文章编号：1003-0077（2017）00-0000-00

基于多任务学习的汉语基本篇章单元和主述位联合识别

葛海柱[1] 孔芳[2]
（苏州大学 计算机科学与技术学院，江苏 苏州 215006）

摘要：基本篇章单元（Elementary Discourse Units，EDU）识别是构建篇章结构的基础，对篇章分析意义重大。从篇章衔接性视角看，篇章话题结构理论认为，每个 EDU 都由要表达信息的起始点（主位）和传达的新信息（述位）两部分构成。因此，EDU 识别与主述位识别任务间关系密切。基于此，本文给出了一个基于多任务学习的汉语基本篇章单元和主述位联合识别方法。该方法利用双向长短时记忆网络和图卷积网络对基本单元进行序列化和结构化拓扑信息的表征，再利用多任务学习框架让两个任务共享参数，借助不同任务间的相关性来提升模型的性能。实验结果表明，基于多任务学习的 EDU 和主述位识别性能均优于单任务学习模型中各自的性能，其中基本篇章单元识别的 F1 值达到 91.9%，主述位识别的 F1 值达到了 85.65%。

关键词：多任务学习；基本篇章单元；主位；述位

中图分类号：TP391 **文献标识码：**A

Chinese Elementary Discourse Unit and Theme-Rheme Jointly Detection based on Multi-task Learning

GE Haizhu[1], KONG Fang[2]

(School of Computer Sciences and Technology, Soochow University, Suzhou, Jiangsu 215006 ,China)

Abstract : Elementary discourse unit (EDU) recognition is a fundamental task of discourse analysis. From the perspective of discourse cohesion,the the theoretical system of discourse topic structure holds that each EDU is composed of theme and theme. Theme represents the starting point of the information and theme is the new information being transmitted. Therefore, EDU detection is closely related to rheme-theme recognition task.Thus this paper proposes a Chinese elementary discourse unit and theme-rheme joint detection method based on multi-task. This method uses Bidirectional Long Short-Term Memory networks and Graph Convolutional networks to represent the elementary discourse unit's serialized and structured topological information, and uses multi-task learning framework to make the two tasks share parameters, and improves the performance of the model with the help of the correlation between different tasks. The experimental results show that the performance of EDU and theme-rheme detection based on multi-task is better than that of the single-task learning model.The F1-score of elementary discourse unit detection reached 91.9% and the F1-score of theme-rheme detection reached 85.65%.

Key words: multi-task; elementary discourse unit; theme; rheme

0 引言

随着句子级词法、句法研究的日臻成熟，篇章级别的任务逐渐进入研究者的视线。篇章分析正是篇章级任务的研究热点之一，它在信息抽

收稿收稿日期：2017-03-16；**定稿日期**：2017-04-26
基金项目：国家自然科学基金面上项目（61876118），国家自然科学基金人工智能应急管理项目(61751206）

取、机器翻译、指代消解等自然语言处理领域中的应用越来越广泛，成为自然语言理解的核心问题之一。

篇章分析作为自然语言处理的一个研究热点，主要任务就是从整体上分析出篇章结构及其构成单元之间的语义关系，并从上下文角度理解篇章。篇章也称语篇，通常是由一系列连续的子句、句子和句群构成的语言整体单位[1]。根据不同的篇章分析目的，基本篇章单元及其关系可以表示为不同的篇章基本结构，主要包括修辞结构、话题结构、指代结构、功能结构、事件结构等范畴[1]。因此，篇章分析的一般步骤包括，（1）识别基本篇章单元。基本篇章单元（Elementary Discourse Units，EDU）是句子中具有独立语义和独立功能的最小单位，是进行篇章分析的基本单位。（2）篇章结构及关系的解析。将识别出的基本篇章单元依据一定的关系（例如修辞关系）构建形成特定结构，常见的结构有树和图等。因此，无论进行哪种篇章结构的分析，EDU 识别都是一项基础工作，它的识别性能会对后续篇章结构的解析产生极大的影响。

本文提出了一种基于多任务深度学习的 EDU 和主述位自动识别联合模型，该模型利用双向长短时记忆网络和图卷积网络进行特征抽取，利用多任务学习方法对 EDU 和主、述位识别进行联合学习。实验结果表明，多任务学习模型取得了一定的效果。

1 相关工作

1.1 基本篇章单元识别的相关研究

随着 RST-DT（Rhetorical Structure Theory Discourse Treebank）[2]与 PDTB（Penn Discourse Treebank）[3]英文篇章语料库的发布，针对英文基本篇章单元识别的研究受到了很多研究人员的关注。代表性工作包括：Sporleder 和 Lapata[4]第一个引入神经网络模型，将基本篇章单元识别作为序列化标注问题。Xuan Bach 等[5]在 RST-DT 语料中进行的 EDU 识别实验得到目前的最优性能，F1 值为 93.7%。Chloe Braud[6]采取序列化标注的方法，使用自动词法、句法信息作为输入特征，F1 值为 86.8%。

