文章编号：1003-0077(2013)05-0137-07

基于句法结构约束的模糊限制信息范围检测

周惠巍，杨 欢，黄德根，李 瑶，李丽双

（大连理工大学 计算机科学与技术学院，辽宁 大连 116024）

摘 要：模糊限制信息检测用于区分模糊限制信息与事实信息，提高抽取信息的真实性和可靠性。模糊限制信息范围的界定具有依赖于语义和句法结构的特点，是模糊限制信息检测的一个难点。该文提出一种基于句法结构约束的模糊限制信息范围检测方法，基于依存结构树和短语结构树构建决策树，获取句法结构约束集，用于产生句法结构约束特征，并加入到条件随机域模型中进行模糊限制信息范围检测。实验采用 CoNLL-2010 共享任务数据集，在标准的模糊限制语标注语料上，获得了 70.28% 的 F 值，比采用普通的句法结构特征提高了 4.22%。

关键词：模糊限制信息范围检测；句法结构约束；决策树；条件随机域

中图分类号：TP391 **文献标识码**：A

Hedge Scope Detection Based on Syntactic Structural Constraints

ZHOU Huiwei, YANG Huan, HUANG Degen, LI Yao, LI Lishuang

(School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning 116024, China)

Abstract: Hedge scope detection is used to distinguish factual information and uncertain information, which could improve the authenticity and reliability in information extraction. Hedge scope detection is a difficult task because of its dependency of the semantic and syntactic structures. In this paper, we propose a hedge scope detection method based on syntactic structural constraints. First, two decision trees are constructed on dependency structure and phrase structure respectively to build the syntactic constraint set. And then the hedge scope detection results based on the syntactic constraint set are used as the syntactic constraint features for Conditional Random Fields (CRF) models. Experiments on the CoNLL-2010 corpus achieve the 70.28% F-score on the golden standard hedge cues, which is 4.22% higher than the system with the common syntactic construction features.

Key words: hedge scope detection; syntactic structural constraints; decision tree; conditional random fields

1 引言

模糊限制语最早是由 G. Lakoff 提出的，用来指“把一些事情弄得模模糊糊的词语”，表示的是不确定性和临时性的观点[1]。由模糊限制语所引导的信息为模糊限制性信息(Hedge Information)。

统计表明，在线生物医学文献数据库 MEDLINE 的论文摘要中，11%的句子包含模糊限制信息[2]；在用于模糊限制信息检测研究的 BioScope 语料库[3] 中，正文中 22.29% 的句子，和摘要中 17.69%的句子包含模糊限制信息。在 Medlock 和 Briscoe 标注的语料中，32.41%的基因名出现在模糊性的句子中[4]。在生物医学领域进行模糊限制信息检测，能提高抽取信息的可靠性和真实性。

近年来，模糊限制信息检测引起了国内外研究人员的广泛关注，国际计算语言学协会将模糊限制性句子识别和模糊限制信息范围检测定为 2010 年 CoNLL (Conference on Computational Natural Language Learning) 共享任务[5]。共享任务包含生物医学和维基百科两个领域，其中生物医学领域训练语料源自 BioScope 语料库[3]。BioScope 语料库

收稿日期：2013-06-22 定稿日期：2013-08-21

基金项目：国家自然科学基金资助项目(61272375；61173100；61173101)

作者简介：周惠巍(1969—)，女，副教授，主要研究方向为句法分析、生物医学信息处理等；杨欢(1988—)，女，硕士研究生，主要研究方向为生物医学信息处理；黄德根(1964—)，男，教授，主要研究方向为机器翻译、跨语言信息检索等。

对模糊限制语及其范围进行了标注，如例句(1)，模糊限制语“appear”的模糊限制范围为“the Ras/Raf/ERK pathway did not appear to mediate the effect of the antioxidant”。基于BioScope的模糊限制性句子识别研究已经取得了一定的进展，由模糊限制语所引导的模糊限制信息范围检测仍然是一个难点。目前模糊限制信息范围检测方法主要有基于规则的方法、基于统计的方法和基于规则和统计相结合的方法。

例句(1) Transfection of trans-dominant negative expression vectors of ras and raf, together with AP-1-dependent reporter constructs, as well as Western blot analysis using anti-ERK (extracellular signal-regulated kinase) antibodies, indicated that <*xcope*>the Ras/Raf/ERK pathway did not <*cue*>appear</*cue*> to mediate the effect of the antioxidant</xcope> .

