基于SVM的中文名词短语指代消解研究

高俊伟 孔 芳 朱巧明 李培峰
(苏州大学计算机科学与技术学院 苏州 215006) (江苏省计算机信息处理技术重点实验室 苏州 215006)

摘 要 指代消解是自然语言处理领域中要研究的关键问题之一。在自然语言中,为了使语言简明,减少冗余,往往对同一意思的单词、句子或某一事件用不同的单词来代替。相对于人而言,计算机理解这些指代现象就比较困难,因此近年来关于指代消解的研究越来越多。由于中文指代消解研究起步较晚,因此关于中文名词短语指代消解的研究还比较少,大多研究是关于英文指代消解的。给出了一个基于SVM的中文名词短语指代消解平台并详细介绍了整个实现过程,语料库采用OntoNotes 3.0的中文新闻语料。利用3种评测算法对系统性能进行了评测,结果表明本系统是一个比较好的中文指代消解平台。

关键词 指代消解,名词短语,自然语言处理,SVM
中图法分类号 TP391 **文献标识码** A

Research of Chinese Noun Phrase Anaphora Resolution: A SVM-based Approach

GAO Jun-wei KONG Fang ZHU Qiao-ming LI Pei-feng
(School of Computer Science & Technology, Soochow University, Suzhou 215006, China)
(Key Lab of Computer Information Processing Technology of Jiangsu Province, Suzhou 215006, China)

Abstract Coreference resolution is an important subtask in natural language processing systems. In natural language, to make the natural language clear and explicit illusions, it is common that two or more words, sentences or events which have the same meaning are replaced by different words. In compare to people, it is difficult to understand these phenomenon by using the computer, so more and more researchers focus on noun phrases coreference resolution. A great deal of research has been done on this task in English. In Chinese, because the research about coreference resolution starts late, much less work is done in this area. We presented a Chinese noun phrase coreference resolution system based on a SVM approach and gave the details of the platform in the paper. We adopted three tools to evaluate the performance of the platform. Experiments on the Chinese portion of OntoNotes 3.0 show that the platform achieves a good performance.

Keywords Coreference resolution, Noun phrase, Natural language processing, SVM

1 概述

指代现象是自然语言领域中广泛存在的现象,它是指两个实体是否指向现实世界的同一实体;这对简化语言,减少冗余有很大的作用,使自然语言看起来更加紧凑。指代消解在篇章理解、人机对话、机器翻译、信息抽取、文本摘要等领域中的相关研究也越来越多。指代消解已经成为自然语言理解领域中的热点和难点之一。指代分为两种:回指(Anaphora)和共指(Coreference),其中回指是指当前指代词与文章中出现的词或句子有意义关联性,依存于上下文环境中,在不同的语言环境中可能指代不同的实体;共指则指当前指代词与文章中不同位置出现的词指向同一实体,比如中国、CHINA和中华人民共和国都指向同一个实体。本文主要关注共指消解的研究。

一般来说,指代有多种不同的类型,常见的有代词指代、指示名词短语指代、专有名词短语指代、事件指代、零指代等等。本文关注的是中文名词短语指代消解研究,它包括像代词、指示性名词短语、专有名词、普通名词短语等各类名词短语的指代消解研究。

2 相关工作

近几年,针对指代消解研究的方法不断被提出,并取得了不错的性能。其中早期的指代消解研究主要侧重理论上的探索,运用一些规则进行消解,性能不是很好。典型的有:Hobbs(1976)[1]提出的一种不依赖语义知识和篇章知识,只依赖语法规则和树图信息的算法;Lappin和Leass(1994)[2]

到稿日期:2011-11-02 返修日期:2012-03-09 本文受国家自然科学基金(90920004,60970056,61070123,61003153),江苏省高校自然科学重大基础研究项目(08KJA520002)资助。

高俊伟(1986—),男,硕士生,主要研究方向为自然语言处理,E-mail:gaojunwei100@163.com;**孔 芳**(1977—),女,副教授,主要研究方向为自然语言处理;**朱巧明**(1963—),男,教授,博士生导师,主要研究方向为中文信息处理、网格计算;**李培峰**(1971—),男,副教授,主要研究方向为中文信息处理与自然语言理解。

