基于领域迁移和任务迁移相结合的试验鉴定文本命名实体识别

徐建[1]，阮国庆[1]，吴蔚[1]，李晓冬[1]，梁木[1]

[1]中电莱斯信息系统有限公司
461629348@qq.com

摘要：试验鉴定领域的命名实体识别是指从文本中抽取出试验鉴定相关的关键文本片段，包括试验目的、被试对象的特点和武器系统的特性。然后由于该领域的特殊性以及复杂性，开放语料非常匮乏，标注过程极其繁琐，如何在小样本的情况下提高该领域的命名实体识别效果是本次 ccks 评测的重要目的。针对这项任务，我们提出了基于领域迁移和任务迁移相结合的命名实体识别模型，有效的融入了外部知识；并且采用了基于实体替换和伪标签的两种数据增强的技术，有效的提高了小样本情况下的命名实体识别效果。

关键词：命名实体识别; 小样本；领域迁移；任务迁移；数据增强

1 引言

命名实体识别是抽取文本中特定的文本片段并对该文本片段进行分类的过程。该任务是相对比较成熟一个任务，有众多的解决方案，主要包括基于 crf 的实体抽取，和基于指针网络的信息抽取模型。前者主要包括 bert 向量表示层、双向 lstm 层和 crf 层，但是如果源领域和目标领域具有不同的实体类型，需要替换 CRF 层，而且该模型抽取过程也不能很好的引入实体类型等的先验知识。

基于阅读理解抽取模型可以针对给定问题在文本中抽取答案片段。该任务可以看做是两个分类任务，预测答案的开始位置和结束位置. 近年来，许多任务都被转化为阅读理解任务。比如 McCann et al. [1] 将多种 nlp 任务转化为阅读理解模型 ，并取得不错的效果。Li et al., 2019a[2] 针对命名实体识别任务，他们利用机器阅读理解模型回答实体相关的问题。由于指针抽取网络与实体类型无关，所以很适合领域迁移的问题，我们采用此模型作为我们的骨架(backbone)模型。

以 BERT[3]为代表的模型包含预训练过程和微调过程，前者能够充分利用无监督预训练阶段学习到的语言先验知识，在微调时将其迁移到下游 NLP 任务上，开启了自然语言处理的预训练新范式。这种预训练加微调的机制有如下几种好处： 1、无监督数据很大，不需要标注语料，整个网络上所有文本都是潜在的训练数据（BERT 预训练使用的数据集共有 33 亿个字，其中包含维基百科）；2、训练好的语言模型能够学会语法结构、解读语义甚至可以理解指代消解。通过特征提起或者微调能够有效的训练下游任务并提升其表现；3、简化了下游任务的网络结构，比较容易完成下游任务的开发。

但是如果直接将 bert 在样本量较小的目标领域微调 finetune，会面临两个问题[4]：领域适配问题，因为 bert 的预训练主要是基于维基百科语料，其目标函数包含掩码语言模型和下一个句子预测，该过程主要是根据训练语料上下文学习单词表示。由于 bert 预训练主要是针对百科知识领域，该过程很可能对试验鉴定领域文本单词或者短语一无所知 ，这就会造成领域的不匹配；另一个是任务适配问题，此处我们利用阅读理解的方法来做命名实体识别，针对任务的适配主要是两层全连接网络，如果微调时候样本过少，模型并不能够充分学习这两层全连接网络的权重。

传统的文本数据增强[5]包括：基于单词（字符）的替换；新增单词（字符）；删除单词（字符）；交换单词（字符）和回译。虽然这些技术对于文章或者句子整体的分类有效，但是对于字符层面的标注或者 span 层面的分类这些技术需要特殊处理，因为所有的操作都不能破坏文本片段（span）的意思。为此我们将文本中的字符分为两种，标记为 O 的表示称之为非目标字符，反之未标记为 O 的乘坐目标字符。为了保证文本片段（span）的正确性，我们的所有数据增强操作只针对目标字符和非目标字符，不会针对二者的交叉处理。我们所采用的增强操作符包括:针对非目标字符的替换（REP）、增加（INS）、删除（DEL）、交换（SW）;和针对目标字符的 replacement（SP_REP）。

