软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(12):5649−5669 [doi: 10.13328/j.cnki.jos.006750]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于语义先验知识与类型嵌入的复杂实体识别*

姜小波, 何 昆, 阎广瑜

(华南理工大学 电子与信息学院, 广东 广州 510641)

通信作者: 何昆, E-mail: hk15616172426@163.com

摘 要: 实体识别是信息抽取的关键任务. 随着信息抽取技术的发展, 研究人员从简单实体的识别转向复杂实体的识别. 然而, 复杂实体缺乏明显的特征且在句法结构与词性组成上更加复杂多样, 给实体识别带来了巨大挑战. 此外, 现有模型广泛采用基于跨度的方法来识别嵌套实体, 在实体边界检测方面呈现出模糊化, 影响识别的性能. 针对这些问题和挑战, 提出了一种基于语义先验知识与类型嵌入的实体识别模型 GIA-2DPE. 该模型使用实体类别的关键词序列作为语义先验知识来提升对实体的认知, 并通过类型嵌入捕获不同实体类型的潜在特征, 然后通过门控交互注意力机制将先验知识与类型特征相融合以辅助复杂实体识别. 另外, 模型通过 2D 概率编码来预测实体边界, 并利用边界特征和上下文特征来增强对边界的精准检测, 从而提升嵌套实体的识别效果. 在 7 个英文数据集和 2 个中文数据集上进行了广泛实验. 结果表明, GIA-2DPE 超越了目前最先进的模型; 并且在 ScienceIE 数据集的实体识别任务中, 相对基线 *F*1 分数取得了最高 10.4% 的提升.

关键词: 信息抽取; 复杂实体识别; 门控交互注意力机制; 2D 概率编码

中图法分类号: TP18

中文引用格式: 姜小波, 何昆, 阎广瑜. 基于语义先验知识与类型嵌入的复杂实体识别. 软件学报, 2023, 34(12): 5649–5669. http://www.jos.org.cn/1000-9825/6750.htm

英文引用格式: Jiang XB, He K, Yan GY. Complex Entity Recognition Based on Prior Semantic Knowledge and Type Embedding. Ruan Jian Xue Bao/Journal of Software, 2023, 34(12): 5649–5669 (in Chinese). http://www.jos.org.cn/1000-9825/6750.htm

Complex Entity Recognition Based on Prior Semantic Knowledge and Type Embedding

JIANG Xiao-Bo, HE Kun, YAN Guang-Yu

(School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China)

Abstract: Entity recognition is a key task of information extraction. With the development of information extraction technology, researchers turn the research direction from the recognition of simple entities to the recognition of complex ones. Complex entities usually have no explicit features, and they are more complicated in syntactic constructions and parts of speech, which makes the recognition of complex entities a great challenge. In addition, existing models widely use span-based methods to identify nested entities. As a result, they always have an ambiguity in the detection of entity boundaries, which affects recognition performance. In response to the above challenge and problem, this study proposes an entity recognition model GIA-2DPE based on prior semantic knowledge and type embedding. The model uses keyword sequences of entity categories as prior semantic knowledge to improve the cognition of entities, utilizes type embedding to capture potential features of different entity types, and then combines prior knowledge with entity-type features through the gated interactive attention mechanism to assist in the recognition of complex entities. Moreover, the model uses 2D probability encoding to predict entity boundaries and combines boundary features and contextual features to enhance accurate boundary detection, thereby improving the performance of nested entity recognition. This study conducts extensive experiments on seven English datasets and two Chinese datasets. The results show that GIA-2DPE outperforms state-of-the-art models and achieves a 10.4% *F*1 boost compared with the baseline in entity recognition tasks on the ScienceIE dataset.

Key words: information extraction; complex entity recognition; gated interactive attention; 2D probability encoding

---

* 基金项目: 国家自然科学基金 (U1801262); 广东省科技计划 (2019B010154003)
收稿时间: 2021-12-02; 修改时间: 2022-02-25, 2022-06-14; 采用时间: 2022-07-23; jos 在线出版时间: 2023-02-15
CNKI 网络首发时间: 2023-02-16

© 中国科学院软件研究所 http://www.jos.org.cn

实体识别是信息抽取的关键任务, 其目的是从文本中识别出特定类型的实体并将它们正确分类. 目前, 研究人员已经在简单实体的识别任务中取得了较大成功. 例如, Eberts 等人[1]对医疗报告中的药物与副作用类型的实体进行了识别, 取得了 89.3% 的 $F1$ 分数; Friedrich 等人[2]识别了材料科学文献中的材料、器件等类型的实体, $F1$ 分数达到了 81.5%; Li 等人[3]则以 84.8% 的 $F1$ 分数识别了新闻文本中的人物、机构等类型的实体. 然而, 这些简单实体通常仅由几个名词组成, 包含的信息太少, 如图 1(a) 所示.

随着自动信息抽取的不断发展, 研究人员不仅需要识别简单实体, 还需要识别信息量更大的复杂实体. 例如任务、方法等类型的实体, 如图 1(b) 所示. 复杂实体通常由短语组成, 在语义、句法结构和词性组成上都更加复杂, 给相关模型带来了巨大挑战. 例如, Sahrawat 等人[4]使用先进的预训练模型对材料学和计算机科学等领域文献中的任务、处理方式等类型的实体进行了识别, 得到的 $F1$ 分数为 52.2%. Luan 等人[5]则使用精心设计的模型 SCIIE 对 AI 领域文献中的任务、方法等类型的实体进行了识别, 结果为 64.2%.

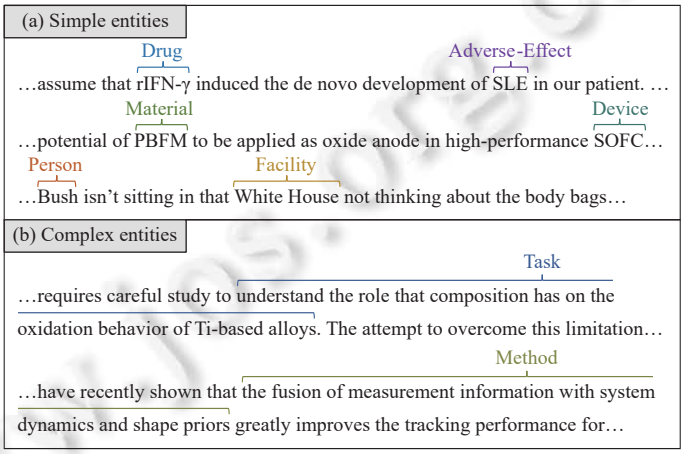

图 1 简单实体与复杂实体样例[2,5–8]

与简单实体相比, 复杂实体的识别主要存在两方面挑战. 一方面, 简单实体普遍具备一些明显的特征, 有助于简单实体的准确识别. 例如, 人物类实体由首字母大写的单词组成且一般充当句子主语; 药物类实体则经常包含一些特殊字符. 然而, 复杂实体通常缺乏明显的特征, 采用表层显著特征的方法对识别复杂实体效果较差. 本文通过一种深层语义理解的方法来识别复杂实体. 具体地, 利用实体类别 (如任务、方法等) 的语义先验知识, 结合上下文信息, 来提高对实体的语义认知, 从而提升复杂实体识别性能. 另一方面, 简单实体通常仅由几个名词构成, 其句法结构简单且词性组成单一, 如图 1(a) 所示. 而复杂实体在句法结构和词性组成上具有多样性, 进一步加剧了识别难度. 例如图 1(b) 中的任务类型实体样例包含了多种词性的词语, 且句法结构为定语从句. 本文发现, 捕获并利用不同实体类型蕴含的潜在特征来辅助识别, 可以提升复杂实体识别的性能.

此外, 复杂实体和简单实体一样可能在内部包含了嵌套实体, 如图 2 所示. 为了避免信息丢失, 现有模型广泛采用基于跨度的方法来识别嵌套实体: 首先枚举出所有的跨度, 使嵌套实体从外围实体中分离出来, 例如“NF-chi B site”的跨度为“NF-”“chi”“B”“site”“NF-chi”“chi B”“B site”“NF-chi B”和“chi B site”; 然后对所有跨度进行分类来判断哪些是实体, 并在相关数据集[7,9,10]上展现了先进性能. 不过, 这些模型专注于学习跨度自身的表征, 通常难以分辨一些具有细微差别的跨度[11]. 例如, 两个跨度“chi B”和“chi B site”只有微小的边界差异, 两者的表征具有相似性, 但前者是嵌套实体而后者不是, 相关模型凭借表征来分辨这两个跨度容易造成混淆 (即边界模糊化); 但如果根据结束边界是“B”还是“site”来进行分辨则不容易混淆. 本文采用实体边界的精准检测方法替代基于跨度的方法, 以避免这种模糊化, 从而提高识别性能.

针对上述挑战和问题, 本文提出了一种端到端的实体识别模型, 称为 GIA-2DPE (gated interactive attention and 2D probability encoder). 首先, 为了增强模型对实体的语义认知, 我们为每个实体类别设计了一段关键词序列作为

© 中国科学院软件研究所 http://www.jos.org.cn

语义先验知识, 然后将语义先验知识与原文本进行拼接作为模型的输入. 例如, AI 领域的“任务”实体类别对应的关键词序列为{task, processing, image, speech, video, information, translation, classification, recognition}. 其次, 我们在模型中添加了一个可训练的专用嵌入矩阵 ETE (entity type embedding), 以支持模型自适应地学习不同实体类别对应的潜在特征向量.

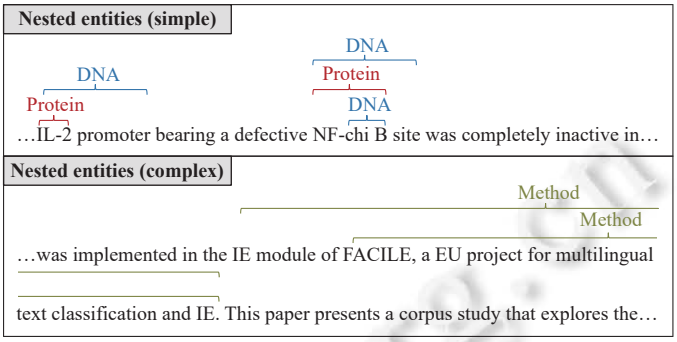

图 2 嵌套实体样例[5,10]

为了综合利用上述先验知识与类别特征来辅助复杂实体的识别, GIA-2DPE 模型使用了提出的门控交互注意力 (GIA) 机制. 该方法通过注意力机制将先验知识表示和类别特征向量分别与上下文表示进行交互, 并将交互结果通过一个“融合门”函数进行结合, 来获取上下文中各词语与识别内容的相关度, 从而缓解了复杂实体因缺乏明显的特征而难以被识别的问题.

此外, 为了增强对实体边界的精准检测, GIA-2DPE 模型使用提出的一种 2D 概率编码 (2DPE) 机制来识别嵌套实体. 该方法将跨度分类任务转化为实体边界的检测任务, 使模型能够预测出作为实体边界的词语, 从而实现对实体边界的监督. 同时, 该方法通过捕获实体的边界特征以及全局上下文特征来进一步辅助模型对实体边界的精准检测, 从而提升对嵌套实体的识别性能.

