

基于语境交互感知和模式筛选的隐式篇章关系识别

郭凤羽^{1),2)} 贺瑞芳^{1),2)} 党建武^{1),2),3)}

¹⁾(天津大学智能与计算学部 天津 300350)

²⁾(天津市认知计算与应用重点实验室 天津 300350)

³⁾(日本北陆先端科学技术大学院大学 石川县 日本 9231292)

摘 要 隐式篇章关系识别是篇章分析(Discourse Analysis)中一项具有挑战性的子任务,旨在推断出同一篇章内相邻文本片段(称为论元)之间潜在的语义连接关系,例如:时序关系、因果关系等.如何有效地表征篇章论元以及挖掘论元间的交互信息是实现该任务的核心要素.传统研究注重篇章中人工总结的表层语言特征(即情感极性、位置特征和动词类型等),存在数据稀疏和预处理错误级联的问题,导致机器学习模型性能不高.新近的深度神经网络模型则自动提取篇章论元中的特征,利用注意力或记忆机制等捕获论元的重要信息,并组合不同神经网络提取大量关系特征,以提升模型识别性能.然而,其忽略了表示过程中论元间双向非对称的交互信息,以及识别过程中论元间交互模式的稀疏特性.受认知学相关理论的启发,本文提出基于语境交互感知和模式筛选的隐式篇章关系识别方法(MATS).首先,通过双向长短期记忆网络(Bidirectional Long Short-Term Memory, BiLSTM)分别编码两个论元,以获取带有上下文语境信息的论元表示;其次建模其动态交互注意力机制,以自动学习论元之间的非对称关联矩阵,进而得到融合语境交互感知信息的论元表示;最后,利用带有稀疏约束的张量神经网络捕捉具有篇章关系指示性的深层交互模式,从而提升模型的识别性能. Penn Discourse Treebank (PDTB)语料库上的实验结果表明,本文提出方法的精确率在其四分类上改善了2.36%.

关键词 隐式篇章关系识别;双向长短期记忆网络;交互注意力机制;稀疏约束;张量神经网络

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2020.00901

Implicit Discourse Relation Recognition Based on Contextual Interaction Perception and Pattern Filtering

GUO Feng-Yu^{1),2)} HE Rui-Fang^{1),2)} DANG Jian-Wu^{1),2),3)}

¹⁾(College of Intelligence and Computing, Tianjin University, Tianjin 300350)

²⁾(Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin 300350)

³⁾(Japan Advanced Institute of Science and Technology, Ishikawa 9231292, Japan)

Abstract Recognizing implicit discourse relation is a challenging sub-task in discourse analysis, which aims to understand and annotate the latent relations between two adjacent text units (called discourse arguments), such as temporal, comparison, etc. It is beneficial to many downstream natural language processing (NLP) applications, e. g., machine translation, information retrieval, conversation system and so on. The key factors lie in appropriately representing two discourse arguments as well as modeling their interactions. Traditional methods focus on surface features of discourse (for instance, position feature, sentiment polarity, syntactic pattern, and so forth.), rule-based method and statistical algorithms. These approaches heavily rely on hand-crafted features to obtain rich linguistic information extracted from two discourse arguments, which usually suffer

收稿日期:2019-08-23;在线出版日期:2020-02-06. 本课题得到国家自然科学基金(61976154)、天津市自然科学基金(18JCYBJC15500)、国家重点研发计划项目(2018YFB1305200)、天津市科技项目(18ZXZNGX00330)资助. 郭凤羽,博士研究生,中国计算机学会(CCF)学生会员,主要研究方向为自然语言处理、篇章关系识别. E-mail: fengyuguo@tju.edu.cn. 贺瑞芳(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为自然语言处理、社交媒体挖掘及机器学习. E-mail: rfhe@tju.edu.cn. 党建武,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为语音识别、语音对话、言语认知脑机理、言语理解.

from data sparsity problem and are prone to error propagation due to the unsuitable preprocessing. Although previous basic neural network-based models could capture the semantic information to some extent by learning dense vector representation, and alleviate the data sparsity problem simultaneously, they model two discourse arguments separately and neglect the pair specific clues. Various complicated neural network architectures with different attention or memory mechanisms are proposed to receive the significant information of the arguments. However, they ignore the bidirectional asymmetry interactions between the arguments in the representation process. Further researches explore to mine more complicated semantic features which may include some redundant and noisy information. These studies rarely consider the sparsity of pair patterns indicating discourse relation in the identification stage. In addition, inspired by relevant cognitive linguistic theories, we have a similar feeling intuitively during judging the discourse relation; a more reasonable strategy is that people may read two arguments back and forth, and find some relevant and indicative clues to make a decision. Therefore, we propose a novel Mutual Attention-based Neural Tensor Network with Sparse Constraint framework (MATS) to excavate the specific indicative pair patterns. Specifically, we model two discourse arguments by basic bidirectional long short-term memory (BiLSTM) network to preserve the contextual information around the word embeddings. In order to catch the pair cues between the arguments, we integrate mutual attention mechanism into BiLSTM for dynamically obtaining the reciprocal information. It is aware of the asymmetry interactions between the two arguments, in a way that the bidirectional information from the arguments can directly influence the computation of their representations. Furthermore, we employ neural tensor network (NTN) model to capture the different aspects of semantic interactions between the arguments. Since there exists some redundant or irrelevant noise affecting the performance of the system, we design a sparse constraint for NTN model in training process, which could remove the noisy features and select the deeper and indicative pair patterns to identify discourse relations efficiently. The experimental results on Penn Discourse Treebank (PDTB) show that our MATS model is effective in implicit discourse relation recognition. It suggests that utilizing mutual attention mechanism to dynamically learn the bidirectional asymmetry connections between the arguments helps improve the semantic representations. Excavating the indicative pair patterns by adding a sparse constraint to NTN also plays an important role in our MATS model.

Keywords implicit discourse relation recognition; bidirectional long short-term memory; mutual attention mechanism; sparse constraint; neural tensor network

1 引 言

篇章关系识别是篇章分析的核心任务之一^[1],旨在自动判别同一篇章内部两个相邻或跨度在一定范围内的文本片段之间的语义连接关系,例如:时序关系、扩展关系等.这里文本片段指具有完整语义的文字序列,包括有子句、句子或句群等.该任务在自然语言处理的其它研究中有广泛的应用,例如:篇章连接词能有效地改善统计机器翻译^[2]的性能;对话系统^[3]整合篇章信息以生成具有逻辑关系的回复;

抽取式摘要^[4]可利用篇章关系选取文章中具有重要意义的句子作为候选句子等.

参照命题库^①(PropBank)中的“谓词-论元(Predicate-Arguments)”结构,宾州篇章树库^[5](Penn Discourse Treebank, PDTB)将篇章中含有语义关系的文本片段标记为“连接词-论元(Connective-Arguments)”结构.具体地,由连接词连接的两个文本片段称为论元,连接词所引导的其中一个论元记作 Arg2,另一个为 Arg1,Arg1-Arg2 作为整体论

① 在宾州树库(Penn Treebank)的基础上,对其动词的配价关系进行标注,建立了“命题树库”(Proposition Bank, PropBank).

元对。根据论元之间是否存在显式连接词,篇章关系可以分为显式和隐式篇章关系^[6]。其中,隐式篇章关系则缺乏直观明确的连接线索^[7],需要从论元对中抽取句法、语义等相关特征进行间接地分析推断。由于篇章论元本身结构的复杂性、语义的歧义性以及上下文信息的不确定性问题,使得隐式篇章关系识别难度加大,成为篇章分析领域的研究瓶颈。因此,本文专注于隐式篇章关系的识别。

传统特征工程和机器学习^[8-9]的方法大多侧重于人工标注的语言学特征及总结的规则,如从论元对中提取词性信息、布朗聚类词对^[10](Brown Cluster Pairs)和依存句法特征等。然而,这些特征过于依赖人工总结,耗时耗力,且受限于有限的语言学资源,容易引发一系列问题:(1)数据稀疏,如利用词袋模型(Bag of Word)表示篇章论元,当不同的词汇作为完全不同的特征时,由于不常见的词汇在训练数据中出现次数较少,在词表中呈现稀疏化;(2)预处理错误级联,如预处理中生成了错误的依存句法树,则该特征错误会在后续识别过程中存在,从而导致训练不充分,影响篇章关系识别性能。

近年来,深度神经网络被成功应用于隐式篇章关系识别任务中,使其识别性能取得了有效的提升。主要包括:(1)基础的神经网络,如卷积神经网络(Convolutional Neural Network, CNN)、循环神经网络(Recurrent Neural Network, RNN)等,将篇章论元映射成低维的稠密向量,从而计算并识别论元之间的语义关系^[11-13]。但这些方法大多单独建模每个论元,仅挖掘各论元的自身特征,忽略了两者在表示学习过程中的关联线索;(2)利用注意力、记忆或门控机制捕捉论元的重要信息^[14-16],以辅助篇章语义理解。该类方法虽比基础模型获得了更多关键线索^[17-18],但仍未考虑到论元在表示过程中的交互关联特征;(3)基于复杂混合神经网络^[19-20]的模型则是利用不同双线性模型^[21](Bilinear Model)、单层网络^[22](Single Layer Network)或其它门控相关网络等方法,以挖掘论元之间的深层交互特征;另外,部分模型采用 Seq2Seq 网络框架^[23-24]构建论元之间的交互关联性。然而,此类方法未能有效抽取篇章关系的特定模式。

此外,认知学相关理论表明,根据不同的阅读目标,人们会选取不同的阅读策略,进而抽取不同的线索信息^[25]。这里,在判断两个文本片段的语义关系时,人们一般会对其进行前后多次查看,在提高其阅读流畅度的同时,找出相关的线索以加深文本理解,从而确定篇章关系。具体如例 1 所示(一般

地,将 Arg1 以斜体表示,Arg2 以粗体表示),如果仅通过单次阅读关注具有对比情感极性的词对(*lucky*, *earthquake*),该论元对可能被错误地识别为比较关系(Comparison);而通过前后多次反复阅读,查找整体论元信息,能够推断其正确的偶然关系(Contingency)。同时发现,该过程中论元之间的交互是非对称的,即 Arg1-Arg2 和 Arg2-Arg1 两个方向所关注的信息是不同的。

例 1.