模糊限制信息范围是由模糊限制语引导的具有一定语义的连续字符串，往往是句法结构上与模糊限制语相关的一个短语或一个从句。因此，基于规则和基于统计的方法都利用了句法结构树。基于规则的方法是根据模糊限制语的词性及句子的短语结构或依存结构制定模糊限制信息范围检测规则[6-7]。基于统计的方法往往将句法结构信息平面化，用于模糊限制信息范围检测[8-9]。Morante等[8]基于依存结构特征确定模糊限制信息范围，在CoNLL-2010测试集上取得了57.32%的F值，获得了模糊限制信息范围评测的第一名。为减少结构信息平面化产生的数据稀疏问题，Zhou等[10]研究了模糊限制信息范围的结构化表达方法，定义了最短路径包含树等多种短语结构特征，并与平面特征相结合，采用树核和多项式核的复合核方法，在CoNLL-2010测试集上取得57.47%的F值。为实现规则方法和统计方法的优势互补，Rei和Briscoe[11]提出了基于规则和统计相结合的方法，将基于规则的检测结果作为特征引入基于统计的检测模型，取得了55.6%的F值，在模糊限制信息范围检测任务中获得了第二名。Velldal等[12]分别构建了基于依存结构的规则子系统和基于短语结构的统计子系统，并通过融合两个子系统，实现规则方法和统计方法的结合，在标准的模糊限制语标注语料上获得69.60%的F值。基于句法结构规则的模糊限制信息范围检测系统缺乏灵活性，而基于句法结构特征的统计系统不但需要进行繁琐而艰苦的特征选择，而且容易产生数据稀疏问题，难以挖掘有效的句法结构信息。

本文提出一种基于句法结构约束的模糊限制信息范围检测方法。首先利用依存结构和短语结构构建决策树模型，获取句法结构约束集，然后基于句法结构约束集产生句法结构约束特征，用于模糊限制信息范围检测。本文中使用的句法结构约束集是由决策树算法自动产生的，比人工制定规则更具灵活性和准确性，有效地将依存结构和短语结构信息用于模糊限制范围检测，提高了模糊限制信息范围检测性能。

2 模糊限制信息范围检测系统

2.1 系统概述

模糊限制信息范围检测问题可以转化为序列标注问题，即F-scope表示模糊限制信息序列的第一个词，L-scope表示模糊限制信息序列的最后一个词，而NONE表示其他词。

