计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201000099

基于知识蒸馏的隐式篇章关系识别

俞 亮 魏永丰 罗国亮 邬昌兴
华东交通大学软件学院 南昌 330013
(yu3liang@qq.com)

摘 要 由于缺少连接词信息,隐式篇章关系识别模型需要基于两个论元(子句或者句子)的语义来推导它们之间的篇章关系,但目前性能还比较低。对于语料标注人员而言,隐式篇章关系的标注是很困难的,他们通常先插入一个合适的连接词用于辅助隐式篇章关系的标注。基于上述情况,文中提出了一种基于知识蒸馏的隐式篇章关系识别方法,其目的是利用语料标注时插入的连接词信息来提高识别的性能。具体地,先构建一个连接词增强的模型用于融合连接词信息,然后基于知识蒸馏的方式把连接词增强模型学到的知识迁移到隐式篇章关系识别模型中。实验结果表明,在常用的PDTB数据集上,所提方法取得了比同类基准方法更好的识别性能。

关键词:隐式篇章关系识别;知识蒸馏;连接词;篇章结构分析;深度学习
中图法分类号 TP391.1

Knowledge Distillation Based Implicit Discourse Relation Recognition

YU Liang,WEI Yong-feng,LUO Guo-liang and WU Chang-xing
School of Software,East China Jiaotong University,Nanchang 330013,China

Abstract Due to the lack of connectives,implicit discourse relation recognition models infer the semantic relations (e.g.,causal) between two arguments (clauses or sentences) based on their semantics. The performance of these models is still relatively low. It is also very difficult for corpus annotators to annotate implicit discourse relations. They usually insert an appropriate connective to assist the annotation of an implicit discourse relation instance. Considering the above,a knowledge distillation based method is proposed for implicit discourse relation recognition to take use of the connectives inserted during corpus annotating. Specifically,a connective-enhanced model is constructed to integrate the connective information,and then the integrated connective information is transferred to the implicit discourse relation recognition model via knowledge distillation. Experimental results on the commonly used PDTB dataset show that the proposed method achieves better performance than the baselines.

Keywords Implicit discourse relation recognition,Knowledge distillation,Connective,Discourse structure analysis,Deep learning

1 引言

篇章一般指由一系列结构衔接、语义连贯的语言单位(句子或子句),按照一定的语义关系或者层次结构组成的整体语言单位[1]。通常把句子或子句之间的语义关系称作篇章关系,如因果关系、转折关系等。篇章关系识别指的是自动判断两个论元(句子或子句)之间的语义关系,它是篇章结构分析的核心子任务之一,也是其性能瓶颈所在[2-3]。篇章关系识别性能的提高不但能够促进篇章结构分析的发展,也有利于众多下游的自然语言处理任务,如机器翻译、情感分析、问答系统和文本摘要等。

篇章连接词(如因为、但是等)是篇章关系识别中最重要的特征之一。当两个论元有篇章连接词相连时,显式篇章关系识别仅使用连接词作为特征就能达到90%以上的分类准确率[4]。反之,当两个论元之间省略了篇章连接词时,隐式篇章关系识别需要根据两个论元的语义推导它们之间的关系,这是一个极具挑战性的任务,其准确率目前仅在60%左右[5]。如图1所示,隐式篇章关系实例的两个论元之间省略了连接词[1)]"所以",则需要基于文本"积水"和"没去打篮球"来推导它们之间语义上的"因果关系",而这是非常困难的。实际上,即使是语料标注人员也常利用连接词信息来辅助隐式篇章关系的标注。例如,目前规模最大的宾州篇章树库(Penn Discourse TreeBank,PDTB)[6]在标注时,要求标注人员首先在隐式篇章关系实例的两个论元之间插入一个合适的

1) 为了表达的简洁,在不造成理解歧义的情况下,文中把"篇章连接词"简写为"连接词"

到稿日期:2020-10-18 返修日期:2021-03-10 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:国家自然科学基金项目(61866012);江西省自然科学基金项目(20181BAB202012);江西省教育厅科学技术研究项目(GJJ180329)
This work was supported by the National Natural Science Foundation of China(61866012),Natural Science Foundation of Jiangxi Province (20181BAB202012) and Science and Technology Research Project of Jiangxi Education Department(GJJ180329).
通信作者:邬昌兴(wuchangxing@ecjtu.edu.cn)

连接词，然后综合论元和插入的连接词两个方面的信息来判断该实例的篇章关系。也就是说，篇章语料标注人员常使用(插入的)连接词信息来辅助隐式篇章关系的标注。

论元1：操场有积水，
论元2：我们没去打篮球。
插入的篇章连接词：所以
标注的篇章关系：因果关系

图1 隐式篇章关系实例

Fig. 1 Instance of implicit discourse relation

从以上的分析可知：一方面，基于连接词的显式篇章关系识别与基于论元语义的隐式篇章关系识别之间存在巨大的性能差距(90%与60%)；另一方面，语料的标注过程也说明了连接词信息对隐式篇章关系识别是有帮助的。因此，一些研究人员尝试在隐式篇章关系识别模型中利用连接词信息，以提高识别的性能。早在2010年，Zhou等[7]就提出了一种两阶段的方法来模拟篇章语料的标注过程，其首先基于语言模型预测适用于连接两个论元的连接词，然后综合预测的连接词信息和论元来判断篇章关系。为了利用篇章语料标注时插入的连接词信息，Qin等[8]提出了一种基于对抗学习的方法，通过特征模仿的方式把连接词信息迁移到隐式篇章关系识别模型中。Bai等[9]基于多任务学习的方法利用插入的连接词信息，把隐式篇章关系识别和连接词分类看作两个不同但高度相关的任务，通过在这两个任务之间共享输入层和特征抽取层以达到信息共享、共同提高的目的。Nguyen等[10]在多任务学习框架下利用连接词与篇章关系之间存在的映射关系迁移知识。最近，Wu等[11]提出把语料中标注的第一级篇章关系、第二级篇章关系和插入的连接词(可看作第三级篇章关系)看作一个序列，然后基于层次多任务学习神经网络在三级篇章关系之间共享信息，最后叠加一个神经网络CRF(Conditional Random Fields)层用于预测篇章关系序列。上述研究工作中提出的方法都取得了比相应基准方法更好的性能，这进一步证明了连接词信息对隐式篇章关系识别的重要性。

