·第 5 代指挥信息系统信息服务技术专题·　　doi：10.15908/j.cnki.cist.2019.05.003

面向特定标注数据稀缺领域的命名实体识别*

刘哲宁　朱聪慧　郑德权　赵铁军

（哈尔滨工业大学计算机科学与技术学院　哈尔滨 150001）

摘　要：针对传统命名实体识别需要大量标注数据的问题，提出了一种标注语料稀缺条件下的命名实体识别方法。首先，基于远程监督思想，使用 2 个特殊字典对特定领域文本进行伪标注；然后，使用 BERT（来自 Transformer 的双向编码器表征）模型进行语义平滑扩展，并在含有噪音的伪标注语料中训练 AutoNER（自动伪标注的命名实体识别）模型；最后，通过与传统机器学习方法条件的随机场进行试验对比，验证了该方法的有效性。

关键词：命名实体识别；远程监督；语义向量；数据稀缺

中图分类号：TP391.1　　**文献标识码**：A　　**文章编号**：1674-909X(2019)05-0014-05

Named Entity Recognition for Specific Field with Annotated Data Scarcity

LIU Zhening　ZHU Conghui　ZHENG Dequan　ZHAO Tiejun

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Aimed at the problem of requiring a large amount of annotated data in traditional named entity recognition (NER), a NER method in condition of specific field with annotated data scarcity is proposed. Firstly, based on the idea of distant supervision, two specific dictionaries are used to pseudo-annotate texts in the specific fields. Then, the bidirectional encoder representations from Transformer (BERT) model is adopted to smoothly extend the semantic, and the automatic NER (AutoNER) model is trained in the noised pseudo-annotated corpus. Finally, experiment compared with the traditional machine learning method, conditional random field (CRF), verifies the validity of the method.

Key words: named entity recognition (NER); distant supervision; semantic vector; data scarcity

0　引　言

近年来，面向特定领域的垂直问答和知识图谱技术应用场景越来越多。与该领域匹配的命名实体识别工具成为后续智能信息处理技术能否实际应用的关键。大多数特定领域的命名实体识别面临如下相似现状[1-2]：1）难以建立足够多针对特定领域的标注语料，若通过人工构建标注语料，则标注成本高昂，这是因为特定领域有其特有背景知识，普通人很难直接、准确识别该领域实体；2）特定领域内常积累了部分未标注数据和领域词典，当需构建面向特定领域的问答和推理时，常在该领域构建领域词典，这使得很多特定领域存在特有的领域词典。

命名实体识别（NER）[3]又称专名识别，指识别

*基金项目：国家重点研发计划（2017YFB1002102）资助项目。

收稿日期：2019-08-14

引用格式：刘哲宁，朱聪慧，郑德权，等. 面向特定标注数据稀缺领域的命名实体识别[J].指挥信息系统与技术，2019，10(5)：14-18.

LIU Zhening, ZHU Conghui, ZHENG Dequan, et al. Named entity recognition for specific field with annotated data scarcity[J]. Command Information System and Technology, 2019, 10(5): 14-18.

出文本中实体的命名性指称项，并标明其类别。传统命名实体识别涉及的命名实体一般包括 3 大类（实体、时间和数字）和 7 小类（人名、地名、组织机构名、时间、日期、货币和百分比），需在海量无结构文本中确定最小语义逻辑单元。后续的智能信息处理过程关系抽取和属性分析均在实体基础上进行。实体识别是大多数智能信息处理系统的必备基础流程，如知识图谱的构建[4]、自动问答[5]和对话[6]等。

传统 NER 方法基于数据驱动，即标注的数据量越大，学习算法性能越好。但在某些特定领域，经常没有足够多的命名实体识别标注语料，故传统命名实体识别方法不适用。针对该情况，本文提出了一种不依赖于标注数据的 NER 工具训练构建方法，并引入预训练语言表征进行语义平滑，大幅提高了整体方案在不同垂直领域的推广能力，可以在标注数据不充足情况下，构建特定领域 NER 工具。本文选择医疗电子病历数据的 NER 任务作为试验的特定领域，试验结果表明了该方法的有效性。

1　总体解决流程

该方法处理流程如图 1 所示，分为训练和预测 2 个阶段。具体步骤如下：

1）训练阶段

（1）非人工伪标注语料构建：将远程监督的思想应用于 NER 任务，即构建一个知识库（字典），知识库中存储关系二元组（实体类型和实体名），使用该知识库对特定领域语料进行非人工伪标注；伪标注指机器按照知识库的知识自动为无标签文本进行实体自动标注，伪标注语料将代替人工标注语料作为模型监督学习的训练数据。

