实际却被标记为"HCOOH"，以及手动输入错误导致标注了前后若干个不相关字符，如将"Cu," 标注为 M 等。

4. 初赛数据集漏标严重，重复出现的应被标注的单词以及短语并没有被标出。

基于上述分析，仅使用初赛的 243 份数据集，同时数据集存在大量问题，本项目难以顺利开展。因此，需要重新人工构建新的数据集并建立统一的标注规则。

3.3 数据收集

从 web of science 网站，以"metal" OR "cu" OR "copper" (主题) and catalyzed reduction (主题) and "carbon dioxide" OR "CO2" (主题) 为检索条件，检索与初赛数据集主题十分相关的新样本数据，作为我们下一步人工标注的数据来源，共计 1661 份（截止 5.17 日)。新数据集中仅与初赛数据集一共存在 17 篇重复。目前已完成人工完成 150 余份样本标注。

另外，从 web of science 网站，以 cu (主题) and copper (主题) or co2 (主题) or carbon (主题) and reduction (主题) and cata (主题) 为检索条件，检索到 1,178,425 份结果。依据相关性排序，下载了共 12,869 份的样本数据集，作为模型词向量的来源。

后续可以继续增加数据集，改进模型。

3.4 新数据集与初赛数据集比较分析

初赛数据集的 243 份数据中，出现了 6161 种单词。除了介词、冠词，以及已经是标签的除外，频数最高的分别是 ('CO2', 844) ('reduction', 463) ('electrochemical', 270) ('selectivity', 266) ('catalysts', 208) ('catalyst', 198)('surface', 190)('reaction', 164)；

新数据集的 1616 份数据中，出现了 49926 种单词。除了介词、冠词，以及已经是标签的除外，频数最高的分别是：('CO2', 2735)('reduction', 1664)('reaction', 1315)('catalytic', 895)('catalysts', 878)('catalyst', 818)('surface', 606)('catalyzed', 508)('reactions', 466)。

3.5 标签说明

本项目的关键大标签分别是材料种类（Materials types: M），调控因素 (Regulatory factors: R) , 产品种类 (Product categories, P) , Faradaic efficiency(F) 。其中材料种类就是催化剂的种类，有单金属、金属氧化物、金属硫化物、双金属、双金属合金、MOF 金属-有机框架材料等等；调控因素主要是指催化剂材料的形态结构的影响因素，目标数据集中结构控制、表面因素等都是主要影响因素，通过影响材料的结构决定材料的物理化学催化性能；产物种类分为一碳产物、二碳产物和多碳产物[3]，以及大量碳氢、碳氢氧化物，均为附加值较高的二氧化碳转化下游产品。

此外本项目添加了额外的四种大标签分：二氧化碳（CO2：CO2, carbon dioxide），还原（RE：reduction），电化学相关词语 (ELE: electrochemical, electrodes, electrode, electroreduction, etc.) 以及催化相关词语 (CA: cat-

alytic, catalysts, catalyst, catalyzed)。

3.6 标注规则

由于初赛数据集存在许多不利于模型训练的问题，诸如：每个标签标注的字符串存在逻辑不统一，有较多误标注和错标注以及标注密度过低等问题。因此，建立统一、逻辑合理且规范的标注规则十分必要。

首先，由于小样本的数量量过低，仅有数百份，每份的平均标记次数为 3.5，而小标签共计 32 种。因此，预测每个小标签几乎是难以做到事情。故出于保证模型有效的目标下，以预测四种大标签为目前的主要目标，后续得出的结果辅以其他模型或算法来判别小标签。

其次，由于一篇文章可能出现多种催化原料的相互比较或是多种催化产物，但文章往往强调某一种催化原料或催化产物是最可取方案，而其它物质是作为较差的对比项出现，即模型需要根据上下文推理或识别出正确的物质，如 Cu 在某一篇文章中可能是最佳的方案，但在另一篇文章中可能不是。因此，新标注规则不仅需要对样本中正确的物质或因素进行标记，也需要对样本中出现的“次要物质”或“次要因素”也进行标注。