本文提出了一种基于知识蒸馏的隐式篇章关系识别方法，利用篇章语料标注时插入的连接词信息来提高识别的性能。目前，知识蒸馏(knowledge distillation)相关方法已经成功应用于自然语言处理和计算机视觉等领域的众多任务中[12-13]。知识蒸馏的核心是知识迁移，即通常把一个大模型/强模型(常称作教师模型)的知识迁移到小模型/弱模型(常称作学生模型)中[14]。具体地，本文首先构建一个连接词增强的教师模型，以“两个论元＋插入的连接词”作为输入进行分类，用于把连接词信息融合到模型中；然后，构建一个隐式篇章关系识别模型作为学生模型，仅以“两个论元”作为输入，其训练时不仅从标注的训练数据中学习知识，而且还从教师模型中学习融合了连接词信息的知识。文献[8]也使用了知识迁移的方法利用插入的连接词信息，本文方法的不同之处主要体现在以下3个方面：1)文献[8]仅在特征层迁移知识，而本文方法同时在特征层和分类层迁移知识。2)文献[8]基于对抗训练的方法迁移知识，而本文基于知识蒸馏的方法迁移知识；相比而言，基于知识蒸馏的方法更简单，只需要对模型训练的代价函数稍作修改即可。3)本文提出的方法的性能明显优于文献[8]提出的方法的性能。

在常用的PDTB数据集上的实验结果表明，本文提出的基于知识蒸馏的隐式篇章关系识别方法在第一级和第二级篇章关系上都取得了比同类基准方法更好的性能效果。

2 相关工作

2.1 隐式篇章关系识别

隐式篇章关系识别通常被看作一个多分类问题。早期，研究人员一般先设计大量的人工特征，然后基于支持向量机等统计机器学习方法训练分类模型。常用的人工特征包括词相关特征[15-16]、词对和句法特征[17]、实体特征[18]和词类对特征[19]等。近年来，研究人员提出了大量基于神经网络自动学习分布式特征的方法。例如，为了更好地学习论元的语义表示，研究人员提出了浅层卷积神经网络模型[20]、融合实体信息的递归神经网络模型[21]、融合段落级别上下文信息的模型[22]、基于注意力机制的模型[23-24]、融合多级别信息的模型[9]和基于Bert的模型[5]等。为了更好地建模论元之间的交互，研究人员提出了门控相关性神经网络[25]、重复读取神经网络[26]、词级别交互式的注意力机制[27]和基于注意力机制的张量网络[28-29]等。基于神经网络自动学习分布式特征的隐式篇章关系识别方法的性能全面超过了传统基于人工特征的方法，成为当前研究的热点。

针对隐式篇章关系训练语料比较少的问题，Marcu等[30]认为大量显式篇章关系实例可看作由连接词自然标注的语料(例如，“但是”表示转折关系)，移除连接词并映射成相应篇章关系后就可以作为额外的训练语料。然而，Sporleder等[31]指出直接使用这些额外的训练语料会导致识别性能的下降，因为显式篇章关系数据和隐式篇章关系数据之间具有较大的差异，而且移除连接词可能会使语义关系发生改变。随后，为了有效地利用大量的显式篇章关系数据，一些研究人员采用多任务学习的方法[32-33]。具体地，他们定义显式篇章关系实例上的连接词分类任务作为辅助任务，用于提高隐式篇章关系识别这个主任务的性能。一些研究人员基于数据筛选的方法挑选显式篇章关系实例，以扩充训练语料[34-35]。具体地，他们基于某种标准选择与隐式篇章关系实例相似的显式实例，直接用于扩充训练语料。还有一些研究人员首先基于大量显式篇章关系数据学习专用词向量，然后将其用作隐式篇章关系识别模型的输入[36-37]。本文方法的特点是利用隐式篇章关系语料标注时插入的连接词信息，可以与上述利用显式篇章关系数据的研究工作互相补充。

2.2 知识蒸馏

近年来，知识蒸馏被广泛应用于计算机视觉和自然语言处理领域，用于在不同的模型之间迁移知识。例如，Hinton等[14]把知识蒸馏用于模型压缩，即把强大但复杂的模型(教师模型)的知识迁移到简单的模型(学生模型)中。其首先训练好教师模型，然后在训练学生模型时，要求学生模型尽量拟合训练数据的真实类别标记和复杂模型的预测结果，整个训练过程是在分类层之间迁移知识。随后，Yim等[38]提出在两个深度学习模型的对应特征层之间迁移知识，该模型在图片

分类等任务上取得了更快的收敛速度和更好的识别性能。Zhang 等[39]突破了教师-学生的知识蒸馏模式，提出了一种互学习模型，用于在多个模型之间互相迁移知识。Zeng 等[12]基于知识蒸馏把领域外的机器翻译知识迁移到领域内的翻译模型中。Liu 等[40]基于知识蒸馏在多个领域的机器阅读理解任务之间相互迁移知识。本文首次把知识蒸馏应用于隐式篇章关系识别，并同时在特征层和分类层之间迁移知识，取得了比同类基准方法更好的性能。

