文章编号：

面向不平衡数据的隐式篇章关系分类方法研究 *

朱珊珊，洪宇，丁思远，姚建民，朱巧明

（苏州大学 计算机科学与技术学院，江苏省 苏州市 215006）

摘要：隐式篇章关系分类是篇章分析领域的一个重要研究子任务，大部分已有研究都假设参与分类的正类样本和负类样本数量相等，采用随机欠采样等不平衡数据处理方法保持训练样本中数据平衡，然而，在实际语料中正类样本和负类样本的分布是不平衡的，这一现象往往制约隐式篇章关系分类性能的有效提升。针对该问题，本文提出一种基于框架语义向量的隐式篇章关系分类方法，该方法借助框架语义知识库，将论元表示成框架语义向量，在此基础上，从外部数据资源中挖掘有效的篇章关系样本，对训练样本进行扩展，解决数据不平衡问题。在宾州篇章树库（Penn Discourse Treebank, PDTB）语料上的实验结果表明，相较于目前主流的不平衡数据处理方法，本文方法能够明显提高隐式篇章关系分类性能。

关键词：隐式篇章关系分类；不平衡数据；框架语义向量

中图分类号：TP391　　**文献标识码：**A

Research on Implicit Discourse Relation Recognition for Imbalanced Data

ZHU Shanshan, HONG Yu, DING Siyuan, YAO Jianmin, ZHU Qiaoming

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

Abstract: Implicit discourse relation recognition is an important subtask in the discourse analysis field. Most existing studies assume the balance between the numbers of positive and negative samples, and employ random under-sampling method to keep the training data well balanced. However, the training data has imbalanced distribution in reality that affect the recognition performance of the implicit discourse relation. To solve this problem, we propose a novel implicit discourse relation recognition method based on the frame semantic vectors. Firstly, we represent the argument as a frame semantic vector using the FrameNet resource, and then mine a number of effective discourse relation samples from the external data resources based on this new representation. Finally, we add the mined samples into the origin training data sets and perform experiment on this extended data sets. Evaluation on the Penn Discourse Treebank (PDTB) show that the proposed method perform better than the current mainstream imbalanced classification methods.

Key words: Implicit Discourse Recognition; Imbalanced Data; Frame Semantic Vectors

1　引言

篇章关系分类研究旨在自动推测同一篇章内两个文本片段（即“论元”, argument）之间的语义连接关系。宾州篇章树库（Penn Discourse Treebank, PDTB）[1][2]是2008年发布标注具体篇章关系类型的语言学资源，其将篇章关系类型分成三层（如图1所示）：Class层、Type层和Subtype层。Class层包括：Expansion（扩展关系）、Contingency（偶然关系）、Comparison（对比关系）和Temporal（时序关系）；Type层和Subtype层则分别针对上一层进行细分。

此外，依据“论元对”关系类别的不同识别方式，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）两种类型。在显式篇章关系类型中，两个“论元”之间存在连接词（例如连接词“*but*”，“*because*”等），可直接根据连接词判定篇章关系；而在隐式篇章关系类型中，两个“论元”之间缺少连接词

收稿日期：　　**定稿日期：**

基金项目：国家自然科学基金项目(61373097, 61272259, 61272260)

等直观推理线索，无法直接判定篇章关系，须结合上下文、句子语义结构等其它信息间接推理。在PDTB语言学资源中，标注者通过在隐式“论元对”中插入一个连接词表示具体的篇章关系类型。本文主要专注于Class层隐式篇章关系分类问题的研究。例1是从PDTB语料中抽取的具有隐式篇章关系的文本片段，图2给出标注的连接词及其对应的篇章关系类别。

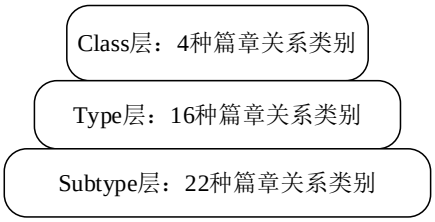


图1 PDTB篇章关系体系

例1 [*Mrs. Tom was fired and prosecuted under a South Carolina law that makes it a crime to breach test security.*]$_{arg1}$ [**Implicit=then**] [*In September, she pleaded guilty and paid a $500 fine*]$_{arg2}$ [**Implicit=but**] [*She never complained to school officials that the standardized test was unfair*]$_{arg3}$ [**Implicit=therefore**] [*Do I have much sympathy for her*]$_{arg4}$ [**Implicit=in fact**] [*Not really*]$_{arg5}$.