我们在 7 个英文数据集和 2 个中文数据集上进行了广泛的实体识别实验. 结果显示, GIA-2DPE 模型在性能上超越了目前最先进的模型; 并且与基线 $F1$ 分数相比, 取得了最高 10.4% 的大幅度提升.

本文的主要贡献如下.

(1) 针对复杂实体识别困难的问题, 提出利用语义先验知识与类别嵌入来辅助识别, 结合提出的 GIA 机制, 提升了复杂实体的识别性能.

(2) 针对嵌套实体识别中的实体边界模糊化问题, 提出了 2DPE 机制来增强对实体边界的精准检测, 提升了嵌套实体的识别性能.

(3) 在上述工作的基础上构建了实体识别模型 GIA-2DPE, 在 9 个相关数据集上取得了最先进的性能.

1 相关工作

1.1 简单实体的识别

随着信息抽取技术的不断发展, 研究人员提出了各种基于神经网络的实体识别模型, 并且成功地应用于生物医学、材料科学以及新闻等领域[12].

这些研究识别的实体大部分是一些简单实体, 它们通常仅由几个名词组成. 例如, Friedrich 等人[2]使用预训练模型 SciBERT[13]与双向长短期记忆网络 (BiLSTM), 通过序列标注方式对材料学文献中的材料、器件等类型的实体 (如 PBFM、SOFC) 进行了识别, 取得了 81.5% 的 $F1$ 分数. Eberts 等人[1]利用预训练模型 BERT[14]与最大池化来对输入序列中的所有跨度进行分类, 实现了医疗报告中的药物与症状类实体 (如 rIFN-γ、SLE) 的识别, 结果达到 89.3%. Li 等人[3]则使用 BERT 模型与前馈神经网络 (FFNN), 通过回答不同的问题来识别不同类型的实体, 以

© 中国科学院软件研究所 http://www.jos.org.cn

84.8% 的 $F1$ 分数从新闻文本中识别了人物、机构等类型的实体 (如 Bush、White House).

1.2 复杂实体的识别

简单实体包含的信息太少, 无法完全满足信息抽取的需求. 尤其在科学技术领域, 研究人员还需要信息量更大的复杂实体, 例如任务、方法等类型的实体, 它们反映了领域的发展和研究现状.

然而, 这些复杂实体通常具有短语结构, 不仅更长且在语义、句法结构和词性组成上更加复杂. 例如, 图 1(b) 所示的任务类实体不仅包含动词、名词和形容词等多种词性的词语, 而且具有定语从句结构, 给现有实体识别模型带来了极大的挑战. 例如, Sahrawat 等人[4]使用高性能的 BERT 模型与 BiLSTM, 通过序列标注方式对材料学、物理学以及计算机科学等领域文本中的任务、处理方式等类型的实体进行了识别, $F1$ 分数为 52.2%. Lai 等人[15]使用 SciBERT 代替 BERT, 进行了与 Sahrawat 等人相同的工作, 结果为 54.6%. Jain 等人[16]在 BERT 与 BiLSTM 的基础上, 进一步结合了条件随机场 (CRF), 并通过序列标注方式对 AI 领域文献中的任务、方法等类型的实体进行了识别, $F1$ 分数为 63.8%. Luan 等人[5]则使用精心设计的 SCIIE 模型, 通过对输入序列中的所有跨度进行分类, 实现了对 AI 领域文献中的任务、方法等类型实体的识别, 结果为 64.2%.

与上述模型相比, 本文模型使用提出的一种门控交互注意力机制来提升复杂实体的识别性能. 该机制利用额外的实体类别的语义先验知识来辅助识别, 以增强模型对实体的语义认知, 从而明确上下文中哪些部分与识别内容相关, 一定程度上弥补了复杂实体特征不明显的缺陷. 同时, 该机制还利用了不同实体类别的潜在特征, 有利于模型识别出具有复杂结构的实体.

1.3 嵌套实体的识别

无论在简单实体还是复杂实体中都有可能存在嵌套实体 (即实体内部包含的实体), 例如图 2 中的“IL-2”“NF-chi B”“chi B”和“FACILE, a EU project for multilingual text classification and IE”. 为了避免信息丢失, 这些嵌套实体也需要被准确、无遗漏地识别出来.

目前, 研究人员针对嵌套实体的识别提出了许多解决方案. 一部分工作致力于使用复杂的转化机制将嵌套结构转化为扁平结构. 例如 Shibuya 等人[17]提出的次佳路径解码机制和 Huang 等人[18]设计的超图机制等. 但 Li 等人[19]指出复杂的转化步骤会带来额外的错误或偏差. 另一部分工作采用一种更直接有效的方法, 其通过对文本中的所有跨度 (即子序列) 进行分类来判别嵌套实体. 例如, Eberts 等人[1]提出了 SpERT, 它由 BERT 模型和一个对跨度进行分类的前馈层组成. Wang 等人[20]提出了一种结合 BERT 与卷积神经网络 (CNN) 的模型 SPE, 通过融合局部信息来获得更好的跨度表征. Shen 等人[21]则使用 BERT 与 BiLSTM 来获取包含丰富信息的跨度表征, 并通过 SoftNMS 算法来增强了对跨度的判别. 不过, 基于跨度的模型侧重于学习跨度自身的表征, 通常难以分辨一些具有细微边界差别的跨度[11].

与上述模型相比, 本文模型提出了一种 2D 概率编码方法来识别嵌套实体. 该方法将跨度分类任务转化为实体边界的检测任务, 使模型预测那些作为实体边界的词语, 并且通过捕获实体的边界特征和全局上下文特征来进一步辅助边界的精准检测, 从而提升了嵌套实体识别的效果.

2 GIA-2DPE 模型

2.1 模型整体框架

定义 1. 给定输入文本 T 和实体类型集合 $C=\{c_i \mid i=1, 2, \ldots, n\}$($n$ 为实体类别数量). 实体识别旨在从 T 中识别出全部的实体并将其正确划分到 C 中的某个类别, 得到一个实体集合 $E=\{(e_i, c_j) \mid e_i \in T, c_j \in C\}$.

由定义 1 可知, 实体识别有两种实现方式. 方式 1 先将所有实体抽取出来, 形成实体集 $E=\{e_i \mid e_i \in T\}$, 再对 E 中每个实体进行分类, 得到最终结果 E. 而方式 2 依次以 $c_1, c_2, \ldots, c_n$ 为目标类型, 得到 n 个满足条件的实体集 $E_1, E_2, \ldots, E_n$. E_i 的所有实体都被划分为 c_i 类型. 因此, 方式 2 同样可以得到最终结果 E. 显然, 方式 2 更具针对性, 且分类误差小于方式 1, 但需要额外信息以使模型明确每一次识别的目标类型是什么.

© 中国科学院软件研究所 http://www.jos.org.cn

本文采用方式 2 进行实体识别. 为了增强模型对目标类型实体的认知, 我们为每个实体类别 c_i 设计了一段关键词序列 $K_i=\{k_j \mid j=1, 2, \ldots, m\}$($m$ 为关键词个数) 作为语义先验知识, 得到先验知识集合 $I=\{K_i \mid i=1, 2, \ldots, n\}$. 例如, AI 领域的“任务”实体类型的关键词序列 $K=$ {task, processing, image, speech, video, information, translation, classification, recognition}. 先验知识将被用于后续与上下文进行交互.

在自然语言处理中, 通常先对输入文本进行分词, 得到输入序列 $X=\{w_i \mid i=1, 2, \ldots, L\}$($L$ 为序列长度). 为了对实体边界进行检测, 我们使用实体在 X 中的起始与结束边界构成的坐标来表示该实体, 即 $e_i \Leftrightarrow (p_{s,i}, p_{e,i})$. 例如, 在序列 $X=$ {It, kills, B, cells, in, the, blood}中, 细胞类型实体“B cells”表示为坐标 (3, 4).

在提出的 GIA-2DPE 模型中, 输入包括输入序列 X、关键词序列 K 以及实体类别标签 c. 输出包括 p_s、p_e 与 m_{2D}. 其中, p_s 和 p_e 为长度等于 L 的向量, p_s 的第 i 个元素代表的是实体起始边界等于 i 的概率, p_e 的第 j 个元素代表的是实体结束边界等于 j 的概率; 而 m_{2D} 是一个 $L\times L$ 的矩阵, 其第 i 行第 j 列元素代表的是坐标 (i, j) 为目标实体的概率.

GIA-2DPE 模型的整体框架如图 3 所示. 模型包含 4 个主要模块: 嵌入模块、门控交互注意力模块、2D 概率编码模块以及过滤模块.

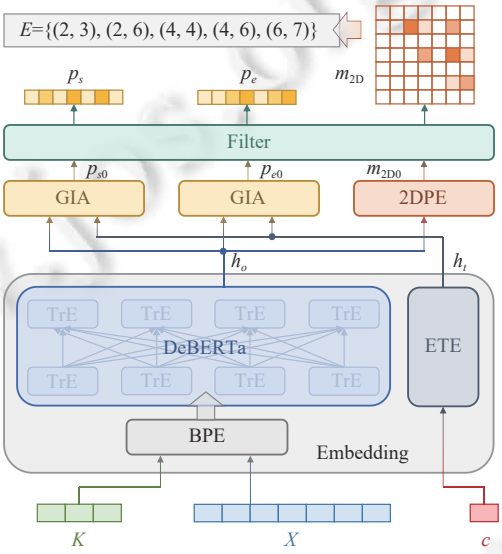

图 3 GIA-2DPE 模型整体框架

(1) 嵌入模块 (embedding): 对 K 与 X 进行拼接和词嵌入, 得到上下文表示 h_o; 并对实体类别 c 进行类别嵌入, 得到 c 类别实体对应的结构特征向量 h_t.

(2) 门控交互注意力模块 (GIA): 利用 h_o 与 h_t 进行门控交互注意力计算, 并将计算结果通过概率化函数映射成向量, 得到实体起始与结束边界的初始概率分布向量 p_{s0} 和 p_{e0}.

(3) 2D 概率编码模块 (2DPE): 利用 h_o 以及实体的边界特征 (包括起始和结束边界) 进行 2D 概率编码, 并将结果通过概率化函数映射成矩阵, 得到实体边界的初始 2D 概率分布矩阵 m_{2D0}.

(4) 过滤模块 (filter): 对 p_{s0}、p_{e0} 和 m_{2D0} 进行过滤和掩膜, 得到最终的输出 p_s、p_e 与 m_{2D}.

2.2 嵌入模块

嵌入模块包括两部分: 使用预训练模型将单词编码成向量 (即词嵌入); 以及自适应地学习不同实体类别 c 对应的潜在特征向量 (即实体类别嵌入).

