基于迁移学习的细粒度实体分类方法的研究

冯建周[1, 2] 马祥聪[1, 2]

摘 要 细粒度实体分类 (Fine-grained entity type classification, FETC) 旨在将文本中出现的实体映射到层次化的细分实体类别中. 近年来, 采用深度神经网络实现实体分类取得了很大进展. 但是, 训练一个具备精准识别度的神经网络模型需要足够数量的标注数据, 而细粒度实体分类的标注语料非常稀少, 如何在没有标注语料的领域进行实体分类成为难题. 针对缺少标注语料的实体分类任务, 本文提出了一种基于迁移学习的细粒度实体分类方法, 首先通过构建一个映射关系模型挖掘有标注语料的实体类别与无标注语料实体类别间的语义关系, 对无标注语料的每个实体类别, 构建其对应的有标注语料的类别映射集合. 然后, 构建双向长短期记忆 (Bidirectional long short term memory, BiLSTM) 模型, 将代表映射类别集的句子向量组合作为模型的输入用来训练无标注实体类别. 基于映射类别集中不同类别与对应的无标注类别的语义距离构建注意力机制, 从而实现实体分类器以识别未知实体分类. 实验证明, 我们的方法取得了较好的效果, 达到了在无任何标注语料前提下识别未知命名实体分类的目的.

关键词 细粒度实体分类, 迁移学习, 双向长短期记忆模型, 注意力机制

引用格式 冯建周, 马祥聪. 基于迁移学习的细粒度实体分类方法的研究. 自动化学报, 2020, **46**(8): 1759−1766

DOI 10.16383/j.aas.c190041

Fine-grained Entity Type Classification Based on Transfer Learning

FENG Jian-Zhou[1, 2] MA Xiang-Cong[1, 2]

Abstract The aim of fine-grained entity type classification (FETC) is that mapping the entity appearing in the text into hierarchical fine-grained entity type. In recent years, deep neural network is used to entity classification and has made great progress. However, training a neural network model with precise recognition requires a great quantity labeled data. The labeled dataset of fine-grained entity classification is so rare that hard to classify unlabeled entity. This paper proposes a fine-grained entity classification method based on transfer learning for the task of entity classification with lack labeled dataset. Firstly, we construct a mapping relation model to mining the semantic relationship between labeled entity type and unlabeled entity type, we construct a corresponding labeled entity type mapping set for each unlabeled entity type. Then, we construct a bidirectional long short term memory (BiLSTM) model, the sentence vector combination representing the mapping type set is used as the input of the model to train the unlabeled entity type. Lastly, the attention mechanism is constructed based on the semantic distance between different types in the mapping type set and corresponding unlabeled type, so as to realize entity classifier to recognize the classification of unknown entities. The experiment shows that our method have achieved good results and achieved the purpose of identifying unknown named entity classification with unlabeled dataset.

Key words Fine-grained entity type classification (FETC), transfer learning, bidirectional long short term memory model (BiLSTM), attention mechanism

Citation Feng Jian-Zhou, Ma Xiang-Cong. Fine-grained entity type classification based on transfer learning. *Acta Automatica Sinica*, 2020, **46**(8): 1759−1766

命名实体识别 (Named entity recognition, NER) 任务最初是在 MUC-6[1] 上被提出的, 它的目的是识别文本中涉及的实体, 并将它们分类到预定义的类型列表中. 它有两个关键的任务: 一是要识别出文本中是否有命名实体, 二是要判断出命名实体具体所指的目标类型. 作为自然语言的信息承载单位, 命名实体识别属于文本信息处理的基础研究领域, 是信息抽取、信息检索、机器翻译、问答系统等多种自然语言处理技术中必不可少的组成部分.

命名实体识别是自然语言处理 (Natural Language Processing, NLP) 领域中一些复杂任务的基础, 因此一直以来都是 NLP 领域中的研究热点. 现有的命名实体识别研究方法有基于规则的方法, 基于传统机器学习的方法 (又称为统计的方法), 以及近年来流行的基于深度学习的方法. 基于规则的方法[2] 由于手工构造规则, 系统能够达到较好的性能, 但构造规则时太依赖于专业领域知识, 费时费力且系统的可移植性较差. 基于机器学习的方法中, 命名实体识别被看作是序列标注问题, 传统的机器学习方法有许多适用于序列标注问题的模型, 比如隐马尔可夫模型 (Hidden Markov model, HMM)[3]、最大熵 (Maximum entropy, ME)[4]、条件随机场 (Conditional random field algorithm, CRF)[5] 等. 基于统计的方法对语料库的依赖比较大, 而可以用来建设和评估命名实体识别系统的大规模通用语料库又比较少, 这成为此方法的一大制约. 近年来, 随着深度学习算法的普及, 很多学者开始将深度学习算法应用在命名实体识别领域, 而且已经取得了卓越的效果. Athavale 等[6] 提出了 BiLSTM模型的方法, 通过 BiLSTM 网络将上下文结合起来, 进行 NER 的训练, 取得了良好的效果. 后来又出现了 BiLSTM+CRF[7]、CNN+BiLSTM+CRF[8]、CNN+Attention+Bi-LSTM+CRF[9] 等算法, 都取得了较好的效果. 但是深度学习算法和统计学方法一样都属于有监督的学习方法, 需要大量标注语料的支持.

