

基于外联关系的隐式篇章关系推理

洪 宇 朱珊珊 丁思远 姚建民 朱巧明 周国栋
(苏州大学计算机科学与技术学院 江苏苏州 215006)
(tianxianer@gmail.com)

Implicit Discourse Relation Inference Based on the External Relation

Hong Yu, Zhu Shanshan, Ding Siyuan, Yao Jianmin, Zhu Qiaoming, and Zhou Guodong
(School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006)

Abstract The main task of the discourse relation recognition is to automatically identify the relationships between two text spans. Currently, the performance of the explicit discourse relation can reach 93.09% because of its direct clues, however the performance of the implicit discourse recognition is still far from satisfactory since it has no direct clues. For the resolution of this issue, this paper proposes a novel implicit discourse relation inference approach based on the external relation. The method follows the existing inference pattern that uses the explicit relation to infer the implicit relation. Firstly, it searches the explicit reference arguments that have the similar content with the test arguments in the large scale of external data, then it uses the standard sorting algorithm to rank the explicit reference arguments. Finally, it predicts the implicit discourse relation based on the ranking results. Especially, the method focuses on mining the text fragments which can synergistically trigger the discourse relation between two arguments (called external elements), and predicts the implicit discourse relation of the arguments with reference to the relation between two external elements. Experiments on the Penn discourse treebank (PDTB) show an accuracy of 54.12%, which is a significant improvement of 11.82% over the current state-of-the-art system.

Key words implicit discourse relation; external relation; explicit relation infer implicit relation; relation detection; discourse analysis

摘 要 针对隐式篇章关系(implicit discourse relation)分类性能较低的问题,提出一种基于“外联”关系的无监督隐式篇章关系推理方法.该方法继承“显式指导隐式”的关系推理模式,针对每个待测“论元对”,在大规模外部数据资源中挖掘与其内容近似的显式“参考对”,借助“参考对”的显式关系推理隐式关系.特别地,该方法侧重挖掘2个论元中能够协同触发篇章关系的文字片段(即“外联”成分),以“外联”成分间的关系为参考,推理“论元对”整体的篇章关系.利用宾州篇章树库(Penn discourse treebank, PDTB)对这一推理方法进行评测.实验结果显示,该方法在隐式篇章关系推理性能上获得显著提升,识别精确率达到54.12%,与现有主流推理方法性能对比,识别精确率提升11.82%.

关键词 隐式篇章关系;外联关系;显式指导隐式;关系检测;篇章分析

中图法分类号 TP391

篇章关系研究旨在推断同一篇章内论元(即具有独立语义的文字片段,包括子句、句子或文本块等)之间的语义连接关系.依据宾州篇章树库(Penn discourse treebank, PDTB)^[1-2]的任务定义,篇章关系类型分为3层:第1层包括4种主关系:时序关系(Temporal)、扩展关系(Expansion)、对比关系

(Comparison)和偶然关系(Contingency);第2层和第3层分别针对上一层进行细分。此外,根据2个论元(即“论元对”)之间是否存在连接词,PDTB又将篇章关系分成显式篇章关系(explicit discourse relation)和隐式篇章关系(implicit discourse relation)2种模式。下述例1和例2分别给出显式和隐式篇章关系的实例。其中,例1为显式篇章关系样例,通过连接词“Because”可直接判定“论元对”的篇章关系为“Contingency”;例2为隐式篇章关系样例,其论元之间不存在连接词作为关系的显式线索,需通过上下文语义及结构等信息进行推理,本例中的连接词可推理为“But”,表征“Comparison”关系。

例1. Arg1: I get very nervous.

〈译文:我非常紧张。〉。

Arg2:【Explicit=Because】I'm using a lot of expensive equipment.

〈译文:【因为】我正在使用很多昂贵的设备。〉。

篇章关系 = “Contingency”。

例2. Arg1: He likes to eat durian.

〈译文:他喜欢吃榴莲。〉。

Arg2:【Implicit=But】I dislike it and I feel it is difficult to eat.

〈译文:【但是】我不喜欢吃榴莲,感觉它很难吃。〉。

篇章关系 = “Comparison”。

篇章关系研究在自然语言处理领域具有广泛的应用价值。比如,篇章中的偶然关系可应用于自动问答系统和事件因果关系抽取(Riaz等人^[3]和Do等人^[4]);对比关系可应用于情感分析研究,辅助实现特定句子内部的情感极性判断(Zhou等人^[5]);扩展关系可应用于自动文摘、篇章关键词识别(Wang等人^[6])等。

目前,对于显式篇章关系研究推理精确率已达90%左右,而隐式篇章关系推理精确率仍然较低。分析原因可知,显式篇章关系中论元之间不存在连接词,且绝大部分连接词表意单一且明确,能够直接指向一种关系类型,如“now(现在)”唯一地指向Temporal,使借助连接词的关系判定既简易又精准。相对地,隐式篇章关系样本中,论元之间缺失连接词,无法直接判定篇章关系,需通过上下文、语义结构及句子特征等其他信息间接推理隐式篇章关系。然而,上下文信息的不确定性、语义结构的复杂性以及句子特征的歧义性,往往制约隐式篇章关系的有效判定。针对这一情况,本文专注于隐式篇章关系推理问题,尝试提出一种更为有效的关系推理机制。

传统的隐式篇章关系检测方法主要采用基于全监督或半监督学习的分类方法:如Pitler等人^[7]提取论元中的情感词极性、动词短语、首尾单词等特征,建立关系分类器,其分类结果优于随机分类的性能。Wang等人^[8]基于树核函数实现隐式篇章关系的分类,整体分类性能达到40.00%。Zhou等人^[9]采用语言模型,实验分类性能有一定程度的提升。上述方法主要借助自然语言处理技术抽取论元的语言学特征,采用分类器判别篇章关系类型,然而实验性能并不理想,究其原因,可发现如下3个问题:

1) 待测论元往往是篇幅较短的子句或文字片段,包含的语言学特征匮乏,符合特定规律的语言学现象难以有效学习,导致模型训练出现偏差,影响最终的关系分类性能。

2) 某些论元在篇章中并非毗连,而是离散分布,两两之间并不存在本源的句法、依存、修辞甚或语义关联的纯粹语言学特性,从而,依赖底层的自然语言特征进行关系学习与判断,略显牵强,且有效性无从保证。

3) 论元并非作为整体与外部产生语义关系,其成分可分为“内联”与“外联”2种。“内联”主要用于保证论元自身文法、语义和修辞的逻辑性;“外联”主要用于保证论元与外部语言单位相互关联的逻辑性。比如2个论元“塔利班遭受空袭”与“塔利班发动报复性偷袭”,其中,“遭受”与“空袭”、“发动”与“偷袭”等作为“内联”成分构成内联关系(例如:依存关系);“遭受空袭”与“报复性偷袭”构成外联关系。现有方法往往忽视论元局部成分的“外联”分析,而使用论元整体进行关联分析,容易受到非“外联”成分的误导。

针对上述问题,需建立一种有效利用外部数据、结合统计信息,并且关联焦点明确的篇章关系分析方法。由此,本文提出一种基于全局“外联”关系的隐式篇章关系推理方法。该方法首先以待测“论元对”为标杆,抽取“论元对”的关键“外联”成分。在此基础上,从大规模同领域数据中挖掘内容近似且显式关联的参考“论元对”(简称“参考对”),形成待测“论元对”的参考样本集,该集合中参考样本集的全局关系分布能有效改善待测“论元对”自身关系线索匮乏的局限性;其次,该方法利用参考样本集的关系分布,通过极大似然估计,进行论元关系推理,避开底层语言学特征的依赖性,有益于非毗连论元之间的关系检测;最后,该方法既考虑待测“论元对”与“参考对”是否具有一致的“外联”成分,又考虑参考样本集中

不同“外联”成分的关系分布,以“外联”一致且“外联”关系分布最大为准绳,形成基于全局“外联”关系线索的论元关系推理方法.