最后，由于标注密度过低，非实体标签占比过高，即样本的标签分布极不平衡，这对模型训练存在影响很大。因此，需要补充额外的大标签。经过新数据集与初赛数据集比较分析后，结果表明两数据集中以 CO_2 ，$reduction$ ，$catalytic/catalysts/catalyst/catalyzed$ 以及 $(electrochemical, electrodes, electrode, electroreduction, electrocatalysts, electrocatalytic$ 这四类词出现频次最高，且与的样本主题（“基于 Cu 催化 CO_2 还原”）存在较强的

相关性，因此，选取这四类词作为额外的大标签，缓解标注密度过低的问题，同时能够辅助模型进行上下文推理，提高模型最终的预测效果。

进一步的统计分析表明，这四类额外大标签几乎可以使用字典进行自动标注。相似的，初赛数据集的分析结果表明 32 种小标签中，特别是部分的催化原料和产物种类，可以用穷举法进行完全枚举。因此，在实际的新样本标注过程中，首先使用初赛数据集的标注结果与基于词频统计的结果生成字典，采用字典对新数据集进行自动标注，再由人工进行核对与修改。

4 项目细节

4.1 任务介绍

本次项目任务小样本下材料领域实体识别是在自然语言处理领域很常见的序列标注任务，序列标注的的输入是一个序列，他的输出也是一个序列，他的典型的例子就是词性标注任务（pos tagging）和命名实体识别任务（ner）。对于序列标注任务这种时序问题最主要使用的模型就是循环神经网络 RNN，而 RNN 中的 LSTM 长短时记忆网络以及它的变种 BiLSTM 双向 LSTM 是对序列长时依赖更有效的模型，同时最近几年由谷歌公司开发的 BERT 模型也在此任务方向上取得了非常不错的效果。

4.2 模型选择

本次项目目前选择的模型为 BERT+CRF，传统 baseline 模型选择为 Word2Vec+LSTM+CRF 和 Word2Vec+BiLSTM+CRF。后续会根据项目进度实现更新的研究成果和模型。

4.3 模型介绍

4.3.1 LSTM/BiLSTM 模型

LSTM 的全称是 Long Short Term Memory，顾名思义，它具有记忆长短期信息的能力的神经网络。LSTM 首先在 1997 年由 *Hochreiter* 和 *Schmidhuber* 提出，由于深度学习在 2012 年的兴起，LSTM 又经过了若干代大牛的发展，由此便形成了比较系统且完整的 LSTM 框架，并且在很多

领域得到了广泛的应用[4] 。

LSTM 提出的动机是为了解决上面我们提到的长期依赖问题，较长的序列在传统 RNN 中循环输入输出，较早的序列信息对后续序列的影响较小，而 LSTM 引入了门（gate）机制用于控制特征的流通和损失，包含记忆 Cell、遗忘门、输入门、输出门等结构[4]。

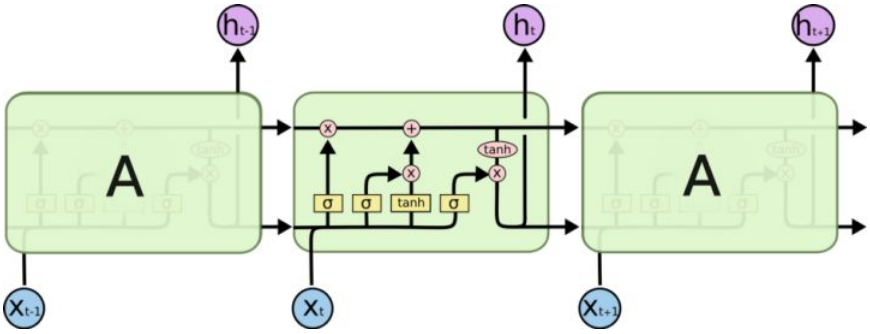

图 5: LSTM 模型结构

同时我们注意到无论是传统的 RNN 还是 LSTM，都是从前往后传递信息，这在很多任务中都有局限性，比如词性标注任务，一个词的词性不止和前面的词有关还和后面的词有关。为了解决该问题，设计出前向和方向的两条 LSTM 网络，被称为双向 LSTM，也叫 BiLSTM。其思想是将同一个输入序列分别接入向前和先后的两个 LSTM 中，然后将两个网络的隐含层连在一起，共同接入到输出层进行预测[5] 。

