融合迁移学习与主动学习的金融科技实体识别方法

石教祥 朱礼军 魏 超 张玄玄
（中国科学技术信息研究所，北京 100038）

摘要：命名实体识别为推动智能系统建设和科技情报服务起到重要作用。针对领域实体识别存在的标注成本高、识别准确率不高问题，从引入通用领域信息、削减孤立点影响的角度出发，设计基于语义相似度与不确定性度量的主动迁移学习方法。该方法结合预训练迁移学习模型来提高分类准确性，通过融合主动学习采样策略来减少标注成本。利用金融科技和通用领域语料库进行一系列实验，实验结果表明该方法能够有效地提高识别准确率，减少标注成本。

关键词：命名实体识别；少样本；主动学习；迁移学习；BERT
DOI：10.3772/j.issn.1674-1544.2022.02.005 **CSTR：**15994.14.issn.1674.1544.2022.02.005
中图分类号：G35 **文献标识码：**A

FinTech Named Entity Recognition Based on Transfer Learning and Active Learning

SHI Jiaoxiang, ZHU Lijun, WEI Chao, ZHANG Xuanxuan
(Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract: Named Entity Recognition (NER) plays an important role in promoting the construction of intelligent systems and scientific and technical information services. Aiming to solve the problems of high labeling cost and low recognition accuracy in named entity recognition in special fields, we propose a novel sampling framework called Active Transfer Learning based on Semantic Similarity and Uncertainty (ATL-SSU) from the perspective of adding extra-semantic information in general field and reducing the impact of outliers. This method combines a pre-trained transfer learning (TL) model to improve classification accuracy, and integrates active learning (AL) sampling strategies to reduce labeling costs. We perform a series of experiments on FinTech corpus and general corpus. The results show that our method can effectively improve the performance and reduce the annotation costs.

Keywords: Named Entity Recognition, Few-shot, active learning, transfer learning, BERT

0 引言

命名实体识别（Named Entity Recognition，NER），又称作专名识别、命名实体，是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词以及时间、数量、货币、

作者简介：石教祥（1995—），男，中国科学技术信息研究所硕士生，研究方向为知识工程与知识发现；**朱礼军**（1973—），男，博士，中国科学技术信息研究所研究员，研究方向为Semantic Web、Web Service和知识技术在科技信息服务中的应用；**魏超**（1985—），男，硕士，中国科学技术信息研究所助理研究员，研究方向为文本表示、知识图谱、自动问答（通信作者）；**张玄玄**（1995—），女，硕士，中国科学技术信息研究所研究实习员，研究方向为科技政策与科技战略。

基金项目：国家重点研发计划项目“颠覆性技术感知响应平台研发与应用示范”课题“地平线扫描系统”（2019YFA0707202）；中国博士后科学基金第 65 批面上项目“流形正则化自编码政策文本表示及主题词抽取方法研究”（2019M650804）。
收稿时间：2021 年 7 月 6 日。

比例数值等。NER是语义知识库、知识图谱的基本组件，旨在发现自然文本中的专有名词并将其归类到预定类别中。在大数据时代，面向公开领域的NER研究相对成熟，已作为知识库构建的一项关键技术为机器翻译、自动问答等应用系统提供底层支撑，但在如金融科技（Financial Technology，FinTech）、生物医药、军事等专业领域，往往缺乏可直接用于模型训练的数据集；在特定领域，由于专业性强，重新标注数据依赖领域专家，这种劳动密集且耗时的缺陷制约了NER的快速发展。因此，针对特定领域，如何利用少量标注样本进行NER研究就显得十分必要。

在现有的研究中，学者们通常利用少样本（Few−shot）学习思路来解决少量标注情况下的分类任务，少样本NER方法从变换特征和增强数据质量的角度划分为迁移学习（transfer learning，TL）和主动学习（active learning，AL）两大类[1−2]。TL的核心思想是将在源域数据上建立的知识模型复用至目标领域，以实现模型共建和知识共享。AL则利用渐进式采样抽取“不确定性”（uncertainty）高、易混淆的样本进行标注，迭代训练模型进而提高拟合能力[3]。TL在通用领域NER任务中表现良好，但是TL依赖领域之间的强相似性，当源域与目标域数据差异较大时，仅仅通过TL模型很难捕获到丰富的领域信息，模型之间知识迁移适应性较差。AL通过计算样本的“不确定性”程度进行标注，从而提升单一领域数据的质量，但是基于“不确定性”原则选取的样本没有考虑领域实体中的离群孤立点现象。此外，由于面向单一领域数据，AL难以充分利用领域外海量数据中蕴含的知识信息，限制了模型效果的进一步提升。

在金融科技等专业领域仅仅使用TL或者AL方法不足以实现模型最优，为此本文提出一种全新设计的主动学习采样策略，并与TL方法进行融合形成统一框架，即基于语义相似度与不确定性的主动迁移学习方法（Active Transfer Learning method based on Semantic Similarity and Uncertainty，ATL−SSU）。该方法在提升单一领域内数据的信息量的同时，将域外海量知识进行整合，提升NER效果。

本文的主要贡献是提出了融合迁移学习和主动学习的统一框架，并提出了更加全面的主动学习采样策略。在NER任务中，迁移学习利用海量的外部知识来训练模型，主动学习通过增强同一领域数据质量提高分类器性能，两者结合将充分利用领域内和领域之间的信息。因此，本文提出了一种融合的分类框架：基本分类器由BERT（bidirectional encoder representations from transformers）[4]和Bi−LSTM−CRF（bidirectional long−short term memory with a conditional random field）[5]串联组成，其中BERT是基于海量通用领域知识构建的预训练语言模型，Bi−LSTM−CRF是特征学习器。之后，利用主动学习采样策略迭代输入语料对模型进行微调以提高模型性能。此外，本文针对主动学习面临的“不完全特征描述”“离群孤立点”等问题，提出了基于联合语义相似度的主动学习采样策略。在度量计算中，联合考虑未标注样本和已标注样本的信息含量和语义距离，这种联合利用较少的样本可以更加充分地拟合模型，进而减少标注成本。