3　基于知识蒸馏的隐式篇章关系识别方法

在篇章语料的标注过程中，标注人员通常在隐式篇章关系实例的两个论元之间插入一个合适的连接词，用于辅助篇章关系的识别。提出基于知识蒸馏的隐式篇章关系识别方法的目的是利用语料标注时插入的连接词信息，以引导隐式篇章关系识别模型的训练，从而提高识别的性能。如图 2 所示，本文方法包括以下两步：1）以“两个论元＋插入的连接词”作为输入，构建一个连接词增强的隐式篇章关系识别模型作为教师模型，用于把连接词信息融合到模型中；2）仅以“两个论元”作为输入，构建一个普通的隐式篇章关系识别模型作为学生模型，训练时学生模型不仅从标注的训练数据中学习知识，而且还基于知识蒸馏的方式从教师模型中学习融合了连接词信息的知识。下面分别介绍教师模型、学生模型及其训练过程。

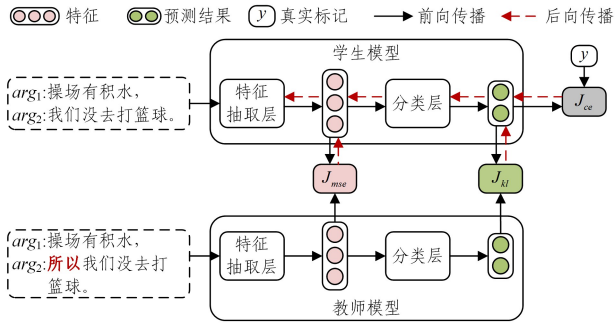


图 2　基于知识蒸馏的隐式篇章关系识别方法

Fig. 2　Knowledge distillation based implicit discourse relation recognition method

3.1　教师模型

训练语料 D 中的任一隐式篇章关系实例均可表示为 (x, c, y)，其中，$x=(arg_1, arg_2)$ 表示该实例的两个论元，c 表示插入的连接词，y 表示标注的隐式篇章关系，即真实的类别标记。

教师模型是一个连接词加强的隐式篇章关系识别模型，以论元 x 和标注时插入的连接词 c 为输入。经过特征抽取层后得到的特征可表示为 $v_t(x,c)$，经过分类层后得到的最终预测结果可表示为 $p_t(x,c)$。训练教师模型时，在语料 D 上最小化交叉熵分类代价函数：

$$J_t(\theta_t)=-\sum_{(x,c,y)\in D} E_y[\log p_t(x,c;\theta_t)] \tag{1}$$

其中，θ_t 为教师模型的参数，y 是真实标记的独热编码（One-hot Encoding），$E[\cdot]$ 表示预测结果关于真实标记的期望值。

教师模型模拟了人类标注隐式篇章关系的过程，在插入连接词 c 的辅助下，其识别性能远高于仅以论元 x 作为输入的学生模型（例如，其在 PDTB 语料的第一级隐式篇章关系分类任务上的准确率可以达到 85％以上）。这说明教师模型能很好地融合语料标注时插入的连接词信息。

3.2　学生模型

学生模型是一个普通的隐式篇章关系识别模型，仅以论元 x 为输入。经过特征抽取层后得到的特征可表示为 $v_s(x)$，经过分类层后得到的最终预测结果可表示为 $p_s(x)$。在训练学生模型时，为了使模型能尽可能地拟合训练数据 $(x,y)\in D$，最小化交叉熵分类代价函数如式（2）所示：

$$J_{ce}(\theta_s)=-\sum_{(x,y)\in D} E_y[\log p_s(x;\theta_s)] \tag{2}$$

其中，θ_s 为学生模型的参数。也就是说，通过最小化 $J_{ce}(\theta_s)$，可以使学生模型基于训练数据 $(x,y)\in D$ 学习，可用于隐式篇章关系识别的知识。

为了从教师模型中学习融合了连接词信息的分类知识，我们采用知识蒸馏的方法，其基本思想是让学生模型尽可能地模拟教师模型的行为。

一方面，希望学生模型和教师模型学到的特征 $v_s(x)$ 和 $v_t(x,c)$ 能尽可能地接近，从而实现两个模型在特征层的知识迁移。从教师模型在 PDTB 数据集上的识别性能远高于学生模型可以看出，特征 $v_t(x,c)$ 含有比特征 $v_s(x)$ 更多的对隐式篇章关系识别有用的信息。具体地，依据文献[41]中的方法，定义对应于特征层知识蒸馏的代价函数为：

$$J_{mse}(\theta_s)=\sum_{(x,y,c)\in D} MSE(v_t(x,c;\theta_t), v_s(x;\theta_s)) \tag{3}$$

其中，$MSE(\cdot)$ 表示均方误差。

另一方面，希望学生模型和教师模型最终的预测结果 $p_s(x)$ 和 $p_t(x,c)$ 能尽可能地接近，从而实现两个模型在分类层的知识迁移。以独热编码表示的真实类别标记 y 可看作是一种硬标记（hard label）；教师模型的预测结果 $p_t(x,c)$ 可以看作是一种软标记（soft label），通常认为其含有更多的类别信息，例如，类别之间的相似度信息[14]。具体地，定义对应于分类层知识蒸馏的代价函数为：

$$J_{kl}(\theta_s)=\sum_{(x,y,c)\in D} KL(p_t(x,c;\theta_t)\,|\,p_s(x;\theta_s)) \tag{4}$$