2 问题定义

输入：给定试验鉴定相关自然语言文本集合。

$$D = \{d_1, d_2 \ldots d_N\}, d_i =< w_{i1}, \ldots . w_{in} > \quad (1)$$

输出：实体提及和所属类别对的集合：

$$\{< m_1, c_{m1} >, < m_2, c_{m2} >, \ldots . < m_p, c_{mp} >\} \quad (2)$$

其中实体提及 mi =$\langle d_i, b_i, e_i \rangle$是出现在文档 d_i 中的试验鉴定实体提及（mention），b_i 和 e_i 分别表示 mi 在 di 中的起止位置，$c_{mi} \in C$ 表示所属的预定义类别。要求实体提及之间不重叠，即 ei < bi+1。

3 方法

3.1、实体类型无关的骨架模型

我们的网络结构采用基于双指针的信息抽取模型。根据实体类型构建问题，然后拼接段落作为输入，经过双指针网络，抽取答案。因为给定一个实体类型，段落中可以有多个答案，所在这里采用 bce loss，利用阈值 0.5，而不是采用交叉熵损失函数

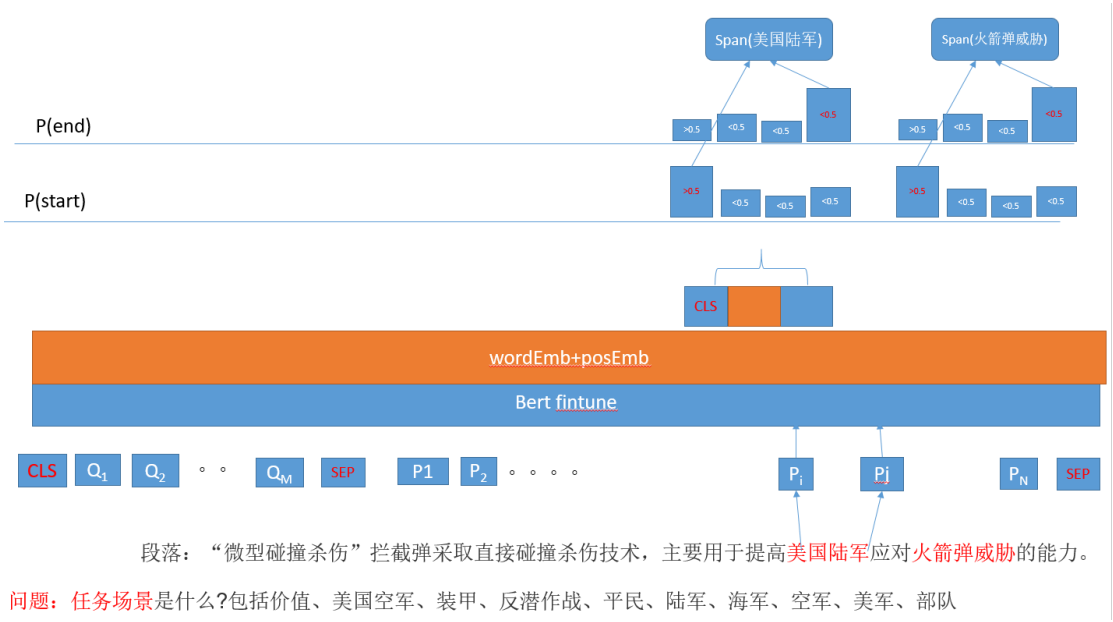

图 1.基于指针网络的信息抽取模型

我们知道神经网络是表征学习，而单独的 bert 是以字为单位进行编码，所以我们在这里引入了分词信息和词性标注信息，具体做法就是对段落进行分词和词性标注，每个单词内的所有字符共享该单词的词向量和词性标注信息，其中词向量来自于腾讯开源的词向量，只不过我们取了前一百万个词语的词向量；词性标注信息是随着网络训练的可学习参数其中维度固定为 8；此外我们将 CLS 向量拼接到每个字符中，使得每个字符带有全文信息。