1　相关的研究

NER是一项较为成熟的研究，相关的模型层出不穷。近年来，深度学习的兴起带来了一波又一波的技术浪潮，它们在通用的NER任务中表现出色，然而这些模型的训练往往依赖于大规模标注数据集，在缺乏足量标注数据集的专业领域上容易发生过拟合现象。因此，也有许多研究集中在少样本学习，甚至零样本学习（zero−shot learning）中[6]。

在少样本NER任务中，迁移学习利用领域相似性，利用分布式词表示构建词共享语义空间，然后再迁移神经网络的参数至目标领域，实现领域之间数据共享和模型共建。按照迁移知识表现形式的不同，迁移学习NER方法大致可分为基于微调、元学习和特征变换的方法。基于微调的迁移学习方法是通过训练大规模语料构建共

享语义空间来实现知识的迁移，如Giorgi等[7]基于LSTM进行网络权重的迁移，首先将源域模型参数迁移至目标领域初始化，之后进行微调使适应任务需要。最近，也有不少学者利用预训练（pre−trained）迁移学习模型来实现微调，预训练模型充分利用了词义和语义特性，能强有力地捕捉潜在语义和句子关系，这种语境化的词嵌入在NER任务中表现突出[6]。基于元学习的迁移学习方法将学习水平从数据提升至任务层面，学习归纳有关跨任务数据更一般的规律性，这种方式试图建立一种在不同任务间都具有良好表现的模型。而基于特征变换的迁移学习方法主要解决领域适配性差的问题，这种方式通过特征互相转移或者特征映射来减少领域之间差异。也有不少学者从跨领域、跨应用、跨语言等角度测试迁移的可行性，还有利用诸如本体库、知识库、启发式规则等外部知识来解决少样本NER问题[8−9]。

此外，主动学习也可被用来解决少样本NER问题。主动学习通过一定的度量方式对语料进行精炼以提高模型拟合效率。在学习过程中，经过种子语料训练过的基础分类器用来预测未标注数据，而选择器从预测样本池中选择出信息量大的样本交给领域专家进行人工标注，这些新样本被加入初始种子语料中进行新一轮的模型训练[10]。在整个过程中，选择器的采样策略最为关键。在当前的研究中，基于不确定性（uncertainty）的样本选择方法是最常用的策略。其基本思想是选择当前模型易混淆、置信度低的样本。如在二分类任务中基于不确定的策略倾向于选择后验概率接近0.5的样本，而对于多分类序列标注任务通常利用信息熵（information entropy，IE）来度量样本的不确定性程度，熵值大的优先被挑选。如Chen等[11]在生物医学文本上利用不确定性标准度量样本的信息量，这种方式通过降低统计学习的期望误差对未标记样本进行优化选择，能够有效减少标注数据的工作量。基于主动学习的NER本质上增强同一分布数据质量，选择出信息量最大的一部分样本进行训练，在缺乏标注数据时能节省一部分标注成本。

整体上，基于TL和AL的方法都能在一定程度上解决少样本问题。其中，TL利用海量外界通用领域知识来辅助NER任务，AL则是通过增强同一领域数据质量以提高模型性能。但在专业领域，仅仅使用一种方法很难达到预期效果，如仅利用TL获取的外部知识不足以拟合模型，要实现良好效果仍需要一定量标注数据来进行微调。而AL基于不确定标准选择样本，这些样本包含丰富的领域信息，但AL策略忽略了大量的外部知识。一种可行的思路是将TL和AL结合起来形成统一框架。在通用领域，已经有一些学者尝试两种方法结合，但针对金融科技等特殊领域，相关的研究还较少[1]。为此，本文拟构建一种融合迁移学习的主动学习框架，并且为进一步提高模型的准确性，改进基于不确定性标准的采样策略，通过增加语义相似性权值削减采样中离群孤立点影响，以实现最佳的NER效果。

2 研究方法

2.1 研究目的

针对少样本NER任务，大多数工作考虑使用TL和AL方法，迁移学习利用领域相似性实现模型共建和数据共享。这种方式可以利用海量互联网文本信息在通用领域实现良好的效果，但在专业领域中，仅利用通用领域信息不足以训练模型，导致NER性能偏低。而主动学习通常基于“不确定性”标准，通过挖掘实体内蕴信息来增强同一领域数据的质量，这种方式从领域数据包含的信息量出发，优先选择信息量丰富的样本，但对于金融科技等特殊领域，部分实体属于未登录词、实体差异性较大，基于“不确定性”的主动学习采样策略对于实体特征的描述不完全，领域数据之间存在的离群孤立点未被充分考虑。针对迁移学习领域实体识别性能偏低、基于“不确定性”的主动学习采样策略特征描述不完全的问题，提出了一种可行的思路，就是融合迁移学习和主动学习，考虑更加全面的采样策略。鉴于此，借助预训练TL模型辅助语义表示，并通过主动学习采样策略增强领域数据。这种经过精心

挑选的样本包含丰富的信息量（informativeness），对模型性能的提升效果明显。本文将整个框架命名为基于语义相似度与不确定性的主动迁移学习方法Active Transfer Learning method based on Semantic Similarity and Uncertainty，ATL−SSU）。该框架从提高基础分类器性能和全面采样入手，将为少样本NER提供新的解决方案。