其中，$KL(\cdot\,|\,\cdot)$ 表示两个概率分布之间的 KL（Kullback-Leibler）距离。

最后，学生模型总的训练代价函数定义为交叉熵分类代价函数、对应于特征层的代价函数和对应于分类层的代价函数的线性求和：

$$J_s(\theta_s)=J_{ce}(\theta_s)+\mu * J_{mse}(\theta_s)+\lambda * J_{kl}(\theta_s) \tag{5}$$

其中，μ，λ 是相应部分代价的权重系数。

3.3　训练过程

算法 1 描述了基于知识蒸馏的隐式篇章关系识别方法的训练过程。整个训练过程分成两个阶段：第一阶段基于式（1）所示代价函数训练教师模型（步骤 1－5），第二阶段基于式（5）所示代价函数训练学生模型（步骤 6－13）。为了简洁，算法 1 中省略了基于验证数据集判定模型是否收敛的步骤。最终训练好的学生模型就是本文所提的隐式篇章关系识别模型。

算法 1　训练算法

输入：训练数据集 D，最大训练轮数 K_t，K_s。

输出：训练好的学生模型

1. 构造教师模型，并随机初始化参数 θ_t

2. 重复以下步骤：

3. 　从训练数据集 D 中取出一批实例{(x,c,y)}

4. 　最小化教师模型代价函数 $J_t(\theta_t)$，更新参数 θ_t

5. 直到：模型收敛或达到最大训练轮数 K_t

6. 构造学生模型，并随机初始化参数 θ_s

7. 重复以下步骤：

8. 　从训练数据集 D 中取出一批实例{(x,c,y)}

9. 　基于教师模型计算对应的特征$\{v_t(x,c;\theta_t)\}$

10. 　基于教师模型计算对应的预测$\{p_t(x,c;\theta_t)\}$

11. 　移除训练实例中的连接词得{(x,y)}

12. 　最小化学生模型代价函数 $J_s(\theta_s)$，更新参数 θ_s

13. 直到：模型收敛或达到最大训练轮数 K_s

4 基于 BiAttention 的分类模型

理论上，任何可用于隐式篇章关系识别的神经网络模型都可看作第 3 节中的教师模型和学生模型。本文实验中采用了一种基于双向注意力机制(Bi-directional Attention，BiAttention)的分类模型[42]，其常用于建模两个句子之间的语义关系，例如，文本蕴含识别、自动问答和句子语义匹配等。为了文章的完整性，下面对其进行简要介绍。

如图 3 所示，基于 BiAttention 的分类模型以 $x=(arg_1, arg_2)$ 为输入，预测它们之间的篇章关系。需要说明的是，教师模型把连接词 c 和论元 arg_2 拼接成一个整体作为输入。该模型主要包含基于 BiLSTM(Bi-directional Long Short Term Memory)的编码层、基于 BiAttention 的交互层、聚合层和分类层。

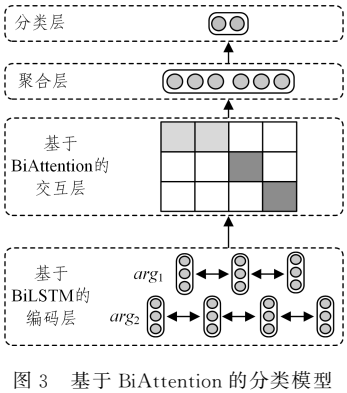


图 3 基于 BiAttention 的分类模型

Fig. 3 Classification model based on BiAttention

基于 BiLSTM 的编码层用于学习论元中的词在上下文中的表示，如下式所示：

$$h_1^1,\cdots,h_i^1,\cdots,h_m^1=\mathrm{BiLSTM}_1(x_1^1,\cdots,x_i^1,\cdots,x_m^1)$$
$$h_1^2,\cdots,h_j^2,\cdots,h_n^2=\mathrm{BiLSTM}_2(x_1^2,\cdots,x_j^2,\cdots,x_n^2) \tag{6}$$

其中，x_i^1，h_i^1 分别为论元 1 中的第 i 个词的词向量及其在上下文中的表示；x_j^2，h_j^2 分别为论元 2 中的第 j 个词的词向量及其上下文中的表示；m 和 n 分别是两个论元中词的个数。

基于 BiAttention 的交互层用于建模两个论元间的交互，即学习局部的篇章语义关系表示，如下式所示：

$$v_i^1=G([h_i^1;\tilde{h}_i^1]),\forall i\in[1,\cdots,m]$$
$$v_j^2=G([h_j^2;\tilde{h}_j^2]),\forall j\in[1,\cdots,n]$$
$$\tilde{h}_i^1=\sum_{j=1}^{n}\frac{\exp(e_{ij})}{\sum_{k=1}^{n}\exp(e_{ik})}h_j^2$$
$$\tilde{h}_j^2=\sum_{i=1}^{m}\frac{\exp(e_{ij})}{\sum_{k=1}^{m}\exp(e_{kj})}h_i^1$$
$$e_{ij}=F(h_i^1)^{\mathrm{T}}F(h_j^2) \tag{7}$$

其中，F 为一个多层前馈神经网络；e_{ij} 为论元 1 中第 i 个词和论元 2 中第 j 个词的相关性权重；$\tilde{h}_i^1$ 为论元 2 中与论元 1 中第 i 个词相关部分的表示；$\tilde{h}_j^2$ 为论元 1 中与论元 2 中第 j 个词相关部分的表示；G 为另一个多层前馈神经网络；[;]表示向量的拼接操作；v_i^1 和 v_j^2 可看作学到的局部语义关系表示。

