文章编号：1003-0077(2021)08-0028-10

基于图卷积神经网络的隐式篇章关系识别

阮慧彬，孙 雨，洪 宇，吴成豪，李 晓，周国栋

（苏州大学 计算机科学与技术学院，江苏 苏州 215006）

摘 要：隐式篇章关系识别是篇章关系识别的子任务，其挑战性在于难以学习到具有丰富语义信息和交互信息的论元表示。针对这一难点，该文提出一种基于图卷积神经网络(Graph Convolutional Network，GCN)的隐式篇章关系分类方法。该方法采用预训练语言模型 BERT(Bidirectional Encoder Representation from Transformers)编码论元以获取论元表示，再分别拼接论元表示和注意力分数矩阵作为特征矩阵和邻接矩阵，构造基于图卷积神经网络的分类模型，从而根据论元自身信息以及交互信息对论元表示进行调整，以得到有助于隐式篇章关系识别的论元表示。该文利用宾州篇章树库(Penn Discourse Treebank，PDTB)语料进行实验，实验结果表明，该方法在四大类关系上分类性能优于基准模型 BERT，且其在偶然(Contingency)关系和扩展(Expansion)关系上优于目前先进模型，F_1 值分别达到 60.70%和 74.49%。

关键词：隐式篇章关系识别；图卷积神经网络；自注意力机制；交互式注意力机制

中图分类号：TP391 **文献标识码**：A

Graph Convolutional Network Based Implicit Discourse Relation Recognition

RUAN Huibin, SUN Yu, HONG Yu, WU Chenghao, LI Xiao, ZHOU Guodong

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

Abstract: Implicit discourse relation recognition is a challenging task in that it is difficult to obtain semantic informative and interaction-informative argument representations. The paper proposes an implicit discourse relation recognition method based on the Graph Convolutional Network (GCN). With the arguments encoded by fine-tuned BERT, the GCN is designed by concatenating the argument representations as feature matrix, and concatenating the attention score matrixes as adjacent matrix. It is hoped that the argument representations can be optimized by the self-attention and inter-attention information to improve implicit discourse relation recognition. Experimental results on the Penn Discourse Treebank (PDTB) show that the proposed method outperforms BERT in recognizing the four of implicit discourse relations, and it outperforms the state-of-the-art methods on Contingency and Expansion with 60.70% and 74.49% on F_1 score, respectively.

Keywords: implicit discourse relation recognition; graph convolutional network; self-attention mechanism; inter-attention mechanism

0 引言

篇章关系识别旨在研究同一篇章内两个文本片段(短语、子句、句子或段落，简称论元)间的逻辑关系。作为自然语言处理(natural language processing，NLP)领域的一项基础研究，篇章关系识别在上层自然语言处理应用中具有重要价值[1]，如情感分析[2-3]、机器阅读理解[4]、文摘提取[5]和机器翻译[6-8]等。篇章关系识别的任务框架如图 1 所示，给定一个论元对(Arg1，Arg2)，使用篇章关系分类模型来识别两者间的篇章关系。

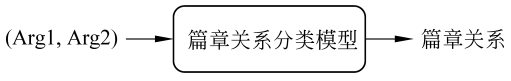

图 1 任务框架

收稿日期：2019-01-08 **定稿日期**：2020-04-08

基金项目：国家自然科学基金(61672368，61751206，61703293)

目前，篇章关系识别研究领域最大的权威语料库是宾州篇章树库[9]（penn discourse treebank，PDTB），其根据不同粒度，将篇章关系定义为一个三层的语义关系类型体系。其中，最顶层的四类语义关系是：比较（comparison）关系、偶然（contingency）关系、扩展（expansion）关系以及时序（temporal）关系。同时，根据两个论元表述间是否具有连接词（也称为线索词，如“because（因为）”等）作为衔接手段，PDTB 将篇章关系分为两类：显式篇章关系（explicit discourse relation）和隐式篇章关系（implicit discourse relation）[1]。其中，显式篇章关系是可直接通过显式连接词推理得到的篇章关系类型。如例 1 所示，此显式偶然关系论元对包含显式连接词“so（所以）”，这一线索指明 Arg2 是由 Arg1 导致的结果，因此，我们可直接推理出例 1 中的论元对具有偶然关系。

例 1［**Arg1**］：and will take measures

（译文：并将采取措施）

［**Arg2**］：**so** this kind of thing doesn't happen in the future

（译文：**所以**这类事情不会再发生）

［**篇章关系**］：Contingency.Cause.Result

相对地，隐式篇章关系论元对中缺少显式连接词，所以其更依赖于词法、句法、语义以及上下文等特征，如下述例 2 中的“hurricane（飓风）”是需要落实“precautionary mechanisms（预防机制）”的原因，因此，可推导出此论元对包含的篇章关系为偶然关系。

例 2［**Arg1**］：With a hurricane you know it's coming

（译文：你知道飓风将要来了）

［**Arg2**］：You have time to put precautionary mechanisms in place

（译文：你有时间把预防措施落实到位）

［**篇章关系**］：Contingency.Cause.Result

显式篇章关系研究目前已取得较高分类性能，Pitler 等[10]采用显式连接词与篇章关系的映射即可达到 93.09％的准确率。然而，隐式篇章关系识别性能相对较低，现有最优方法在四大类关系上的 F_1 值仅达 53％[11]。因此，本文针对隐式篇章关系识别任务展开研究。