预训练模型的输入序列 X_{in} 为关键词序列 K 与序列 X 的拼接:

$$X_{\text{in}}=\{[\text{CLS}], k_1, k_2, \ldots, k_M, [\text{SEP}], w_1, w_2, \ldots, w_L, [\text{SEP}]\} \tag{1}$$

© 中国科学院软件研究所 http://www.jos.org.cn

其中, M 与 L 分别表示关键词序列 K 与序列 X 的长度, 特殊符号“[CLS]”用来表示整个序列 X_{in} 的语义信息, 而特殊符号“[SEP]”用来分隔不同的序列.

在词嵌入之前, 我们使用 BPE 编码[22]算法来对输入序列 X_{in} 进行更加细粒度的分词. BPE 编码实质上是一种基于连续字节对频率统计的 SubWord 算法, 其将生僻词分解为常见的子词, 例如“hypergraph”分解为“hyper”和“graph”, 有效缓解了词嵌入的 OOV (out of vocabulary) 问题.

然后, 我们使用预训练模型 DeBERTa[23]对分词后的序列进行词嵌入, 得到上下文表示 h_o:

$$h_o = \text{DeBERTa}(\text{BPE}(X_{\text{in}})) \in \mathbb{R}^{l\times d} \tag{2}$$

其中, l 为输入序列经过 BPE 分词后的长度, d 为词嵌入的维度. 假设关键词序列 K 和序列 X 经过 BPE 分词后的长度分别为 l_k 和 l_x, 则 $l=l_k+l_x+3$.

本文选择 DeBERTa 而不使用其他常用的预训练模型 (例如 BERT 和 SciBERT 等), 有如下两个主要原因: 第一, DeBERTa 采用内容和位置信息相互分离的自注意机制, 其对于两个词语的注意力权重不仅取决于它们的内容, 而且取决于它们的相对位置, 例如单词“deep”和“learning”相邻出现时, 它们之间的依赖性要比相距较远时强得多. 这种改进有利于普遍较长的复杂实体的识别. 第二, DeBERTa 的输出层采用了一种增强型的掩码解码机制 (EMD), 一定程度上缓解了预训练和微调之间的不匹配.

除了文本语义的词嵌入, 我们还考虑了实体在不同类别上的不同特征, 构建了一个 $n\times d$ 的专用嵌入矩阵 ETE. 该嵌入矩阵是通过反向传播算法进行学习, 其将输入的目标实体类别 c 映射成一个 d 维向量 h_t, 用于表示 c 类型的实体普遍具有的类别上的潜在特征, 即实体类别嵌入:

$$h_t = \text{ETE}(c) \in \mathbb{R}^d \tag{3}$$

对于中文文本, 我们直接以字为单位进行分词, 并使用中文 BERT 进行词嵌入, 其余步骤同上.

2.3 门控交互注意力模块

门控交互注意力模块 (GIA) 旨在利用实体类别的语义先验知识以及实体的类别特征来进行门控交互注意力计算, 使模型增强对实体的认知, 以明确上下文中哪些部分与识别内容相关, 有利于缓解复杂实体难以识别的问题. 该模块的输入为 h_o 和 h_t, 输出为实体边界的初始概率分布向量 p_0, 如图 4 所示.

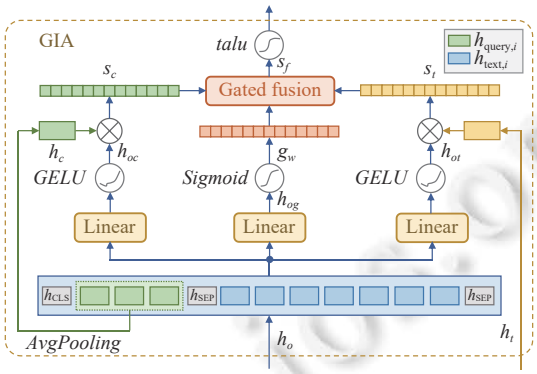

图 4 门控交互注意力模块

首先, 我们对 h_o 中属于关键词序列 K 的所有词向量进行平均池化, 得到语义先验知识表示 h_c:

$$h_c = AvgPooling(\{h_o[i] | i = 2,3,\ldots,l_k+1\}) \in \mathbb{R}^d \tag{4}$$

接着, 通过 2 个不同的线性映射以及高斯误差线性单元 ($GELU$)[24]计算, 将 h_o 映射成 h_{oc} 和 h_{ot}:

$$GELU(x) = 0.5x\left(1+\frac{2}{\sqrt{\pi}}\int_0^{\frac{x}{\sqrt{2}}} \mathrm{e}^{-t^2}\,\mathrm{d}t\right) \tag{5}$$

$$h_{oc} = GELU(h_o \cdot W_c + b_c) \in \mathbb{R}^{l\times d} \tag{6}$$

© 中国科学院软件研究所 http://www.jos.org.cn

$$h_{ot} = GELU(h_o \cdot W_t + b_t) \in \mathbb{R}^{l \times d} \tag{7}$$

再将 h_{oc} 与 h_c 进行矩阵-向量乘法计算, 得到交互注意力分数 s_c, 如公式 (8) 所示. 向量 s_c 中的各元素代表了序列中各词语在语义特征空间中与目标实体的相关度分数. 同样地, 可以得到 h_{ot} 与 h_t 的交互注意力分数 s_t, 如公式 (9) 所示. 向量 s_t 反映了序列中各词语在结构特征空间中与目标实体类型的相关度分数.

$$s_c = h_{oc} \cdot (h_c)^{\mathrm{T}} \in \mathbb{R}^l \tag{8}$$

$$s_t = h_{ot} \cdot (h_t)^{\mathrm{T}} \in \mathbb{R}^l \tag{9}$$

然后, 通过线性变换将 h_o 映射成长为 l 的向量 h_{og}, 并通过 *Sigmoid* 函数将 h_{og} 中的各元素映射到 (0, 1) 范围内来表示融合时的权重, 得到用于门控融合的权重向量 g_w:

$$h_{og} = h_o \cdot v_g^{\mathrm{T}} + b_g \in \mathbb{R}^l \tag{10}$$

$$g_w = \frac{1}{1 + \mathrm{e}^{-h_{og}}} \in \mathbb{R}^l \tag{11}$$

利用 g_w 对 s_c 与 s_t 进行门控融合, 得到融合后的交互注意力分数 s_f, 如公式 (12) 所示. 其中, 运算符号"⊙"表示向量的元素相乘. 向量 s_f 从语义和结构两个角度综合反映了序列中各词语与目标实体的相关度, 给模型提供了可能成为实体边界的词语, 有利于特征不明显的复杂实体的识别.

$$s_f = g_w \odot s_c + (1 - g_w) \odot s_t \in \mathbb{R}^l \tag{12}$$

最后, 通过提出的概率化函数 *talu* 将 s_f 中各元素映射到 (0, 1) 范围内, 目的是将相关度分数转化为概率分布, 使得相关度越高的词语越有可能成为实体的边界词. 转化后的结果即为实体边界的初始概率分布向量 p_0:

$$p_0 = \frac{\mathrm{e}^{s_f}}{\mathrm{e}^{s_f} + \mathrm{e}^{-s_f}} \in \mathbb{R}^l \tag{13}$$

其中, *talu* 函数在 $x = 0$ 处的导数是广泛使用的 *Sigmoid* 函数的 2 倍, 有利于模型区分相关度分数相近的词语.

我们构造了两个独立的 GIA 模块来分别执行上述步骤, 得到了实体的起始边界与结束边界对应的两个初始概率分布向量 p_{s0} 和 p_{e0}. 其中, 向量 p_{s0} 的第 i 个元素代表的是实体起始边界等于 i 的概率, 而向量 p_{e0} 的第 j 个元素代表的是实体结束边界等于 j 的概率.

2.4 2D 概率编码模块

2D 概率编码模块 (2DPE) 旨在使用一维卷积运算和自注意力机制来捕获实体的边界特征 (包括起始和结束边界) 以及全局上下文特征, 并将这些特征映射成一个 2D 概率分布矩阵来对实体的边界进行精准检测. 该模块的输入为 h_o, 输出为实体边界坐标的初始 2D 概率分布矩阵, 如图 5 所示.

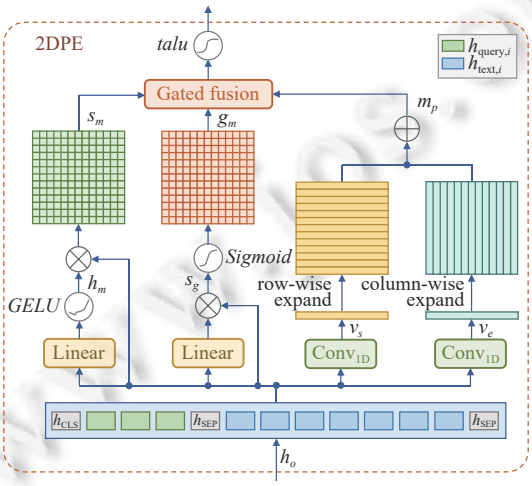

图 5 2D 概率编码模块

© 中国科学院软件研究所 http://www.jos.org.cn

首先, 我们使用大小为 $1\times d$ 的两个不同的卷积核来对 h_o 进行一维卷积操作 (步长为 1), 目的是分别捕获实体起始边界的特征向量 v_s, 以及实体结束边界的特征向量 v_e:

$$v_s = 1D\text{-}Conv_s(h_o) \in \mathbb{R}^l \tag{14}$$

$$v_e = 1D\text{-}Conv_e(h_o) \in \mathbb{R}^l \tag{15}$$

v_s 与 v_e 分别经过逐行扩展和逐列扩展, 得到两个 $l\times l$ 的矩阵. 将它们相加后得到矩阵 m_p:

$$m_p = expand_{\text{row-wise}}(v_s) + expand_{\text{column-wise}}(v_e) \in \mathbb{R}^{l\times l} \tag{16}$$

接着, 通过不同的线性映射以及 *GELU* 计算, 将 h_o 映射成 h_m 和 h_g:

$$h_m = GELU(h_o \cdot W_m + b_m) \in \mathbb{R}^{l\times d} \tag{17}$$

$$h_g = h_o \cdot W_g + b_g \in \mathbb{R}^{l\times d} \tag{18}$$

h_m 和 h_g 分别与 h_o 进行矩阵乘法计算, 得到自注意力分数矩阵 s_m 以及矩阵 s_g:

$$s_m = h_m \cdot (h_o)^{\mathrm{T}} \in \mathbb{R}^{l\times l} \tag{19}$$

$$s_g = h_g \cdot (h_g)^{\mathrm{T}} \in \mathbb{R}^{l\times l} \tag{20}$$

然后, 使用 *Sigmoid* 函数将 s_g 中的元素映射到 (0, 1) 范围内来表示融合时的权重, 得到权重矩阵 g_m; 并利用矩阵 g_m 对 s_m 与 m_p 进行门控融合, 得到融合后的矩阵 m_f:

$$g_m = \frac{1}{1+\mathrm{e}^{-s_g}} \in \mathbb{R}^{l\times l} \tag{21}$$

$$m_f = g_m \odot s_m + (1-g_m) \odot m_p \in \mathbb{R}^{l\times l} \tag{22}$$

最后, 通过 *talu* 函数将矩阵 m_f 概率化, 得到实体边界坐标的初始 2D 概率分布矩阵 m_{2D0}:

$$m_{2D0} = \frac{\mathrm{e}^{m_f}}{\mathrm{e}^{m_f}+\mathrm{e}^{-m_f}} \in \mathbb{R}^{l\times l} \tag{23}$$

矩阵 m_{2D0} 的第 i 行第 j 列元素代表的是坐标 (i, j) 为目标实体的概率, 并且该概率值的计算结合了边界特征和全局上下文特征. 由于横坐标 i 和纵坐标 j 分别代表实体的起始与结束边界. 因此, 通过该矩阵, 我们可以区分任何具有微小边界差异的候选实体, 从而增强对实体边界的精准检测, 有利于嵌套实体的识别.