聚合层基于局部语义关系表示计算全局语义关系表示 v，如下式所示：

$$v=[v_1;v_2]$$
$$v_1=\sum_{i=1}^{m}v_i^1 \tag{8}$$
$$v_2=\sum_{j=1}^{n}v_j^2$$

其中，v 就是经过特征层抽取后得到的特征，在学生模型和教师模型中分别表示为 $v_s(x)$ 和 $v_t(x,c)$。

分类层用于计算最终的分类结果，如下式所示：

$$p=MLP(v) \tag{9}$$

其中，MLP 由一个多层前馈神经网络和一个 softmax 层组成；p 是最终的分类结果，在学生模型和教师模型中分别表示为 $p_s(x;\theta_s)$ 和 $p_t(x,c;\theta_t)$。

5 实验

5.1 数据集

我们在常用的英文 PDTB 数据集上验证所提方法的有效性。为了便于对比，依据文献[8-9]中的实验设置，把 PDTB 数据集分成 3 部分：第 0－1 节中的实例用作验证集，第 2－20 节中的实例用作训练集，第 21－22 节中的实例用作测试集。考虑到数据的不平衡，使用准确率(Acc)和宏平均 F1 值评价方法的整体性能。为了减少模型训练的不稳定性，训练时基于不同的随机数初始化模型，取 5 次结果的平均值。

在 PDTB 的第一级和第二级隐式篇章关系上分别定义一个分类任务。第一级隐式篇章关系包含 Temporal(时序)、Comparison(对比)、Contingency(因果)和 Expansion(解说)，因此相应的任务是一个 4 分类任务。第二级隐式篇章关系中虽然包含 16 种篇章关系，但依据文献[9]的实验设置，忽略其中 5 种在验证集和测试集中没有相应实例的关系，因此相应的任务是一个 11 分类任务。第一级和第二级隐式篇章关系分类实验数据统计情况分别如表 1 和表 2 所列。

表 1 第一级隐式篇章关系分类实验数据统计

Table 1 Statistics of experimental data on the first-level implicit discourse relation classification

篇章关系	训练集	验证集	测试集
Temporal(时序)	689	54	68
Comparison(对比)	1898	191	146
Contingency(因果)	3288	287	276
Expansion(解说)	6900	651	556

表 2　第二级隐式篇章关系分类实验数据统计

Table 2　Statistics of experimental data on the second-level implicit discourse relation classification

篇章关系	训练集	验证集	测试集
Temporal. Asynchronous	541	46	54
Temporal. Synchrony	148	8	14
Contingency. Cause	3 237	281	269
Contingency. Pragmatic cause	51	6	7
Comparison. Contrast	1 717	176	129
Comparison. Concession	181	15	17
Expansion. Conjunction	2 890	266	206
Expansion. Instantiation	1 100	106	118
Expansion. Restatement	2 430	260	211
Expansion. Alternative	150	10	9
Expansion. List	330	9	12

注：表 2 中的第二级篇章关系没有比较统一的中文名称，因此没有列出

5.2　参数设置

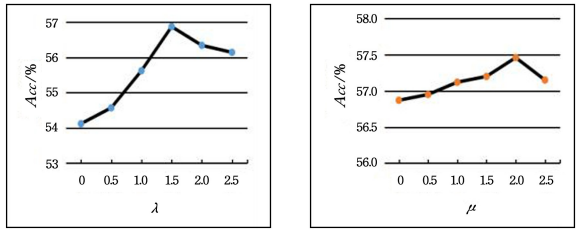
实验中，使用预训练好的 300 维 word2vec 词向量[1)]，且不再进一步优化这些词向量。为了缓解模型训练时的过拟合问题，将 dropout 技术[43]用于输入的词向量上。基于验证集上的最优性能选择模型中超参数的取值，如表 3 所列。实验发现，这些超参数的值同时适用于第一级和第二级隐式篇章关系分类。

表 3　超参数的值

Table 3　Values of super parameters

超参数	值
词向量维度	300
最大论元长度	100
批数据大小	32
优化器	Adam
学习率	0.001
BiLSTM 隐藏层维度	200(前向)，200(后向)
F 中隐藏层维度	200(只有一层)
G 中隐藏层维度	200(只有一层)
MLP 中隐藏层维度	200(只有一层)
非线性函数	ReLU
Dropout 率	0.3
λ	1.5
μ	2.0

表 3 中的超参数 λ 和 μ 是本文基于知识蒸馏的隐式篇章关系识别方法中的关键参数。如式(5)所示，λ 表示对应于分类层知识蒸馏的代价函数，μ 表示对应于特征层知识蒸馏的代价函数。我们在验证集上探索了这 2 个超参数对分类性能的影响。图 4 给出了 λ 和 μ 的不同取值对第一级隐式篇章关系分类性能的影响。图 4(a)为 $\mu=0$ 时，λ 的不同取值对性能的影响。可以看出，随着 λ 值的增加，识别性能逐步提升，且在 $\lambda=1.5$ 时识别性能达到最优。图 4(b)为 $\lambda=1.5$ 时，μ 的不同取值对性能的影响。可以看出，随着 μ 值的增加，识别性能同样逐步提升，且在 $\mu=2.0$ 时识别性能达到最优。在第二级隐式篇章关系分类上，λ 和 μ 的不同取值对性能的影响类似，这里不再列出。从 λ 和 μ 相对大的取值可以看出，基于教师模型计算的知识蒸馏相关代价在训练学生模型时起到了重要的作用。



(a)　(b)

图 4　λ 和 μ 对第一级隐式篇章关系分类性能的影响

Fig. 4　Effect of λ and μ on performance of the first-level implicit discourse relation classification

