计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200800115

基于数据增强的中文隐式篇章关系识别方法

王体爽　李培峰　朱巧明
苏州大学计算机科学与技术学院　江苏 苏州 215006
江苏省计算机信息技术处理重点实验室　江苏 苏州 215006
(20175227090@stu.suda.edu.cn)

摘　要　由于缺乏显式连接词,隐式篇章关系识别是一个具有挑战性的任务。文中提出了一种结合主动学习和多任务学习来间接扩充隐式篇章关系训练数据的隐式篇章关系识别方法,旨在在增强训练数据的同时尽量少地引入伪隐式篇章关系数据中的噪声。首先,基于BERT模型通过主动学习方法的分类不确定性来选择部分显式篇章关系样本;然后,移除显式篇章关系数据中的显式连接词作为伪隐式篇章关系数据;最后,采用多任务学习方法使伪隐式篇章关系数据有助于隐式篇章关系识别。在中文篇章树库(CDTB)上进行的实验的结果显示,相比基准模型,所提方法在宏平均F1、微平均F1值上均得到了提高。

关键词:篇章分析;隐式篇章关系识别;主动学习;多任务学习
中图法分类号　TP391

Chinese Implicit Discourse Relation Recognition Based on Data Augmentation

WANG Ti-shuang,LI Pei-feng and ZHU Qiao-ming
School of Computer Sciences and Technology,Soochow University,Suzhou,Jiangsu 215006,China
Provincial Key Laboratory for Computer Information Processing Technology,Suzhou,Jiangsu 215006,China

Abstract　Due to the lack of connectives,implicit discourse relation recognition is a challenging task,especially in Chinese. This paper proposes a method for Chinese implicit discourse relation recognition,which expands the training data by combining active learning and multi-task learning method. This method aims to reduce the noise as much as possible when it expands the training data set. Firstly,the active learning is used to select some explicit data through the classification uncertainty based on BERT,and then the connectives in the explicit data are removed and regarded as pseudo-implicit training data. Finally,a multi task learning method is used to boost implicit discourse relation recognition by using the pseudo-implicit training data. Experimental results on Chinese discourse treebank (CDTB) show that our method improves the macro-average F1 and micro-average F1 scores,compared with the baselines.

Keywords　Discourse parsing,Implicit discourse relation recognition,Active learning,Multi-task learning

1　引言

近年来,随着自然语言处理研究的重点逐渐从浅层的分词、句法分析延伸到更深层次的语义理解,研究对象也从词语、短语和句子延伸到了段落和篇章。篇章分析是自然语言处理中的一个基本任务,主要目的是分析两个篇章单元(称作论元)Arg1和Arg2之间的关系和层次结构,从而构建完整的篇章结构树。作为篇章分析中的一个重要子任务,篇章关系识别的目的在于识别论元之间的语义逻辑关系。自动识别篇章关系有利于篇章分析向实用化方向发展,也可以为自然语言处理许多下游任务提供帮助,如问答系统[1]、机器翻译[2]等。

篇章关系根据论元之间是否存在显式连接词(如所以、并且等)分为显式和隐式篇章关系。显式连接词是篇章关系识别中的一个重要线索,对于存在显式连接词的显式篇章关系识别而言,简单的显式连接词映射就能达到很高的识别准确率[3]。而对于缺乏显式连接词的隐式篇章关系识别,则完全依赖于对文本的理解,是一个极具挑战性的任务[4]。如例1所示,虽然论元之间没有逻辑连接词,但是可以根据前后两句的内容推断出Arg1是Arg2的原因,从而得出论元之间存在因果关系。

例1
[Arg1]越来越多的韩国企业正在看好大连,
[Arg2]韩国对大连的投资已连续三年保持持续增长。
[篇章关系]因果关系

到稿日期:2020-08-18　返修日期:2021-01-21　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61836007,61772354,61751206);江苏高校优势学科建设工程资助项目(PAPD)
This work was supported by the National Natural Science Foundation of China(61836007,61772354,61751206) and Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions(PAPD).
通信作者:李培锋(pfli@suda.edu.cn)