相比英文，有关汉语基本篇章单元识别的研究相对较少，主流方法是将 EDU 识别任务看作逗号分类问题，代表性工作有：李艳翠[7]分析了逗号与基本篇章单元的关系，并在标注语料上进行了基于逗号的汉语 EDU 识别研究。Nianwen Xue 等[8]将中文子句切分当作逗号分类问题，自动识别汉语句子中表示句号功能的逗号，识别的准确率接近 90%。Jin 等[9]提出利用逗号、谓词等特征分割汉语句子的方法，准确率为 87.1%。

上述汉语 EDU 识别研究都是基于传统机器学习方法，基本思想是将汉语 EDU 识别当作逗号分类问题，虽然取得了不错的识别效果，但也有不足之处。首先，他们的模型均需人工提取特征，而人工建立特征工程往往需要投入大量的人力、物力。而且实验效果依赖标准词法、句法信息，当没有标准信息时，实验结果较差。鉴于此，本文构建了一个基于深度学习的多任务 EDU 与主述位自动识别模型，它可以自动学习和共享每个任务和领域的信息，不需要大量人工的参与。

1.2 主、述位识别的相关研究

主述位结构及推进模式理论是篇章话题结构理论体系中的一种。主述位理论中的主位、述位两个概念，最早来自于布拉格学派提出的功能语句观理论框架，在语言学中得到广泛的应用，但在计算机领域运用较少。Mathesius[10]从功能语句观的角度提出主位、述位信息理论，用于描述句子所传递的信息结构。主位是指在既定语境中已知或至少是明显的信息，是说话人信息的出发点；述位是话语的核心，是说话人对主位的阐发。此后，Halliday[11]认为：句子按主位展开，主位用于表示在上下文语境中已知或是明显的信息，是说话人想要表达信息的起始点；述位代表话题的核心，用于表示说话人扩展或解释主位的信息，往往是说话人要传达的新信息。

受限于语料库，有关主、述位的研究相对较少。Kwanghyun Park[12]提出了一种基于系统功能语言学的主、述位结构自动分析计算系统。该系统以英文文本作为输入，输出为文本中各句子的主、述位结构。奚雪峰等[13]提出了基于主述位理论的篇章微观话题结构表示体系，并依据它标注形成了 500 篇文档的微观话题结构语料库 CDTC

（Chinese Discourse Topic Corpus）。[1]他在 CDTC 上使用基于判定树的主、述位自动识别方法进行实验，主、述位识别准确率为 74.05%。

1.3 多任务学习

多任务学习是指在一个模型中同时完成多个任务，利用不同任务之间的相似性来辅助决策。Caruana[14]于 1997 年使用预测不同道路的特征来辅助学习自动驾驶的方向掌握。Zhang[15]等使用头部姿势估计与面部属性特征推断辅助脸部轮廓检测任务。Liu[16]等同时学习查询分类与网页搜索。Girshick[17]同时预测图像中物体的类别和位置。Arik[18]同时预测文本到语言的过程中音素的持续时间和频率。由此可见，多任务学习能够在一定程度上改进单任务学习的缺陷，提高模型的预测性能。

鉴于此，本文尝试在深度学习的基础上，引入多任务学习方法，充分利用任务之间的互补性及语料数据信息，从而完成汉语基本篇章单元与主述位边界的自动切分任务。尽管多任务学习可以提高模型的预测性能，但是，并非所有任务都可以通过多任务学习组合在一起进行训练。多任务学习的前提是任务之间具有相关性，如果任务之间不相关，可能会使性能降低，产生负迁移现象（Negative Transfer）。

从篇章角度分析，主位是基本篇章单元（EDU）中的第一个构成成分，述位是基本篇章单元中去除主位后遗留的成分[19]。因此，一个完整的句子可看作“主位-述位-主位-述位……”的序列，其中相邻的主位-述位构成一个基本篇章单元。基本篇章单元与主位、述位的关系如例子 1 所示。

例子 1：[[外商投资企业的出口商品]T1 [仍以轻纺产品为主，]R1]EDU1 [[其中，出口额最大的商品]T2 [是服装，]R2]EDU2 [[Φ]T3 [去年为七十六点八亿美元。]R3]EDU3

例子 1 中，句子“外商投资企业的出口商品仍以轻纺产品为主，其中，出口额最大的商品是服装，去年为七十六点八亿美元。”由三个基本篇章单元组成。其中，EDU1“外商投资企业的出口商品仍以轻纺产品为主，”中包含主位 T1“外商投资企业的出口商品”和述位 R1“仍以轻纺产品为主，”。同理 EDU2 中 T2“其中，出口额最大的商品”为主位，R2“是服装。”为述位。但 EDU3 则与 EDU1、EDU2 不同，其主位 T3 被省略，为隐式主位，述位为 R3“去年为七十六点八亿美元。”。