（2）语义向量提取：采用特征集成方式使用 BERT（来自 Transformer 的双向编码器表征）模型提取语料的语义向量；BERT 模型是 Devlin 等[7]于 2018 年提出的面向自然语言处理任务的通用模型架构，该架构在当时的 11 项自然语言处理任务中夺得最佳评测结果。

（3）数据预处理：将特定领域语料的文本、文本表征和实体标注类型对齐，再输入 AutoNER（自动伪标注的 NER）模型；AutoNER 模型是本文为了适应在包含大量噪音的标注样本进行训练而提出使用的一个 NER 模型架构。

（4）模型训练：对 AutoNER 模型在包含噪音的标注语料中进行训练，训练结束后，在本地保存训练完成的模型参数。

2）预测阶段

（1）语义向量提取：采用特征集成方式使用 BERT 模型提取语料的语义向量；

（2）数据预处理：将特定领域待识别语料的文本和文本表征对齐，再输入 AutoNER 模型；

（3）模型预测：待 AutoNER 模型导入训练参数后，对输入模型的文本进行实体识别。

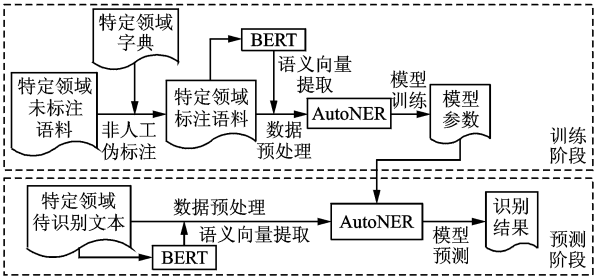

图 1　处理流程

1.1　远程监督

远程监督（distant supervision）由 Mintz 等[8]在关系抽取[9]任务中首次提出。关系抽取（information extraction）是 NER 基础上的一个任务，指自动识别实体间具有的某种语义关系。对于训练一个简单的限定域关系抽取器，需明确待识别关系的 2 个目标实体和出现的文本上下文，且待识别的语义关系是预定义的。因传统的有监督限定域关系抽取模型无法自行完成关系命名，故需依赖人工标注的数据来完成实体间关系命名。当模型训练完成后，若该模型在句中再遇到相应实体对，会自行完成关系抽取。

远程监督属于一种弱监督方式[10]，利用外界庞大的知识库来指导标注或扩充训练样本集。其基本思想是，假设存在 2 个实体，如果这 2 个实体能够和知识库中存在的实体对匹配，那么这 2 个实体同时包含了在知识库中匹配的实体对关系。

将远程监督的思想应用于 NER 任务，即需要自行构建一个知识库，字典中存储标注知识关系二元组（实体类型和实体名）。对于给定一个数据集，需按照一定顺序对数据集进行遍历，当在某一位置的实体与字典中存储的实体名匹配时，则将该命名实体标注于字典中存储的相应实体类型。

1.2　语义向量提取模型

引入 BERT 作为语义向量提取模型的主要目

的是对词典中词条进行语义平滑，提高词典的语义匹配能力。BERT 模型的目标是利用大规模无标注语料训练、获得包含丰富语义信息的文本表征，即文本的语义表示，然后将文本的语义表示在特定自然语言处理任务中进行微调(fine-tuning)或特征集成(feature-ensemble)，最终应用于该特定自然语言处理任务。

1.2.1 特征表示

BERT 模型的输入是文本中各字/词的原始字/词向量，该向量既可随机初始化获得，也可利用 Word2Vec[11] 等模型进行预训练获得；BERT 模型输出是文本中各字/词融合了全文语义信息后的向量表示，BERT 输入层如图 2 所示。

BERT 模型的输入层由子词嵌入(token embeddings)、段嵌入(segment embeddings)、位置嵌入(position embeddings)组成。其中，子词嵌入使用 WordPiece[12] 将文本分词，再查询向量表将文本中每个词转换为一维向量(WordPiece 是一种分词方法，将单词划分成一组有限的公共子词单元，能在单词有效性和字符灵活性间取得一个折中的平衡)；位置嵌入指将词的位置信息编码成特征向量，是向模型中引入单词位置关系重要的一环；段嵌入用于区分文本所属句子，例如，假设 B 是 A 的下文(对话或问答场景等)，对于 A、B 这组句子对，默认第 1 个句子的段嵌入值为 0，第 2 个句子的段嵌入值为 1。

