《中文信息学报》稿件排版格式

文章编号：1003-0077（2011）00-0000-00

基于语言现象的文本蕴涵识别*

任函 [1,2]，冯文贺 [1,2]，刘茂福 [3,2]，万菁 [2]

（1.广东外语外贸大学语言工程与计算实验室，广东省 广州市 510006；2. 武汉大学湖北语言与智能信息处理研究基地，湖北省 武汉市 430072；2. 武汉科技大学计算机学院，湖北省 武汉市 430065；）

摘要：本文提出一种基于语言现象的文本蕴涵识别方法，该方法建立了一个语言现象识别和整体推理判断的联合分类模型，目的是对两个高度相关的任务进行统一学习，避免管道模型的错误传播问题并提升系统精度。针对语言现象识别，设计了 22 个专用特征和 20 个通用特征；为提高随机森林的泛化能力，提出一种基于特征选择的随机森林生成算法。实验结果表明，基于随机森林的联合分类模型能够有效识别语言现象和总体蕴涵关系。

关键词：文本蕴涵识别；语言现象；随机森林

中图分类号：TP391　　**文献标识码：**A

Recognizing Textual Entailment Based on Inference Phenomena

Han Ren[1,2], Wenhe Feng[1,2] ,Maofu Liu[3,2],Jing Wan[2]

(1.Laboratory of Language Engineering and Computing, Guangdong University of Foreign Studies ,Guangzhou, Guangdong 510006, China ; 2. Hubei Research Center for Language and Intelligent Information Processing, Wuhan University, Wuhan, Hubei 430072 ,China ; 3. College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, Hubei 430065, China)

Abstract: This paper introduces an approach of textual entailment recognition based on language phenomena. The approach gives a joint classification model for language phenomenon recognition and entailment recognition, which is to learn with two highly relevant tasks, avoiding error propagation and improving system performances. For language phenomenon recognition, 22 specific and 20 general features are employed, while for enhancing the generalization of random forest, a feature selection method is adopted on building trees of random forest. Experimental results show that the joint classification model based on random forest recognizes language phenomena and entailment relation effectively.

Key words: Recognizing Textual Entailment; Language Phenomena; Random Forest

1 引言

文本蕴涵识别（Recognizing Textual Entailment）是一个判断命题之间逻辑推导关系的挑战任务，其定义为：给定一个语段 T（Text）和一个假设 H（Hypothesis），如果 H 的意义可以从 T 的意义中推断出来，那么就认为 T 蕴涵 H，记为$T \rightarrow H$[1]。文本蕴涵识别是自然语言理解的重要研究课题之一，能够广泛应用于问答系统、多文档自动摘要、信息抽取、机器阅读等自然语言处理应用[2, 3]。

文本蕴涵识别需要考察多种推理关系，例如词义、句法和语义变换。现有文本蕴涵识别研究往往集中于针对某一特定类型的推理问题设计精确的解决方案，这种方式虽然能够提高针对这类问题的推理能力，然而由于文本蕴涵识别涉及的推理关系众多，使得这种方式对于

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金（61402341）；国家社会科学基金重大项目（11&ZD189）；华中师范大学中央高校基本科研业务费教育科学专项资助（ccnu16JYKX014）；教育部人文社科项目（13YJC740022）；河南高校哲社基础研究重大项目（2015-JCZD-022）

文本蕴涵识别的整体性能提升非常有限[4]。为此，一些文本蕴涵识别研究尝试对推理中涉及的语言现象进行分类，并据此建立语言现象的标注方法和资源[5, 6, 7]。例如：

T：乔姆斯基是 20 世纪最伟大的语言学家之一，他提出了形式语法理论。
H：乔姆斯基创立了形式语法理论。

其中，“提出”和“创立”属于词义蕴涵（Lexical Entailment）现象，“他”和“乔姆斯基”属于指代（Coreference）现象。显然，获取这些语言现象将有助于对两个句子的蕴涵关系进行判断。

目前，针对文本蕴涵中语言现象的研究主要集中在资源标注方面，而利用标注的语言现象进行文本蕴涵识别的相关研究则非常缺乏。本文提出一种基于语言现象的文本蕴涵识别方法。该方法建立了一种语言现象识别和整体推理判断的联合分类模型，对两个高度相关的任务进行统一学习，避免了管道模型的错误传播问题。针对语言现象识别，设计了 22 个专用特征和 20 个通用特征；为提高随机森林的泛化能力，提出一种基于特征选择的随机森林生成算法。实验结果表明，基于随机森林的联合分类模型能够有效识别语言现象和总体蕴涵关系。