2.5 过滤模块

过滤模块旨在对初始概率分布 p_{s0}、p_{e0} 与 m_{2D0} 中的元素进行过滤和掩膜 (mask), 使模型不用考虑一些不可能的实体起始或结束边界情况, 如图 6 所示.

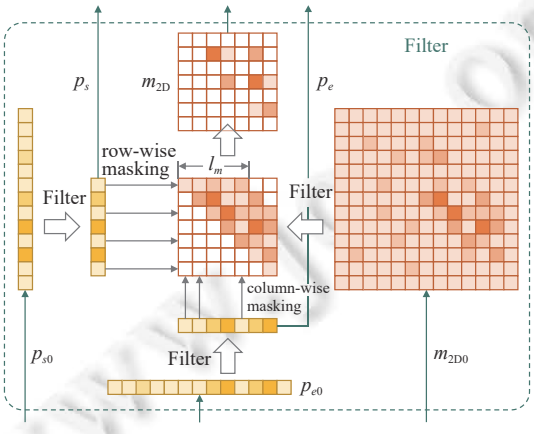

图 6 过滤模块

首先, 由于特殊符号“[CLS]”“[SEP]”以及关键词序列 K 中不包含实体, 我们将 p_{s0}、p_{e0} 与 m_{2D0} 中的相关部分删除, 得到第 1 次过滤后的结果 p_{s1}、p_{e1} 与 m_{2D1}:

© 中国科学院软件研究所 http://www.jos.org.cn

$$p_{s1} = \{p_{s0}[i] | l - l_x \leqslant i \leqslant l - 1\} \in \mathbb{R}^{l_x} \tag{24}$$

$$p_{e1} = \{p_{e0}[i] | l - l_x \leqslant i \leqslant l - 1\} \in \mathbb{R}^{l_x} \tag{25}$$

$$m_{2D1} = \{m_{2D0}[i][j] | l - l_x \leqslant i, j \leqslant l - 1\} \in \mathbb{R}^{l_x \times l_x} \tag{26}$$

接着, 我们对矩阵 m_{2D1} 进行掩膜. 设定一个概率阈值 P_t, 假设 m_{ij} 为矩阵 m_{2D1} 中的任一元素, 该元素位于 m_{2D1} 的第 i 行第 j 列, 若概率值 $p_{s1}[i]$ 和 $p_{e1}[j]$ 均大于阈值 P_t, 则元素 m_{ij} 保持不变, 否则 m_{ij} 被置零. 另外, 考虑到实体的起始边界不可能大于结束边界, 并且实体的长度总在一定的范围之内, 我们设置了一个实体长度的最大值 l_m, 并将不满足条件"$0 \leqslant j - i \leqslant l_m$"的元素 m_{ij} 置零:

$$m_{ij} = \begin{cases} m_{2D1}[i][j], & p_{s1}[i], p_{e1}[j] > P_t, 0 \leqslant j - i \leqslant l_m \\ 0, & \text{otherwise} \end{cases} \tag{27}$$

最后, 对于英文文本, 考虑到我们采用了 BPE 分词, 一些生僻词会被切分成多个片段, 例如"hypergraph"会被切分成"hyper"和"graph"两个词片段. 对于每一个被切分的词 w_i, 我们记录其词片段在序列中的索引区间 $r_i = [a_i, b_i]$, 得到切分区间集 $R = \{r_i \mid i = 1, 2, \ldots, z\}$($z$ 为被切分的词的个数), 并通过算法 1 对 p_{s1}、p_{e1} 与 m_{2D1} 进一步过滤, 得到最终输出结果 p_s、p_e 与 m_{2D}. 其中, p_s、p_e 为 L 维概率分布向量, m_{2D} 为 $L \times L$ 概率分布矩阵.

算法 1. 增强过滤.

输入: 切分区间集 R, 概率分布 p_{s1}、p_{e1} 与 m_{2D1};
输出: 过滤后的概率分布 p_{s1}、p_{e1} 与 m_{2D1}.

```
for i = 1 to z do
    第 i 个被切分词的索引区间 r_i = R[i] = [a_i, b_i];
    词片段对应概率中的最大值 p_m = max(m_2D1[a_i:b_i, a_i:b_i]);
    p_m 的横、纵坐标 x, y = m_2D1.index(p_m);
    for j = a_i to b_i do
        if j ≠ x then
            删除矩阵 m_2D1 的第 j 行;
            删除向量 p_s1 的第 j 个元素;
        end if
        if j ≠ y then
            删除矩阵 m_2D1 的第 j 列;
            删除向量 p_e1 的第 j 个元素;
        end if
    end for
end for
return p_s1, p_e1, m_2D1;
```

2.6 模型训练

本文通过最小化训练集上的损失函数来训练提出的 GIA-2DPE 模型. 损失函数如公式 (28) 所示, 它由 3 部分相加而成, 并由一个超参数 λ ($0 < \lambda < 1$) 来调节各部分的比重:

$$f_{\text{loss}} = \lambda f_{\text{bce}}(m_{2D}, y_{2D}) + \frac{1 - \lambda}{2}[f_{\text{bce}}(p_s, y_s) + f_{\text{bce}}(p_e, y_e)] \tag{28}$$

其中, y_s 表示实体起始边界的真实分布, 它是一个长度等于 L 的二值向量当且仅当 i 为实体的起始边界时 y_s 的第 i 个元素等于 1, 否则等于 0. 二值向量 y_e 为实体结束边界的真实分布. 而 y_{2D} 为实体边界坐标的真实分布, 它是一个大小为 $L \times L$ 的二值矩阵, 其第 i 行第 j 列元素等于 1 当且仅当坐标 (i, j) 是目标实体, 否则等于 0. 函数 f_{bce} 表示

© 中国科学院软件研究所 http://www.jos.org.cn

二值交叉熵 (BCE) 函数, 其表达式如下:

$$f_{\text{bce}}(x,y)=-\frac{1}{L}\sum_{i=1}^{L}[y_i\ln x_i+(1-y_i)(1-\ln x_i)] \tag{29}$$

本文采用反向传播 (BP) 算法来对 GIA-2DPE 模型的参数进行更新. 此外, 为了避免训练过程中出现梯度爆炸的问题, 我们将梯度的 L2 范数限制在 1.0 以内.

2.7 模型推断

在推理阶段, 对于给定的文本 T、类别标签集合 $C=\{c_i\mid i=1, 2, \ldots, n\}$($n$ 为实体类别数量) 以及人工构造的先验知识集合 $I=\{K_i\mid i=1, 2, \ldots, n\}$($K_i$ 为 c_i 对应的) 关键词序列), 执行以下步骤以获取实体识别结果.

Step 1. 对 T 进行分词, 得到文本序列 $X=\{w_i\mid i=1, 2, \ldots, L\}$($L$ 为文本序列长度).

Step 2. 以 c_1 为目标实体类型, 将 X、K_1 和 c_1 送入训练好的 GIA-2DPE 中, 得到 $m_{2\text{D}}$, 并通过算法 2 得到类型为 c_1 的实体集合 $E_1=\{(e_k, c_1)\mid k=1, 2, \ldots\}$. 再以 c_2 为目标实体类型进行相同操作, 得到 E_2. 以此类推.

Step 3. 最终的实体识别结果为 $E=\{E_i\mid i=1, 2, \ldots, n\}$.

算法 2. 获取目标实体集合.

输入: 序列 X, 矩阵 $m_{2\text{D}}$ (规格为 $L\times L$), 概率阈值 P_t 以及实体长度上限 l_m;
输出: 目标类型的实体集合 E.

```
1.  初始化一个空集合 E = {}
2.  for i = 1 to L do
3.    for j = i to i + l_m − 1 do
4.      if m_2D[i][j] ⩾ P_t then
5.        目标类型实体 e = X[i: j + 1];
6.        E.add(e);
7.      end if
8.    end for
9.  end for
10. return E;
```

3 实验与分析

3.1 数据集与评估指标

本文在 9 个具有代表性的实体识别任务的数据集上进行了实验. 这些数据集包括: 2 个包含复杂实体和嵌套实体的英文数据集 SciERC 和 ScienceIE; 3 个包含嵌套实体的英文数据集 GENIA、ACE04 和 ACE05; 2 个以扁平的简单实体为主的英文数据集 ADE 和 SOFC-Exp; 以及 2 个以简单扁平实体为主的中文数据集 MSRA 和 OntoNotes 4.0 中文版.

(1) SciERC: 文本来源于 AI 领域文献的摘要. 实体类别分为 6 种: Task, Method, Material, Metric, Generic 和 Other-Scientific-Term (OST). 其中, Task 和 Method 类别的实体大部分是具有短语结构的复杂实体, 其他类别的实体则以简单实体为主. 另外, 这些实体中还存在少量的嵌套实体.

(2) ScienceIE: 文本来源于材料科学、计算机科学以及物理学领域的文献. 实体类别分为 3 种: Task, Material 和 Processing. 其中, Task 与 Processing 类别的实体大部分为复杂实体; 而 Material 类别的实体则以简单实体为主; 并且这些实体中还包含了大量的嵌套实体.

(3) GENIA: 文本来源于生物医学领域文献的摘要. 该数据集共有 36 种细粒度的实体类别, 相关工作通常将这

© 中国科学院软件研究所 http://www.jos.org.cn

些子类别归纳为 5 种粗粒度的实体类别: Protein, DNA, RNA, Cell Type 以及 Cell Line. 该数据集的实体均为简单实体, 但包含了大量的嵌套实体.

(4) ACE04 和 ACE05: 文本来源于新闻报刊. 实体类型分为 7 种: Person (PER), Location (LOC), Organization (ORG), Facility (FAC), Weapon (WEA), Vehicle (VEH) 以及 Geographical-Entities (GPE). 所有实体均为简单实体, 但包含了大量的嵌套实体, 并且嵌套层数较深.

(5) ADE: 文本来源于电子医疗报告. 该数据集的实体类别只分为两种: Drug 和 Adverse-Effect. 所有实体均为扁平的简单实体.