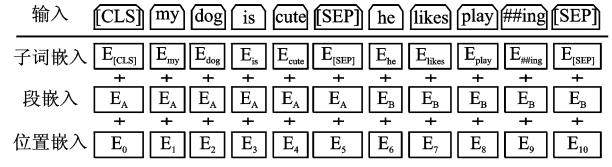

图 2 BERT 输入层[7]

1.2.2 模型训练

BERT 模型使用方式包括特征集成和微调 2 种。2 种方式区别如下：

在下游任务中使用预训练模型时，若采用特征集成方式，即将当前输入文本按照输入要求构造成向量形式，在通过 BERT 模型网络的逐层后，将每个输入单词对应位置的高层长短期记忆单元激活的嵌入值(或将输入单词对应位置的若干隐藏层的激活值进行加权求和获得的嵌入值)作为下游任务特定模型该单词对应的输入。这是一种典型的应用预训练模型方法，更侧重于单词的上下文特征表达方面。

在下游任务中使用预训练模型时，若采用微调方式，即在获得 BERT 模型及对应网络结构(Transformer[13])后，第 2 个阶段仍采用与预训练过程相同的网络结构。当处于训练阶段时，将该任务的部分训练数据直接用在该网络上进行模型训练，从而针对性修正预训练阶段获得的网络参数，一般该阶段称为微调。这是另一种典型应用方法。

特征集成模式是引入普适外部知识的过程，通过在大规模数据上的无监督训练得到更加准确全面的语义表示特征，对未登录词处理性能有提升。微调模式是让引入的外部知识更加适合具体任务，通过在一定规模数据上的有监督训练对外部语义表示进行微调以提升具体任务性能。前者是为了提升模型的推广能力；后者是为了提升性能，是一个相互影响的整体解决方案。因此，本文提出的识别方法选择特征集成方式，即将中文文本直接输入训练好的 BERT 模型，取其每个输入单词对应位置的高层长短期记忆单元激活的嵌入值作为下游任务特定模型的输入文本表征。

1.3 AutoNER 模型

AutoNER 是 Shang 等[14] 提出的命名实体识别模型，AutoNER 可以在非人工标注、包含噪音标注样本的数据上实现模型训练。AutoNER 在当时提出的 3 个数据集上均与有监督学习的模型基准评价效果相当，同时也是当时所有使用字典进行数据回标中效果最好的模型。

1.3.1 标注模式

AutoNER 使用新的标注模式——相关或无关模式来判断实体名边界。该模式具体做法是：

1) 对于相邻 2 个词语，如这 2 个词语指的是同一实体，则将二者间关系标上“相关”标签；

2) 如二者间至少有 1 个词语属于不知道类型但可能是待识别实体的词语，则将二者间关系标上“未知”标签；

3) 如二者关系不符合前 2 种情况的任何一种，则将二者间关系标上“无关”标签。

1.3.2 识别模型

AutoNER 采用字、词向量拼接的向量作为输入；隐藏层为 2 层门控循环神经网络；由于 AutoNER 模型采用 2 阶段方式，即先判断文本中命名实体的边界，再基于命名实体边界结果判断实体类型，

故其输出层包括如下 2 个独立部分：

1）sigmoid 层

sigmoid 层用于估计标签为“无关”(Break)的概率，即判断实体边界。sigmoid 层计算公式如下：

$$p(y_i = \text{Break} \mid \boldsymbol{u}_i) = \sigma(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{u}_i) \tag{1}$$

其中，y_i 为第 i 个词与其前一个词间的标签；$\sigma()$ 为 sigmoid 函数；$\boldsymbol{W}$ 为 sigmoid 层的权重参数。

实体名边界检测的损失可用下式计算：

$$L_{\text{span}} = \sum_{i \mid y_i \neq \text{Unknown}} l(y_i, p(y_i = \text{Break} \mid \boldsymbol{u}_i)) \tag{2}$$