3.2、领域适配和任务适配

我们引入了领域迁移和任务迁移，更好的融入先验知识。其中任务迁移我们利用微软命名实体识别语料[6]做任务迁移种类型的实体：人名+地名+机构名。我们利用图 1 的骨架模型做任务适配，该过程涉及的损失函数如下：

$$Loss^{s}_{boundary} = -\sum_{i=1}^{N}[y_s^i log P_s^i + (1-y_s^i)\log(1-P_s^i)] \quad (3)$$

$$Loss^{e}_{boundary} = -\sum_{i=1}^{N}[y_e^i log P_e^i + (1-y_e^i)\log(1-P_e^i)] \quad (4)$$

$$Loss_{boundary} = Loss^{s}_{boundary} + Loss^{e}_{boundary} \quad (5)$$

其中 $Loss^{s}_{boundary}$表示开始位置的损失函数，$Loss^{e}_{boundary}$表示结束位置的损失函数，$Loss_{boundary}$表示开始位置和结束位置损失函数之和；y_s^i表示字符 i 是否是实体的开始位置，如果是则标记为 1 否则标记为 0，y_e^i 表示字符 i 是否是实体的结束位置，如果是则标记为 1 否则标记为 0；P_s^i和P_e^i表示网络输出，P_s^i代表了位置 i 作为开始位置的概率，P_e^i代表了位置 i 作为结束位置的概率，得到开始和结束位置即能够得到文本片段。

对于领域迁移，主要就是利用掩码语言模型进行训练，为此我们防务快讯网站中的新闻语料，进行掩码语言模型训练。利用网页解析工具 BeautifulSoup 抽取其网页中 div 类型为 newsContent 的标签，得到网页正文，并过滤掉英文文章；将文章分段，并且保证每个段落长度少于 200 个字

符；利用掩码语言模型 masked language model 来预测随机掩盖掉的单词，该过程损失函数$L_{MLM}(\theta;D)$记为：

$$L_{MLM}(\theta;D)=\frac{1}{|D|}\sum_{X\epsilon D}\sum_{t=1}^{|X|}logp(x_t|X\backslash t) \quad (6)$$

其中|D| 表示所有样本个数；X 表示字符组成的单个样本；|X|表示样本单词个数，t 表示样本中每个单词，x_t是单词的 t 的向量表示，X\t 表示句子中去除掉 t 以后的剩余单词；对文章利用开源库 jieba 进行分词，对其中 15%单词进行 替换，具体包括 3 种形式的操作，80%的单词用[MASK]替换，10%的单词用随机的单词替换，10%的单词保持不变。最后我们利用二者损失函数之和作为预训练阶段的联合损失作为优化目标。

$$Loss=Loss_{boundary}+Loss_{MLM} \quad (7)$$

3.3、文本增强

为了应用文本增强操作，首先需要在原始文本中采样一个字符（或者 span），该过程我们称之为前采样；如果操作是替换 replace 或者交换 swap，还需要做一个后采样决定新的字符（或者 span）该过程称之为后采样。

采样做法有 3 种：均匀采样，也就是说每个字符或文本片段采样到的概率是均等的；依据重要程度的采样，根据字符或者文本片段的重要性采样，这个重要性一般是依据 TFIDF 或者其出现频率；基于语义相似度的采样，该采样主要是针对文本片段（span）的后采样来说的，后采样的文本片段主要基于该片段和源片段的语义相似性来度量，选择相似度高的文本片段替换原始文本。针对本次任务我们主要采用基于文本片段（span）的实体替换技术，前采样我们基于均匀分布的采样策略，后采样我们基于 bert 相似度采样策略。

表 1　前后采样策略

操作符	类型	前采样	后采样
基于 span 的替换	替换	均匀分布	Bert 相似度

3.4、伪标签技术

我们首先将数据做 5-折交叉验证，然后各自训练一个模型；然后各自模型抽取出的实体做投票得到融合后的结果。我们利用该融合技术对于测试数据进行抽取，并将抽取的结果作为训练语料的一部分，进行二次训练。具体过程如图所示