本文主要针对中文隐式篇章关系识别展开研究。目前，英文中大部分隐式篇章关系识别工作可以分为两大类：一种是仅使用隐式篇章数据，通过对论元编码之后，使用论元交互层捕获论元之间的语义关系[5-8]；另一种是引入额外数据，直接或间接扩充训练数据，以解决数据稀疏问题[9-10]。而中文隐式篇章关系识别的相关研究较少，大规模语料库匮乏导致的数据稀疏问题是阻碍中文隐式篇章关系识别发展的重要原因。

受 Xu 等[9]和 Liu 等[10]的启发，本文首先使用主动学习方法，基于 BERT 模型通过分类不确定性选择部分显式篇章关系样本，然后移除显式篇章关系样本中的显式连接词作为伪隐式篇章关系数据，最后采用多任务学习方法使得伪隐式篇章关系数据辅助完成隐式篇章关系识别。在中文篇章树库(Chinese Discourse Treebank，CDTB)[11]上进行的隐式篇章关系识别实验的结果显示，本文方法在宏平均 F1 值、微平均 F1 值以及各个篇章关系类别上的 F1 值均得到了提高。

2 相关工作

目前，许多工作都是基于英文修辞结构理论篇章树库(Rhetorical Structure Theory Discourse Treebank，RST-DT)[12]和宾州篇章关系树库(Penn Discourse Treebank，PDTB)[13]的，使用传统机器学习或者神经网络方法对英文隐式篇章关系识别任务进行相关研究。在英文隐式篇章关系识别任务中，无论是传统机器学习方法还是神经网络方法大都可分为以下两种方式。

(1)基于论元交互的方法。首先把论元中的每个词表示为词向量；其次使用编码层(如双向长短时记忆网络 Bi-LSTM，卷积神经网络 CNN 等)把论元编码为语义向量；然后使用一个论元交互层建模论元之间的语义联系；最后融合两个论元表示信息进行篇章关系识别。Chen 等[5]提出了一个带有门控单元的神经网络模型，捕获论元之间的线性和非线性交互信息，从而通过匹配矩阵来识别论元之间的语义关系。Liu 等[6]提出了一种基于多层注意力机制的交互模型，模拟人类在推导篇章语义关系时反复阅读论元的过程。将注意机制与外部记忆相结合，逐步找到有助于判断篇章关系的特定词汇。Jia 等[7]提出了一个记忆网络模型来建模论元之间的语义联系，解决了论元过长时出现的信息遗忘问题。Linh 等[8]意识到隐式连接词的重要性，同时预测关系类型和隐式连接词，并通过连接词和关系类型的嵌入来传递两个任务之间的有用信息。

(2)引入额外数据的半监督方法。隐式篇章关系识别的难点不仅在于论元之间缺乏显式连接词，数据稀疏问题也是制约其发展的原因之一。因此，有研究者将关注点放在了引入额外数据的半监督方法，通过自然标注的显式篇章数据或其他语料库数据直接或间接地扩充隐式篇章关系识别的训练集。Xu 等[9]基于主动学习的方法挑选部分显式篇章样本，并将其加入训练集合直接扩充训练语料。Liu 等[10]基于多任务学习模型，借助其他语料库同时训练 4 个任务，间接地扩充了训练集样本数量，通过共享参数的方式训练 CNN，充分挖掘多个相关数据之间的关联，增强了网络特征提取能力。

相比英文隐式篇章关系识别，由于大规模语料库的缺乏，中文隐式篇章关系识别的相关研究较少，大多是借鉴英文任务采用的方法。Kong 等[14]使用语义相似度、上下文、词汇和依存树等人工构建特征，采用最大熵模型，构建了一个端到端的篇章解析器。Xu 等[15]搭建了一个三层注意力神经网络模型，使用自注意力模型和交互注意力模型模拟人类双向阅读和重复阅读过程来识别中文隐式篇章关系。Xu 等[16]考虑到篇章主题能为篇章关系识别提供高层次的语义线索，提出了一个主题张量网络，通过神经主题模型推断论元的主题分布，捕获论元在句子级别和主题级别不同层面的交互信息。