1 相关工作

目前,篇章关系检测研究以全监督或半监督机器学习方法为基础,包括基于概率统计的篇章关系推理方法和基于语言学特征的篇章关系分类方法.

基于概率统计的篇章关系推理方法中, Marcu 等人^[10]首次将概率统计应用于篇章关系分析,利用连接词与篇章关系映射概率,实现篇章关系判断.在此基础上, Saito 等人^[11]继承 Marcu 等人的方法,加入短语特征,有效提高日文隐式篇章关系推理的性能. Zhou 等人^[9]使用三元文法模型,在相邻论元间插入合适的连接词,将插入的连接词作为辅助分类特征,有效实现隐式篇章关系的分类. 相比于 Saito 等人,该方法不局限于语法的规范,满足词特征相互组合的连贯性和灵活性,但是其性能仅在偶然和时序关系上有所提升,对扩展和比较关系的分类性能仍然偏低.

基于语言学特征的篇章关系分类方法中, Pitler 等人^[7]使用情感词极性、动词短语长度、动词类型和上下文等特征进行关系分类,最终在 PDTB 测试集上获得优于随机关系分类的性能. Lin 等人^[12]继承 Pitler 等人的方法体系,采用句法树的结构特征与依存特征等;同时,结合 Soricut 等人^[13]提出的论元内部结构特征,在 PDTB 第 2 层隐式篇章关系类别上获得 40.2% 的分类精确率. 随后, Wang 等人^[8]基于卷积树核函数提升句法结构特征的区分能力,然而实验性能没有显著提升(精确率约 40.0%). Park

等人^[14]提出特征集合优化的方法,采用前向选择算法对情感词极性、句子首尾单词、产生式规则等特征进行特征选择. Lan 等人^[15]提出基于多任务学习的隐式篇章关系分类方法,在交互结构优化(alternating structure optimization, ASO)多任务学习框架下,引入外部数据资源,训练主分类器及辅分类器,实现隐式篇章关系分类. 近期, Li 等人^[16-17]通过解决训练实例篇章关系类别不平衡以及特征表示稀疏问题,进一步提高隐式篇章关系的推理性能.

总体而言,上述基于语言学特征的分类方法,往往局限于有限的语言学资源,导致分类特征(如句法结构等)的缺失或训练不充分,影响关系分类性能. 此外,上述基于统计信息的推理方法对论元中不同成分赋予等等的权重,任何成分的关系都拟比为论元整体的外部关系,忽视论元焦点和辅助成分的差异,混淆两者在“外联”和“内联”中的不同作用. 对此,本文提出“外联”成分识别以及利用全局“外联”关系的隐式篇章关系推理方法.

2 推理框架

2.1 基本思想

本文主要通过识别论元的有效“外联”成分,结合大规模外部数据的统计信息和排序算法,实现隐式论元关系的推理. 其基本思想是:两两论元之间的关系主要由其各自的“外联”成分指定,并且,大规模数据中往往能够蕴含大量语言学片段的显式篇章关系,从而,通过查询“外联”成分在外部数据中的关联样本,可实现基于全局统计信息的“外联”关系判定,辅助指导论元整体的关系推理. 图 1 给出这一思想的实例化说明.

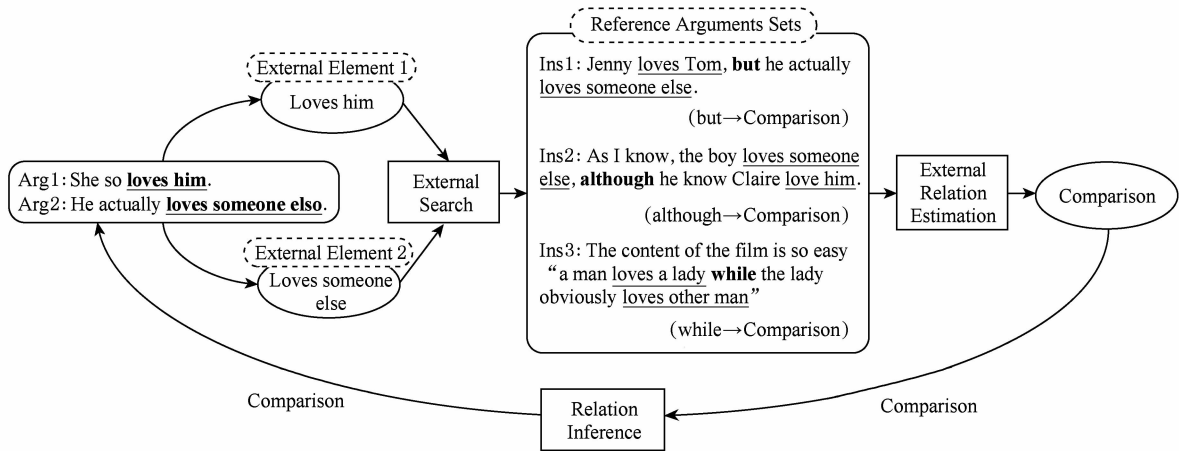


Fig. 1 Example of the external relation determination based on the global statistical information.

图 1 基于全局统计信息的“外联”关系判定实例化说明

2.2 术语定义

全文涉及的新术语包括“外联”成分、“内联”成分、“外联”关系和“内联”关系,在进行后续陈述之前,本文首先给出上述术语的定义。

定义 1. “外联”成分. 论元(句子或子句)中能够反映论元外部语义关系的文字片段或组合,如图 1 中“loves+him”和“loves+someone else”。

定义 2. “内联”成分. 论元(句子或子句)中仅能反映论元自身文法、句法或依存的生成规则的文字片段(而非组合),如图 1 中“someone else”(“else”依存于“someone”)。

定义 3. “外联”关系. 由“外联”成分反映的论元外部关系,受限于不同上下文。

定义 4. “内联”关系. “内联成分”之间的文法、句法和依存关系。

其中,“内联”成分及关系有助于识别论元内衔接紧凑的词或短语,形成“外联”成分的候选语块;“外联”成分则专门用于探测论元的外部关系。事实上,借助“内联”关系,特定论元可以切分为多项语块,成为多个“外联”成分候选。从而,两两论元之间,借助外部数据,可以建立多种“外联”成分候选的关联关系,形成两者的多种“外联”关系候选。因此,基于外联线索的论元关系推理主要需要解决如下关键问题:

- 问题 1. 如何划分论元中的不同成分?
- 问题 2. 如何判定“外联”成分,并评估特定上下文中“外联”关系的确定性?
- 问题 3. 如何确定“外联”成分之间的语义关系?