<译文：*依据南卡罗来纳法：违反安全测试是一种违法行为，汤姆小姐被解雇并同时被起诉。**【随后】** 在九月份，她承认罪行并支付了500美金的罚款。**【但是】** 她从没有向学校官员抱怨标准化测试是不公平的。**【因此】** 我同情她吗？**【实际上】** 并不是这样的。*>

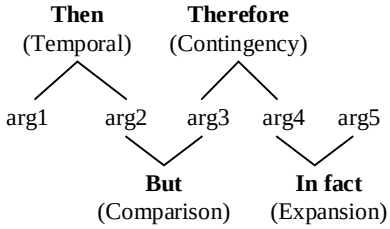


图2 例1中各“论元对”篇章关系标注结果

关于篇章关系分类的研究已开展多年，在显式篇章关系方面，分类精确率已达93.09%[3]。而在隐式篇章关系方面，分类精确率仍然较低。主流研究方法主要采用基于语言学特征的机器学习方法实现关系分类，在这种分类方法中，大多数研究者都假设参与分类的正类样本和负类样本的数量是相等的，通过随机欠采样等方法保持数据平衡。然而随机欠采样方法存在一个明显的缺点（注：为表述清楚，本文将样本中数量较多的一类称为多数类，将样本数量较少的一类称为少数类。）：欠采样过程从多数类样本中随机选择与少数类样本数量相等的样例，致使多数类样本中大量有用的样本被丢弃，在后续分类过程中未能发挥作用，从而影响整体的关系分类性能。

针对该问题，本文提出一种基于框架语义向量的训练样本扩展方法，旨在充分利用已标注的篇章关系样本，对不平衡的训练样本进行扩展，解决基于不平衡数据的隐式篇章关系分类问题。具体实现过程中，该方法借助框架语义知识库，将篇章关系样本表示成框架语义向量，借助框架语义向量，在外部未标注数据资源中挖掘篇章关系样本，实现对训练样本的扩展，从而解决数据不平衡问题。

本文的组织结构如下：第二节介绍相关工作；第三节描述框架语义知识库及框架语义向量；第四节给出基于框架语义向量的隐式训练样本集扩展方法；第五节给出实验结果及相关分析；第六节总结全文。

2 相关工作

2.1 隐式篇章关系分类

基于全监督学习的隐式篇章关系分类是目前的主流研究方法，主要包括：Pilter 等 [4]（2009）首次使用 PDTB 语言学资源，抽取论元的情感极性，动词类型及上下文特征进行关系分类，最终获得优于随机分类的性能，奠定隐式篇章关系分类研究的基础。Zhou 等 [5]（2010）根据语言模型困惑度在两个论元之间插入显式连接词，并将该连接词作为额外的分类特征，进一步提高隐式篇章关系的分类性能。Lin 等 [6]（2009）在 Pilter 的研究基础上，增加句法特征及依存特征。随后，Wang 等 [7]（2010）提出基于树核函数的隐式篇章关系分类方法，有效提升句法特征之间的区分能力，最终在 PDTB 语料上获得 40.0%的关系分类性能。Park 等 [8]（2012）采用前向选择算法对单词对、动词、极性、句法特征等 8 种特征进行特征选择，在每种关系类型上都获得一个最优的特征集合。Wang 等 [9]（2012）通过 SCC（single centroid clustering）聚类算法选择“典型”的训练样例，减少噪音文本。近期，Rutherford 等 [10]（2014）使用布朗聚类对特征代替单词对特征，Li 等 [11]（2014）通过改变句法特征的表示方法，有效解决特征表示的稀疏性问题。

2.2 不平衡数据分类方法

目前，主流的不平衡数据分类方法可分成两大类：采样技术及代价敏感函数方法。

其中，采样技术应用最为广泛，主要包括随机欠采样（Random Under-sampling）和随机重采样（Random Over-sampling）两种方法。详细而言，随机欠采样方法从多数样本中删除部分样例使得样本分不平衡；而随机重采样方法是从少数类样本中随机选择部分样例进行复制，直到多数类和少数类样本数量相等。Mani 等 [12]（2003）提出基于 K 近邻的欠采样方法，与随机欠采样方法相比，该方法通过 K 近邻算法从多数类样本选择需要删除的样例，保留多数类样本中有用的分类信息。Lin 等 [13]（2009）将采样技术与集成学习方法相结合，从多数类样本中抽取子集与少数类样本进行组合，训练多个分类器进行分类决策。Lin 等还提出一种基于平衡-级联算法的不平衡数据分类方法，该方法以监督学习方法为基础，通过训练多个分类器选择多数类样本中需要删除的样例。此外，Chawla 等 [14]（2002）提出基于少数类合成的过采样技术（简称 SMOTE 算法），该方法以少数类样本为种子样例，基于 K 近邻算法生成新的少数类样例，对少数类进行扩展。Han 等 [15]（2005）对 SMOTE 算法进行改进，对少数类样本进行归类，在此基础上，提出一种基于边界-少数类合成的采样方法。