除了有监督的学习方法, 半监督和无监督的学习方法也被用于 NER 任务. 半监督学习是最近兴起的一项技术, 主要技术为 BootStrapping[10]. 半监督的学习方法使用少量的标注语料作为初始种子集, 通过层层迭代的方法从大规模的未标注语料中逐步获取更多的种子实例, 扩充语料规模. 半监督学习在某些领域表现出了可与有监督学习竞争的效果, 但半监督学习依然需要部分标注语料当作种子语料. 无监督学习方法主要是采用设置语法和句法规则对未标注的数据识别实体. 无监督的学习方法不需要人工标注

收稿日期 2019-01-16 录用日期 2019-08-08
Manuscript received January 16, 2019; accepted August 8, 2019
国家自然科学基金 (61602401), 河北省高等学校科学技术研究青年基金 (QN2018074), 河北省自然科学基金 (F2019203157) 资助
Supported by National Natural Science Foundation of China (61602401), Youth Fund for Scientific Technological in Colleges and Universities of Hebei Province (QN2018074), and Nature Scientist Fund of Hebei Province (F2019203157)
本文责任编委 赵铁军
Recommended by Associate Editor ZHAO Tie-Jun
1. 燕山大学信息科学与工程学院 秦皇岛 066004 2. 燕山大学河北省软件工程重点实验室 秦皇岛 066004
1. School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004 2. Software Engineering Key Laboratory of Hebei Province, Yanshan University, Qinhuangdao 066004

的训练语料, 又称为开放式实体识别方法, 比较有代表性的成果如华盛顿大学的开源工具 Reverb、OLLIE[11], 斯坦福大学的 Stanford CoreNLP[12] 等.

随着 NLP 任务的深入, 扩展命名实体的种类渐渐被提上日程. 最近的研究工作表明, 使用更大范围的细粒度命名实体会给包括关系抽取、问答系统在内的自然语言处理任务带来实质性的提升. 因此, 研究细粒度实体分类问题已经逐渐成为业内的又一个研究热点.

1 细粒度实体分类研究现状

传统的命名实体识别只把实体分为人名、地名、组织机构和其他 4 种类型, 而细粒度实体分类会构建一种层次化的分类体系, 比如 "艺术家" 作为一个实体类别还可以分解为 "画家"、"音乐家"、"舞蹈家" 等子类别. 一个细粒度实体类别体系往往包括上百个甚至几百个实体类别, 从而给实体分类带来了更大的困难. 近年来, 细粒度实体分类 (Fine-grained entity type classification, FETC) 的研究逐渐活跃起来. Ling 等[13] 首次引入了一个训练和评估数据集 FIGER(GOLD), 他们使用线性分类器进行多标签分类. Gillick 等[14] 引入了依赖于上下文的 FETC 的概念, 其中, 实体的类型被限制为可以从其上下文可以推断的内容, 并且引入了新的手动注释的评估数据集 OntoNotes. Shimaoka 等[15] 提出了一种基于注意力机制的神经网络模型, 该模型使用长短期记忆网络 (Long short-term memory, LSTM) 来编码实体的上下文, 并使用注意力机制将模型的关注点集中于上下文中的相关表达. 由于 FETC 领域标注语料更加困难, 有研究者提出使用远程监督方法, 通过知识库中的实体链接到维基百科并获取语料, 但是这种方法因为标注时没有考虑句子上下文会引入大量噪声. 并且远程监督不是万能的, 对于许多基于社交媒体或与安全相关的应用, 我们无法访问高覆盖率的知识库, 所以这意味着这种方法有些时候是不合适的. 因此, 无法获取高质量的标注语料, 是一直困扰着 FETC 领域研究的一大难题.

为了解决标注语料缺乏的难题, 有学者开始采用迁移学习的方法开展细粒度实体分类的研究. 在机器学习、深度学习和数据挖掘的大多数任务中, 我们都会假设训练和预测时, 采用的数据服从相同的分布、来源于相同的特征空间. 但在现实应用中, 这个假设很难成立, 往往遇到的问题是带标记的训练样本数量有限, 比如, 处理 A 领域的分类问题时, 缺少足够的训练样本, 同时, 与 A 领域相关的 B 领域, 拥有大量的训练样本, 但 B 领域与 A 领域处于不同的特征空间或样本服从不同的分布. 这时, 知识迁移是一个不错的选择, 即把 B 领域中的知识迁移到 A 领域中来, 提高 A 领域分类效果.

迁移学习的具体做法分为基于资源的迁移学习和基于模型的迁移学习[16]. 基于资源的迁移学习需要借用额外的语料注释, 比如跨语言词典. 基于资源的方法在跨语言迁移方面取得了相当大的成功, 但对额外资源的规模和质量相当敏感. 基于模型的迁移不需要额外的资源, 它通过自适应地修改模型体系结构、训练算法或特征表示, 利用源任务和目标任务之间的相似性和相关性实现迁移学习的目的. 基于模型的迁移学习在实体识别领域已有应用, 它们大都是借助完整的标注语料集训练一个初始模型, 之后借鉴其参数, 用标注语料稀缺的数据集重新训练模型微调模型参数, 获得比直接用稀疏标注语料训练模型更好的效果. Lee 等[17] 提出了一个 TransNER 模型, 在两个拥有相同实体类别的数据集上进行迁移, 他们首先在有标注的源数据集上训练一个 NER 模型, 之后通过迁移模型不同层的参数在目标数据集上基于稀疏标注信息进一步训练模型的高层参数, 从而获得了比不迁移更好的效果. Yang 等[16]在跨领域、跨语言和跨任务三个设定上进行 NER 迁移, 提出了影响迁移学习模型效果的三个因素: 目标任务中标签的丰富度、源任务和目标任务的关系, 可共享的参数数量. Abhishek 等[18] 提出的 FNET 模型是首个将迁移学习用于细粒度实体分类的方法, 他们同样使用了借鉴模型参数的迁移方法获得了更好的效果, 表明迁移学习可以用于细粒度实体分类任务中, 但他们的源数据集和目标数据集都含有大量标注语料.

