软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2022,33(10):3776–3792 [doi: 10.13328/j.cnki.jos.006305]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

动态迁移实体块信息的跨领域中文实体识别模型*

吴炳潮[1,3], 邓成龙[1,3], 关 贝[1], 陈晓霖[1,3], 咎道广[1,3], 常志军[4], 肖尊严[5], 曲大成[5], 王永吉[1,2,3]

[1](中国科学院 软件研究所 协同创新中心, 北京 100190)

[2](计算机科学国家重点实验室(中国科学院 软件研究所), 北京 100190)

[3](中国科学院大学, 北京 100049)

[4](中国科学院 文献情报中心, 北京 100190)

[5](北京理工大学 计算机学院, 北京 100081)

通信作者: 王永吉, E-mail: ywang@itechs.iscas.ac.cn

摘 要: 由于中文文本之间没有分隔符, 难以识别中文命名实体的边界. 此外, 在垂直领域中难以获取充足的标记完整的语料, 例如医疗领域和金融领域等垂直领域. 为解决上述不足, 提出一种动态迁移实体块信息的跨领域中文实体识别模型(TES-NER), 将跨领域共享的实体块信息(entity span)通过基于门机制(gate mechanism)的动态融合层, 从语料充足的通用领域(源领域)动态迁移到垂直领域(目标领域)上的中文命名实体模型, 其中, 实体块信息用于表示中文命名实体的范围. TES-NER 模型首先通过双向长短期记忆神经网络(BiLSTM)和全连接网络(FCN)构建跨领域共享实体块识别模块, 用于识别跨领域共享的实体块信息以确定中文命名实体的边界; 然后, 通过独立的基于字的双向长短期记忆神经网络和条件随机场(BiLSTM-CRF)构建中文命名实体识别模块, 用于识别领域指定的中文命名实体; 最后构建动态融合层, 将实体块识别模块抽取得到的跨领域共享实体块信息通过门机制动态决定迁移到领域指定的命名实体识别模型上的量. 设置通用领域(源领域)数据集为标记语料充足的新闻领域数据集(MSRA), 垂直领域(目标领域)数据集为混合领域(OntoNotes 5.0)、金融领域(Resume)和医学领域(CCKS 2017)这 3 个数据集, 其中, 混合领域数据集(OntoNotes 5.0)是融合了 6 个不同垂直领域的数据集. 实验结果表明, 提出的模型在 OntoNotes 5.0、Resume 和 CCKS 2017 这 3 个垂直领域数据集上的 *F*1 值相比于双向长短期记忆和条件随机场模型(BiLSTM-CRF)分别高出 2.18%、1.68%和 0.99%.

关键词: 命名实体识别; 迁移学习; 跨领域; 动态融合; 双向长短期记忆神经网络

中图法分类号: TP391

中文引用格式: 吴炳潮, 邓成龙, 关贝, 陈晓霖, 咎道广, 常志军, 肖尊严, 曲大成, 王永吉. 动态迁移实体块信息的跨领域中文实体识别模型. 软件学报, 2022, 33(10): 3776–3792. http://www.jos.org.cn/1000-9825/6305.htm

英文引用格式: Wu BC, Deng CL, Guan B, Chen XL, Zan DG, Chang ZJ, Xiao ZY, Qu DC, Wang YJ. Dynamically Transfer Entity Span Information for Cross-domain Chinese Named Entity Recognition. Ruan Jian Xue Bao/Journal of Software, 2022, 33(10): 3776–3792 (in Chinese). http://www.jos.org.cn/1000-9825/6305.htm

Dynamically Transfer Entity Span Information for Cross-domain Chinese Named Entity Recognition

WU Bing-Chao[1,3], DENG Cheng-Long[1,3], GUAN Bei[1], CHEN Xiao-Lin[1,3], ZAN Dao-Guang[1,3], CHANG Zhi-Jun[4], XIAO Zun-Yan[5], QU Da-Cheng[5], WANG Yong-Ji[1,2,3]

[1](Collaborative Innovation Center, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)

[2](State Key Laboratory of Computer Science (Institute of Software, Chinese Academy of Sciences), Beijing 100190, China)

[3](University of Chinese Academy of Sciences, Beijing 100049, China)

* 基金项目: 国家重点研发计划(2017YFB1002303)

收稿时间: 2020-10-16; 修改时间: 2020-12-15; 采用时间: 2021-01-20; jos 在线出版时间: 2021-02-07

[4](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[5](School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

Abstract: Boundaries identification of Chinese named entities is a difficult problem because of no separator between Chinese texts. Furthermore, the lack of well-marked NER data makes Chinese named entity recognition (NER) tasks more challenging in vertical domains, such as clinical domain and financial domain. To address aforementioned issues, this study proposes a novel cross-domain Chinese NER model by dynamically transferring entity span information (TES-NER). The cross-domain shared entity span information is transferred from the general domain (source domain) with sufficient corpus to the Chinese NER model on the vertical domain (target domain) through a dynamic fusion layer based on the gate mechanism, where the entity span information is used to represent the scope of the Chinese named entities. Specifically, TES-NER first introduces a cross-domain shared entity span recognition module based on a bidirectional long short-term memory (BiLSTM) layer and a fully connected neural network (FCN) which are used to identify the cross-domain shared entity span information to determine the boundaries of the Chinese named entities. Then, a Chinese NER module is constructed to identify the domain-specific Chinese named entities by applying independent BiLSTM with conditional random field models (BiLSTM-CRF). Finally, a dynamic fusion layer is designed to dynamically determine the amount of the cross-domain shared entity span information extracted from the entity span recognition module, which is used to transfer the knowledge to the domain-specific NER model through the gate mechanism. This study sets the general domain (source domain) dataset as the news domain dataset (MSRA) with sufficient labeled corpus, while the vertical domain (target domain) datasets are composed of three datasets: Mixed domain (OntoNotes 5.0), financial domain (Resume), and medical domain (CCKS 2017). Among them, the mixed domain dataset (OntoNotes 5.0) is a corpus integrating six different vertical domains. The $F1$ values of the model proposed in this study are 2.18%, 1.68%, and 0.99% higher than BiLSTM-CRF, respectively.

Key words: named entity recognition (NER); transfer learning; cross-domain; dynamic fusion; bidirectional long short-term memory (BiLSTM) neural network

命名实体识别是从非结构化文本中抽取指定实体类型的提及(mentions)[1], 例如, 从新闻领域中识别人名、地名和组织名等通用命名实体, 或者从医疗领域的临床记录中识别疾病、症状和药物等命名实体[2], 是知识图谱、问答系统和关系抽取等下游任务中的关键步骤.

大量研究工作关注于从英文文本中抽取命名实体, Huang 等人[3]通过融合英文的拼写特征(首字母大写)和上下文特征到长短期记忆神经网络和条件随机场(LSTM-CRF)来抽取命名实体. Ma 等人[4]分别通过卷积神经网络和长短期记忆神经网络融合英文文本中的字级别和词级别特征来抽取英文命名实体.

虽然针对英文文本的命名实体抽取模型已取得明显进展[3,4], 中文命名实体识别模型相比于英文命名实体识别模型的难点在于中文文本不存在明显的分隔符, 难以确定中文命名实体的边界[5]. 现有基于管道(pipeline)的中文命名实体识别模型首先对中文文本分词, 然后采用基于词的统计机器学习模型或者神经网络模型抽取中文命名实体. 但是中文分词一方面会耗费大量人力资源, 另一方面会将分词引入的误差传递到中文命名实体模型, 导致错误地识别了中文命名实体的边界.

针对基于管道模型的中文命名实体识别模型中, 中文分词任务引入的弊端, 现有主流的中文命名实体识别模型是采用基于字的端到端(end-to-end)序列标注模型[5−8], 通过对每个中文字分配特定的标签来识别命名实体. 具体例子如图 1 所示.

北	京	欢	迎	您	.
B-LOC	E-LOC	O	O	O	O

图 1 序列标注模型所需标注语料

“北京”被标记为“B-LOC E-LOC”, 其中, “B”和“E”分别表示命名实体的起始字和终止字, “LOC”表示该命名实体是地名类型(LOCATION). “欢迎您.”均被标记为“O”标签, 表示为非实体字(non-entity word).

虽然上述基于字的端到端序列标注的中文命名识别模型在有大量高质量标注语料的通用领域上有较好的表现, 然而在一些垂直领域(医疗领域、金融领域等)或者混合领域(融合多个领域)中, 基于字的端到端序列标注的中文命名实体识别模型的效果相比于通用领域有着明显的下滑, 其主要原因如下.