如图1所示，本文的研究主要包含分类器模块（Classifier）和选择器模块（Selector）。Classifier模块的作用是训练每一轮更新后的数据并提高分类性能，而Selector模块是为了挑选出最有价值的样本。在实践中，维护一个动态样本池迭代地参与训练：首先利用种子语料Initial Corpus训练出基础分类器Classifier；然后使用该分类器对未标记样本Unlabeled pool进行预测；最后通过选择器Selector挑选出置信度最高的一组新示例，并加入Labeled pool样本池中迭代训练。特别地，在Classifier中，引入了BERT预训练语言模型（pre−train model），以充分表征中文字词的语义，并利用Bi−LSTM−CRF进行序列约束以提高分类精度。此外，在选择器Selcetor中，充分考虑离群孤立点（outlier）现象，利用结合语义相似度（semantic similarty）和不确定性（uncertainty）度量的主动采样策略来削减离群孤立点的影响。

2.2　基于BERT-Bi-LSTM-CRF的分类器

在ATL−SSU中，基础分类器Classifier的构建极为关键。为此，借鉴TL微调机制，提出基于BERT−Bi−LSTM−CRF网络结构的中文NER识别方法。这种TL微调机制将预训练模型与循环神经网络模型融合，能够实现较高的准确率，有助于后续Selector进行更精准的样本挑选。该方法由BERT预训练模型、Bi−LSTM神经网络和CRF线性链组合而成，BERT−Bi−LSTM−CRF框架如图2所示。其中，BERT作为语义表示输入，Bi−LSTM抽取特征，CRF获取概率最大标签。与传统的NER模型相比，BERT−Bi−LSTM−CRF关键是BERT预训练语言模型的引入，BERT通过无监督建模的方式学习海量互联网语义信息，能够充分表征实体的语义信息，基于BERT进行TL微调可以有效地提高NER性能。

BERT是一种基于TL微调机制的多层双向

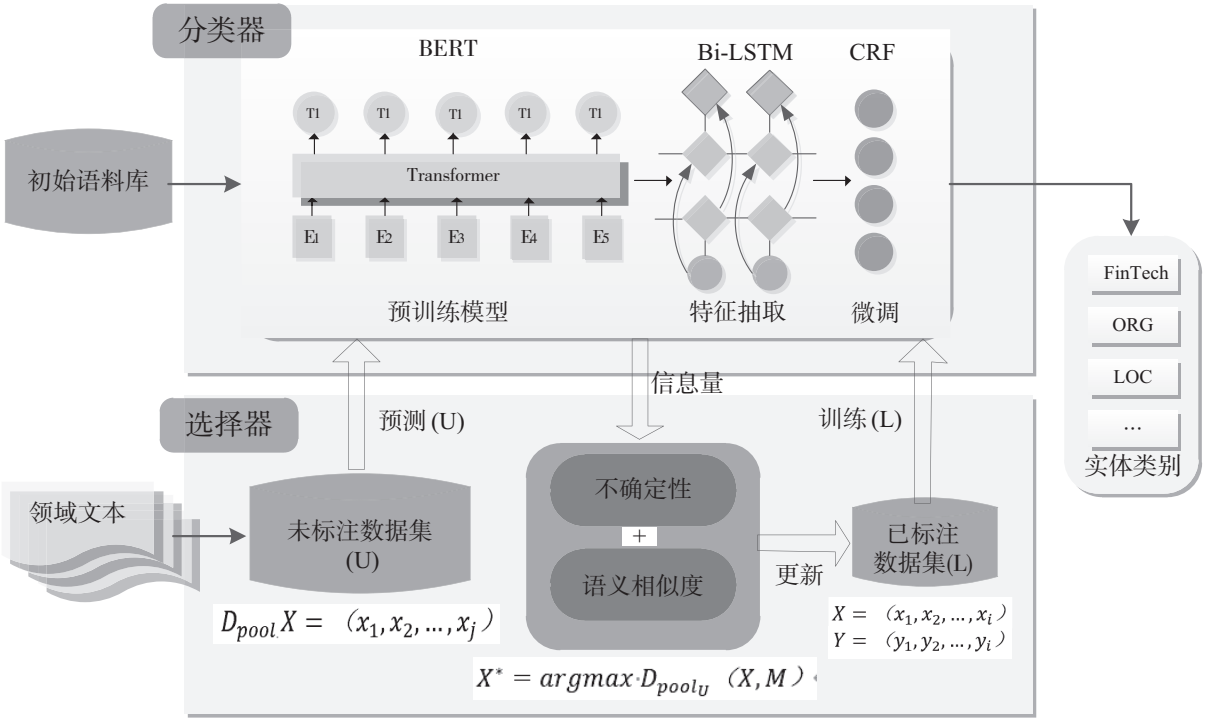


图1　基于深度迁移主动学习的NER框架

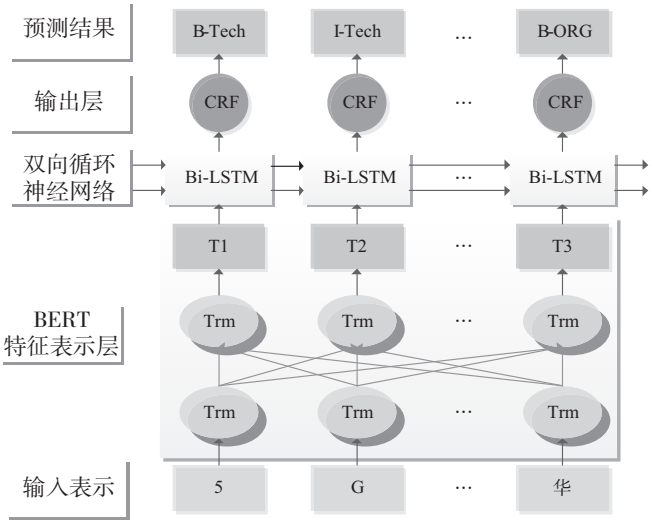


图 2　BERT-Bi-LSTM-CRF模型结构

Transformer编码器，它的特征表示依赖于左右上下文信息。Transformer则利用attention机制对文本建模，如式（1）所示，对于输入的字向量矩阵Query（Q）、Key（K）、Value（V）和向量维度d_k，通过*softmax*归一化获取每个向量的全局权重表示为

$$Attention\left(Q,K,V\right)=softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