(6) SOFC-Exp: 文本来源于材料科学领域的论文. 实体类型分为 4 种: Material, Device, Experiment 和 Value. 所有实体均为简单实体, 且不存在嵌套实体.

(7) MSRA: 文本来源于新闻报刊. 实体类型共 3 种: 人物, 地点和机构. 所有实体均为简单扁平实体.

(8) OntoNotes 4.0 中文版: 文本来源于新闻报刊. 实体类型有 4 种: 人物, 地点, 组织和地理政治实体. 所有实体均为简单扁平实体.

我们对上述 9 个数据集进行了信息探索, 包括实体平均长度、嵌套实体占比和最大嵌套层数等信息, 结果如表 1 所示. 这些信息将被用于后续的实验结果的分析与讨论.

表 1　实体识别任务数据集

数据集	是否包含复杂实体	是否包含嵌套实体	实体平均长度	嵌套实体占比 (%)	最大嵌套层数
SciERC	√	√	4.6	3.4	2
ScienceIE	√	√	5.2	18.2	2
GENIA	×	√	4.1	10.1	3
ACE04	×	√	3.2	24.4	5
ACE05	×	√	2.9	22.3	5
ADE	×	×	2.7	—	—
SOFC-Exp	×	×	3.9	—	—
MSRA	×	×	2.5	—	—
OntoNotes 4.0	×	×	3.4	—	—

实体识别任务的评估指标通常包括精确率 (precision, P)、召回率 (recall, R) 以及 $F1$ 分数:

$$P = N_{\text{True-Prediction}}/N_{\text{Prediction}} \tag{30}$$

$$R = N_{\text{True-Prediction}}/N_{\text{Reality}} \tag{31}$$

$$F1 = 2PR/(P+R) \tag{32}$$

其中, $N_{\text{Prediction}}$ 表示模型识别出的实体总数; N_{Reality} 代表实际的实体总数; $N_{\text{True-Prediction}}$ 代表模型识别正确的实体个数. 注意, 只有当实体内容及类别均正确才算识别正确.

另外, 上述评估指标的计算方式通常有两种: 一种是先计算各个类别的结果, 再取平均 (称为 Macro 方式); 另一种将所有类别的结果进行汇总再计算 (称为 Micro 方式). 为了与相关工作保持一致, 我们在 ADE 和 SOFC-Exp 数据集上采用 Macro 方式进行评估; 在其他数据集上采用 Micro 方式进行评估.

3.2　数据预处理与参数设置

为了增强模型对目标类型实体的语义认知, 我们为每个实体类别设计了一段关键词序列作为解释说明, 如表 2 所示. 这些简短的关键词序列将与数据集中的原始语句进行拼接, 作为模型的输入. 模型可以利用这些关键词明确输入中哪些部分与目标实体相关, 有利于提升识别性能.

实验在 GTX1080Ti GPU 上完成. 我们使用 Spacy 工具来进行初步的英文分词, 并将用于训练的输入序列的长度控制在 64 个词语以内 (受 GPU 显存限制). 使用的英文预训练模型 DeBERTa 的词典大小为 50 265, 词向量的维度为 1 024; 中文 BERT 的词典大小为 21 128, 词向量维度为 768. 在训练过程中, 我们采用一种优化的梯度下降算

© 中国科学院软件研究所 http://www.jos.org.cn

法 AdamW, 并设置权重衰减率为 0.01. 训练集和测试集的批大小 (batch size) 分别为 4 和 16. 损失函数中的系数 λ 为 0.1. 其他的实验参数配置在不同数据集上有所差别, 如表 3 所示.

表 2　实体类别的关键词序列

数据集	实体类别	关键词序列
SciERC	Task	task, processing, image, speech, video, information, translation, classification, recognition
	Method	method, techniques, approach, algorithm, model, framework, network
	Material	structured, annotated, Chinese, English, data, corpus, corpora, text, image, speech, video
	Metric	metrics, accuracy, precision, recall, *F*1, BLEU, evaluation, variation, variance, robustness
	Generic	general, common, scientific, term
	OST	other scientific term
ScienceIE	Task	task, analysis, problems, design
	Material	material, data, particles, surface
	Processing	process, model, method, algorithm, approach
GENIA	Protein	protein, organic, compounds, body, tissues, muscle, hair, collagen, enzymes, antibodies
	DNA	DNA, deoxyribonucleic acid
	RNA	RNA, ribonucleic acid
	Cell Type	cell type category
	Cell Line	cell line group
ACE04 & ACE05	PER	person, human, single, individual, group
	LOC	geographical location, areas, landmasses, mountains, water, geological formations
	ORG	organization, companies, corporations, agencies, institutions, groups of people
	FAC	facility, buildings, man-made structures, airports, highways, bridges
	WEA	weapon, physical devices, instruments, physically harming guns, arms, gunpowder
	VEH	vehicle, devices, move, carry, transported, helicopters, trains, ship, motorcycles
	GPE	geographical, political, countries, nations, regions, cities, states, government, social group
ADE	Drug	drug, interferon, methotrexate, alpha, beta, lithium acid, amiodarone carbamazepine
	Adverse-Effect	severe acute syndrome, symptoms, reaction effects, toxicity, hypersensitivity, disease
SOFC-Exp	Material	material, anode, cathode, electrolyte, fuel, interlayer, support
	Device	device SOFC
	Experiment	experiment evoking word
	Value	value, voltage, current, power, resistance, thickness, temperature
MSRA	人物	人物, 名人, 人, 人类, 个体, 人群, 大众
	地点	地点, 地理位置, 地域, 区域, 景点, 景区, 山区, 河流流域
	组织机构	组织机构, 公司, 企业, 事务所, 学校, 警局, 医院, 馆, 厂
OntoNotes 4.0	人物	人物, 名人, 人, 人类, 个体, 人群, 大众
	地点	地点, 地理位置, 地域, 区域, 景点, 景区, 山区, 河流流域
	组织机构	组织机构, 公司, 企业, 事务所, 学校, 警局, 医院, 馆, 厂
	地理政治实体	地理政治实体, 国家, 国籍, 人种, 宗教, 政府, 省, 市, 区, 镇

表 3　实验参数配置

数据集	训练轮数	学习率	dropout	P_t	l_m
SciERC	10	5×10^{-6}	0.3	0.5	20
ScienceIE	15	5×10^{-6}	0.2	0.5	25
GENIA	5	8×10^{-6}	0.3	0.5	15
ACE04 & ACE05	10	8×10^{-6}	0.3	0.5	15
ADE	10	1×10^{-5}	0.4	0.6	10
SOFC-Exp	15	8×10^{-6}	0.4	0.7	15
MSRA	10	1×10^{-5}	0.5	0.5	10
OntoNotes 4.0	12	1×10^{-5}	0.3	0.6	10

© 中国科学院软件研究所 http://www.jos.org.cn

3.3 对比实验结果与分析

本文在上述 9 个数据集上对提出的 GIA-2DPE 模型分别进行了训练. 训练过程中, 各训练集上的平均损失变化曲线如图 7 所示. 为了验证训练好的 GIA-2DPE 模型的有效性, 我们选择了相关工作中最具代表性的 3 种模型作为比较的基准模型 (在后续的实验结果表格中用“*”标记). 它们分别是:

(1) BERT+BiLSTM: 基于序列标注的一种主流实体识别模型, 在相关工作中被广泛用作基准模型.

(2) SpERT: 一种具有代表性的基于跨度分类的实体识别模型, 由 Eberts 等人[1]提出.

(3) Multi-Turn QA: Li 等人[3]提出的首个基于多轮问答的实体识别模型, 在嵌套实体识别的相关工作中经常作为比较的基准模型.

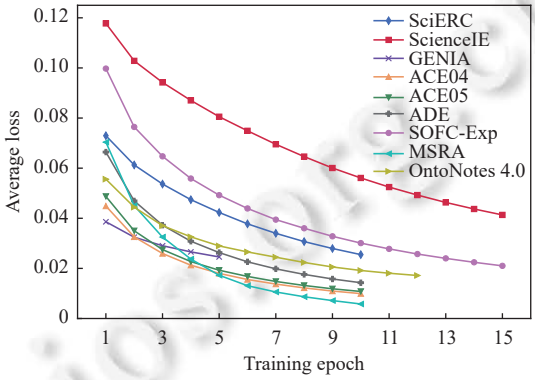

图 7 训练集上的平均损失变化曲线

3.3.1 复杂实体识别结果与分析

我们在包含复杂实体的 SciERC 和 ScienceIE 数据集上进行了实验, 结果如表 4 所示. 可以看出, 本文提出的 GIA-2DPE 模型在性能上超越了其他的代表性和最新的相关模型, 在这两个数据集上均取得了最先进的性能, $F1$ 分数分别达到了 70.8% 和 62.6%. 与基线模型相比, GIA-2DPE 模型在整体性能上分别实现了 7.0% 和 10.4% 的大幅度提升, 证明了其对复杂实体识别的有效性. 另外, 我们发现 ScienceIE 数据集上的性能提升比 SciERC 上的提升更大, 这与前者包含更多复杂实体的事实相一致.

对于表 4 中列举的大部分模型, 一方面, 它们仅使用数据集中的原始语句作为输入, 并没有考虑实体类别的语义先验知识, 因此缺乏对目标类型实体的认知. 在识别不具备明显特征的复杂实体时, 这些模型通常难以明确输入中哪些部分与目标实体相关, 进而影响识别性能. 相比之下, GIA-2DPE 模型使用人工设计的关键词序列与原始语句的拼接作为输入, 在训练之前获取了语义先验知识, 从而增强了对目标类型实体的认知. 这种认知可以帮助模型有目标地去识别实体, 减少对特征的依赖, 有利于复杂实体的识别. 我们注意到, BERT-MRC 模型[25]通过机器阅读理解的方式进行识别, 一定程度上也利用了实体类别的语义先验信息, 但它和其他模型一样忽略了不同实体类别的潜在特征; 而 GIA-2DPE 模型捕获并利用了这些类别上的潜在特征来辅助对复杂实体的识别, 提升了识别性能.

3.3.2 嵌套实体识别结果与分析

我们在包含嵌套实体的 GENIA、ACE04 和 ACE05 这 3 个数据集上也进行了实验, 结果如表 5 所示. 实验结果表明, GIA-2DPE 模型在嵌套实体的识别上, 性能同样超越了绝大多数代表性和最新的相关模型, 分别达到了 80.2%、88.2% 以及 88.5% 的 $F1$ 分数. 与基线模型相比, GIA-2DPE 模型在整体性能上分别实现了 3.4%、4.6% 和 3.7% 的提升, 证明了其对于嵌套实体识别的有效性.