中文隐式篇章关系识别研究工作主要在中文篇章树库(CDTB)上进行，较英文语料资源而言，CDTB 语料规模更小，数据稀疏问题更加严重。而已有的工作大多采用基于论元交互的方法，还没有采用引入额外数据的半监督方法。因此，本文借鉴 Xu 等[9]的工作，通过主动学习的方法挑选部分显式篇章关系样本，删除其显式连接词作为伪隐式篇章关系数据，以扩充隐式篇章关系训练数据。Xu 等是直接合并隐式篇章关系数据和伪隐式篇章关系数据，而 Sporleder 等[17]的研究表明，显式篇章数据和隐式篇章数据中词的分布和关系的分布都存在较大差异，应该看作是来自不同领域的数据，将其直接合并作为训练集合会引入噪声，反而会影响模型的性能。受 Liu 等[10]的激发，本文并不是直接合并隐式和伪隐式篇章关系数据，而是通过多任务学习的方法，让伪隐式篇章关系数据辅助完成隐式篇章关系识别，从而达到间接扩充训练语料的目的。Bi-LSTM 能够捕获全局特征，但对于序列过长的文本来说，仍然会造成数据遗忘问题。因此，Liu 等在多任务学习框架中采用了能捕获局部 n-gram 特征的 CNN。但是，相比英文语料库而言，在中文语料库 CDTB 中只标注了段内篇章关系，论元序列长度较短，平均每个基本篇章单元汉字长度为 22。因此，本文在多任务学习框架中采用特征捕获能力更强的 Bi-LSTM 网络，并且不会造成数据遗忘问题。此外，本文没有引入其他语料库中的数据，且没有引入额外的人工特征。实验结果表明，该方法可以有效提高中文隐式篇章关系的识别准确率。

3 基于主动学习和多任务学习的中文隐式篇章关系识别方法

本节首先介绍了基于 BERT 模型的主动学习方法，通过主动学习方法选取出部分显式篇章关系数据；然后介绍了多任务学习方法，通过多任务学习让显式篇章关系数据辅助完成隐式篇章关系识别。

3.1 基于 BERT 的主动学习方法

在自然语言处理领域，Devlin 等[18]提出的 BERT(Bidirectional Encoder Representation from Transformers)模型已被广泛应用于自然语言处理任务，如文本分类、问答、序列标注等。相比广泛使用的 Word2Vec[19]，BERT 模型采用双向 Transformer Encoder 作为编码器，它的特征表示在所有层中共同依赖于左右两侧上下文，可以获得更好的词向量表示。

而 Word2Vec 本身是一种浅层结构训练的词向量，受到窗口大小的限制，Word2Vec 学习到的语义信息有限。

一般的主动学习方法是先在已有的标注数据上训练基本模型，然后利用基本模型对未标注数据打标签，再迭代地从未标注语料库中优先选择富含有效信息的样本（即当前模型预测最不准确，分类不确定性较大），最后由人工标注后加入训练集重新训练模型。这种方法可以减少对模型帮助不大的样本标注工作，其工作流程如图 1 所示。

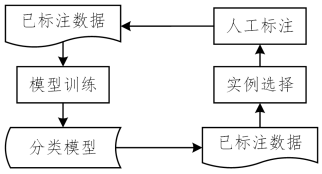

图 1 主动学习工作流程图

Fig. 1 Active learning process

目前使用较多的主动学习方法主要有基于委员会投票和不确定性抽样两种[20]。参考 Xu 等[9]的相关研究，本文采用基于不确定性度量方法选择部分显式篇章样本。该方法根据基本模型对显式篇章数据样本的分类置信度进行选择，样本分类置信度较低，说明基本模型尚不能很好地区分该样本，即基本模型缺乏该样本含有的信息。因此，将该样本加入训练集有助于调整基本模型。另外，选择样本之后不再人工标注数据，而是通过显式到隐式的关系映射，直接利用显式篇章关系样本原有的标签，然后移除显式连接词作为伪隐式篇章关系数据加入训练集。主动学习算法的流程如算法 1 所示。