上述采样技术主要通过调整样本数量保持数据平衡，代价敏感函数方法则是在分类过程中改变误分类的代价函数 [16]，保证在多数类样本中分错的代价大于在少数类样本中分错的代价。在此基础上，后续研究者提出代价敏感决策树和代价敏感神经网络，进一步解决不平衡数据分类问题。

3 框架语义知识库及框架语义向量

3.1 框架语义知识库

框架语义知识库（FrameNet）[1]是基于框架语义学（Frame Semantics）[17]构建的权威英文语义词汇资源。框架语义学由 Fillmore 于 1992 年提出，它是一种通向理解及描写词语和语法结构意义的方法。该理论的核心思想是为了理解语言中词的意义，首先要有一个概念结构，这个概念结构为词在语言及言语中的存在和使用提供背景和动因。表 1 给出 FrameNet 中相关术语定义及标注示例。从表 1 中的标注示例可以看出，两个标注示例包含不同的语义

[1] http://framenet.icsi.berkeley.edu/

信息，但它们具有相同的框架语义，目标词 cooks 和 fry 对应的框架语义均为 APPLY_HEAT，通过框架语义信息，可将两个具有不同语义信息的文本片段关联起来。

本文引入框架语义，主要动机在于框架语义有助于“论元”语义一级的描述，对于后续隐式训练样本的扩展，能够有效提升“论元对”的挖掘精度与广度，并提升其分类效率。目前，框架语义学领域已然形成多种自动框架语义分析与识别工具。本文采用 Dipanjan Das 等人开发的 SEMFOR[2]标注工具进行框架语义标注，该工具对给定的句子进行目标词与框架的有效识别。

表 1 FrameNet 相关术语定义及标注示例

相关术语定义
框架语义（Frame）：由目标词触发的语义场景，表征时间、状态或者实体。
目标词（Target）：触发框架语义的关键词语或者短语。
词元（Lexical Unit）：目标词与框架类型的组合。
框架元素（Frame Element）：除目标词外，描述框架语义的辅助词项。
元素类型（Frame Type）：框架元素对应的具体角色类型。

例 2 **a.** Mary cooks the potato.

<译文：玛丽煮了土豆>

框架语义：**APPLY_HEAT**； 目标词：**cooks**；词元：cooks + APPLY_HEAT；

框架元素 1：Mary → 元素类型：AGENT

框架元素 2：potato → 元素类型：FOOD

b. Fry the breadcrumbs gently.

<译文：轻轻地油炸面包屑>

框架语义：**APPLY_HEAT**； 目标词：**fry**；词元：fry + APPLY_HEAT；

框架元素 1：breadcrumbs → 元素类型：FOOD

框架元素 2：gently → 元素类型：MANNER

3.2 框架语义向量生成方法

本文使用 SEMAFOR 框架语义分析与识别工具对训练样本进行框架语义标注。在此基础上，将“论元”中的所有框架语义进行组合形成框架语义向量，利用该向量表示“论元”，实现“论元”的抽象描述，从而减少隐式篇章关系分类任务的复杂度。例 3 为标注的“论元对”实例，Arg1 中可识别出 3 个目标词：events, took place 和 years，其对应的框架语义分别为 Event, Event 和 Measure_duration，将它们组合起来形成框架语义向量 **Sf1**；同理 Arg2 中可识别出 has 等 5 个目标词，将它们对应的框架语义组合起来形成框架语义向量 **Sf2**。

例 3 Arg1：These events took place 35 years ago.

<译文：这些事件发生在 35 年前>

Sf1: (Event, Event, Measure_duration)

Arg2：It has no bearing on our work force today.

<译文：现在它对工作人员并没有什么影响>

Sf2: (Possession, Objective_influence, Working_on, Military, Calendric_unit)

[2] http://www.ark.cs.cmu.edu/SEMAFOR/

4　基于框架语义向量的隐式训练样本集扩展方法

4.1　隐式篇章关系分类数据分析

本文采用 PDTB 标注的隐式数据集作为实验数据集， 共包含 13,815 个实例。表 2 给出该数据集上四种篇章关系类别的实例数量、在语料中的比例以及正负类别比例。从表中可以看出，四种篇章关系类别的实例数量相差较大，正负不平衡比例介于 0~2。除了 Expansion 类别，其余 3 个关系类别（Comparison、Contingency 和 Temporal）的正例样本数量均小于负例样本数量。这种情况容易导致在这 3 个类别上训练的分类模型更倾向于将测试实例判定为负类，产生较大的误差，影响隐式篇章关系分类的整体性能。基于此，本文借助框架语义知识库，对实例数量较少的 3 个篇章关系类别进行样本扩展，解决隐式篇章关系分类过程中样本数据不平衡的问题。