在表 5 中, Path-BERT[17]、BERT+Seq2Seq[34]和 BERT+TreeCRFs[35]等模型使用复杂的解码机制, 将嵌套结构转化为扁平结构进行识别, 容易造成额外的错误或偏差[19]. Multi-Turn QA[3]和 MRC4ERE++[37]等基于问答或阅读理解的模型则未考虑所有可能的嵌套情形, 遗漏了部分嵌套实体. BERT+SoftNMS 模型[21]采用了更加直接有效的

© 中国科学院软件研究所 http://www.jos.org.cn

基于跨度的方法, 但注重于学习跨度表征, 缺乏对实体边界的精准检测[11], 容易被具有相近边界的候选实体所混淆. 本文的 GIA-2DPE 模型采用一种简单有效的 2D 概率编码机制, 利用实体的边界特征和全局上下文特征来对实体边界进行精准检测, 从而提升了嵌套实体识别的性能.

表 4 复杂实体识别的实验结果 (%)

数据集	模型	*P*	*R*	*F*1
SciERC	*BERT+BiLSTM[16]	65.6	62.1	63.8
	SCIIE[5]	67.2	61.5	64.2
	BERT-MRC[25]	69.5	62.2	65.6
	SPE[20]	67.7	66.1	66.9
	*SpERT[1]	68.5	66.7	67.6
	ENPAR[26]	—	—	67.9
	PURE[27]	—	—	68.9
	RHGN[28]	—	—	69.8
	GIA-2DPE (ours)	**71.1**	**70.5**	**70.8**
ScienceIE	SciBERT+BiLSTM[4]	55.0	49.5	52.1
	*BERT+BiLSTM[4]	55.6	49.2	52.2
	SciBERT+JLSD[15]	—	—	54.6
	BERT+JLSD[15]	—	—	55.4
	BERT-MRC[25]	57.5	54.2	55.8
	SEAL[29]	—	—	56.4
	RoBERTa+CRF[30]	62.3	55.3	58.6
	XLNet+CRF[30]	64.7	56.1	60.1
	GIA-2DPE (ours)	**66.1**	**59.3**	**62.6**

表 5 嵌套实体识别的实验结果 (%)

数据集	模型	*P*	*R*	*F*1
GENIA	HGN+BR+LR[18]	72.9	79.4	75.9
	*BERT+BiLSTM[31]	76.7	76.7	76.8
	Dispatched Attention[32]	**80.9**	73.8	76.8
	Multi-Agent[33]	77.2	76.6	76.9
	Path-BERT[17]	77.8	76.9	77.4
	BERT+Seq2Seq[34]	—	—	78.2
	BioBERT+TreeCRFs[35]	78.2	78.2	78.2
	BERT+BENSC[11]	79.2	77.4	78.3
	GIA-2DPE (ours)	80.1	**80.2**	**80.2**
ACE04	*Multi-Turn QA[3]	84.4	82.9	83.6
	BERT+Seq2Seq[34]	—	—	84.3
	Path-BERT[17]	85.9	85.7	85.8
	BERT-MRC[25]	85.1	86.3	86.0
	BERT+TreeCRFs[35]	86.7	86.5	86.6
	BERT+Seq2Set[36]	**88.5**	86.1	87.3
	BERT+SoftNMS[21]	87.4	87.4	87.4
	GIA-2DPE (ours)	88.4	**88.0**	**88.2**
ACE05	BERT+Seq2Seq[34]	—	—	83.4
	Path-BERT[17]	83.8	84.9	84.3
	*Multi-Turn QA[3]	84.7	84.9	84.8
	BERT+TreeCRFs[35]	84.5	86.4	85.4
	MRC4ERE++[37]	—	—	85.5
	BERT-MRC[25]	87.2	86.6	86.9
	BERT+SoftNMS[21]	86.1	87.3	86.7
	BERT+Seq2Set[36]	87.5	86.6	87.1
	GIA-2DPE (ours)	**88.5**	**88.5**	**88.5**

另外, 考虑到上述 SciERC 和 ScienceIE 数据集中同样存在嵌套实体, 我们将这 5 个数据集上的 *F*1 分数提升值与嵌套实体占比相联系, 发现如下规律: 无论简单的还是复杂的嵌套实体, 随着其占比的增加, *F*1 分数提升值也在增加. 这进一步证明了 GIA-2DPE 模型在嵌套实体识别上的有效性.

3.3.3 简单扁平实体识别结果与分析

为了验证本文的模型具有泛化性, 我们还在以简单扁平实体为主的英文数据集 ADE、SOFC-Exp 以及中文数据集 MSRA、OntoNotes 4.0 上进行了实验, 结果如表 6 所示. 实验结果表明, GIA-2DPE 模型在广受关注的简单扁平实体识别上, 同样优于其他代表性和最新的相关模型, 并在这 4 个数据集上都取得了最高的 *F*1 分数, 分别为 91.4%、85.0%、96.2% 和 83.2%. 与基线模型相比, 我们的模型在整体性能上分别实现了 2.1%、5.3%、1.4% 和 4.0% 的提升, 证明了其具有泛化性.

注意到 GIA-2DPE 模型在 SOFC-Exp 数据集上的提升很大, 我们分析了该数据集中不同类别实体的 *F*1 分数提升值, 如表 7 所示. 我们发现, Experiment 类型的实体的提升最多, 而该类型的实体由动词构成, 与其他类别的名词实体不同. 这证明了本文模型能够有效捕获不同实体类型的潜在特征.

© 中国科学院软件研究所 http://www.jos.org.cn

表 6 简单扁平实体识别的实验结果 (%)

数据集	模型	*P*	*R*	*F*1
ADE	DAPNA[38]	90.8	86.2	88.4
	*SpERT[1]	89.0	89.6	89.3
	CMAN[39]	—	—	89.4
	BERT+FFNN[40]	89.5	89.9	89.6
	BERT+TSE[41]	—	—	89.7
	BERT+TriMF[42]	89.5	91.3	90.4
	$SPAN_{Multi\text{-}Head}$[43]	89.9	91.3	90.6
	KECI[44]	—	—	90.7
	GIA-2DPE (ours)	**91.3**	**91.5**	**91.4**
SOFC-Exp	*BERT+BiLSTM[2]	81.5	78.1	79.7
	SciBERT+BiLSTM[2]	82.7	80.4	81.5
	GIA-2DPE (ours)	**85.2**	**84.9**	**85.0**
MSRA	Lattice-LSTM[45]	93.6	92.8	93.2
	*BERT+BiLSTM[14]	95.0	94.6	94.8
	Glyce-BERT[46]	95.6	95.5	95.5
	BERT-MRC[25]	**96.2**	95.1	95.8
	GIA-2DPE (ours)	96.1	**95.9**	**96.2**
OntoNotes 4.0	Lattice-LSTM[45]	76.4	71.6	73.9
	*BERT+BiLSTM[14]	78.0	80.4	79.2
	Glyce-BERT[46]	81.9	81.4	81.6
	BERT-MRC[25]	83.0	81.3	82.1
	GIA-2DPE (ours)	**83.6**	**82.8**	**83.2**

表 7 不同类别实体 (来自 SOFC-Exp 数据集) 的 *F*1 分数提升 (%)

模型	Material	Device	Experiment	Value
*BERT+BiLSTM[2]	88.1	81.5	76.0	72.9
GIA-2DPE (ours)	**94.9**	**84.6**	**84.8**	**75.6**
*F*1分数提升	6.8	3.1	8.8	2.7

3.4 消融实验结果与分析

GIA-2DPE 模型由 4 部分组成: 嵌入模块 (内含类别嵌入矩阵 ETE)、门控交互注意力模块 (GIA)、2D 概率编码模块 (2DPE) 以及过滤模块 (filter). 为了分析不同模块对实体识别性能的影响, 我们在提出的 GIA-2DPE 模型上进行了消融实验, 结果如表 8 所示.

表 8 GIA-2DPE 模型的消融实验结果 (*F*1) (%)

模块	SciERC	ScienceIE	GENIA	ACE04	ACE05	ADE	SOFC-Exp	MSRA	OntoNotes 4.0
GIA-2DPE	**70.8**	**62.6**	**80.2**	**88.2**	**88.5**	**91.4**	**85.0**	**96.2**	**83.2**
w/o ETE	69.4	61.1	79.2	87.3	87.6	91.2	83.8	95.8	82.7
w/o DeBERTa	69.3	60.5	79.4	86.6	87.1	90.9	84.2	—	—
w/o GIA	69.0	60.2	78.8	87.7	87.9	91.1	84.1	95.8	82.4
w/o 2DPE	69.9	60.7	78.6	85.3	86.2	90.8	84.0	95.9	82.4
w/o Filter	69.1	60.9	79.0	86.8	87.3	90.7	83.5	96.0	82.6

3.4.1 嵌入模块的影响

嵌入模块包括预训练模型和专用嵌入矩阵 ETE, 二者分别用于获取词向量和不同实体类别的潜在特征向量.

© 中国科学院软件研究所 http://www.jos.org.cn

从表 8 可以看出, 删除专用嵌入矩阵 ETE 后, 各数据集上的 $F1$ 分数有不同程度的降低, 证明了 ETE 的有效性, 并且证明了不同实体类别的潜在特征对识别的辅助作用. 具体来看, 简单实体识别的 $F1$ 分数下降了 0.2%–1.2%, 而复杂实体识别的 $F1$ 分数下降相对较多, 为 1.4%–1.6%, 这说明实体越复杂, ETE 的有效性越明显. 另外, 在简单实体的识别中, SOFC-Exp 数据集上的 $F1$ 分数下降最多, 这与第 3.3.3 节最后提到的结论一致, 进一步证明了 ETE 能够帮助模型捕获不同实体类型的潜在特征.

此外, 为了证明本文模型的有效性不完全归功于 DeBERTa, 我们还将该模块中的 DeBERTa 替换为广泛使用的 BERT (中文数据集上的实验不用替换, 因为原本使用的就是 BERT). 结合表 4–表 6 和表 8 可以发现, 虽然模型在性能上下降了 0.5%–2.1%, 但仍然超越了同样使用 BERT 的相关模型. 这证明了本文模型在除了词嵌入之外的其他方面的改进也是有效的.

3.4.2 门控交互注意力模块的影响

表 8 显示, 当删除 GIA 模块后, GIA-2DPE 模型在各数据集上的性能均有不同程度的下降, 这证明了 GIA 模块的有效性. 具体来看, 简单实体识别的 $F1$ 分数下降了 0.3%–1.4%, 而复杂实体识别的 $F1$ 下降较多, 为 1.8%–2.4%, 这说明 GIA 模块对挑战性更大的复杂实体识别反而更加有效. 我们对此进行了如下分析: 复杂实体之所以难以被识别, 是因为它们通常缺乏明显的特征, 导致模型在识别过程中难以定位相关的识别内容. 而 GIA 模块充分利用了实体类别的语义先验知识和实体的结构特征来辅助识别, 弥补了特征的缺乏.

此外, 我们还从另一角度分析了 GIA 模块的有效性. 考虑到 Zheng 等人[47]指出现有模型在长实体的识别上不理想, 我们将各数据集上的 $F1$ 分数下降值与实体平均长度进行了关联, 如图 8 所示. 可以看出, 实体的平均长度越长, $F1$ 分数下降得越多. 这证明了 GIA 模块具有改善长实体的识别性能的潜在优势, 而复杂实体比简单实体更长, 因此更能体现出 GIA 模块的这一优势.