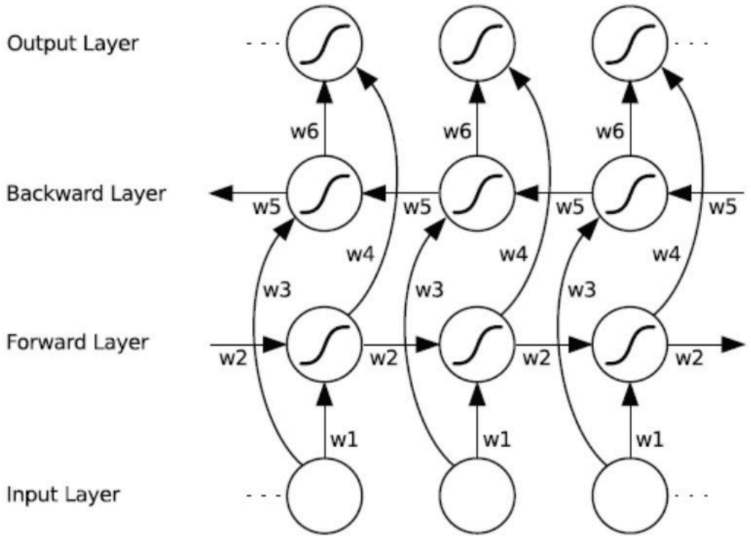

图 6: BiLSTM 模型结构

4.3.2 LSTM/BiLSTM 模型的输入与输出

LSTM 和 BiLSTM 模型的输入结构为 X=(Batch size,Max sentence len,Embedding dim)，由于样本量较小，我们选择较小的 Batch size=1 和较小的 learning rate，经过统计样本数据，我们将 Max sentence len 定为 600，Embedding 层我们使用 word2vec 作为预训练模型，Embedding dim 目前定为 256。

我们需要的输出为 Y=(Batch size,Max sentence len,Len of label)，而 Y(i, j, k)=p 的含义是第 i 个 Batch 下第 j 个单词的标签为 k 的概率。我们有 27 种标签，而 LSTM 的隐藏层 hidden dim=512，因此需要大小为 linear=(512,27) 的全连接层，而 BiLSTM 的隐藏层 hidden dim=64，因此需要大小为 linear=(128,27) 的全连接层。

4.3.3 BERT 模型

BERT 是 2018 年 10 月由 Google AI 研究院提出的一种预训练模型[6] 。BERT 的全称是 Bidirectional Encoder Representation from Transformers，可以知道 BERT 模型实际上是使用 transformer 作为算法的主要框架，双向的 Transformers 模型的 Encoder 部分，是一种典型的双向编码模型。它强调了不再像以往一样采用传统的单向语言模型或者把两个单向语言模型进行浅层拼接的方法进行预训练，而是采用新的 masked language model（MLM）和 next sentence prediction 的多任务训练目标，是一个自监督的过程，不需要数据的标注。使用 tpu 这种强大的机器训练了大规模的语料，是 NLP 的很多任务达到了全新的高度。