本文第二部分简要介绍基于语言现象的文本蕴涵识别相关工作，第三部分介绍基于语言现象的文本蕴涵识别模型，第四部分对实验结果进行分析，第五部分对全文工作进行总结和展望。

2 相关工作

基于语言现象的文本蕴涵识别策略通过分析语言现象获取局部片断的推理关系，再进行整体蕴涵判断。该策略一般涉及资源建设和蕴涵识别两个部分。

2.1 资源建设

现有语言现象的资源建设工作主要基于英语。Garoufi[8]从对齐、上下文及指代三个方面归纳了 23 种现象来标注 T 和 H 的推理关系。他在 RTE-2 的测试数据集上共标注了 400 个蕴涵的文本对，并随机选取了 25%的矛盾类进行标注。Sammons 等[7]定义了 39 类语言现象，并在 RTE-5 中挑选了 210 个文本对上进行标注，然后用标注结果对现有 RTE 参赛系统进行评估。Bentivogli[5]将语言现象归为词汇、句法、词汇-句法关系、篇章及推理五大类，在 RTE-5 数据集上进行了 90 个文本对的标注实践。这一工作与其它工作的区别在于，T 和 H 被分解成一系列推理过程，每次分解的结果用（T, H_i）表示，其中 T 为原始语段，H_i 表示一系列假设，然后通过人工总结这一系列（T，H_i）中所含语言现象里存在的推理关系。

此外，Kaneko 等[6]定义了 26 类推理现象，并用于标注 RITE-2 任务中的日语语料。而第一份中文语言现象标注语料则由 RITE-3 任务[9]给出，其中包括 19 类蕴涵现象和 9 类矛盾现象，共标注了 581 对训练集和 1200 对测试集数据中的语言现象。

从规模上看，这些资源标注数量比较有限，但他们的工作使得语言现象的标注资源在推理中的作用显得更为重要，并也形成了一些可供参考的标注资源。

2.2 蕴涵识别

基于语言现象的蕴涵识别还是一个鲜有涉足的研究领域。Huang 等[10]对推理现象识别进行了初步探索。他们考察了矛盾类语言现象，并为每类现象总结出启发式规则。为考察语言现象的识别效果，他们设计了两个实验，第一个实验分别统计机器和人工识别语言现象的准确率；在第二个实验中，他们将自动识别的语言现象作为特征，放入 SVM 进行训练。第一个实验结果显示，机器标注的结果（52.38%）与人工结果（95.24%）的性能相去甚远，但第二个实验结果显示，仅利用 5 个矛盾类语言现象作为特征进行学习得到的分类器性能与 RTE-5 全部参评系统的平均准确率相当。这在一定程度上体现语言现象对文本推理系统的有

效性。然而，到目前为止，还没有利用语言现象进行文本蕴涵识别的大规模研究。

3 基于语言现象的文本蕴涵识别

本文提出一种基于语言现象的文本蕴涵识别方法。该方法建立了一种语言现象识别和整体推理判断的联合分类模型，并利用改进的随机森林方法进行训练和预测。

3.1 语言现象类别

本文实验基于中文蕴涵语料，为此，我们以 RITE-3 评测任务中定义的汉语语言现象为基础定义本实验中的语言现象类别。RITE-3 语料包括 19 类蕴涵现象和 9 类矛盾现象，共标注了 581 对训练集和 1200 对测试集数据中的语言现象。我们对其定义的语言现象进行如下改进：

1）将 Relative_clause 与 Clause 合并，称为 Clause 现象，原因是两者所表示的语言现象非常相近，都是 T 中包含了 H 中没有的句法成分。

2）将 Antonym、Exclusion:modality 和 Exclusion:modifier 合并，称为 Antonym 现象，原因是后两者所表示的语言现象属于意义相对的成分，与 Antonym 包含对义关系相似。

3）去掉 Paraphrase、Inference 和 Exclusion:common_sense 三类语言现象，原因是这三类现象体现了对文本的解释和重写，而非仅仅是词汇或句法的替换，识别这类语言现象已相当于对整体进行推理判断。因此，我们将包含这三类现象的文本对直接利用推理判断模型进行识别，不再为其指定语言现象类别。