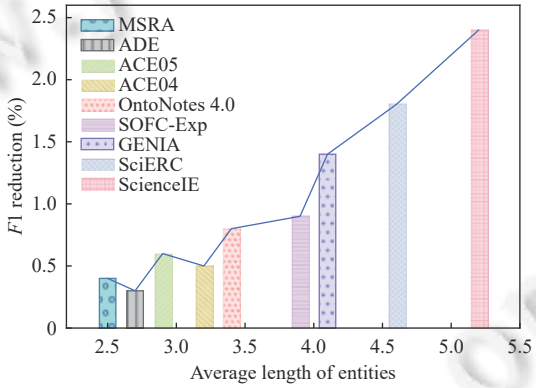

图 8 不同的实体平均长度下的 $F1$ 分数下降

3.4.3 2D 概率编码模块的影响

从表 8 可以看出, 在删除了 2DPE 模块之后, GIA-2DPE 模型在各数据集上的性能均有不同程度的下降. 其中, 扁平实体识别的 $F1$ 分数下降了 0.3%–1.0%, 而嵌套实体识别的 $F1$ 分数下降更多, 为 0.9%–2.3%. 这证明了 2DPE 模块对于嵌套实体识别的有效性.

另外, 我们还将各数据集上的 $F1$ 分数下降值与嵌套实体占比进行了关联, 得到了图 9 所示的结果. 注意到 ADE、SOFC-Exp、MSRA 和 OntoNotes 4.0 的嵌套实体占比均为 0, 因此我们选择了 ADE 数据集作为代表. 从图 9 可以看出, 嵌套实体的占比越大, $F1$ 分数下降得越多, 进一步证明了 2DPE 模块具有提升嵌套实体识别性能的优势. 我们对此进行了分析: 嵌套实体识别任务的主流方法是基于跨度的分类方法, 其专注于学习跨度本身的表征; 而 2DPE 使用实体的边界特征与全局上下文特征来对嵌套实体的边界进行精准检测, 有利于增强模型对具有细微差别的候选实体的辨识能力, 从而提升嵌套实体的识别性能.

© 中国科学院软件研究所 http://www.jos.org.cn

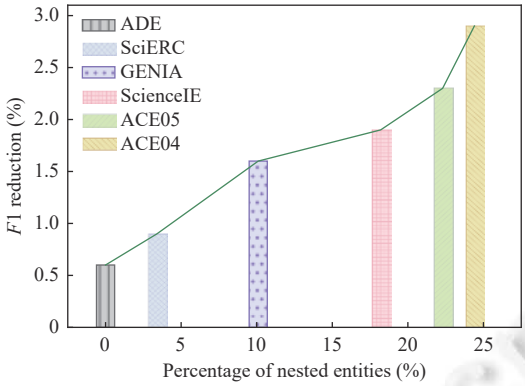

图 9　不同的嵌套实体占比下的 $F1$ 分数下降

3.4.4　过滤模块的影响

表 8 显示, 在删除了 filter 模块之后, GIA-2DPE 模型在各数据集上的 $F1$ 分数下降了 0.2%–1.7%, 证明了该模块的有效性. 我们分析了 filter 模块在实体识别中的作用: 由于我们使用实体在序列中的起始与结束边界来表示一个实体, 实体识别任务被转化为二分类任务, 即判断输入序列中的每个词语是否为实体的边界词, 若是, 则该词语的标签为 1, 否则为 0. 当文本中的实体较少时, 标签为 0 的词的数量远多于标签为 1 的词的数量, 造成不平衡问题, 直接影响了模型的训练效果. 而 filter 模块可以过滤或掩盖 (mask) 掉输出概率分布中的大部分标签为 0 的词语对应的结果, 从而提升了训练效果.

3.5　案例分析与可视化结果

为了更加直观地解释提出的 GIA-2DPE 模型的工作流程和效果, 本文对预处理好的 GENIA 数据集中的一个典型样例进行了案例分析. 该样例如图 10 所示, X 为该样例在数据集中的原始语句, c 为目标实体类型, K 为对应的关键词序列, 红色括号下方是官方标注的实体.

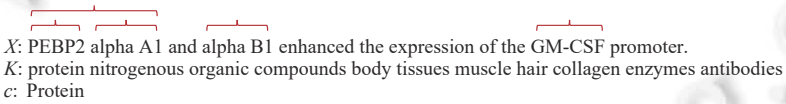

图 10　典型样例 (来源于 GENIA 数据集)

首先, 我们将 X 与 K 进行拼接后送入 GIA-2DPE 模型中, 得到词嵌入矩阵 h_o. 为了验证词向量的在语义上的准确性, 我们从 h_o 中取出了 X 的词向量和 K 的词向量, 然后计算了两者的交互注意力分数, 并对其进行了可视化, 如图 11 所示. 可以看出, “PEBP2”与“protein”的注意力分数最高, “alpha A1”和“alpha B1”与“antibodies”的注意力分数也很高, 这与实际相符. 事实上, “PEBP2”就是一种蛋白质, 而“alpha A1”和“alpha B1”就是两种抗体. 除此之外, “expression”与“body”和“muscle”的注意力分数同样很高, 这也符合生物医学知识. 由此可见, GIA-2DPE 模型获取的词向量在语义上具有较高的准确性.

其次, 我们对 filter 模块中的 3 个计算结果: 起始边界概率分布 p_{s1}、结束边界概率分布 p_{e1} 以及 2D 概率分布 m_{2D1} 进行了可视化, 分别如图 12(a)–图 12(c) 所示. 其中, 向量 p_{s1} 显示了识别出的蛋白质实体的 4 个起始边界: 1、4、8 和 16; 而向量 p_{e1} 显示了 4 个结束边界: 3、6、10 和 19. 如果不考虑 2D 概率分布矩阵, 则我们最多只能得到 4 个蛋白质类型的实体: (1, 3)、(4, 6)、(8, 10) 和 (16, 19), 分别对应“PEBP2”“alpha A1”“alpha B1”和“GM-CSF”. 当考虑 2D 概率分布时, 矩阵 m_{2D1} 显示了 5 个识别出的蛋白质类型的实体, 除上述 4 个之外还有一个 (1, 6), 对应“PEBP2 alpha A1”. 实际上, “PEBP2 alpha A1”确实也是官方标注的蛋白质实体. 由此可见, 本文提出的 GIA-2DPE 模型在实体边界的精准检测方面具备有效性.

© 中国科学院软件研究所 http://www.jos.org.cn

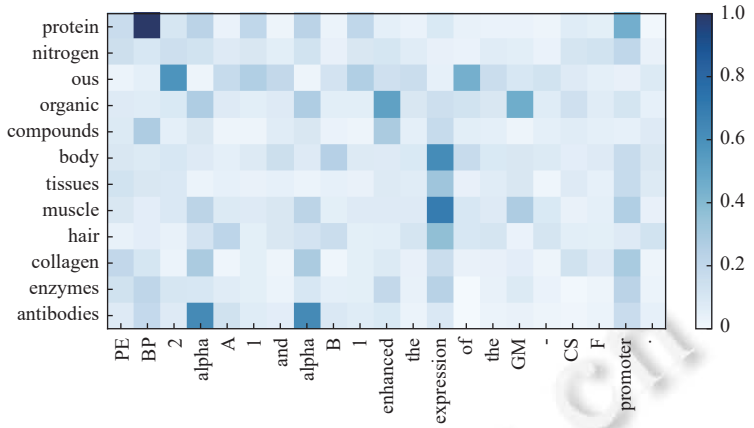

图 11 交互注意力分数的可视化

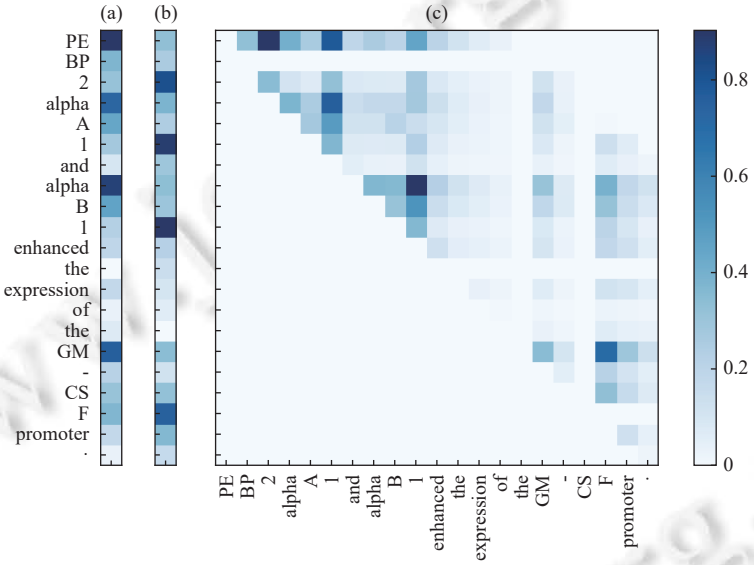

图 12 概率分布的可视化

4 总 结

本文提出了一种能够有效提升复杂实体识别性能的神经网络模型 GIA-2DPE. 该模型利用实体类别的语义先验知识增强了对实体的认知, 并通过专用嵌入矩阵自适应地捕获了不同实体类别的潜在特征, 同时利用提出的 GIA 机制将先验知识与类别特征相结合来辅助识别, 在复杂实体识别任务的基线 *F*1 分数上取得了最高 10.4% 的大幅度性能提升, 超越了目前最先进的模型. 其次, 本文模型还通过提出的 2D 概率编码机制来预测作为实体边界的词语, 利用边界特征与全局上下文特征增强了对实体边界的精准检测, 在嵌套实体识别任务的基线 *F*1 分数上也取得了最高 4.6% 的性能提升. 最后, 为了验证本文模型的泛化性, 我们在简单扁平实体识别任务上也进行了实验, 同样实现了最先进的性能.