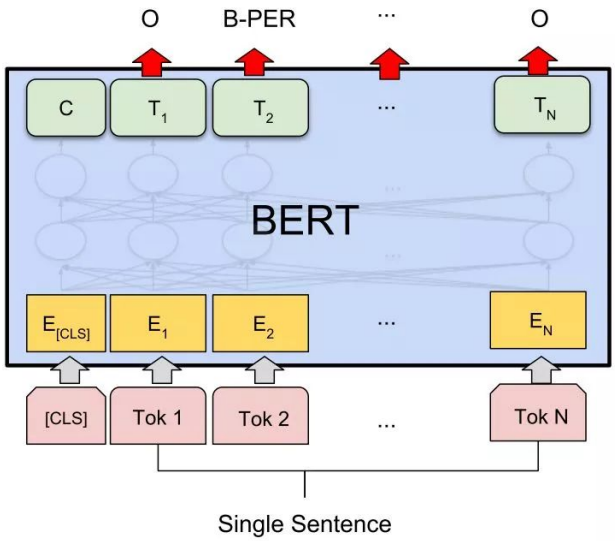

图 7: BERT 模型结构

BERT 模型十分巨大，模型分为 base 和 large 两个版本，base 版本由

12 层 Transformers 模型的 Encoder 部分组成，有 768 个隐藏层参数，总参数量有 1.1 亿个，而 large 版本由 24 层 Transformers 模型的 Encoder 部分组成，有 1024 个隐藏层参数，总参数达到了 3.4 亿个。

BERT 如此巨大的模型也需要庞大的数据量和计算资源，因此 BERT 模型一般是由大机构或研究所预训练完成后上传至 HuggingFace 作为开源预训练模型使用[7] ，我们可以比较方便地在自己的数据集上进行 fine-tune 微调。

BERT 模型的每个目标词是直接与句子中所有词分别计算相关度 (attention) 的，所以解决了传统的 RNN 模型中长距离依赖的问题。通过 attention，可以将两个距离较远的词之间的距离拉近为 1 直接计算词的相关度[8] ，而传统的 RNN 模型如 LSTM/BiLSTM 中，随着距离的增加，词之间的相关度会被削弱。

4.3.4 BERT 模型的输入与输出

BERT 模型的输入与 LSTM/BiLSTM 模型不同，输入的向量是由三种不同的 embedding 组合而成[6] ，分别是：

1. wordpiece embedding：单词本身的向量表示。WordPiece 是指将单词划分成一组有限的公共子词单元，能在单词的有效性和字符的灵活性之间取得一个折中的平衡。

2. position embedding：将单词的位置信息编码成特征向量。因为我们的网络结构没有 RNN 或者 LSTM，因此我们无法得到序列的位置信息，所

以需要构建一个 position embedding。构建 position embedding 有两种方法：BERT 是初始化一个 position embedding，然后通过训练将其学出来；而 Transformer 是通过制定规则来构建一个 position embedding

3. segment embedding：用于区分两个句子的向量表示。这个在问答等非对称句子中是用区别的。

虽然输入较 LSTM/BiLSTM 复杂，但 BERT 模型有其自带的 tokenizer 工具[6]，因此在输入方面比较简单，只需要将句子输入 tokenizer 中，就能将其分成 input ids、attention mask 和 offset mapping，后续只需要将 offset mapping 转化为我们的标签信息即可将这三项输入模型中。

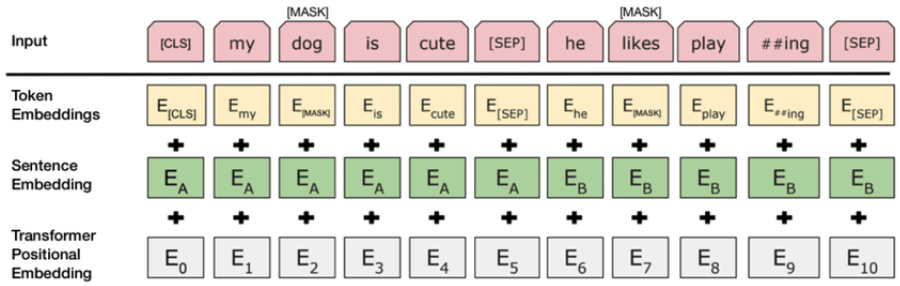

图 8: BERT 模型的输入与输出

BERT 模型的输出与传统 LSTM/BiLSTM 模型的输出相同，为 Y=(Batch size,Max sentence len,Len of label)，而 Y(i,j,k)=p 的含义是第 i 个 Batch 下第 j 个单词的标签为 k 的概率。