改进后的语言现象包括 16 类蕴涵现象和 6 类矛盾现象，如表 1 所示。

3.2 语言现象识别

语言现象识别的任务是，找出 T 和 H 中包含的语言现象。一种方法是，为每类语言现象设计对应的规则，若 T 和 H 中存在符合规则的文本片断对，则认为存在该语言现象。例如：

T：水蕴草为雌雄异株的植物。

H：水蕴草为雌雄异株的生物。

该语言现象为“上下位关系”，可为其制定启发式规则：若 T 存在某一词语，H 中存在其上位词，则认为该文本对包含“上下位关系”这一语言现象。该方法对于比较简单的词汇类语言现象具有一定的识别能力。然而，对于比较复杂的文本，简单的规则往往导致准确率不高；若编制比较复杂的规则，又会面临召回率降低的问题，其原因在于约束条件过多。Huang 等[10]的实验也表明，采用规则方法难以获得理想的识别性能。

本文提出一种基于机器学习的方法，将语言现象识别看作一个学习问题，即首先通过训练数据获得语言现象识别知识，再对测试数据进行预测。为此，我们定义了一组专用特征，如表 1 所示。这些专用特征所覆盖了本文定义的语言现象。

专用特征可分为两类，一类为词汇类特征，另一类为句法、语义类特征。绝大多数词汇类特征都需要利用世界知识进行判断。如缩略词、上下位、同义词等。我们使用同义词词林、HowNet、百度汉语[1]、金山词霸汉语[2]等词典识别同义、反义、上下位关系、整体-部分关系等语言现象。对于缩略语现象，除采用以上资源进行识别外，还利用规则从中文维基百科中抽取缩略语集合进行识别。对于对义关系，采用一种基于 HowNet 词汇语义相似度的方法[11]进行计算，该方法利用了义原的反义、对义关系和义原信息计算词汇相似度。对于词汇蕴涵关系，采用一种基于词向量的方法[12]进行计算，该方法从中文维基百科语料上训练出 100 维的词向量，并利用分类的方法识别词汇蕴涵关系。对于 Spatial 现象，利用已抽取的地理

[1] http://hanyu.baidu.com/
[2] http://hanyu.iciba.com/

信息资源[13]进行识别。

表 1 语言现象专用特征

ID	特征名称	取值范围	描述
1	Abbreviation	0/1	T 和 H 中是否存在缩略词及对应的全称
2	Apposition	0/1	若 T/H 中某一词在 H/T 中是否存在对应同位语，则该特征为 1
3	Case_alternation	0/1	T 和 H 中是否存在使成式、处置式、被动式变化
4	Clause	0/1	T 中是否包含 H 中不存在的句法成分
5	Coreference	0/1	T 中代词所指是否也出现在 H 中
6	Hypernymy	0/1	若 T 中某一词的上位词存在于 H 中，则该特征为 1
7	Lexical_entailment	[0,1]	若 T 中某一词与 H 中某一词存在蕴涵关系，则该特征值为蕴涵程度；若不存在蕴涵关系，则该值为 0
8	List	0/1	若 H 中的并列成分（由顿号、逗号或“和”连接的具有平等和相同句法关系的成分）均存在于 T 中，则该特征为 1
9	Meronymy	0/1	若 T 中某个词与 H 中某个词存在部分-整体关系，则该特征为 1
10	Modifier	0/1	若 T 中的某个中心词修饰语在 H 中被省略，则该特征为 1
11	Quantity	0/1	若 T 和 H 中的数量存在正确换算关系或数量一致，则该特征为 1
12	Scrambling	0/1	若 T 和 H 中的词汇相同，语序不同，则该特征为 1
13	Spatial	0/1	若 T 和 H 中存在空间地理信息指示词，且 H 中的词包含或从属于 T 中的词，则该特征为 1
14	Synonymy	[0,1]	T 和 H 中若存在同义词，则该特征值为相似程度；若不存在同义词，则该值为 0
15	Temporal	0/1	若 T 和 H 中存在时间表示，且 T 中的时间表示等价于或包含于 H 中的词，则该特征为 1
16	Transparent_head	0/1	若 T 中的某一词的中心词在中的句法位置被该词替代，则该特征为 1
17	Antonym	[0,1]	T 和 H 中若存在意义相反或相对的词，则该特征值为反义或对义程度；若不存在反义词，则该值为 0
18	Exclusion:predicate_argument	0/1	T 和 H 中是否存在谓词和相关论元不一致
19	Exclusion:quantity	0/1	若 T 和 H 中的数量存在错误换算关系或数量不一致，则该特征为 1
20	Exclusion:spatial	0/1	若 T 和 H 中存在空间地理信息指示词，且 H 中的词不包含和从属于 T 中的词，则该特征为 1
21	Exclusion:temporal	0/1	若 T 和 H 中存在时间表示，且 T 中的时间表示不等价和包含于 H 中的词，则该特征为 1
22	Negation	0/1	若 T 和 H 中，一方存在否定词而另一方没有，则该特征为 1