2.3 推理方法与流程

针对上述问题,本文分别给出相应解决方法,主要包括 3 个部分:1)“显式指导隐式”的关系推理模式;2)“外联”成分的抽取方法;3)基于全局“外联”关

系的推理优化算法。其中,方法 1 为前人工作^[18],在本文中仅简略介绍。相对地,本文侧重方法 2 和方法 3 的分析与检验。

基于全局“外联”关系的隐式论元关系推理流程可分成 3 个步骤:

Step1. 将每个待测论元切分为“内联”成分,并在两两待测论元之间,跨论元组合成不同片段对,计算每个片段对的相关性(即关联度),根据关联度排序,选择 Top *M* 片段对,作为存在“外联”关系的片段对,简称“外联”关系对。

Step2. 将每个“外联”关系对作为查询,借助检索技术,在 Gigaword 语料中搜索内容近似的显式“论元对”,形成关系推理的参考集合,将该集合中“论元对”的显式关系作为关系推理线索,由此形成“外联”关系对及其实例的全局资源。

Step3. 针对每个“外联”关系对,计算其触发论元关系的可信度。额外地,计算待测“论元对”与“参考对”的语义相似度。在此基础上,利用线性加权,综合衡量关联度、可信度以及相似度的统一指标,对“外联”关系对的实例进行排序,选择 Top *N* 显式关系实例,将其中分布最广的关系判定为待测“论元对”的篇章关系。

3 “显式指导隐式”的关系推理模式

针对隐式篇章关系推理任务,本文继承“显式指导隐式”的关系推理模式,利用显式关系易于判定的优势,借助显式关系推理隐式关系。本节对“显式指导隐式”关系推理模式的基本思想进行简单概述。

“显式指导隐式”的关系推理模式的基本思想是:如果一对隐式论元与一对显式论元在语义以及结构上具有一致性,则两对论元的篇章关系也一致。

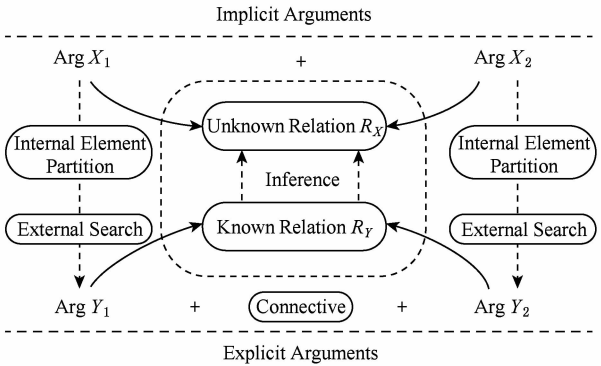


Fig. 2 Structure of “use explicit relation to infer implicit relation” inference pattern.

图 2 “显式指导隐式”的关系推理模式结构图

通过以隐式“论元对”与显式“论元对”之间的一致性为桥梁,在外部数据中可挖掘大量包含连接词等直观推理线索的显式“论元对”,进而可通过显式关系推理隐式关系.图2为“显式指导隐式”的关系推理模式结构图.详细而言,给定隐式“论元对” X ,通过外部检索搜索出与其在语义上以及结构上一致的显式“论元对” Y ,则隐式论元对 X 的篇章关系 R_X 与显式论元对 Y 的篇章关系 R_Y 一致,可直接通过“论元对” Y 的显式关系推理“论元对” X 的隐式关系.

外部检索的显式“论元对”与待测“论元对”之间的语义一致性强弱主要通过整体相似度 Sim 予以度量,计算方法如式(1)所示:

$$Sim = \frac{PreSim + PosSim}{2e^{|PreSim - PosSim|}}, \quad (1)$$

其中, $PreSim$ 表示前置论元 $Arg1$ 之间的相似度; $PosSim$ 表示后置论元 $Arg2$ 之间的相似度;分子为 $PreSim$ 与 $PosSim$ 的和,表征论元之间的整体相似度;分母为指数平滑.此外, $PreSim$ 和 $PosSim$ 均采用基于向量空间模型的余弦相似度予以度量.

4 基于全局“外联”关系的隐式篇章关系推理

上述基于“显式指导隐式”关系推理模式的方法

将“内联”成分的关系比拟为“论元对”整体的外部关系,混淆了“外联”成分和“内联”成分的不同作用,影响篇章关系推理性能.因此,本文提出“外联”成分的识别方法,在此基础上,继承“显式指导隐式”关系推理模式的基本思想,利用全局“外联”关系实现隐式篇章关系的推理.

4.1 “外联”成分挖掘

“外联”成分是论元(句子或子句)中能够反映论元外部语义关系的文字片段或组合,将“外联”成分之间的关系作为线索,可有效推理出“论元对”整体的篇章关系.为了获取“外联”成分,本文首先对待测“论元对”进行“内联”成分划分,借助“内联”关系,将论元切分成多项语块.在此基础上,将两两待测论元之间的“内联”成分,跨论元地交叉组合成多个不同片段对,形成“外联”成分的候选集合.

本文使用语块分析(chunk parsing,也称为浅层句法分析)方法进行论元“内联”成分划分,标注工具为Senna-V3.0^①,标注准确率达到94.32%.

如表1所示,对“论元对”实例进行语块分析,可分别获得 $Arg1$ 和 $Arg2$ 的语块标注标签 $S1_{Chunk_Label}$ 和 $S2_{Chunk_Label}$,将其中的语块单元作为论元的“内联”成分,在两两论元之间交叉组合成25个不同的片段对,形成“论元对”实例的“外联”成分候选集合.

Table 1 External Elements Mining of the Arguments Instance
表1 “论元对”实例“外联”成分挖掘

Annotation Status	Examples
Arguments Instance	$Arg1$: We have established the quality of the clothing in the last six months. $Arg2$: There is no fire sale in that clothing store.
Chunk Annotation Results	$S1_{Chunk_Label} = \{B-NP\ I-NP\ I-NP\ E-NP\ B-VP\ E-VP\ B-NP\ E-NP\ S-PP\ B-NP\ E-NP\ O\}$ $S2_{Chunk_Label} = \{S-NP\ S-VP\ B-NP\ I-NP\ E-NP\ S-PP\ B-NP\ I-NP\ E-NP\ O\}$
Internal Elements	$Arg1$: {We, have established, the quality, of, the clothing, in, the last six months} $Arg2$: {There, is, no fire sale, in, that clothing store}
External Elements Candidate Sets	{we+there, have established+is, the quality+no fire sale,..., have established+no fire sale, ..., the clothing+that clothing store}

4.2 “外联”关联度估计

本文分别采用基于向量空间模型以及基于协同测量的关联度估计方法,获得“外联”成分候选集合中每个片段对的关联度,根据关联度排序选择Top M 片段对,作为存在“外联”关系的片段对(即“外联”关系对).下面详述2种关联度估计方法:

1) 基于向量空间模型的关联度估计

本文将 $Arg1$ 与 $Arg2$ 分别表示成向量: $S_1 =$

$(C_{11}, C_{12}, \dots, C_{1m})$ 和 $S_2 = (C_{21}, C_{22}, \dots, C_{2n})$,其中 C_{ij} 表示论元 Arg_i 中的第 j 个“内联”成分.交叉组合 $Arg1$ 和 $Arg2$ 的“内联”成分,获得跨论元片段对向量 $S_{chunk_pair} = (C_{11_C_{21}}, C_{12_C_{22}}, \dots, C_{1x_C_{2y}}, C_{1(x+1)_C_{2(y+1)}}, \dots, C_{1m_C_{2n}})$,其中 $x = 1, 2, \dots, m; y = 1, 2, \dots, n$.通过Jaccard相关系数法计算 S_{chunk_pair} 中所有片段对的关联度.