5.3　实验结果

为了验证本文方法在隐式篇章关系识别任务上的有效性，我们对比了以下几类基准方法。

第一类为使用 word2vec 或 Glove[44]等普通词向量作为输入的模型，具体如下。

(1)Qin 等(2017)[8]：基于对抗学习的方法，通过特征模仿的方式把连接词信息迁移到隐式篇章关系识别模型中。

(2)Cai 等(2017)[45]：基于一种 Attention 机制建模两个论元之间的交互信息，用于隐式篇章关系识别。

(3)Guo 等(2020)[46]：把外部资源 WordNet 中词之间的同义、反义等关系融入模型中，用于隐式篇章关系识别。

第二类为使用融合了上下文信息的词向量 ELMo[47]作为输入的模型，具体如下。

(1)Bai 等(2018)[9]：以 ELMo 词向量增强模型的输入，联合学习词级、短语级和论元级等信息用于分类，并使用连接词分类作为辅助任务来提高隐式篇章关系识别的性能。

(2)Wu 等(2020)[11]：以 ELMo 词向量增强模型的输入，把连接词作为第三级篇章关系，基于神经网络 CRF 层预测三级篇章关系序列。

第三类为本文方法的基本模型和一些变种，分别用于验证本文改进的效果，具体如下。

(1)BiAttention：本文第 4 节介绍的基于 BiAttention 的分类模型。

(2)BiAttention-MTL：采用基于 BiAttention 的分类模型，在多任务学习框架下把连接词分类任务作为辅助任务，用于提高隐式篇章关系识别的性能。

(3)BiAttention-D1：本文方法的一种简化，即把式(5)中的 λ 设为 0，用于验证特征层上知识蒸馏的有效性。

(4)BiAttention-D2：本文方法的一种简化，即把式(5)中的 μ 设为 0，用于验证分类层上知识蒸馏的有效性。

(5)BiAttention-D12：本文提出的一种方法，同时利用特征层和分类层的知识蒸馏。

(6)ELMo-*：$*$ 可以是以上 5 种方法中的任一种，表示使用 ELMo 增强了的相应方法。具体地，把预训练好的 ELMo 的输出与基于 BiLSTM 编码层的输出拼接后作为基于 BiAttention 的交互层的输入。训练时，ELMo 中的参数保持不变。

(7)BERT-*：$*$ 可以是以上前 5 种方法中的任一种，表

1) https://code.google.com/archive/p/word2vec/

示使用预训练模型 BERT[48]增强了的相应方法。近年来，基于超大规模语料预训练的 BERT 等模型成功应用于众多自然语言处理任务中，且取得了较好的效果。在这组方法中，两个论元按“[CLS]+[论元 1 中的词]+[SEP]+[论元 2 中的词]+[SEP]”的方式拼接，用作模型的输入，[CLS]和[SEP]是 BERT 中预定义的标识符。使用预训练好的 BERT 代替基于 BiLSTM 的编码层，即用 BERT 最后一层的输出(分成两部分)作为基于 BiAttention 的交互层的输入。训练时，BERT 中的参数保持不变。

表 4 列出了第一级隐式篇章关系上的分类性能，表 5 列出了第二级隐式篇章关系上的分类性能。“☆”表示运行作者提供代码的结果。从表 4 和表 5 中的实验结果可以看出：1)与基于对抗学习利用插入的连接词信息的方法相比，本文提出的模型的性能在两个级别的分类任务上都有明显的提高(行 7 vs. 行 1)。2)与基于多任务学习利用插入的连接词信息的方法相比，本文提出的模型的性能在两个级别的分类任务上都有明显的提高(行 7 vs. 行 4、行 14 vs. 行 8、行 14 vs. 行 11)。3)Wu 等(2020)基于 CRF 预测三级篇章关系序列方法的同时利用了标注的第一级、第二级篇章关系及插入的连接词。与之相比，本文提出的方法仅同时利用了第一级(或第二级)篇章关系及插入的连接词，但还是取得了不错的性能(行 14 vs. 行 9)。4)BiAttention-D1 和 BiAttention-D2 方法的性能都明显优于基本的 BiAttention 方法(行 5 vs. 行 3、行 6 vs. 行 3)，说明在特征层和分类层蒸馏知识都是有效的；方法 BiAttention-D12 的性能优于方法 BiAttention-D1 和 BiAttention-D2 的性能，说明联合使用特征层和分类层蒸馏知识能进一步提高识别的性能。ELMo 和 BERT 增强的方法(ELMo-* / BERT-*)的性能也具有相同的趋势。5)在使用强大的预训练模型 ELMo 或 BERT 的基础上，本文所提方法的性能仍有实质性的提高(行 14 vs. 行 10、行 19 vs. 行 15)，这进一步说明了所提方法的有效性。从以上分析可知，本文提出的方法是有效的，在同类方法(仅使用插入的连接词作为辅助信息)中取得了较好的识别性能。

表 4 第一级隐式篇章关系分类性能

Table 4 Performance of the first-level implicit discourse relation classification

(单位：%)

方法	*Acc*	*F1*
Qin 等(2017)☆	57.08	47.37
Guo 等(2020)	57.25	47.90
BiAttention	56.42	46.18
BiAttention-MTL	57.11	46.88
BiAttention-D1	57.76	47.20
BiAttention-D2	57.95	47.33
BiAttention-D12	58.64	47.95
Bai 和 Zhao(2018)☆	60.23	51.08
Wu 等(2020)	61.74	52.62
ELMo-BiAttention	59.10	50.21
ELMo-BiAttention-MTL	59.87	50.93
ELMo-BiAttention-D1	60.55	51.58
ELMo-BiAttention-D2	60.80	51.96
ELMo-BiAttention-D12	61.28	52.70
BERT-BiAttention	63.07	54.16
BERT-BiAttention-MTL	63.53	54.58
BERT-BiAttention-D1	63.57	54.84
BERT-BiAttention-D2	63.98	55.02
BERT-BiAttention-D12	64.21	55.13