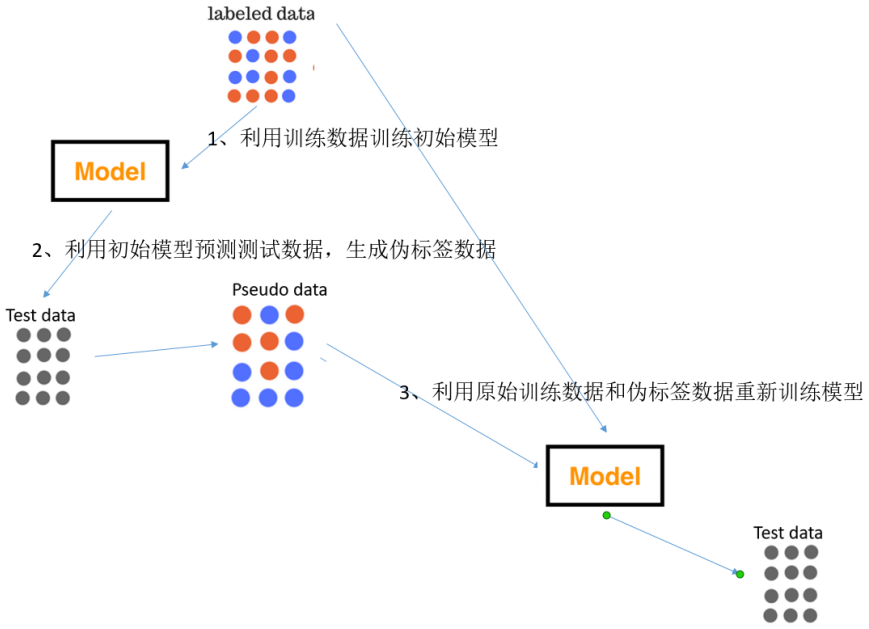

图 2.伪标签生成过程

4、试验

4.1 试验数据

本次评测面向试验鉴定的命名实体识别任务包含 400 条训练数据，其中包含 4 个实体类别分别是试验要素、性能指标、系统组成、任务场景。其中单个实体长度在 20 以内，单个样本长度在 276 个字符以内。

4.2 实验设置

对于不同类型问题构建，我们按照该类型词语的词频降序排列，然后将这些实体拼接使得长度不超过 100 个字符；对于段落我们限制字符个数在 280 以内，超过这个长度则截断；所以总计文本长度限制在 380 个字符。

对于预训练阶段：对于任务适配我们采用微软开源数据集 *SIGHAN06*，我们只取前 10000 个样本；对于领域适配我们爬取防务快讯网站 http://www.dsti.net/Information/NewsList/ 中的新闻语料，并选择其中 10000 个新闻样本。

对于微调 finetune 截断，我们采取 5 折交叉验证的策略，模型融合采用实体层面的投票策略。

4.3、试验结果分析

表 2　模型效果分析

多模型	领域适配	任务适配	数据增强	伪标签	F1(百分比)
√	×	×	×	×	69
√	√	√	×	×	70.18（+1.2）
√	√	√	√	×	70.98（+0.8）
√	√	√	√	√	71.75（+0.7）

试验都是基于多模型融合的基础上做得，可以看到领域适配和任务适配对于最后的结果提升有1个多点的提升，说明我们选择的骨架网络以及领域和任务的适配缺失提升了小样本的识别效果。此外数据增强和伪标签技术也可以有效提升模型识别效果。

5、结论

本文提供了一种基于领域迁移和任务迁移相结合的试验鉴定文本命名实体识别模型，主要包括以下4点：1、提出了一种通用的命名实体识别抽取模型，统一了源领域和目标领域因为实体类型不同所带来的的网络结构差异；2、充分利用现有可用的实体抽取标注数据和目标领域大量的无标注数据，实现领域适配和任务适配，解决了基于预训练+微调模型在小样本情况下面临的问题；3，利用实体替换技术实现训练增强，有效的增加了训练数据数量；4，结合伪标签技术实现两阶段训练的策略，有效提高了模型在测试数据的识别效果。

参考文献

- [1] B. Mccann, N. S. Keskar, C. Xiong, and R. Socher, “The Natural Language Decathlon: Multitask Learning as Question Answering,” arXiv: Computation and Language, 2018
- [2] Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2019a. A unified MRC framework for named entity recognition. CoRR, abs/1910.11476.
- [3] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
- [4] Xu H, Liu B, Shu L, et al. Bert post-training for review reading comprehension and aspect-based sentiment analysis[J]. arXiv preprint arXiv:1904.02232, 2019.

- [5] Miao Z, Li Y, Wang X, et al. Snippext: Semi-supervised opinion mining with augmented data[C]//Proceedings of The Web Conference 2020. 2020: 617-628.
- [6] Bob Carpenter. 2006. Character Language Models for Chinese Word Segmentation and Named Entity Recogntion. Proceedings of the 5th International Chinese Language Processing Workshop (SIGHAN). Sydney.