表 2 PDTB 隐式数据集四种篇章关系分布

类别	实例个数	实例比例（%）	正负比例
Expansion	7,535	54.53	1.20
Comparison	2,076	15.03	0.24
Contingency	3,464	25.07	0.33
Temporal	741	5.36	0.06

4.2　未标注篇章关系样本挖掘方法

本文采用的外部数据资源为GIGAWORD纽约时报语料，共包含1,298,498篇新闻文本。在进行训练样本扩展之前，本文对 GIGAWORD 中所有文本进行切分，为了验证本文的方法能够有效地选择与测试样本语义相近的隐式“论元对”，本文将 GIGAWORD 样本分别切分成显式篇章关系样本和隐式篇章关系样本，下面详述这两种切分方法。

1)　显式篇章关系样本切分

该方法以 PDTB 语言学资源中的 Golden 连接词为基础，从 GIGAWORD 文本中切分获得显式篇章关系样本，切分后的文本符合以下两个条件：

- 以“论元对”为单元，即包含前置论元 Arg1 和后置论元 Arg2。
- Arg2 中的第一个单词为 Golden 连接词 [3]，且将 Golden 连接词作为未标注“论元对”的先验知识，“论元对”具有显式篇章关系。

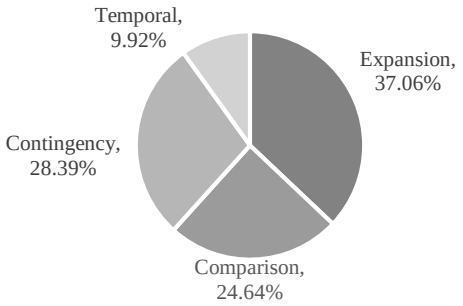


图 3 GIGA-Explicit 显式样本四种篇章关系分布情况

[3] Golden 连接词：指向某一特定篇章关系的概率大于 96%，例如连接词“*now*”唯一地指向 Temporal（时序关系），PDTB 共统计得出 87 个 Golden 连接词。

按照上述切分条件，本文共获得 2,520,777 个显式“论元对”（简写为 GIGA-Explicit），四种篇章关系分布比例如图 3 所示。图 4 为显式“论元对”数量较多的 Top10 Golden 连接词，从图中可以看出，包含“*or*”，“*so*”，“*for*”等连接词的显式“论元对”在语料中所占比例较大，导致 Expansion 篇章关系类别在语料中的比例最大（如图 3 中 Expansion 在所有挖掘的 GIGA-Explicit 篇章关系样本中的比例为 37.06%）。

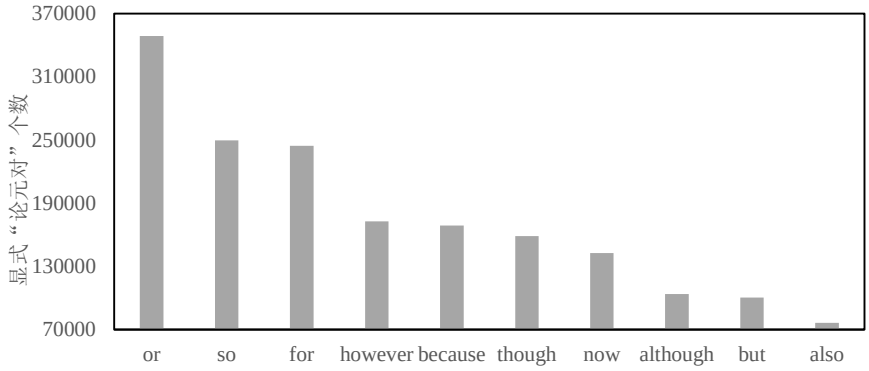


图 4 GIGA_Explicit 样本中显式“论元对”数量较多的 Top10 Golden 连接词

2) 隐式篇章关系样本切分

与显式篇章关系样本切分方法类似，该方法将将 GIGAWORD 文本切分成隐式篇章关系样本，切分后的文本须满足以下两个条件：

- 以“论元对”为单元，即包含前置论元 Arg1 和后置论元 Arg2。
- “论元对”中不存在连接词，即“论元对”具有隐式篇章关系。

与显式篇章关系样本切分方法的唯一不同的是，该方法不以 Golden 连接词为先验知识，“论元对”的篇章关系类别不确定。此外，在文本切分过程中，本文通过句法分析确保挖掘到的隐式“论元对”符合自然语言规律。本文最终切分获得 9,08,142 个隐式“论元对”（简写为 GIGA-Implicit）。