表 5 第二级隐式篇章关系分类性能

Table 5 Performance of the second-level implicit discourse relation classification

(单位：%)

方法	*Acc*	*F1*
Qin 等(2017)	46.23	—
Cai 等(2017)	45.81	—
BiAttention	45.14	25.02
BiAttention-MTL	45.90	25.36
BiAttention-D1	45.95	25.24
BiAttention-D2	46.81	25.56
BiAttention-D12	47.21	26.05
Bai 和 Zhao(2018)☆	48.27	30.07
Wu 等(2020)	49.76	32.11
ELMo-BiAttention	47.37	29.15
ELMo-BiAttention-MTL	48.03	29.97
ELMo-BiAttention-D1	48.50	30.58
ELMo-BiAttention-D2	49.10	30.96
ELMo-BiAttention-D12	49.56	31.85
BERT-BiAttention	50.36	32.19
BERT-BiAttention-MTL	50.84	32.47
BERT-BiAttention-D1	51.17	32.63
BERT-BiAttention-D2	51.10	32.82
BERT-BiAttention-D12	51.64	33.05

结束语 本文提出了一种简单有效的基于知识蒸馏的隐式篇章关系识别方法。具体地，把连接词增强模型中的知识从特征层和分类层迁移到相应的隐式篇章关系识别模型中，以利用语料标注时插入的连接词信息。本文方法在常用 PDTB 数据集的第一级和第二级隐式篇章关系上，都取得了比同类方法更好的识别性能。在未来的工作中，我们将探索知识蒸馏相关方法在其他具有额外信息的自然语言处理任务中的应用，例如，带有用户和产品等属性信息的评论文本情感分析。

参 考 文 献

[1] LI Y, FENG W, SUN J, et al. Building Chinese Discourse Corpus with Connective-driven Dependency Tree Structure[C]// Proceedings of EMNLP 2014. 2014: 2105-2114.

[2] ZHANG L, XING Y, KONG F, et al. A Top-down Neural Architecture towards Text-level Parsing of Discourse Rhetorical Structure[C]// Proceedings of ACL 2020. 2020: 6386-6395.

[3] HU C W, YANG Y L, WU C X. An Overview of Implicit Discourse Relation Recognition Based on Deep Learning[J]. Computer Science, 2020, 47(4): 157-163.

[4] PITLER E, NENKOVA A. Using Syntax to Disambiguate Explicit Discourse Connectives in Text[C]// Proceedings of ACL-IJCNLP 2009. 2009: 13-16.

[5] KISHIMOTO Y, MURAWAKI Y, KUROHASHI S. Adapting BERT to Implicit Discourse Relation Classification with a Focus on Discourse Connectives[C]// Proceedings of the 12th Language Resources and Evaluation Conference. 2020: 1152-1158.

[6] PRASAD R, DINESH N, LEE A, et al. The Penn Discourse TreeBank 2.0[C]// Proceedings of the Sixth International Conference on Language Resources and Evaluation. 2008.

[7] ZHOU Z M, XU Y, NIU Z Y, et al. Predicting Discourse Connectives for Implicit Discourse Relation Recognition[C]// Proceedings of COLING 2010. 2010: 1507-1514.

[8] QIN L，ZHANG Z，ZHAO H，et al. Adversarial Connective-exploiting Networks for Implicit Discourse Relation Classification [C]// Proceedings of ACL 2017. 2017：1006-1017.

[9] BAI H，ZHAO H. Deep Enhanced Representation for Implicit Discourse Relation Recognition [C]// Proceedings of COLING 2018. 2018：571-583.

[10] NGUYEN L T，NGO L V，THAN K，et al. Employing the Correspondence of Relations and Connectives to Identify Implicit Discourse Relations via Label Embeddings [C]// Proceedings of ACL 2019. 2019：4201-4207.

[11] WU C，HU C，LI R，et al. Hierarchical Multi-task Learning with CRF for Implicit Discourse Relation Recognition [J]. Knowledge-Based Systems，2020，195(5-6).

[12] ZENG J，LIU Y，SU J，et al. Iterative Dual Domain Adaptation for Neural Machine Translation [C]// Proceedings of EMNLP 2019. 2019：845-855.

[13] LIU Y，CHEN K，LIU C，et al. Structured Knowledge Distillation for Semantic Segmentation [C]// Proceedings of CVPR 2019. 2019：2604-2613.

[14] HINTON G，VINYALS O，DEAN J. Distilling the Knowledge in a Neural Network [C]// Proceedings of NIPS 2014 Deep Learning Workshop. 2015：1-9.

[15] PITLER E，LOUIS A，NENKOVA A. Automatic Sense Prediction for Implicit Discourse Relations in Text [C]// Proceedings of ACL 2009. 2009：683-691.

[16] LI S，KONG F，ZHOU G D. Implicit Discourse Relation Recognition Based on PDTB System [J]. Journal of Chinese Information Processing，2016，30(4)：81-89.

[17] LIN Z，KAN M Y，NG H T. Recognizing Implicit Discourse Relations in the Penn Discourse Treebank [C]// Proceedings of EMNLP 2009. 2009：343-351.

[18] LOUIS A，JOSHI A，PRASAD R，et al. Using Entity Features to Classify Implicit Discourse Relations [C]// Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2010：59-62.