前人将注意力机制用于论元表示的计算[12-16]，来评估论元间词义信息的关联性，借以捕获重要的词义特征来辅助隐式篇章关系识别。然而，相关研究仅关注论元自身或论元间的词义特征关联性，因此，这种单一特征无法全面地表征论元语义信息。若仅关注论元交互信息，如例 3 中的词对信息“good-wrong（好的-错误的）”和“good-ruined（好的-毁坏的）”，其很容易导致此论元对被识别为对比关系[12]。但是如果论元捕获了自身信息，关注到 Arg1 中的词“not（不）”和“good（好的）”，再结合论元间的交互信息，关注到 Arg2 中的词“ruined（毁坏的）”，那么基于词“not（不）”和“ruined（毁坏的）”的双重否定[17]可推理出此论元对包含的篇章关系为偶然关系。

例 3［**Arg1**］：Psyllium's not a good crop

（译文：车前草没有好收成）

［**Arg2**］：You get a rain at the wrong time and the crop is ruined

（译文：错误时间下的雨毁坏了庄稼）

［篇章关系］：Contingency.Cause.Reason

为了捕获论元自身信息和论元间的交互信息，借以辅助隐式篇章关系识别，本文提出了一种基于自注意力和交互式注意力机制的图卷积神经网络（self-attention and inter-attention based graph convolutional network，SIG），用于构建隐式篇章关系分类模型。此模型基于自注意力机制（self-attention）及交互式注意力机制（inter-attention）来构建邻接矩阵，因此，这一模型可利用论元自身的语义特征，同时还能够捕获论元之间的交互信息，以编码出更好的论元表征，来提升隐式篇章关系识别性能。

本文采用 PDTB 2.0[9]数据集进行实验和测试，结果证明本文所提模型 SIG 在隐式篇章关系分类上的表现优于基准模型，且其在多个关系上优于目前的隐式篇章关系识别模型。

1　相关工作

现有的隐式篇章关系识别研究主要分为两个方向：构建复杂的分类模型和挖掘大量的训练数据。其中，模型构建主要包括基于特征工程的机器学习模型以及基于论元表示的神经网络模型。

前人采用多样化的语言学特征来构建统计学习模型。在 PDTB 数据集上，Pitler 等[18]第一次尝试使用多种语言学特征对顶层四类隐式篇章关系进行识别，其实验性能超越随机分类方法；Lin 等[19]基于上下文特征、词对特征、句法结构特征以及依存结构

特征设计篇章关系识别模型；Rutherford 和 Xue[20] 提取布朗聚类特征来缓解词对稀疏性问题。Braud 和 Denis[21] 基于浅层词汇特征，使用现有的无监督词向量，训练最大熵模型来进行隐式篇章关系分类；Lei 等[17] 挖掘每类关系的语义特征，结合话题连续性和论元来源这两种衔接手段，训练朴素贝叶斯模型，在四路分类上达到 47.15% 的 F_1 值，其性能超过大部分现有的神经网络模型。

现今的隐式篇章关系识别研究大多构建复杂的神经网络模型来提升分类性能。Ji 和 Eisenstein[22] 基于论元以及实体片段的向量表示，使用两个递归神经网络(recursive neural network，RNN)进行隐式篇章关系识别。Zhang 等[23] 提出了仅包含一个隐藏层的浅层卷积神经网络，避免了过拟合问题；Chen 等[12] 基于双向长短时记忆网络(bidirectional long short-term memory network，Bi-LSTM)获取词向量表征，使用门控相关网络(gated relevance network)捕捉词对间的语义交互信息。Qin 等[24] 在卷积神经网络的基础上，增加了门控神经网络(gated neural network，GNN)来捕捉论元之间的交互信息(如词对)；Lan 等[16] 采用基于多任务注意力机制的神经网络模型，使用未标注外部语料库 BLLIP 生成伪隐式篇章关系语料，来识别隐式篇章关系，将其作为辅助任务以提升 PDTB 隐式篇章关系识别性能。Bai 和 Zhao[13] 构造了复杂的论元表征模型，融合不同粒度词向量、卷积、递归、残差和注意力机制抽取论元特征；Nguyen 等[11] 采用了 Bai 和 Zhao[13] 的模型，此外，基于知识迁移对关系表示及连接词表示进行映射，使其处于同一向量空间，从而辅助隐式篇章关系识别。

针对隐式篇章关系语料不足的问题，前人使用不同手段来扩充 PDTB 的隐式语料。朱等[25] 通过论元向量，从其他数据资源里挖掘在语义和关系上与原始语料一致的实例；Wu 等[26] 发现双语语料中存在显隐式不匹配的情况，即英文语料中没有连接词，但其对应的中文语料中却有显式连接词，基于此，Wu 等[26] 从 FBIS 和 HongKong Law 语料库中提取了伪隐式篇章关系语料；Xu 等[27] 用显式篇章关系语料构造伪隐式样例，基于主动学习方法挑选含高信息量的样例，来扩充隐式篇章关系语料；Ruan 等[28] 采用问答语料库中的 WHY 式问答对，基于“问句陈述句转换”生成伪隐式论元对，以扩充隐式因果关系语料。

2 方法

本文提出的基于自注意力和交互式注意力的图卷积神经网络(self-attention and inter-attention based graph convolutional network，SIG)框架如图 2 所示。