之后再利用MultiHead位置嵌入来实现高速并行计算，Multi−Attention机制由几个按比例缩放的点积注意力组成，每个注意力从不同的维度和表示空间学习语义信息，计算方式如式（2）、式（3）所示，W_i^K，W_i^Q，W_i^V为权重矩阵。

$$Head_i=Attention\left(QW_i^Q,KW_i^K,VW_i^V\right) \qquad (2)$$

$$MultiHead\left(Q,K,V\right)=Concat\left(Head_1,\ldots,Head_n\right) \qquad (3)$$

为了训练Transformer，BERT采用Masked LM和Next Sentence Prediction方式。Masked LM的目的是根据上下文来预测masked字词的原始词语，而Next Sentence Prediction用来判断两个语句对是否连续，如表1所示。

与其他LM相比，BERT这种语境化的词嵌入在NLP中表现突出[4]，在中文NER任务中，谷歌的Chinese BERT−Base应用最为广泛。该模型由海量中文Wikipedia页训练而成，具有良好的语义表征能力。鉴于此，本文引入Chinese BERT−Base模型参与训练。

2.3　基于主动迁移学习的选择器

在Selecor模块中，基础采样是基于不确定性（Uncertainty）标准的采样策略，这种策略通过计算样本的信息熵（Information Entropy）来衡量样本的不确定性程度[3]。然而，在专业领域，数据差异性大，仅利用不确定性采样策略存在不完全特征描述问题，在采样中会挑选出大量离群孤立点（Outliers），而Outiliers会降低模型的分类性能。为此，本文从样本的代表性和不确定性的角度出发，考虑更加全面特征描述，提出结合语义相似度和不确定性度量的主动采样策略。

2.3.1　基于不确定性的主动学习

在主动学习中，基于不确定性标准的采样策略最为常见。其基本思想是挑选当前模型最不能确定的样本进行人工标注。如在二分类任务中基于不确定的策略倾向于选择后验概率接近0.5的样本，如果用SVM来训练模型的话，可

表 1　下一句预测

输入	标签
[CLS]近年来[SEP]金融科技[MASK]龙头招商银行已具[MASK]实力[SEP]	IsNext（是下一句）
[CLS]在数字[MASK]时代[SEP]科技[MASK]银行业务[SEP]	NotNext（不是下一句）

以挑选距离分类面最近的一些样本进行标注。而对于多分类序列标注任务，可以用信息熵来度量样本的不确定性程度。如对于给定的序列$X=(x_1,x_2,\ldots,x_i)$和标记序列$Y=(y_1,y_2,\ldots,y_i)$，x被预测为Y的不确定性，可以用式（4）、式（5）、式（6）来度量。

基于最低置信度原则（Least Confidence，LC）[12]：

$$\Phi^{LC}(x)=1-P(y^*\mid x) \quad (4)$$

其中，$P(y^*\mid x)$表示的是序列样本x对应的最可能标签序列，如在使用LSTM−CRF模型时，表示当前序列x属于标签y^*的概率。

最大归一化样本采样策略（Maximum Normalized Log−Probability，MNLP）[1]：

$$\Phi^{MNLP}(x)=\max_{y_1,\ldots,y_n}\frac{1}{n}\sum_{i=1}^{n}\log P\left(y_i\mid y_1,\ldots,y_{n-1},\{x_{ij}\}\right) \quad (5)$$

LC策略倾向于选择长句子，因此经过归一化的采样策略逐渐成为主流。在本文中，针对NER这种序列标注任务，利用概率计算信息熵（Information Entropy，IE）的最大归一化样本采样策略进行挑选，具体公式是：

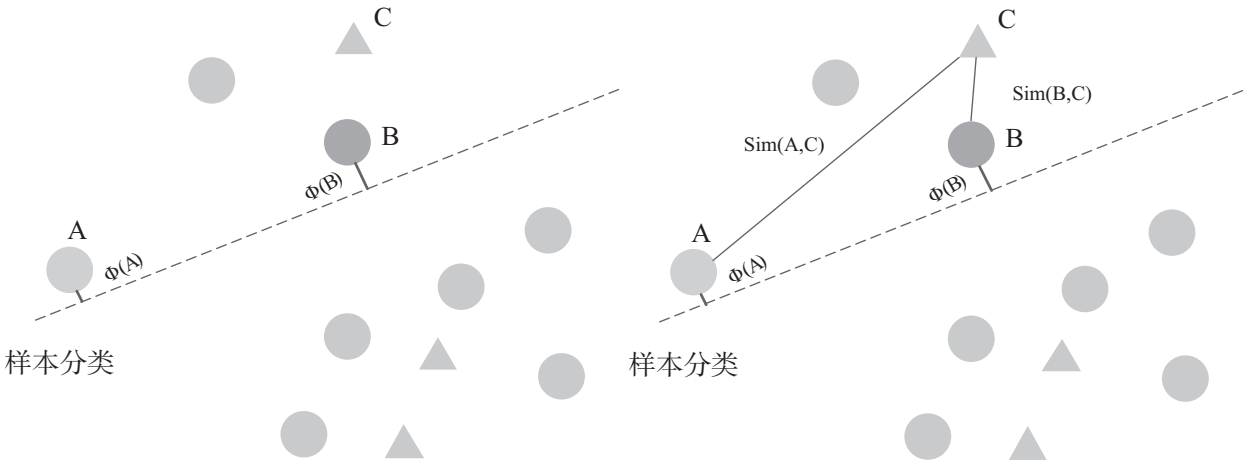
$$\Phi^{IE}(x)=-\frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M}P(y_i=m)\log P(y_i=m) \quad (6)$$