[19] RUTHERFORD A，XUE N. Discovering Implicit Discourse Relations through Brown Cluster Pair Representation and Coreference Patterns [C]// Proceedings of EACL 2014. 2014：645-654.

[20] ZHANG B，SU J，XIONG D，et al. Shallow Convolutional Neural Network for Implicit Discourse Relation Recognition [C]// Proceedings of EMNLP 2015. 2015：2230-2235.

[21] JI Y，EISENSTEIN J. One Vector is Not Enough：Entity-Augmented Distributed Semantics for Discourse Relations [J]. Transactions of the Association for Computational Linguistics，2015，3：329-344.

[22] DAI Z，HUANG R. Improving Implicit Discourse Relation Classification by Modeling Inter-dependencies of Discourse Units in a Paragraph [C]// Proceedings of NAACL 2018. 2018：141-151.

[23] FAN Z W，ZHANG M，LI Z H. Implicit Discourse Relation Classification Based on BiLSTM Combined with Self -Attention Mechanism and Syntactic Information [J]. Computer Science，2019，46(5)：221-227.

[24] ZHANG B，XIONG D，SU J，et al. Learning Better Discourse Representation for Implicit Discourse Relation Recognition via Attention Networks [J]. Neurocomputing，2018，275：1241-1249.

[25] CHEN J，ZHANG Q，LIU P，et al. Implicit Discourse Relation Detection via a Deep Architecture with Gated Relevance Network [C]// Proceedings of ACL 2016. 2016：1726-1735.

[26] LIU Y，LI S. Recognizing Implicit Discourse Relations via Repeated Reading：Neural Networks with Multi-Level Attention [C]// Proceedings of EMNLP 2016. 2016：1224-1233.

[27] LEI W，WANG X，LIU M，et al. SWIM：A Simple Word Interaction Model for Implicit Discourse Relation Recognition [C]// Proceedings of IJCAI 2017. 2017：4026-4032.

[28] 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 COLING 2018. 2018：547-558.

[29] GUO F Y，HE R F，DANG J W. Implicit Discourse Relation Recognition Based on Context Interaction Perception and Pattern Selection [J]. Chinese Journal of Computers，2020，43(5)：901-915.

[30] MARCU D，ECHIHABI A. An Unsupervised Approach to Recognizing Discourse Relations [C]// Proceedings of ACL 2002. 2002：368-375.

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

[32] WU C，SHI X，CHEN Y，et al. Bilingually-constrained Synthetic Data for Implicit Discourse Relation Recognition [C]// Proceedings of EMNLP 2016. 2016：2306-2312.

[33] LAN M，WANG J，WU Y，et al. Multi-task Attention-based Neural Networks for Implicit Discourse Relationship Representation and Identification [C]// Proceedings of EMNLP 2017. 2017：1310-1319.

[34] WU C，SHI X，SU J，et al. Co-training for Implicit Discourse Relation Recognition Based on Manual and Distributed Features [J]. Neural Processing Letters，2017，46(1)：233-250.

[35] XU Y，HONG Y，RUAN H，et al. Using Active Learning to Expand Training Data for Implicit Discourse Relation Recognition [C]// Proceedings of EMNLP 2018. 2018：725-731.

[36] BRAUD C，DENIS P. Learning Connective-based Word Representations for Implicit Discourse Relation Identification [C]// Proceedings of EMNLP 2016. 2016：203-213.

[37] WU C，SU J，CHEN Y，et al. Boosting Implicit Discourse Relation Recognition with Connective-based Word Embeddings [J]. Neurocomputing，2019，369：39-49.

[38] YIM J，JOO D，BAE J，et al. A Gift from Knowledge Distillation：Fast Optimization，Network Minimization and Transfer Learning [C]// Proceeding of CVPR 2017. 2017：7130-7138.

[39] ZHANG Y，XIANG T，HOSPEDALES T M，et al. Deep Mutual Learning [C]// Proceeding of CVPR 2018. 2018：4320-4328.

[40] LIU X，LIU K，LI X，et al. An Iterative Multi-Source Mutual Knowledge Transfer Framework for Machine Reading Comprehension [C]// Proceedings of IJCAI 2020. 2020：3794-3800.

[42] PARIKH A, TÄCKSTRÖM O, DAS D, et al. A Decomposable Attention Model for Natural Language Inference[C]// Proceedings of EMNLP 2016. 2016: 2249-2255.

[43] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[44] PENNINGTON J, SOCHER R, MANNING C. Glove: Global Vectors for Word Representation[C]// Proceedings of EMNLP 2014. 2014: 1532-1543.

[45] CAI D, ZHAO H. Pair-Aware Neural Sentence Modeling for Implicit Discourse Relation Classification[C]// Proceedings of Advances in Artificial Intelligence: From Theory to Practice 2017. 2017: 458-466.

[46] GUO F, HE R, DANG J, et al. Working Memory-Driven Neural Networks with a Novel Knowledge Enhancement Paradigm for Implicit Discourse Relation Recognition[C]// Proceeding of AAAI 2020. 2020: 10-18.

[47] PETERS M, NEUMANN M, IYYER M, et al. Deep Contextualized Word Representations[C]// Proceedings of NAACL 2018. 2018: 2227-2237.

[48] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [C]// Proceedings of NAACL-HLT. 2019: 4171-4186.

YU Liang, born in 1996, postgraduate, is a member of China Computer Federation. His main research interests include natural language processing and deep learning.

WU Chang-xing, born in 1981. Ph. D, lecturer, is a member of China Computer Federation. His main research interests include nature language processing and deep learning.