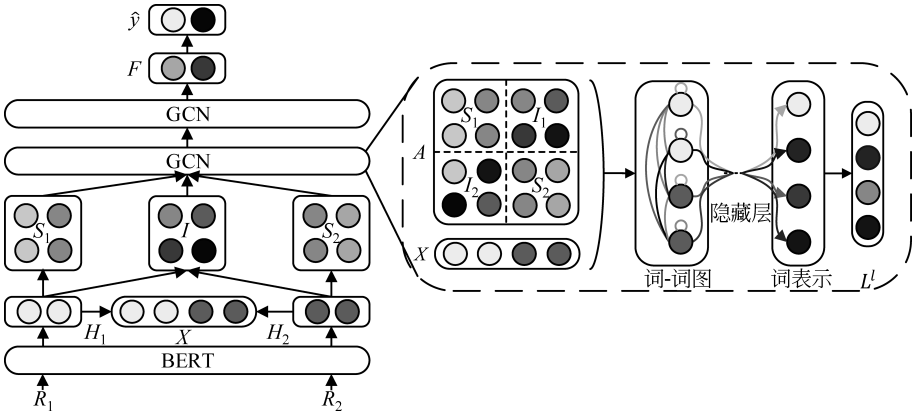

图 2 SIG 模型框架图

首先，通过微调的 BERT 语言模型[29] 获取两个论元的论元表示；其次，通过拼接构造特征矩阵和邻接矩阵，从而得到全连接的“词-词”图，作为图卷积神经网络(graph convolutional network，GCN)的初始特征，通过双层 GCN 的隐藏层对词特征进行卷积和非线性变换操作，以得到最终的词表示；最后，将词表示送入全连接层进行降维，并使用 softmax 函数对其进行归一化，得到最终分类结果。

2.1 向量表示层

给定论元表示 $R_1=(x_1^1,x_2^1,\cdots,x_L^1)$ 和 $R_2=(x_1^2,x_2^2,\cdots,x_L^2)$，本文使用微调的 BERT 预训练语

言模型对其进行编码。具体地，拼接 R_1 和 R_2 作为模型输入，以得到论元分布式表示 $\boldsymbol{H}=(h_1,h_2,\cdots,h_{2L+3})$，其中，$h_i\in\mathbb{R}^{d_k}$ 表示拼接后第 i 个词经过 BERT 编码后的向量表示。最后，根据论元最大长度 L 从 H 中截取出编码后的论元表示 $\boldsymbol{H}_1$ 和 $\boldsymbol{H}_2$，具体计算如式(1)～式(3)所示。

$$\boldsymbol{H}=\mathrm{BERT}([\mathrm{CLS},R_1,\mathrm{SEP},R_2,\mathrm{SEP}]) \tag{1}$$

$$\boldsymbol{H}_1=(h_2,h_3,\cdots,h_{L+1}) \tag{2}$$

$$\boldsymbol{H}_2=(h_{L+3},h_{L+4},\cdots,h_{2L+2}) \tag{3}$$

其中，CLS 为专用分类符号，可使用其经过 BERT 编码的向量表示，作为整个输入序列的向量表示；SEP 为专用符号，用于分隔输入序列中的两个论元。

2.2　图卷积神经网络

本节简单介绍图卷积神经网络，这一模型架构由 Kipf 和 Welling[30] 设计并在 2016 年提出，可对图结构数据直接进行计算。具体地，给定一个图 $G=(V,E)$，V 是包含 N 个节点的顶点集，E 是包括自循环边(即每个顶点都与自身相连)的边集。Kipf 和 Welling[30] 使用 $\boldsymbol{X}\in\mathbb{R}^{N\times d_k}$ 作为特征矩阵，其中，每个节点的特征维度为 d_k，矩阵中第 i 行向量 $\boldsymbol{x}_{v_i}\in\mathbb{R}^{d_k}$ 表示第 i 个节点 v_i 的特征。其邻接矩阵 $\boldsymbol{A}\in\mathbb{R}^{N\times N}$ 中的元素 a_{ij} 表示图中第 i 个节点与第 j 个节点间是否存在连接。一般情况下，若两个节点之间存在连接，则 a_{ij} 值为 1，否则为 0[31]。在实际应用中，多层 GCN 表现往往优于单层，由于其可融合更广范围的节点信息。具体地，第 l 层对第 $l-1$ 层的输出进行编码，计算如式(4)所示[31]。

$$\boldsymbol{L}^l=f(\boldsymbol{A}\boldsymbol{L}^{l-1}\boldsymbol{W}_l+\boldsymbol{b}_l) \tag{4}$$

其中，$\boldsymbol{W}_l\in\mathbb{R}^{d_k\times d_k}$ 是可学习的参数矩阵，$\boldsymbol{b}_l\in\mathbb{R}^{d_k}$ 是偏置项，f 为激活函数，其可对输出进行非线性变换。l 表示 GCN 的层数，第 0 层的 GCN 输出为节点特征矩阵 $\boldsymbol{X}$，即 $\boldsymbol{L}^0=\boldsymbol{X}$。

图卷积神经网络通过共享参数 $\boldsymbol{W}_l$ 对特征矩阵进行卷积操作。由于共享局部参数，GCN 在一定程度上能够防止过拟合。在对文本进行处理时，构建以词特征表示为节点的 GCN，则可通过节点 A 感受野范围内的邻居节点来对节点 A 的语义特征向量进行更新，以得到包含邻居节点语义信息的特征表示。

2.3　基于自注意力和交互式注意力的图卷积层

本文使用多层 GCN 对论元表示矩阵进行更新。具体地，拼接两个编码后的论元表示作为节点特征矩阵，同时，拼接论元的注意力分数矩阵来构造邻接矩阵。

• 节点特征矩阵

给定两个编码后的论元表示 $\boldsymbol{H}_1$ 和 $\boldsymbol{H}_2$，本文将其拼接作为节点特征矩阵 $\boldsymbol{X}\in\mathbb{R}^{2L\times d_k}$，即 $\boldsymbol{X}=[\boldsymbol{H}_1,\boldsymbol{H}_2]$。在此基础上，可对两个论元表示同时进行图卷积操作，借以得到富含论元自身信息和交互信息的特征矩阵。

• 邻接矩阵

考虑到篇章关系依赖于深层次的文本理解和论元间的信息交互，本文基于论元的自注意力分数矩阵和交互式注意力分数矩阵，来构造图卷积神经网络的邻接矩阵，以得到一个以论元表示为节点的全连接图。下面分别介绍本文所用的自注意力机制和交互式注意力机制的计算方法。