其中，$P(y)$为预测标签的条件分布概率，M为标签的个数，n为序列的长度。在NER等这种序列标注任务中，MNLP考虑平均信息熵，能够准确表示句子所含信息量。因此，在本文研究中，使用这种方法作为baseline方法。

2.3.2　结合语义相似度的不确定性

基于不确定性的采样策略可以优先挑选出最有标注价值的样本。但是针对特定领域的数据，仅仅考虑不确定性标准作为度量对特征描述是不完全的。因此，有必要考虑更加全面的采样策略。如在图3所示的样本分布中（图中圆形表示未标注样本，三角形为已标注样本，虚线表示初始的分类线）。如果根据不确定性标准采样，应该计算样本的信息熵大小，优先选择信息熵最大的样本。如图3中左图所示，当信息熵时，离分类面最近的样本A被优先选择。但在实际中，样本A在整个样本分布中属于离群样本，这种样本所包含的信息不具有代表性，为避免出现孤立点的消极影响，有必要考虑全局样本的信息量。

如图3中右图所示，在基于不确定性采样得到样本的信息熵后，添加语义相似度的权值来消减离群孤立点的负面影响，也即当$\Phi B\cdot Sim(B,C)>\Phi A\cdot Sim(A,C)$时，优先选择样本B。这是一种计算信息密度（Information Density，ID）的度量方式，它对于样本的描述更为全面，有助于Selector挑选出信息量更丰富的样本。对于给定的不确定性值Φ^{SE}，信息密度Φ^{ID}的计算如



（1）如果 Φ(A) > Φ(B)，则样本A被选中　　（2）如果 Φ(B)*Sim(B,C) >Φ(A)·Sim(A,C)，则样本B被选中

图3　基于信息密度的采样策略

式（7）：

$$\Phi^{ID}=\Phi^{SE}\cdot sim\left(X^{(U)},X^{(L)}\right) \quad (7)$$

其中，$sim(U,L)$为待标注样本$\bar{X}^{(U)}$与已标注样本$\bar{X}^{(L)}$之间的语义相似度。在本文研究中使用采用余弦值来度量语义相似度：对于序列$\bar{X}^{(U)}=(x_1,x_2,\ldots,x_i)$和序列$\bar{X}^{(L)}=(x_1,x_2,\ldots,x_j)$，其计算见式（8）：

$$sim\left(X^{(U)},X^{(L)}\right)=\frac{\vec{x}^{(U)}\cdot\vec{x}^{(L)}}{\vec{x}^{(U)}\cdot\vec{x}^{(L)}} \quad (8)$$

整体上结合语义相似度与不确定性度量的算法（ATL−SSU）计算步骤是：

输入：Labed samples(L)和Unlabeled samples (U)。

输出：更新后的ATL−SSU模型。

While的终止条件是：①训练基础模型 $ATL-SSU_i(L)$；②利用 $ATL-SSU_i$ 预测U；③计算U的不确定性ΦU；④计算U与L的相似度$\Phi U\cdot sim\left(X^{(U)},X^{(L)}\right)$；⑤挑选出Top K的无标注样本集K；⑥更新样本池，$L=L+K$，$U=U-K$，$i=i+1$。

3 实验与分析

3.1 数据集和参数设置

本文实验数据选用《人民日报》（Chinese Daily News）、金融科技文本（FinTech Corpus）两种不同的语料。其中，Chinese Daily News是公开的数据集，而FinTech Corpus是2019年12月从人民网科技板块以“金融科技”为关键词检索到的68篇文本经专家标注形成的实验语料。本文研究中的金融科技命名实体指的是与金融科技行业紧密相关的各种实体的统称。语料库包括：技术（FT_technology）、产品（FT_product）、金融行为（FT_behavior）、金融现象（FT_phenomenon）、金融事件（FT_event）、法律法规（FT_norm）等6类领域实体和人名（Person）、组织机构名（Organization）、地名（Location）、时间（Time）等4类通用实体。两种语料的实体统计信息如表2所示。

在实验之前，将标注好的数据随机划分为4个子数据集，分别为InitTrain、IterTrain、Valid、Tests。其中，InitTrain数据集共有500句，用于模型的初始训练；IterTrain数据集共有1 000句，可看作是未标注数据集，供模型每轮迭代使用；Valid数据集共有200句，用作交叉验证；Test数据集共有400句，用作测试集。在实验过程中，每次从IterTrain数据集中选出100句数据添加到InitTrain数据集中，之后进行迭代训练，迭代一共进行10轮。

3.2 基于BERT-Bi-LSTM-CRF分类器的有效性验证

本轮实验为验证结合预训练TL的有效性，也即对比BERT−Bi−LSTM−CRF与Bi−LSTM−CRF的性能。具体地，从Chinese Daily News和FinTech Corpus语料库中各选择1 500句训练集，500句作为测试集，对比在相同规模训练数据下，BERT−Bi−LSTM−CRF和Bi−LSTM−CRF模型在中文NER中的准确率（P）、召回率（R）和F1值。为避免随机性，在每种语料中实验3次取平均值。实验结果如表3所示。