(1) 有限的高质量标注语料[9]. 由于垂直领域的数据标注需要较强的专业知识, 标注同等数量的语料需

耗费更多的资源;

(2) 命名实体类别数量的增多. 通用领域中命名实体类型数量一般少于垂直领域或者混合领域, 例如, 微软亚洲研究院数据集 MSRA (http://sighan.cs.uchicago.edu/bakeoff2006/)(通用领域)中包含人名、地名与组织名这 3 类命名实体, 而在垂直领域或者混合领域中, 由于抽取的命名实体需满足复杂的应用场景, 命名实体类别数量相比于通用领域会有所增多. 例如, OntoNotes 5.0 数据集(http://cemantix.org/data/ontonotes.html)(混合领域)包含 18 种命名实体类别, CCKS 2017 数据集(http://www.ccks2017.com/en/index.php/sharedtask/)(医疗领域)包含 5 种命名实体类型.

针对上述因语料不足而导致实体识别效果下降的问题, 领域迁移能够缓解中文实体识别模型性能的下降, 是一种领域自适应(domain adaption)的迁移学习, 通过利用具有充足标记语料的源领域(source domain)上的知识来迁移到标记语料不足的目标领域(target domain)[9]. 现有应用到中文命名实体抽取任务上的领域迁移方法主要分为基于参数初始化的领域迁移方法和基于多任务的领域迁移方法[9], 其中, 基于参数初始化的领域迁移方法首先利用在源领域数据集上预训练好的网络模型参数初始化目标领域上的模型, 然后根据目标领域上的数据对初始化后的参数进行自适应更新; 基于多任务的领域迁移方法通过共享浅层网络模型来同时训练源领域和目标领域上的网络模型.

虽然上述两种基于领域迁移的中文命名实体识别方法在缓解高质量标注数据不足的问题上取得了不错的效果, 但是也忽略了源领域中特有的命名实体语义对目标领域上中文实体识别模型的影响. 例如, 通用领域中地名(LOCATION)实体的语义会影响中文命名实体识别模型在医疗领域中识别出疾病(DISEASE)实体, 尤其是当目标领域的实体类别数量多于源领域时, 迁移源领域上特有的语义信息到目标领域会严重影响目标领域上命名实体识别模型的性能.

为解决上述问题, 本文提出一种动态迁移实体块信息的跨领域中文命名实体识别模型(TES-NER), 通过显性指定领域之间迁移内容为实体块信息, 并且利用基于门机制的动态融合层, 将实体块信息从源领域动态迁移到目标领域上的中文命名实体识别模型. 其中, 实体块信息只考虑命名实体的提及(mention)而忽略了实体类型. TES-NER 模型首先采用共享的字向量嵌入层与长短期记忆神经网络模块抽取跨领域共享的上下文信息; 其次引入实体块识别模块, 通过长短期记忆神经网络和全连接层获取领域之间迁移的实体块信息; 再次, 利用源领域和目标领域上的中文命名实体识别模块通过长短期记忆神经网络和条件随机场(BiLSTM-CRF)来识别领域指定的中文命名实体; 然后采用基于门机制(GateMechanism)的动态融合层(如后文图 2 中蓝色圆圈部分所示), 根据源领域(目标领域)中上下文表征来动态决定融合到源领域(目标领域)上中文命名实体识别模型的实体块信息的量; 最后, 采用多任务的方式来训练 TES-NER 模型. 模型的整体框架如后文图 2 所示.

TES-NER 模型的优势有: (1) 显性指定实体块信息有助于更精确地识别中文命名实体的边界; (2) 迁移跨领域共享的实体块信息能够缓解源领域中特有的命名实体语义信息对目标领域上模型的影响, 提高目标领域上中文命名实体识别的性能; (3) 动态融合层通过控制领域间迁移实体块信息的量来避免模型过于关注中文文本中的非实体字(non-entity word).

本文提出的模型在混合领域(OntoNotes 5.0)、金融领域(Resume)和医学领域(CCKS 2017)这 3 个数据集上的性能均优于双向长短期记忆神经网络和条件随机场模型(BiLSTM-CRF), 分别在 $F1$ 值上达到 75.21%、95.49%和 90.82%, 且构建统计验证实验结果也表明本文提出的方法在统计上是有意义的. 上述 3 个数据集被视为目标领域上中文命名实体识别的数据集, 源领域的数据集则采用通用领域的 MSRA 语料. 此外, 通过可视化归一化后的实体块信息相关的门向量发现: 命名实体识别模型更加关注于共享实体块中的起始字, 进一步验证了本文提出方法的有效性.

本文第 1 节介绍关于中文实体抽取的相关工作. 第 2 节介绍本文提出的模型和方法. 第 3 节列出对比实验结果和分析. 第 4 节是结论和下一步工作.

1 相关工作

基于机器学习方法的中文实体识别模型是基于手动构造的特征, 其中包含最大熵模型(ME)[10]、条件随机场(CRF)[11]和隐马尔可夫(HMMs)[12]. 随着深度神经网络的流行, 大量工作提出基于神经网络的中文实体抽取模型, 通过自动提取特征来完成中文实体识别任务, 缓解因手动提取特征带来的人力消耗. Huang 等人[3]提出采用双向长短期记忆神经网络和条件随机场来识别命名实体, 其中, 双向长短期记忆神经网络用于获取句子的上下文特征, 条件随机场用于建立标签之间的依赖关系. Chen 等人[13]提出采用卷积神经网络模型来解决基于循环神经网络无法并行化和计算效率低的问题, 然而, 因无法良好地建模文本的位置信息, 其性能相比于基于循环神经网络的模型略有下降. Zhu 等人[14]首先融合卷积神经网络和自注意力机制提取文本的局部上下文特征, 然后采用循环神经网络获取文本的长期上下文特征, 以进一步提升模型的性能.

近期, 基于迁移学习的中文实体抽取方法受到研究人员的关注. 基于迁移学习的中文实体抽取模型主要分为:

(1) 多任务的实体抽取模型

由于中文分词任务中包含丰富的实体边界信息, Peng 等人[15]通过联合中文实体分词任务和中文实体抽取任务的多任务模型来利用分词任务中的边界信息. 然而, 上述方法存在中文分词边界和实体边界不一致的问题. 为解决上述问题, Cao 等人[16]提出了对抗迁移神经网络来获取任务之间共享的边界信息, 同时抑制中文分词特有的边界信息. 于此同时, Peng 等人[17]迁移语言模型中的知识到实体识别模型中来, 进一步学习到句子的语义信息.

(2) 基于跨领域的实体抽取模型

Yang 等人[18]在不同领域之间设计深度分层的多任务循环神经网络, 通过共享隐藏层来迁移领域之间的知识. Lin 等人[19]设计了 3 个自适应层来解决领域偏移的问题. 虽然上述方法都取得了一定的效果, 但是只通过共享的隐藏层来迁移领域之间或者不同任务之间的知识, 即未显性指定不同领域之间迁移的语义信息. 本文提出的动态迁移实体块信息的跨领域中文实体抽取方法显性指定领域之间迁移的实体块信息, 且通过动态融合层来控制其迁移信息的量.

为充分利用标注的命名实体语料, 研究人员关注于如何利用实体块信息或者实体边界信息来提升命名实体识别的性能. Aguilar 等人[20]提出了命名实体分割和命名实体抽取的多任务网络, 其中, 命名实体分割采用 sigmoid 网络, 命名实体抽取采用条件随机场. 为了更好地建模命名实体块信息, Aguilar 等人[21]将 sigmoid 网络替换为条件随机场来获取实体块信息, 即将建模实体块信息的原有二分类任务替换为多分类任务, 且采用 BIO 标记体系. Zhai 等人[22]首先通过指针网络(pointer network)来显性提取实体块信息, 然后分类到预定义的命名实体集合上. Xiao 等人[23]通过相似度模型将语义标签向量融合到命名实体抽取模型, 用于判别字是否为实体字(entity word), 其中, 语义标签向量由两个随机初始化的向量构成, 分别用于表示实体字与非实体字(non-entity word). 同时, 设计了辅助函数来学习实体块信息. Zheng 等人[24]首先通过识别实体边界来获取实体块信息, 然后采用 softmax 网络将其分类到预定义的嵌套命名实体类型集合. 然而, 上述利用实体块信息的方法未考虑如何在跨领域之间共享实体块信息, 本文提出: 基于迁移学习从语料充足的源领域中显性引入实体块信息到语料稀有的目标领域来提升命名实体模型的性能.

2 模型框架

本节将详细阐述 TES 模型的细节和训练过程. 首先介绍共享的词向量嵌入层和双向长短期记忆神经网络模块, 用于获取不同领域不同任务之间通用的共享特征; 其次, 设计实体块识别模块来抽取源领域与目标领域共享的实体块信息, 其包含用于提取领域共享的实体块信息相关的上下文表征的双向长短期记忆神经网络和实体块任务指定的输出层; 然后, 目标领域(源领域)的中文命名实体识别模块去抽取目标领域(源领域)中实体识别任务相关的特征, 并将融合实体块信息到模型中, 其包含用于提取上下文表征的双向长短期记忆神经

网络、用于动态融合实体块信息的动态融合层和用于建模实体标签之间依赖关系的命名实体识别任务指定的输出层；最后，通过联合实体块抽取任务和实体识别任务来训练网络模型. 模型的框架如图 2 所示(左边部分和右边部分表示分别从源领域和目标领域中抽取中文命名实体，中间部分是用于提取共享的实体块表示(紫色部分)，并通过动态融合层(蓝色圆圈部分)将其分别融合到源领域和目标领域的命名实体抽取模型).