本文对论元表示 $\boldsymbol{H}_1$ 和 $\boldsymbol{H}_2$ 分别使用自注意力机制[32]，来衡量其自身每个单词表示的重要程度，以得到论元的自注意力分数矩阵 $\boldsymbol{S}\in\mathbb{R}^{L\times L}$。以 Arg1 为例，具体计算如式(5)～式(7)所示。

$$\boldsymbol{Q}_1=\boldsymbol{H}_1\boldsymbol{W}_{Q1} \tag{5}$$

$$\boldsymbol{K}_1=\boldsymbol{H}_1\boldsymbol{W}_{K1} \tag{6}$$

$$\boldsymbol{S}_1=\mathrm{softmax}\left(\frac{Q_1K_1^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{7}$$

其中，$\boldsymbol{W}_{Q1}\in\mathbb{R}^{d_k\times d_k}$ 和 $\boldsymbol{W}_{K1}\in\mathbb{R}^{d_k\times d_k}$ 是可学习的参数矩阵，以 $\sqrt{d_k}$ 为分母防止内积过大。同理，可计算得到 Arg2 的自注意力权重分布矩阵 $\boldsymbol{S}_2$。

同时，在得到两个论元的向量表示 $\boldsymbol{H}_1$ 和 $\boldsymbol{H}_2$ 后，本文对其使用交互式注意力机制[12]，来计算得到论元对的交互注意力矩阵 $\boldsymbol{I}\in\mathbb{R}^{L\times L}$。具体地，对 $\boldsymbol{I}$ 进行归一化可得到 Arg1 对 Arg2 中每个词的交互式注意力分数 $\boldsymbol{I}_1$，同理，对 $\boldsymbol{I}^{\mathrm{T}}$ 进行归一化可得到 Arg2 对 Arg1 中每个词的交互式注意力分数 $\boldsymbol{I}_2$，具体计算如式(8)～式(10)所示。

$$\boldsymbol{I}=\boldsymbol{H}_1\boldsymbol{W}_I\boldsymbol{H}_2^{\mathrm{T}} \tag{8}$$

$$\boldsymbol{I}_1=\mathrm{softmax}(\boldsymbol{I}) \tag{9}$$

$$\boldsymbol{I}_2=\mathrm{softmax}(\boldsymbol{I}^{\mathrm{T}}) \tag{10}$$

其中，可学习的参数矩阵 $\boldsymbol{W}_I\in\mathbb{R}^{d_k\times d_k}$ 是 Arg1 和 Arg2 信息交互的媒介。

通过上述计算可得到自注意力分数矩阵 $\boldsymbol{S}_1$ 和 $\boldsymbol{S}_2$ 以及交互式注意力分数矩阵 $\boldsymbol{I}_1$ 和 $\boldsymbol{I}_2$。基于此，本文拼接 $\boldsymbol{S}_1$、$\boldsymbol{S}_2$、$\boldsymbol{I}_1$ 和 $\boldsymbol{I}_2$，以得到融合论元自身信息和交互信息的邻接矩阵 $\boldsymbol{A}\in\mathbb{R}^{2L\times 2L}$，具体拼接方式如式(11)所示。

$$\boldsymbol{A}=\begin{pmatrix}\boldsymbol{S}_1 & \boldsymbol{I}_1\\ \boldsymbol{I}_2 & \boldsymbol{S}_2\end{pmatrix} \tag{11}$$

• **图卷积操作**

基于以上公式得到图卷积神经网络的节点特征矩阵 $\boldsymbol{X}$ 和邻接矩阵 $\boldsymbol{A}$，我们参照公式(4)来计算节点特征矩阵 $\boldsymbol{X}$ 的图卷积特征[31]，此处采用的 GCN 层数为 2，具体计算如式(12)所示。

$$\boldsymbol{L}^2=f(\boldsymbol{A}f(\boldsymbol{AXW}_1+\boldsymbol{b}_1)\boldsymbol{W}_2+\boldsymbol{b}_2) \tag{12}$$

2.4 全连接层

本文通过多层 GCN 得到更新后的特征表示 $\boldsymbol{L}^l=\{\boldsymbol{g}_1^l,\boldsymbol{g}_2^l,\cdots,\boldsymbol{g}_{2L}^l\}$，其中，$\boldsymbol{g}_i^l\in\mathbb{R}^{d_k}$ 表示第 l 层 GCN 更新得到的第 i 个节点的特征表示。本文对最后一层 GCN 输出的每个节点的特征表示求和，以得到最终的论元对特征表示 $\boldsymbol{F}\in\mathbb{R}^{d_k}$，具体计算如式(13)所示。

$$\boldsymbol{F}=\sum_{i}^{2L}\boldsymbol{g}_i^l \tag{13}$$

通过将 $\boldsymbol{F}$ 输入全连接层，计算 Arg1 和 Arg2 间具有关系 r 的概率，具体计算如式(14)所示。

$$\hat{\boldsymbol{y}}=\mathrm{softmax}(\boldsymbol{WF}^{\mathrm{T}}+\boldsymbol{b}) \tag{14}$$

其中，$\boldsymbol{W}\in\mathbb{R}^{n\times d_k}$，$\boldsymbol{b}\in\mathbb{R}^n$ 是可学习的参数，$\boldsymbol{W}$ 可对最终特征表示 $\boldsymbol{F}$ 进行降维。$\hat{\boldsymbol{y}}\in\mathbb{R}^n$ 是预测此论元对是否具有关系 r 的概率。