以片段对 $C_{11_C_{21}}$ 的关联度计算为例,将片段 C_{11}

① <http://ml.nec-labs.com/senna/>

和 C_{21} 分别表示成向量 $\mathbf{A}=(w_{11},w_{12},\cdots,w_{1k},w_{1(k+1)},\cdots,w_{1i})$ 和 $\mathbf{B}=(w_{21},w_{22},\cdots,w_{2k},w_{2(k+1)},\cdots,w_{2j})$, 通过计算向量 $\mathbf{A}$ 与 $\mathbf{B}$ 的 Jaccard 相关系数, 获得片段对 $C_{11}_C_{21}$ 的关联度. 其中, 向量 $\mathbf{A}$ 中的每维特征表示, 在所有 Arg1 包含 C_{11} 的显式关系实例中, Arg2 的所有“内联”成分; 向量 $\mathbf{B}$ 中的每维特征表示, 在所有 Arg2 包含 C_{21} 的显式关系实例中, Arg1 的所有“内联”成分. 下面介绍具体计算步骤:

Step1. “内联”成分 C_{11} 和 C_{21} 分别对应的向量 $\mathbf{A}$ 与向量 $\mathbf{B}$ 中, w_{1k} 与 w_{2k} 均表示第 k 维特征的权重. 本文给出如下 2 种度量方法: 共现频数 (co-occurrence frequency, CF) 和共现频数-逆向语块频数 (co-occurrence frequency-inverse chunk frequency, CF-ICF). CF 计算方法中, 对于 $w_{1k}(w_{2k})$, CF 表示在所有 Arg1(Arg2) 包含 $C_{11}(C_{21})$ 的显式关系实例中, Arg2(Arg1) 的某个“内联”成分与 $C_{11}(C_{21})$ 的共现频数. CF-ICF 计算方法中, ICF 计算方法如式 (2) 所示:

$$ICF_k = \text{lb} \left[\frac{\sum_{i=1}^k CF_i}{CF_k} \right] + 1, \tag{2}$$

其中, ICF_k 表示第 k 个特征的逆向语块频数, 由 k 维特征的 CF 总数除以第 k 个特征的 CF, 并将商取对数, 通过加 1 操作平滑计算结果.

Step2. 针对每个片段对向量(如向量 $\mathbf{A}$ 与 $\mathbf{B}$), 分别以 CF 和 CF-ICF 作为每维特征的权重, 采用相关系数度量方法(例如 Jaccard 距离、Wilson 距离、Pearson 距离等)计算片段对的关联度.

2) 基于协同测量的关联度估计

基于协同测量的关联度估计方法的主要思路是: 对于 2 个待测变量, 如果它们的期望频数分布与实际观察频数分布越一致, 则说明其中一个变量对另一个变量的影响越小, 即 2 个变量相关性越小, 关联度越小; 反之, 2 个变量相关性越大, 关联度越大. 对于任意一个片段对, 将它们看作 2 个变量 U 和 V , 则可将关联性估计问题转化成变量 U 和 V 的独立性分析问题.

图 3 为观察频数和期望频数统计结果. 变量 U 表示前置“内联”成分, 变量 V 表示后置“内联”成分, u 和 v 分别表示 U 与 V 的具体“内联”成分信息. 图 3(a) 和图 3(b) 中 O_{ij} 和 $E_{ij}(i, j=1, 2)$ 分别表示 u 与 v 的观察频数和期望频数. 当 $i=1$ 或 2 时,

分别表示 u 出现与不出现的情况; 同理, $j=1$ 或 2 时, 分别表示 v 出现与不出现的情况, 例如 O_{12} 表示 u 出现且 v 不出现的观察频数, E_{22} 表示 u 与 v 均不出现的期望频数. 图 3(b) 中各变量 C_1, C_2, R_1, R_2, N 满足:

$$C_1 = O_{11} + O_{21}, \tag{3}$$

$$C_2 = O_{12} + O_{22}, \tag{4}$$

$$R_1 = O_{11} + O_{12}, \tag{5}$$

$$R_2 = O_{21} + O_{22}, \tag{6}$$

$$N = C_1 + C_2 = R_1 + R_2. \tag{7}$$

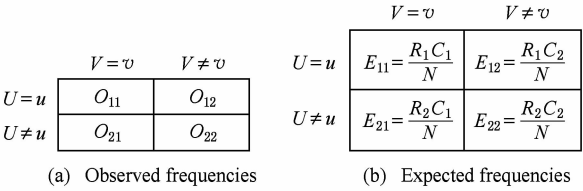


Fig. 3 Statistical results of observed frequencies and expected frequencies.

图 3 观察频数以及期望频数统计结果

经过上述计算获得观察频数 O_{ij} 及期望频数 E_{ij} , 采用卡方检验(如式 (8) 所示) 计算获得观察频数与期望频数之间的独立性强弱, 计算结果值越小, 则说明变量 U 与 V 越独立, 相关性越小; 反之, 计算结果值越大, 相关性越大.

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}. \tag{8}$$

其中, $i, j=1, 2$.

4.3 “外联”可信度估计

本节针对 4.2 节构建的“外联”关系对, 结合检索技术在本地 Gigaword 语料中搜索内容近似的显式“参考对”, 形成“外联”关系的参考集合. 在此基础上, 针对每个“外联”关系对, 计算其触发当前待测“论元对”的可信度. 本节首先介绍显式“参考对”的挖掘方法, 在此基础上, 给出具体的“外联”可信度的计算方法.

4.3.1 显式“参考对”挖掘

为了获取大规模显式“参考对”, 本文将从 PDTB 语料中统计出的 87 个 Golden 连接词(指向某一篇章关系的概率在 96% 以上) 作为索引, 利用 Apache Lucene 4.1^① 检索本地 Gigaword 语料, 抽取包含连接词的显式“论元对”, 并根据 Golden 连接词及其最大似然关系映射表(如表 2 所示), 将对应的最大似然关系作为“论元对”的篇章关系.