其中，$l()$ 为 logistic 损失函数。计算时忽略“未知”标注位置。

2）softmax 层

在得到候选实体边界后(即经过 sigmoid 层计算后)，softmax 层用于计算相应命名实体对应各种实体类型的概率。softmax 层计算公式如下：

$$p(\boldsymbol{t}_j \mid \boldsymbol{v}_i) = \exp(\boldsymbol{t}_j^{\mathrm{T}}\boldsymbol{v}_i) \Big/ \sum_{t_k \in L} \exp(\boldsymbol{t}_k^{\mathrm{T}}\boldsymbol{v}_i) \tag{3}$$

其中，$\boldsymbol{t}_j$ 表明一个实体类型；L 为包含无标签在内的所有实体类型集合。

实体类型预测的损失可由下式计算：

$$L_{\text{type}} = H(\hat{p}(\cdot \mid \boldsymbol{v}_i, L_i), P(\cdot \mid \boldsymbol{v}_i)) \tag{4}$$

其中，$H()$ 为交叉熵损失函数；$\hat{p}(\cdot \mid \boldsymbol{v}_i, L_i)$ 为软监督分布(soft supervision)，定义为如下形式：

$$\hat{p}(\boldsymbol{t}_j \mid \boldsymbol{v}_i, L_i) = \frac{\delta(\boldsymbol{t}_j \in L_i)\exp(\boldsymbol{t}_j^{\mathrm{T}}\boldsymbol{v}_i)}{\sum_{t_k \in L}(\delta(\boldsymbol{t}_k \in L_i)\exp(\boldsymbol{t}_k^{\mathrm{T}}\boldsymbol{v}_i))} \tag{5}$$

其中，$\delta(\boldsymbol{t}_j \in L_i)$ 为布尔函数，表明在远程监督中的候选实体 i 是否标注为实体类型 $\boldsymbol{t}_j$。

2　试验验证与分析

2.1　试验设置

本文提出的解决方案适用于标注数据稀缺且易于构建知识库的特定领域。选择医疗领域数据进行试验验证方案性能。试验使用的医疗领域标注数据来源于 2017 年举行的全国知识图谱与语义计算大会(CCKS)中电子病历 NER 评测任务的公开数据集。该任务要求是对于给定的一组电子病历文档(纯文本文件)，识别并抽取出与医学临床相关的实体指称(entity mention)，并将它们归类至预定义的类别(predefined categories)，如疾病、症状和检查等。该公开数据集包括训练集和测试集 2 个部分，其中训练集有 50 余万字，测试集有 4 万余字，其中共有 5 种待识别实体类型，分别是药品名、疾病名、症状名、治疗方式名和身体部位名。在进行远程指导学习时，去掉了所有标注信息，作为无标注数据使用。

试验包含 3 种基线：Shang 等[14]提出的试验结果、条件随机场算法试验结果(以字为输入单位)[15]以及不使用 BERT 模型生成的文本表征作为输入的 AutoNER 模型试验结果。

条件随机场试验将文本以字作为基本输入单位，使用开源程序 CRF＋＋ 0.58 进行试验。AutoNER without BERT 试验使用“字＋词”作为基本输入，模型输入为随机初始化后字、词向量的拼接向量。BERT＋AutoNER 试验使用“字＋词”作为基本输入，模型输入为 BERT 提取的文本表征和随机初始化后的词向量的拼接向量。试验超参设置如表 1所示。其中，RNN 为循环神经网络；若损失值连续 5 轮学习后仍不下降，则学习率进行衰减；分词工具 pkuseg[16]使用医药领域语料训练得到的模型参数。

表 1　试验超参设置

超参名称	数值
随机失活(dropout)[17]	0.5
学习率	0.05
学习衰减率	0.9
学习算法	基于动量的随机梯度下降
RNN 基本单元	门控循环单元[18]
隐藏层单元数量	400
词向量维度	400
字向量维度	768
是否使用归一化	是
分词工具	pkuseg

2.2　试验结果分析

医疗领域语料试验评测结果如表 2 所示。从条件随机场与 BERT＋AutoNER 的试验结果对比看，在人工标注语料上训练、作为监督学习方式的 CRF 算法的试验结果表现最好，远程监督非人工伪标注语料上训练的 AutoNER 模型试验效果虽不如条件随机场，但效果不差，这主要因为该试验的训练数据量不够多。一般地，深度学习在大数据集上训练会有优势，而在小数据集上训练效果可能不如传统机器学习算法。此外，知识库知识质量也会直接影响试验结果，若知识库的标注知识仅包含待标注语料所属领域知识，这将大大提升该方法针对特定领域 NER 试验结果的精确率；若知识库的标注知识包含一些可能的、适量的、与其他领域交叉的标注知识，这将提升该方法试验结果的召回率。

