文章编号：1003-0077（2008）06-0000-00

句法与词义相结合的中文代词消解*

宋巍，秦兵，郎君，刘挺

（哈尔滨工业大学，计算机学院，信息检索研究室，哈尔滨 150001）

摘要：句法知识对代词消解有很大的帮助。近年来依存句法由于其利于描述语言中词与词之间的关系、突出核心词的特点日益得到重视。本文提出了一种中文第三人称代词消解方法，直接利用依存句法分析器的结果，构建有效的句法角色特征和名词短语的支配词之间的词义相似和词语相关特征，采用支持向量机作为分类器，在ACE2005语料上的实验证明了这些特征的有效性。

关键词：代词消解；依存句法；句法角色；词义相似

中图分类号：TP391　　**文献标识码：**A

Combining Syntax and Word Sense for Chinese Pronoun Resolution

Song Wei, Qin Bing, Lang Jun, Liu Ting

(Information Retrieval Laboratory, School of Computer Science and Technology,
Harbin Institute of Technology, Harbin 150001)

Abstract: Syntactic knowledge is important for pronoun resolution. In recent years, research on dependency parsing becomes active, because dependency grammar benefits to represent the relations between terms. We propose a dependency parsing based method for Chinese pronoun resolution, design effective syntactic role features and word sense similarity and word relevance features in related to the dependent words. Support Vector Machine is used as the classifier. The experimental result on the ACE 2005 training data shows that these dependency parsing based features are effective.

Key words: pronoun resolution; dependency parsing; syntactic role; word sense similarity

1 引言

指代是指当前的指示语与上文出现的短语（先行语）存在语义关联，指代消解的过程即是对当前指示语确定先行语的过程。指代消解是自然语言理解与处理领域的核心问题之一，在信息抽取、机器翻译等应用中，都发挥重要作用[1]。

早期的指代消解算法基于语言学知识，以Hobbs算法[2]和中心理论[3]为代表。近年来，研究者们尝试使用机器学习方法来解决。基于机器学习的指代消解方法一般可以分为两类：有指导方法和无指导方法。前者的主要思想是将指代消解问题视为二元分类问题，首先利用标注有指代关系的训练数据训练一个分类器，而后利用这个分类器判断两个名词短语是否具有指代关系。应用于指代消解的有指导的机器学习方法有贝叶斯[4]、决策树[5]和最大熵[6]等。基于无指导方法的指代消解算法研究相对较少。Cardie等提出一种基于聚类的名词短语共指消解方法[7]，采用特征向量来表示各个名词短语，然后用聚类算法来实现名词短语的共指消解。

中文指代消解的研究相对较少[8,9]。与英文相比，中文浅层词汇处理难度更大，主要体现在句子需要分词，名词短语没有明确的性别、单复数特征，代词没有明确的主格和宾格特征，省略非常普遍等方面。这些难点都给中文指代消解的特征提取带来了很大困难。

* **收稿日期：**　　**定稿日期：**

基金项目：本文受到国家自然科学基金(60575042, 60503072)和863项目(2006AA01Z145)资助

作者简介：宋巍（1983-），男，博士研究生，主要研究方向为自然语言处理、信息检索；秦兵（1968-），女，副教授，主要研究方向为自然语言处理；郎君（1981-），男，博士研究生，主要研究方向为自然语言处理；刘挺（1972-），男，教授，博士生导师，主要研究方向为自然语言处理、信息检索。

2 相关工作

句法分析一直是研究者依靠的“武器”之一。Hobbs提出了两种指代消解的算法：一种是简单Hobbs算法，通过自左向右先广搜索，层次遍历句法树来消解代词，另外一种在句法知识基础上加入语义约束。1994 年，Lappin和Leass提出了句法与约束规则相结合的方法[10]，首先使用槽文法分析器分析句子结构，继而通过约束规则过滤掉不满足条件的候选先行语，最后计算候选先行语权值来评判其作为先行语的可能程度。Xiaofeng Yang提出基于Tree-Kernel的方法[11]，将句法分析树结构作为特征，利用Convolution Tree Kernel[12]计算两棵句法树之间的相似程度，取得了很好的效果。

近年来，依存语法和依存句法分析[13]越来越受到关注。依存语法建立起句子中词和词的“依存”关系，每一个关系将上下两项联系起来，上项称为支配词，下项称为所属词，其主要目的在于描述与揭示构成语言的元素与元素之间的关系，因此可能更适于语言结构的描述与更深层次的分析。本文利用自动生成的依存句法分析器的结果，构建句法角色特征，利用Hownet[14]计算指代词与候选先行语的支配词的词义相似度和词汇相关性作为特征，采用支持向量机作为学习算法在ACE2005 的中文训练语料上，针对第三人称代词进行了实验。