4.3 基于框架语义向量的训练样本集扩展方法

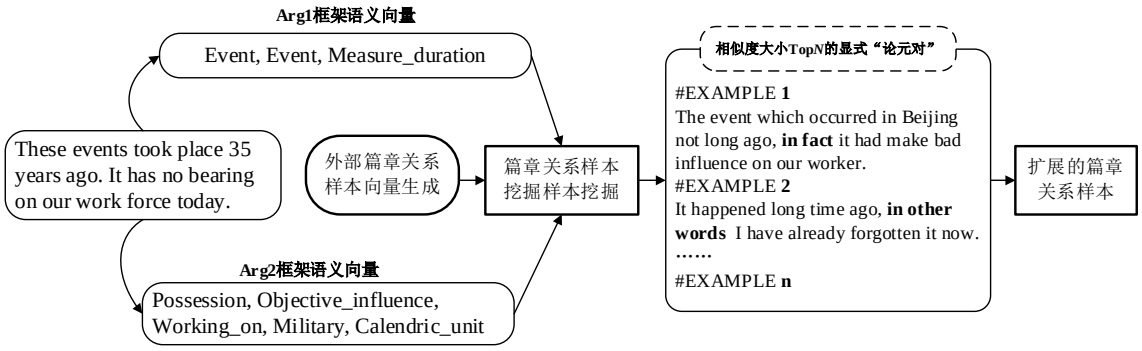


图 5 基于 GIGA-Explicit 显式样本扩展训练样本的实例化说明

对 4.2 节构建的两个篇章关系样本 GIGA-Explicit 和 GIGA-Implicit，本文使用 SEMFOR 语义框架标注工具对所有样本进行标注，获得样本的框架语义向量。在此基础上，以 PDTB 中隐式训练样本为种子样例，分别计算每个种子样例的框架语义向量与两个篇章关系样本中样例对应的框架语义向量之间的语义相似度，根据相似度计算结果排序，选择与当前种子样例最相似的 Top*N*“论元对”作为扩展的训练样本。其中，语义相似度计算方法如公式（1）

所示，Arg_1Sim 表示前置论元 Arg1 框架语义向量之间的余弦相似度，Arg_2Sim 表示后置论元 Arg2 框架语义向量之间的余弦相似度，特征权重使用框架语义在论元中的出现频数。图 5 为基于 GIGA-Explicit 显式样本扩展训练样本方法的实例化流程图。

$$Sim = \frac{\mathrm{Arg}_1Sim + Arg_2Sim}{2e^{|\mathrm{Arg}_1Sim - Arg_2Sim|}}, \quad (1)$$

特别地，由于显式“论元对”和隐式“论元对”之间存在不同的语义特性，在篇章关系样本 GIGA-Explicit 和 GIGA-Implicit 中挖掘扩展“论元对”时，存在以下两点不同之处：

- 在 GIGA-Explicit 显式篇章关系样本中，“论元对”的篇章关系类别是确定的。在挖掘过程中，由于存在“噪音”文本，与种子样例最相似的 Top*N* 显式“论元对”中可能会出现篇章关系类别不一致的情况，即当前种子样例的篇章关系类别为 R_x，挖掘到的“论元对”的先验篇章关系类别为 R_y，$R_x \neq R_y$。针对这种情况，本文在选择扩展“论元对”之前，删除与种子样例篇章关系类别不一致的显式“论元对”，在此基础上，选择与种子样例最相似的 Top*N* 显式“论元对”作为扩展样本。
- 在 GIGA-Implicit 隐式篇章关系样本中，“论元对”的篇章关系类别不确定。根据 Hong 等[18]（2012）提出的“平行推理机制”理论， 与种子样例最相似的 Top*N* 隐式“论元对”在关系上是平行的，即 Top*N* 隐式“论元对”的篇章关系与种子样例的篇章关系相同，可直接将挖掘到的隐式“论元对”作为扩展样本。

5 实验

5.1 实验设置

本文使用 PDTB 隐式数据集中 Section 02-20 作为训练数据集，Section 21-22 作为测试数据集，Section 00-01 作为验证数据集。各数据集在四种篇章关系类别上的分布情况如表 3 所示。本文使用词向量（Semantic Vector）[4]作为分类特征，向量维度设定为 100 维。

表 3 实验数据集四种篇章关系分布

数据集	实例个数 (所占比例 %)				实例个数
	Expansion	Comparison	Contingency	Temporal	总数
Section 00-01	656(55.45)	193(16.31)	279(23.58)	55(4.65)	1,183
Section 02-20	6,878(54.45)	1883(14.91)	3,185(25.21)	686(5.43)	12,632
Section 21-22	562(53.72)	145(13.86)	269(25.72)	70(6.69)	1,046