① <http://lucene.apache.org>

Table 2 Mapping Table of PDTB Golden Connectives and Their Most Possible Relation (Only List A Part)

表 2 PDTB 中 Golden 连接词及其最大似然关系映射表 (仅列出部分)

Discourse Relations	Golden Connectives (Probability of the Most Possible Relation/%)
Expansion	alternatively (100.00), in fact (96.42), indeed (96.03), unless (98.95)...
Contingency	because (99.28), so (99.06), so that (100.00), hence (100.00)...
Comparison	however (99.33), but (96.95), by contrast (98.86), although (98.67)...
Temporal	now (100.00), afterward (100.00), later (97.17), before (100.00)...

本文将“外联”关系对作为查询,在上述获得的“外联”关系参考集合中搜索与待测“论元对”语义一致的显式“论元对”(显式“论元对”的 Arg1 和 Arg2 须分别包含“外联”关系的前置“内联”成分和后置“内联”成分),将所有检索结果作为“外联”关系的参考样本集合,形成“外联”关系及其实例的全局资源.

4.3.2 “外联”可信度计算

针对每个“外联”关系对,会挖掘到多种“外联”关系.但相同方式构建的“外联”关系对,对当前待测“论元对”的篇章关系具有不同的可信程度,本文将这种可信程度定义为“外联”关系对可信度.

本文主要通过衡量“外联”关系参考样本集合中的篇章关系分布与外部数据资源中的篇章关系分布之间的一致性程度来表征可信度.基本思想如下:对每个“外联”关系对,统计其对应参考样本集合中的篇章关系分布比例及检索的外部数据资源中的篇章关系分布比例.如果两者之间的篇章关系分布比例越一致,则说明该“外联”关系对的“外联”关系在语料中分布比较普遍,对当前待测“论元对”的篇章关系判定影响越小,可信度越低;反之,如果分布比例差异性越大,则说明该“外联”关系对的“外联”关系在某些篇章关系类别上分布比重较大、在其他类别上分布比重较小,偏离原有语料中的篇章关系分布,对当前待测“论元对”的篇章关系判定影响越大,可信度越高.

对于某一“外联”关系对,假设其对应的“外联”关系分布比例为 $R_a:R_b:R_c:R_d$,语料中各篇章关系分布比例为 $R_A:R_B:R_C:R_D$,通过 KL 散度量 2 个概率分布之间的差异性.计算结果值越大,2 个概率分布差异越大,可信度越大;反之,可信度越小.

4.4 “显式指导隐式”关系推理模式的优化模型

本文在推理方法上继承“显式指导隐式”关系推理模式的基本思想.对每个待测“论元对”,区分出“内联”成分和“外联”成分,借助“外联”成分挖掘与

待测“论元对”在语义上以及结构上具有一致性的显式“论元对”,并在相似度模型基础上融合“外联”关联度及“外联”可信度模型,对挖掘到的显式“参考对”进行综合评价,评价指标如式(9)所示:

$$ComScore(R_k)=\alpha\times Linear(Sim_k)+\beta\times Linear(Rele_k)+\gamma\times Linear(Con_k),\quad (9)$$

其中, R_k 表示某“外联”关系, Sim_k 表示“外联”关系对的关系实例与待测“论元对”的相似度(见第 3 节), $Rele_k$ 表示“外联”关系对关联度(见 4.2 节), Con_k 表示“外联”关系对可信度(见 4.3 节), $Linear$ 表示采用线性函数转化法对数据进行标准化; α, β, γ 是 3 种模型对应的参数权重,并满足 $\alpha, \beta, \gamma \in [0, 1]$, $\alpha + \beta + \gamma = 1$.

通过上述评价指标对每个“参考对”进行综合质量评估,根据计算结果排序,选择 Top N 显式“参考对”,将其中分布最广的关系判定为待测“论元对”的篇章关系.

5 实验设计

本节主要介绍实验中涉及的实验数据集、评价标准以及参与实验的各隐式篇章关系推理系统.

5.1 数据集及评价标准

本实验选择 PDTB 语料中 Section 21-22 作为测试集 1(简称为 DP1),共包含 1 046 个隐式实例; Section 23-24 作为测试集 2(简称为 DP2),共包含 1 192 个隐式实例.本文对于包含 2 种或 2 种以上篇章关系的测试实例,选择最主要的关系类别作为其正确的篇章关系.此外,实验中还涉及在本地 Gigaword 语料中挖掘的显式关系样本(简称为 DG),并按照自然分布比例从 DG 中抽取显式关系样本(为避免偶然性,随机进行 3 次抽取,分别简称为 DG1, DG2, DG3).表 3 为各数据集 4 种篇章关系的分布情况.

Table 3 Discourse Relation Distribution of Data Sets Used in the Experiment

表 3 各数据集篇章关系分布

Data Sets	Number of Instances (Ratio of Instances/%)				Total Number of Instances
	Expansion	Comparison	Contingency	Temporal	
DP1	562(53.72)	145(13.86)	269(25.72)	70(6.69)	1 046
DP2	606(50.84)	208(17.45)	338(28.36)	40(3.36)	1 192
DG	924243(20.44)	2621028(57.98)	715546(15.83)	249960(5.53)	4 520 777
DG1,DG2,DG3	934243(54.53)	257431(15.03)	429548(25.07)	91887(5.36)	1 713 109

为了评估推理系统的分类性能,本文使用精确率(accuracy)作为评价标准衡量每个类别的推理性能,计算方法如式(10)所示:

$$Accuracy=\frac{TruePositive+TrueNegative}{N}, \quad (10)$$

其中, *TruePositive* 表示被正确分为正例的个数, *TrueNegative* 表示被正确分为负例的个数, *N* 为待测“论元对”总数.

该评测方法也作为整体性能的评价标准,计量方法是将 *TrueNegative* 设置为恒定值 0,只检验每个待测“论元对”是否获得正确的关系推理,即只计算 *TruePositive* 指标与 *N* 的比值.整体关系推理准确率统称为“Four-Way”.

5.2 实验系统

本节主要描述实验中参与的基准系统、基于“外联”关系的隐式篇章关系推理系统以及涉及到的对比系统,简单介绍如下:

1) 基准系统

① RANB. 给定测试样本,随机判定关系类别为 4 种篇章关系类别中的任意一种,识别准确率为 25%.

② MAXB. 给定测试样本,判定所有待测“论元对”都具有 1 种关系且为出现概率最大的关系,概率最大关系通过训练语料获得为“Expansion”(如表 3 所示).

③ EPIB. 给定测试样本,识别出待测“论元对”的“内联”成分,在此基础上通过外部检索搜索出与待测“论元对”在语义上以及结构上具有一致性的显式“论元对”,通过显示关系推理隐式关系.“论元对”整体的一致性强弱通过式(1)(见第 3 节)度量,单一论元的一致性强弱则通过相余弦相似度度量.