因此，主、述位识别任务与基本篇章单元识别的任务之间具有明显的相关性。（1）两者任务定义都是将一个句子分成多个完整的片段；（2）篇章基本单元由一个主位和一个述位构成，当两个任务共享编码层时，主述位边界的确定有助于 EDU 边界的确定，同理，确定了 EDU 边界又为主、述位的识别确定了主位和述位的切分点的范围。

2 基于多任务学习的 EDU 及主述位识别

本文构建了一个基于深度学习的多任务模型，如图 1 所示。它可以自动学习和共享每个任务和领域的信息，而不需要人工的努力。与传统的 EDU 识别方法相比，进行了如下改进：（1）利用多任务学习方法学习不同任务的特定深度特征；（2）使用 Pointer Network 代替传统的条件随机场（CRF）进行解码；（3）使用图卷积神经网络（Graph Convolutional Network，以下简称 GCN）模型融入句子的依存句法信息。

2.1 编码层

本模型基于多任务模型实现，共享编码层。编码层由双向长短时记忆网络（BiLSTM）和图卷积神经网络模型实现，用于提取文本的语法与句法特征。

2.1.1 BiLSTM

在编码阶段，我们将含有 n 个词的句子记作：$(x_1, x_2, x_3, ..., x_n)$，其中 x_i 表示句子的第 i 个词在词表中的 id。然后，利用预训练的 Embedding 矩阵将句子中的每个词 x_i 映射为低维稠密的词向量，最终将该词向量与随机初始化的词性向量拼接作为 BiLSTM 的输入。前向 LSTM 与后向 LSTM 的隐藏状态在 LSTM hidden 处串联表示整个序列的全局信息，最终将此信息作为 GCN Layer 层的输入。

[1] 李艳翠等标注的汉语连接词驱动的篇章树库 CDTB 与奚雪峰等标注的微观话题结构语料库 CDTC 使用了一致的基本篇章单元 EDU，它们从不同篇章视角进行汉语篇章分析。

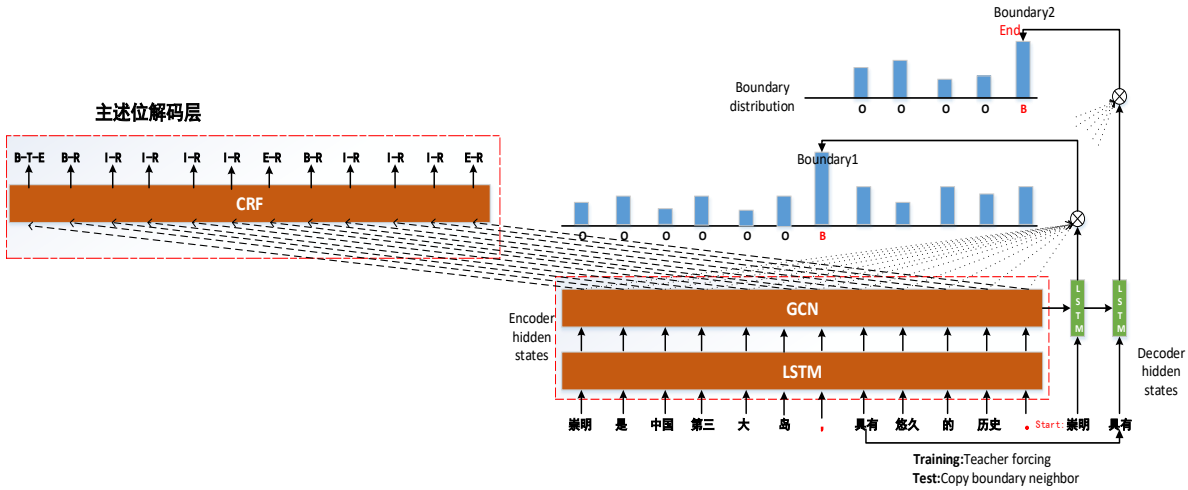

图 1 基于多任务学习的基本篇章单元识别模型

2.1.1 GCN

众所周知，BiLSTM 模型在序列化标注模型中取得了极好的效果，目前的 state-of-the-arts 基本都是基于 BILSTM-CRF 模型。但是该模型仅利用 LSTM 捕获输入文本中的序列信息，实际上可以有更多的信息可以利用，比如句法分析。我们认为早期的研究者们没有使用这些句法信息的原因在于缺乏一种简单、有效的方法将句法信息纳入序列神经网络模型。因此，本文采用图卷积模型[20]来解决这个限制。