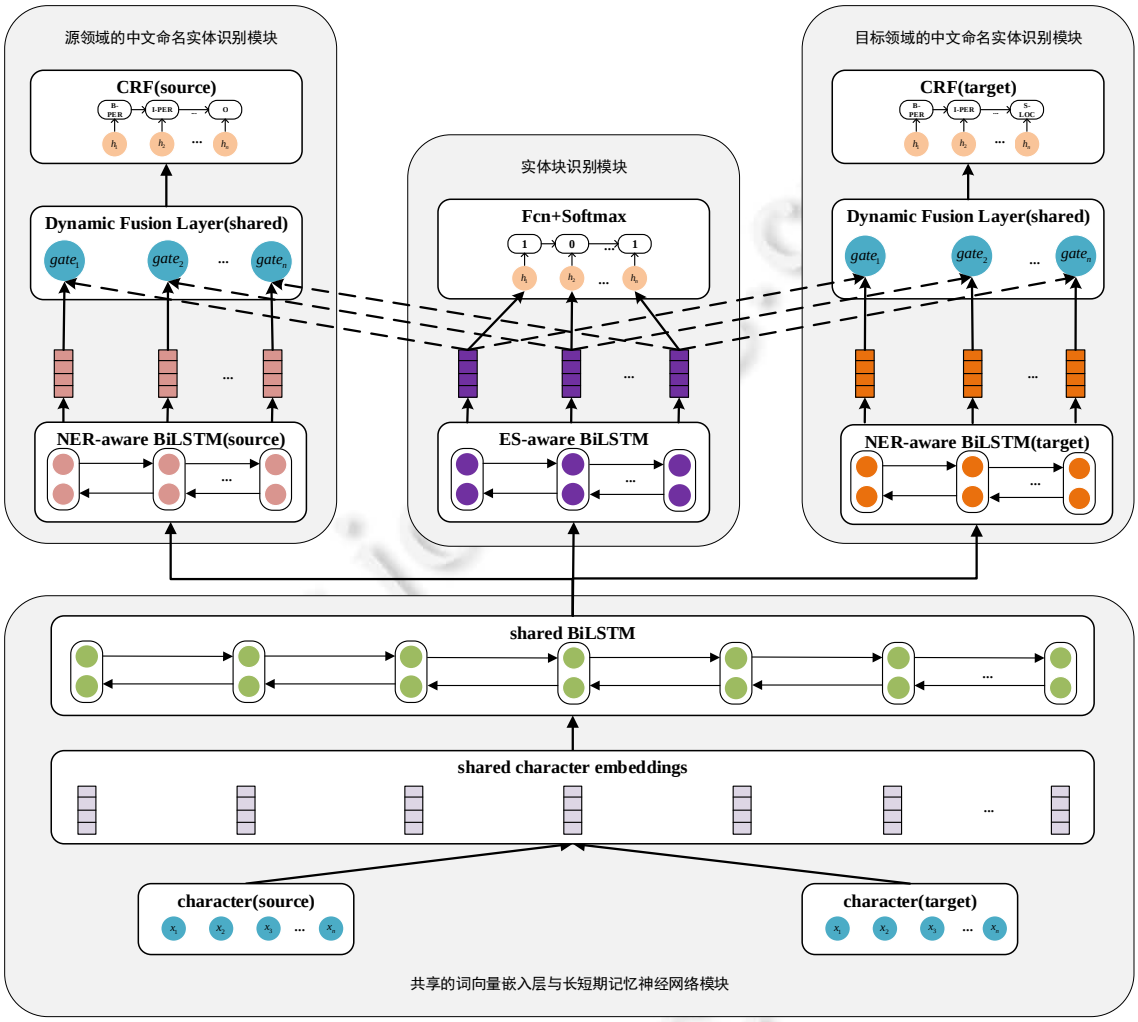


图 2 TES-NER 模型的整体框架

2.1 共享的词向量嵌入层与长短期记忆神经网络模块

为了在跨领域和跨任务之间迁移共享的语义信息，本节设计了一个共享的词向量嵌入层和共享的双向长短期记忆神经网络，以分别将字映射为稠密的向量和提取共享的长距离上下文特征.

给定一个源领域或者目标领域的句子 $\boldsymbol{c}=(c_1,c_2,\ldots,c_n)$，其中，$c_n$ 表示一个中文字，共享的词向量嵌入层通过字向量映射表将字映射为字向量 $\boldsymbol{W}=(\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_n)$:

$$\boldsymbol{w}_i=T(c_i) \tag{1}$$

其中，T 为字向量的映射表. 然后输入到共享的双向长短期记忆神经网络，输出长距离依赖关系的上下文表征特征 $\boldsymbol{H}=(\boldsymbol{h}_1,\boldsymbol{h}_2,\ldots,\boldsymbol{h}_n)$:

$$\vec{\boldsymbol{h}}_i=\overrightarrow{LSTM}(\boldsymbol{w}_i) \tag{2}$$

$$\overleftarrow{\boldsymbol{h}_i} = \overleftarrow{LSTM}(\boldsymbol{w}_i) \tag{3}$$

$$\boldsymbol{h}_i = \overrightarrow{\boldsymbol{h}_i} \oplus \overleftarrow{\boldsymbol{h}_i} \tag{4}$$

其中, $\overrightarrow{LSTM}$ 和 $\overleftarrow{LSTM}$ 分别表示前向长短期记忆神经网络和反向长短期记忆神经网络, 分别用于获取当前特征 $\boldsymbol{h}_i$ 的历史信息和未来信息. ⊕表示向量拼接操作.

2.2 实体块识别模块

实体块识别模块是源领域和目标领域共享的网络模块, 用于识别各自领域的实体块信息, 为领域之间迁移实体块信息提供基础. 实体块识别模块由实体块指定的长短期记忆神经网络(ES-aware BiLSTM)与实体块指定的输出层(FNN+softmax)构成.

2.2.1 实体块识别任务指定的双向长短期记忆神经网络

实体块识别任务指定的双向长短期记忆神经网络是受Cao等人[16]提出的对抗迁移神经网络中任务判别器启发, 通过双向长短期记忆神经网络(BiLSTM)来抽取不同领域之间共享的实体块信息. 具体而言, 将第 2.1 节中得到的上下文表征特征 $\boldsymbol{H}=(\boldsymbol{h}_1,\boldsymbol{h}_2,\ldots,\boldsymbol{h}_n)$输入到双向长短期记忆神经网络(BiLSTM), 输出实体块指定的上下文表征 $\boldsymbol{S}=(\boldsymbol{s}_1,\boldsymbol{s}_2,\ldots,\boldsymbol{s}_n)$:

$$\boldsymbol{s}_i = BiLSTM_{seg}(\boldsymbol{h}_i) \tag{5}$$

其中, $BLSTM_{seg}$ 表示实体块信息相关的双向长短期记忆网络.

需强调的是: 该模块的网络模型参数在领域之间共享, 通过共享模型参数的方式来迁移不同领域之间的实体块信息.

2.2.2 实体块识别任务指定的输出层

为显性指定获取领域之间共享的实体块信息, 参考 Aguilar 等人[20]的工作, 将实体块指定的上下文表征 $\boldsymbol{S}=(\boldsymbol{s}_1,\boldsymbol{s}_2,\ldots,\boldsymbol{s}_n)$依次输入到全连接神经网络层(fully connected neural network)和 Sigmoid 激活函数, 得到实体块指定的输出向量 $\boldsymbol{O}^s=(o_1^s,o_2^s,\ldots,o_n^s)$, 用于区别文本中的实体字与非实体字:

$$\boldsymbol{o}_i^s = \sigma(\boldsymbol{W}^s\boldsymbol{s}_i + b^s) \tag{6}$$

其中, σ为 Sigmoid 激活函数, 将输出值限制在[0,1]之间; $\boldsymbol{W}^s$ 和 b^s 为模型的可训练参数.

2.3 目标领域(源领域)的中文命名实体识别模块

目标领域的中文命名实体识别模块用于识别目标领域中的命名实体, 包含用于获取实体识别任务指定的上下文表征的双向长短期记忆网络、用于融合实体块信息的动态融合层和用于建立命名实体标签之间依赖关系的输出层. 由于源领域与目标领域上的中文命名实体识别模型相同, 本节只阐述目标领域上的中文命名实体识别模型.