4.3.5 CRF 模型

条件随机场即 CRF 模型可以在在新的观测序列上找出一条概率最大最可能的隐状态序列[5] 。

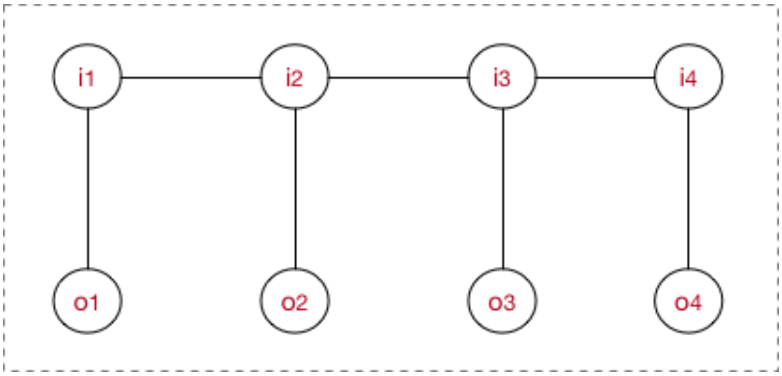

图 9: CRF 模型

CRF 与 HMM 隐马尔可夫模型结构类似，但 CRF 是无向图，HMM 为有向图，CRF 是判别式模型，HMM 为生成式模型，CRF 在如今的 NLP 领域应用更加广泛。

单独的 LSTM/BiLSTM 抑或 BERT 也能够通过输出各个 token 的各个 label 概率来预测标签的序列，但是它们都不能学习到标签之间的条件转移，而 CRF 是全局范围内统计归一化的条件状态转移概率矩阵，再预测出一条指定的 sample 的每个 token 的 label，因为 CRF 的特征函数的存在就是为了对 given 序列观察学习各种特征（n-gram，窗口），这些特征就是在限定窗口 size 下的各种词之间的关系，然后一般都会学到这样的一条规律（特征），因此加入 CRF 模型会大大提升整体的预测准确率和合理性[5] 。

CRF 的建模公式如下：

$$P(I \mid O)=\frac{1}{Z(O)} \prod_{i} \psi_{i}\left(I_{i} \mid O\right)=\frac{1}{Z(O)} \prod_{i} e^{\sum_{k} \lambda_{k} f_{k}\left(O, I_{i-1}, I_{i}, i\right)}=\frac{1}{Z(O)} e^{\sum_{i} \sum_{k} \lambda_{k} f_{k}\left(O, I_{i-1}, I_{i}, i\right)}$$

其分子为路径分数的指数，分母为归一化的整体路径分数的指数之和，即可以理解为:

$$p(l \mid s)=\frac{\exp [\text{score}(l \mid s)]}{\sum_{l^{\prime}} \exp \left[\text{score}\left(l^{\prime} \mid s\right)\right]}=\frac{\exp \left[\sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_{j} f_{j}\left(s, i, l_{i}, l_{i-1}\right)\right]}{\sum_{l^{\prime}} \exp \left[\sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_{j} f_{j}\left(s, i, l_{i}^{\prime}, l_{i-1}^{\prime}\right)\right]}$$

我们当然是想让正确路径的概率越接近于 1 越好，因此可以令 $loss = -\log(p(l \mid s)$ 作为损失函数训练模型。

4.4 Word2Vec

word2vec 是 Google 研究团队里的 Tomas Mikolov 等人于 2013 年的《Distributed Representations ofWords and Phrases and their Compositionality》以及后续的《Efficient Estimation of Word Representations in Vector Space》两篇文章中提出的一种高效训练词向量的模型，基本出发点和 Distributed representation 类似: 上下文相似的两个词，它们的词向量也应该相似，比如香蕉和梨在句子中可能经常出现在相同的上下文中，因此这两个词的表示向量应该就比较相似[9] 。