由表3可知，无论是通用领域数据集还是专业领域数据集，使用BERT预训练语言模型的框架其准确率（P）、召回率（R）、F1值都比不使用预训练模型的效果有显著提升。在Chinese Daily News和FinTech Corpus数据集中Bi-LSTM−CRF模型的F1值比较低，分别为41.58%和39.32%，而使用BERT−Bi-LSTM−CRF模型，F1值分别为85.03%和62.97%。这是因为在BERT−Bi−LSTM−CRF模型中，引入了BERT这种经过大规模语料训练的预训练语言模型，BERT采用了双向Transformer结构，可表征的语义空间足够大，并且self−attention机制有效克服了长距离依赖问题，能够对上下文语义进行充分学习，因此在NER任务中相较于常规词向量或者单独的one−hot编码效果显著。这也表明采用大规模语料预训练的特征向量包含更加丰富的信息，模型的刻画能力更强，有助于识别效果的提升。

对于通用领域数据集Chinese Daily News而

表 2　语料中实体数目分布情况

单位：条

数据集	类型	初始数据集	迭代数据集	验证数据集	测试数据集
《人民日报》语料	PER（人名）	189	358	78	155
	LOC（地名）	386	745	160	308
	ORG（组织机构）	249	438	96	182
金融科技语料	FT_technology（技术）	356	526	115	153
	FT_product（产品）	121	331	62	105
	FT_behavior（金融行为）	217	414	81	115
	FT_phenomenon（金融现象）	139	293	53	108
	FT_event（金融事件）	12	126	8	17
	FT_norm（法律法规）	77	55	10	17
	PER（人名）	62	187	41	108
	ORG（组织机构）	250	567	108	191
	LOC（地名）	100	534	85	188
	Time（时间）	97	172	30	47

表 3　两种模型的对比实验结果

单位：%

方法	《人民日报》语料			金融科技语料		
	精准率（P）	召回率（R）	调和数（F1）	精准率（P）	召回率（R）	调和数（F1）
Bi−LSTM−CRF	40.34 ± 0.78	42.90 ± 0.25	41.58 ± 0.43	40.46 ± 1.77	38.41 ± 1.92	39.32 ± 0.32
BERT−Bi−LSTM−CRF	83.70 ± 1.03	86.41 ± 0.51	85.03 ± 0.43	59.69 ± 1.45	66.63 ± 1.23	62.97 ± 0.36

言，使用BERT对F1值提升43.45%，提升近一倍。而对于金融科技领域数据集FinTech Corpus，F1值提升23.65%，提升效果为60.3%。金融科技领域性较强，数据差异性较大，领域数据常常含有不规范用语，而BERT采用的是大规模通用语料训练而成，因此在金融科技领域基于BERT模型NER的性能提升不如在人民日报新闻通用语料。但从标准差的角度来看，无论是Bi-LSTM−CRF模型还是BERT−Bi-LSTM−CRF模型，在金融科技数据集中其标准差普遍高于通用领域数据集。这在一定程度上表明，在金融科技等专业领域，数据的差异性较大，存在着较多的奇异点数据。从整体上来看，使用BERT预训练语言模型对NER任务有显著提升，但是不容忽视的是在实验中选取的是一次挑取1 500句样本参与训练。因此，有必要在不损失模型精度的同时进一步减少标注量。鉴于此，将利用当前有效融合预训练迁移学习的主动学习NER框架进行实验。

3.3　结合语义相似度与不确定性的有效性验证

本实验为验证结合相似度的深度主动学习的有效性，即对比结合相似度的主动学习和单独的主动学习方法的差异。其中，模型均为BERT−Bi-LSTM−CRF，主动学习迭代次数为10次，初始训练集为InitTrain，共500句，每次迭代从IterTrain中随机挑选100句样本，加入到InitTrain中进行训练，用这种方式模拟人工标注的过程。用F值评价模型的标注效果，为消减误差，每轮实验进行3次取平均值作为结果。具体地，设置如下对比实验。

Baseline all方法（简称ALL）：采用完全标注数据集，即一次训练完成InitTrain和IterTrain中的所有数据。

Baseline random方法（简称Random）：采用随机主动学习方法，即每次从IterTrain中随机挑选100句样本，添加进InitTrain中。

Baseline active方法（简称Active−U）：采用基于不确定性的主动学习方法，即每次训练中通过基于不确定性的采样方法[3]，挑选出Top 100的数据进行迭代。

Active Transfer Learning method based on

Semantic Similarity and Uncertainty方法（简称ATL−SSU）：是本文提出的方法，采用基于不确定性和相似度结合的主动迁移学习方法。

分别在Chinese Daily News和FinTech Corpus语料上进行实验，如图4、图5所示。Chinese Daily News数据集中仅使用初始的500句训练数据训练模型的F1值为0.784 2，而全部的1 500条训练数据F1值为0.850 3。对于FinTech Corpus数据集，初始的500句训练数据的F1值为0.521 7，而10轮迭代后的1 500句的F1值为0.629 5。对于Active−U和ATL−SSU方法而言，10轮迭代中的F1值显著高于Random方法。如在第一轮迭代，同样的600条数据，Active−U和ATL−SSU方法F1值皆高于Random方法。这证明在相同样本数量情况下，经过主动学习挑选的数据具有更多的信息量，能让模型尽快收敛。

另外，随着主动学习迭代次数的增加F1值也在缓慢增加。其中，在Chinese Daily News数据集中，Active−U方法在第7次迭代时就能达到所有数据量训练的效果，而ATL−SSU方法（本文方法）达到最佳的模型训练效果时次数是6。对于FinTech Corpus数据集而言，Active−U方法在第8次达到最优的效果，ATL−SSU方法在第7次。如图6所示，在两数据集中，通用语料