Arg1: *You are really lucky.*

Arg2: **The earthquake suddenly came two hours after you left.**

源于此阅读策略的启发,我们提出一种新颖的隐式篇章关系识别模型(Mutual Attention-based Neural Tensor Network with Sparse Constraint Model, MATS)。具体来讲,我们侧重在表示阶段建模论元之间非对称的交互关联,设计交互注意力机制以得到具有交互特性的论元表示;针对识别过程中交互模式的有效筛选,我们采用张量神经网络(Neural Tensor Network, NTN)抽取论元之间丰富的语义关联特征,并加入稀疏约束以过滤冗余或噪声信息,从而获取特定的篇章关系模式以辅助任务识别。本文主要贡献如下:

(1)从人们双向阅读策略的认知角度建模篇章论元;从交互模式的稀疏性角度挖掘关系特征;将两者融入一个新颖的隐式篇章关系识别框架中;

(2)利用 BiLSTM 进行篇章论元编码,建模动态交互注意力机制,自动计算论元之间的双向注意力向量,得到具有非对称语境交互感知的论元表示;

(3)采用张量神经网络在更高维度上捕获丰富的关系特征,并在训练过程中加入稀疏约束,从而筛选出更具有指示性的交互模式,以提升识别性能。

本文第 2 节综合分析和总结了相关工作;第 3 节详细论述本文提出的基于语境交互感知和模式筛选的隐式篇章关系识别方法;第 4 节介绍实验数据的准备、模型评估方法以及对比实验设置;第 5 节给出实验结果的分析和讨论;第 6 节进行总结和展望。

2 相关工作

自 2008 年发布人工标注的英文篇章关系语料库^[5](Penn Discourse Treebank, PDTB)以来,涌现出大量篇章关系研究相关的工作:从传统机器学习方法到目前的深度神经网络模型。传统方法包括基于语言学特征、规则的和基于概率统计等不同方法,

例如:Pitler 等人^[7]构建了相关词法、句法特征,引入情感词极性、动词短语长度、动词类型等特征识别篇章关系类型. Lin 等人^[8]在 Pitler 的工作基础上,构建论元的句法树和依存句法树等特征,以改善其模型的识别性能. Rutherford 等人^[10]使用布朗聚类词对等语言特征进一步提高隐式篇章关系识别能力,但此类方法过于依赖人工标注的表层语言特征和所总结的语言规则,存在数据稀疏性问题,且难以捕捉篇章丰富的语义信息. 而深度神经网络将论元对中的单词进行分布式表示,得到论元的向量表征,进而可在高维空间中体现论元之间的相关性,从而有效地识别篇章关系. 其中,与本文相关的工作有以下几个方面.

2.1 篇章论元表示

篇章论元的有效表示是隐式篇章关系识别任务的首要条件. 大多神经网络模型的前提是将论元表征为低维、连续稠密向量^[26]. Ji 等人^[27]首次将论元的表层语言特征进行词向量编码以改善模型识别性能;继而在 2015 年利用句法分析和共指实体等语言特征以增强论元表示^[28]. Lei 等人^[29]通过分析各个篇章关系的语言特性信息,并将其编码为复杂的特征,作为论元的语义表示. 然而,这类研究单独编码论元,没能反映出表示过程中需要关注的重点信息.

针对存在的问题,研究者们将注意力或门机制整合到神经网络模型中,以捕获篇章论元的核心特征,如 Li 等人^[30]提出一种层级注意力机制,用以捕获不同粒度上的重要语义信息. 但这些方法未考虑在表示过程中两者之间的交互关联线索.

2.2 论元间交互模式的挖掘

在获得篇章论元重要信息的基础上,大量工作聚焦于如何利用复杂神经网络的方法挖掘论元间丰富的交互模式特征:Chen 等人^[19]设计了新颖的门机制相关的深度神经网络模型,通过门控来组合双线性模型和单层网络,以捕获更多的交互模式. Lei 等人^[20]计算论元中所有单词对的线性关系和二次关系,以体现论元之间的交互特征. Lan 等人^[31]利用基于注意力的神经网络捕获论元间的交互特征,同时使用多任务学习模型抽取辅任务中无标注数据的相关特征,将其作为关联知识以增强主任务中论元表示,从而提升识别性能. Dai 等人^[32]认为论元不能独立于段落或其相关段落级结构来理解,从而设计段落级神经网络以获取论元之间的相互依赖、关系之间的连续性和关系模式.

此外,注意到张量神经网络能够有效地建模数

据的多重关系,并在自然语言处理的不同任务中得到成功验证,如实体关系抽取^[33]、中文分词^[34]和情感分析^[35]等. 在隐式篇章关系识别中, Li 等人^[36]将张量层融合到多视角框架以捕获更高维度的交互特征. 但他们并未区分所获取的交互关系特征中存在的冗余或噪声信息,忽略了交互模式的稀疏特性. 尽管已有研究通过稀疏正则化过滤冗余权重来获取紧凑神经网络^[37-38],但由于仅修正了网络结构,未针对任务本身特性进行改善,因此交互模式中冗余或噪声信息仍未得到有效解决.

2.3 认知理论的启发

篇章理解属于人工智能中的认知智能,人们理解篇章含义和语义关系的认知行为给本文任务带来了新的启发. 研究者们从认知角度分析,模拟人们的阅读行为,从而建模各自的任务. Yee 等人^[39]的研究展示了人们大脑如何存储和唤醒记忆; Zhang 等人^[40]在 Yee 等人工作基础上,构建语义记忆单元,以分布式形式存储篇章的相关知识,辅助篇章关系识别. Liu 等人^[17]则设计多层注意力机制模拟反复阅读策略,挖掘可判断篇章关系的特定词汇以辅助识别隐式篇章关系. Yue 等人^[18]将最初编码所得的论元表示作为指导,模仿多次阅读过程,动态获得论元的关键线索,以达到逐步深入理解篇章语义的目的.

综上,前人工作在论元表示阶段未考虑论元之间非对称的语义关联特性,在关系识别阶段未区分所获取的交互关系的稀疏特性. 实际阅读过程中,两个论元在不同的阅读顺序中具有不同的语义信息,而该信息在篇章关系识别过程中可能会产生不同的语义关系特征,即论元之间的交互信息是非对称的. 本文将从这两个角度深入探索,以建模交互感知的论元表示和稀疏关系模式筛选,从而改善隐式篇章关系的识别性能.

3 MATS 研究框架

本文将隐式篇章关系识别任务形式化为多分类问题. 本节首先给出 MATS 模型的整体框架,如图 1 所示,主要包括两大部分:(1)语境交互感知的论元表示,实现保留论元自身上下文信息的同时,建模论元间非对称的交互信息表示;(2)稀疏交互模式筛选,获取论元对不同层面的关系特征. 其中,在模型训练过程中实现了对关系特征稀疏特性的处理. 下面将详细介绍模型中的各个部分.

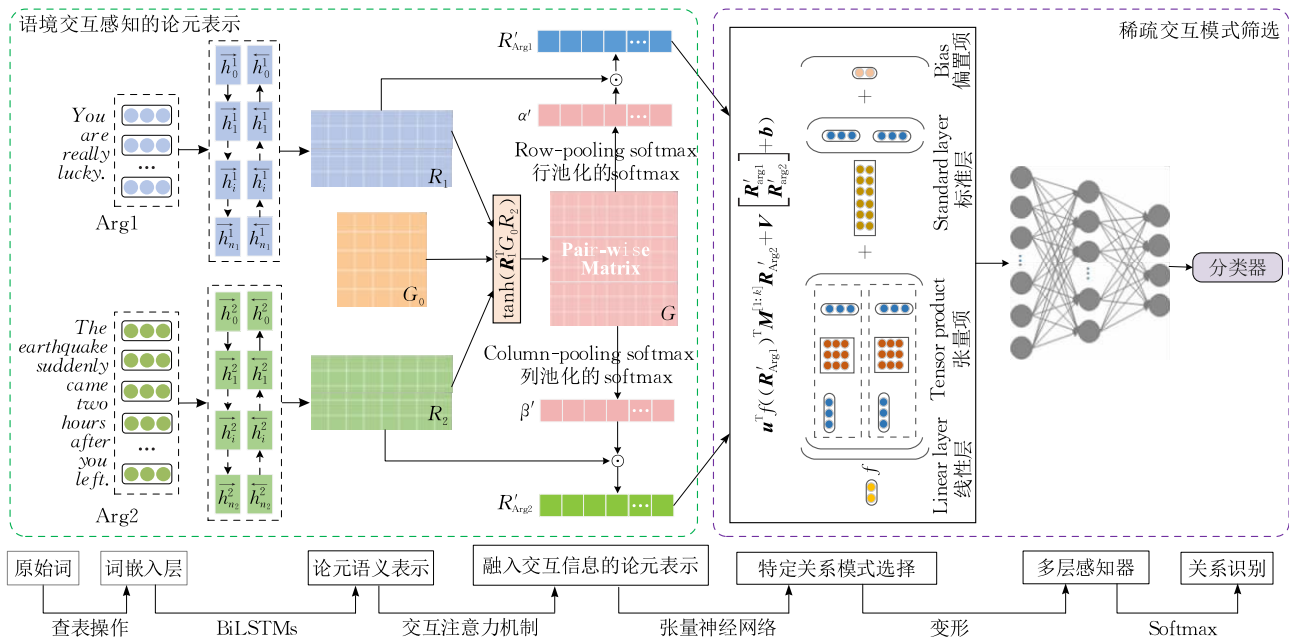


图1 基于语境交互感知和模式筛选的隐式篇章关系识别框架(MATS)