本部分参考 Diego Marcheggiani[21]提出的方法，将传统 GCN 模型改进为一种基于句法的 GCN 编码器，从而能够基于自动预测的依存句法[2]结构对 BiLSTM 的输出进行重新编码。而且 Diego Marcheggiani 发现 GCNs 与 LSTMs 是相互补充的，虽然 BILSTM 能够在没有提供句法信息的情况下捕获一定程度上的句法信息，但是 LSTM 对于相距较远的词对之间关系的获取并不好，距离越远，其效果越差[22]。而 GCN 可以帮助缩减两个词之间的距离。

本部分使用的是简化版的 GCN 模型，定义如下：对于一个图 $G=(V,E)$，V 表示图中的节点（$v_1,v_2,...,v_N$）（本文为句子中每个词 $w_1,w_2,...,w_N$），每个节点都携带一个特征常量或者特征向量。同 LSTM 一样，GCN 也可以叠加多层。多层 GCN 可以合并更高程度的邻域，获得更丰富的信息。经过 GCN 后，节点 i 携带的特征向量可由以下公式形式化表达：

$$h_v^{(k)}=\mathrm{Re}LU\Big(\sum_{u\in N(v)}W_{L(u,v)}^{(k-1)}h_u^{(k-1)}+b_{L(u,v)}^{(k-1)}\Big) \quad (1)$$

其中，k 表示第 k 层 GCN，当 k 等于 1 时，h_u^0 为 LSTM 的输出（$h_1,h_2,h_3,...,h_N$），$L(u,v)$ 包括两个词之间的依存关系和依存弧方向。图 2 给出了 GCN 模型的具体实现方法。

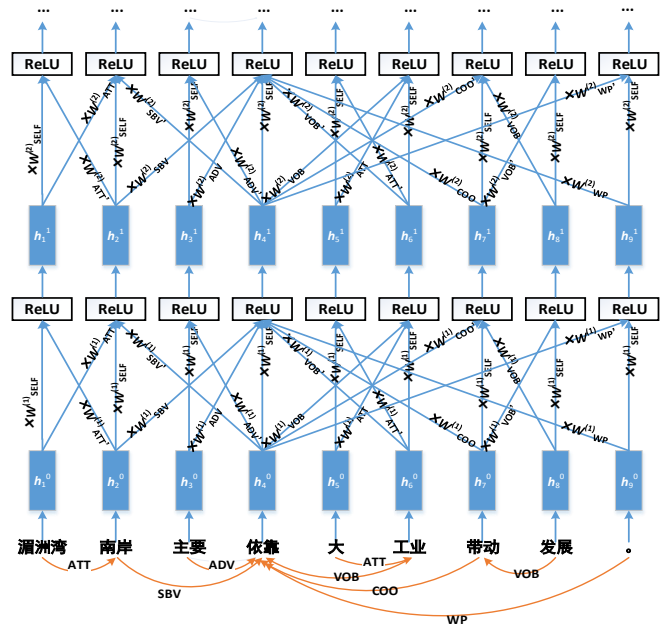

图 2 基于依存句法的图卷积模型

由图 2 和公式（1）可知，经过一层 GCN 后，每个节点携带的特征向量变成其依存句法图中邻接点特征向量的加权平均和。但是这样的做法存在两大问题：

（1）pyltp 工具生成的依存句法中每个节点都不存在自连接（自身与自身有一条边），如图 2 中“南岸”这个依存节点包含一条以其为起点

[2] 依存句法使用 pyltp 工具包自动生成

的边 $\overrightarrow{SBV}$ 和一条以其为终点的边 $\overleftarrow{ATT}$。因此经过 GCN 后会导致每个节点丢失自身所携带的信息。

（2）如果 $L(u,v)$ 中即考虑两个词之间的依存关系又考虑信息流动的方向，这会导致模型 over-parameterized。

为解决上述问题，我们参考 Diego Marcheggiani 提出的方法，制定了如下约束：

约束 1：为了避免丢失节点自身所携带的信息，我们对图中每个节点添加一个指向自己的特殊的边，指定其标签为 SELF。

约束 2：在依存句法图中，并不能假设信息仅沿着依存弧的方向流动，我们同样允许信息沿着反方向流动。

约束 3：由于我们将标签扩展为正反向和 SELF，则原来 pyltp 使用的 14 中依存关系被扩展为 29 种，每种依存关系对应一个矩阵 W 和向量 b，如图 2 中参数 W 所示。这会使得模型参数过多，导致模型 over-parameterized。因此在本文中我们不关心依存关系的具体类别，只保留依存关系的三种方向，即只有三种邻接点组合方式：（1）$W_{ARC}^{(i)}$ 表示节点为依存弧的起点；（2）$W_{ARC'}^{(i)}$ 表示节点为依存弧的终点；（3）$W_{SELF}^{(i)}$ 表示指向自己的边。以图 2 中节点“南岸”为例，图中 $W_{ATT'}^{(1)}$ 被转化为 $W_{ARC}^{(1)}$，$W_{SBV}^{(1)}$ 被转化为 $W_{ARC}^{(1)}$。