当前的迁移学习方法主要针对稀疏标注语料的数据集, 对于完全无标注语料的数据集并不适用. 然而, 当前很多领域根本就没有任何标注语料, 如何解决无标注语料的 FETC 问题, 是摆在研究者面前的一个难题. 本文提出了一种基于迁移学习的细粒度实体分类方法, 通过构建不同领域实体类别间的映射关联关系, 采用注意力机制, 实现对无标注语料的领域实体的类别分类.

2 基于实体类别关系映射与注意力机制的迁移学习模型

对于完全没有标注语料的领域数据集, 要想实现细粒度实体分类显然是一件困难的事情, 完全套用之前的迁移模型也是不合适的, 但是, 不同领域的实体类别间往往存在各种关联关系, 如果能够充分利用这种实体类别间的关系, 构建实体类别映射模型, 便可以实现知识的适度迁移, 本文从构建不同领域的实体类别关系入手, 构建迁移模型.

2.1 迁移模型的整体框架

针对无标注语料细粒度实体分类任务, 本文提出了一种基于实体类别关系映射与注意力机制的迁移模型 (Transfer learning model of entity type relationship mapping and attention mechanism, TLERMAM). 该模型由实体类别映射模型和实体类别分类器两部分组成. 实体类别映射模型通过计算无标注领域数据集 A' 中的实体类别与有标注领域数据集 A 中的实体类别间的词向量相似度得到二者语义相似程度. 根据语义相似度由大到小对与 A' 中实体类别 t'_m $(1 \le m \le M$, 假设 A' 中共有 M 个实体类别) 语义相关的所有 A 中的实体类别进行排序后, 取与 t'_m 最相似的 n 个实体类别 $t_1, t_2, \cdots, t_n$, 组合成映射类别集 $(1 \le n \le N$, 假设 A 中共有 N 个实体类别). 之后向 BiLSTM 分别输入实体类别 $t_1, t_2, \cdots, t_n$, 各自所在的句子向量 $\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_n$, 通过 BiLSTM 获取 n 个实体各自的上下文语义特征. 最后将 $t_1, t_2, \cdots, t_n$ 与 t'_m 的语义相似度作为权重, 应用实体级别注意力机制将 n 个语义特征向量进行结合得到 t'_m 的近似语义特征 $C_{t'_m}$, 再将 $C_{t'_m}$ 输入 Softmax 层得到其在各类标签上的概率, 最终训练模型收敛到在 t'_m 类别上的概率最高, 整个迁移模型的总体架构如图 1 所示.

2.2 实体类别映射模型

本文采用通用词向量来表示各种实体类别, 为了获得无标注领域实体类别的近似语义特征, 选取语义特征与其

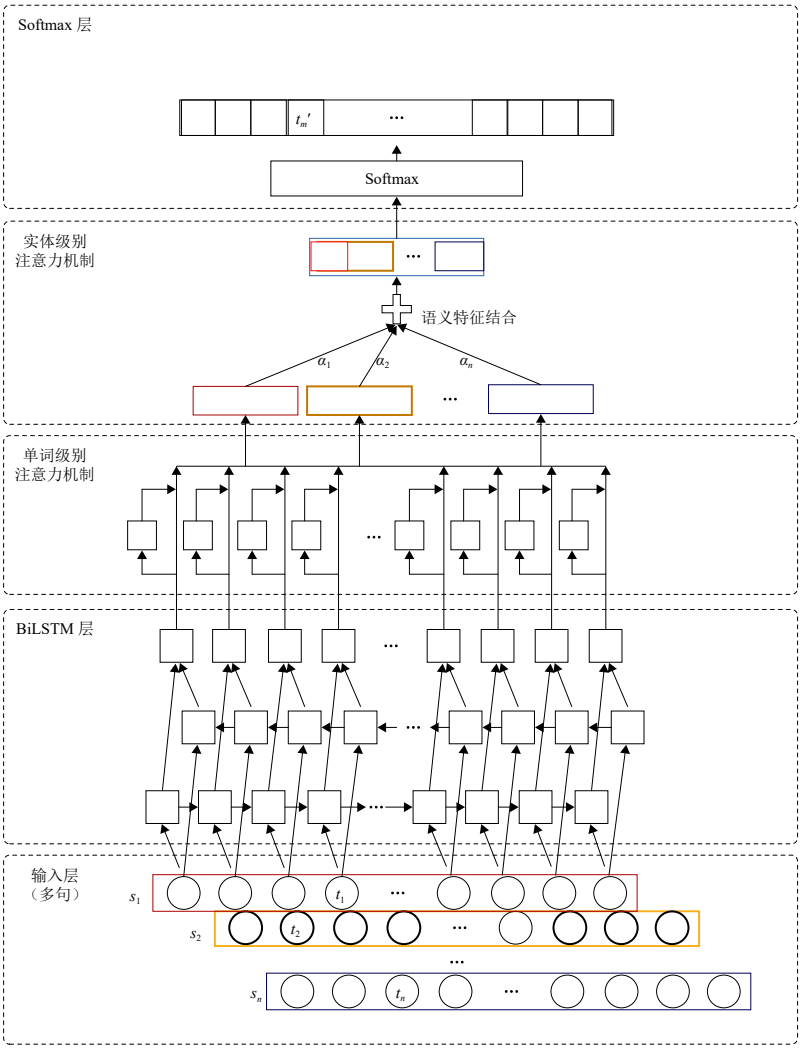

图 1　基于实体类别关系映射与注意力机制的迁移模型结构

Fig.1　The transferring model based on entity type relationship mapping and attention mechanism