3.1 语境交互感知的论元表示

如何有效编码篇章论元是篇章关系识别任务的核心之一。虽然许多深度神经网络模型能够编码篇章论元,但大多为单独建模两个论元,无法获取丰富的语义特征。同时,源于认知相关理论^[25]的启发,考虑到人们的阅读策略,即通过前后文查看并寻找篇章的重要线索,进而判断具体的篇章关系。其中,不同的阅读顺序(从前向后、从后向前)可能关注到不同的篇章焦点线索,帮助关系的决策;而且前后方向的篇章线索也是非对称的。因此,我们需要从两个方向建模论元之间的非对称语义信息。

3.1.1 嵌入层

数据或特征的分布式表示是神经网络建模的前提。为了利用神经网络编码两个论元,我们将论元中原始单词的独热(One-hot)编码转换为分布式表示。嵌入层可以看作是一个简单的映射层,通过查找表操作索引实现单词嵌入,以捕捉单词的内在属性。我们将词汇表中的每个单词与向量表示 $x \in \mathbb{R}^d$ 关联起来,其中 d 是词向量的维度。每个论元被视为词向量序列,表示为

$$\text{Arg1}: [x_1^1, x_2^1, \dots, x_{n_1}^1], \text{Arg2}: [x_1^2, x_2^2, \dots, x_{n_2}^2] \quad (1)$$

其中, x_i^1, x_j^2 分别是 Arg1 中第 i 个单词和 Arg2 中第 j 个单词, n_1, n_2 则是 Arg1、Arg2 句子长度。

3.1.2 基本论元表示

考虑到循环神经网络(RNN)适合建模序列化数据,但存在梯度消失或爆炸,且难以处理长距离依赖问题。因此,我们采用长短期记忆网络^[41-42](Long

Short-Term Memory, LSTM)建模篇章论元。给定论元中词向量表示, LSTM 通过式(2)~(7)进行每个位置 t 上状态序列计算:

$$i_t = \sigma(W_i[x_t, h_{t-1}] + b_i) \quad (2)$$

$$f_t = \sigma(W_f[x_t, h_{t-1}] + b_f) \quad (3)$$

$$o_t = \sigma(W_o[x_t, h_{t-1}] + b_o) \quad (4)$$

$$\tilde{c}_t = \tanh(W_c[x_t, h_{t-1}] + b_c) \quad (5)$$

$$c_t = i_t \odot \tilde{c}_t + f_t \odot c_{t-1} \quad (6)$$

$$h_t = o_t \odot \tanh(c_t) \quad (7)$$

其中,向量 i_t, f_t, o_t, c_t, h_t 分别表示在位置 t 处输入门、遗忘门、输出门、记忆单元和隐藏状态; W 和 b 是网络中相对应的权重和偏置; σ 表示 sigmoid 激活函数, $\odot$ 表示元素级乘法运算。具体来讲,输入门控制当前输入哪部分被存储,遗忘门控制所保留的历史信息,输出门决定有多少信息作为输出。

由于 LSTM 仅考虑到单一方向序列的历史信息,而未来语境信息也有助于序列建模。故我们采用 BiLSTM^[43] 编码篇章论元,分别从前向和反向编码输入序列,可以捕获历史和未来的上下文信息,从而得到位置 t 上的两个表示:前向隐藏状态 $\vec{h}_t$ 和反向隐藏状态 $\overleftarrow{h}_t$,然后将两者连接起来得到中间表示 $h_t = [\vec{h}_t, \overleftarrow{h}_t]$ 。同理,给定论元对 Arg1-Arg2,得到其中每个单词的中间表示,进而获取其整体表示(R^1, R^2)。

显然,通过基本 BiLSTM 编码所得的两个论元表示之间没有任何关联,也就是说,即使两者之间存在篇章关系,目前却是相互独立的。为了充分挖掘论元表示阶段中二者之间的语义关联,我们探索一种

新的论元表示。

3.1.3 语境交互感知的注意力机制

针对前向和后向论元阅读的不对称语义信息,我们设计两种注意力机制:(1)静态交互注意力,根据论元中所有词对信息,手动构建论元之间的关联矩阵;(2)动态交互注意力,在建模过程中,通过参数更新而自动学习到相应的非对称关联矩阵。

(1) 静态交互注意力

首先,我们通过内积(Dot Product)操作构建篇章论元之间的交互联系,利用式(8)计算论元对中所有词对的语义联系,从而建立关联矩阵(Pair-wise Matrix),矩阵中每个元素体现了 Arg1 中第 i 个词和 Arg2 中第 j 个词的关联性得分。

$$\mathbf{G}_S(i, j) = (\mathbf{R}_i^1)^T \cdot \mathbf{R}_j^2 \quad (8)$$

其中, $\mathbf{G}_S \in \mathbb{R}^{n_1 \times n_2}$, $\mathbf{R}_i^1$ 和 $\mathbf{R}_j^2$ 分别是 Arg1、Arg2 的中间表示。

其次,针对 Arg2 中每个词,我们在矩阵 $\mathbf{G}_S$ 中进行逐列 softmax 以获得对应 Arg1 上的概率分布 α_i ; 同样地,采用逐行 softmax 以得到考虑 Arg1 中每个词时所对应 Arg2 上的 β_j , 如下公式所示。

$$\alpha_i = \text{softmax}(\mathbf{G}_S(1, t), \dots, \mathbf{G}_S(n_1, t)) \quad (9)$$

$$\beta_j = \text{softmax}(\mathbf{G}_S(t, 1), \dots, \mathbf{G}_S(t, n_2)) \quad (10)$$

其中, $\alpha_i = [\alpha_i^1, \alpha_i^2, \dots, \alpha_i^{n_1}]$, α_i^t 表示 Arg1 中第 i 个词在位置 t 上的注意力值; $\beta_j = [\beta_j^1, \beta_j^2, \dots, \beta_j^{n_2}]$, β_j^t 表示 Arg2 中第 j 个词在位置 t 上的注意力值。我们将 $\alpha_i \in \mathbb{R}^{n_1}$ 称作在位置 t 上的“Arg2-to-Arg1”的注意力, $\beta_j \in \mathbb{R}^{n_2}$ 是“Arg1-to-Arg2”的注意力。这一过程的注意力向量是根据 $\mathbf{G}_S$ 计算所得,故称为静态注意力(Static Interactive Attention),如图 2 所示。

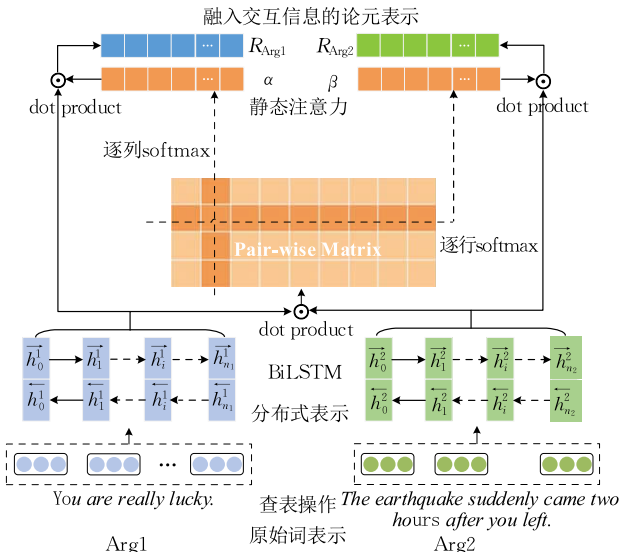


图 2 整合静态注意力机制的论元表示

然后,通过论元对的整体交互信息来表征其语义联系,我们对 α_i 和 β_j 分别进行平均操作,得到两个论元各自的注意力值:

$$\alpha = \frac{1}{n_2} \sum_{t=1}^{n_2} \alpha_t, \quad \beta = \frac{1}{n_1} \sum_{t=1}^{n_1} \beta_t \quad (11)$$

最后,整合论元相应的注意力值到由 BiLSTM 所编码的论元表示中,得到蕴含有论元对的非对称关联特征和上下文信息的最新论元表示:

$$\mathbf{R}_{\text{Arg1}} = \mathbf{R}^1 \alpha, \quad \mathbf{R}_{\text{Arg2}} = \mathbf{R}^2 \beta \quad (12)$$

然而,这里的静态交互注意力向量是根据 BiLSTM 编码论元的中间状态所产生,具有一定的局限性,不能够捕获更多论元之间所隐含的关系特征。为此,我们设计面向篇章关系的动态注意力机制^[44-45],以深入挖掘论元之间的关联性。下面将给出详细的阐述。