2.5 训练

本文为 PDTB 语料四大类关系中的每一类分别构造一个二分类器。在训练过程中，本文采用交叉熵损失函数作为目标函数，使用 Adam[34] 优化算法更新所有模型参数。对于给定论元对(Arg1，Arg2)及其关系标签 y_i，其损失函数计算如式(15)所示。

$$L(y,\hat{\boldsymbol{y}})=-\sum_{i=1}^{n}y_i\log(\hat{\boldsymbol{y}}_i) \tag{15}$$

其中，$\hat{\boldsymbol{y}}_i$ 指论元对间是否具有关系 r 的概率，由于本文采用了 softmax 激活函数，因此，$\hat{\boldsymbol{y}}_i>0$ 且 $\sum_{i=1}^{n}\hat{\boldsymbol{y}}_i=1$。$y_i\in[0,1]$ 是指明论元对是否具有关系 r 的真实标签。n 表示类别数量。

3 实验

3.1 实验数据

本文在宾州篇章树库[9](penn discourse treebank，PDTB)语料上用 SIG 模型进行隐式篇章关系识别实验。PDTB 由 Prasad 等在 2008 年提出，其来源于《华尔街日报》(Wall Street Journal，WSJ)的 2 304 篇文章，共标注了 40 600 个篇章关系样本，其中，隐式篇章关系实例占 16 224 个[1]。为了与前人工作保持一致，本文以 section 02-20 为训练集，section 00-01 为开发集，section 21-22 为测试集。顶层四类语义关系 Comparison(COM.)、Contingency(CON.)、Expansion(EXP.)和 Temporal(TEM.)的数据分布如表 1 所示[1]。

表 1 PDTB 四大类隐式篇章关系数据分布

关系类型	训练集	开发集	测试集
COM.	1 855	189	145
CON.	3 235	281	273
EXP.	6 673	638	538
TEM.	582	48	55
总计	12 345	1 156	1 011

由表 1 可知，在 PDTB 数据集中，除 EXP.之外的其他三类篇章关系数据量都较少[18]，类间不平衡问题使得研究者通常为各个关系类型单独训练二分类器[11-17,23-24]以进行评估。所以，本文参照前人工作，基于不同篇章关系的训练集分别训练二分类模型，一共得到 4 个二分类器，分别用于判断样例是否包含该篇章关系，并通过 F_1 值对其性能进行评估。本文跟随前人工作[11-17,23,24,36]未对同一个样例的四次二分类结果进行整合，在进行二分类时仅讨论单类篇章关系的是或否问题。此外，由于 PDTB 数据集存在正负例样本不均衡的问题，本文对负例进行随机下采样[24]，来构造正负例平衡的训练数据集。同时，为了更好地与前人工作进行比较，本文在 PDTB 数据集上进行了四路分类实验，基于训练集训练一个四分类器，并采用 Macro-F_1 值和准确率(Accuracy)对其进行评估。

3.2 实验设置

为了证明使用 GCN 融合自注意力和交互式注意力机制有助于隐式篇章关系识别，本文设置了以下六个对比系统。

• **BERT**(Baseline)：通过微调 BERT 模型得到 Arg1 和 Arg2 的隐层输出后，分别对其进行裁剪，以得到两个论元表征。然后通过逐词求和获取句级论元表征，通过拼接这两个句级表征得到最终特征，

并输入至全连接层进行分类。

• **Self**：使用 BERT 获取 Arg1 及 Arg2 的论元表征后，分别计算其自注意力分数，并将自注意力权重作用到论元表征上；然后分别对更新后的论元表征进行逐词求和，以获取句级表征；最后拼接句级表征作为全连接层的输入。

• **Inter**：得到 BERT 输出的论元表征后，对其使用交互式注意力机制，来获取交互式注意力权重分布矩阵，并作用于论元表征；然后再通过对新论元表征逐词求和及拼接得到句对级论元表征，输入全连接层进行隐式篇章关系分类。

• **Concatenate**：通过拼接上述 **Self** 和 **Inter** 系统分别生成的句级表征得到句对级论元表征，输入到全连接层进行隐式篇章关系分类。

• **Transformer**：拼接通过 BERT 编码得到的 Arg1 和 Arg2 的论元表征，作为具有 8 头注意力机制的双层 Transformer[32] 的输入，再对 Transformer 编码后的词特征进行逐词求和，借以得到论元对的句级表征，将其输入到全连接层进行隐式篇章关系分类。

• **SIG**：使用 BERT 得到 Arg1 和 Arg2 的论元表征后，分别计算两者的自注意力权重分布矩阵和交互式注意力权重分布矩阵；然后拼接两个论元表征得到特征矩阵，再拼接注意力权重分布矩阵得到邻接矩阵，来构建双层 GCN；对最后一层 GCN 的输出进行逐词求和，以得到两个论元的句级表征，并将其输入全连接层进行隐式篇章关系分类。

3.3　参数设置

本文使用微调的 BERT[29] 隐层输出作为论元表示，其中，我们设置隐层向量维度 d_k 为 768，论元最大长度 L 为 80。基于论元表示构造的特征矩阵，本文拼接论元自注意力和交互式注意力权重分布矩阵得到邻接矩阵，构造 2 层（$l=2$）GCN 神经网络，并使用 tanh 函数作为模型的激活函数。构建 Transformer 模型时，我们采用了 Vaswani 等[32] 工作中 Transformer 的编码器作为本文的一层 Transformer。本文采用了双层 Transformer 对编码后论元表示进行变换，且设置其前馈神经网络的隐层维度为 768，并采用 GeLU[33] 作为激活函数。在训练过程中，使用交叉熵作为损失函数，采用基于 Adam[34] 的批梯度下降法优化模型参数，其中，批大小为 32，学习率为 5e-5。本文在最后一层 GCN 后进行了 dropout 计算，其随机丢弃的概率为 0.1。