最相似的 n 个有标注实体类别作为映射实体类别集, 如图 2 中的 $t_1, t_2, \cdots, t_n$. 图 2 中, e_i $(i = 1, 2, \cdots, n)$ 表示语义特征相似度, $t_m^{'}$ 为无标注数据集 A' 中的实体类别, t_i $(i = 1, 2, \cdots, n)$ 为有标注数据集 A 中的实体类别, 从左到右语义相似度逐渐减小. 目前有许多方法可以用来计算语义相似度, 例如余弦相似度 (Cosine similarity)、欧几里得距离 (Euclidean distance)、曼哈顿距离 (Manhattan distance), 我们采用余弦相似度计算语义特征相似度. 余弦相似度通过计算两个向量夹角的余弦值来评估他们的相似度. 给定两个词向量 $\boldsymbol{A}$ 和 $\boldsymbol{B}$, 其余弦相似度 θ 由点积和向量长度[19] 给出:

$$Semantics \quad Similarity = \cos\theta \tag{1}$$

其中, $\cos\theta = \dfrac{\sum_{i=1}^{n}(\mathbf{A}_i \times \mathbf{B}_i)}{(\sqrt{\sum_{i=1}^{n}(\mathbf{A}_i)^2}) \times (\sqrt{\sum_{i=1}^{n}(\mathbf{B}_i)^2})}$.

本文利用与实体类名相同单词的词向量代表该类用于计算语义相似度. 对于多词组成的实体类别, 如 body_part, 则采用组成该类名的所有单词的词向量平均值代表该类参与相似度计算. 根据余弦相似度度量无标注实体类别的词向量与每一个有标注实体类别的词向量相似度并排序, 将前 n 个实体类别加入到 $t_m^{'}$ 的实体类别映射集合 $D_{t_m'}$ 中.

$$D_{t_m'} = (t_1, t_2, \cdots, t_n) \tag{2}$$

假设无标注领域数据集 A' 中有 M 个实体类别, 对每一个无标注实体类别 $t_m^{'}$ $(m = 1, 2, \cdots, M)$, 我们都计算其与所有有标注实体类别的语义相似度, 最终得到所有无标注实体类别的映射实体类别集合.

2.3　迁移学习模型的数据输入

1) 词嵌入. 近年来许多研究证明词嵌入 (词向量) 能够很好地捕捉到单词的语义, 许多 NLP 任务通常都采用词向量分布式地表示自然文本中的单词. 本文采用预训练的词向量而不是随机初始化的词向量, 因为实验证明从大量的未标记数据中训练出的词向量比随机映射初始化的词向量效果更令人满意.

本文首先将上下文中的每个单词转化成能提供语义特征的实值向量. 给定一个句子 $s = \{x_1, x_2, \cdots, x_m\}$, 通过映射词向量矩阵 $E \in R^{|V| \times d_w}$, 每个单词 x_i 表示为 d_w 维实值向量, V 是词汇表的大小.

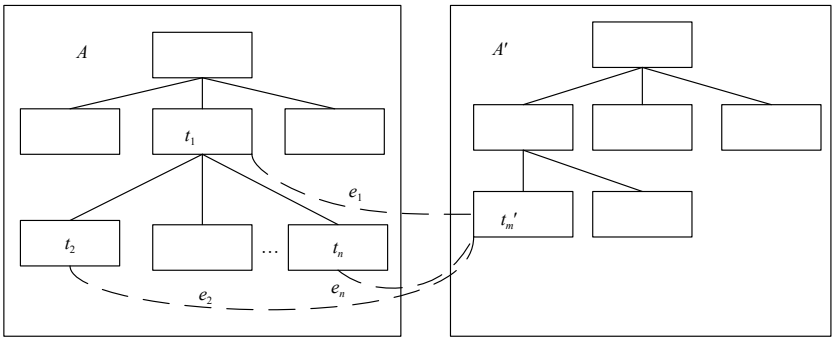

图 2 实体类别映射关系图

Fig. 2 Entity type mapping relation chart

2) 位置嵌入. 词嵌入虽然能很好地捕捉到单词的词义信息, 但它无法捕捉句子的结构信息. 为了增加更多的上下文语义, 本文引入位置信息向量, 位置信息体现句子中的每个单词 x_i 与关注实体的距离, 每个距离都被映射到一个随机初始化的 d_p 维的位置向量 $\boldsymbol{x}_i^p$.

最后, 我们将得到的两个向量结合, 于是句子中第 i 个单词的总嵌入向量为: $\boldsymbol{x}_i^E = \left[(\boldsymbol{x}_i^d)^{\mathrm{T}}, (\boldsymbol{x}_i^p)^{\mathrm{T}}\right]^{\mathrm{T}}$.

3) 迁移输入. 本文的迁移思想主要体现在用多个相似的有标注实体类别组合表征无标注实体类别, 从而可以借鉴多条有标注实体类别句子的组合向量来训练一个无标注实体类别. 因此在模型的输入端, 将无标注实体类别的映射类别集所在的句子向量作为向量集输入, 即所谓按组输入. 由式 (2) 可知, $t_1, t_2, \cdots, t_n$ 组成了无标注类别 $t_m^{'}$ 的映射类别集合, 假设在有标注数据集 A 中有 v_1 条句子标注为 t_1 类别, 形成集合 S_{t_1}, 有 v_2 条句子标注为 t_2 类别, 形成集合 S_{t_2}, 以此类推, 有 v_n 条句子标注为 t_n 类别形成集合 S_{t_n}, 则为了训练类别 $t_m^{'}$, 输入端从对应的 $S_{t_1}, S_{t_2}, \cdots, S_{t_n}$ 中分别任选一条句子组成输入句子向量集合 $\{\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_n\}$, 其中, $\boldsymbol{s}_1 \in S_{t_1}$, $\boldsymbol{s}_2 \in S_{t_2}, \cdots, \boldsymbol{s}_n \in S_{t_n}$.