此外，本文使用 LIBSVM（Chang 等[19]）作为分类器，核函数选用线性核函数。针对每种篇章关系类别，分别训练一个二元分类器，计算获得每个篇章关系类别的分类精确率（*Accuracy*）（如公式（2）所示），公式（2）中 *TP* 和 *TN* 分别表示被正确分为正例和负例的个数。整体性能评价标准使用精确率的宏平均（*Micro-average Accuracy*）（如公式（3）所示），其中 *R*={Expansion, Comparison, Contingency, Temporal}。

$$Accuracy = \frac{TP + TN}{NumOfInstances} \quad (2)$$

$$Micro\text{-}average\ Accuracy = \frac{\sum_{r \in R} Accuracy(r)}{Num(R)} \quad (3)$$

[4] http://nlp.stanford.edu/software/lex-parser.shtml

5.2 实验系统

表 4 列出参与实验的各分类系统，编号 2-9 为基于主流不平衡分类方法的实验系统，编号 10-11 为本文提出的基于框架语义向量的不平衡隐式篇章关系分类系统，其中 Expand-Explicit 系统使用 GIGA-Explicit 显式篇章关系样本，Expand-Implicit 系统使用 GIGA-Implicit 隐式篇章关系样本。

表 4 实验系统

编号	实验系统	描述
1	Baseline	直接在原始训练样本上进行模型训练
2	Random-US	在多数类样本中使用随机欠采样方法删除部分样例
3	Random-OS	在少数类样本中使用随机重采样方法复制部分样例
4	KNN-US[12]	在多数类样本中使用基于 K 近邻的欠采样方法删除部分样例
5	Easy-Ensemble[13]	基于集成学习方法训练多个分类器进行分类决策
6	Balance-Cascade[13]	基于平衡-级联算法实现关系分类
7	SMOTE[14]	少类样本合成过采样技术
8	Borderline-SMOTE[15]	边界区域少类样本合成过采样技术
9	Meta-Cost-Sensitive[16]	基于代价敏感函数的关系分类方法
10	Expand-Explicit	基于框架语义向量的训练样本扩展方法（GIGA-Explicit）
11	Expand-Implicit	基于框架语义向量的训练样本扩展方法（GIGA-Implicit）

5.3 实验结果及分析

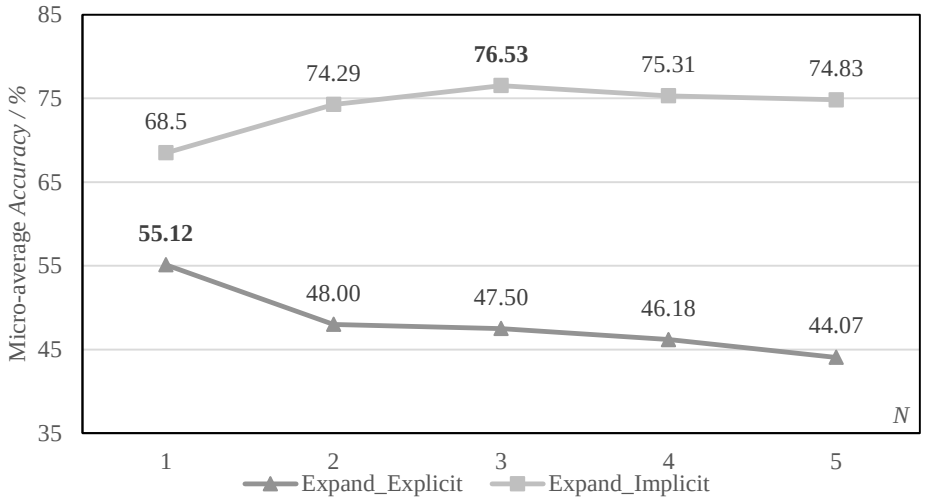


图 6 *N* 的不同取值对应的整体分类性能

针对每个种子样例，本文在外部篇章关系样本中选择与其最相似的 Top*N*“论元对”作为扩展样本，图 6 分别给出在 Expand-Explicit 和 Expand_Implicit 两个实验系统中参数 *N* 的不同取值对应的整体分类性能变化图。从图 6 中可以看出，*N* 取值分别为 3 和 1 时，两个实验系统能够获得最优的分类性能，且在参数调整过程中，Expand_Implicit 实验系统的整体分类性能均明显优于 Expand-Explicit 实验系统。

图 7 为各实验系统的实验性能对比情况，从图中可以看出，与 Baseline 系统对比，Expand-Explicit 实验系统的实验性能获得小幅度的提升，整体分类精确率提升 6.75%，Expand_Implicit 实验系统的分类性能提升幅度较大，整体分类精确率提升 28.16%。结合图 6 和图 7，分析原因可知，Expand-Explicit 实验系统扩展的训练样本来自 GIGA-Explicit 篇章