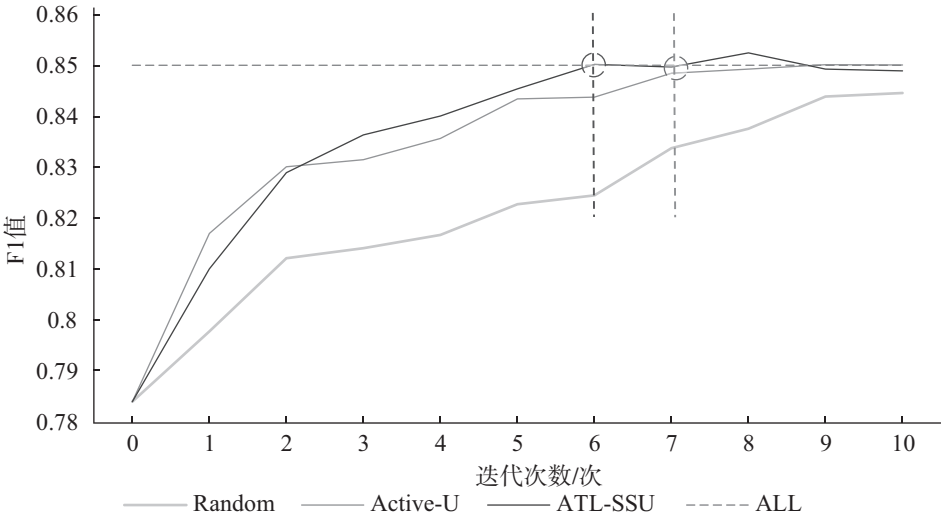


图4　在Chinese Daily News数据集上对比实验结果

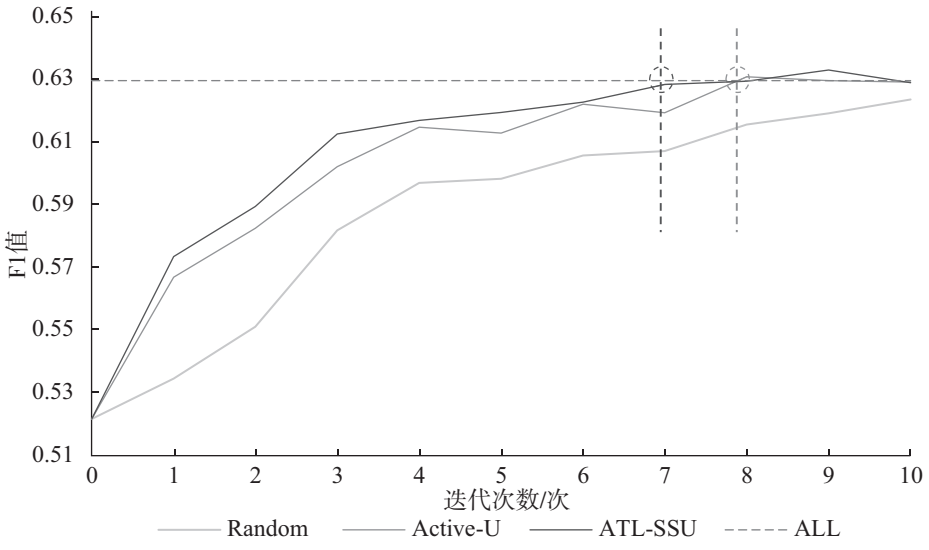


图5　在FinTech Corpus数据集上对比实验结果

Chinese Daily News的ATL−SSU方法仅用1 100条数据就能接近使用全部1 500条数据的效果，而在FinTech Corpus中要接近最优效果是1 200条。这表明在通用领域数据的差异性要小于专业领域，经过样本选择策后能更快挑选出代表性样本。如表4所示，Active−U和ATL−SSU方法都能减少一定的标注成本。其中在Chinese Daily News数据集中，Active−U方法能节省20.00%的标注成本，ATL−SSU则为26.67%；在FinTech Corpus数据集中，Active−U方法能节省13.33%的标注成本，ATL−SSU为20.00%。从整体来看，ATL−SSU方法节省的成本更多，相较于只要基于不确定性标准的Active−U方法，ATL−SSU方法可进一步减少6.67%的标注成本。此外，如图4、图5所示，结合语义相似度的ATL−SSU方法F1值曲线普遍高于Active−U方法，也更加平滑稳定。这表明ATL−SSU方法能有效规避离群样本点的影响，并且呈现出更为稳定的F1表现，证明了结合不确定性和相似度权值的主动学习方法的有效性。

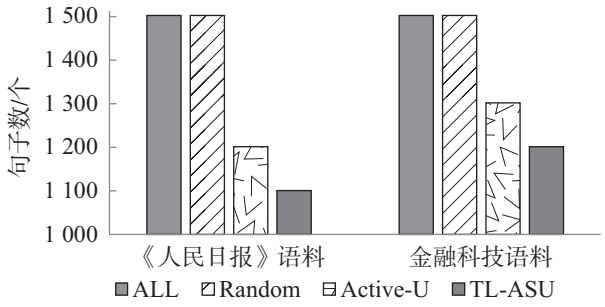


图6　模型达到最优时所需句子数

表4　Active和ATL-SSU方法节省的差额数据占整体比例

单位：%

方法	《人民日报》语料	金融科技语料
Active−U	20.00	13.33
ATL−SSU	26.67	20.00

4　结语

实体识别是文本挖掘中的一个阶段。本文研究面向特定领域标注数据少问题，利用融合迁移学习和主动学习的方法抽取出领域实体，减少了大规模语料中的人工成本，提高了实体识别效率，有助于颠覆性技术识别、热点事件发现、地平线扫描等工作的进行。本文提出的融合预训练迁移学习模型的中文领域主动学习NER框架，主要包含TL分类器和AL选择器两部分，重点解决领域NER中特征描述不完全导致的准确率低的问题。在金融科技领域，本文的方法F1值相较于LSTM−CRF提高23.65%，这表明富含外部语义信息的BERT表征能有效提升领域实体特征的广度，进而显著提高识别精度。此外，本文的方法还可以节省26.67%的标注成本，相较于基线方法提高6.67%，这表明语义相似度的添加能有效削减离群点的影响。本文提出的基于预训练语言模型的中文领域主动学习NER框架能够有效节省标注成本，同时能够提升F1值。因此，该模型是有效的。

中文领域NER较通用领域更加困难，本文研究中相同标注数据，模型在金融科技领域的数据集中的F1值为62.97%，而通用领域语料的F1值为85.03%，相差22.07%，这表明在面向特定领域时还需要考虑更多的领域特性，在后续的研究中还要充分考虑领域词边界、未登录词等问题，以期进一步提升模型的性能。

参考文献

[1] 姚明海, 黄展聪. 基于主动学习的半监督领域自适应方法研究[J]. 高技术通讯, 2020, 30(8): 783−789.