后续内容组织如下，第 3 章将详细介绍本文指代消解系统框架和基于依存句法分析的特征构建。第 4 章介绍实验的设计与结果的分析。最后对全文内容做出总结与展望。

3 依存句法与词义结合的代词消解方法

3.1 系统框架

本文采用支持向量机作为分类器，利用自动获得的依存句法分析器的结果构建特征。系统的框架如图 1 所示。对训练语料中每一个待消解的代词首先确定候选先行语的范围，而后进行一致性约束过滤淘汰掉不符合一致性约束的候选得到最终的候选集合，继而构建训练实例训练得到一个分类模型。测试时对每一个待消解的代词确定其候选先行语的集合，每个候选先行语与代词构成一个分类实例，利用训练得到的模型进行分类，确定最终的先行词。

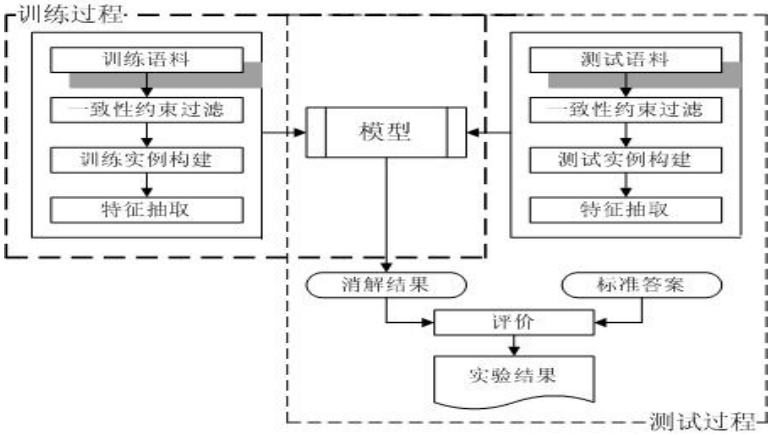

图 1 系统框架

3.2 一致性约束过滤和训练、测试实例构建

代词和其先行语均代表现实中的同一实体。据此，不符合性别、单复数和名词短语的类别一致性约束的名词短语之间是不能互指的。例如：“她”不能指向布什先生，因为不符合性别一致性；“他”不能指向“同学们”，因为不符合单复数一致性；同样，“他”不能指向“摩天大楼”，因为前者指人，后者指建筑，两者不符合类别一致性。在一致性约束过滤阶段，将过滤掉不符合一致性约束的候选，缩减候选先行语的规模。这需要对名词短语的性别、单复数和类别进行识别。单复数和类别的识别依靠标注语料提供的信息，将在实验设计部分阐述。

性别识别将名词短语的性别表示为 Male，Female，Unknown 之一。对人名和非人名名词短语采取不同策略。对于人名，利用尾字作为性别指示词，按照它们在男名和女名中出现的比例作

为判断标准。对出现在男名或者女名中的比例大于一定标准（实验中设为 0.75）的尾字对应的人名标注为相应的性别，其余标注为 Unknown。对于非人名名词短语，借助 Hownet 提供的语义信息识别。Hownet 中有词语的性别指示信息，“male|男”表示男性指示词，“female|女”表示女性指示词，“#male|男”表示男性相关指示词，“#female|女”表示女性相关指示词。在 Hownet 中查询名词短语的核心词（Head），如果核心词的知识元定义中含有包含性别指示的义项则将其置为相对应的性别。若 Hownet 不包含核心词或对应的义项中不包含性别指示信息则置为 Unknown。最终对于两个名词短语 M1 和 M2，如果具有明确（Male 或 Female）但不相同的性别信息，如：M1 为 Male，M2 为 Female，则认为二者不具有性别一致性。

构建训练实例时，对于每一个代词将在其之前出现的三句（包括代词所在句）以内的，与该代词属于同一实体的名词短语与该代词作为一个训练正例，对于不属于同一实体的名词短语进行一致性约束过滤，余下的候选名词短语分别与该代词构成一个训练反例。构建测试实例时，将代词前三句内的符合一致性约束的名词短语作为候选，与代词构成一个分类实例。

3.3 特征选择

通常的指代消解算法构造特征时，只考虑代词和候选先行语两者本身的属性或者两者之间的联系，如：词汇距离、词语类别、是否具有同位或并列关系等，即使语义特征，也仅仅考虑两个名词短语之间的语义相关性，而没有利用上下文中其他词语的语义特征。这种策略下有一些指代现象难以区分，例如：