关系样本，样本中的实例包含连接词，而待扩展的原始训练样本均不包含连接词，连接词的缺失导致两种篇章关系样本在语义上存在差异，随着扩展的训练样本的增加，实验系统的分类性能有所下降。而在 Expand_Implicit 实验系统中，本文方法借助框架语义向量，从 GIGA-Implicit 篇章关系样本中挖掘隐式“论元对”加入训练样本中，在各个篇章关系类别上引入了更多的分类信息，有效地提升了篇章关系分类性能。

从图 7 中还可以看出，相较于各主流不平衡数据分类方法的实验系统，本文性能较优的 Expand-Implicit 实验系统有效提升了整体分类精确率，与主流方法性能最优的基于代价敏感函数的 Meta-Cost-Sensitive 实验系统进行对比，整体分类精确率提升 5.19%。分析原因可知，各主流不平衡数据分类方法侧重通过采样或者改变错误权重等方法解决训练样本数据不平衡问题，这些方法往往局限在有限的数据资源中，忽略了不平衡样本数据本身存在信息不充分的问题，影响篇章关系分类性能。针对这一问题，本文借助框架语义向量，利用大规模外部数据资源，挖掘有效的隐式篇章关系样本，对样例数量较少的三个篇章关系类别进行样本扩展，提升了整体篇章关系分类性能。实验结果也证明本文提出的基于框架语义向量的方法能够从外部数据资源中有效的挖掘隐式篇章关系样本，从而对原始训练样本进行扩展，辅助篇章关系分类任务。

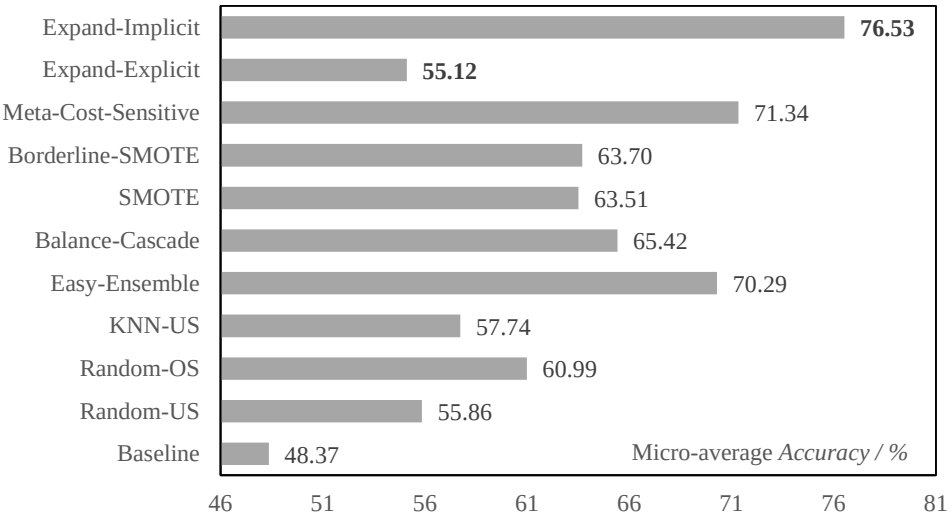


图 7 各实验系统性能对比

表 5 各隐式篇章关系推理系统性能

实验系统	不平衡分类方法	微平均精确率（%）
Pilter-SYS	随机欠采样	65.40
Zhou-SYS	随机欠采样	62.16
Park-SYS	随机欠采样	73.80
Li-SYS	随机欠采样	64.69
Wang-SYS	基于 SCC 聚类算法的样本选择方法	67.03
Expand-Implicit	基于框架语义向量的训练样本扩展	**76.53**

此外，表 5 给出本文性能最优的 Expand_Implicit 实验系统以及各主流隐式篇章关系分类系统的实验性能对比，从表中可以看出，本文提出的基于框架语义向量的隐式训练样本扩展方法性能提升明显，相较于性能较优的 Park-SYS 实验系统，整体分类精确率提升 2.73%，这也进一步证明了本文基于框架语义向量进行训练样本扩展的方法具有一定的有效性和可

行性，与主流方法采用的随机欠采样方法相比，能够获得更优的分类性能。

6 总结

本文研究隐式篇章关系分类任务中的不平衡数据分类问题，提出一种基于框架语义向量扩展训练样本的分类方法。实验结果显示，本文方法能够很好的解决隐式篇章关系分类任务中数据不平衡的问题，相较于传统的基于原始训练样本的采样方法以及代价敏感函数方法，实验性能获得显著提升。