(2) 动态交互注意力

动态交互注意力(Dynamic Mutual Attention)关注输入的论元对,使其中一个论元的语义信息在某种程度上会影响另一个论元的语义表示计算,反之亦然。其主要思路是根据 BiLSTM 产生的两个论元的中间状态进行相关联特征的度量,由模型自动学到其注意力向量,如图 1 的表示阶段示意。

正如前面所述,我们得到两个论元对应的表示矩阵 Arg1: $\mathbf{R}^1 \in \mathbb{R}^{d \times n_1}$, Arg2: $\mathbf{R}^2 \in \mathbb{R}^{d \times n_2}$, 然后通过计算式(13)学习到关联矩阵 $\mathbf{G} \in \mathbb{R}^{n_1 \times n_2}$:

$$\mathbf{G} = \tanh((\mathbf{R}^1)^T \mathbf{G}_0 \mathbf{R}^2) \quad (13)$$

其中, $\mathbf{G}_0 \in \mathbb{R}^{d \times d}$ 作为由神经网络所学习的参数矩阵,通过从 $[-0.1, 0.1]$ 均匀分布中随机采样来初始化,并在模型训练过程中进行参数更新;我们使用 tanh 作为激活函数。而矩阵 $\mathbf{G}$ 的元素 $G_{i,j}$ 是列向量 $\mathbf{R}_i^1$ 和 $\mathbf{R}_j^2$ 相关性得分,即篇章论元之间单词对的隐藏向量的对齐得分。之后,我们分别沿着横向和纵向对 $\mathbf{G}$ 进行池化操作以生成重要的特征向量,其形式化为式(14)与(15)。根据实验验证,表明平均池化(Mean-pooling)的效果比最大池化(Max-pooling)更好,故我们采用平均池化操作。

$$g_i^{R^1} = \text{mean}(\mathbf{G}_{i,1}, \mathbf{G}_{i,2}, \dots, \mathbf{G}_{i,n_2}) \quad (14)$$

$$g_j^{R^2} = \text{mean}(\mathbf{G}_{1,j}, \mathbf{G}_{2,j}, \dots, \mathbf{G}_{n_1,j}) \quad (15)$$

其中, $g_i^{R^1}$ 解释为关于 Arg2 的 Arg1 中第 i 单词上下文的重要性得分, $g_j^{R^2}$ 是关于 Arg1 的 Arg2 中第 j 单词上下文的重要性得分。继而,我们得到两个论元对应的池化向量:

$$\mathbf{g}^{R^1} = [g_1^{R^1}, g_2^{R^1}, \dots, g_{n_1}^{R^1}]^T \quad (16)$$

$$\mathbf{g}^{\mathbf{R}^2} = [\mathbf{g}_1^{\mathbf{R}^2}, \mathbf{g}_2^{\mathbf{R}^2}, \dots, \mathbf{g}_{n_2}^{\mathbf{R}^2}]^T \quad (17)$$

然后,利用 softmax 函数将向量 $\mathbf{g}^{\mathbf{R}^1}$ 和 $\mathbf{g}^{\mathbf{R}^2}$ 概率化,分别得到注意力向量 α' 和 β' ,其具体的每个元素由式(18)和(19)计算所得.

$$\alpha'_i = \frac{\exp(\mathbf{g}_i^{\mathbf{R}^1})}{\sum_{j=1}^{n_1} \exp(\mathbf{g}_j^{\mathbf{R}^1})} \quad (18)$$

$$\beta'_i = \frac{\exp(\mathbf{g}_i^{\mathbf{R}^2})}{\sum_{j=1}^{n_2} \exp(\mathbf{g}_j^{\mathbf{R}^2})} \quad (19)$$

最后可得融合论元间交互信息和上下文语境的新论元表示 $\mathbf{R}'_{\text{Arg1}}$ 和 $\mathbf{R}'_{\text{Arg2}}$. 此表示学习过程在一定程度上反映了人们的双向阅读策略.

$$\mathbf{R}'_{\text{Arg1}} = \mathbf{R}^1 \alpha', \mathbf{R}'_{\text{Arg2}} = \mathbf{R}^2 \beta' \quad (20)$$

需要注意的是,本文 MATS 模型选择动态注意力机制以自动学习出论元之间更丰富的关联特征.

3.2 基于张量神经网络的论元交互模式挖掘

不同篇章关系存在其自身特定的关系特征,这些特征由论元之间的交互模式所体现,可以隐含地标识出篇章关系.然而,如何捕获有效的交互模式是篇章关系识别的关键所在.

一般衡量论元之间的交互模式方法有双向性模型和单层网络等.其中,双线性模型通过特定关系(这里指论元之间的语义关联关系)的双线性形式,简单且有效地结合两个向量之间的强线性交互,具体如式(21)所示,

$$\mathbf{S}_{\text{bilinear}}(\mathbf{R}'_{\text{Arg1}}, \mathbf{R}'_{\text{Arg2}}) = (\mathbf{R}'_{\text{Arg1}})^T \mathbf{M}_{bi} \mathbf{R}'_{\text{Arg2}} \quad (21)$$

其中仅有参数 $\mathbf{M}_{bi} \in \mathbb{R}^{d \times d}$.但其不能获取非线性交互特征.

单层网络通过标准非线性操作隐式地连接两个论元,捕获其非线性交互,被定义为

$$\mathbf{S}_{\text{single}}(\mathbf{R}'_{\text{Arg1}}, \mathbf{R}'_{\text{Arg2}}) = (\mathbf{u}_{si})^T f \left(\mathbf{v}_{si} \left[\mathbf{R}'_{\text{Arg1}} \right] + \mathbf{b}_{si} \right) \quad (22)$$

其中, f 是一个标准非线性函数, $\mathbf{u}_{si} \in \mathbb{R}^n$, $\mathbf{v}_{si} \in \mathbb{R}^{n \times 2d}$, $\mathbf{b}_{si} \in \mathbb{R}^n$, n 是隐藏层的大小.但其以论元之间的弱交互信息为代价,未能挖掘丰富的交互特征.

张量神经网络(NTN)通过整合双线性模型和单层网络,可以捕获更高维度中的线性和非线性交互信息.同时,Socher 等人^[33]利用 NTN 代替标准神经网络,直接计算两个实体向量在多个维度上关联得分,进行实体关系推理以证实 NTN 的有效性.与其相似,本文旨在建模两个篇章论元之间的语义关系.因此,我们利用 NTN 挖掘论元之间深层次的

语义交互特征,以促进隐式篇章关系识别.

张量是一种描述向量、标量或其它高维表征各自间关系的几何对象.参考前人工作^[33-34],我们采用三阶张量直接建模论元之间的语义连接关系,如式(23)所示:

$$\mathbf{S}(\mathbf{R}'_{\text{Arg1}}, \mathbf{R}'_{\text{Arg2}}) = \mathbf{u}^T f \left((\mathbf{R}'_{\text{Arg1}})^T \mathbf{M}^{[1:k]} \mathbf{R}'_{\text{Arg2}} + \mathbf{V} \left[\mathbf{R}'_{\text{Arg1}} \right] + \mathbf{b} \right) \quad (23)$$

其中, f 是标准非线性函数, $\mathbf{M}^{[1:k]} \in \mathbb{R}^{d \times d \times k}$ 表示由 k 个矩阵构成的张量,反映了两个论元之间的二维交互.式中 $(\mathbf{R}'_{\text{Arg1}})^T \mathbf{M}^{[1:k]} \mathbf{R}'_{\text{Arg2}}$ 表示双线性张量项(Bilinear Tensor Product),计算可得向量 $\mathbf{m} \in \mathbb{R}^k$,而 $\mathbf{m}$ 的元素是根据对应的张量片(Tensor Slices)计算所得: $m_i = (\mathbf{R}'_{\text{Arg1}})^T \mathbf{M}^{[i]} \mathbf{R}'_{\text{Arg2}}$ ($i=1, 2, \dots, k$),其等效于 k 个双线性模型同时捕获篇章论元之间的线性交互关联;在模型训练过程中, $\mathbf{M}^{[i]}$ 也是在均匀分布 $[-0.1, 0.1]$ 中随机采样得到初始值,通过模型训练进行参数更新.随着 i 的增加,模型的参数和计算难度也随之增加,因此选择合适数目的张量片十分重要.其它参数 $\mathbf{V} \in \mathbb{R}^{k \times 2d}$, $\mathbf{u} \in \mathbb{R}^k$, $\mathbf{b} \in \mathbb{R}^k$ 是标准形式神经网络的参数,具体设置在实验部分进行介绍.这里,每个张量片均可看作为一个“特征抽取器”,以抽取论元之间复杂的语义连接特征.

3.3 稀疏交互模式筛选和篇章关系识别

由张量神经网络所输出的复杂关系矩阵,体现了两个篇章论元之间的语义关系得分,我们将其视为关系特征,重新转换为向量,将其作为多层感知器的输入.具体地,将转换的向量输入到全连接隐藏层,得到更加抽象的表示,然后连接到输出层,利用 softmax 函数计算不同关系类别的概率,从而得到最终的识别结果.

给定包含 n 个任务实例 $\{(\mathbf{x}, \mathbf{y})\}_{r=1}^n$, $\mathbf{x}, \mathbf{y}$ 分别表示一个论元对及其关系标签.我们采用交叉熵损失来评估预测关系表示真实关系的程度,定义为

$$L(\hat{\mathbf{y}}, \mathbf{y}) = - \sum_{j=1}^C \mathbf{y}_j \log(\Pr(\hat{\mathbf{y}}_j)) \quad (24)$$

其中, $\Pr(\hat{\mathbf{y}}_j)$ 表示第 j 个标签的预测概率, C 是篇章关系类别的数量.