2) 基于“外联”关系的隐式篇章关系推理系统

① VSM_SYS. 该系统继承“显式指导隐式”关系推理模式的基本思想.对每个待测“论元对”,通过基于向量空间模型的关联度估计方法,有效识别出“外联”关系对,在 Gigaword 显式关系样本中挖掘

内容近似的显式“参考对”.在此基础上,针对每个“外联”关系对,计算其触发当前待测“论元对”的可信度.利用线性加权,综合衡量关联度(见 4.2 节)、可信度(见 4.3 节)以及相似度(见第 3 节)的统一指标,对“外联”关系对进行综合评价,选择排序 Top*N* 的显式“参考对”,将其中分布最广的关系判定为待测“论元对”的篇章关系.

② MA_SYS. 与 VSM_SYS 系统方法相同,但采用基于协同测量的关联度估计方法.

另外,本文还将推理性能与各主流方法性能进行对比,选择的对比对象主要为:Wang 等人^[8]于 2010 年提出的基于树核函数的隐式篇章关系推理系统(标记为 WANG_SYS)以及 Lan 等人^[15]于 2013 年提出的基于多任务学习的隐式篇章关系推理系统(标记为 LAN_SYS).

6 实验结果与分析

6.1 参数选择

实验中,需要估计的参数主要分成 3 类:1)Top *M* “外联”关系对权重 *M*;2)Top *N* 显式“参考对”权重 *N*;3)相似度、关联度、可信度权重 α, β, γ .

首先调整参数 *M* 和 *N* 的取值大小,令 $\alpha=\beta=\gamma=1/3$.图 4 显示在各推理系统上调整参数 *M* 和 *N* 对推理性能的影响.分类结果显示,当 *M* 大小不变时,分类精度首先随着 *N* 的增大而增大;当 *N* 达到一定值后,分类精度基本保持不变.原因在于,*N* 取值较小,则不能提供足够的显式关系信息推理隐式关系,导致分类精度偏低;当 *N* 取值达到一定时,“外联”关系参考集合中的篇章关系分布可以代表当前“外联”关系对的自然篇章关系分布,增大 *N* 值会增加冗余信息,而对分类精度没有影响.同理,当 *N* 取值固定不变,随着 *M* 取值的增大,分类精度首先不断增大,当 *M* 取值达到一定时,“外联”关系对的

数量已经足以推理待测“论元对”的篇章关系,增大 M 值也不会改变分类精度大小. 此外,在实验过程中,对于 VSM_SYS 实验系统,本文尝试采用多种相关系数度量方法(例如 Jaccard 距离、Wilson 距

离、Pearson 距离等),通过实验证明采用 Jaccard 距离度量方法能够获得更优的推理性能,下文涉及的 VSM_SYS 实验系统,均默认采用 Jaccard 距离度量方法.

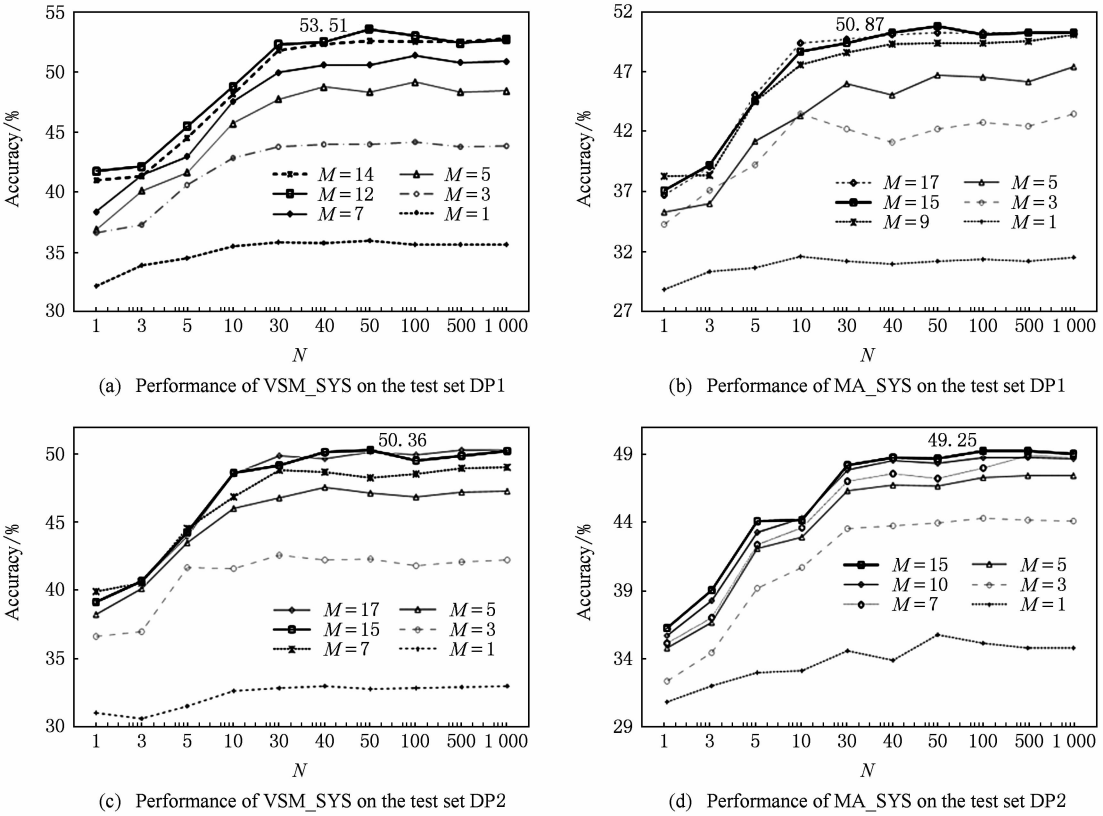


Fig. 4 Classification accuracies with the changing parameters M and N .

图 4 参数 M 和 N 的变化对分类精度的影响

从图 4 还可以得出,VSM_SYS 的整体分类性能优于 MA_SYS. 对比图 4(a)与图 4(b),在测试集 DP1 上,VSM_SYS 获得最高 53.51% 的分类精度 ($M=12,N=50$),MA_SYS 获得最高 50.87% 的分类精度 ($M=15,N=50$);对比图 4(c)与图 4(d),在测试集 DP2 上,VSM_SYS 获得最高 50.36% 的分类精度 ($M=15,N=50$),MA_SYS 获得最高 49.25%

的分类精度 ($M=15,N=100$).

为了获取各推理系统的最优分类性能,对参数 α, β, γ 的权重进行调整,固定 M 与 N 为获取最优性能时的取值(如图 4(a)中 VSM_SYS,设定 $M=12,N=50$). 基于此,调整 α, β, γ 的权重,满足 $\alpha, \beta, \gamma \in [0,1], \alpha + \beta + \gamma = 1$. 针对测试集 DP1 和 DP2,选取 4 组较优的实验结果及其对应的权重.