算法 1 主动学习算法

Input：I：隐式篇章关系数据

E：显式篇章关系数据

m：迭代次数

φ(x)：效用函数，用于衡量样本可信度的函数

θ：预测置信度阈值

While(m)：

利用数据 I 训练模型 C

利用模型 C 预测 E 中的所有样本

根据效用函数 φ(x)从 E 中选择伪隐式篇章样本集合 E′

if (E′为空)：

停止迭代

更新数据集

E=E−E′

本文选择熵作为不确定性度量机制[9,21]，以此来指导样本的选择，效用函数如式(1)所示：

$$\varphi(x_i)=\sum_{r_j\in R} I_{r_j}(x_i;M)=\sum_{r_j\in R}-P(r_j\mid x_i)\log P(r_j\mid x_i) \quad (1)$$

其中，r_j 表示关系类别 j，对于一个样本 x_i，利用基本模型 M 对其进行分类，$P(r_j\mid x_i)$ 表示样本 x_i 分类为类别 j 的概率，然后计算该样本的熵（即混乱度），熵越大表明模型对该样本分类的不确定性越大，那么将该样本加入训练集，对模型调整的贡献就越大。

3.2 多任务学习方法

目前，大多数在 PDTB 语料库上基于显式篇章数据的半监督方法的相关研究都是直接利用显式数据来调整模型参数，Mikolov 等[19]指出，直接合并显式篇章数据和隐式篇章数据作为训练语料容易引入噪声，在集成大量的显式篇章数据时效果并不理想。因此，在中文隐式篇章关系识别任务中，如何使用伪隐式篇章数据成为了一个难点。本文尝试使用多任务学习方法，借助伪隐式篇章关系数据来辅助完成隐式篇章关系识别。

多任务学习指多个任务同时进行学习，在模型训练时，网络模型通过共享底层特征，相互促进学习，从而提高模型的泛化能力。另外，多任务学习可以间接扩充训练数据，这也是语料增强的一种方案。

基于 1.2 节中的主动学习方法选择出的伪隐式篇章关系数据和原始隐式篇章关系数据，搭建了图 2 所示的网络模型结构。

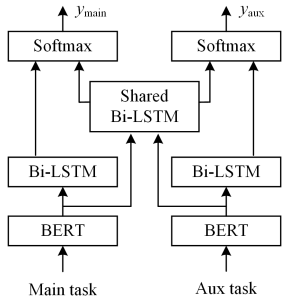

图 2 多任务学习网络结构图

Fig. 2 Multi-task learning structure

图 2 中，主任务（Main task）为隐式篇章关系识别，辅助任务（Aux task）为伪隐式篇章关系识别。通过 BERT 模型获取词向量表示后，每个任务有一个私有的 Bi-LSTM，用于提取各自任务的私有特征，此外还有一个共享的 Bi-LSTM 用于提取主任务和辅助任务的共享通用特征，以达到利用伪隐式篇章关系数据辅助完成隐式篇章关系识别的目的，从而间接扩充训练语料。

首先通过 BERT 模型获取词向量矩阵 $\boldsymbol{X}=[x_1,x_2,\cdots,x_n]$，其中 n 为词序列长度，然后将其分别输入到私有 Bi-LSTM 和共享 Bi-LSTM 以提取私有特征和共享通用特征，如式(2)－式(4)所示：

$$h_t^{\text{main}}=\text{Bi-LSTM}(h_{t-1}^{\text{main}},X^{\text{main}}) \quad (2)$$

$$h_t^{\text{aux}}=\text{Bi-LSTM}(h_{t-1}^{\text{aux}},X^{\text{aux}}) \quad (3)$$

$$h_t^{\text{share}}=\text{Bi-LSTM}(h_{t-1}^{\text{share}},X^{\text{main}}\oplus X^{\text{aux}}) \quad (4)$$