2.3.1 命名实体任务指定的长短期记忆神经网络

类似第 2.2.1 节的实体块识别任务指定的双向长短期记忆神经网络, 本节采用双向长短期记忆神经网络来抽取目标领域上中文命名实体任务相关的上下文表征向量. 具体而言, 给定第 2.1 节中共享双向长短期记忆网络输出的上下文表征特征 $\boldsymbol{H}=(\boldsymbol{h}_1,\boldsymbol{h}_2,\ldots,\boldsymbol{h}_n)$, 输入到目标领域上命名实体任务指定的双向长短期记忆神经网络(NER-aware BiLSTM (target)), 得到目标领域中命名实体的上下文表征特征 $\boldsymbol{R}_t=(\boldsymbol{r}_1^t,\boldsymbol{r}_2^t,\ldots,\boldsymbol{r}_n^t)$:

$$\boldsymbol{r}_i^t = BiLSTM_{ner}^t(\boldsymbol{h}_i) \tag{7}$$

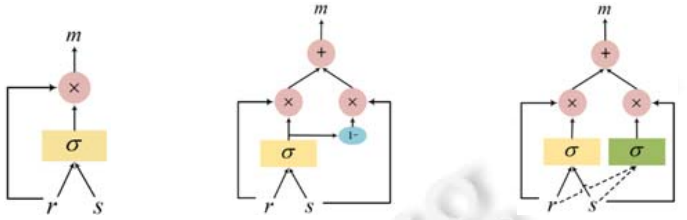
其中, $BiLSTM_{ner}^t$ 分别表示目标实体识别相关的双向长短期记忆网络.

2.3.2 动态融合层

将实体块信息融合到命名实体识别任务, 有助于提升命名实体的性能[23]. Aguilar 等人[20]提出了融合实体块分类任务与命名实体识别任务的多任务学习, 通过共享双向长短期记忆神经网络来隐性地利用实体块信息, 其存在两个问题: (1) 仅仅通过共享隐藏层无法有效地获取领域之间共享的语义信息[19], 尤其是当实体类别数量增多时, 该问题会变得愈发棘手; (2) 命名实体字与非命名实体字是融合相同权重大小的实体块信息,

同时，非实体字相比于实体字在文本中占比高，由此导致模型更加关注非实体字.

为解决上述不足，本节提出 3 种基于门机制的动态融合层，通过动态决定融合实体块信息的量来提升目标领域上命名实体抽取任务的性能. 首先，根据实体块的上下文表征向量 $\boldsymbol{S}$ 和目标领域中命名实体的上下文表征向量 $\boldsymbol{R}_t$ 来计算得到门向量(gate)；然后，设计单一融合、共享融合和分离融合这 3 种不同的融合策略，将实体块信息动态融合到命名实体相关的上下文表征向量，如图 3 所示(单一融合策略和共享融合策略采用一个门向量，而分离融合策略采用两个不同的门向量去融合实体块识别任务相关的上下文表征信息和命名实体识别任务相关的上下文表征信息).



(a) 单一融合策略 (b) 共享融合策略 (c) 分离融合策略

图 3 3 种不同的融合策略

具体而言，给定第 2.2 节的实体块相关的上下文表征向量 $\boldsymbol{S}$ 和第 2.3.1 节的目标领域中命名实体的上下文表征向量 $\boldsymbol{R}_t$，首先通过门机制计算得到实体块相关的门向量 $\boldsymbol{g}_{es}$ 和目标领域上命名实体相关的门向量 $\boldsymbol{g}_{ner}^t$：

$$\boldsymbol{g}_{es}=\sigma(\boldsymbol{W}_1\boldsymbol{S}+\boldsymbol{W}_2\boldsymbol{R}_t+\boldsymbol{b}^1) \tag{8}$$

$$\boldsymbol{g}_{ner}^t=\sigma(\boldsymbol{W}_3\boldsymbol{S}+\boldsymbol{W}_4\boldsymbol{R}_t+\boldsymbol{b}^2) \tag{9}$$

其中，$\boldsymbol{W}_1$、$\boldsymbol{W}_2$、$\boldsymbol{b}^1$、$\boldsymbol{W}_3$、$\boldsymbol{W}_4$ 和 $\boldsymbol{b}^2$ 是可训练的网络模型参数；σ表示 sigmod 激活函数，将输出值的取值范围限制在[0,1].

然后，设计 3 种不同的融合策略来得到目标领域上的融合向量 $\boldsymbol{M}^t=(\boldsymbol{m}_1^t,\boldsymbol{m}_2^t,...,\boldsymbol{m}_n^t)$.

(1) 单一融合策略

只有实体抽取相关的上下文特征参与融合，实体块信息只用于计算得到命名实体相关的门向量 $\boldsymbol{g}_{ner}^t$：

$$\boldsymbol{M}^t=\boldsymbol{g}_{ner}^t\circ\boldsymbol{R}_t \tag{10}$$

其中，$\circ$表示矩阵中元素之间的乘积.

(2) 共享融合策略

采用共享的门向量去融合实体块的上下文表征向量 $\boldsymbol{s}$ 和目标领域中命名实体的上下文表征向量 $\boldsymbol{R}_t$:

$$\boldsymbol{M}^t=(1-\boldsymbol{g}_{ner}^t)\circ\boldsymbol{S}+\boldsymbol{g}_{ner}^t\circ\boldsymbol{R}_t \tag{11}$$

(3) 分离融合策略

与共享融合策略的不同之处在于，分离融合策略假设实体块信息和命名实体相关的信息不是互斥的. 由此，该融合策略分别采用 $\boldsymbol{g}_{es}$ 和 $\boldsymbol{g}_{ner}^t$ 不同的门向量计算得到目标领域上的融合向量 $\boldsymbol{M}^t$:

$$\boldsymbol{M}^t=\boldsymbol{g}_{es}\circ\boldsymbol{S}+\boldsymbol{g}_{ner}^t\circ\boldsymbol{R}_t \tag{12}$$

2.3.3 命名实体任务指定的输出层

由于自然语言句子中存在句法限制，因此标签之间的依赖关系十分重要[25]. 例如，“I-LOCATION”标签不会出现在“B-PERSON”标签之后. 与此同时，条件随机场(CRF)是广泛应用到序列标注任务去获取标签之间的依赖关系[4]，例如，命名实体抽取任务和中文分词任务等. 因此，本节采用条件随机场作为命名实体任务指定的输出层，并将其建立在动态融合层之上，用于获取目标领域中输出的命名实体标签之间的依赖关系.

将第 2.3.2 节中动态融合层输出的融合向量 $\boldsymbol{M}^t=(\boldsymbol{m}_1^t,\boldsymbol{m}_2^t,...,\boldsymbol{m}_n^t)$ 输入到条件随机场中，输出命名实体标签序列 $\boldsymbol{y}^t=(y_1^t,y_2^t,...,y_n^t)$:

$$p(\boldsymbol{y}^t \mid \boldsymbol{M}^t)=\frac{\exp\left(\sum_{i=1}^{n}\boldsymbol{W}_{y_i^t,y_{i-1}^t}\boldsymbol{m}_i^t+\boldsymbol{b}_{y_i,y_{i-1}}\right)}{\sum_{y'\in Y^t}\exp\left(\sum_{i=1}^{n}\boldsymbol{W}_{y_i',y_{i-1}'}\boldsymbol{m}_i^t+\boldsymbol{b}_{y_i',y_{i-1}'}\right)} \tag{13}$$

其中, $p(\boldsymbol{y}^t|\boldsymbol{M}^t)$是在给定融合向量$\boldsymbol{M}^t$的基础上输出序列$\boldsymbol{M}^t$的条件概率, $\boldsymbol{W}_{y_i^t,y_{i-1}^t}$和$\boldsymbol{b}_{y_i,y_{i-1}}$是条件随机场中可训练的参数, 其下标表示相邻的标签对(y_i,y_{i-1}), Y^t表示目标领域中实体标签的集合.

2.4 多任务的跨领域迁移学习

多任务的跨领域迁移学习已在命名实体任务中获取优异效果, 通过将不同领域和不同任务的知识迁移到目标领域的命名实体抽取任务. Lee 等人[26]通过共享隐藏层的方式将不同领域的知识迁移到目标领域, 然而, 由于不同领域之间实体类型数量的差异较大, 例如, OntoNotes 4.0 数据集中存在 18 类实体, 而 MSRA 数据集中仅存在 3 类实体, 导致不同领域间转移的语义知识模糊不清.

为解决上述问题, 本节设计了实体块识别的辅助任务, 通过构造的二值交叉熵损失函数来训练得到不同领域迁移的实体块信息, 解决因只采用共享层而导致迁移知识模糊的问题. 给定第 2.2.2 节得到的实体块指定的输出向量$\boldsymbol{O}^s=(o_1^s,o_2^s,...,o_n^s)$, 将二值交叉熵损失函数作为实体块抽取任务的损失函数:

$$loss_{es}(\boldsymbol{O}^s)=-\sum_{i=1}^{n}y_i'\log o_i^s+(1-y_i')\log(1-o_i^s) \tag{14}$$

是真实的实体块标签, 其中, 1 表示字为实体字, 0 表示字为非实体字.