1. *李老师病倒了，小明说他要去看一看。*
2. *李老师病倒了，小明说他的病是累出来的。*

对句中的“他”进行消解。显然，句 1 中的“他”应当指代“小明”，句 2 中的“他”指代李老师。然而，传统的指代消解系统对这两句的处理一般是没有区别的，很可能给出相同的结果。观察句 2 中的代词“他”和候选先行语“李老师”的支配词都为“病”，给予很大的指示性：“他”指向“李老师”的可能性将更大一些。又例如：

“我行使权力加重对王某的刑罚，判他坐牢 7 年和鞭打 12 下。”

其中“王某”和“他”分别依存于“刑罚”和“坐牢”，这两个词具有很大的相关性。这给我们一个启示：代词与候选先行语所在的上下文内，与它们关系密切的词语之间的语义关联性可以帮助确定复杂情形下的指代关系。依存句法不仅本身可以提供丰富的句法信息而且这种词和词之间的依存关系了提供一种途径来确定与代词和候选先行语具有密切关系的词语。这即是我们将句法与支配词词义结合的初衷。以下，令P为待消解的代词，A为候选的先行语，D_P为P依存的支配词，D_A为A依存的支配词，这里我们限定支配词只是动词或名词。根据依存分析的特点设计了以下特征。

3.3.1 支配词的词义特征

（1） 支配词词义相似：我们寄希望于支配词的词义相似性可以帮助代词消解，利用 Hownet 提供的 API：HowNet_Get_Concept_Similarity，来计算候选先行语的支配词与代词的支配词之间的概念相似度。HowNet_Get_Concept_Similarity 可以给出两个概念（即义项）之间的相似度，计算方法综合考虑了概念的类的相似度，框架的相似度，定义的相似度等。一个词在 Hownet 中往往具有多个义项，令 WordSence_Similarity(X, Y)表示词语间的词义相似度并定义如下：

定义 1：设词X具有义项（x_0, …, x_i），Y具有义项（y_0, …, y_j），i >= 0, j >= 0, 则X与Y的词义相似度WordSence_Similarity为：

$$\text{WordSence\_Similarity}(X, Y) = \max_{i,j}(\text{HowNet\_Get\_Concept\_Similarity}(x_i, y_j)) \quad (1)$$

如果WordSence_Similarity(D_A, D_P)大于指定阈值（设为 0.8），将此特征置为True；

（2） 支配词词语相关：Hownet提供的API：HowNet_Get_Concept_Relevance，可以得到一个

概念的相关概念场的词语，利用该API把一个词所有义项对应的相关词语合并作为该词的相关概念词集。如果D_P在D_A的相关概念词集内或相反，认为支配词相关，将该特征置为True。

（3） 支配词字符串一致：D_P和D_A字符串相同则置为True，否则置为False。

3.3.2 句法角色特征

句法角色特征主要考虑名词短语处在句子（子句）的主语部分或是宾语部分。以往利用句法角色的系统，都简单地将名词短语的句法角色设置为句中主语或句中宾语。然而中文代词与英文不同，并没有区分主格、宾格。例如代词“他”，在句中既可以充当句法角色的核心成分，如：“他很高兴。”，也可以充当句法角色的修饰成分，如：“他的老师很高兴。”。基于这样的考虑，将句法角色特征进一步细分为主语（宾语）修饰词或是主语（宾语）核心词。对一个名词短语其可能角色为主语核心词、主语修饰词、宾语核心词和宾语修饰词。本文分别对候选先行语和代词确定其句法角色特征。

如果D_A为句中动词，且依存关系为主谓关系，认为A为句中主语核心词，若依存关系为为动宾关系，认为A为动词宾语核心词；如果D_A是名词短语，且是主语的核心词，则认为A为句中主语的修饰语；如果D_A是名词短语，且是宾语的核心词，则认为A为句中宾语的修饰语。相应地可确定代词P的句法角色。

3.3.3 一般特征

表 1 给出了本文采用的一般特征，在实验部分一般特征将作为一个 Baseline，来考察句法及词义特征对系统性能的影响。

表 1 一般特征

特征名称	取值
候选先行语是否为代词	是、否
候选先行语单复数	单数或复数
候选先行语性别	男、女或不确定
代词性别	男、女或不确定
距离	本句、前一句或前两句及以上
两者是邻近 NP	是、否
候选先行词为句中第一个 NP	是、否