2.2 解码层

在解码环节，我们对 EDU 识别和主述位识别两个任务分别采用不同的解码方式。

2.2.1 EDU 识别的解码

对于 EDU 识别任务，传统的基于机器学习的方法严重依赖于精心设计的特征。近年来，神经网络模型已经成功地应用于 EDU 识别任务。例如，在 EDU 识别任务中使用长短时记忆网络（LSTM）对输入序列进行编码，使用条件随机场（CRF）模型对标签序列进行解码[4][5][6]。但是基于序列标注的神经网络模型虽然不需要手工制定特征工程，但它们也存在相应的缺陷。比如：模型容易受到 EDU 边界稀疏的影响、不能很好的处理可变大小输出词汇等问题[23]。

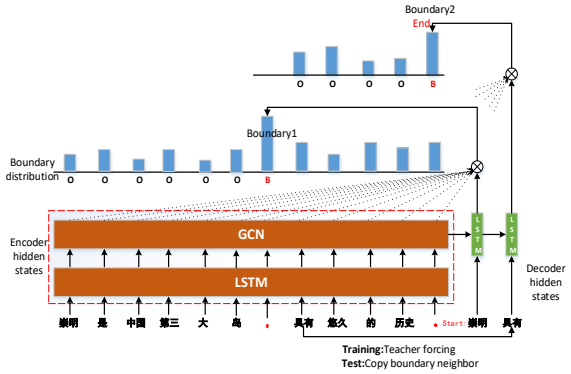

图 4 基于多任务学习的 EDU 识别模型[3]

为解决上述问题，我们采用 Pointer Network[24]作为解码器。如图 4 所示，该解码器由两部分构成：Decoding Phase 和 Pointing Phase。

由于解码时的边界数量随输入而变化，因此在 Decoding Phase 中使用循环神经网络的变种 LSTM 模型进行解码工作。训练时，Decoding Phase 将输入序列中的每个 EDU 的第一个词 U_m 作为启动单元。如图 4 Decoder hidden states 部分所示，输入句子为“崇明 是 中国 第三 大 岛 ，具有 悠久 的 历史 。”，其中包含两个 EDU，启动单元 U_m 分别为：“崇明”和“具有”。然后，利用预训练的 Embedding 矩阵将 U_m 映射为低维稠密的词向量 x_m。最后，将 x_m 作为一个单向 LSTM 的输入，Decoding Phase 的隐层状态计算方式如下：

$$d_m = LSTM(x_m, \theta) \qquad (2)$$

其中，θ 为 LSTM 中隐藏层的参数。如果输入序列中包含 M 个 EDU，即 U_m 包含 M 个启动单元，则 Decoding Phase 产生的隐层状态 $d \in \mathbf{R}^{M \times H}$，$H$ 为隐藏层的维度。

传统的 seq2seq 模型的输出词汇表是固定的，而在我们的示例中，每个解码步骤中输入序列中边界的数量都会发生变化。为了解决这个问题，我们在解码器中使用了一个指向机制[23]（pointing mechanism）来计算输入序列中分割点的位置：

$$u_j^m = v^T \tanh(W_1 h_j + W_2 d_m)，\text{for} \quad j \in (m,...,M)$$

[3] 为制图方便及页面大小的限制，图 4 中仅体现了 EDU 识别任务的模型，但图 4 与图 5 的 Encoder 部分是共享的（共享 BiLSTM+GCN 部分）。

$$p(y_m \mid x_m) = soft\max(u^m) \quad (3)$$

其中，$h \in \mathbf{R}^{N\times 2H}$ 和 $d \in \mathbf{R}^{M\times H}$ 分别为编码层和 Decoding Phase 隐藏状态的输出，$j \in [m, M]$ 表示输入序列中分割点位置。$p(y_m \mid x_m)$ 表示起始序列为 U_m 时，U_j 为 EDU 边界的概率。

在训练时，我们采用“teacher forcing[25]”机制来训练模型，即训练时在 Decoding Phase 中为模型提供正确的启动单元 U_m 和分割边界“B”。teacher forcing 的工作原理是：在训练过程的 t 时刻，使用训练数据集的期望输出或实际输出 $y(t)$ 作为下一时间步的输入 x_{t+1}，而不是使用模型生成的输出 $h(t)$。这一机制可以保证在训练时解码层的输出接近正确的启动单元和分割边界。