基于句法结构约束的检测模型训练过程如图1

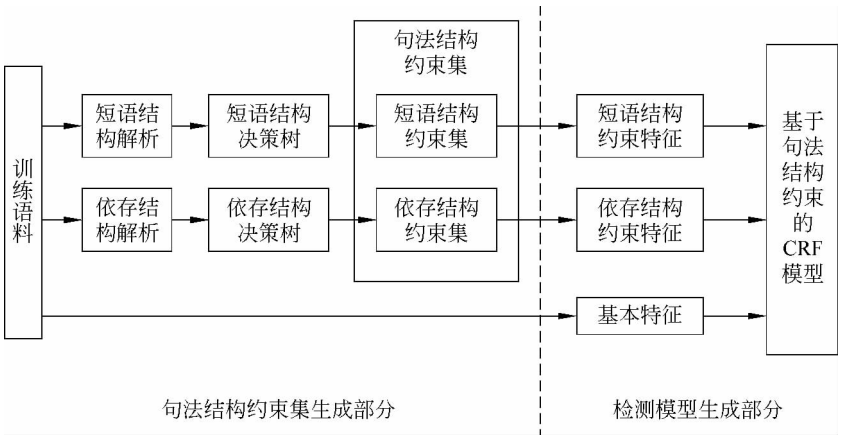

图1 句法结构约束检测模型的训练过程

所示，包括句法结构约束集生成和模糊限制信息范围检测模型生成两个部分。

句法结构约束集生成部分基于训练语料的短语结构和依存结构，利用决策树算法分别训练获得两个决策树，产生基于短语结构和依存结构的句法约束集。

模糊限制信息范围检测模型生成部分基于句法结构约束集获得模糊限制信息范围检测结果，作为句法约束特征，与基本特征共同训练获得基于句法结构约束的 CRF 检测系统。

基于句法约束的模糊限制信息范围检测过程与训练过程相似，只是无需训练获得决策树，直接基于句法约束集即可获得句法约束特征，用于模糊限制信息范围检测。

2.2　基本特征

有的句子包含多个模糊限制语，每个模糊限制语对应一段模糊限制范围，一段模糊限制信息可能是另一段模糊限制信息的子串，模糊限制信息范围存在嵌套关系。为明确界定每个模糊限制语的限制范围，我们把句子按照模糊限制语的数量进行复制，使得每个句子有且只有一个模糊限制语（包含单个词模糊限制语和多词短语模糊限制语），再进行训练和检测。

CRF 模型是一种判别式的序列标注模型，在模糊限制信息范围检测任务中得到广泛使用，本文采用 CRF 模型标注模糊限制信息范围，选取的基本特征包括：

- 单词特征：$word(i)(i=-3,-2,-1,0,+1,+2,+3)$
- 词干特征：$stem(i)(i=-3,-2,-1,0,+1,+2,+3)$
- 词性特征：$pos(i)(i=-3,-2,-1,0,+1,+2,+3)$
- 组块特征：$chunk(i)(i=-3,-2,-1,0,+1,+2,+3)$
- 模糊限制语特征：$hedge(i)(i=-3,-2,-1,0,+1,+2,+3)$，当前句子的模糊限制语作为模糊限制信息范围检测的重要特征，模糊限制语采用 IOB2 标注模式。
- 模糊限制语词干链特征：$hedgeStem(i)(i=-3,-2,-1,0,+1,+2,+3)$
- 模糊限制语词性链特征：$hedgePos(i)(i=-3,-2,-1,0,+1,+2,+3)$
- 当前词与模糊限制语的距离：$DH(i)(i=-3,-2,-1,0,+1,+2,+3)$，从当前词到模糊限制语的单词个数。

为比较句法结构约束与常用的句法特征的区别，实验中我们分别引入依存结构和短语结构两种句法结构特征如下：

- 依存标记特征：$dependencyRel(i)(i=-3,-2,-1,0,+1,+2,+3)$
- 短语路径特征：$hedgePath(i)(i=-2,-1,0,+1,+2)$，当前词到模糊限制语的短语路径。

2.3　句法结构约束特征

决策树方法[13]能够从训练实例中归纳出一组树形结构表示的分类规则，分类时基于树形分类规则从根节点逐步对样本属性进行测试，沿着相应的分支向下走，直至某个叶子节点，该叶子节点即为样本类型。决策树方法广泛用于自然语言处理任务，并取得了较好的分类效果[14-15]，属性选择是利用决策树算法进行分类的关键。

模糊限制语的限制范围与模糊限制语本身具有很大关系，因此我们选取了模糊限制语属性。除此之外，分别选取短语树属性和依存结构树属性构建短语结构决策树和依存结构决策树。以例句(1)为例，分别介绍各类属性。假设当前词为“mediate”，Y 表示“是”，N 表示“不是”，L 表示“在左边界上”、R 表示“在右边界上”、I 表示“不在边界上，但在边界内”、O 表示“在边界外”。

(1) 模糊限制语属性

- 模糊限制语是否是单个词：此例为“Y”。
- 模糊限制语是单个词时的词性：此例为“VB”。
- 模糊限制语是多词时首词的词性：此例为“NULL”。
- 模糊限制语是多词时尾词的词性：此例为“NULL”。

(2) 短语结构属性

例句(1)的短语结构树片段如图 2 所示，其中“appear”是模糊限制语，“the”是模糊限制信息范围的左边界，“antioxidant”是模糊限制信息范围的右边界。