我们选择由 12,000 余篇相关领域的论文数据训练出 Word2Vec 预训练模型代替随机初始化的 Embedding 层，使得我们的词向量包含更大的信息

量帮助我们的模型学习更深层次的信息。后续也会继续增大数据量，提升模型的标注效果。

4.5 其他训练算法

4.5.1 Self-training/Co-training

半监督学习是一种介于监督式学习和无监督学习之间的学习范式，我们都知道，在监督式学习中，样本的类别标签都是已知的，学习的目的找到样本的特征与类别标签之间的联系。一般来讲训练样本的数量越多，训练得到的分类器的分类精度也会越高。但是在很多现实问题当中，一方面由于人工标记样本的成本十分高昂，导致了有标签的样本十分稀少。而另一方面，无标签的样本很容易被收集到，其数量往往是有标签样本的上百倍。半监督学习就是要利用大量的无标签样本和少量的有标签样本来训练分类器，解决有标签样本不足这个难题[10][11]。

将初始的有标签数据集作为初始的训练集，根据训练集训练得到一个初始分类器。利用初始分类器对无标签数据集中的样本进行分类，选出最有把握的样本，如本次项目实验中可以选择以路径分数为判断依据。而后将选择出的样本加入到有标签数据集中对模型进行训练，随后根据新的训练集训练新的分类器，重复步骤 2 到 5 直到满足停止条件（例如所有无标签样本都被标记完了）最后得到的分类器就是最终的分类器。

但由于试验结果不稳定，本次报告并没有加入此项训练方法，后续研读更多最新论文后可以继续改进。

4.5.2 负采样

NER 数据会存在大量漏标，实体标注应该算是 NLP 中比较复杂的，需要专业标注知识、需要统一标注规范。NER 数据中存在大量实体，标注员想要把所有实体都标注出来是不现实的，因此数据存在漏标也不可避免。特别是在专业领域小样本下的命名实体识别，在本身数据量较小的情况下更容易收到此类噪声的影响。把未标注的实体当作“负样本”就是一种噪声，因为漏标的实体不应当做标签为 O 的负样本来看待。

未标注实体问题会导致 NER 指标下降。主要有 2 个原因：一是实体标注量减少；二是把未标注实体当作负样本。其中第二个原因起主要作用。因此需要对所有非实体片段进行负采样（下采样）[12] 。这也很好理解：所有非实体片段中，有一部分可能是真正的、但未标注的实体（也就是未标注实体），但我们把能把它们都当作“负样本”看待，因此需要对所有非实体片段进行负采样。我们可令一个小概率（5%）将 O 随机标注为 M、R、P、F 产生一定量的类似于噪声的数据，同时保证在整个数据集上的概率归一化以保证不会产生过大的偏差。

后续对负采样的优化和改进仍在研究当中。

5 实验数据

项目实验平台为 python 3.6.13 torch 1.10.2 cuda 11.6

项目实验目前标签识别的整体测试的 F1 值如下：

模型	F1 score
W2V+LSTM+CRF	0.8793
W2V+BiLSTM+CRF	0.8958
BERT+CRF	0.8971
W2V+LSTM+CRF+ 负采样	0.8897
W2V+BiLSTM+CRF+ 负采样	0.8847
BERT+CRF+ 负采样	0.8979
BERT+CRF+ 负采样 +official data	0.9803

表 1: 各模型整体 F1 值

由表 1 可以看出各个模型学习效率较好，平均 F1 值可达到 0.88 以上，但组委会数据的训练结果 F1 值达到 0.98 以上十分反常，猜测是由于非实体标签 O 过多导致的部分过拟合现象，大量标签被标注为 O 使得 F1 值虚高。