3.4　实验结果

本文采用六种不同结构的神经网络模型，来分别对 PDTB 四大类隐式篇章关系进行分类，具体分类性能如表 2 所示。其中，本文所提模型 SIG 在多个关系上的表现优于其他五个对比模型。其原因主要在于 SIG 融合了两种注意力机制的优点，在关注两个论元自身信息的同时，还能够关注到两者间的交互信息，并通过这样的信息来更新论元表示。因此，SIG 能够生成更符合隐式篇章关系分类任务特性的论元表示。

表 2　不同模型在四大类篇章关系上的分类结果

（单位：%）

Models	COM.	CON.	EXP.	TEM.
BERT (Baseline)	41.14	55.67	73.39	35.34
Self	41.10	57.06	73.40	38.94
Inter	43.75	56.20	73.47	39.20
Concatenate	41.53	53.56	73.18	37.24
Transformer	46.83	56.76	**74.62**	41.60
SIG	**48.08**	**60.70**	74.49	**42.00**

然而，模型 Transformer 采用了 8 头注意力机制来捕捉多方面的论元自身信息和论元间的交互信息。但是，Transformer 在模拟论元之间的信息交互时，仅采用论元点积矩阵作为注意力分数矩阵，而 SIG 可使用的注意力机制则较为灵活，本文采用了双线性模型来模拟两个论元之间的线性交互。此外，Transformer 使用 8 头注意力机制，而 SIG 仅采用单头自注意力机制；同时，Transformer 的注意力分数矩阵在不同层数值不一致，而 SIG 中不同层的 GCN 共享同一邻接矩阵，其元素值的大小表示不同词节点之间连接的强弱；且每层 Transformer 在使用注意力机制更新论元特征后，还需使用包含两个全连接层的前馈神经网络对其进行变换，并采用了残差机制。相较之下，SIG 模型结构更为简单，在一定程度上防止了过拟合。因此，Transformer 在数据量较多的 Expansion 关系上表现优于 SIG，而在其他关系上表现稍弱。

此外，模型 Concatenate 在几乎所有篇章关系上，性能劣于 Self 和 Inter，我们认为主要由以下两方面原因造成：其一，仅拼接的方式过于简单，难以模拟两个论元之间的复杂关系和两种注意力机制间的平衡；其二，此模型存在一定的过拟合问

题。相对地,本文所提模型SIG应用GCN来权衡两种注意力机制。其中,GCN模型固有的权重共享特性在一定程度上能够防止过拟合情况的发生,因此SIG几乎能够在四类篇章关系上分类性能超越其他模型。

为了证明本文所提模型SIG的有效性,我们与现有先进模型进行了对比(见表3)。其中,Bai和Zhao[13]使用字符级(Character)、子词级(Subword)和基于ELMo[35]的词级(Word)表示构建多粒度论元表示,结合卷积操作、残差机制、交互式注意力机制和多任务学习思想构建复杂的深度神经网络。在Bai和Zhao[13]的基础上,Nguyen等[11]基于知识迁移思想,映射关系向量与连接词向量到同一向量空间。此外,Lan等[16]借助BLLIP等外部数据训练多任务模型。在同一篇章内,从上而下的篇章关系间存在一定关系,Dai和Huang[36]深入挖掘这一特点,利用集成学习的方法构造隐式篇章关系分类器。

表3 SIG与现有先进模型对比结果

(单位:%)

模型	二分类				四路分类	
	COM.	CON.	EXP.	TEM.	Macro-F_1	Acc
Zhang[23]	33.22	52.04	69.59	30.54	—	—
Chen[12]	40.17	54.76	—	31.32	—	—
Qin[24]	41.55	57.32	71.50	35.43	—	—
Liu[15]	37.91	55.88	69.97	37.17	44.98	57.27
Lan[16]	40.73	58.96	72.47	38.50	47.80	57.39
Dai[36]	46.79	57.09	70.41	**45.61**	48.82	57.44
Lei[17]	43.24	57.82	72.88	29.10	47.15	—
Guo[14]	40.35	56.81	72.11	38.65	47.59	59.06
Bai[13]	47.85	54.47	70.60	36.97	51.06	—
Nguyen[11]	**48.44**	56.84	73.66	38.60	**53.00**	—
SIG	48.08	**60.70**	**74.49**	42.00	52.51	**60.18**

相较于前人工作,本文所提模型SIG较为简单,仅使用了标准PDTB数据集进行训练,而其能在多个关系上分类性能超越目前最优方法。其原因主要在于:①BERT预训练语言模型中已含有大量先验知识,其对需要常识知识的隐式篇章关系识别具有一定帮助;②前人工作通常使用交互式注意力机制抽取论元间的交互信息,但忽略了论元自身信息的重要性,而SIG融合了自身信息以及交互信息。

表4展示了本文所用的PDTB四大类篇章关系上的词汇分布。其中,每类关系中都含有大量未登录词(Out-Of-Vocabulary,OOV)。研究者通常将这些未登录词用特殊符号"UNK"表示,并统一初始化得到一致的词向量,这虽能打破未登录词词向量查找的困境,但是其削减了一定的信息量,且对隐式篇章关系识别带来一定影响。

表4 四大类篇章关系上的词汇分布

数据集	COM.	CON.	EXP.	TEM.
训练集	17 005	22 518	30 612	10 694
测试集	6 616	6 616	6 616	6 616
未登录词	2 010	1 668	1 276	2 837

如例4中,未登录词"steamed(推进)"在训练集中未出现过,在没有词"steamed(推进)"的情况下,仅靠"paused(暂停)"和"reaching its high(到达高点)"难以推导出因果关系。然而,BERT能够使用词的上下文信息为未登录词进行词向量初始化,且"steamed forward(向前推进)"是"reaching its high(到达高点)"的原因,因此,可推导出此论元对包含的篇章关系为偶然关系。