Table 4 Partial Weight Adjustment Experimental Results of α, β, γ (Explicit Data Sets Using DG1)

表 4 α, β, γ 权重调整部分实验结果(显式关系样本使用数据集 DG1)

Test Sets	DP1($M=12,N=50$)				DP2($M=14,N=50$)			
	Result1	Result2	Result3	Result4	Result1	Result2	Result3	Result4
α	0.05	0.35	0.20	0.20	0.35	0.20	0.20	0.05
β	0.20	0.45	0.45	0.60	0.45	0.60	0.75	0.75
γ	0.75	0.20	0.35	0.20	0.20	0.20	0.05	0.20
Four-Way Probability/%	53.10	54.02	54.12	53.92	50.80	50.89	50.98	50.89

从表 4 可以看出,对于测试数据集 DP1,VSM_SYS 获得最高 54.12% 的推理性能 ($M=12,N=$

50, $\alpha=0.2, \beta=0.45, \gamma=0.35$);对于测试数据集 DP2,VSM_SYS 获得最高 50.98% 的推理性能

($M=14$, $N=50$, $\alpha=0.2$, $\beta=0.75$, $\gamma=0.05$). 两者达到最优性能时,参数 β 的取值均大于其他参数,即关联度权重对篇章关系的推理性能影响最大,这也进一步验证本文基于“外联”关系的隐式篇章关系推理方法的可行性和有效性.

6.2 分类结果

基于上述参数指标,对实验中各推理系统进行性能评测.表 5 给出 2 项分类结果:1)对单一篇章关系的二元分类结果;2)推理系统整体分类结果.

从表 5 可以看出,本文各推理系统均具有较好的稳定性,在整体上均获得较优的分类性能.结果显示,测试集 DP1 的推理性能达到 54.12%(Four-

Way),单一关系的二元分类性能最优为 90.96%(如表 5 中的 Temporal 关系),最劣为 54.15%(如表 5 中的 Expansion 关系);测试集 DP2 的推理性能达到 51.29%,单一关系的二元分类性能最优为 96.42%,最劣为 50.75%.

另外,从表 5 可以看出,VSM_SYS 性能均优于 MA_SYS 性能.由此得出,计算“外联”关系对关联度权重时,基于向量空间模型的关联度估计方法优于基于协同测量的方法,能更好地获取“外联”关系对和提升隐式篇章关系的推理性能.本节将性能最高的推理系统 VSM_SYS 与基准系统以及各主流推理系统进行对比.

Table 5 Classification Results of Each Inference System

表 5 各推理系统分类结果											%
System	DP1					DP2					
	Exp	Cont	Comp	Temp	Four-Way	Exp	Cont	Comp	Temp	Four-Way	
VSM_SYS+DG1	56.87	72.33	84.64	90.95	54.12	53.82	69.63	81.62	95.91	50.98	
VSM_SYS+DG2	56.46	72.02	84.33	90.95	53.61	54.48	68.50	80.83	96.18	50.49	
VSM_SYS+DG3	56.30	72.66	84.15	90.96	53.76	53.97	69.76	81.80	96.07	51.29	
MA_SYS+DG1	54.15	72.71	83.84	89.87	51.62	50.94	69.68	80.89	96.42	49.44	
MA_SYS+DG2	54.57	71.87	80.59	89.73	50.21	52.31	67.62	79.48	95.73	48.04	
MA_SYS+DG3	54.71	72.01	82.84	89.73	51.48	50.75	69.68	80.89	96.42	49.34	

Notes: Exp stands for Expansion; Cont stands for Contingenay; Comp stands for Comparison; Temp stands for Temporal.

6.3 推理系统性能对比

本节将性能最高的推理系统 VSM_SYS 与基准系统以及各主流推理系统进行对比,如表 6 所示.对比系统为 WANG_SYS,LAN_SYS,在没有重现系统的情况下,为了与它们进行性能对比,本文使用与这 2 个系统相同的测试数据集:Section 21-22 (DP1)和 Section 23-24(DP2),分别在这 2 个数据集上进行性能评测.基准系统为 RANB,MAXB,EPIB,其中 RANB 和 MAXB 推理系统可直接通过统计方法获得分类性能;对于 EPIB,本文在 2 个测试数据集上重现该系统的推理方法,分别获得 50.56%和 46.53%的推理性能.

从表 6 可以看出,本文性能较优的推理系统 VSM_SYS在测试集DP1上最优性能达到54.12%,在测试集 DP2 上最优性能达到 51.29%.各对比系统中,WANG_SYS 采用基于树核函数的分类性能为 40.00%,LAN_SYS 基于多任务学习的方法推理性能达到 42.30%.实验结果显示,本文方法在推理隐式篇章关系中具有明显优势,相较于各对比系统,推理精确率分别获得 11.82%和 11.29%的性能提

升;相较于 EPIB 推理系统,推理精确率分别获得 3.56%和 4.76%的性能提升,这进一步验证本文提出的“显式指导隐式”优化模型的有效性和可行性.

Table 6 Performance of Each Implicit Discourse Relation Inference System

表 6 各隐式篇章关系推理系统性能				%
DP1(Section 21-22)		DP2(Section 23-24)		
Systems	Four-Way	Systems	Four-Way	
LAN_SYS	42.30	WANG_SYS	40.00	
RANB	25.00	RANB	25.00	
MAXB	53.72	MAXB	50.83	
EPIB	50.56	EPIB	46.53	
VSM_SYS	54.12	VSM_SYS	51.29	

虽然与 MAXB 相比,本文的最优性能仍然偏低,但 LAN_SYS 及 WANG_SYS 的方法推理性能远低于 MAXB 系统,这反映隐式篇章关系推理难度仍然较大,主要是因为隐式“论元对”本身具有较强的主观性和歧义性,从不同的角度考虑具有不同的语义关系.比如“He worked all night yesterday”和

“He slept all day today”,“论元对”之间可表示偶然关系也可表示时序关系. 此外,本文中使用的语块分析标注工具的精度也会影响实验结果.

实际上,本文实验中各推理系统的整体分类性能均已超过 LAN_SYS 和 WANG_SYS(如表 6 所示),这证明基于“外联”关系的隐式篇章关系推理模型的可行性. 此外,为了进一步验证本文提出的方法能够有效推理非毗连隐式“论元对”的篇章关系,本文进行了补充实验,借助 Gigaword 语言学资源,从中抽取 1 000 对非毗连隐式“论元对”作为测试样本进行人工标注,使用 EPIB, VSM_SYS, MA_SYS 实验系统进行实验. 图 5 为 3 个实验系统在该测试语料上的实验性能对比情况,可以看出,本文提出的基于外联关系的推理方法能够有效提升非毗邻隐式“论元对”的篇章关系推理性能.

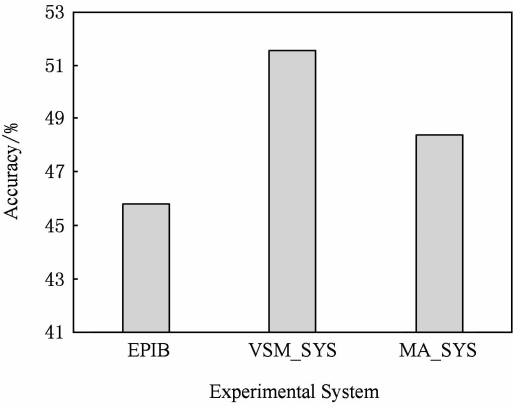


Fig. 5 Performance comparison of each inference system on the test data.