经过动态注意力增强论元表示,张量神经网络捕获论元间不同方面的语义交互关联特征.然而,并非所有交互关联征都是有效的,可能存在一些影响系统性能的冗余或噪声交互信息,也就是说,高维度的关系特征具有稀疏性.一般情况下,机器学习方法通过规则项来约束模型,使其具有一定特性,例如:

稀疏、低秩或平滑等,规则项可以是模型参数的范数,不同范数的选择对参数的约束效用也不一样.其中, ℓ_0 和 ℓ_1 范数可以实现稀疏约束,前者是指向量中非零元素的个数,后者是指向量中元素的绝对值之和.然而,对 ℓ_0 的求解是一个 NP 难问题,研究者们大多采用其最优凸近似 ℓ_1 来进行求解,以稀疏化相应的参数.因此,我们选取 ℓ_1 稀疏规则算子以实现张量抽取特征的稀疏性.为了移除不相关交互特征,利用 ℓ_1 对张量项进行约束,使每个张量片 $\mathbf{M}^{[i]}$ 中大部分值为 0,从而筛选出具有指示性的交互模式.另外,利用 ℓ_2 正则化避免模型过拟合的问题,具体训练目标为

$$J(\theta)=\frac{1}{n}\sum_{r=1}^nL(\hat{\mathbf{y}}^r,\mathbf{y}^r)+Q(\theta)\tag{25}$$

$$Q(\theta)=\lambda_M\|\theta_M\|_1+\frac{\lambda_o}{2}\|\theta_o\|_2\tag{26}$$

其中, $Q(\theta)$ 是关于参数 θ 的正则化项,包含两部分(见式(26)): θ_M 是张量片的参数, θ_o 是模型中其它参数.特别地, $\|\theta_M\|_1$ 是对张量片的稀疏约束,为了获取各个方面的重要值.

ℓ_1 正则化项在 0 处不可微,故而为了最小化目标函数,我们采用近端梯度下降法^[46]将目标优化为平滑和非平滑项的组合,具体计算如式(27)~(30)所示.

$$\theta_i^t{}'=\theta_i^t-\gamma\left(\frac{\partial L}{\partial \theta_i}+\lambda\frac{\partial Q}{\partial \theta_i}\right)|_{\theta_i=\theta_i^t}\tag{27}$$

$$\frac{\partial Q}{\partial \theta_i}=\begin{cases}2\theta_i^t,&\|\theta\|_2\\ \text{sign}(\theta_i^t),&\|\theta\|_2\text{ and }\theta_i^t\neq 0\end{cases}\tag{28}$$

$$\theta_i^{t+1}=prox_{\lambda}(\theta_i^{t'})=\tau(\theta_i^{t'},\gamma\lambda)\tag{29}$$

$$\tau=\begin{cases}a-z,&a>z\\ a+z,&a<-z\\ 0,&\text{其他}\end{cases}\tag{30}$$

其中, $prox_{\lambda}$ 是近似数学操作, τ 是软阈值方法, γ 表示学习率.

4 实验准备

4.1 实验数据

近年来,不同架构的篇章语料构建使篇章级相关研究受到越来越多研究者的关注.宾州篇章树库(PDTB)于 2008 年发布,是目前规模最大的英文篇章关系语料库,其借鉴篇章词汇化树形连接语法(Discourse Lexical Tree Adjunct Grammar,D-LTAG)理论^[47],从《华尔街日报》的 2312 篇文章中标注了

40 600 个篇章关系.该篇章关系被定义为一个 3 级层次意义(Sense)结构:最上层是类别(Class),第二层是类型(Type),第三层是子类型(Sub-type).

本文仅针对第一层的四大类隐式篇章语义关系进行识别,即比较关系(Comparison),偶然关系(Contingency),扩展关系(Expansion)和时序关系(Temporal),其实例分别对应如例 2 所示.

例 2.

(a) [In the U. S. , one-fifth of the office PCs are hooked up to some sort of network]_{Arg1}. {Implicit= while} [In Japan, about 1% are linked]_{Arg2}.

(Comparison; Contrast; juxtaposition)

(b) [Mr. Yeargin is lying]_{Arg1}. {Implicit = because} [They found students in an advanced class a year earlier who said he gave them similar help]_{Arg2}.

(Contingency; Pragmatic Cause; justification)

(c) [An enormous turtle has succeeded where the government has failed]_{Arg1}. {Implicit = specifically} [He has make speaking Filipino respectable]_{Arg2}.

(Expansion; Restatement; specification)

(d) [New York Stock Exchange volume was 237,960,000 shares]_{Arg1}. {Implicit = then} [One cell was teased out, and its DNA extracted]_{Arg2}.

(Temporal; Asynchronous; precedence)

其中,比较关系一般是突显两种情况的差异;偶然关系主要描述论元之间主观或客观的影响;扩展关系是涵盖扩大的论元论述范围,并推动其叙述向前发展的语义关系;时序关系是反映论元之间在时间上的关联特性.这些篇章关系中没有显式的语言线索,且可能存在歧义表述,导致其识别的难度增加.

与前人工作一致^[17,19,48],我们选用语料中 02-20 小章作为训练集,21-22 小章作为测试集,00-01 小章作为验证集.注意,扩展关系的数据设置遵循了 Rutherford 等人^[10]的方法,未将实体关系(EntRel)关系合并作为扩展关系.表 1 给出了数据在各个隐式篇章关系中的分布情况.

表 1 PDTB 中隐式篇章关系数据统计.

关系类型	训练集	验证集	测试集
比较关系(Comparison)	1842	188	144
偶然关系(Contingency)	3139	272	266
扩展关系(Expansion)	6658	635	537
时序关系(Temporal)	579	47	55
总数	12218	1142	1002

为了与目前篇章关系识别的先进模型进行比较,我们采用两种实验设置:(1)多元分类,即四分类;(2)四个独立二分类.前者用于评估 MATS 模

型的整体性能,符合任务本身四个关系的真实设置. 后者评估单一关系的二分类性能,即单一关系为正例,其它关系均为负例,可在一定程度上解决语料不平衡的问题. 在二分类中,除了扩展关系(Expansion)之外,其它关系数量较少,故每个关系的训练数据均使用相同数量的正例和负例,负例从相应的集合中随机选择. 同时,验证集和测试集不作改变,保持自身的自然状态.

4.2 参数设置

在本文 MATS 模型训练过程,我们首先对语料进行预处理,例如将语料中的单词均转换为小写等. 我们使用 GloVe^[49] 预训练的词嵌入表示,并尝试其不同维度[50,100,200]的向量表示,最终选择维度为 50 的词向量;如果预训练的词向量中未出现当前单词,则该单词由 $[-1,1]$ 均匀分布随机初始化而生成向量表示. 每个论元的长度被设置或填充为 50;同时参照已有研究^[17,19],通过比较不同的 LSTM 中间层表示长度的实验结果,且选择将其设置为 50. 其它参数从 $[-0.1,0.1]$ 均匀分布中随机初始化生成.

为了保证实验结果的稳定性,我们在模型训练过程中,取 5 次结果的平均值,进而使用简单网格搜索算法进行参数选择. 根据在验证集上的调整,确定模型最终参数,部分超参的设置如表 2 所示.

表 2 MATS 模型中部分超参设置

超参数	数值
初始化学率	0.01
Mini-batch 大小	32
Dropout	0.1
张量片的数量	3

4.3 评估方法

为了有效评估本文 MATS 模型对四种隐式篇章关系的识别性能,我们采用篇章关系识别中通用的评测标准:F1 值和精确率(Accuracy). 其中,F1 值是准确率(Precision)和召回率(Recall)的调和均值,可以更好地衡量模型性能.

4.4 对比实验设计

4.4.1 对比模型

为了验证本文所提 MATS 模型的有效性,我们从篇章论元表示、交互模式挖掘和认知启发三个方面,选择以下与本文相关的具有代表性的系统作为对比模型.

(1) 篇章论元表示

① Ji2015^[28] 采用递归神经网络建模论元的句

法分析树,并沿树状结构得到实体表示,将论元与实体组合作为最终论元表示.

② Qin2016^[14] 提出一种堆叠式神经网络模型,包括卷积神经网络获取“不同粒度”局部特征;协同门控网络特征控制重要信息的比重,以辅助篇章关系识别. 注意,尽管在一定程度上其捕获论元之间交互特征,但从模型整体分析,本文认为其主要体现了论元表示的建模.

③ Lei2018^[29] 详细分析四类关系的语言特性,总结其独特语义特征和衔接手段,即主题连续性(Topic Continuity)和归因信息(Attribution)针对性,通过建模这些语言特征作为论元的语义表示,采用朴素贝叶斯分类器进行隐式篇章关系的识别.

(2) 论元之间交互特征

① Chen2016^[19] 设计门控相关的神经网络模型(GRN),以便捕获篇章论元中单词对之间的线性和非线性语义交互,使用池化操作选择最具信息性的交互特征.

② Lei2017^[20] 采用简单的神经网络模型(SWIM)识别隐式篇章关系,将词向量加权平均作为论元表示,计算论元中词对的交互得分作为论元之间交互特征.

③ Lan2017^[31] 在利用基于注意力机制的神经网络学习论元表示的同时,考虑两个论元的交互,并采用多任务学习从辅任务中获取标注及未标注语料的相关知识.