例4 [**Arg1**]: Instead, the rally only paused for about 25 minutes and then **steamed** forward as institutions resumed buying

(译文:反而,股票价格的涨势仅停了25分钟左右,然后股价的涨势便随着机构恢复购买股票而加速前进)

[**Arg2**]: The market closed minutes after reaching its high for the day of

(译文:股市在股票交易量达到当日高点的几分钟后就关闭了)

[**篇章关系**]: Contingency.Cause.Result

为了证明模型SIG的有效性,本文使用模型Self、Inter和SIG对例3进行注意力权重分布计算,并对注意力权重数值逐词求平均来绘制灰度色块,分别获取三个模型通过例3计算得到的注意力分布灰度图(图3)。由图3可知,模型Self和SIG都关注到了Arg1中的单词"not(不)"和"good(好的)"。然而,只有模型SIG对Arg2中的单词"ruined(毁坏的)"赋予了较高权重。因此,模型SIG能够通过单词"not(不)"和"ruined(毁坏的)"的双重否定[17]来推理得到这两个论元之间包含的隐式篇章关系为偶然关系。

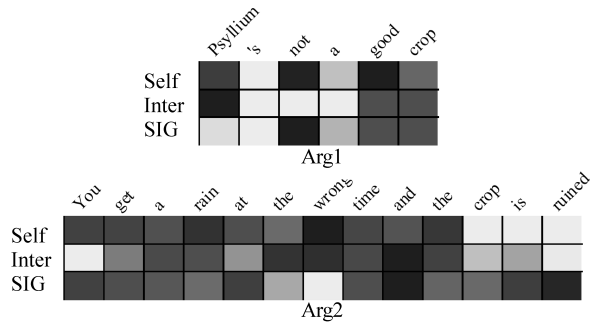

图 3　例 3 由不同系统得到的注意力分布灰度图

本文对使用不同层数 GCN 构造的模型进行了实验，其性能如表 5 所示。

表 5　基于不同层数 GCN 的模型分类性能

（单位：%）

层数	二分类				四路分类	
	COM.	CON.	EXP.	TEM.	Macro-F_1	Acc
GCN1	44.27	57.99	74.18	41.70	50.38	57.49
GCN2	**48.08**	**60.70**	**74.49**	**42.00**	52.51	60.18
GCN3	47.30	58.69	74.10	41.74	53.47	**61.28**
GCN4	47.43	56.89	74.14	41.70	**53.86**	59.48
GCN5	44.29	57.18	73.84	39.82	52.45	59.28
GCN6	44.60	56.47	73.58	38.42	51.43	59.68

其中，在 GCN 层数为 2 时（即 GCN2），二分类器在 F_1 值上达到最大值，而在 GCN 层数为 4 时，四路分类的 Macro-F_1 值和准确率分别是 53.86% 和 59.48%。这主要是由于二分类模型训练集样本量低于四分类模型，因此，当 GCN 层数较多时，二分类器易于出现过拟合现象。

4　结论

本文针对隐式篇章关系识别展开研究，提出了基于自注意力和交互式注意力机制的图卷积神经网络模型，用以对隐式篇章关系进行识别。实验结果表明，本文所提模型 SIG 表现优于基准模型 BERT，且其在多类关系上性能优于现有先进方法。

从实验结果可知，隐式篇章关系识别任务仍具有极大挑战性，除 EXP.外的其他三大类关系的分类性能皆较低，远达不到实际应用需求。下一步工作中，我们将从两个方面展开研究：①针对数据不平衡问题，从外部挖掘高质量的隐式篇章关系语料；②构建更复杂且符合隐式篇章关系识别任务特性的分类模型。

参考文献

[1] 阮慧彬. 基于数据增广与论元表征的隐式篇章关系识别方法研究[D].苏州：苏州大学硕士学位论文，2020.

[2] Somasundaran S，Namata G，Wiebe J，et al. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing，2009：170-179.

[3] Zhou L，Li B，Gao W，et al. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities[C] //Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics，2011：162-171.

[4] Narasimhan K，Barzilay R. Machine comprehension with discourse relations[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing，2015：1253-1262.

[5] Yoshida Y，Suzuki J，Hirao T，et al. Dependency-based discourse parser for single-document summarization[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing，2014：1834-1839.

[6] Meyer T，Popescu-Belis A. Using sense-labeled discourse connectives for statistical machine translation[C]//Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation and Hybrid Approaches to Machine Translation. Association for Computational Linguistics，2012：129-138.

[7] Xiong D，Ding Y，Zhang M，et al. Lexical chain based cohesion models for document-level statistical machine translation[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing，2013：1563-1573.

[8] Meyer T，Webber B. Implicitation of discourse connectives in machine translation[C]//Proceedings of the Workshop on Discourse in Machine Translation，2013：19-26.

[9] Prasad R，Dinesh N，LeeA，et al. The Penn Discourse TreeBank 2.0[C]//Proceedings of the International Conference on Language Resources and Evaluation，2008：2961-2968.

[10] Pitler E, Raghupathy M, Mehta H, et al. Easily identifiable discourse relations[R]. Technical Reports (CIS), 2008: 884.

[11] Linh The Nguyen, Linh Van Ngo, Khoat Than, et al. Employing the correspondence of relations and connectives to identify implicit discourse relations via label embeddings[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,2019: 4201-4207.

[12] Chen J, Zhang Q, Liu P, et al. Implicit discourse relation detection via a deep architecture with gated relevance network[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016, 1: 1726-1735.

[13] Bai H, Zhao H. Deep enhanced representation for implicit discourse relation recognition[J]. arXiv preprint arXiv: 1807.05154, 2018.