此外, 本节构建目标领域和源领域的两个负对数似然函数作为命名实体抽取任务的损失函数, 其计算方式如下所示:

$$loss_t(\boldsymbol{M}^t)=-\log_{y^t\in Y^t}(p(\boldsymbol{y}^t\mid\boldsymbol{M}^t)) \tag{15}$$

$$loss_s(\boldsymbol{M}^s)=-\log_{y^s\in Y^s}(p(\boldsymbol{y}^s\mid\boldsymbol{M}^s)) \tag{16}$$

其中, $p(\boldsymbol{y}^t|\boldsymbol{M}^t)$是第 2.3.3 节中条件随机场的条件概率, Y^t和Y^s分别表示目标领域和源领域的命名实体类型集合. 由此, 模型最终的损失函数是将上述 3 类损失函数按照权重相加:

$$Loss_{total}=loss_t+\alpha loss_{es}+\beta loss_s \tag{17}$$

其中, α和β分别是模型的超参数, 分别用于控制实体块识别任务和源领域中命名实体识别任务的重要程度.

3 实 验

本节首先介绍源领域和目标领域上的数据集, 接着介绍算法评估和模型超参数等实验设置, 然后汇报 TES-NER 模型在不同目标领域数据集上的实验结果, 最后详细分析 TES-NER 模型, 例如, 不同融合策略下的实验分析等. TES-NER 模型的源代码将在 https://github.com/paulpig/TES-NER 中公开.

3.1 数据集

本文将 MSRA 数据集[11]作为源领域上中文命名实体识别任务的语料, 将中文 OntoNotes 5.0 数据集[27]、Resume 数据集[28]和 CCKS 2017 数据集(http://www.ccks2017.com/en/index.php/sharedtask/)作为目标领域的中文实体识别任务的语料, 其中, MSRA 语料是属于新闻领域; OntoNotes 数据集属于混合领域, 包含新闻热线、广播新闻、广播对话、杂志、电话通话和网络数据这 6 个子领域; Resume 数据集属于金融领域; CCKS 2017 数据集属于医学领域, 通过人工标注电子病历得到. 实验结果建立在目标领域的中文实体识别任务上, 不考虑源领域上中文命名实体任务的性能. 目标领域和源领域上的语料统计信息见表 1.

由于 CCKS 2017 数据集与 MSRA 数据集不存在验证集, 本文将各自的训练集中随机抽取 20%作为验证集, 其余数据作为训练集.

表 1 目标领域与源领域上的语料统计信息

数据集	领域	句子数量	字数量	句子的平均长度
OntoNotes	新闻热线	4 180	171 860	41.1
	广播新闻	9 483	459 830	48.5
	广播对话	10 960	251 314	22.9
	杂志	4 801	239 791	49.9
	电话通话	9 607	130 839	13.6
	网络数据	8 011	264 503	33.0
Resume	金融领域	4 761	153 089	32.2
CCKS 2017	医学领域	1 597	348 754	218.4
MSRA	新闻领域	55 289	2 342 480	42.3

OntoNotes 5.0 数据集、Resume 数据集、CCKS 2017 数据集和 MSRA 数据集分别包含 18 类命名实体类型、8 类命名实体类型、5 类命名实体类型和 3 类命名实体类型, 详情的实体类别信息见表 2.

表 2 目标领域与源领域上语料中的命名实体类型详情

数据集	类别数量	命名实体类型
OntoNotes	18 类	人名(person)、宗教团体(norp)、服务设施(facility)、组织(organization)、地理(gpe)、地名(location)、产品(product)、事件(event)、艺术品(work of art)、法律(law)、语言(language)、日期(date)、时间(time)、百分比(percent)、钱(money)、数量(quantity)、序数词(ordinal)、其余数字(cardinal)
Resume	8 类	国家(country)、教育机构(educational institution)、地名(location)、人名(personal name)、组织名(organization)、专业(profession)、种族背景(ethnicity background)和职位名称(job title)
CCKS 2017	5 类	身体部位(body)、症状(symptoms)、检查(check)、疾病(disease)和治疗(treatment)
MSRA	3 类	人名(person)、机构(organization)和地名(location)

3.2 实验设置

- 算法评估

本文实验采用准确率(P)、召回率(R)和 $F1$ 值评估模型的性能.

$$P=\frac{TP}{TP+FP} \tag{18}$$

$$R=\frac{TP}{TP+FN} \tag{19}$$

$$F1=\frac{2\times P\times R}{P+R} \tag{20}$$

其中, TP、FP 和 FN 分别为正阳性(true positive)、假阳性(false positive)和假阴性(false negative).

- 超参数设置

表 3 展示了 TES-NER 模型的超参数, 其中, 100 维的字向量是通过 word2vec 方法从百度百科和微博语料中训练得到[16]. 不同模块中的双向长短期记忆神经网络的层数和隐藏层均设置为 2 和 300. 为了防止模型过拟合, 设置 *Dropout* 层的比率为 0.5. 此外, 设置初始学习率为 0.001, 且采用 Adam 优化器来训练模型. 最后, 设置实体块识别任务的权重α和源领域命名实体任务的权重β分别为 0.5 和 0.8.

表 3 超参数的值

参数	值
字向量大小	100
LSTM 网络的隐藏层	300
LSTM 网络的层数	2
Dropout 层的比率	0.5
初始学习率	0.001
优化器	Adam
α	0.5
β	0.8

- 标注体系

采用 BIOES 的标注方式来标注命名实体, 例如“B-PER I-PER E-PER O”标签序列, “B-PER”“I-PER”和“E-PER”分别表示人名实体的起始字、内部字和终止字, “O”表示非实体字.

- 实验环境

实验环境基于 Ubuntu18.04 操作系统的 PyTorch 1.6 深度学习框架.硬件配置为: Inter(R) Core(TM) i7-8700 K CPU@3.7 GHz, NVIDIA GeForce GTX 1080Ti, 32 GB 内存.

- 融合策略

后续实验均采用分离融合策略, 假设实体块信息和命名实体信息是非互斥, 不同融合策略的详细分析见第 2.3.2 节.

3.3 实验结果

在本节中, 我们将本文提出的模型分别在 OntoNotes 5.0 数据集、Resume 数据集和 CCKS 2017 数据集上与之前的方法进行比较, 以验证本文提出的模型的有效性.

3.3.1 OntoNotes 5.0 数据集

如表 4 所示, 在 OntoNotes 5.0 数据集上的整体性能分析分为 3 部分: 第 1 部分列出之前研究在 OntoNotes 5.0 数据集上的中文实体识别模型和方法, 第 2 部分列出基于迁移学习的中文实体识别模型和方法, 第 3 部分为本文提出的 TES-NER 模型. 由于之前研究提出的基于迁移学习的中文命名实体识别方法未在 OntoNotes 5.0 数据集上验证, 本节通过在 OntoNotes 5.0 数据集上运行其开源的代码或者复现论文中的模型, 得到最终的实验结果.

表 4 OntoNotes 5.0 数据集上的整体性能

模型	*P*	*R*	*F*1
Pradhan 等人[29]	78.20	66.45	71.85
Lattice LSTM[28]	76.34	77.01	76.67
Word-based BiLSTM-CRF[30]	**77.94**	75.33	76.61
DGLSTM-CRF[30]	77.40	**77.41**	**77.40**
Lee 等人[26]	74.03	74.3	74.16
Yang 等人[18]	**75.53**	74.6	75.06
Cao 等人[16]	73.67	74.93	74.30
Char-based BiLSTM-CRF	73.15	72.88	73.02
TES-NER	75.21	**75.22**	**75.21**

在表 4 的第 1 部分中, Pradhan 等人[29]采用斯坦福开源的命名实体识别工具(https://nlp.stanford.edu/software/CRF-NER.shtml)在 OntoNotes 5.0 数据集上的 *F*1 值达到 71.85%. LatticeNER 是由 Zhang 等人[28]提出的, 通过融合句子中所有潜在的词向量信息, 在 OntoNotes 5.0 数据集上的 *F*1 值达到 76.67%. 然而 LatticeNER 模型的效率较低, 由于受到自身网络结构的限制, 导致其运行的批大小(batch size)只能设置为 1. 此外, LatticeNER 除利用预训练好的字向量之外, 还需利用预训练好的词向量. Jie 等人[30]提出了 DGLSTM-CRF, 通过融合词向量的上下文信息与句法信息, 在 OntoNotes 5.0 数据集上的 *F*1 值达到 77.40%. 虽然 DGLSTM-CRF 模型目前在 OntoNotes 5.0 数据集上达到了最优性能, 但其本质上是基于词的命名实体抽取模型, 需要人工对句子进行分词操作, 因此会耗费大量的人力资源.