在捕获到私有特征和共享通用特征后，将其向量表示进行拼接，然后输入到 Softmax 函数进行篇章关系分类，如式(5)－式(6)所示：

$$y^{\text{main}}=\text{Softmax}([h^{\text{main}},h^{\text{share}}]) \quad (5)$$

$$y^{\text{aux}}=\text{Softmax}([h^{\text{aux}},h^{\text{share}}]) \quad (6)$$

相比直接合并隐式篇章关系数据和伪隐式篇章关系数据，多任务学习方法不仅可以学习到不同任务之间的共同特征，还能避免引入辅助任务数据中过多的噪声，而且可以更好地学习到主任务原始语料的私有特征。

4 实验

本节首先介绍实验用到的语料库 CDTB 以及详细的实验设置，随后给出并分析实验结果。

4.1 CDTB 语料库

针对中文篇章的一般特点，Li 等[11]结合 RST-DT 的树形结构和 PDTB 的显式连接词处理方法，基于宾州大学中文树库(Penn Chinese Treebank，CTB)[22]建立了中文篇章树库 CDTB。在 CDTB 语料库中，每一个段落解析为一棵链接依存树，称作篇章结构树。

CDTB 共标注了 5496 个隐式篇章关系和 1 812 个显式篇章关系，关系类型分为两层，包含 4 大类和 17 小类。从大类来看，并列类占比最大，显式篇章关系中并列类占 40.7%，隐式篇章关系中并列类占 50.3%，可见 CDTB 语料库存在较严重的数据稀疏问题。CDTB 语料库包括 500 篇新闻文档，共有 2342 个段落，10650 个子句，每个子句的平均汉字长度为 22。

4.2 实验设置

参考 Xu 等[15]的相关工作，本文选择相同的 450 篇文章作为训练集，50 篇文章作为测试集，并将所有的非二叉树转换为左连接的二叉树。转换后的 4 类篇章关系统计信息如表 1 所列。由于转折类在所有篇章关系样本中的占比极低，在隐式篇章关系中，转折类训练集样本数量只有 40 个，这显然很难通过深度学习框架学习到相应的语义特征。另外，转折类测试集样本数量只有 1 个，对模型的评估也不客观。因此，本文跟随之前的工作[12-13]在实验时移除了占比极低的转折类，只在上层的三大类篇章关系上做相关实验分析，报告各个篇章关系类别的 F1 值，以及宏平均 F1 值和微平均 F1 值。本文在主动学习方法和多任务学习方法上使用的显式篇章关系样本和隐式篇章关系样本均来自 CDTB 语料库。

表 1 CDTB 中篇章关系的统计信息

Table 1 Statistical information of discourse relations in CDTB

关系	训练集			测试集		
	显式	隐式	总计	显式	隐式	总计
因果	426	787	1213	40	79	119
并列	1030	3588	4618	118	397	515
解说	190	1275	1465	11	140	151
转折	165	40	205	10	1	11

本文使用的 BERT 均为哈尔滨工业大学提供的预训练好的 BERT 模型，网络结构一共有 12 层，隐藏层有 768 维，采用 12 头模式，共有 110M 个参数。另外，为了防止过拟合，在输入到 Softmax 层之前，对其进行 dropout 操作[23]，dropout 设置为 0.2。

4.3 实验结果

为了检验本文方法的有效性，本文选取了 BERT 模型作为基准，此外还包括 3 个已有工作中的先进模型作为基准。1)Kong 等[14]使用语义相似度、上下文、词汇和依存树等人工构建特征，采用最大熵模型，构建了一个端到端的篇章解析器。2)Liu 等[6]提出了一种基于多层注意力机制的交互模型，模拟人类在推导篇章语义关系时反复阅读论元的过程，将注意机制与外部记忆相结合，逐步找到有助于判断篇章关系的特定词汇。3) Xu 等[15]搭建了一个三层注意力神经网络模型，使用自注意力模型和交互注意力模型模拟人类双向阅读和重复阅读过程，以识别中文隐式篇章关系，该模型在 CDTB 语料库三大类隐式篇章关系识别任务上取得了目前最优的性能。4)BERT：仅使用隐式篇章关系训练数据，采用基本的 BERT 模型。在 CDTB 上进行的三分类实验结果如表 2 所列。