但是当使用解码层进行测试时，模型无法获取正确的启动单元与分割边界。我们采用与传统 seq2seq 解码器类似的方法，根据前一步的输出确定输入。以图 4 为例，输入序列为“崇明 是 中国 第三 大 岛 ， 具有 悠久 的 历史 。”，解码器将输入序列的第一个词“崇明”送入 Decoding Phase 的 LSTM 得到 d_0，然后通过公式（3）计算“崇明”到输入序列末端“。”中所有位置的边界分布概率，得到“，”为分割边界，如图所示得到第一个分割边界“Boundary1”。然后，将上一步得到的分割边界“，”的下一个词“具有”送入 Decoding Phase 的前向 LSTM 得到 d_1，同样计算“具有”到序列末端“。”的边界分布概率。

2.2.2 主位、述位识别的解码

主述位识别的解码器采用 CRF 模型进行句子级别的序列化标注。CRF 模型接收编码层传来的全局信息作为特征，借助解码环节为每个词分配标签。如果记一个长度等于句子 x 中词的个数的标签序列为 $y = (y_1, y_2, y_3, \ldots, y_n)$，那么模型对于句子 x 的标签等于 y 的打分为：

$$score(x, y) = \sum_{i=1}^{n} P_{i,y_i} + \sum_{i=1}^{n+1} A_{y_{i-1}, y_i} \quad (4)$$

其中，A_{ij} 表示的是从第 i 个标签到第 j 个标签的转移得分，从公式（4）可以看出整个序列的打分等于各个位置的打分之和，而每个位置的打分由编码器输出的 p_i 和 CRF 的转移矩阵 A 决定。对所有的得分使用 Softmax 进行归一化后的概率：

$$P(y \mid x) = \frac{\exp(score(x, y))}{\sum_{y'} \exp(score(x, y'))} \quad (5)$$

其中，x 为训练样本，分子上的 y 为正确的标注序列，下面对真实标记序列 y 的概率取 log，得到损失函数：

$$\log(P(y \mid x)) = score(x, y) - \log(\sum_{y'} \exp(score(x, y'))) \quad (6)$$

最终的目标就是最大化公式（6），因此对公式（6）取负，然后最小化，这样就可以使用梯度下降等优化方法来求解参数。

模型训练完毕，使用动态规划的 Viterbi 算法解码，求解最优路径：

$$y^* = \arg\max score(x, y') \quad (7)$$

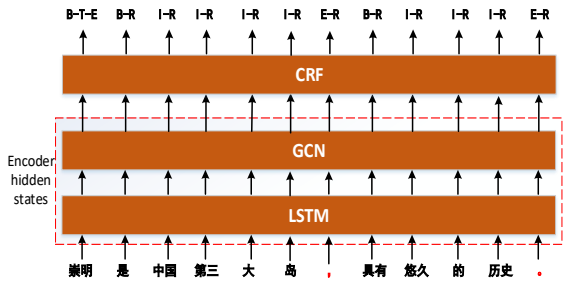

图 5 基于多任务学习的主述位识别模型

最终，将 y^* 作为预测结果输出。如图 5 输出所示，模型的输入为“崇明 是 中国 第三 大 岛 ， 具有 悠久 的 历史 。”，预测结果为“B-T-E B-R I-R I-R I-R I-R E-R B-R I-R I-R I-R E-R”，预测结果中的每一个标签对应输入句子中相应位置的词，由标签我们可将输入句子分为 3 个片段，如图 1 所示：第一个片段为 Theme1，即主位“崇明”；第二个片段为 Rheme1，即述位“是中国第三大岛，”；第三个片段为 Rheme2，即述位“具有悠久的历史。”。当主位为单个词时，我们设计了一个特殊的标签“B-T-E”来处理这种情况。一般情况下主位的标签与述位类似，由“B-T”、“I-T”、“E-T”构成。

2.2 损失函数

本文采用多任务学习的方式，所以，在联合训练的过程中，最终的损失函数由 EDU 识别任务的损失 $Loss_{EDU}$ 和主、述位识别任务的损失 $Loss_{TR}$ 加和得到：

$$Loss_{total} = Loss_{EDU} + Loss_{TR}$$

3 实验结果与分析

3.1 实验设置

实验选用的语料是基于微观话题结构（Micro-Topic Scheme）的汉语篇章话题结构语料库（Chinese Discourse Topic Corpus，CDTC）[26]。

表 1 实验超参数设置

Parameter	Value
Batch size	32
Hidden_dim	128
Word_emb_dim	300
Pos_emb_dim	20
GCN layer	1
LSTM layer	1
Learning_rate	0.5
Optimizer	AdaDelta
Dropout	0.65
Iteration	200

该语料库从 CTB6.0 中抽取了 500 篇文档进行标注，标注了基本篇章单元、主位、述位、篇章话题链等信息。该语料中所有标注项的 Kappa 值均大于 0.75，其中基本篇章单元的 Kappa 值为 0.91、主述位的 Kappa 值为 0.83。关于该语料的详细介绍，可参考文献[26]。实验超参数设置如表 1 所示。本文模型如未特殊说明，均在自动句法树下进行。