3.4 支持向量机

本文采用的分类器为支持向量机(Support Vector Machine)[15]。这是一种基于统计学习理论，由最优分类界面分类器发展而来的学习算法，具有良好的推广能力。提出以来，在各个领域，包扩自然语言处理领域得到了广泛重视，在很多成功的应用中，都取得了比其他统计学习算法更好或相当的效果。在实验中，使用开源的SVM-Light [1]工具包，参数使用默认值。对每一个测试实例分类器给出类别标记及置信值（confidence）。若一个代词的有N个候选先行词，将构建N个分类实例，选取置信值最大的实例对应的候选先行词作为最终的消解结果。如果最大置信值对应多个候选，我们选择距离代词较近的候选作为消解结果。

4 实验设计与分析

4.1 语料

使用ACE2005评测的中文训练数据作为实验数据。ACE评测由美国国家标准技术研究院（NIST）组织。其研究的主要内容为自动内容抽取，包括抽取语料中的事件、关系、实体等内容。ACE2005中文训练语料分为三个部分，分别为：Broadcast News（BN），Newswire（NWire）和 Weblog（WL）。语料标注了实体、实体属性、实体关系、关系属性，以及对应于同一个现实实体的各个名词短语之间的共指关系。

ACE2005 中文语料提供了实体的 Type 信息，其中 Person(PER)表示人。利用这一信息判断名

[1] http://svmlight.joachims.org/

词短语的类别一致性，过滤掉类别不是 Person（PER）的候选。

此外在 ACE2005 语料的实体标注中，如果实体为单数，被标识为：SUBTYPE="Individual"，如果为复数则被标识为：SUBTYPE="Group"。利用该信息来判断两个实体的单复数一致性。

4.2 评价方法

采用成功率 success 作为评价准则，其计算方法为：

$$\text{success} = \frac{\text{正确消解的代词数}}{\text{待消解的代词数}} \times 100\% \quad (2)$$

若一个代词的消解结果在标准答案中与代词表示的是同一实体，即认为是正确消解。

4.3 实验分析

在三部分语料分别抽取了 501， 566 和 343 个代词做实验。由于语料规模有限，可能存在着训练不充分和特征分布不均匀的问题，采取五等份交叉验证策略，最终取平均值作为消解结果。进一步，将三部分语料合并，构成全部语料。依然采取五等份的交叉验证方式，以检验训练语料更充分的条件下系统的表现。依存句法分析器由哈尔滨工业大学信息检索研究室语言技术平台（LTP）[16]提供。

表 2 加入句法角色特征和支配词词义特征的系统与 Baseline 系统的比较

	BNews(%)	NWire(%)	WL(%)	Total (%)
Norm	80.77	80.61	76.23	81.66
Norm + DpRole	**80.98**	82.61	78.41	83.86
Norm + DpWS	80.75	**83.18**	77.84	82.70
Norm + DpRole + DpWS	80.79	82.81	**78.44**	**84.38**

表 2 展示了采取不同特征的系统在各语料上的成功率。第一行给出了只采用一般特征的 Baseline 系统的表现。DpRole 为依存句法角色特征，DpWS 为支配词词义特征。第二、三行分别表示在一般特征基础上添加依存句法角色特征或依存词义特征后系统的表现，第四行为将两个依存特征全部加入后系统的表现。我们看到，加入句法角色特征后 BNews 语料库上的系统表现达到最佳，尽管提升仅有 0.21。而在加入支配词词义特征后，NWire 语料库上系统的表现提升了 2.57% 达到最佳。将句法角色特征与支配词词义特征全部加入后，WL 上系统成功率提高了 2.21%达到最佳。

表 2 最后一列给出在全部语料上采取不同特征的实验结果。在一般特征基础上只增加依存句法角色特征后系统成功率提高 2.20%，只增加支配词词义特征系统成功率提高 1.04%，将两者全部加入后，系统成功率提高 2.72%，提升比较显著。这证明了基于依存的句法角色特征和支配词词义特征的有效性。并且有理由相信在足够规模的语料上，有效性可能更加明显。

具体考察了单独的支配词词义特征对系统的贡献。表 3 列举了从全部特征中除去欲考察的特征，系统重新运行的表现。All 表示采取所有特征，对于支配词词义特征，具体考察了三个子特征：DpStrEqual 表示支配词字符串相同特征，DpWSenceSimilarity 表示支配词词义相似性特征，DpRelevance 则代表支配词相关性特征。

表 3 单独支配词词义特征对系统影响

	Bnews(%)	Nwire(%)	WL(%)	Total (%)
ALL	80.79	82.81	78.44	84.38
ALL - DpStrEqual	80.58	82.80	78.10	83.63
ALL - DpWSenceSimilarity	80.56	82.98	77.81	84.15
ALL - DpRelevance	80.34	82.44	78.13	84.45