References:

[1] Eberts M, Ulges A. Span-based joint entity and relation extraction with transformer pre-training. In: Proc. of the 24th European Conf. on Artificial Intelligence. Santiago de Compostela: IOS Press, 2020. 2006–2013. [doi: 10.3233/FAIA200321]

© 中国科学院软件研究所 http://www.jos.org.cn

[2] Friedrich A, Adel H, Tomazic F, Hingerl J, Benteau R, Marusczyk A, Lange L. The SOFC-Exp corpus and neural approaches to information extraction in the materials science domain. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 1255–1268. [doi: 10.18653/v1/2020.acl-main.116]

[3] Li XY, Yin F, Sun ZJ, Li XY, Yuan A, Chai D, Zhou MX, Li JW. Entity-relation extraction as multi-turn question answering. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 1340–1350. [doi: 10.18653/v1/P19-1129]

[4] Sahrawat D, Mahata D, Zhang HM, Kulkarni M, Sharma A, Gosangi R, Stent A, Kumar Y, Shah RR, Zimmermann R. Keyphrase extraction as sequence labeling using contextualized embeddings. In: Proc. of the 42nd European Conf. on Information Retrieval. Lisbon: Springer, 2020. 328–335. [doi: 10.1007/978-3-030-45442-5_41]

[5] Luan Y, He LH, Ostendorf M, Hajishirzi H. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018. 3219–3232. [doi: 10.18653/v1/D18-1360]

[6] Gurulingappa H, Rajput AM, Roberts A, Fluck J, Hofmann-Apitius M, Toldo L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. Journal of Biomedical Informatics, 2012, 45(5): 885–892. [doi: 10.1016/j.jbi.2012.04.008]

[7] Walker C, Strassel S, Medero J, Maeda K. ACE 2005 Multilingual Training Corpus. Philadelphia: Linguistic Data Consortium. 2006. [doi: 10.35111/mwxc-vh88]

[8] Augenstein I, Das M, Riedel S, Vikraman L, McCallum A. SemEval 2017 task 10: ScienceIE—Extracting keyphrases and relations from scientific publications. In: Proc. of the 11th Int'l Workshop on Semantic Evaluation. Vancouver: ACL, 2017. 546–555. [doi: 10.18653/v1/S17-2091]

[9] Doddington GR, Mitchell A, Przybocki MA, Ramshaw LA, Strassel SM, Weischedel RM. The automatic content extraction (ACE) program-tasks, data, and evaluation. In: Proc. of the 4th Int'l Conf. on Language Resources and Evaluation. Lisbon: European Language Resources Association, 2004.

[10] Ohta T, Tateisi Y, Kim JD. The GENIA corpus: An annotated research abstract corpus in molecular biology domain. In: Proc. of the 2nd Int'l Conf. on Human Language Technology Research. San Diego: Morgan Kaufmann Publishers Inc., 2002. 82–86.

[11] Tan CQ, Qiu W, Chen MS, Wang R, Huang F. Boundary enhanced neural span classification for nested named entity recognition. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI Press, 2020. 9016–9023. [doi: 10.1609/aaai.v34i05.6434]

[12] Li J, Sun AX, Han JL, Li CL. A survey on deep learning for named entity recognition. IEEE Trans. on Knowledge and Data Engineering, 2022, 34(1): 50–70. [doi: 10.1109/TKDE.2020.2981314]

[13] Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing. Hong Kong: ACL, 2019. 3615–3620. [doi: 10.18653/v1/D19-1371]

[14] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]

[15] Lai T, Bui T, Kim DS, Tran QH. A joint learning approach based on self-distillation for keyphrase extraction from scientific documents. In: Proc. of the 28th Int'l Conf. on Computational Linguistics. Barcelona: ACL, 2020. 649–656. [doi: 10.18653/v1/2020.coling-main.56]

[16] Jain S, van Zuylen M, Hajishirzi H, Beltagy I. SciREX: A challenge dataset for document-level information extraction. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 7506–7516. [doi: 10.18653/v1/2020.acl-main.670]

[17] Shibuya T, Hovy E. Nested named entity recognition via second-best sequence learning and decoding. Trans. of the Association for Computational Linguistics, 2020, 8: 605–620. [doi: 10.1162/tacl_a_00334]

[18] Huang HY, Lei M, Feng C. Hypergraph network model for nested entity mention recognition. Neurocomputing, 2021, 423: 200–206. [doi: 10.1016/j.neucom.2020.09.077]

[19] Li F, Wang Z, Hui SC, Liao LJ, Zhu XH, Huang HY. A segment enhanced span-based model for nested named entity recognition. Neurocomputing, 2021, 465: 26–37. [doi: 10.1016/j.neucom.2021.08.094]

[20] Wang YJ, Sun CZ, Wu YB, Yan JC, Gao P, Xie GT. Pre-training entity relation encoder with intra-span and inter-span information. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. ACL, 2020. 1692–1705. [doi: 10.18653/v1/2020.emnlp-main.132]

[21] Shen YL, Ma XY, Tan ZQ, Zhang S, Wang W, Lu WM. Locate and label: A two-stage identifier for nested named entity recognition. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language

© 中国科学院软件研究所 http://www.jos.org.cn

Processing. ACL, 2021. 2782–2794. [doi: 10.18653/v1/2021.acl-long.216]

[22] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. In: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: ACL, 2016. 1715–1725. [doi: 10.18653/v1/P16-1162]

[23] He PC, Liu XD, Gao JF, Chen WZ. DeBERTa: Decoding-enhanced BERT with disentangled attention. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021.

[24] Hendrycks D, Gimpel K. Gaussian error linear units (GELUs). arXiv:1606.08415, 2016.

[25] Li XY, Feng JR, Meng YX, Han QH, Wu F, Li JW. A unified MRC framework for named entity recognition. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 5849–5859. [doi: 10.18653/v1/2020.acl-main.519]

[26] Wang YJ, Sun CZ, Wu YB, Zhou H, Li L, Yan JC. ENPAR: Enhancing entity and entity pair representations for joint entity relation extraction. In: Proc. of the 16th Conf. of the European Chapter of the Association for Computational Linguistics. ACL, 2021. 2877–2887. [doi: 10.18653/v1/2021.eacl-main.251]

[27] Zhong ZX, Chen DQ. A frustratingly easy approach for entity and relation extraction. In: Proc. of the 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. ACL, 2021. 50–61. [doi: 10.18653/v1/2021.naacl-main.5]

[28] Wan Q, Wei LN, Chen XH, Liu J. A region-based hypergraph network for joint entity-relation extraction. Knowledge-based Systems, 2021, 228: 107298. [doi: 10.1016/j.knosys.2021.107298]

[29] Garg A, Kagi SS, Singh M. SEAL: Scientific keyphrase extraction and classification. In: Proc. of the 2020 ACM/IEEE Joint Conf. on Digital Libraries. ACM, 2020. 527–528. [doi: 10.1145/3383583.3398625]

[30] Chernyavskiy A, Ilvovsky D, Nakov P. Transformers: "The end of history" for Natural Language Processing? In: Proc. of the 2021 Joint European Conf. on Machine Learning and Knowledge Discovery in Databases. Bilbao: Springer, 2021. 677–693. [doi: 10.1007/978-3-030-86523-8_41]

[31] Jiang D, Ren HP, Cai Y, Xu JY, Liu YX, Leung HF. Candidate region aware nested named entity recognition. Neural Networks, 2021, 142: 340–350. [doi: 10.1016/j.neunet.2021.02.019]

[32] Fei H, Ren YF, Ji DH. Dispatched attention with multi-task learning for nested mention recognition. Information Sciences, 2020, 513: 241–251. [doi: 10.1016/j.ins.2019.10.065]

[33] Li CG, Wang GH, Cao J, Cai Y. A multi-agent communication based model for nested named entity recognition. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2021, 29: 2123–2136. [doi: 10.1109/TASLP.2021.3086978]

[34] Straková J, Straka M, Hajic J. Neural architectures for nested NER through linearization. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 5326–5331. [doi: 10.18653/v1/P19-1527]

[35] Fu Y, Tan CQ, Chen MS, Huang SF, Huang F. Nested named entity recognition with partially-observed TreeCRFs. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI Press, 2021. 12839–12847.

[36] Tan ZQ, Shen YL, Zhang S, Lu WM, Zhuang YT. A sequence-to-set network for nested named entity recognition. In: Proc. of the 30th Int'l Joint Conf. on Artificial Intelligence. Montreal: ijcai.org, 2021. 3936–3942. [doi: 10.24963/ijcai.2021/542]

[37] Zhao TY, Yan Z, Cao YB, Li ZJ. Asking effective and diverse questions: A machine reading comprehension based framework for joint entity-relation extraction. In: Proc. of the 29th Int'l Joint Conf. on Artificial Intelligence. Yokohama: ijcai.org, 2020. 3948–3954. [doi: 10.24963/ijcai.2020/546]

[38] Kong LY, Lai QH, Liu S. End-to-end drug entity recognition and adverse effect relation extraction via principal neighbourhood aggregation network. Journal of Physics: Conference Series, 2021, 1848: 012110. [doi: 10.1088/1742-6596/1848/1/012110]

[39] Zhao S, Hu MH, Cai ZP, Liu F. Modeling dense cross-modal interactions for joint entity-relation extraction. In: Proc. of the 29th Int'l Joint Conf. on Artificial Intelligence. 2020. 4032–4038. [doi: 10.24963/ijcai.2020/558]

[40] Giorgi J, Wang XD, Sahar N, Shin WY, Bader GD, Wang B. End-to-end named entity recognition and relation extraction using pre-trained language models. arXiv:1912.13415, 2019.

[41] Wang J, Lu W. Two are better than one: Joint entity and relation extraction with table-sequence encoders. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. ACL, 2020. 1706–1721. [doi: 10.18653/v1/2020.emnlp-main.133]

[42] Shen YL, Ma XY, Tang YC, Lu WM. A trigger-sense memory flow framework for joint entity and relation extraction. In: Proc. of the 2021 Web Conf. Ljubljana: ACM, 2021. 1704–1715. [doi: 10.1145/3442381.3449895]

[43] Ji B, Yu J, Li SS, Ma J, Wu QB, Tan YS, Liu HJ. Span-based joint entity and relation extraction with attention-based span-specific and contextual semantic representations. In: Proc. of the 28th Int'l Conf. on Computational Linguistics. Barcelona: Int'l Committee on Computational Linguistics, 2020. 88–99. [doi: 10.18653/v1/2020.coling-main.8]

[44] Lai T, Ji H, Zhai CX, Tran QH. Joint biomedical entity and relation extraction with knowledge-enhanced collective inference. In: Proc. of

© 中国科学院软件研究所 http://www.jos.org.cn

the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing. ACL, 2021. 6248–6260. [doi: 10.18653/v1/2021.acl-long.488]

[45] Zhang Y, Yang J. Chinese NER using lattice LSTM. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: ACL, 2018. 1554–1564. [doi: 10.18653/v1/P18-1144]

[46] Meng YX, Wu W, Wang F, Li XY, Nie P, Yin F, Li MY, Han QH, Sun XF, Li JW. Glyce: Glyph-vectors for Chinese character representations. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver, 2019. 247.

[47] Zheng HY, Qin B, Xu M. Chinese medical named entity recognition using CRF-MT-Adapt and NER-MRC. In: Proc. of the 2nd Int'l Conf. on Computing and Data Science. Stanford: IEEE, 2021. 362–365. [doi: 10.1109/CDS52072.2021.00068]

姜小波(1972－), 男, 博士, 副教授, 主要研究领域为智能人机交互, 自然语言处理, 知识图谱.

阎广瑜(1999－), 男, 硕士, 主要研究领域为自然语言处理, 信息抽取, 数据挖掘.

何昆(1995－), 男, 硕士, 主要研究领域为自然语言处理, 信息抽取, 知识图谱.

© 中国科学院软件研究所 http://www.jos.org.cn