项目实验目前关键标签（M、R、P、F）识别的测试值 F1 如下：

模型	B-M	I-M	B-R	I-R	B-P	I-P	B-F	I-F	average
W2V+LSTM+CRF	0.4667	0.0606	0.1034	0.2000	0.5882	0.2500	Nan	Nan	0.2086
W2V+BiLSTM+CRF	0.6218	0.1270	0.0364	Nan	0.6452	0.2609	Nan	Nan	0.2114
BERT+CRF	0.6667	0.2952	0.2295	Nan	0.6377	0.2449	Nan	Nan	0.2588
W2V+LSTM+CRF+负采样	0.6087	0.2192	0.1194	0.1463	0.6452	0.2609	Nan	Nan	0.2500
W2V+BiLSTM+CRF+负采样	0.4615	0.0869	0.0689	0.0357	0.6557	0.3333	0.2667	0.3636	0.2840
BERT+CRF+ 负采样	0.6949	0.2857	0.1852	0.1322	0.6769	0.2449	0.2789	0.3356	0.3543
BERT+CRF+ 负采样 +official data	0.5882	0.3636	Nan	Nan	0.7500	Nan	Nan	Nan	0.2127

表 2: 各模型关键标签以及平均 F1 值

由表 2 可以看出试验结果基本符合最初设想，W2V+LSTM+CRF 到 W2V+BiLSTM+CRF 再到 BERT+CRF 模型复杂程度增加，关键数据的识别率依次提高，同时在整体 F1 值变化不大的情况下，增加了负采样的模型在小样本下对少数标签的识别率和 F1 值，同时也验证了上一个对组委会数据训练结果的猜想，其 F1 值虚高，而对关键数据的识别率和 F1 值较低。

6 总结分析

横向对比各个模型，BERT+CRF 模型的 F1 值明显高于另外两种传统 W2V+LSTM+CRF/W2V+BiLSTM+CRF 模型，并且负采样技术可以增加标签的识别率，特别是对极少出现的标签识别率有质的提高，如 R、F 标签。

纵向来分析各个模型的数据可以得出目前三种模型的平均 F1 值在 0.88 以上，模型的学习效率较好。但大多是由于无关且变化较少的标签如 ELE、CA、CO2 等以及非实体标签 O 识别率高的影响，而关键标签 M、R、P、F 的识别率由于数据量过于稀少，并不十分理想。特别是由于官方数据标签稀少，如果只看整体的 F1 值官方数据非常漂亮，但是其关键数据识别率较我们的数据训练结果差很多，因此我们重新对数据进行处理是非常有必要的。

后续项目应继续完善数据标注，继续增加数据集，同时可以在半监督学习、不完全实体标注问题、机器阅读理解 MRC 模型等方向进一步研究学习，提升算法效率。

参考文献

[1] Manoj B Gawande, Anandarup Goswami, François-Xavier Felpin, Tewodros Asefa, Xiaoxi Huang, Rafael Silva, Xiaoxin Zou, Radek Zboril, and Rajender S Varma. Cu and cu-based nanoparticles: synthesis and applications in catalysis. *Chemical reviews*, 116(6):3722–3811, 2016.

[2] Hiroyuki Takeda, Claudio Cometto, Osamu Ishitani, and Marc Robert. Electrons, photons, protons and earth-abundant metal complexes for molecular catalysis of co2 reduction. *ACS Catalysis*, 7(1):70–88, 2017.

[3] Kun Jiang, Robert B Sandberg, Austin J Akey, Xinyan Liu, David C Bell, Jens K Nørskov, Karen Chan, and Haotian Wang. Metal ion cycling of cu foil for selective c–c coupling in electrochemical co2 reduction. *Nature Catalysis*, 1(2):111–119, 2018.

[4] Jürgen Schmidhuber, F Gers, and Douglas Eck. Learning nonregular languages: A comparison of simple recurrent networks and lstm. *Neural computation*, 14(9):2039–2041, 2002.

[5] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language un-

derstanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[8] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

[9] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[10] Bayu Aryoyudanta, Teguh Bharata Adji, and Indriana Hidayah. Semi-supervised learning approach for indonesian named entity recognition (ner) using co-training algorithm. In *2016 International seminar on intelligent technology and its applications (ISITIA)*, pages 7–12. IEEE, 2016.

[11] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

[12] Yangming Li, Lemao Liu, and Shuming Shi. Empirical analysis of unlabeled entity problem in named entity recognition. *arXiv preprint arXiv:2012.05426*, 2020.