2.4 注意力机制

2.4.1 单词级注意力机制

本文使用 BiLSTM 网络结合单词级别注意力机制的方法获得有标注实体的上下文表示. 长短期记忆网络 (LSTM) 的 "门" 设计可以学习长期依赖信息, 通过门来控制输入信息流入记忆细胞的比例, 以及从以前的状态中忘记信息的比例. BiLSTM 网络包含两个子网络, 一个称为前向 LSTM, 另一个称为后向 LSTM, 这是两个不同参数的网络. 对于给定的包含 q 个单词的句子 $s = (x_1, x_2, \cdots, x_q)$, 用两个 LSTM 结构分别计算句子中每个单词的左上下文表示 $\overrightarrow{h_i}$ 和右上下文表示 $\overleftarrow{h_i}$. h_i 为单词 x_i 的上下文信息 $(1 \le i \le q)$, 连接其左右上下文表示为: $h_i = \left[\overrightarrow{h_i}; \overleftarrow{h_i}\right]$. 将 BiLSTM 网络产生的向量集合 $\{\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_q\}$ 组合成矩阵 H. 根据文献 [20] 中所提的方法, 我们应用单词级别的注意力机制, 让模型选择信息更丰富的单词进行训练. 最终语义上下文表示 C 由这些向量的加权和组成:

$$G = \tanh(H) \tag{3}$$

$$\alpha = \mathrm{Softmax}(w^{\mathrm{T}} G) \tag{4}$$

$$C = H\alpha^{\mathrm{T}} \tag{5}$$

$H \in \mathbf{R}^{d_s \times q}$, 其中 d_s 是 LSTM 网络隐藏单元的个数, q 是句子单词个数, w 是一个预训练的参数向量. w、α、C 的维度分别是 d_s、q、d_s.

2.4.2 实体级注意力机制

根据之前获得的映射类别集合, 组合有标注实体类别的语义特征得到无标注实体类别的近似语义特征. 这里我们采用实体级别的注意力机制进行特征组合. 根据类别语义相似的实体其上下文语境也应该相似的特点, 我们采用第 2.2 节的方法选择 n 个有标注实体类别组合成无标注实体类别的映射类别集合, 每个有标注实体类别在集合中的权重根据相似度大小分配, 与无标注实体类别最相似的实体类别在最后形成的总特征中权重最大. 相似度权重 α 计算方法如下:

$$\alpha_i = \frac{\exp(e_i)}{\sum_{k=1}^{n} \exp(e_k)}, \qquad 1 \le i \le n \tag{6}$$

其中, e_i 是无标注实体类别 $t_m^{'}$ 的的映射类别集中的实体类别 t_i 与 $t_m^{'}$ 的语义相似度. α_i 为归一化后的语义相似度权重值. 获得每个有标注实体类别的权重后, 将有标注实体类别语义特征进行组合, 用来近似表示无标注实体类别的语义特征. 语义特征组合公式如下:

$$C_{t_m'} = \sum_{i=1}^{n} \alpha_i \times C_{t_i} \tag{7}$$

其中, $C_{t_m'}$ 是无标注实体类别 $t_m^{'}$ 的上下文语义特征, C_{t_i} 是 $t_m^{'}$ 的映射类别集合中的实体 t_i 基于上节的单词级注意力机制得到的上下文特征. 最后将基于实体级注意力机制得到的 $t_m^{'}$ 的近似上下文特征 $C_{t_m'}$ 输入 Softmax 分类器预测其实体类别 $\hat{y}$:

$$\hat{p}(y|C_{t_m'}) = \mathrm{softmax}(WC_{t_m'} + b) \tag{8}$$

$$\hat{y} = \mathrm{argmax}\, \hat{p}(y|C_{t_m'}) \tag{9}$$

实验中本文采用 L2 正则化方法防止过拟合, Softmax 的损失函数如下, 其中 λ 代表 L2 正则化参数, θ 代表模型的所有参数.

$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log_2(\hat{p}(y_i|C_{t_m'})) + \lambda \|\theta\|^2 \tag{10}$$

3　实验

为了证明本文提出方法的优越性, 本节设置了几组对比实验. 通过比较 TLERMAM 模型与其他模型在实现跨领域迁移细粒度实体分类任务时的性能, 以及 TLERMAM 模型在不同迁移规模下的效果来说明本文提出方法的有效性.

3.1　数据集

本文使用公开的 FIGER 数据集[13] 对模型进行评价. FIGER 是用于英文细粒度命名实体分类的数据集, 包含标准的训练集和测试集. FIGER 数据集包含 113 种细分的细粒度实体类别. FIGER 数据集的训练数据的获取使用了远程监督方法, 通过映射语料库 (维基百科) 的语料与 Freebase 等知识库中的相同实体自动生成. FIGER 数据集的测试数据主要是由 Ling 等手动标注的新闻报道句子组成.