对于句法、语义类特征，首先利用 Stanford CoreNLP[3]工具对 T 和 H 进行句法和语义分析，再利用结果进行识别。特别地，对于 Coreference 特征，利用上述工具进行指代消解，再进行识别；对于 Case_alternation、List 特征，首先为每种句式制定相应匹配规则，再结合句法分析结果进行结构匹配。

定义专用特征的目的是描述特定语言现象，即每一个特征描述一种特定的语言现象。然

[3] http://nlp.stanford.edu/

而，仅凭专用特征难以描述完整地描述语言现象。为此，我们加入了通用特征，这些通用特征包括词汇、句法和语义的相关性特征，目的是联合专用特征进行语言现象识别。通用特征有助于语言现象的识别，例如当词汇相似度较高、句法相似度较低，并且 Case_alternation 特征为真时，表明文本对存在句式变换的可能性较高。

通用特征利用了我们提出的 15 种蕴涵识别特征，包括重叠特征、相似度特征、结构特征和语言学特征[13]。此外，还利用了以下 5 种特征：Jaro-Winkler 距离、Manhattan 距离、切比雪夫距离、欧式距离和 Jaccard 相似度。

3.3 文本蕴涵识别

文本蕴涵识别的任务是，利用语言现象识别结果对文本对(T, H)进行整体推理判断。这一步骤是必要的，因为蕴涵或矛盾语言现象存在并不代表 T 和 H 具有蕴涵或矛盾关系。例如：

T：美国疾病控制与预防中心通报美国首宗爱滋病感染案例。

H：美国疾病控制与预防中心通报全球首宗爱滋病感染案例。

尽管“美国”包含于“全球”，但 T 和 H 并不具有蕴涵关系，理由很明显：局部推理关系并不能代表总体推理关系。因此，除语言现象识别结果外，我们还需结合上下文才能进行整体推理判断。

文本蕴涵识别的一种主要策略是分类的方法，即将文本对(T, H)表示成特征向量，然后利用机器学习方法进行分类，输出蕴涵或非蕴涵的判断结果。基于此，我们可以将语言现象识别结果作为向量的一维，加入到现有特征向量中参与训练。然而，这一方法存在以下问题：1）语言现象识别结果仅占特征向量的一维，比重过小；2）语言现象识别的错误可能会造成错误传播，影响整体推理判断的性能。

基于此，本文提出一种语言现象识别与整体推理判断的联合分类模型，其目的是用一个统一的模型解决两个高度相关的任务，能够在一定程度上避免上述问题。模型的输入为文本对(T, H)，输出为蕴涵或不蕴涵的判断，以及文本对中存在的语言现象。

本文采用随机森林（Random Forest, RF）作为联合分类器，理由如下：

1）RF 适合处理特征较多的问题。语言现象识别需要利用 42 种特征，蕴涵判断需要用到 20 种特征，尽管通用特征既可用于识别语言现象，也可用于进行推理判断，但总体特征数仍较多。而 RF 能够处理高维数据，不用进行特征选择，因此适合本任务。

2）RF 适合处理输出较多的任务。本模型的输出为语言现象类别（22 种）和蕴涵判断结果（蕴涵/非蕴涵），共有 44 种组合，远多于一般分类问题的类别个数。对于一般文本蕴涵识别而言，只需获得最终蕴涵判断结果即可；本文定义组合类别的目的在于获得语言现象的识别结果并进行分析，同时该结果也可与其他文本蕴涵识别模型结合以改进蕴涵识别性能，或对其他文本蕴涵识别系统进行评估。

3）RF 对于分布不均衡的数据能够保持稳定的性能。从 RITE-3 的语料统计[9]上看，在训练集中出现较多的语言现象，如 Inference 出现次数多达 75 次，而 Meronymy 语言现象则仅出现 4 次，存在明显的样本偏置。