表 2 对比实验结果

Table 2 Comparative experimental results

模型	因果	并列	解说	宏 F1	微 F1
Kong	32.4	77.3	51.8	54.8	67.5
Liu	29.6	79.0	53.9	54.4	68.0
Xu	31.6	79.4	57.6	56.2	68.7
BERT	50.0	84.4	68.2	67.8	77.0
Ours	**56.9**	**85.7**	**71.9**	**72.1**	**79.4**

本文共选取了 4 个公开的有向网络数据集进行实验。从表 2 可以看出，得益于 BERT 的强大，基本的 BERT 模型在性能上已经大幅超越了其他 3 个基本模型，相比之前最好的模型 Xu，BERT 模型在宏平均 F1 上提升了 11.6%，微平均 F1 提升了 8.3%。虽然 BERT 模型明显优于其他 3 个基本模型，但是本文方法相比 BERT 模型在性能上也有一定的提升，宏平均 F1 值和微平均 F1 值分别提升了 4.3%和 2.4%。

4.4 实验分析

为了更好地检验本文方法的有效性，本文构建变体模型来进行如下实验。1)BERT：仅使用隐式篇章关系训练数据来训练 BERT 模型。2)Blender：将所有的显式篇章关系样本移除显式连接词得到伪隐式篇章关系样本后，全部直接加入训练集。3)AL：通过主动学习方法，选择部分对模型调整有帮助(分类不确定性较大)的样本，移除显式连接词后直接加入训练集。4)BM：使用全部移除显式连接词后的伪隐式篇章关系数据，通过多任务学习方法进行隐式篇章关系识别。5)AM：对于隐式篇章关系数据和通过主动学习方法选择出的伪隐式篇章关系数据，采用多任务学习方法进行隐式篇章关系识别。为了公平地对比不同的变体模型，实验中所有的变种模型均采取相同的超参数。实验结果如表 3 所列。

表 3 变体模型的实验结果

Table 3 Experimental results of variant model

模型	因果	并列	解说	宏 F1	微 F1
BERT	50.0	84.4	68.2	67.8	77.0
Blender	45.3	84.2	64.7	66.3	76.5
AL	52.4	84.9	70.9	69.5	77.4
BM	51.5	85.2	70.7	70.1	77.9
AM	**56.9**	**85.7**	**71.9**	**72.1**	**79.4**

从表 3 可以看出，基本的 BERT 模型已达到较高的识别准确率。对比 BERT 和 Blender 的实验结果可以看出，简单地将所有显式篇章关系样本全部加入训练集，不仅对隐式篇章关系的识别没有帮助，反而会使其性能下降。这主要是因为显式篇章数据和隐式篇章数据之间存在差异，容易引入噪声。

对比 BERT 和 AL 两个模型可以发现，本文使用主动学习方法挑选部分显式篇章关系样本并将其加入训练集的方式，在一定程度上可以减少噪声的引入，并帮助调整基本模型，其宏 F1 提升了 1.7%，微 F1 提升了 0.4%。通过主动学习方法共从显式篇章数据中选取出 38.6%的显式篇章关系样本作为伪隐式篇章关系数据，具体数据如表 4 所列。

表 4 主动学习方法选择样本的详细数据

Table 4 Sample data selected by active learning method

关系	伪隐式	显式	占比/%
因果	318	426	74.6
并列	267	1 030	25.9
解说	51	190	26.8
总计	636	1 646	38.6

从表 4 可以看出，因果关系选择出较大比例的显式篇章样本的主要原因是，基本模型本身对因果关系的识别性能较差，这使得在使用基本模型对显式篇章关系样本进行识别时，也会对因果关系的识别带来较大的不确定性。根本原因是隐式篇章因果关系训练的样本较少，难以通过训练捕获足够的语义特征，而并列关系有足够多的训练数据，可以使基本模型达到较高的识别率，因此选择出的样本占比相对来说较低。受语料规模的影响，基本模型在解说关系上的识别率不如并列关系高，但是由于解说关系的两个论元之间通常存在更强的语义关联，解说关系本身就是一个论元对另一个论元的进一步的解释说明，因此其识别不确定性较低，故选择出的伪隐式解说关系样本的占比也较低。