④ Dai2018^[32] 将篇章语言单元由从句扩大到段落,引入一种段落级神经网络,以建模论元之间的相互依赖性、篇章关系的连续性,并预测段落中论元之间语义关系的顺序.

(3) 认知角度

① Zhang2016^[40] 受人们理解篇章语义的启发,构建语义记忆网络,其中浅层编码器学习论元的表示,语义编码器进一步建模论元的深层语义特征,并利用注意力检索出最相关特征.

② Liu2016^[17] 通过设计基于多层注意力的神经网络模型(NNMA)来模仿人们反复阅读策略,选择识别隐式篇章关系过程中可能关注的词汇信息. 注意,本文参照其论文设置,选取 2 层和 3 层注意力.

需要说明的是,文中选用的代表性系统模型为直接引用原始论文中的结果. 由于所有模型都是遵循相同语料和数据划分,因此具有可比性.

4.4.2 消融模型

(1) LSTM 分别编码两个论元,将两者的表示

直接拼接起来作为后续全连接隐藏层的输入,之后用 softmax 进行篇章关系分类.

(2) BiLSTM 弥补了 LSTM 的不足,采用两个 LSTM 保留历史和未来上下文信息(如 3.1.2 小节所述)以更好地建模论元对.

(3) BiLSTM+Static Interactive Attention(BSIA)静态构建论元对的交互关联矩阵,整合注意力向量于论元表示,见式(8)~(11).

(4) BiLSTM+Dynamic Mutual Attention(BDMA)改进了 BSIA 的单一性,由模型自动学习到论元对之间的交互关联性,将其融入到基础论元表示中,具体见式(13)~(19).

(5) BiLSTM+Tensor Layer(BT)在 BiLSTM 编码论元后,利用张量神经网络来捕捉不同方面的丰富的语义关联特征.

(6) BSIA+Tensor Layer(BSIAT)融合 BSIA 和张量神经网络,采用 k -max 池化操作从丰富的交互特征中选择前 k 个重要关系模式特征.

(7) BDMA+Tensor Layer(BDMAT)与 BSIAT 相似,BDMA 和张量层相结合, k -max 池化操作选择前 k 个重要关系模式特征.

5 实验结果与分析

本节首先给出 MATS 模型、对比系统和消融模型分别在四分类和二分类上各自的实验结果;然后通过对比分析,探讨本文模型的优缺点;最后对模型的主要模块进行分析验证.

5.1 与对比模型的比较

表 3 展示了对比系统及本文模型在四分类任务上的实验结果,通过观察可以得到以下结论:

(1)整体上,侧重篇章论元表示的系统性能低于侧重论元间交互特征的和受认知理论启发的方法. 导致这种现象的原因可能是侧重篇章论元表示的系统对论元的独立并行编码,忽略了论元之间的关联特性. 另外,与侧重论元间交互的方法相比,认知理论启发的系统尽管从人类阅读角度设计模型,但其 $F1$ 值并未得到明显的提升. 这可能是因为其焦点均集中于论元各自的重要信息,而未探索这些信息之间的关联性,而恰好这些关联信息在一定程度上可以反映出语义关系. 本文 MATS 模型则在考虑论元交互式表示的同时,挖掘论元对之间的语义交互关系,取得了比其它系统好的结果. 其中, MATS 模型自动学习到交互感知的新颖论元表示,取得了更高的性能.

(2)具体来讲, Ji2015 作为侧重篇章论元表示的方法代表,其识别性能最低. 其他两类中, Liu2016⁽³⁾ 和 Dai2018 获得了较好的精确率,表明通过深入挖掘篇章中重要线索、扩大篇章语言单元的粒度均能够增强篇章论元的理解,以改善隐式篇章关系的识别性能. 而 Liu2016⁽³⁾ 的精确率都稍高于 Dai2018,其原因可能是:①篇章本身核心线索的指示作用高于更大语言粒度的语义表示;②利用注意力机制捕获许多重要特性信息的同时,过多的注意力层会产生更多参数,从而导致过拟合问题. Liu2016⁽²⁾ 的 $F1$ 值高于 Liu2016⁽³⁾ 的 1.34%,而精确率却低于 Liu2016⁽³⁾; Dai2018 在 $F1$ 和精确率方面的性能均超过 Liu2016⁽²⁾,这些结果也在一定程度上反映了原因②的可能性.

(3)值得注意的是,与在 $F1$ 和精确率上取得最高值的 Dai2018 和 Liu2016⁽³⁾ 相比,本文 MATS 模型的 $F1$ 值和精确率均获得了提升,特别是精确率比 Liu2016⁽³⁾ 高了 2.36%. 这可能是论元之间非对称的双向交互表示和稀疏关系模式选择的共同作用,从而有效地提升了隐式篇章关系识别性能.

表 3 各个系统在四分类任务上的性能.

模型	$F1/\%$	Accuracy/ $\%$
Ji2015	38.52	43.56
Lei2018	47.15	—
Lei2017	46.46	—
Lap2017	47.8	57.39
Dai2018	48.82	57.44
Liu2016 ⁽²⁾	46.29	57.17
Liu2016 ⁽³⁾	44.95	57.57
Our MATS	48.94	59.93

注:(1)未列出系统模型表明没有获取该项结果;
(2) Liu2016⁽²⁾ 和 Liu2016⁽³⁾ 分别表示 2 层、3 层注意力^[17].

关于二分类,表 4 给出各个模型在四个关系上的 $F1$ 性能结果,通过观察可以得到以下结论:

(1)在四大类关系中,所有模型在时序关系上的 $F1$ 值均最低. 其原因可能是时序关系在语料库中所占的比例最小,仅有 5%. 而随着不同关系类别中实例数量的增加,其 $F1$ 值也随之上升,这表明语料的规模大小对隐式篇章关系的识别是至关重要的.

(2)本文 MATS 模型在扩展关系识别上取得了最好性能. 其中的原因可能有两个方面:①扩展关系实例中可能存在一些容易混淆的词对,通过非对称交互注意力的编码,而获得不同信息;②对于复杂的论元对,需要进一步加深理解,从而捕获具有指示性和交互特性的关系特征. MATS 模型则能有效融合这两个方面.

(3)偶然关系方面, MATS 虽然未取得最好性能

能,其 $F1$ 值低于 Lan2017 的 1.07%,但也取得了与 Lan2017 可比较的结果. 这可能是因为 Lan2017 利用辅任务添加了未标注数据的相关特征作为知识,扩大了语料库,增强了论元的表示,即不仅考虑到篇章内部语义,还引入了相关知识. 另外,Chen2016 和 Liu2016 均与本文工作十分相关,不同的是我们在表示层面加入交互感知,以及训练过程加入稀疏约束,这可能是 MATS 取得更好性能的原因(分别获得了 3.13%和 3.41%的改善).

(4) Lei2018 在比较关系上得到了最优性能,其 $F1$ 值超过 MATS 模型的 1.67%. 这是由于 Lei2018 针对 PDTB 语料进行详细分析,发现了①篇章关系的独特语义信息;②主题连续性和归因的论元衔接手段. 这些特征能够反映各个关系的特定模式,从其他三类关系的结果也表现这一现象. 然而,其需要复杂的特征工程进行描述,不具有强泛化能力,特别是时序关系由于实例数量较少,Lei2018 不能有效获取其特征,导致性能较低.

表 4 各个系统在二分类任务上的性能				
模型	比较/%	偶然/%	扩展/%	时序/%
Ji2015	35.93	52.78	—	27.63
Qin2016	41.55	57.32	71.50	35.43
Lei2018	43.24	57.82	72.88	29.10
Chen2016	40.17	54.76	—	31.32
Lei2017	40.47	55.36	69.50	35.34
Lan2017	40.73	58.96	72.47	38.50
Dai2018	37.72	49.39	67.45	40.70
Zhang2016	34.82	53.70	70.95	36.16
Liu2016 ⁽²⁾	36.70	54.48	70.43	38.84
Liu2016 ⁽³⁾	39.86	53.69	69.71	37.61
Our MATS	41.57	57.89	73.89	39.71

5.2 与消融模型的比较

设计消融模型以验证 MATS 中动态注意力机制、张量神经网络及其稀疏约束的有效性. 实验结果见表 5 所示,针对二分类,我们得到以下结论:

(1)表 5 首先给出了 LSTM 和 BiLSTM 模型实验结果,作为基线方法. 其中,LSTM 在四类关系上

的实验性能均较差;虽然 BiLSTM 保留了历史和未来上下文信息,优于仅存储历史信息的 LSTM,但结果却并未得到很高的提升. 其原因可能是这两个方法对论元的编码都是独立、并行的,没有从任务本质出发,从而设计具有篇章特征的编码方法.

(2)与 LSTM 和 BiLSTM 相比,基于注意力机制的方法(BISA、BDMA)实现了更好的性能. 特别地,BISA 的 $F1$ 值在比较、偶然和时序三个关系上分别比 LSTM 提高了 2.48%,2.54%,1.70%. 这表明通过建立论元对中所有词对的关联以描述论元之间的交互特性,能够获取特定的语义关系线索,改善篇章关系识别的性能. 而 BDMA 则在比较、扩展关系上取得了更大的改进,即分别高于 LSTM 模型达 1.25%,1.51%. 这证明交互感知的论元表示有助于隐式篇章关系识别,且由动态注意力机制捕获的非对称双向关联特征更加丰富,优于以词对为基础所构建的静态注意力矩阵. 注意,我们对实验结果进行了显著性检验,表明其在单侧 t 检验中均显著($p < 0.05$). 此外,BT 模型在整体上取得了与 BDMA 不相上下的实验性能. 这种现象说明张量层所挖掘的丰富且复杂的交互关系特征对该任务是有效的.