提出的一种 RAP 算法，它通过 McCord's 提出的一种槽方法解析获得文档的语法结构，再通过计算先行语的突显性和过滤规则进行消解。

随着 Internet 的发展及一些标注语料库的出现，一些基于机器学习的方法被提出。其中 Soon(2001)[3] 首次给出了一个基于分类的指代消解系统的完整实现步骤。他采用了 12 个特征，用决策树算法进行训练和分类，在 MUC-6 和 MUC-7 语料上 F 指数分别为 70.4% 和 63.4%。Ng 等(2002)[4] 对 Soon 的系统进行了扩充，将其特征集合从 12 个扩充到 53 个，另外改变在候选项中按照从右到左的顺序查找先行语，而是从候选项中选择最有可能是先行语的作为结果，试验结果显示其比 Soon 的系统有一定的提高。Yang 等人(2003)[5] 认为以往系统采用的是单候选模型，其不能利用候选先行语之间的信息，因此提出一种双候选模型，即通过学习各先行语候选之间的竞争关系更好地确定先行语。Yang 等人(2004)[6] 又探讨了全匹配特征对名词短语指代消解的影响。Kong(2009)[7,12] 从语义角度出发探讨了中心理论在代词指代消解中的作用且分析了在中心理论指导下语义角色在指代消解中的应用。

与国际上英文指代消解的长期研究相比，中文指代消解的研究起步较晚，并且主要集中在代词的消解研究。相关的研究有王厚峰[8,9] 对人称代词互指对消解的研究。李国臣(2005)[10] 提出的一种利用决策树算法和优先选择策略进行指代消解的方法，通过考虑距离属性和频次属性，对候选互指对进行优先选择的人称代词的消解。周俊生(2007)[11] 利用了一种无监督的聚类算法来实现名词短语的指代消解，他引入了一个带权图对指代消解问题进行建模，并取得了不错的性能。

3 名词短语指代消解基本框架

本系统是基于机器学习的中文名词短语指代消解平台，语料采用的是 OntoNotes 3.0 中文新闻语料。其思想是将指代消解问题理解为一个二元分类问题，系统采用 SVM[1] 作为分类器，通过分类器判断照应语与每个候选的先行语是否具有指代关系。整个平台的流程如图 1 所示。

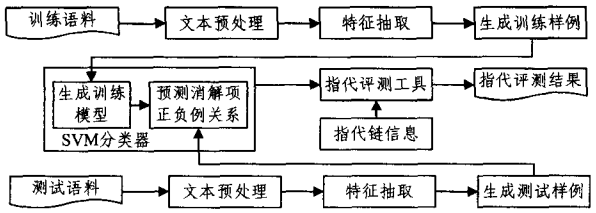

图 1 中文名词短语指代消解平台

整个平台可以分为预处理、特征向量的选择、训练样例和测试样例的生成。

3.1 预处理

整个预处理模块由分句、分词、命名实体识别、名词短语及中心词的识别、语义角色信息的获得等部分组成。由于分词的准确度直接影响到待消解项识别的准确度，对指代消解性能影响较大，因此本文采用的是语料库中标注的标准分词结果。根据这些分词结果，用 Stanford Parser[2] 工具生成文章中每一句的句法树及依存关系信息。从句法树信息中抽取出所有以 NP 开头的名词短语列表。然后利用依存关系信息获得所有名词短语的中心词，中心词是名词短语的核心词，是名词短语的主干部分。在依存关系中，每条记录的格式为：依存关系(中心词-位置，依存词-位置)，因此利用依存关系可以获得每一个名词短语的中心词。例如"外商 投资 企业 成为 中国 外贸 重要 增长点 "，该句的依存关系如图 2 所示。

```
nn(企业-3，外商-1)
nn(企业-3，投资-2)
nsubj(成为-4，企业-3)
nn(外贸-6，中国-5)
nn(增长点-8，外贸-6)
zmod(增长点-8，重要-7)
dobj(成为-4，增长点-8)
```

图 2 用工具生成例句的依存关系

例如"nn(企业-3,外商-1)"中，"企业"是中心词，"外商"是依存词，每个词后面的数字是该词在句子中的位置。"nn"指两个词之间的依存关系为名词修饰关系。因此对应于名词短语"外商投资企业"，根据依存关系找出它的中心词为"企业"。命名实体信息是指代消解中重要的一部分，其语义类别一致性在指代消解的过滤阶段和特征选择阶段都具有很重要的意义，其任务主要是对文本中的专有名称，包括人名、地名、机构名以及时间表达式和数字表达式进行识别。语义角色信息是句中的名词短语在相应动词的驱动下所承担的句法成分。核心的语义角色为 Arg0－Arg5 6 种，Arg0 通常表示动作的施事，Arg1 通常表示动作的影响等。其余的语义角色为附加语义角色。本文中只考虑 Arg0 和 Arg1 两种形式，其中 Arg0 表示该名词短语在某一动词驱动下充当施事者，Arg1 表示该名词短语在某一动词驱动下充当受事者。例如"[外商投资企业 Arg0][成为 V][中国外贸重要增长点 Arg1]"，在这句话中，"成为"是谓词，"外商投资企业"是其施事者，而"中国外贸重要增长点"是其受事者。本文中命名实体信息及语义角色信息由实验室[3] 开发的 MYNER 及 SRL 工具生成。