对比 Blender 和 BM 两个模型可以发现，多任务学习方法较直接混合训练集在各个指标上均有明显的提升，这是因为多任务学习方法通过伪隐式篇章关系数据辅助完成隐式篇章关系识别，通过多任务学习框架可以保留原始私有特征，并且可以学习到伪隐式篇章关系数据中的共享信息。

从表 3 可以看出，结合主动学习方法和多任务学习方法的 AM 模型的整体性能最优。该方法能有效提升因果关系识别的准确率，这是因为对于因果关系而言，训练集中的隐式样本占比最少(13.9%)，而伪隐式样本的占比最多(50.0%)，若不区分两种训练语料，直接将其混合(模型 AL)作为训练集，则会由于两个数据集的差异，使得因果关系引入的噪声更多。另外，使用全部移除显式连接词后的伪隐式篇章关系数据，通过多任务学习方法进行隐式篇章关系识别(模型 BM)会使得部分数据产生偏差，对于某些伪隐式样本而言，删除显式连接词后的关系类型可能会发生变化，如例 2 所示，移除显式连接词“从而”后，篇章关系类型也可以为并列关系，该样本可能会带来噪声。因此，结合主动学习和多任务学习方法，不仅可以过滤掉一部分容易产生歧义的样本，而且可以使得模型能够区分两种训练数据的不同，从而更好地使用伪隐式篇章样本来帮助隐式篇章关系识别。

例 2

[Arg1] 影响美国商人的对华投资信心，

[Arg2] 从而也影响到美国人的就业机会。

[篇章关系] 因果关系

在主动学习方法中，通过实验发现，由于数据规模不大，迭代次数超过 5 次时，便不会有新的伪隐式篇章关系样本被选择出来，因此本文将迭代次数 m 设置为 5。另外，若预测置信度阈值 θ 设置得越小，选择样本就越多，则容易引入过多噪声；若 θ 设置得越大，选择样本越少，模型则无法学习到决策边界的样本，使得模型泛化能力较弱。通过对比实验，本文将预测置信度阈值 θ 设置为 0.9，在所有变体模型中均使用相同的超参数。

表 5 列出了本文方法在中文隐式篇章关系识别任务上错误识别的关系类型分布。可以看出，36.7% 的因果关系和 20.0% 的解说关系被错误地识别为并列关系，主要原因有以下两点：1)并列关系占到了一半以上；2)一些样本之间虽然不是并列关系，但是在语义层面较为相似，容易被识别为并列关系。如例 3 所示，例子中 Arg1 和 Arg2 语义的相似度较高，模型容易错误地捕获到其语义相似度特征，从而将其错误地识别为并列关系。

表 5 识别错误样本比例

Table 5 Proportion of wrong samples identified

(单位：%)

关系	因果	并列	解说
因果	—	36.7	13.9
并列	3.8	—	10.6
解说	2.9	20.0	—

例 3

[Arg1] 俄罗斯 1997 年估计增长百分之零点五，

[Arg2] 1998 年预计增长百分之一点五。

[篇章关系] 顺承关系

结束语 在英文隐式篇章关系识别任务上，相关研究证明了语料增强方法的有效性，而在中文微观隐式篇章关系识别任务上还没有语料增强方法的相关研究。相比英文语料库，中文语料库的规模较小，数据稀疏问题更加严重。另外，由于移除显式连接词后，伪隐式篇章关系样本的关系类型可能会发生变化，且显式篇章关系样本和伪隐式篇章关系样本的数据分布不同。因此，本文提出基于主动学习方法和多任务学习方法来间接扩充训练语料。首先使用主动学习方法，基于 BERT 模型，通过分类不确定性来选择部分显式篇章关系样本，然后移除显式篇章关系样本中的显式连接词作为伪隐式篇章关系数据，最后采用多任务学习方法使得伪隐式篇章关系数据辅助完成隐式篇章关系识别。实验结果表明，本文方法在宏平均 F1、微平均 F1 上达到了目前最好的性能。在今后的研究中，可以尝试结合其他语料资源来进一步扩充训练数据，另外也可以尝试使用迁移学习方法来间接扩充训练数据。

参 考 文 献

[1] LIAKATA M, DOBNIK S, SAHA S, et al. A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task [C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 747-757.