基于领域迁移的中文实体识别方法在保证模型的效率和无需人工进行分词操作的前提下, 通过迁移源领域中的语义知识来提升目标领域上中文实体识别模型的性能. 在表 4 的第 2 部分中, Lee 等人[26]提出了基于参数初始化的领域迁移学习, 将源领域实体识别模型中的双向长短期记忆神经网络的参数来初始化目标领域中对应的模型, 其在 OntoNotes 5.0 数据集上的 *F*1 值达到了 74.16%. Yang 等人[18]提出了基于多任务的领域迁移模型, 通过不同领域之间共享双向长短期记忆神经网络来达到迁移语义知识的目的, 其在 OntoNotes 5.0 数据集上的 *F*1 值达到了 75.06%. 由于上述方法未在原始论文中汇报 OntoNotes 5.0 数据集上的性能, 本文将上述方法应用到 MSRA 数据集作为源领域和 OntoNotes 5.0 数据集作为目标领域中得到实验结果. Cao 等人[16]通过

迁移中文分词任务中的语义知识到中文实体抽取任务, 本文将中文分词任务替换, 以 MSRA 作为源领域的中文实体识别任务, 从而修改为领域自适应的中文命名实体抽取模型, 并在 OntoNotes 5.0 数据集上的 $F1$ 值达到了 74.30%.

从表 4 中的第 3 部分可观察得到: 本文提出的动态迁移实体块信息的跨领域中文实体识别模型(TES-NER)在 OntoNotes 5.0 数据集上的 $F1$ 值不仅超越了上述基于领域迁移的中文实体识别方法, 且比基于字的基线模型(char-based BiLSTM-CRF)高出 2.19%(本文采用成对 t 检验(pairwise t-test)的 p 值小于 0.01 来表明 TES-NER 模型在统计学上显著优于基线模型(char-based BiLSTM CRF), 详情见后文第 3.4.4 节). 模型性能的提升可能是因为: (1) 迁移源领域的实体块信息能够缓解命名实体识别模型同时识别实体块(命名实体的范围)和实体类型的压力, 让目标领域的命名实体识别模型更加关注于对实体类型的识别; (2) 动态融合层能够根据模型自身通过门机制来动态确定融合的实体块信息的量, 以减少噪声的引入.

3.3.2 Resume 数据集和 CCKS 数据集

表 5 展示了在 Resume 数据集上的实验结果, 其中, Lattice LSTM[28]模型和 LR-CNN[31]模型分别采用双向长短期记忆神经网络和卷积神经网络来融合所有潜在的词向量, 分别在 Resume 数据集上的 $F1$ 值达到 94.46%和 95.11%. 本文提出的 TES-NER 模型相比于上述两种模型可在 $F1$ 值上分别超出 1.03%和 0.38%, 而且在 $F1$ 值上比基于字的基线模型(char-based BiLSTM-CRF)高出 1.68%.

表 5 Resume 数据集上的性能

模型	P	R	$F1$
Lattice LSTM[28]	94.81	94.11	94.46
LR-CNN[31]	95.37	94.84	95.11
Char-based BiLSTM-CRF	94.13	93.5	93.81
TES-NER	**95.57**	**95.4**	**95.49**

表 6 展示了在 CCKS 2017 数据集上不同模型的性能, 其中, Zhang 等人[32]提出的模型融合多任务学习、自注意力(self-attention)和多步训练方法去获取丰富的语义特征, 在 CCKS 2017 数据集上达到 90.52%的 $F1$ 值. Zhao 等人[33]提出的模型采用对抗训练和所有潜在的词向量语义信息来识别命名实体, 其在 CCKS 2017 数据集上的性能达到 89.64%的 $F1$ 值. 从表 6 可以观察到: 本文提出的模型在 $F1$ 值上不仅超越了上述研究提出的模型, 而且优于基于字的基线模型.

表 6 CCKS 2017 数据集上的性能

模型	P	R	$F1$
Zhang 等人[32]	89.24	91.83	90.52
Zhao 等人[33]	88.98	90.28	89.64
Char-based BiLSTM-CRF	89.27	90.4	89.83
TES-NER	**90.29**	**91.35**	**90.82**

3.3.3 消融实验

为进一步验证本文提出的模型中动态融合层和领域迁移学习的有效性, 本节分别在目标领域中的 3 个数据集上构建消融实验, 表 7–表 9 分别展示了在 OntoNotes 5.0 数据集、Resume 数据集和 CCKS 2017 数据集上的消融实验结果. 本节构造以下模型来分析每个组件的有效性.

(1) TES-NER 模型: 本文提出的动态迁移实体块信息的跨领域中文实体识别模型;

(2) –Fusion: 在 TES-NER 模型的基础上忽略动态融合层, 即在目标领域的中文命名实体识别模块中的条件随机场直接建立在命名实体任务指定的长短期记忆神经网络之上;

(3) –SE: 在 TES-NER 模型的基础上忽略实体块识别任务, 具体而言, 将实体块识别任务对应的损失函数的权重α设置为 0;

(4) –Transfer: 在 TES-NER 模型基础上忽略源领域的中文命名实体识别模块, 只由共享词向量嵌入层与长短期记忆神经网络模块、实体块识别模块和目标领域的中文命名实体抽取模块这 3 个模块构成.

表 7 在 OntoNotes 5.0 数据集上的消融实验

模型	*P*	*R*	*F*1
TES-NER	75.21	**75.22**	**75.21**
–Fusion	74.92	74.28	74.6
–SE	**75.39**	74.26	74.82
–Transfer	73.74	73.74	73.74

表 8 在 Resume 数据集上的消融实验

模型	*P*	*R*	*F*1
TES-NER	**95.57**	**95.4**	**95.49**
–Fusion	94.83	94.6	94.72
–SE	95.42	94.66	95.04
–Transfer	94.13	93.5	93.81

表 9 在 CCKS 2017 数据集上的消融实验

模型	*P*	*R*	*F*1
TES-NER	90.29	91.35	**90.82**
–Fusion	89.55	**91.81**	90.67
–SE	**90.69**	90.58	90.64
–Transfer	89.47	91.29	90.37

从表 7–表 9 可以观察到: 若忽略动态融合层(–Fusion), 则模型的性能在 OntoNotes 5.0、Resume 和 CCKS 2017 数据集上分别下降 0.61%、0.77%和 0.15%的 *F*1 值, 说明引入动态融合层能够提升模型的性能. 当忽略实体块识别任务时, 模型在 3 个数据集上的性能有所下降, 可能是因为引入实体块识别任务一方面能够指导模型学习到融合到命名实体任务中的实体块信息, 另一方面也能起到正则化的作用, 防止目标领域的中文命名实体识别模型陷入过拟合. 此外, 从表 7–表 9 可以观察到: 当忽略源领域的中文命名实体识别模块时, 模型的性能下降得较为明显, 说明源领域中的语义知识能够提升目标领域中的中文命名实体识别模型的性能.

3.4 实验分析

3.4.1 不同融合策略的分析

本节分析动态融合层中不同融合策略对 TES-NER 模型性能的影响, 在 OntoNotes 5.0 数据集上的分析结果见表 10. 从表 10 可观察到, 单一融合策略的性能低于融合策略和分离策略, 说明仅仅利用实体块信息构造门向量无法充分利用不同领域的实体块信息, 需进一步引入实体块信息的上下文表征特征. 从表 9 还可以观察到, 分离融合策略在 *F*1 值上比共享融合策略提升 0.28%. 其不同之处在于: 共享融合策略是采用共享门向量融合(见式(11)), 而分离融合策略是独立的门向量(见式(12)). 其可能的原因是, 实体块信息的上下文特征和命名实体的上下文特征不是互斥的.

表 10 在 OntoNotes 5.0 数据集上不同融合策略的分析

策略	*P*	*R*	*F*1
单一融合	74.84	74.75	74.78
共享融合	74.11	75.76	74.93
分离融合	**75.21**	**75.22**	**75.21**

3.4.2 TES-NER 模型与 BERT 模型的兼容性分析

为了验证 TES-NER 模型与 BERT 模型的兼容性, 即融合 BERT 模型对 TES-NER 模型性能的影响, 本节通过将 BERT 模型输出的最后一层语义向量在字级别上拼接到第 2.1 节中的共享词向量嵌入层输出的字向量上, 且在模型训练过程中禁止 BERT 模型微调, 实验结果见表 11.

从中可以观察到, 融合 BERT 模型在 OntoNotes 5.0、Resume 和 CCKS 2017 数据集上均能进一步提升 TES-NER 模型的性能. 其中, 在 OntoNotes 5.0 数据集上的 *F*1 值提升 4.83%. 由此说明, 本文提出的 TES-NER 模型能够有效地兼容 BERT 模型, 且融合 BERT 模型在不同领域的中文命名实体数据集上的性能提升均有效.