[14] Guo F, He R, Jin D, et al. Implicit discourse relation recognition using neural tensor network with interactive attention and sparse learning[C]//Proceedings of the 27th International Conference on Computational Linguistics, 2018: 547-558.

[15] Liu Y, Li S, Zhang X, et al. Implicit discourse relation classification via multi-task neural networks [C]//Proceedings of the AAAI Conference on Artificial Intelligence,2016: 2750-2756.

[16] Lan M, Wang J, Wu Y, et al. Multi-task attention-based neural networks for implicit discourse relationship representation and identification[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2017: 1299-1308.

[17] Lei W, Xiang Y, Wang Y, et al. Linguistic properties matter for implicit discourse relation recognition: Combining semantic interaction, topic continuity and attribution[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018.

[18] Pitler E, Louis A, Nenkova A. Automatic sense prediction for implicit discourse relations in text[C]// Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Association for Computational Linguistics, 2009: 683-691.

[19] Lin Z, Kan M Y, Ng H T. Recognizing implicit discourse relations in the Penn Discourse Treebank [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2009: 343-351.

[20] Rutherford A, Xue N. Discovering implicit discourse relations through brown cluster pair representation and coreferencepatterns[C]//Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014: 645-654.

[21] Braud C, Denis P. Comparing word representations for implicit discourse relation classification[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2015: 2201-2211.

[22] Ji Y, Eisenstein J. One vector is not enough: Entity-augmented distributional semantics for discourse relations[J]. arXiv preprint arXiv: 1411.6699, 2014.

[23] Zhang B, Su J, Xiong D, et al. Shallow convolutional neural network for implicit discourse relation recognition[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2015: 2230-2235.

[24] Qin L, Zhang Z, Zhao H. A stacking gated neural architecture for implicit discourse relation classification [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016: 2263-2270.

[25] 朱珊珊，洪宇，丁思远等. 基于训练样本集扩展的隐式篇章关系分类[J]. 中文信息学报，2016，30(5)：111-120.

[26] Wu C, Chen Y, Huang Y. Bilingually-constrained synthetic data for implicit discourse relation recognition[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016: 2306-2312.

[27] Xu Y, Hong Y, Ruan H, et al. Using active learning to expand training data for implicit discourse relation recognition[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018: 725-731.

[28] Ruan H, Hong Y, Sun Y, et al. Using WHY-type-question-answer pairs to improve implicit causal relation recognition[C]//Proceedings of the International Conference on Asian Language Processing, 2019: 355-360.

[29] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171-4186.

[30] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv: 1609.02907, 2016.

[31] Marcheggiani D, Titov I. Encoding sentences with graph convolutional networks for semantic role labeling[J]. arXiv preprint arXiv: 1703.04826, 2017.

[32] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of Advances in Neural

Information Processing Systems，2017：5998-6008.
[33] Hendrycks D，Gimpel K. Gaussian error linear units (gelus)[J].arXiv preprint arXiv：1606.08415，2016.
[34] Kingma D，Ba J. Adam：A Method for stochastic optimization[J].arXiv preprint arXiv：1412.6980，2014.
[35] Peters M E，Neumann M，Iyyer M，et al. Deep contextualized word representations[J]. arXiv preprint arXiv：1802.05365，2018.
[36] Dai Z，Huang R. Improving Implicitdiscourse relation classification by modeling inter-dependencies of discourse units in a paragraph[J]. arXiv preprint arXiv：1804.05918，2018.

阮慧彬(1995—)，硕士研究生，主要研究领域为篇章分析。
E-mail：huibinnguyen@gmail.com

孙雨(1996—)，硕士研究生，主要研究领域为篇章分析。
E-mail：sunyu41679@gmail.com

洪宇(1978—)，通信作者，博士，教授，主要研究领域为信息检索、信息抽取。
E-mail：tianxianer@gmail.com

CAIL 2021 | 中国法律智能技术评测正式开启

法律智能研究旨在赋予机器理解法律文本的能力。近年来，随着以裁判文书为代表的司法大数据不断公开，以及自然语言处理技术的不断突破，如何将人工智能技术应用在司法领域，辅助司法工作者提升案件处理的效率和公正性，逐渐成为法律智能研究的热点。CAIL 旨在为研究者提供交叉学科的学术交流平台，推动自然语言理解与处理、智能信息检索等人工智能技术在法律领域的应用，共同促进中国法律智能技术的创新发展，为科技赋能社会治理作出贡献。

为了促进智能技术赋能司法，实现更高水平的数字正义，在最高人民法院和中国中文信息学会的指导下，从 2018 年起，CAIL 已连续举办了三届中国法律智能技术评测，先后吸引了来自海内外高校、企业和组织的 3010 支队伍参赛，成为中国法律智能技术评测的重要平台。CAIL 2018 年设置了罪名预测、法条推荐、刑期预测三个任务，并提供了包含 268 万刑事法律文书的数据集；CAIL 2019 年设置了阅读理解、要素识别、相似案例匹配三个任务；CAIL 2020 年设置了阅读理解、司法摘要、司法考试、论辩挖掘四个任务。随着智能技术与法律需求交叉融合的不断深入，CAIL 的任务设置更加符合司法需求，任务难度也逐年升级。

CAIL 2021 一共设置了七个任务，分别为：**阅读理解、类案检索、司法考试、司法摘要、论辩理解、案情标签预测以及信息抽取**，同时将提供海量司法文书数据作为数据集。CAIL 2021 已于 2021 年 8 月 1 日全面开启报名通道，总体赛程将持续至 2021 年 12 月初，并预计于 2021 年 12 月在北京举办颁奖会暨法律智能技术研讨会。七个任务的具体赛程安排敬请关注 CAIL 官网：http://cail.cipsc.org.cn/。诚邀学术界、工业界的研究者与开发者积极参与和支持评测！