[2] 潘崇煜, 黄健, 郝建国, 等. 融合零样本学习和小样本学习的弱监督学习方法综述[J]. 系统工程与电子技术, 2020, 42(10): 2246−2256.

[3] SHEN Y, YUN H, LIPTON Z C, et al. Deep active learning for named entity recognition[C] // Proceedings of the 2nd Workshop on Representation Learning for NLP. Stroudsburg: Association for Computational Linguistics, 2017: 252−256.

[4] JACOB D, MING W C, KENTON L, et al. BERT: pre−training of deep bidirectional transformers for language understanding[C] // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Tech−

nologies. Stroudsburg: Association for Computational Linguistics, 2019: 4171−4186.

[5] HUANG Z, XU W, YU K. Bidirectional LSTM−CRF models for sequence tagging[J]. arXiv preprint arXiv, 2015: 1508.01991.

[6] MA Y, CAMBRIA E, GAO S. Label embedding for zero−shot fine−grained named entity typing[C] // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Stroudsburg: Association for Computational Linguistics, 2016: 171−180.

[7] GIORGI J M, BADER G D. Transfer learning for biomedical named entity recognition with neural networks [J]. Bioinformatics, 2018, 34(23): 4087−4094.

[8] 李舟军, 范宇, 吴贤杰. 面向自然语言处理的预训练技术研究综述[J]. 计算机科学, 2020, 47(3): 162−173.

[9] DANDAPAT S, WAY A. Improved named entity recognition using machine translation−based cross−lingual information[J]. Computacion Y sistemas, 2016, 20(3): 495−504.

[10] YANG Y, CHEN W, LI Z, et al. Distantly supervised NER with partial annotation learning and reinforcement learning[C] // Proceedings of the 27th International Conference on Computational Linguistics. Strouds−burg: Association for Computational Linguistics, 2018: 2159−2169.

[11] CHEN Y K, LASKO T A, MEI Q Z, et al. A study of active learning methods for named entity recognition in clinical[J]. Journal of biomedical informatics, 2015, 58: 11−18.

[12] BANERJEE P S, CHAKRABORTY B, TRIPATHI D, et al. A information retrieval based on question and answering and NER for unstructured information without using SQL[J]. Wireless personal communications, 2019, 108(3): 1909−1931.

（上接第26页）

horizon−europe/how−horizon−europe−was−developed_en.

[3] 苏文明, 王政书. 强化军民融合平台颠覆创新能力的对策建议: 以四川省为例[J]. 军民两用技术与产品, 2019(2): 15−19.

[4] 于川信, 刘志伟. 军民融合: DARPA的创新之路[M]. 北京: 国防工业出版社, 2018.

[5] 梁偲, 王雪莹, 常静. 欧盟“地平线2020”规划制定的借鉴和启示[J]. 科技管理研究, 2016, 36(3): 36−40.

[6] 张翼燕. 使命导向型的创新[EB/OL]. (2018−06−20)[2021−11−30]. https: //epaper.gmw.cn/gmrb/html/2018−06/20/nw.D110000gmrb_20180620_2−14.htm.

[7] 汪凌勇.发达国家科技计划管理机制研究[M].北京: 科学出版社, 2016.

[8] European Commission.LAB−FAB−APP Investing in the European future we want[EB/OL]. [2021−11−15]. https: //ec.europa.eu/info/events/shaping−our−future−2017−jul−03_en.

[9] 毛振芹, 程桂枝, 唐五湘. 部分科技发达国家科技计划项目的管理模式及启示[J]. 武汉工业学院学报, 2003, 22(3): 100−103.

[10] 慈元卓, 廖小刚. NASA战略技术投资组合管理研究系列之四NASA战略技术投资规划及其制定过程研 究[EB/OL]. [2021−11−12]. https://baijiahao.baidu.com/s?id=1599221841674413737&wfr=spider&for=pc.

[11] 汪江桦, 冷伏海, 王海燕. 美国科技规划管理特点及启示[J]. 科技进步与对策, 2013, 30(7): 106−110.

[12] Ministry of Education. Culture, Sports, Science and Technology National Institute of Science and Technology Policy.National Institute of Science and Technology Policy’s Key Activities[EB/OL]. [2021−11−09]. https: //www.nistep.go.jp/en/?p=4921.

[13] KUROGI Y, KOSHIBA H.ST Foresight 2019 – Comparative analysis of prediction by affiliation and age[EB/OL]. [2021−11−09]. https: //nistep.repo.nii.ac.jp/?action=pages_view_main&active_action= repository_view_main_item_detail&item_id=6734&item_no=1&page_id=13&block_id=21.

[14] 陈峰. 日本第八次科学技术预见项目的竞争情报学解析[J]. 竞争情报, 2007 (1): 2−5.

[15] 韩秋明, 袁立科, 王革. 韩国第五次技术预测实践及对我国的启示[J]. 全球科技经济瞭望, 2017, 32(8): 35−44.

[16] 陈炳硕.韩国科技计划评估模式分析[J].全球科技经济瞭望, 2017, 32(10): 39−44.