3.2 特征向量的选择

在机器学习方法中，特征向量的选择对指代消解的性能有很重要的影响，选择好的特征向量能明显提高指代消解的性能。选择的特征要不能互相冲突，不能局限于某一领域，要有一定的通用性，能反映语言学的共同特征。本系统所采用特征信息及其获得方法如下所示。

1. Distance：先行语和照应语之间的距离，在一句之内为 0，相差一句为 0.1，相差 2 句为 0.2，以此类推。如果照应语为代词，往前搜索 4 句；如果是专有名词，则往前搜索 8 句；其它情况下的名词短语，往前搜索 9 句，所以该特征取值范围为 0～0.9。

2. StringMatch：若先行语和照应语完全匹配相等，则该特征值为 1，否则为 0。

1) http://svmlight.joachims.org/

2) http://nlp.stanford.edu/software/lex-parser.shtml

3) http//nlp.suda.edu.cn

3. Alias：若先行语或照应语有一个是另一个的别名，则该特征值为1，否则为0。本文在判断该特征时，若有一个名词短语是从另一个名词短语中抽取出来的，则就认为此名词短语是另一个的别名。

4. Appositive：如果照应语和先行语是同位语，则该特征值为1，否则为0。本文规定如果两个短语并列出现充当句中某一成分，且其中一个为专有名词，这种情况下认为它们之间有同位语关系。

5. i-Pronoun：若照应语为代词，则该特征值为1，否则为0。如果词性为PN，则该名词短语为代词，此特征值为1。

6. j-Pronoun：若先行语为代词，则该特征值为1，否则为0。即先行语词性如果为PN，则该名词短语为代词，特征值为1，否则为0。

7. DemonstrativeNP：若照应语为指示性名词短语，则该特征值为1，否则为0。若词性为DP，则该特征值为1，否则为0。

8. Semantic Class Agreement：若照应语和先行语实体类别一致，则该特征为1，否则为0，如照应语实体类别为Person，先行语实体类别也为Person，则该特征值为1。

9. i-ProperNP：若候选项是专有名词，则该特征值为1，否则为0，如果该照应语的词性为NR，则该特征值为1，否则为0。

10. j-ProperNP：若先行语是专有名词，则该特征值为1，否则为0。

11. i-arg0：若照应语在句中充当某一动词的施事者，则该特征值为1，否则为0。即如果其语义角色信息为Arg0，则该特征值为1。

12. i-arg1：若照应语在句中充当某一动词的受施者，则该特征值为1，否则为0。即如果其语义角色信息为Arg1，则该特征值为1。

13. j-arg0：如果先行语语义角色信息为Arg0，则该特征值为1，否则为0。

14. j-arg1：如果先行语语义角色信息为Arg1，则该特征值为1，否则为0。

15. Similarity：该特征的值是照应语和候选先行语之间的相似度的值与它们中心词之间的相似度的值中的最大者。相似度算法是根据刘群的基于知网的语义相似度算法获得的，其定义如下：

$$\mathrm{Sim}(W1,W2)=\frac{a}{Dis(W1,W2)+a}$$

$Dis(W1,W2)$是词$W1$与$W2$之间的距离，是一种基于世界知识的计算方法，a是一个可调节的参数。

16. Nest in：照应语是否嵌套在某一名词短语内，若是，则该特征值为1，否则为0。

17. Nest out：照应语是否嵌套其它名词短语，若嵌套，则该特征值为1，否则为0。

3.3 训练样例和测试样例的生成

在生成训练实例的过程中，从第二个名词短语开始，往后的名词短语都是潜在的照应语，在该名词短语之前的名词短语都可作为该照应语的候选先行语。往前找到该照应语的最近先行语组成正例，在最近先行语和照应语之间的名词短语和该照应语构成负例。比如该平台从一篇文章中识别出的名词短语是A1、A2、A3、A4、A5从$Ai(i\geqslant 2)$开始，先从语料库中查找Ai是否在指代链中，若Ai不在指代链中，即不属于任何指代链，则将它视为非消解项，无需查找它的先行语；若在指代链中，则该名词短语作为待消解项，分别与其前面的名词短语$A(i-1)$至A0进行组对直到找到其最近先行语。然后文中定义了一些规则，对这些组对进行过滤，将那些单复数不一致、性别不一致、实体类别不一致、照应语与先行语是嵌套关系、先行语和照应语实体类别同为other且相似度小于0.7的等都过滤掉。实验结果表明，当相似度阈值设为0.7时系统有较好的性能。