表 11 TES-NER 模型与 BERT 模型的兼容性分析

数据集	模型	*P*	*R*	*F*1
OntoNotes 5.0	TES-NER	75.21	75.22	75.21
	TES-NER+BERT	**77.75**	**82.47**	**80.04**
Resume	TES-NER	95.57	95.4	95.49
	TES-NER+BERT	**95.72**	**96.13**	**95.93**
CCKS 2017	TES-NER	90.29	91.35	90.82
	TES-NER+BERT	**91.3**	**93.4**	**92.34**

3.4.3 混合领域中不同子领域的实验分析

本节分析本文提出的 TES-NER 模型和 BiLSTM-CRF 模型在混合领域数据集(OntoNotes 5.0 数据集)上的不同领域的性能分析, 如图 4 所示. 从中可以观察到: 在 OntoNotes 5.0 数据集的 6 个子领域中, TES-NER 模型的性能均优于 BiLSTM-CRF 模型. 一方面说明本节提出的 TES-NER 能够利用源领域的语义知识来提升目标领域中文命名实体识别的性能, 另一方面说明模型的泛化能力好, 能够适用于多种不同的子领域, 包括电话广播对话、通话领域和网络数据等领域.

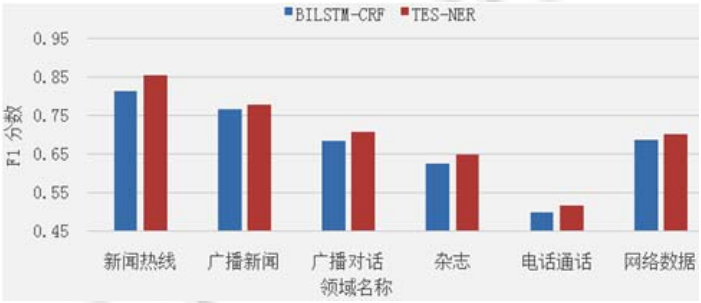


图 4 TES-NER 模型和 BiLSTM-CRF 模型在 OntoNotes 5.0 数据集中不同领域上的性能

3.4.4 TES-NER 模型的统计验证分析

为验证本文提出的 TES-NER 模型在统计上显著优于基于字的双向长短期记忆和随机条件模型(char-based BiLSTM CRF), 本节依据文献[34], 通过重复(100 次)从测试集中抽取 300 个句子计算 *F*1 分数, 然后采用文献[28]中的配对 t 检验(pairwise t-test)(*p* 值小于 0.01)来表示统计显著性. 具体流程如下.

(1) 在训练集上分别训练 Char-based BiLSTM CRF 模型和 TES-NER 模型;

(2) 将步骤(1)中训练完成的两个模型分别应用到测试集上, 随机抽取的 300 条数据计算得到 *F*1 分数;

(3) 重复步骤(2)100 次得到两组 *F*1 分数, 其中每组包含 100 个 *F*1 分数;

(4) 采用配对 t 检验计算两组 *F*1 分数的 *p* 值(*p* 值小于 0.01), 以验证本文提出的方法在统计上是有意义的. 实验结果见表 12.

表 12 在不同数据集下配对 t 检验的 *p* 值

数据集	*P* 值
OntoNotes 5.0	2.05e−13
Resume	4.65e−30
CCKS 2017	2.93e−22

注: 由于 CCKS 2017 测试集的数量不足 300 条, 采用有放回的采样, 对其余两个测试集采用无放回采样

由表 12 可知: TES-NER 模型与 Char-bsed BiLSTM CRF 模型在 OntoNotes 5.0、Resume 和 CCKS 2017 数据集上通过配对 t 检验得到的 *p* 值分别为 2.05e−13、4.65e−30 和 2.93e−22, 均小于 0.01. 由此说明, 本文提出的 TES-NER 模型在统计上明显优于 Char-based BiLSTM CRF 模型.

3.4.5 针对不同目标数据集规模的实验分析

本节分析 TES-NER 模型与 BiLSTM-CRF 模型在不同 OntoNotes 5.0 数据集大小条件下的性能, 从图 5 可以观察到, TES-NER 模型在不同数据集大小条件下的性能均高于 BiLSTM-CRF 模型, 进一步说明本文提出模型的有效性. 此外, 从图 6 可以观察到: 当目标领域的数据集大小占目标数据集的 10%–50%时, TES-NER 模型比 BiLSTM-CRF 模型提升的性能随着数据集大小的增多而下降, 可能是因为, 当目标领域的数据集数量较少时, 目标领域的中文实体识别模型本身不足以学习到命名实体的语义信息, 因此该模型会利用更多源领域的语义信息(实体块信息). 当数据集大小在 50%–100%时, TES-NER 模型比 BiLSTM-CRF 模型在 *F*1 上的提升值趋近于平缓, 且小于小数据量下的提升值, 这可能是因为, 在目标领域数据量上能够学习到较充分的命名实体语义信息, 减少了对源领域上命名实体语义信息的依赖.

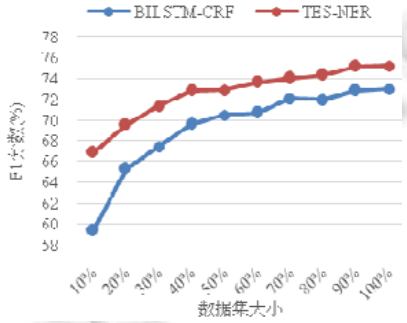


图 5 TES-NER 模型与 BiLSTM-CRF 模型在不同 OntoNotes 5.0 数据集大小条件下的性能

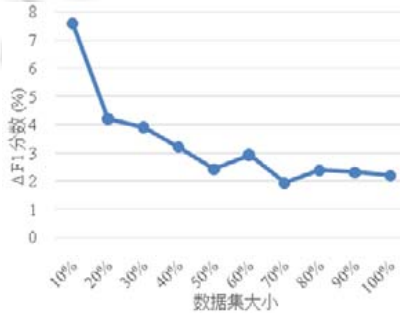


图 6 在不同 OntoNotes 5.0 数据集大小条件下, TES-NER 模型比 BiLSTM-CRF 模型高出的 *F*1 值

3.4.6 实体块信息的可视化分析

由于本文提出的动态融合层是通过动态选择实体块信息的量来提升命名实体识别模型的性能, 本节通过观察句子中每个字上门向量的权重来衡量动态融合层融合实体块信息的量. 本节首先归一化实体块相关的门向量(见式(8))到[0,1]; 之后, 通过热图的可视化方式展示, 从而能够显性观察到实体块中起关键作用的部分. 图 7 展示了一个可视化的例子, 其中, *x* 轴表示句子中的字, *y* 轴表示归一化后实体信息相关的门向量.

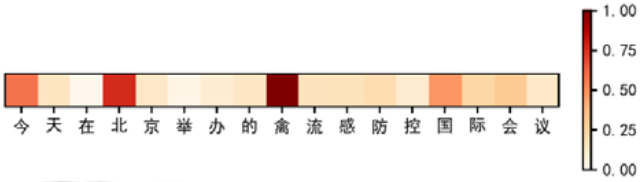


图 7 实体块信息的可视化展示

从图 7 可以发现, 命名实体模型更加关注于实体块中的起始字. 例如, “今天”是一个实体块, 其中, “今”字

的权重高于“天”，说明模型更加关注于实体块的起始字. 再例如: 在“北京”实体块中，模型更加关注于起始字“北”; 在“禽流感防控国际会议”实体块中，模型更加关注于起始字“禽”. 这一结果与先前的声明一致，即本文提出的动态融合层通过迁移不同领域的实体块信息来缓解命名实体识别模型，同时确定命名实体范围和命名实体类型的压力.

3.4.7 案例分析

表 13 展示了 TES-NER 模型与 BiLSTM-CRF 模型的两个案例. 从第 1 个案例中可以观察到: 本节提出的 TES-NER 模型正确识别“开滦矿务局”为组织实体，而 BiLSTM-CRF 模型则无法正确识别出来. 说明在源领域迁移的实体块信息的帮助下，可在目标领域中正确识别命名实体的范围. 从第 2 个案例中可以观察到: TES-NER 模型能够正确识别“突尼斯”为地理实体，而 BiLSTM-CRF 模型将“突尼斯”错误识别为人名实体. 说明本文提出的动态迁移源领域的实体信息能够缓解目标领域的中文命名实体识别模型的压力，让模型更加关注于对命名实体类型的识别.