另一方面，RF 泛化能力的一个决定因素是随机树的平均相关度，相关度越低则泛化能力越强。我们可以通过特征选择提高树之间的差异性，以此改进 RF 的分类性能。对于本问题而言，树之间的差异性体现在语言现象的识别，即专用特征；而通用特征主要分析 T 和 H 的相关程度，不同蕴涵现象的文本对可能体现出现相同的相关程度，若某些建树过程都使用了通用特征而未使用专用特征，可能导致生成的树的差异程度过小。因此，有必要在建树时分配一定数量的专用特征和通用特征。为此，本文提出一种改进特征选择的随机森林生成算法，算法描述如图 1 所示。

输入：训练集 D，随机森林规模 K，专用特征集 F_S，通用特征集 F_G，决策树的特征个数 n

输出：随机森林 RF

算法：

1.采用 Bootstrap 抽样，从训练集中有放回地抽取 K 个训练子集；

2. for i = 1 to K

a) 从 F_S 中随机选择 m 个特征，从 F_G 中随机选择 n-m 个特征，组合形成新的特征子集 f，$n<<|F_S|+|F_G|$；

b)利用第 i 个训练子集和特征子集 f 建立分类树，并加入 RF。

图 1 随机森林生成算法

在预测阶段，由 K 个决策树分别对测试数据进行投票，计算所有投票数，找出票数最高的类别即可得到测试数据的蕴涵关系及包含的语言现象。

4 实验结果及分析

4.1 数据准备

实验采用 RITE-3 中文任务的训练和测试语料，包括 581 对训练数据和 1200 对测试数据。每条数据包括一个语段 T 和一个假设 H，并标注了一个语言现象和整体蕴涵关系（蕴涵/非蕴涵）。其中，训练集包含 370 对具有蕴涵关系的文本对，211 对具有非蕴涵关系的文本对；测试集分别包含 600 对蕴涵关系与非蕴涵关系的文本对。为方便处理，首先对数据进行以下规范化操作：

1）将文本中的中英文标点符号统一替换成中文标点符号；

2）统一度量单位，如长度为米，重量为千克；

3）将汉字大写数字转换为阿拉伯数字；

4）将全角字符转换为半角字符；

5）将分数统一转换为汉语表示，如“X 分之 X”；

6）将日期统一转换为 XXXX 年 XX 月 XX 日格式。

4.2 实验结果

本实验评估了本文提出的随机森林方法对语言现象和整体蕴涵关系的识别性能。实验评估手段为准确率（Precision）、召回率（Recall）和 F1 值。

实验设置了四个系统，第一个系统（svm_combined）直接利用专用特征和通用特征建立特征空间，并利用 SVM 进行学习和预测；第二个系统（svm_cascaded）采用两阶段识别方法，首先利用专用特征进行语言现象识别，再将识别结果作为特征，和通用特征一起建立特征空间（实验中提高了识别特征的权重），利用 SVM 进行训练和预测；第三个系统（RF-FS）采用基于随机森林的联合分类模型，但树的构建采用完全随机特征选择的方法；第四个系统（RF+FS）在第三个系统基础上采用改进的随机森林生成算法，即本文方法。基准系统（baseline）采用我们在 NTCIR-11 上的参赛系统[13]。该系统采用分类方法，利用字串、相似度、结构和语言学共 15 种特征构建基于 SVM 的分类系统。实验结果如表 2 所示。

实现结果表明：

1）识别语言现象能够有效提高文本蕴涵识别系统的性能。从文本方法与基准系统的性能对比上看，蕴涵关系识别的准确率、召回率和 F1 值分别高出 13.89%、2.5%和 9.2%，非蕴涵关系识别的三个指标分别高出 6.06%、8.33%和 7.74%，显示出本文方法的性能显著优于基准系统；从 svm_cascaded 和基准系统的性能对比上看，准确率和 F1 值在蕴涵类关系识

别上分别提高 3.42%和 2%，在非蕴涵类关系识别上分别提高 1.53%和 3.46%，说明仅加入语言现象识别结果，也能在一定程度上改进蕴涵识别系统的性能。

表 2 文本蕴涵识别结果

	Entailment			Non-entailment		
	P	R	F1	P	R	F1
svm_combined	0.5588	0.6833	0.6148	0.5794	0.4183	0.4858
svm_cascaded	0.5692	0.6967	0.6265	0.5816	0.435	0.4977
RF-FS	0.6514	0.7117	0.6802	0.6108	0.4533	0.5204
RF+FS	**0.6739**	**0.725**	**0.6985**	**0.6269**	**0.475**	**0.5405**
baseline	0.535	0.7	0.6065	0.5663	0.3917	0.4631