- 当前词是否是模糊限制语的首词：此例为“N”。
- 当前词是否在短语句法成分的边界上：“mediate”在“VP_{81}”短语的左边界上，此例为“L”。

• 模糊限制语父亲节点的短语类型：此例为“VP”。

• 当前词是否在以模糊限制语的父亲节点为根的子树内：此例为“Y”。

• 模糊限制语和当前词的最小包含树的短语类型：此例中最小包含树为“VP_{56}”，因此属性值为“VP”。

• 模糊限制语和当前词的最小包含树的根节点与其父亲节点是否具有相同的短语类型：此例中最小包含树的根节点“VP_{56}”与其父亲节点“VP_{42}”短语类型相同，因此属性值为“Y”。

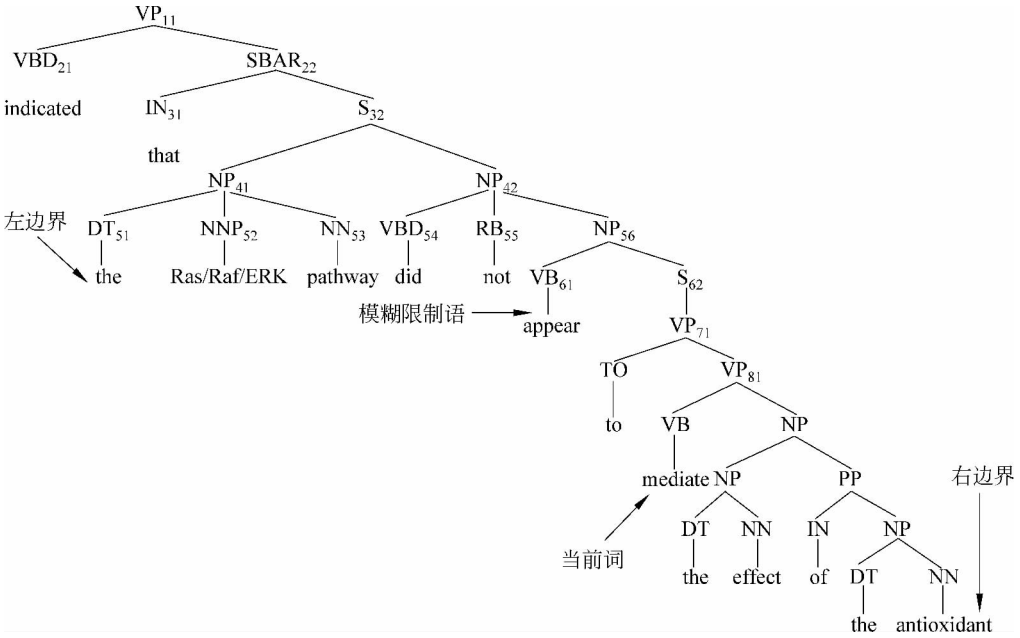

图2 例句(1)的短语结构片段

(3) 依存树属性

图3为例句(1)的依存树片段，W表示单词，R表示依存关系标记，“[]”中的为词性。

• 当前词是否是模糊限制语的首词：此例为“N”。

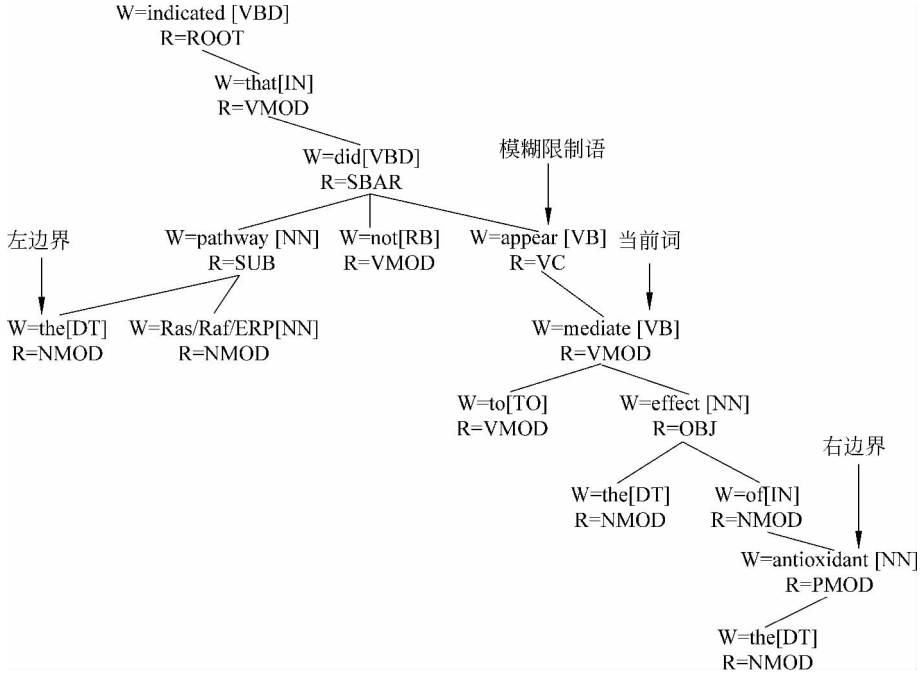

图3 例句(1)的依存树片段

• 模糊限制语的依存关系类型：此例为“VC”。

• 当前词是否在以模糊限制语为根的子树的边界上：此例为“I”。

• 模糊限制语是否是依存树的根节点：此例为“N”。

• 模糊限制语的父亲节点的依存关系标记：此例为“SBAR”。

• 当前词是否在以模糊限制语的父亲节点为根的子树的左右边界上：此例为“I”。

基于模糊限制语属性和短语结构属性训练获得短语结构决策树，基于模糊限制语属性和依存树属性训练获得依存结构决策树。选取短语结构决策树和依存结构树中准确率较高的决策规则，构建短语结构约束集和依存结构约束集。应用短语结构约束集和依存结构约束集检测句子的模糊限制信息范围，将其检测结果作为短语结构约束特征和依存结构约束特征用于 CRF 检测系统，提高系统检测性能。短语结构约束特征和依存结构约束特征分别为：

• 短语结构约束特征：$phraseCh(i)(i=-3,-2,-1,0,+1,+2,+3)$，特征值为基于短语结构约束集的规则检测结果。

• 依存结构约束特征：$dependencyCh(i)(i=-3,-2,-1,0,+1,+2,+3)$，特征值为基于依存结构约束集的规则检测结果。