(3)BSIAT 和 BDMAT 模型比前两类方法的性能都更好,这意味着交互感知的表示学习和丰富关系模式的融合可以更好地改善篇章关系识别的结果;其中,BDMAT 在四类关系上的实验性能比 BSIAT 均略有提升,这表明神经网络动态学习的注意力向量可能包含更多不同方面的信息. 然而,BDMAT 在各个关系上的性能均低于本文 MATS 模型. 导致这一结果的原因是前者利用 k -max 池化操作选取前 k 个重要特征,无法保证从全局角度可得到有效的重要交互模式. 而 MATS 在 BDMAT 的基础上,增加了对张量项的稀疏约束,从而筛选出更具有指示性的交互关系模式. 另外,也表明注意力机制、张量神经网络及其稀疏约束的融合,有效提高了隐式篇章关系识别性能.

表 5 消融模型在 PDTB 数据上的实验性能						
模型	二分类				四分类	
	比较关系/%	偶然关系/%	扩展关系/%	时序关系/%	$F1$ /%	Accuracy/%
LSTM	32.95	43.38	68.10	30.80	36.40	54.50
BiLSTM	34.01	44.68	68.53	31.27	36.54	55.31
BSIA	35.43	45.92	68.57	32.50	43.27	55.68
BDMA	36.68	46.44	70.08	32.99	43.75	56.03
BT	37.36	46.73	69.81	33.89	43.61	55.97
BSIAT	39.78	55.13	70.96	37.65	46.77	57.83
BDMAT	40.90	56.37	71.87	38.30	47.39	59.17
Our MATS	41.57	57.89	73.89	39.71	48.94	59.93

与此同时,观察发现消融模型四分类结果趋势与二分类基本一致,证明了 MATS 各模块的有效性.

5.3 不同组件对模型性能的影响

在 MATS 模型中,动态交互注意力机制和张量神经网络均可改善隐式篇章关系的识别. 本小节将对两者分别进行详细地讨论分析.

5.3.1 动态交互注意力机制的影响

为了直观地理解两个篇章论元之间的交互关联性,我们对前文中例 1 进行可视化表示,如图 3 所示,其中展示了三个模型相应的交互关联矩阵:仅用 BiLSTM、融入静态注意力机制和融入动态注意力机制. 图中两个论元中词对所对应位置的背景颜色呈现了其相关性,颜色越深则表明所对应的词对的相关性越强.

图 3 中的三个子图充分展示了上述例子所描述

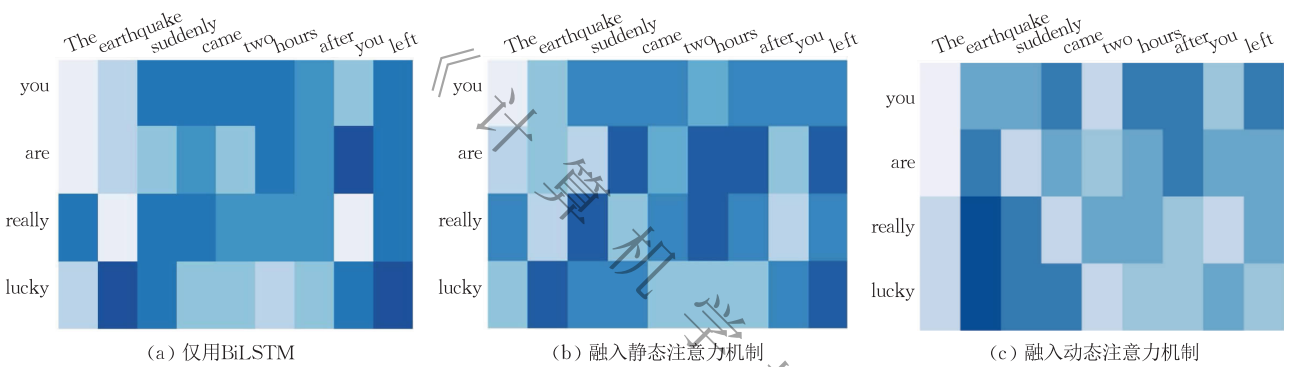


图 3 不同方法所获取的篇章论元之间关联性

5.3.2 张量片数量的影响

张量神经网络探究了多个维度上两个向量的关联性,其多个维度是由张量片的数量所体现. 故我们分析了张量片不同数量对四类隐式篇章关系识别的影响,即在 MATS 模型张量层中设置[0,1,2,3,4]不同大小的张量片,得到其在四种关系上对应的 F1 值. 注意,当片数为 0 时,其直接退化为一个简单的单层网络模型.

从图 4 可知,随着张量片数量的增加,MATS 在四类关系上的 F1 值均有所提升. 这表明复杂的

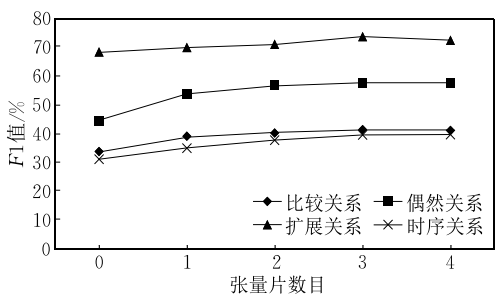


图 4 不同数量的张量片对四类关系的影响

的现象. 具体分析,图 3(a)中,词对(are, you), (lucky, earthquake), (lucky, left)获得了较高的相关性得分,而其它词对的得分显得十分随机. 这表明仅用 BiLSTM 模型可能受到语料库中词对频率的影响,即“are”和“you”这些词经常在语料库中共存,故得分会相对较高. 同时,图中没有反映出一些关联性较强的词对信息.

图 3(b)、(c)作为对比,观察发现其存在更多较高得分的词对,其中一些并未在图 3(a)中出现. 这表明交互注意力在一定程度上捕获了论元之间的关联特征. 进一步,图 3(c)更好地显示了其关联信息,即与“lucky”和“earthquake”相关联的词对得分比图 3(b)较为平均,可能意味其语义理解范围更加广泛,这有助于加深对篇章的理解,从而准确推断其隐式的篇章关系.

张量层能够捕获更多表征论元之间交互模式的特征. 然而,张量片的增加可能会使模型待学习参数变多,消耗更多计算资源,延长其计算时间,还可能导致模型过拟合. 因此,通过实验分析,我们在 MATS 模型中选择张量片数为 3,保证合理计算的同时,获得有效的识别性能.

6 总结与展望

本文抓住隐式篇章关系的主要挑战,受认知科学相关理论启发,提出了基于语境交互感知和模式筛选的隐式篇章关系识别方法(MATS),以便更好地理解篇章论元和选择重要交互模式进行高效识别. 通过设计动态交互注意力机制以刻画论元间双向非对称的内在关联,也就是将人们的双向阅读策略通过自动学习建模到融合语境交互感知的论元表示中;由于识别过程中隐含的篇章论元认知交互具有海量和稀疏特性,本文采用带有稀疏约束的张量神经网络挖掘具有篇章关系指示性的有效交互模

式,从而提升模型的识别性能。在 PDTB 语料库上的实验结果显示了论元之间双向非对称的交互感知信息和稀疏模式筛选的有效性。与此同时,我们分析了张量神经网络中不同数量的张量片对模型性能的影响;同时,采用实例分析展示了交互注意力机制的特点,进一步表明了各模块对隐式篇章关系识别的促进作用。

然而,本文工作也存在一些问题:(1) 尽管 MATS 模型可以筛选出更具有指示性的交互模式,但其不具有可解释性。这也是神经网络模型面临的主要困难之一;(2) MATS 模型仅仅关注于篇章论元间的关联性信息,忽略了更广范围的篇章语境信息。因此,未来研究工作中,我们将尽可能探索指示性模式的可解释性;建模更高层级篇章单元(例如:句群、段落)之间的关系或利用外部关联知识来增强篇章语义理解,实现更准确的隐式篇章关系识别。

致 谢 衷心感谢各位审稿专家对本文工作的指导!