2）在随机森林的建树过程中进行特征选择，能够提高模型的泛化能力，从而改进蕴涵识别的性能。对比 RF+FS 与 RF-FS 的实验结果，在准确率、召回率和 F1 值三个指标上，蕴涵关系识别分别高出 2.25%、1.33%和 1.83%，非蕴涵关系识别分别高出 1.61%、2.17%和 2.01%，表明模型的分类性能在经过特征选择后有了一定程度的提高。事实上，语言现象识别和整体推理判断属于相互关联的两个问题，因此所建的分类树要能对两个问题进行判断，采用特征选择方法则对分类树特征集合中的专用特征和通用特征进行了一定比例的分配，避免了分类树特征类别单一的问题。

3）与 SVM 相比，随机森林能够更有效地处理语言现象识别和整体推理判断的联合分类问题。对比 RF-FS 与 svm_cascaded 的实验结果，在准确率、召回率和 F1 值三个指标上，蕴涵关系识别分别高出 8.22%、1.5%和 5.37%，非蕴涵关系识别分别高出 2.92%、1.83%和 2.27%，说明随机森林能够更有效地处理多特征、多类别的分类问题；另一方面，与随机森林的蕴涵类识别准确率比较，SVM 的准确率过低，表明很多数据都被错误地识别为蕴涵类，其中的大部分原因是由于数据不均衡导致的。这也表明，随机森林方法具有更稳定的性能。

此外，从实验结果上看，svm_combined 的性能不如 svm_cascaded，其原因在于，尽管 svm_combined 使用了更多的特征，但由于数据集中每个文本对只包含一种语言现象，因此这些特征具有排斥性，导致数据稀疏，从而影响分类性能。

我们还对本文定义的 22 类语言现象识别结果进行了统计。统计数据来自 RF+FS 与 RF-FS 的语言现象识别结果。此外，我们还建立了一个基于 SVM 的分类系统，用于识别语言现象。该系统使用专用特征和通用特征进行训练和预测，输出为语言现象类别。实验评估指标为 F1 以及 Marco-F1 值[9]。实验结果如表 3 所示。

实验结果表明：

1）对于语言现象识别而言，随机森林的性能要优于 SVM。从总体性能上看，RF-FS 的 Macro-F1 比 SVM 方法高 3.89%，而 RF+FS 比 SVM 方法高 4.75%。从具体的语言现象上看，对 RF-RS 和 RF+FS 的大部分语言现象的 F1 值均高于 SVM 方法。

2）相对于 SVM，随机森林方法能够显著提高部分语言现象识别性能。对比 RF+FS 与 SVM 方法，前者识别 Lexical_entailment、Modifier、Antonym 等语言现象的 F1 值均高于后者 10%以上。其原因在于，语言现象识别与整体推理判断具有一定的关联性，例如一个矛盾类现象出现在整体为蕴涵关系的文本对中的可能性较低。而随机森林方法为联合分类方法，两个任务在训练中相互影响，有助于各自识别性能的改进。本实验中的 SVM 方法则未将整体推理关系用于识别。

表 3 语言现象识别结果

	RF+FS	RF-FS	SVM+F_S+F_G
Abbreviation	0.7203	0.6955	0.6801
Apposition	0.7197	0.7197	0.6398
Case_alternation	0.3704	0.3704	0.1852
Clause	0.7579	0.7368	0.6737
Coreference	0.4583	0.4583	0.4167
Hypernymy	0.6296	0.6667	0.6667
Lexical_entailment	0.5517	0.5172	0.4483
List	0.5405	0.4865	0.4324
Meronymy	0.6087	0.6087	0.6522
Modifier	0.7252	0.7023	0.5649
Quantity	0.6552	0.6552	0.6207
Scrambling	0.6571	0.6571	0.6857
Spatial	0.6429	0.6429	0.6190
Synonymy	0.7059	0.6863	0.6471
Temporal	0.6614	0.6614	0.6502
Transparent_head	0.6538	0.6538	0.7308
Antonym	0.5377	0.5189	0.4151
Exclusion:predicate _argument	0.5263	0.5263	0.4474
Exclusion:quantity	0.6552	0.6552	0.6207
Exclusion:spatial	0.5625	0.5313	0.5625
Exclusion:temporal	0.6471	0.6471	0.6176
Negation	0.4286	0.4286	0.3929
Macro-F1	0.6098	0.6012	0.5623