3.2 实验结果及分析

本文提出了一种多任务的 EDU 识别方法，首先验证该方法的有效性。为了验证多任务的有效性，我们的实验分两组进行，分别是单任务实验和多任务实验。单任务的 EDU 识别模型与多任务学习对应的 EDU 识别部分的模型相同、参数也相同，均采用 BiLSTM+GCN 模型进行编码、Pointer Network 模型进行解码，具体结果如表 2 所示。

表 2 两种不同 EDU 识别方法对比

System	P/%	R/%	F/%
Single_task_edu	89.05	91.31	90.15
Multi_task_edu	91.22	92.58	**91.90**

同理，我们对主述位识别进行了相同的对比实验，均采用 BiLSTM+GCN 模型进行编码、CRF 模型进行解码，结果如表 3 所示：

表 3 两种不同主、述位识别方法对比

System	P/%	R/%	F/%
Single_task_tr	83.22	87.52	85.31
Multi_task_tr	83.32	88.11	**85.65**

从表 2、表 3 可以看出，相对于单任务学习，多任务学习取得了一定的进步，EDU、主述位识别的 F 值分别提高了 1.7%、0.3%。此结果验证了多任务学习的有效性。EDU 识别与主述位识别都是进行句子级别的序列化标注工作，二者具有明显的相关性，并行学习，参数共享，改变了权值动态更新的动态特性，可能使网络更适合多任务学习。

已有的一些中文 EDU 识别研究都是把 EDU 识别看作逗号消歧问题，通过人工提取逗号所在上下文的多种信息对逗号进行分类，从而完成 EDU 的识别，代表性的工作有李艳翠等[27]。由于评价指标不尽相同，本文复现了李艳翠的基于 SVM 逗号分类的自动 EDU 识别模型，同时也复现了传统的基于 BiLSTM+CRF 的自动 EDU 识别模型，实验结果如表 4 所示。

表 4 Multi-task/SVM/BiLSTM EDU 识别结果

System	P/%	R/%	F/%
SVM 标准句法树	90.0	88.3	89.1
SVM 自动句法树	86.9	87.0	87.0
BiLSTM+CRF	83.5	88.6	85.6
Multi_task_edu	91.2	92.5	**91.9**

通过实验结果的对比可以发现，使用自动句法树的情况时，基于多任务的深度学习模型取得较好的实验结果，相比于使用自动句法树的 SVM

分类器，系统 F1 值提高了约 2.8%，其中 BiLSTM+CRF 模型未使用句法信息。

虽然 BiLSTM 模型在序列化标注模型中取得了较好的效果，但是该模型仅利用了句子中的语义信息，实际上可以有更多的信息可以使用，比如句法分析。本文采用图卷积（GCN）模型将句法信息纳入序列神经网络模型。为了验证 GCN 的有效性，本文分别使用了 BiLSTM+Pointer Network 模型和 BiLSTM+GCN+Pointer Network 模型进行 EDU 识别，实验结果如表 5 所示。

表 5 两种不同 EDU 识别方法对比

System	k	P/%	R/%	F/%
BiLSTM+PN	0	87.38	90.98	89.15
BiLSTM+GCN+PN	1	89.05	91.31	**90.15**
BiLSTM+GCN+PN	2	88.57	91.17	89.85
BiLSTM+GCN+PN	3	88.76	90.76	89.74

表中 k 表示 GCN 的层数。实验结果表明，在相同条件下，加入 GCN 的 EDU 自动识别模型的性能更好。当 k 为 1 时，模型取得最好的性能。GCN 可以用来解决一般神经网络不容易处理图等结构化拓扑结构的问题，本文使用 GCN 从依存句法中抽取拓扑特征，提高 EDU 分割的性能。

同理，为了验证 EDU 识别中解码器 Pointer Network 模型的效果，本文分别使用条件随机场和 Pointer Network 作为解码器进行了对比实验，实验结果如表 5 所示。

表 6 两种不同 EDU 识别方法对比

System	P/%	R/%	F/%
BiLSTM+GCN+PN	89.05	91.31	**90.15**
BiLSTM+GCN+CRF	87.82	91.32	89.53

表 6 中的结果显示了两种模型的 EDU 识别性能。在两者使用相同参数的情况下，Pointer Network 解码器优于 CRF。

4 总结

本文提出一种基于多任务学习的 EDU 与主述位联合识别方法，其基本思想是利用辅助任务--主位、述位的自动识别，帮助提高主任务--EDU 的自动识别性能。实验结果表明，相对于单任务学习的 EDU 自动识别，多任务学习能够有效地提升识别性能。

本文的工作主要体现在 3 个方面：（1）利用多任务学习方法提升 EDU 和主述位的识别性能。（2）采用图卷积（GCN）模型将句法信息纳入序列神经网络模型。（3）采用 Pointer Network 模型作为解码器，解决了传统模型容易受到 EDU 边界稀疏的影响、不能很好的处理可变大小输出词汇的问题。

参考文献

[1] Beaugrande Robert-Alain De, Dessler Ulrich Wolfgang. Introduction to text linguistics. London and New York: Longman Paperback, 1981.

