句法与词义相结合的中文代词消解[1]

宋巍　秦兵　郎君　刘挺
(哈尔滨工业大学信息检索研究室, 哈尔滨 150001)
E-mail: {wsong, bqin, bill_lang, tliu}@ir.hit.edu.cn

摘要：句法知识对代词消解有着很大的支持。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相关性特征，采用支持向量机作为分类器，在 ACE2005 语料上的取得了满意的效果。

关键词：代词消解；依存句法；句法角色；词义相似；支持向量机

Combining Syntax and Word Sense for Chinese Pronoun Resolution

Wei Song　Bing Qin　Jun Lang　Ting Liu
(Information Retrieval Laboratory, Harbin Institute of Technology, Harbin 150001)
E-mail: { wsong, bqin, bill_lang, tliu}@ir.hit.edu.cn

Abstract: Syntactic knowledge is important for pronoun resolution. In recent years, research on dependency parsing becomes active, because dependency grammar benefits to represent the relation between words. 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 our approach gives a good performance and is effective for Chinese pronoun resolution.

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

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

[1]本文受到国家自然科学基金(60575042, 60503072)和 863 项目(2008AA01Z144)资助

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]、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。

在中文上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，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、宾格特征等。这些难点都给指代消解的特征提取带来了很大困难。

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了语义约束。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，第二章将详细介绍我们的指代消解系统框架和基于依存句法分析的特征构建。第三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

我们的方法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测试实例，利用训练得到的模型进行分类，对每一个测试实例分类器给出类别标记及置信值（confidence），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。

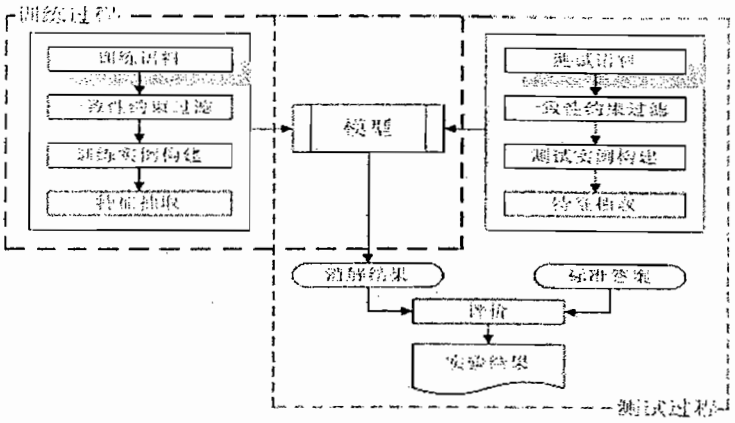

图 1 系统框架

3.2 一致性约束过滤和训练、测试实例构建

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。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”的支配词分别为“病”，“病倒”，给予我们很大的指示性：“他”指向“李老师”的可能性将更大一些。又例如：

“我行使权力加重他的刑罚，判他坐牢7年和鞭打12下。”

其中两个“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我们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，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） DependWordsSimilarity：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，计算两个词语词义相似度。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：

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

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） DependWordRelevant：Hownet提供的API：HowNet_Get_Concept_Relevance可以得到与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。

3.3.2 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。

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

一般特征包括：距离特征，表示代词与候选先行语之间的句子距离；相邻性特征，如果代词与候选先行语相邻；同位和并列结构特征、绑定约束和字符串匹配等词形特征。

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，我们使用开源的 SVM-Light [2]工具包，参数使用默认值。

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

我们实现了一个的 Baseline 系统作为对比，Baseline 系统选择代词之前距离最近的符合类别、性别和单复数一致性的候选作为消解结果。这个方法虽然简单，但综合考虑了人称、性别、单复数和距离因素，仍然不失为一个有效的方法。

在三个领域上分别抽取了 502， 566 和 353 个代词做实验。由于语料规模有限，可能存在着

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

训练不充分和特征分布不均匀的问题，采取交叉验证策略，将代词分为五等份，每次取一份做测试，其余四份做训练，最终取平均值作为消解结果。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1 算法与Baseline系统对比结果

Corpus	BNews(%)	NWire(%)	WL(%)
Baseline	74.9	73.5	68.8
Our Method	83.3	83.2	85.4

表1给出了我们方法与Baseline系统的对比。可以看出，我们的方法在ACE三个语料库上的表现都要远超过Baseline系统。这说明，依存句法分析可以为代词消解提供有效的特征选择。

此外，考察了各个组合特征对系统的贡献。表2列举了从全部特征中除去欲考察的特征，重新运行之后系统的表现。All表示采取所有特征，DpRole为依存句法角色特征，DpSema为支配词语义特征。对于支配词语义特征，具体考察了三个子特征：DpEqual表示相同支配词特征，DpWSenceSimilar表示支配词词义相似性特征，DpRelevance则代表支配词相关性特征。距离、词形特征等一般特征已经证明对指代消解作用明显，这里将不再讨论。

表2 单个特征和组合特征对系统影响

Corpus	BNews(%)	NWire(%)	WL(%)
ALL	83.3	83.2	85.4
ALL - DpRole	79.7	81.9	82.7
ALL - DpSema	82.0	81.7	86.1
ALL - DpEqual	81.9	81.3	84.9
ALL - DpWSenceSimilar	81.6	81.5	85.1
ALL - DpRelevance	81.5	81.9	83.6

表2结果表明，依存句法角色特征的作用明显，去除该特征后在BNews上系统的成功率下降了3.6%，在NWire和WL上分别下降了1.3%和2.7%。

在去除支配词语义相似、相关性特征后，在BNews和NWire上系统成功率分别下降1.3%和1.5%，然而在WL上系统性能上升了0.7。进一步分析三个子特征，单独去除每个子特征后，在三个语料库上，系统的表现都有下降。这说明支配词之间的语义联系作为代词和候选先行语的特征是可以支持代词消解的，但会受噪音影响。特征间的组合依然非常重要，单独有效的特征，组合起来效果却未必更好。

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更有效的关系，包括支配词提供的信息，结合这些信息能够构建更多的特征来支持指代消解。

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。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，如利用Wordnet[17]和Hownet来判断名词短语的语义属性兼容性，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语义特征，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，依存句法角色特征对代词消解系统的作用很大而这种基于支配词词义相似、相关的特征也能够对代词消解提供一定的帮助。但是采用的衡量词语相似、相关性的计算方法相对来说比较粗糙，倘若能够设计出更合理，更有针对性的词义、语义相似算法将对指代消解提供更大的语义支持，同时我们也发现如何去除噪声，如何选择最好的特征组合有待进一步的分析。

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, J.R., Resolving pronoun references, In: B.J. Grosz, K. Sparck-Jones, B.L. Webber, eds. Readings in natural language processing. San Francisco, CA, USA: Morgan Kaufmann Publishers, 1978. 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

[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), page 335-342

[13] Joakim Nivre. 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] V.Vapnik. 1995. The nature of Statistical Learning Theory. Springer

[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.