然而本文方法仍存在不足之处，将论元表示成框架语义向量，可能存在数据稀疏问题：统计发现，论元中识别出的框架语义平均数量为 6 个，在某些文本较短的论元中，由于识别出的框架语义较少，形成的框架语义向量并不能很好的表示该论元，影响后续训练样本扩展的精确率。基于此，在未来工作中，我们将对本文方法进行细化，根据论元的框架语义数量对论元进行筛选，选择符合要求的“论元对”，并尝试采用 Stacked Learning、Tri-training 等多分类器的学习方法实现隐式篇章关系分类任务。

参 考 文 献

[1] R. Prasad, N. Dinesh, A. Lee, et al. The Penn Discourse TreeBank 2.0[C] //Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC), 2008:2961-2968.

[2] E. Miltsakaki, L. Robaldo, A. Lee, et al. Sense Annotation in the Penn Discourse Treebank[C] //Proceedings of the Computational Linguistics and Intelligent Text Processing. Springer Berlin Heidelberg, 2008:275-286.

[3] E. Pitler, M. Raghupathy, H. Mehta, et al. Joshi. Easily Identifiable Discourse Relations[R]. Technical Reports (CIS), 2008:87-90.

[4] E. Pitler, A. Louis, and A. Nenkova. Automatic Sense Prediction for Implicit Discourse Relations in Text[C] //Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-AFNLP), 2009, Volume 2:683-691.

[5] Z. M. Zhou, Y. Xu, Z. Y. Niu, M. Lan, J. Su, and C. L. Tan. Predicting Discourse Connectives for Implicit Discourse Relation Recognition[C] //Proceedings of the 23rd International Conference on Computational Linguistics (COLING): Posters, 2010:1507-1514.

[6] Z. H. Lin, M. Y. Kan, and H. T. Ng. Recognizing Implicit Discourse Relations in the Penn Discourse Treebank[C] //Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2009, Volume 1:343-351.

[7] W. T. Wang, J. Su, and C. L. Tan. Kernel Based Discourse Relation Recognition with Temporal Ordering Information[C] //Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL), 2010:710-719.

[8] J. Park and C. Cardie. Improving Implicit Discourse Relation Recognition Through Feature Set Optimization[C] //Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), 2012:108-112.

[9] X. Wang, S. J. Li, J. Li, et al. Implicit Discourse Relation Recognition by Selecting Typical Training Examples[C]// Proceedings of the 24rd International Conference on Computational Linguistics (COLING). 2012: 2757-2772.

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

[11] J. J. Li, M. Carpuat, A. Nenkova. Cross-lingual Discourse Relation Analysis: A corpus study and a semi-supervised classification system[C] //Proceedings of the 25th International Conference on Computational Linguistics (COLING), 2014: 577-587.

[12] I. Mani, J. P. Zhang. KNN approach to unbalanced data distributions: a case study involving information extraction[C]//Proceedings of Workshop on Learning from Imbalanced Datasets. 2003.

[13] X. Y. Liu, J. Wu, Z. H. Zhou. Exploratory under-sampling for class-Imbalance learning [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2009, 2(39): 539-550.

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, et al. SMOTE: synthetic minority over-sampling technique [J]. Journal of artificial intelligence research, 2002: 321-357.

[15] H. Han, W. Y. Wang, B. H. Mao. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning [M] //Advances in intelligent computing. Springer Berlin Heidelberg, 2005: 878-887.

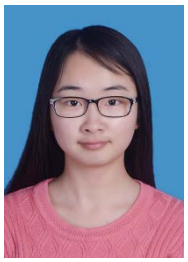
[16] C. Elkan. The foundations of cost-sensitive learning[C]//International joint conference on artificial intelligence (IJCAI). Lawrence Erlbaum Association Ltd, 2001, 17(1): 973-978.

[17] C. Fillmore. Frame semantics [J]. Linguistics in the morning calm, 1982: 111-137.

[18] Y. Hong, X. P. Zhou, T. T. Che, et al. Cross-argument inference for implicit discourse relation recognition[C] //Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM), 2012: 295-304.

[19] C. C. Chang, C. J. Lin. LIBSVM: a library for support vector machines [J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2001, 2(3): 389-396.

作者简介：朱珊珊（1992—），女，硕士研究生，主要研究领域为篇章分析。 Email: zhushanshan063@gmail.com；洪宇（1978—），通讯作者，男，副教授，主要研究领域为信息抽取，信息检索，事件关系检测等。 Email: tianxianer@gmail.com； 丁思远（1992—），男，硕士研究生，主要研究领域为事件关系检测。Email: dsy.ever@gmail.com



朱珊珊

洪宇

丁思远