图 5 各实验系统在测试样本上的实验性能对比

本文提出的基于外联关系的隐式篇章关系推理方法的主要优势在于,该方法区分“内联”成分和“外联”成分,通过基于向量空间模型以及基于协同测量的关联度估计方法,有效识别出“外联”关系对,将“外联”关系比拟为“论元对”的篇章关系,从而将复杂的隐式篇章关系推理问题转化成“外联”关系评估与判定问题. 此外,本文借助大规模外部数据资源,构建“外联”关系的参考样本集合,将显式“参考对”作为关系推理线索,对“外联”关系进行评估与判定,有效避免“论元对”关系推理过程中自身线索匮乏的问题,最终实现隐式篇章关系的推理.

7 总 结

本文提出基于“外联”关系的隐式篇章关系推理

方法. 该方法继承“显式指导隐式”的关系推理模式,在此基础上通过基于向量空间模型以及基于协同测量的关联度估计方法识别出“外联”关系对,以“外联”关系对之间的关系为准绳,结合检索技术在大规模外部数据资源中挖掘包含连接词等直观推理线索的显式“论元对”,通过显式关系推理隐式关系. 其中,基于向量空间模型的关联度估计方法的系统推理性能较优,相比于对比系统,在 2 个测试数据集上分别获得 11.82%和 11.29%的性能提升.

然而,本文提出的方法也存在一些不足之处. 比如在大规模数据资源中挖掘显式“论元对”比较耗时,效率不高;通过“内联”成分,构建的“外联”关系对准准确率不高.

未来工作中,我们将对本文方法进行深入和细化,结合 LDA 模型以及篇章上下文等信息探索更加有效的方法提升“外联”成分的识别准确率,完善隐式篇章关系推理模型,并在此基础上扩展第 2 层的篇章关系识别,实现更细粒度的篇章关系推理.

参 考 文 献

[1] Prasad R, Dinesh N, Lee A, et al. The Penn discourse treebank 2.0 [C] //Proc of the 6th Int Conf on Language Resources and Evaluation (LREC). New York: ACM, 2008: 2961-2968

[2] Miltsakaki E, Robaldo L, Lee A, et al. Sense annotation in the Penn discourse treebank [G] //LNCS 4919: Journal of Computational Linguistics and Intelligent Text Processing. Berlin: Springer, 2008: 275-286

[3] Riaz M, Girju R. Another look at causality: Discovering scenario-specific contingency relationships with no supervision [C] //Proc of the 4th Int Conf on Semantic Computing (ICSC). Piscataway, NJ: IEEE, 2010: 361-368

[4] Do Q X, Chan Y S, Roth D. Minimally supervised event causality identification [C] //Proc of the Conf on Empirical Methods in Natural Language Processing (EMNLP). New York: ACM, 2011: 294-303

[5] Zhou Lanjun, Li Binyang, Gao Wei, et al. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities [C] //Proc of the Conf on Empirical Methods in Natural Language Processing (EMNLP). New York: ACM, 2011: 162-171

[6] Wang Jicheng, Wu Gangshan, Zhou Yuanyuan, et al. Research on automatic summarization of Web document guided by discourse [J]. Journal of Computer Research and Development, 2003, 40(3): 398-405 (in Chinese)

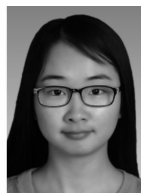
- (王继成, 武港山, 周源远, 等. 一种篇章结构指导的中文 Web 文档自动摘要方法[J]. 计算机研究与发展, 2003, 40 (3): 398-405)
- [7] Pitler E, Raghupathy M, Mehta H, et al. Easily identifiable discourse relations [C] //Proc of the 22nd Int Conf on Computational Linguistics (COLING). New York: ACM, 2008: 87-90
- [8] Wang Wenting, Su Jian, Tan C L. Kernel based discourse relation recognition with temporal ordering information [C] //Proc of the 48th Annual Meeting of the Association for Computational Linguistics (ACL). New York: ACM, 2010: 710-719
- [9] Zhou Zhimin, Xu Yu, Niu Zhengyu, et al. Predicting discourse connectives for implicit discourse relation recognition [C] //Proc of the 23rd Int Conf on Computational Linguistics (COLING). New York: ACM, 2010: 1507-1514
- [10] Marcu D, Echihiabi A. An unsupervised approach to recognizing discourse relations [C] //Proc of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). New York: ACM, 2002: 368-375
- [11] Saito M, Yamamoto K, Sekine S. Using phrasal patterns to identify discourse relations [C] //Proc of the 2006 Human Language Technology Conf of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL). New York: ACM, 2006: 133-136
- [12] Lin Ziheng, Kan M Y, Ng H T. Recognizing implicit discourse relations in the Penn discourse treebank [C] //Proc of the Conf on Empirical Methods in Natural Language Processing (EMNLP). New York: ACM, 2009: 343-351
- [13] Soricut R, Marcu D. Sentence level discourse parsing using syntactic and lexical information [C] //Proc of the 2003 Human Language Technology Conf of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL). New York: ACM, 2003: 149-156
- [14] Park J, Cardie C. Improving implicit discourse relation recognition through feature set optimization [C] //Proc of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL). New York: ACM, 2012: 108-112
- [15] Lan Man, Xu Yu, Niu Zhengyu. Leveraging synthetic discourse data via multi-task learning for implicit discourse relation recognition [C] //Proc of the 51st Annual Meeting of the Association for Computational Linguistics (ACL). New York: ACM, 2013: 476-485
- [16] Li J J, Nenkova A. Addressing class imbalance for improved recognition of implicit discourse relations [C] //Proc of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL). New York: ACM, 2014: 142-150
- [17] Li J J, Nenkova A. Reducing sparsity improves the recognition of implicit discourse relations [C] //Proc of the

15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL). New York: ACM, 2014: 199-207

- [18] Hong Yu, Zhou Xiaopei, Che Tingting, et al. Cross-argument inference for implicit discourse relation recognition [C] //Proc of the 21st ACM Int Conf on Information and Knowledge Management (CIKM). New York: ACM, 2012: 295-304



Hong Yu, born in 1978. PhD, associate professor in Soochow University. Member of China Computer Federation. His main research interests include personal information retrieval, topic detection and tracking, discourse analysis.



Zhu Shanshan, born in 1992. Master in Soochow University. Student member of China Computer Federation. Her main research interests include discourse analysis (zhushanshan063@gmail.com).



Ding Siyuan, born in 1992. Master in Soochow University. Student member of China Computer Federation. His main research interests include event relation detection (soochowlichking@sina.com).



Yao Jianmin, born in 1971. PhD, professor in Soochow University. His main research interests include machine translation (jmyao@szkj.gov.cn).



Zhu Qiaoming, born in 1963. PhD supervisor, professor in Soochow University. Senior member of China Computer Federation. His main research interests include Chinese information processing (qmqzhu@suda.edu.cn).



Zhou Guodong, born in 1967. PhD supervisor, professor in Soochow University. Senior member of China Computer Federation. His main research interests include natural language understanding, information extraction, statistical machine translation, machine learning (gdzhou@suda.edu.cn).