3　实验结果与分析

3.1　实验设置

本文仅研究模糊限制信息范围检测任务，因此采用标准标注的模糊限制语评测模糊限制信息范围检测性能。实验采用 CoNLL-2010 共享任务 2 语料，训练语料包含有 3 327 个模糊限制语，存在于 2 620 个模糊性句子中；测试语料包含 1 033 个模糊限制语，存在于 790 个模糊限制性句子中。在预处理中，使用 GENIA Tagger[①] 工具包获得词干、词性和组块信息；分别使用依存关系解析器 GDep Parser[②] 和短语结构解析器 Berkeley Parser[③] 解析句子，获得依存结构树和短语结构树。采用 C4.5R8[④] 工具包来构建决策树，CRF 模型使用 CRF++-0.54[⑤] 工具包获得。

模糊限制信息范围是一段连续的字符串，然而，分类器的输出结果不能保证只识别出一个左边界和一个右边界。因此，分类器的输出必须经过处理才能得到完整的模糊限制信息范围。实验采用文献[10]的后处理算法。模糊限制信息范围采用 CoNLL-2010 共享任务组织者提供的评测工具进行评测，包括召回率、准确率和 F 值三个评价指标。因采用正确标注的模糊限制语进行测试，所以召回率等于准确率，也等于 F 值，在此仅表示 F 值。

3.2　实验结果及分析

实验时，我们首先建立一个基于基本特征的系统，称为基础系统；然后分别把依存标记特征和短语路径特征加入基础系统，得到普通的句法结构特征对系统性能的影响，如表 1 所示。由实验结果可以看出，短语路径特征和依存标记特征对系统检测性能均有提高，短语路径特征比依存标记特征更有效。

表 1　普通的句法结构特征对系统检测性能的影响

特　　征	F 值/%
基础系统	63.08
＋短语路径	65.87
＋依存标记	63.37
＋依存标记＋短语路径	66.06

选取短语结构决策树和依存结构树中准确率较高的决策规则，构建短语结构约束集和依存结构约束集。

由 C4.5 决策树生成的每条规则都有一个准确率来反映这个规则的准确程度，实验设定一个规则准确率阈值 p，选取具有一定准确率的决策规则，构建句法结构约束集，检测具有不同可靠性的句法约束规则对模糊限制信息检测性能的影响。