我们使用 Shimaoka 等[15] 提供的处理后的 FIGER 数据集进行实验. 每条训练数据包含实体 mention 的起始位置、终止位置、句子、mention 和实体 label. 由于远程监督方法在为实体打标签时未考虑实体的上下文, 导致一个实体可能有多个标签, 这种噪声数据会影响模型的性能, 因此我们在原始数据集的基础上过滤掉了含有多个标签的句子, 实体类数同时减少.

3.2　评价指标

本文采用准确率 (Accuracy, Acc), 宏平均 F1 值 (Macro-averaging F1-Measure, Macro F1) 和微平均 F1 值 (Micro-averaging F1-Measure, Micro F1) 评价模型性能. 现存的细粒度实体分类模型中广泛使用以上三种评价方法.

根据样例的真实类别与模型预测类别的组合, 可以分为真正例、假正例、真反例和假反例, 如表 1 所示.

表 1　混淆矩阵

Table 1　Confusion matrix

		预测情况	
		正例	反例
真实情况	正例	TP (真正例)	FN (假反例)
	反例	FP (假正例)	TN (真反例)

精确率 (Precision, P)、召回率 (Recall, R)、$F1$ 值和准确率 (Acc) 的计算公式如下:

$$P=\frac{TP}{TP+FP}, R=\frac{TP}{TP+FN}, F1=2\times\frac{P\times R}{P+R} \tag{11}$$

$$Acc=\frac{TP+TN}{TP+TN+FP+FN} \tag{12}$$

宏平均 $F1$ 值, 先计算每一个实体类的统计指标值, 然后再对所有类求算术平均值, 计算公式如下:

$$F1_{Macro}=\frac{1}{n}\sum_{i=1}^{n}F1_i \tag{13}$$

其中, n 为实体类别的数量, $F1_i$ 表示第 i 类实体的 $F1$ 值.

微平均 $F1$ 值, 对测试数据中的每一个进行不分类统计, 然后计算相应指标. 计算公式如下:

$$P_{Micro}=\frac{\sum_{i=1}^{n}TP_i}{\left(\sum_{i=1}^{n}TP_i+\sum_{i=1}^{n}FP_i\right)} \tag{14}$$

$$R_{Micro}=\frac{\sum_{i=1}^{n}TP_i}{\left(\sum_{i=1}^{n}TP_i+\sum_{i=1}^{n}FN_i\right)} \tag{15}$$

$$F1_{Micro}=2\times\left(\frac{P_{Micro}\times R_{Micro}}{P_{Micro}+R_{Micro}}\right) \tag{16}$$

3.3　实验设置

本文实验采用 Tensorflow 框架, 并使用 NVIDIA 的 1080 显卡进行了加速. 其他实验设置情况如下:

1) 预训练词向量. 与随机初始化的词向量相比, 使用预训练的词向量可以取得更好的效果. 本文采用了 300 维的实值向量表示单词的词向量. Glove 词向量模型[21] 是一种广泛使用的词向量模型, 其基于词共现结构以无监督的方式学习单词的向量表示, 具有比较优秀的准确性, 因此本次实验我们使用了 Glove 词向量模型.

2) 参数设置. 模型的超参数包括 Adam Optimizer 的学习速率 L_r, 位置嵌入维度 D_p, LSTM 层 D_s 的状态大小, 批大小 B, LSTM 层的输入丢失保持概率 P_i 和输出丢失保持概率 P_o, L2 正则化参数 λ. 通过评估每个数据集的开发集上的模型性能获得的这些超参数的值可以在表 2 中找到.

表 2　超参数设置

Table 2　Hyper-parametric settings table

L_r	D_w	D_p	B	P_i	P_o	λ
0.0002	180	85	256	0.7	0.9	0.0

3.4　相似领域的算法对比

我们将 FIGER 集中每一个实体大类中的多个子类别随机分成了两组, 得到了两个完全不同的实体类别集合 A 和 A'. 接着我们将 FIGER 数据集的句子语料根据划分的两个实体类别集进行分割, 由于已经过滤掉包含标签的噪声数据, 所以此时不会出现无法归类的句子. A 作为源领域, 其所对应的句子集合作为有标注数据集, 隐去 A' 对应的句子集的标注信息, 将其作为无标注的目标领域, 数据集规模见表 3.

表 3　数据集规模表

Table 3　Datasets size table

	有标注数据集 (源领域)	无标注数据集 (目标领域)
类别数量	50	30
mention 数量	896 914	229 685
Token 数量	15 284 525	3 929 738

本节将本文提出的 TLERMAM 模型与其他先进的命名实体识别模型进行了对比. TransNER 模型是 Lee 等[17]提出的一种用于迁移命名实体识别的模型. FNET 模型是 Abhishek 等[18] 等提出的一个可用于细粒度实体分类的迁移 NER 模型. 当目标领域的句子集合完全无标注时, 我们用实体类别集 A 对应的标注句子集训练 TransNER 和 FNET 模型. 在训练 TLERMAM 模型时, 我们将映射实体的句子语料、各映射实体类别与目标实体类别的相似度信息以及目标实体类别标签一同输入模型, 采用第 2 节的方法训练模型, 然后在测试集上对三种算法进行对比, 对比结果如表 4 所示.

表 4 无标注领域不同模型对比实验

Table 4 Comparative experiment of different models in unlabeled field

模型	Acc	Macro F1	Micro F1
TransNER	0.051	0.035	0.041
FNET	0.026	0.027	0.028
TLERMAM	0.369	0.290	0.355

我们从 A' 中抽取少量句子增加标注 (大约 10%) 作为稀疏标注语料, 借助稀疏标注语料, 在之前已有的模型基础上, 继续训练 TransNER、FNET 和 TLERMAM 模型, 并在测试集上测试效果. 实验对比结果如表 5.