[2] TU M, ZHOU Y, ZONG G. Enhancing grammatical cohesion: Generating transitional expressions for SMT[C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014: 850-860.

[3] XUE N, TOU H, PRADHAN S. The CoNLL-2016 shared task on shallow discourse parsing[C]// Proceedings of the 20th Conference on Computational Natural Language Learning-Shared Task. 2016: 1-19.

[4] PILER E, LOUIS A, NENKOVA A. Autimatic sense prediction

for implicit discourse relation 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. 2009:683-691.

[5] CHEN J, ZHANG Q, LIU P. Implicit Discourse Relation Detection via a Deep Architecture with Gated Relevance Network [C]// Proceedings of the 54nd Annual Meeting of the Association for Computational Linguistics. 2016:1726-1735.

[6] LIU Y, LI S. Recognizing Implicit Discourse Relations via Repeated Reading Neural Networks with Multi-Level Attention Building Chinese discourse corpus with connective-driven dependency tree structure[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016:1224-1233.

[7] JIA Y, YE Y, FENG Y, et al. Modeling discourse cohesion for discourse parsing via memory network[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018:438-443.

[8] NGUYEN L T, NGO L U, THAN K, 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.

[9] XU Y, HONG Y, RUAN H, et al. Using Active Learning to Expand Training Data for Implicit Discourse Relation Recognition [C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018:725-731.

[10] LIU Y, LI S, ZHANG X. Implicit Discourse Relation Classification via Multi-task Neural Networks[C]// Proceedings of the 2016 AAAI Conference on Artificial Intelligence. 2016: 2750-2756.

[11] LI Y, KONG F, ZHOU G. Building Chinese discourse corpus with connective-driven dependency tree structure[C]// In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014:2105-2114.

[12] CARLSON L, OKUROWSKI M E, MARCU D. RST discourse treebank[M]. Linguistic Data Consortium, University of Pennsylvania, 2002.

[13] RASHMI P, ELENI M, NIKHIL D, et al. The Penn Discourse Treebank 2. 0 Annotation Manual[OL]. https://www. seas. upenn. edu/~pdtb/PDTBAPI/pdtb-annotation-manual. pdf.

[14] KONG F, ZHOU G. A CDT-styled end-to-end Chinese discourse parser[J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2017, 16(4):26.

[15] XU S, WANG T S, LI P F, et al. Multi-Layer Attention Network Based Chinese Implicit Discourse Relation Recognition [J]. Journal of Chinese Information Processing, 2019, 27(3): 12-19.

[16] XU S, LI P, ZHU Q, et al. Topic tensor network for implicit discourse relation recognition in Chinese[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019:608-618.

[17] SPORLEDER C, LASCARIDES A. Using Automatically Labelled Examples to Classify Rhetorical Relations: an Assessment [J]. Natural Language Engineering, 2008, 14(3):369-416.

[18] DEVLIN J, CHANG M, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [J]. arXiv:1810. 04805.

[19] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]// Advances in Neural Information Processing Systems. 2013:3111-3119.

[20] SELLTES B. Active Learning Literature Survey[R]. Computer Sciences Technical Report, University of Wisconsin-Madison, 2009.

[21] RAMIREZ-LOAIZA M E, SHARMA M, KUMAR G, et al. Active learning: an empirical study of common baselines[J]. Data Mining and Knowledge Discovery. 2017, 31(2):287-313.

[22] XUE N, XIA F, CHIOU F, et al. The Penn Chinese treebank: Phrase structure annotation of a large corpus[J]. Natural Language Engineering, 2005, 11(2):207-238.

[23] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. arXiv:1207. 0580, 2012.

WANG Ti-shuang, born in 1993. His main research interests include natural language processing and machine learning.

LI Pei-feng, born in 1971, Ph.D, professor, Ph. D supervisor, is a member of China Computer Federation. His main research interests include natural language processing and machine learning.