表 13 案例分析

	Case1	Case2
原始句子	唐山大地震中开滦矿务局井下万名矿工成功脱险的奇迹	突尼斯再次呼吁国际良知采取行动
正确的命名实体	事件 组织 唐山大地震中开滦矿务局井下万名矿工成功脱险的奇迹	地理 突尼斯再次呼吁国际良知采取行动
BiLSTM-CRF 模型	事件 唐山大地震中开滦矿务局井下万名矿工成功脱险的奇迹	人名 突尼斯再次呼吁国际良知采取行动
TES-NER 模型	事件 组织 唐山大地震中开滦矿务局井下万名矿工成功脱险的奇迹	地理 突尼斯再次呼吁国际良知采取行动

4 结束语

本文针对垂直领域中有标注的命名实体语料不足的问题，提出了动态迁移实体块信息的跨领域中文实体识别模型. 该模型采用基于多任务学习的命名实体识别领域迁移来利用源领域中的语义信息，且利用实体块识别模型，以字是否为实体字作为判断依据来抽取领域之间迁移的实体块信息. 此外，本文设计了动态融合层，以动态控制从源领域迁移到目标领域的实体块信息的量. 实验结果表明，本文提出的模型在混合领域、金融领域和医学领域上相比于标准的双向长短期记忆和条件随机场模型(BiLSTM-CRF)高出 2.18%、1.68%和 0.99%，从而验证了本文提出的模型的有效性. 并且，通过可视化归一化后的实体块信息相关的门向量发现，中文命名实体识别模型更关注于实体块的起始字. 在未来的工作中，将本文提出的模型应用到其他序列标注任务，例如词性标注、中文分词等，并且扩展到更多垂直领域.

References:

[1] Behrang M. Named Entity Recognition. 2014. [doi: 10.1007/978-3-642-45358-8_7]

[2] Yang JF, Yu QB, Guan Y, Jiang ZP. An overview of research on electronic medical record oriented named entity recognition and entity relation extraction. Acta Automatica Sinica, 2014, 40(8): 1537−1562 (in Chinese with English abstract). [doi: 10.3724/SP.J.1004.2014.01537]

[3] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv: 1508.01991, 2015.

[4] Ma X, Hovy E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Vol.1: Long Papers). 2016. 1064−1074.

[5] Wu F, Liu J, Wu C, *et al*. Neural Chinese named entity recognition via CNN-LSTM-CRF and joint training with word segmentation. In: Proc. of the World Wide Web Conf. 2019. 3342−3348.

[6] Jia Y, Xu X. Chinese named entity recognition based on CNN-BiLSTM-CRF In: Proc. of the 9th IEEE Int’l Conf. on Software Engineering and Service Science (ICSESS). IEEE, 2018. 1−4.

[7] Zhong Q, Tang Y. An attention-based BiLSTM-CRF for Chinese named entity recognition. In: Proc. of the 5th IEEE Int’l Conf. on Cloud Computing and Big Data Analytics (ICCCBDA). IEEE, 2020. 550−555.

[8] Zhang HN, Wu DY, Liu Y, *et al*. Chinese named entity recognition based on deep neural network. Journal of Chinese Information Processing, 2017, 31(4): 28−35 (in Chinese with English abstract).

[9] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans. on Knowledge and Data Engineering, 2009, 22(10): 1345−1359.

[10] Bender O, Och FJ, Ney H. Maximum entropy models for named entity recognition. In: Proc. of the 7th Conf. on Natural Language Learning at HLT-NAACL 2003. 2003. 148−151.

[11] Chen W, Zhang Y, Isahara H. Chinese named entity recognition with conditional random fields. In: Proc. of the 5th SIGHAN Workshop on Chinese Language Processing. 2006. 118−121.

[12] Zhou GD, Su J. Named entity recognition using an HMM-based chunk tagger. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics. 2002. 473−480.

[13] Chen H, Lin Z, Ding G, *et al*. GRN: Gated relation network to enhance convolutional neural network for named entity recognition. In: Proc. of the AAAI Conf. on Artificial Intelligence, Vol.33. 2019. 6236−6243.

[14] Zhu Y, Wang G. CAN-NER: Convolutional attention network for Chinese named entity recognition. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol.1 (Long and Short Papers). 2019. 3384−3393.

[15] Peng N, Dredze M. Improving named entity recognition for Chinese social media with word segmentation representation learning. In: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Vol.2: Short Papers). 2016. 149−155.

[16] Cao P, Chen Y, Liu K, *et al*. Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. 2018. 182−192.

[17] Peng DL, Wang YR, Liu C, *et al*. TL-NER: A transfer learning model for Chinese named entity recognition. In: Proc. of the Information Systems Frontiers. 2019. 1−14.

[18] Yang Z, Salakhutdinov R, Cohen WW. Transfer learning for sequence tagging with hierarchical recurrent networks. arXiv preprint arXiv: 1703.06345, 2017.

[19] Lin BY, Lu W. Neural adaptation layers for cross-domain named entity recognition. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. 2018. 2012−2022.

[20] Aguilar G, Maharjan S, López-Monroy AP, *et al*. A multi-task approach for named entity recognition in social media data. In: Proc. of the 3rd Workshop on Noisy User-generated Text. 2017. 148−153.

[21] Aguilar G, López-Monroy AP, González FA, *et al*. Modeling noisiness to recognize named entities using multitask neural networks on social media. In: Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol.1 (Long Papers). 2018. 1401−1412.

[22] Zhai F, Potdar S, Xiang B, *et al*. Neural models for sequence chunking. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. 2017. 3365−3371.

[23] Xiao S, Ouyang Y, Rong W, *et al*. Similarity based auxiliary classifier for named entity recognition. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). 2019. 1140−1149.

[24] Zheng C, Cai Y, Xu J, *et al*. A boundary-aware neural model for nested named entity recognition. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). 2019. 357−366.

[25] Dong C, Wu H, Zhang J, *et al*. Multichannel LSTM-CRF for named entity recognition in Chinese social media. In: Proc. of the Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Cham: Springer, 2017. 197−208.

[26] Lee JY, Dernoncourt F, Szolovits P. Transfer learning for named-entity recognition with neural networks. In: Proc. of the 11th Int'l Conf. on Language Resources and Evaluation (LREC 2018). 2018.

[27] Weischedel R, Palmer M, Marcus M, *et al*. Ontonotes release 5.0 ldc2013t19. In: Proc. of the Linguistic Data Consortium. 2013. 23.

[28] Zhang Y, Yang J. Chinese NER using lattice LSTM. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Vol.1: Long Papers). 2018. 1554−1564.

[29] Pradhan S, Moschitti A, Xue N, *et al*. Towards robust linguistic analysis using ontonotes. In: Proc. of the 17th Conf. on Computational Natural Language Learning. 2013. 143−152.

[30] Jie Z, Lu W. Dependency-guided LSTM-CRF for named entity recognition. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). 2019. 3853−3863.

[31] Gui T, Ma R, Zhang Q, *et al*. CNN-based Chinese NER with lexicon rethinking. In: Proc. of the IJCAI. 2019. 4982−4988.

[32] Zhang Q, Li Z, Feng D, *et al*. Multitask learning for chinese named entity recognition. In: Proc. of the Pacific Rim Conf. on Multimedia. Cham: Springer, 2018. 653−662.

[33] Zhao S, Cai Z, Chen H, *et al*. Adversarial training based lattice LSTM for Chinese clinical named entity recognition. Journal of Biomedical Informatics, 2019, 99: 103290.

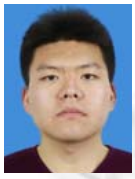
[34] Koehn P. Statistical significance tests for machine translation evaluation. In: Proc. of the 2004 Conf. on Empirical Methods in Natural Language Processing. 2004. 388−395.

附中文参考文献:

[2] 杨锦锋, 于秋滨, 关毅, 蒋志鹏. 电子病历命名实体识别和实体关系抽取研究综述. 自动化学报, 2014, 40(8): 1537−1562. [doi: 10.3724/SP.J.1004.2014.01537]

[8] 张海楠, 伍大勇, 刘悦, 等. 基于深度神经网络的中文命名实体识别. 中文信息学报, 2017, 31(4): 28−35.

吴炳潮(1995－), 男, 学士, 主要研究领域为人工智能, 自然语言处理, 信息抽取.



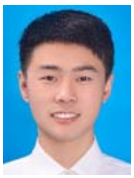
邓成龙(1993－), 男, 硕士, 主要研究领域为人工智能, 计算机辅助诊断.

关贝(1986－), 男, 博士, 高级工程师, 主要研究领域为人工智能, 大数据, 网络安全技术, 虚拟化技术, 操作系统技术, 云计算.

陈晓霖(1996－), 男, 博士生, 主要研究领域为人工智能, 联邦学习, 隐私保护.

訾道广(1997－), 男, 学士, 主要研究领域为自然语言处理.

常志军(1981－), 男, 副研究馆员, 主要研究领域为大数据平台建设, 知识图谱, 文本实体关系识别.



肖尊严(1999－), 男, 学士, 主要研究领域为人工智能, 推荐系统.

曲大成(1974－), 男, 副研究员, 主要研究领域为社交网络, 推荐系统, 生物信息学.

王永吉(1962－), 男, 博士, 研究员, 博士生导师, 主要研究领域为人工智能, 大数据分析, 智能制造, 云计算, 知识工程, 虚拟化技术, 隐蔽信道, 高可信网络技术, 系统仿真, 实时系统, 传感器网络, 数据挖掘, 软件工程, 图像处理.