表 5 稀疏标注领域不同模型对比实验

Table 5 Comparison experiment of different models in the field of sparse annotation

模型	Acc	Macro F1	Micro F1
TransNER	0.500	0.337	0.534
FNET	0.523	0.329	0.447
TLERMAM	0.805	0.487	0.805

从表 4 和表 5 中可以看出, 无论是在无标注语料的情况还是有稀疏标注语料的情况, TLERMAM 模型表现出的效果比 TransNER 和 FNET 都要好. 在目标领域完全无标注语料时, 我们的模型从跨领域的映射实体的上下文中提取到了无标注实体的近似语义特征, 将其用于训练目标领域的实体分类, 从而取得了比其他算法更好的效果. 在目标领域具有稀疏标注语料时, 我们采用迁移算法构建的模型比其他两种算法 (先基于其他领域的语料训练一个模型, 再基于稀疏语料调参) 同样具有更好的效果. 另外, 从实验结果可以看出 TLERMAM 模型宏平均值均明显低于微平均值, 这是因为 FIGER 数据集各类实体语料数量并不平均. 部分目标实体映射实体语料不足, 导致其特征表示不能有效表达目标实体的语义, 模型对其的识别效果较差, 拉低了整体的平均值.

为了验证两个领域规模的差异大小对模型性能的影响, 本节做了以下对比实验. 迁移规模指源领域与目标领域之间实体数量的比例, 我们设置了几种不同的迁移规模, 逐渐减少目标无标注实体集合的大小, 增大有标注领域实体集合的大小.

我们还比较了不同映射类别集合大小 (window) 对模型性能的影响. 同时, 为了与固定的 window 对比, 我们采用了设置阈值选取映射实体的方法, 只选用与目标实体类别的语义相似度大于某一阈值的映射实体类别. 实验结果如图 3.

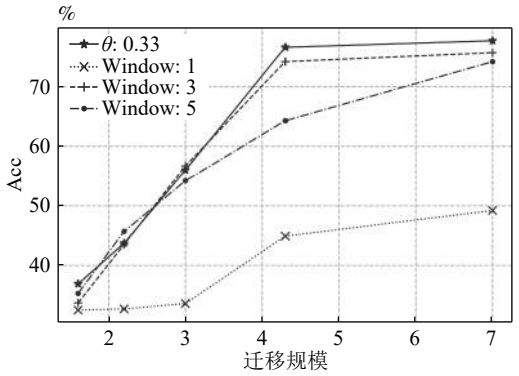

图 3 迁移规模与实体类别映射集规模对比图

Fig. 3 Transfer scale and entity type mapping set scale contrast chart

从图 3 中得出的实验效果可以看出, 随着有标注领域数据集比例的增加, 模型效果也在提高, 因为源领域的有标注实体类别越多, 就可以为目标领域提供更多的相似实体类别, 为目标领域提供更接近的语义特征, 使无标注实体类别的近似语义特征与真实语义接近.

另外, 当 window 为 3 (即目标实体类的映射类别集的类别数为 3) 时比 window 为 1 和 5 时, 实验效果都要更好, 说明单纯的扩大映射类别集的元素数量并没有使模型性能提升, 反而使效果下降, 原因是新增加的映射实体类别与目标实体类别相似度较低, 增加较低相似度的映射实体类别会给目标实体类别带来噪声, 使模型训练时偏移目标实体类别的真实语义, 导致预测准确度下降. 在当前数据集情况下, 经过反复实验证明, 在 window 为 3 时, 效果最好.

最后, 设置语义相似度阈值 θ 后模型效果较固定窗口数量总体上更好, 这表明只选取与目标实体类别更相似的有标注实体类别组合得到的目标实体类别的语义特征更接近真实, 这虽然会导致每个无标注实体类别的映射实体类别集合大小不固定, 但映射集合中的实体类别更为可靠, 它们的语义特征与目标实体类别更接近, 有利于获得与目标实体类别更相近的语义特征. 在当前数据集情况下, 经过反复实验证明, 在阈值 $\theta = 0.33$ 时, 效果最好.

3.5 无关领域的算法对比

为了验证本模型的通用性, 我们构造了两个看似无关的领域集合, 从 FIGER 数据集中抽取了一个军事领域集合作为无标注目标领域, 又抽取了一个文化领域数据集作为源领域. 其中军事领域共有 9 种实体类别, 文化领域共有 25 种实体类别. 各领域实体类别见表 6, 数据集规模见表 7.

以军事领域作为迁移的目标领域, 文化领域作为源领域. 对军事领域的每类实体, 使用文化领域的实体对其做映射, 构建映射类别集, 映射 window 选取为 3. 使用 TLERMAML 模型与其他先进的命名实体识别模型 (TransNER 模型、FNET 模型) 进行对比. 对比实验见表 8 和表 9.