分别把依存结构约束特征和短语结构约束特征加入到基础系统中，得到句法结构约束特征对系统性能的影响，如表 2 所示。短语约束特征均取得了与短语路径特征相当的检测效果，当阈值 p 为 60% 和 70%时，检测性能高于短语路径特征。依存约束

① Available at http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

② Available at http://people.ict.usc.edu/~sagae/parser/gdep/

③ Available at http://code.google.com/p/berkeleyparser/

④ Available at http://www2.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tutorial.html

⑤ Available at http://crfpp.googlecode.com/svn/trunk/doc/index.html

特征对检测系统性能的提高均较明显，其对系统检测性能的提高远远高于依存标记特征。两种句法结构约束共同作用时，系统性能进一步提高。这说明模糊限制信息范围检测系统对短语结构和依存结构均具有依赖性，且对依存结构的依赖较强。最后将句法结构特征与约束特征都加入基本系统中，最高可达到70.28%的F值，比基础系统提高7.2%。阈值p在30%到85%的广泛范围内，基于句法结构约束方法的检测性能，均明显优于普通的基于句法结构特征的方法。因此，本文方法对于未知的数据，具有较强的鲁棒性。

表2 句法结构约束特征对系统检测性能的影响

特征 / 阈值 p/%	基本特征+短语约束特征	基本特征+依存约束特征	基本特征+依存约束特征+短语约束特征	基本特征+句法特征+约束特征
30	65.63	**67.18**	68.25	69.80
50	65.73	67.18	68.25	70.28
60	66.22	67.09	68.44	**70.28**
70	**66.51**	66.89	**68.64**	70.28
80	65.54	66.89	67.47	69.60
85	65.54	66.60	67.09	69.41

单纯基于句法约束规则的检测结果如表3所示，由检测结果可以看出，基于决策树方法获得的句法结构约束规则具有较好的检测效果，因此将其检测结果作为特征用于CRF检测模型取得了更好的检测性能。

表3 基于句法结构约束规则的检测结果

系统 / 阈值 p/%	短语规则	依存规则	依存规则+短语规则
30	58.86	62.79	61.57
50	59.44	62.79	62.34
60	60.02	62.89	63.50
70	60.21	**63.08**	**63.89**
80	**60.31**	62.21	63.70
85	60.02	60.77	62.25

表4为本文方法的检测结果与Rei和Briscoe[11]、Velldal等[12]方法的对比。Rei和Briscoe的系统是CoNLL-2010共享评测排名第二的系统，排名第一的系统没有公布在标准的模糊限制语标注语料上的模糊限制信息范围检测结果。

表4 与其他系统的比较

系统	Ours	Rei和Briscoe[11]	Velldal等[12]
F值/%	70.28	66.3	69.60

本文系统F值比Rei和Briscoe[11]高3.98%，比Velldal等[12]结果略高。本文通过构建决策树自动产生约束规则，理论上比Velldal等[12]人工制定的规则更具有适应性。

4 结论

本文提出了一种基于句法结构约束的模糊限制信息范围检测方法。采用决策树算法分别学习获得短语结构决策树和依存结构决策树，选取具有一定精确度的决策规则构建句法结构约束集，用于产生句法结构约束特征，并加入到CRF模型中进行模糊限制信息范围检测。相比于传统的基于句法结构特征的检测模型，本文方法利用决策树模型学习获得有效的模糊限制信息范围检测规则，辅助模糊限制信息范围检测，有效地提高了系统检测性能。应用决策树算法构建检测规则时，仅需进行属性选择，相对于人工制定规则的方法，处理相对简单，并具有较强的适应性和鲁棒性。本文仅研究了基于不同准确率的决策规则对检测性能的影响，如何挖掘决策规则间的相互关系，分别将不同的决策规则用于句法结构特征的生成和对分类器输出进行后续处理，将是本文下一步主要研究工作之一。

参考文献

[1] George L. Hedges: a study in meaning criteria and the logic of fuzzy concepts [J]. Journal of Philosophical