参 考 文 献

- [1] Xi Xue-Feng, Sun Qing-Ying, Zhou Guo-Dong. Research and prospect of discourse topic structure analysis for discourse intentionality. *Chinese Journal of Computers*, 2017, 40(83): 1-26(in Chinese)
(奚雪峰, 孙庆英, 周国栋. 面向意图性的篇章话题结构分析研究与展望. *计算机学报*, 2017, 40(83): 1-26)
- [2] Meyer T, Popescu-Belis A. Using sense-labeled discourse connectives for statistical machine translation//*Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation and Hybrid Approaches to Machine Translation*. Avignon, France, 2012: 129-138
- [3] Ma M D, Bowden K K, Wu J, et al. Implicit discourse relation identification for open-domain dialogues//*Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy, 2019: 666-672
- [4] Gerani S, Mehdad Y, Carenini G, et al. Abstractive summarization of product reviews using discourse structure//*Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar, 2014: 1602-1613
- [5] Prasad R, Diesh N, Lee A, et al. The penn discourse TreeBank 2.0//*Proceedings of the 6th Edition of the Language Resources and Evaluation Conference*. Marrakech, Morocco, 2008: 2961-2968
- [6] Pitler E, Raghupathy M, Mehta H, et al. Easily identifiable discourse relations. University of Pennsylvania, Philadelphia, USA: Technical Report (CIS): No. MS-CIS-08-24, 2008
- [7] Pitler E, Louis A, Nenkova A. Automatic sense prediction for implicit discourse relations in text//*Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language*. Singapore, 2009: 683-691
- [8] Lin Z, Kan M Y, Ng H T. Recognizing implicit discourse relations in the penn discourse TreeBank//*Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore, 2009: 343-351
- [9] Zhou Z M, Xu Y, Niu Z Y, et al. Predicting discourse connectives for implicit discourse relation recognition//*Proceedings of the 23rd International Conference on Computational Linguistics*. Beijing, China, 2010: 1507-1514
- [10] Rutherford A T, Xue N. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns//*Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden, 2014: 645-654
- [11] Braud C, Denis P. Comparing word representations for implicit discourse relation classification//*Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal, 2015: 2201-2211
- [12] Zhang B, Su J, Xiong D, et al. Shallow convolutional neural network for implicit discourse relation recognition//*Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal, 2015: 2230-2235
- [13] Liu Y, Li S, Zhang X, Sui Z. Implicit discourse relation classification via multi-task neural networks//*Proceedings of the 30th AAAI Conference on Artificial Intelligence*. Phoenix, USA, 2016: 2750-2756
- [14] Qin L, Zhang Z, Zhao H. A stacking gated neural architecture for implicit discourse relation classification//*Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, USA, 2016: 2263-2270
- [15] Cianflone A, Leila K. Attention for implicit discourse relation recognition//*Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan, 2018: 1946-1951
- [16] Cai D, Zhao H. Pair-aware neural sentence modeling for implicit discourse relation classification//*Proceedings of the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Montreal, Canada, 2017: 458-466
- [17] Liu Y, Li S. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention//*Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, USA, 2016: 1224-1233

- [18] Yue X, Fu L, Wang X. Externally controllable RNN for implicit discourse relation classification//Proceedings of the 6th Conference on Natural Language Processing and Chinese Computing. Dalian, China, 2017: 158-169
- [19] Chen J, Zhang Q, Liu P, et al. Implicit discourse relation detection via a deep architecture with gated relevance network //Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 1726-1735
- [20] Lei W, Wang X, Liu M, et al. SWIM: A simple word interaction model for implicit discourse relation recognition//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia, 2017: 4026-4032
- [21] Jenatton R, Roux N L, Bordes A, Obozinski G R. A latent factor model for highly multi-relational data//Proceedings of the 26th Annual Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2012: 3167-3175
- [22] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multi-task learning//Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland, 2008: 160-167
- [23] Shi W, Demberg V. Learning to explicitate connectives with Seq2Seq network for implicit discourse relation classification//Proceedings of the SIGDial 2019 Conference. Stockholm, Sweden, 2019: 188-199
- [24] Varia S, Hidey C, Chakrabarty T. Discourse relation prediction: Revisiting word pairs with convolutional networks//Proceedings of the SIGDial 2019 Conference. Stockholm, Sweden, 2019: 442-452
- [25] Wren S. The cognitive foundations of learning to read: A framework. Southwest Educational Development Laboratory, Austin, USA: Descriptive Report (ERIC): No. MS-CIS-08-24, 2000
- [26] Qin L, Zhang Z, Zhao H. Shallow discourse parsing using convolutional neural network//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 70-77
- [27] Ji Y, Eisenstein J. Representation learning for text-level discourse parsing//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, USA, 2014: 13-24
- [28] Ji Y, Eisenstein J. One vector is not enough: Entity-augmented distributional semantics for discourse relations. Transactions of the Association for Computational Linguistics, 2015, 3: 329-344
- [29] Lei W, Xiang Y, Wang Y, et al. Linguistic properties matter for implicit discourse relation recognition: Combining semantic interaction, topic continuity and attribution//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Louisiana, USA, 2018: 4848-4855
- [30] Li Q, Li T, Chang B. Discourse parsing with attention-based hierarchical neural networks//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, USA, 2016: 362-371
- [31] Lan M, Wang J, Wu Y, et al. Multi-task attention based neural networks for implicit discourse relationship representation and identification//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 1299-1308
- [32] Dai Z, Huang R. Implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph //Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Louisiana, USA, 2018: 141-151
- [33] Socher R, Chen D, Manning C D, Ng A. Reasoning with neural tensor networks for knowledge base completion//Proceedings of the 27th Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2013: 926-934
- [34] Pei W, Ge T, Chang B. Max-margin tensor neural network for Chinese word segmentation//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, USA, 2014: 293-303
- [35] Socher R, Perelygin A, Wu J, et al. Recursive deep models for semantic compositionality over a sentiment Treebank//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, USA, 2013: 1631-1642
- [36] Li H, Zhang J, Zhou Y, Zong C. Predicting implicit discourse relation with multi-view modeling and effective representation learning//Proceedings of the National CCF Conference on Natural Language Processing and Chinese Computing. Kunming, China, 2016: 374-386
- [37] Yoon J, Hwang S J. Combined group and exclusive sparsity for deep neural networks//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia, 2017: 3958-3966
- [38] Wen W, He Y, Rajbhandari S, et al. Learning intrinsic sparse structures within long short-term memory. arXiv preprint arXiv:1709.05027, 2017
- [39] Yee E, Chrysikou E G, Thompson-Schill L. The cognitive neuroscience of semantic memory//Ochsner K, Kosslyn S eds. The Oxford Handbook of Cognitive Neuroscience. Oxford, UK: Oxford University Press, 2014
- [40] Zhang B, Xiong D, Su J. Neural discourse relation recognition with semantic memory. arXiv preprint arXiv:1603.03873, 2016
- [41] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780
- [42] Gers F A, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with LSTM//Proceedings of the 9th International Conference on Artificial Neural Networks. Edinburgh, UK, 1999: 850-855

[43] Schuster M, Paliwal K. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 1997, 45(11): 2673-2681

[44] Santos C D, Tan M, Xiang B, Zhou B. Attentive pooling networks. *arXiv preprint arXiv:1602.03607v1*, 2016

[45] Tu C, Liu H, Liu Z, Sun M. CANE: Context-aware network embedding for relation modeling//*Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vancouver, Canada, 2017: 1722-1731

[46] Parikh N, Boyd S. Proximal algorithms. *Foundations and Trends in Optimization*, 2014, 1(3): 127-239

[47] Webber B. D-LTAG: Extending lexicalized TAG to discourse. *Journal of Cognitive Science*, 2004, 28(5): 751-779

[48] Rutherford A T, Demberg V, Xue N. Neural network models for implicit discourse relation classification in English and Chinese without surface features. *arXiv preprint arXiv:1606.01990*, 2016

[49] Pennington J, Socher R, Manning C D. GloVe: Global vectors for word representation//*Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar, 2014: 1532-1543



GUO Feng-Yu, Ph. D. candidate. Her main research interests include natural language processing, discourse relation recognition.

HE Rui-Fang, Ph. D. , professor. Her main research interests include natural language processing, social media mining, and machine learning.

DANG Jian-Wu, Ph. D. , professor. His main research interests include speech recognition, speech conversation, speech cognitive brain mechanism and speech comprehension.

Background

Discourse Relation Recognition belongs to the field of discourse analysis. It aims to understand and annotate the latent relations between two discourse arguments. Yet implicit discourse relation recognition without explicit connectives, which needs to infer the relation from a specific context, is still a challenging problem. Existing implicit discourse relation recognition approaches are roughly divided into two categories: traditional feature-based and neural network-based models. The former adopt the artificially designed linguistic features and complicated rules. However, implicit discourse relations are rooted in the semantics and difficult to recognize from the surface features.

Recent neural network-based models have shown success in recognizing implicit discourse relation. (1) Basic neural networks, such as convolution neural network (CNN) and recurrent neural network (RNN), can learn the dense vector representations of two discourse arguments, which could capture the semantic information to some extent. (2) Further studies exploit different attention or memory mechanisms to capture the significant information of the arguments. (3) More complicated neural network architectures utilize various methods (namely, gated relevance model, single linear layer, etc.) to mine the connections between the arguments. Although they acquire the key information and some interactive clues of discourse arguments, they still have the following drawbacks: (1) ignore the pair specific clue due to encoding two discourse arguments separately; (2) neglect the bidirectional asymmetry interactions between the two arguments during the represen-

tation stage; (3) rarely consider the sparsity of the specific interactive patterns indicating discourse relation.

In addition, we explore the people’s daily reasonable reading strategy to bring inspiration for this task. We propose a novel mutual attention-based neural tensor network with sparse constraint, called MATS, to deepen the semantic understanding of both two discourse arguments and select the significant pair patterns for recognizing implicit discourse relation. Research ideas and innovations in the paper are as follows. The mutual attention-based BiLSTM model learns the semantic representations of two discourse arguments, which imitates the human-like bidirectional reading strategy from a cognitive perspective, and uses the reciprocal influence between the arguments to dynamically compute the attention vectors. We further utilize a sparse-constrained neural tensor network to excavate the indicative and informative interactions between two discourse arguments.

We conduct various experiments on PDTB and the results show that dynamically incorporating mutual attention into discourse argument representations and adopting sparse-constrained neural tensor network to excavate specific pair patterns are effective.

This work is supported by the National Natural Science Foundation of China (61976154), the Tianjin Natural Science Foundation (18JCYBJC15500), the National Key R&D Program of China(2018YFB1305200), the Tianjin Municipal Science and Technology Project (18ZXZNGX00330).