[2] Carlson Lynn, Marcu Daniel, Okurowski Mary Ellen. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In J. van Kuppevelt and R. Smith (eds), Current Directions in Discourse. New York: Kluwer, 2003: 85-112.

[3] Prasad Rashmi, Dinesh Nikhil, Lee Alan, et al. The penn discourse treebank 2.0. In Proceedings of the International Conference on Language Resources and Evaluation. Marrakech, Morocco, 2008: 2961-2968.

[4] Caroline Sporleder, Mirella Lapata. Discourse chunking and its application to sentence compression[C]//Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, British Columbia, Canada：EMNLP，2005：257-264.

[5] Ngo Xuan Bach, Nguyen Le Minh, Akira Shimazu. A Reranking Model for Discourse Segmentation using Subtree Features[C]//Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL). Seoul, South Korea：SIGDIAL，2012：160-168.

[6] Chloe Braud, Ophelie Lacroix, Anders Søgaard. Does syntax help discourse segmentation? Not so much[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark：EMNLP，2017：2432-2442.

[7] 李艳翠，冯文贺，周国栋. 基于逗号的汉语子句识别研究[J]. 北京大学学报（自然科学版），2013，29(1)：7-14.

[8] Xue Nianwen, Yang Yaqin. Chinese sentence segmentation as comma classification[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics.Portland：ACL，2011：631-635.

[9] Meixun Jin, Mi-Young Kim, Dongil Kim, et al. Segmentation of Chinese long sentences using commas[C]//Proceedings of 3rd ACL SIGHAN Workshop.Barcelona：ACL，2004：1-8.

[10] 朱伟华. 马泰休斯（1882—1945）[J]. 国外语言学，

1987(02)：86-88.
[11] Halliday M. A. K. and Christian M. I. M. Matthiessen. An Introduction to Functional Grammar[M].Hodder Education, London, 2004.
[12] Park K, Lu X. Automatic analysis of thematic structure in written English[J]. International Journal of Corpus Linguistics, 2015, 20(1): 81-101.
[13] 奚雪峰. 汉语篇章话题结构：表示体系、资源构建及其分析研究[D]. 苏州大学博士学位论文，2017.
[14] Caruana R. Multitask learning[J]. Machine learning, 1997, 28(1): 41-75.
[15] Zhang Z, Luo P, Loy C C, et al. Facial landmark detection by deep multi-task learning[C]//European conference on computer vision. Springer, Cham, 2014: 94-108.
[16] Liu X, Gao J, He X, et al. Representation learning using multi-task deep neural networks for semantic classification and information retrieval[J]. 2015.
[17] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.
[18] Arik S Ö, Chrzanowski M, Coates A, et al. Deep voice: Real-time neural text-to-speech[C]//Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017: 195-204.
[19] 徐凡，王明文，谢旭升，等. 基于主位-述位结构理论的英文作文连贯性建模研究[J]. 中文信息学报，2016，30(01)：115-123.
[20] Bastings, J., Titov, I., Aziz, W., et al. Graph convolutional encoders for syntax-aware neural machine translation.[C]EMNLP. 2017: 1957-1967.
[21] Marcheggiani D, Titov I. Encoding sentences with graph convolutional networks for semantic role labeling[J]. arXiv preprint arXiv:1703.04826, 2017.
[22] Linzen T, Dupoux E, Goldberg Y. Assessing the ability of LSTMs to learn syntax-sensitive dependencies[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 521-535.
[23] Li J, Sun A, Joty S. SegBot: A Generic Neural Text Segmentation Model with Pointer Network[C]//IJCAI. 2018: 4166-4172.
[24] Vinyals O, Fortunato M, Jaitly N. Pointer networks[C]//Advances in Neural Information Processing Systems. 2015: 2692-2700.
[25] Lamb A M, Goyal A G A P, Zhang Y, et al. Professor forcing: A new algorithm for training recurrent networks[C]//Advances In Neural Information Processing Systems. 2016: 4601-4609.
[26] Xue-feng Xi, Guodong Zhou. Building a Chinese Discourse Topic Corpus with Micro-Topic Scheme based on Theme-Rheme Theory[C]//Proceedings of Big Data Analytics (EI),2017.
[27] 李艳翠. 汉语篇章结构表示体系及资源构建研究[D]. 苏州大学博士学位论文，2015.

葛海柱（1994—），硕士研究生，主要研究领域为自然语言处理。
E-mail：20175227029@stu.suda.edu.cn

孔芳（1977—），通信作者，博士，教授，主要研究领域为机器学习，自然语言理解、篇章分析。
E-mail：kongfang@suda.edu.cn