表 3 结果表明，在去除各支配词词义子特征后，各个语料上系统表现基本呈下降趋势，去除支配词字符串相同特征后在全部语料上成功率下降 0.75% 。去掉支配词相关特征，在三个小语料

库上系统的表现均有降低，但在全部语料上略有小幅上升。可能由于语料扩大，引入了更多的噪声，加之词语相关特征的计算相对简单，抗噪能力有限导致。总体来说，三个支配词词义子特征对系统有积极作用。

5 结论

本文的目的在于考察如何直接使用依存句法分析的结果，利用依存语法能够反映句子中词和词的依存关系的特点来帮助代词消解。从实验结果来看，根据依存句法分析结果，可以从句法结构中发现有效的关系，利用支配词及其信息能够构建更多的特征来支持代词消解。我们利用依存信息构建了非常有效的依存句法角色特征。此外，之前的工作中只有在共指消解中考察了语义知识，例如利用Wordnet[17]和Hownet来判断名词短语的语义属性，在代词消解中则没有涉及，因为代词本身并不能反应足够的语义特征。本文中，没有直接考察代词和候选先行词本身而是考察它们的支配词之间的词义相似、词语相关特征，利用Hownet提供的概念相似度和相关词汇的接口来加强代词消解。在ACE2005 中训练语料上的实验结果证明，依存句法角色特征和这种基于支配词词义相似、词语相关的特征都能够对代词消解提供一定的帮助。在扩大语料规模后，系统消解成功率提升更加明显。

本文采用的词义相似、词语相关性的计算方法相对来说比较简单也有一定局限性。倘若能够设计出更合理、更有针对性的词义、语义相似算法将对代词消解提供更大的语义支持。目前指代消解的研究亟需引入更多的背景知识及全局优化策略来寻求突破，词义、语义特征将是值得今后进一步研究的内容。

参考文献：

[1] Jun Lang, Bing Qin, Ting Liu, Sheng Li, Intra-document Coreference Resolution: The state of the art, Journal of Chinese Language and Computing, 2007,17(4):227-253

[2] Hobbs, Jerry R. 1978 "Resolving pronoun references". Lingua, 44, 339-352.

[3] Grosz, B.J., A.K. Joshi, S. Weinstein, Centering: A framework for modelling the local coherence of discourse. Computational Linguistics, 1995.21(2):203-225.

[4] Ge, N., J. Hale, E. Charniak. A statistical approach to anaphora resolution. In: E. Charniak ed. Proc. of the Sixth Workshop on Very Large Corpora. Montreal,Canada: Association for Computational Linguistics, 1998.161-170.

[5] McCarthy, J.F. and W.G. Lehnert. Using decision trees for coreference resolution. In: C.R. Perrault ed. Proc. of the Fourteenth International Joint Conference on Artificial Intelligence. Québec, Canada: Springer, 1995.1050-1055.

[6] 钱伟, et al., 基于最大熵模型的英文名词短语指代消解. 计算机研究与发展, 2003.40(9):1337-1342.

[7] Cardie, C. and K. Wagstaf. Noun phrase coreference as clustering. In: P. Fung and J. Zhou eds. Proc. of the 1999 Joint SIGDAT conference on Empirical Methods in NLP and Very Large Corpora. College Park, MD, USA: Association for Computational Linguistics, 1999.82-89.

[8] 王厚峰, 指代消解的基本方法和实现技术. 中文信息学报, 2002.16(6):9-17.

[9] 王厚峰,梅铮, 鲁棒性的汉语人称代词消解. 软件学报, 2005:700-707.

[10] Lappin, S. and H.J. Leass, An algorithm for pronominal anaphora resolution. Computational Linguistics, 1994.20(4):535--561.

[11] Xiaofeng Yang, Jian Su, Kernel-Based Pronoun Resolution with Structured Syntactic Knowledge. In proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics(ACL’ 06), 41-48.

[12] A.Moschitti. 2004. A study on convolution kernels for shallow semantic parsing. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics(ACL’04),335-342.

[13] Joakim Nivre. 2005. Dependency Grammar and Dependency Parsing. MSI report 05133. Växjö University: School of Mathematics and Systems Engineering.
[14] 董振东 董强, 知网和汉语研究. 当代语言学, 2001.3(1):33-44.
[15] Vapnik V N. The Nature of Statistical Learning Theory [M]. New York: Springer-Verlag,1995.
[16] 基于 XML 的开放式语言技术平台：LTP 郎君, 刘挺, 李生, 张会鹏, 中国中文信息学会成立二十五周年学术年会, 2006 年 11 月, 北京.
[17] Fellbaum, C. ed. Wordnet. An electronic lexical database: MIT Press, 1998.