从表 8 和表 9 中可以看出, 在两种对比实验中 TLERMAM 模型均取得更好的效果. 由于军事领域和文化领域实体类别的相似度较低, 模型获取得的近似特征有限, 因

表 6 军事领域和文化领域的实体类别集
Table 6 Entity type set of military field and culture field

领域	实体类别
军事	terrorist_organization, weapon, attack, soldier, military, terrorist_attack, power_station, terrorist, military_conflict
文化	film, theater, artist, play, ethnicity, author, written_work, language, director, music, musician, newspaper, election, protest, broadcast_network, broadacast_program, tv_channel, religion, educational_institution, library, educational_department, educational_degree, actor, news_agency, instrument

表 7 军事领域和文化领域的数据集规模表
Table 7 Dataset size of military field and culture field

	有标注数据集 (文化领域)	无标注数据集 (军事领域)
类别数量	25	9
mention 数量	226 734	126 036
Token 数量	3 927 700	2 104 890

表 8 无标注语料的军事领域实体识别效果比较
Table 8 Comparison of entity recognition in unlabeled military field

模型	Acc	Macro F1	Micro F1
TransNER	0.040	0.023	0.012
FNET	0.013	0.014	0.029
TLERMAM	0.257	0.339	0.339

表 9 稀疏标注语料的军事领域识别对比
Table 9 Comparison of entity recognition in military field with sparse annotated corpus

模型	Acc	Macro F1	Micro F1
TransNER	0.338	0.204	0.285
FNET	0.460	0.424	0.537
TLERMAM	0.572	0.504	0.559

此实体类别的识别效果低于第 3.4 节相似领域的模型效果, 但是仍然较其他算法有很大优势, 从而也证明了本文算法的适用范围不仅局限在相似领域.

4 结论

本文提出了一种基于迁移学习的跨领域细粒度实体分类的方法, 通过构建有标注领域的实体类别与无标注领域的实体类别间的语义映射关系, 借助源领域的标注信息构造目标领域的映射语义特征, 并结合注意力机制实现迁移学习. 通过对比实验证明, 当源领域与目标领域越接近且源领域语料更丰富时, 在目标领域的实体分类效果越好. 同时, 实验证明我们的方法不仅可以在完全没有标注语料领域进行命名实体识别任务, 在有稀疏标注语料的领域同样可以更好地完成实体分类任务.

References

1 MUC-6. The sixth in a Series of Message Understanding Conferences [Online], available: https://cs.nyu.edu/cs/faculty/grishman/muc6.html, 1995.

2 Grishman R. The NYU system for MUC-6 or where's the syntax? In: Proceedings of the 6th conference on Message understanding. Maryland, USA: ACL, 1995. 167−175

3 Zhou G D, Su J. Named entity recognition using an hmmbased chunk tagger. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, USA: ACL, 2002. 473−480

4 Borthwick A, Grishman R. A maximum entropy approach to named entity recognition [Ph. D. dissertation], New York University, 1999.

5 Lafferty J D, Mccallum A, Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th Internaional Conference on Machine Learning. Williamstown, MA, USA: ICML, 2001. 282−289

6 Athavale V, Bharadwaj S, Pamecha M, et al. Towards deep learning in hindi NER: an approach to tackle the labelled data scarcity. arXiv preprint arXiv: 1610.09756, 2016.

7 Huang Z, Xu W, Yu K. Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv: 1508.01991, 2015.

8 Ma X, Hovy E. End-to-end sequence labeling via bidirectional lstm-cnns-crf. arXiv preprint arXiv: 1603.01354, 2016.

9 Bharadwaj A, Mortensen D, Dyer C, et al. Phonologically aware neural model for named entity recognition in low resource transfer settings. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Texas, USA: EMNLP, 2016. 1462−1472

10 Putthividhya D P, Hu J. Bootstrapped named entity recognition for product attribute extraction. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, UK: EMNLP, 2011. 1557−1567

11 Schmitz M, Bart R, Soderland S, et al. Open language learning for information extraction. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, Korea: ACL, 2012. 523−534

12 Manning C, Surdeanu M, Bauer J, et al. The Stanford CoreNLP natural language processing toolkit. In: Proceedings of the 52nd annual meeting of the association for computational linguistics. Maryland, USA: ACL, 2014. 55−60

13 Ling X, Weld D S. Fine-Grained entity recognition. In: Proceedings of the 26th AAAI Conference on Artificial Intelligence. Toronto, Canada: AAAI, 2012. 94−100

14 Gillick D, Lazic N, Ganchev K, et al. Context-dependent fine-grained entity type tagging. arXiv preprint arXiv: 1412.1820,

2014.

15 Shimaoka S, Stenetorp P, Inui K, et al. An attentive neural architecture for fine-grained entity type classification. In: Proceedings of the 5th Workshop on Automated Knowledge Base Construction. San Diego, USA: AKBC, 2016. 69−74

16 Yang Z, Salakhutdinov R, Cohen W W. Transfer learning for sequence tagging with hierarchical recurrent networks. arXiv preprint arXiv: 1703.06345, 2017.

17 Lee J Y, Dernoncourt F, Szolovits P. Transfer learning for named-entity recognition with neural networks. arXiv preprint arXiv: 1705.06273, 2017.

18 Abhishek A, Anand A, Awekar A. Fine-grained entity type classification by jointly learning representations and label embeddings. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia, Spain: ACL, 2017. 797−807

19 Cosine_similarity. Cosine_similarity [Online]. available: https://en.wikipedia.org/wiki/Cosine_similarity, 2018.

20 Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany: ACL, 2016. 207−212

21 Pennington J, Socher R, Manning C. Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing. Doha, Qatar: EMNLP, 2014. 1532−1543

冯建周 燕山大学信息科学与工程学院副教授. 主要研究方向为知识图谱, 语义 web. 本文通信作者.
E-mail: fjzwxh@ysu.edu.cn
(**FENG Jian-Zhou** Associate professor at the School of Information Science and Engineering, Yanshan University. His research interest covers knowledge graph and semantic web. Corresponding author of this paper.)

马祥聪 燕山大学信息科学与工程学院硕士研究生. 主要研究方向为知识图谱.
E-mail: maxiangcong@126.com
(**MA Xiang-Cong** Master student at the School of Information Science and Engineering, Yanshan University. His main research interest is knowledge graph.)