表 2 医疗领域语料试验评测结果 %

方法	测试集 F1 值	精确率	召回率
Word2Vec+AutoNER	84.80	88.96	81.00
条件随机场	92.06	92.35	91.78
AutoNER	50.61	60.98	41.53
BERT+AutoNER	88.56	87.08	90.10

3 结束语

本文研究了标注数据稀缺的特定领域 NER 方法。针对大部分特定领域中文标注语料匮乏、人工标注代价昂贵等问题，采用远程监督的思想，引入 BERT 方法进行语义平滑，为标注数据缺失领域构建 NER 工具提供了完整解决方案。该方案可应用于军事指挥等标注数据稀缺领域。

参考文献(References)：

[1] 刘浏，王东波. 命名实体识别研究综述[J]. 情报学报，2018，37(3)：329-340.

[2] 薛天竹. 面向医疗领域的中文命名实体识别[D]. 哈尔滨：哈尔滨工业大学，2017.

[3] 赵军，刘康，何世柱，等. 知识图谱[M]. 北京：高等教育出版社，2018.

[4] 刘峤，李杨，段宏，等. 知识图谱构建技术综述[J]. 计算机研究与发展，2016，53(3)：582-600.

[5] 曾帅，王帅，袁勇，等. 面向知识自动化的自动问答研究进展[J]. 自动化学报，2017，43(9)：1491-1508.

[6] LÓPEZ-CÓZAR R，CALLEJAS Z，GRIOL D，et al. Review of spoken dialogue systems[J]. Loquens，2014，1(2)：12.

[7] DEVLIN J，CHANG M W，LEE K，et al. BERT：pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of Conference on the North American Chapter of the Association for Computational Linguistics. Minneapolis：NAACL，2019：4171-4186.

[8] MINTZ M，BILLS S，SNOW R，et al. Distant supervision for relation extraction without labeled data[C]//Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Stroudsburg：ACL，2009：1003-1011.

[9] 鄂海红，张文静，肖思琪，等. 深度学习实体关系抽取研究综述[J]. 软件学报，2019，30(6)：1793-1818.

[10] 王政，朱礼军，徐硕. 实体关系的弱监督学习抽取方法[J]. 中国科技资源导刊，2018，50(2)：103-110.

[11] MIKOLOV T，CHEN K，CORRADO G，et al. Efficient estimation of word representations in vector space[C]//Proceedings of the International Conference on Learning Representations. Scottsdale：ICLR，2013：1-12.

[12] WU Y H，SCHUSTER M，CHEN Z F，et al. Google′s neural machine translation system：bridging the gap between human and machine translation[EB/OL]. (2016-09-26)[2019-07-29]. https://arxiv.org/pdf/1609.08144.

[13] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[C]//Proceedings of Conference on Neural Information Processing Systems. Long Beach：NIPS，2017：5998-6008.

[14] SHANG J B，LIU L Y，RENX，et al. Learning named entity tagger using domain-specific dictionary[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg：ACL，2018：2054-2064.

[15] LAFFERTY J D，MCCALLUM A，PEREIRA F C N. Conditional random fields：probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 18th International Conference on Machine Learning. San Francisco：Morgan Kaufmann Publishers Inc.，2001：282-289.

[16] LUO R X，XU J J，ZHANG Y，et al. PKUSEG：a toolkit for multi-domain Chinese word segmentation[EB/OL]. (2019-06-27)[2019-07-29]. https://arxiv.org/pdf/1906.11455.

[17] SRIVASTAVA N，HINTON G，KRIZHEVSKY A，et al. Dropout：a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research，2014，15(1)：1929-1958.

[18] CHO K，Van MERRIŠNBOER B，GULCEHRE C，et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg：ACL，2014：1724-1734.

作者简介：

刘哲宁，男(1997—)，硕士研究生，研究方向为自然语言处理。

朱聪慧，男(1979—)，博士，讲师，研究方向为自然语言处理和机器翻译。

郑德权，男(1968—)，博士，副教授，研究方向为自然语言处理和信息抽取。

赵铁军，男(1962—)，博士，教授，研究方向为自然语言处理和人工智能。

(本文编辑：李素华)