3）某些语言现象比较复杂，识别这类现象需要用到更多知识，系统识别性能也有待提高。例如，在 RF+FS 系统上，Case_alternation 现象的 F1 值仅有 37.04%，其原因在于语言形式变化多样，仅通过定义一些匹配模板难以得到准确的包含句式转换的文本片断。又如，Antonym 现象的 F1 值较低的原因之一是许多对义关系并未识别出来，其原因在于本实验中仅采用了 HowNet 以及一些汉语词典作为反义词资源，知识非常有限。

5 结论

本文提出一种基于语言现象的文本蕴涵识别方法。该方法建立了一种语言现象识别和整体推理判断的联合分类模型，并利用改进的随机森林方法进行训练和预测。为识别语言现象，本文设计了 22 类专用特征和 20 类通用特征；为提高随机森林的泛化能力，本文提出一种基于特征选择的随机森林生成算法，通过在建树时分配一定数量的专用特征和通用特征，以增加生成的树的差异度。实验结果表明，识别语言现象能够有效提高文本蕴涵识别系统的性能；同时，在随机森林的建树过程中进行特征选择，能够提高模型的泛化能力，从而改进语言现象识别和整体推理判断的性能。

参考文献

[1] Dagan I. and Glickman O. Probabilistic Textual Entailment: Generic Applied Modeling of Language Variability[C]//In proceedings of PASCAL Workshop on Learning Methods for Text Understanding and Mining. 2004.

[2] Androutsopoulos I. and Malakasiotis P. A Survey of Paraphrasing and Textul Entailment Methods[J]. Journal of Artificial Intelligence Research, 2010, 38(1): 135-187.

[3] Dagan I. and Dolan B. Recognizing textual entailment: Rational, evaluation and approaches[J]. Natural Language Engineering, 2009, 15(4): i-xvii.

[4] Magnini B. and Cabrio E. Combining Specialized Entailment Engines[C]//In Proceedings of LTC'09. 2009.

[5] Bentivogli L., Cabrio E., Dagan I., et al. Building textual entailment specialized data sets: a methodology for isolating linguistic phenomena relevant to inference[C]//Proceedings of the International Conference on Language Resources and Evaluation. 2010.

[6] Kaneko K., Miyao Y. and Bekki D. Building Japanese Textual Entailment Specialized Data Sets for Inference of Basic Sentence Relations[C]//In proceedings of the 51st Annual Meeting of the Association of Computational Linguistics 2013.

[7] Sammons M., Vydiswaran V. G. V. and Roth D. "Ask not what Textual Entailment can do for you..."[C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2010.

[8] Garoufi K. Towards a better understanding of applied textual entailment: Annotation and evaluation of the RTE-2 dataset. Germany, Saarland University. Master Thesis. 2007.

[9] Matsuyoshi S., Miyao Y., Shibata T., et al. Overview of the NTCIR-11 Recognizing Inference in TExt and Validation (RITE-VAL) Task[C]//In Proceedings of the 11th NTCIR Conference. 2014.

[10] Huang H.-H., Chang K.-C. and Chen H.-H. Modeling Human Inference Process for Textual Entailment Recognition[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. 2013.

[11] 江敏, 肖诗斌, 王弘蔚, 等. 一种改进的基于知网的词语语义相似度计算[J]. 中文信息学报, 2008, 22(5).

[12] 张志昌, 周慧霞, 姚东任, 等. 基于词向量的中文词汇蕴涵关系识别[J]. 计算机工程, 2016, 42(2).

[13] Ren H., Wu H., Tan X., et al. The WHUTE System in NTCIR-11 RITE Task[C]//In Proceedings of the 11th NTCIR Conference. 2014.

任函（1980—），男，博士，助理研究员，主要研究领域为自然语言处理。Email: hanren@whu.edu.cn；

冯文贺（1976—），博士，讲师，主要研究领域为理论语言学、计算语言学，本文通讯作者。Email: wenhefeng@gmail.com。

刘茂福（1977—），男，博士，教授，主要研究领域为自然语言处理。Email: liumaofu@wust.edu.cn。

万菁（1981—），女，硕士，主要研究领域为理论语言学。Email: jingwan@whu.edu.cn。