Logic，1973，2(4)：458-508.
[2] Marc L，Qiu X Y，Pandmini S. The language of bioscience：facts，speculations，and statements in between[C]//Proceedings of the BioLINK，Boston，2004，17-24.
[3] Szarvas G，Vincze V，Farkas R，et al. The BioScope corpus：biomedical texts annotated for uncertainty，negation and their scopes[J]. BMC Bioinformatics，2008，9(11)：S9.
[4] Medlock B，Briscoe T. Weakly supervised learning for hedge classification in scientific iterature[C]//Proceedings of ACL，the 45th Annual Meeting of the Association of Computational Linguistics，2007，992-999.
[5] Farkas R，Vincze V，Móra G，et al. The CoNLL 2010 Shared Task：Learning to detect hedges and their scope in natural language text[C]//Proceedings of the CoNLL，Uppsala，Sweden，2010，1-12.
[6] Özgür A，Radev D R. Detecting speculations and their scopes in scientific text[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing，Singapore，August，Association for Computational Linguistics. 2009：1398-1407.
[7] Velldal E，Ovrelid L，Oepen S. Resolving speculation：MaxEnt cue classification and dependency-based scope rules[C]//Proceedings of the CoNLL，Uppsala，Sweden，2010，48-55.
[8] Morante R，Asch V V，Daelemans W. Memory-based resolution of In-Sentence scopes of hedge cues[C]//Proceedings of the CoNLL，Uppsala，Sweden，2010：40-47.
[9] Qiaoming Zhu，Junhui Li，Hongling Wang，et al. A unified framework for scope learning via simplified shallow semantic parsing[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing，2010：714-724.
[10] Huiwei Zhou，Degen Huang，Xiaoyan Li，et al. Combining Structured and Flat Features by a Composite Kernel to Detect Hedges Scope in Biological Texts[J]. Chinese Journal of Electronics，2011，20(3)：476-482.
[11] Rei M，Briscoe T. Combining manual rules and supervised learning for hedge cue and scope detection[C]//Proceddings of the CoNLL，Uppsala，Sweden，2010，56-63.
[12] Velldal E，Øvrelid L，Read J，et al. Speculation and Negation：Rules，Rankers，and the Role of Syntax[J]. Association for Computational Linguistics，2012，38(2)：369-410.
[13] Quinlan J R. C4.5：Programs for Machine Learning[M]. San Mateo，CA：Morgan Kaufman，1993.
[14] 刘玲玲，梁颖红，张永刚，等. 基于决策树的关键短语抽取[J]. 江南大学学报，2010，9(1)：71-74.
[15] 徐鹏，林森. 基于 C4.5 决策树的流量分类方法[J]. 软件学报，2009，20(10)：2692-2704.

(上接第 121 页)

[9] 闫泽华. 基于 LDA 的新闻线索抽取研究[D]. 上海交通大学硕士论文，2012.
[10] Teh Y W，Jordan M I，Beal M J，et al. Hierarchical dirichlet processes[J]. Journal of the American Statistical Association，2006，101(476).
[11] Blei D M，Lafferty J D. Visualizing topics with multi-word expressions[J]. arXiv preprint arXiv：0907.1013，2009.
[12] Wallach H M. Topic modeling：beyond bag-of-words[C]//Proceedings of the 23rd international conference on Machine learning. ACM，2006：977-984.
[13] Wang X，McCallum A，Wei X. Topical n-grams：Phrase and topic discovery，with an application to information retrieval[C]//Proceedings of Data Mining. ICDM 2007. Seventh IEEE International Conference on. IEEE，2007：697-702.
[14] Nallapati R，Feng A，Peng F，et al. Event threading within news topics[C]//Proceedings of the thirteenth ACM international conference on Information and knowledge management. ACM，2004：446-453.
[15] Lau J H，Newman D，Karimi S，et al. Best topic word selection for topic labelling[C]//Proceedings of the 23rd International Conference on Computational Linguistics：Posters. Association for Computational Linguistics，2010：605-613.
[16] Carmel D，Roitman H，Zwerdling N. Enhancing cluster labeling using wikipedia[C]//Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. ACM，2009：139-146.
[17] Song Y，Pan S，Liu S，et al. Topic and keyword re-ranking for LDA-based topic modeling[C]//Proceedings of the 18th ACM conference on Information and knowledge management. ACM，2009：1757-1760.