测试实例的构建过程与训练实例大体一致，不同的是测试阶段由于没有指代链信息，因此不知道一个名词短语是不是照应语，所以在测试阶段将所有识别出的名词短语都作为照应语，都往前寻找其先行语进行组对。另外根据实验统计结果，若照应语为代词，则仅需对照应语前面4句以内的名词短语组对，若是专有名词短语，则对照应语前面8句以内的名词短语进行组对。其它类别的名词短语则对前面9句以内的名词短语进行组对。“候选先行语-照应语”组对生成完以后，按照文中定义的规则进行过滤。然后按照预先设定的特征向量获得这些测试实例的特征向量的值，构建测试实例。指代所占比例与距离之间的关系如表1所列。

表1 指代与距离之间的关系

照应语为代词		照应语为专有名词短语		其它名词短语	
句子距离	所占比例	句子距离	所占比例	句子距离	所占比例
≤1	84.51%	≤2	81.87%	≤2	74.10%
≤2	92.93%	≤4	93.68%	≤4	87.48%
≤3	96.18%	≤6	96.82%	≤6	93.05%
≤4	97.13%	≤7	97.79%	≤8	96.34%
		≤8	98.26%	≤9	97.30%

4 实验结果及分析

本系统采用OntoNotes 3.0中文语料中的新闻语料，一共有325篇文章，文中将其分为5等份，采用5倍交叉验证方法，每次取4份作为训练语料，1份作为测试语料，最后取其平均值作为最终的结果。本文采用多种评测方法对平台进行评测并给出了该系统的准确率(P)和召回率(R)，F值。其中，P代表准确率，是指代消解结果中正确消解的对象数目占实际消解的对象数目的百分比，反映指代消解系统的准确程度。R代表召回率，是指代消解结果中正确消解的对象数目占消解系统应消解对象总数的百分比，反映了指代消解系统的完备性。F值是准确率和召回率这两个值的综合值，即：$F=2\times R\times P/(R+P)$。

实验得出的相关数据如表2所列。

表2 实验结果

	MUC			BCUB			CEAFE		
	P	R	F	P	R	F	P	R	F
Auto	52.53	59.83	55.94	74.33	78.74	76.47	51.58	46.28	48.79
Golden	78.58	65.74	71.59	87.83	75.32	81.1	49.42	61.36	54.75

实验结果表明，通过对平台的预处理等一系列操作后，平台的F值在MUC评测下可以达到55.94%，当所有的信息都是从语料库中获得时，即标准情况下，平台的F值在MUC评测可以达到71.59%。对比Auto和golden两种情况下的性能，发现在MUC评测下F值相差15个点左右，一方面是

由于中文指代消解研究起步较晚，目前没有很好的命名实体识别工具和语义角色识别等工具；另外在名词短语识别过程中，本文是从句子的句法树信息中抽取出名词短语的，发现在名词短语及中心词识别过程中存在名词短语识别错误等问题，会引入很多噪音。另一方面在用相似度对实例对进行过滤的过程中，尽管可以过滤掉很大一部分负例，但也存在正例被过滤掉的情况，这样也会影响平台的性能。从实验结果可以看出，该平台是处理中文指代消解的一个较好的平台。在指代消解中，特征的选择对平台性能影响较大，为了探讨单个特征对分类性能的贡献度，本文对各个特征所起的作用分别进行了详细实验，结果如表 3 所列。

表 3　单个特征对系统影响

	MUC			BCUB			CEAFE		
Auto	P	R	F	P	R	F	P	R	F
Distance	0	0	0	100	31.21	47.58	17.29	55.4	26.36
String Match	48.94	52.4	50.61	74.41	75.97	75.18	49.01	46.41	47.68
Alias	34.19	55.36	42.27	60.12	82.66	69.61	53.61	37.36	44.03
Appositive	67.16	8.02	14.33	89.56	36.73	52.1	20.25	61.29	30.44
i-Pronoun	0	0	0	100	31.21	47.58	17.29	55.4	26.36
j-Pronoun	0	0	0	100	31.21	47.58	17.29	55.4	26.36
DemonstrativeNP	0	0	0	100	31.21	47.58	17.29	55.4	26.36
Semantic Class Agreement	0	0	0	100	31.21	47.58	17.29	55.4	26.36
i-ProperNP	0	0	0	100	31.21	47.58	17.29	55.4	26.36
j-ProperNP	0	0	0	100	31.21	47.58	17.29	55.4	26.36
i-arg0	0	0	0	100	31.21	47.58	17.29	55.4	26.36
i-arg1	0	0	0	100	31.21	47.58	17.29	55.4	26.36
j-arg0	0	0	0	100	31.21	47.58	17.29	55.4	26.36
j-arg1	0	0	0	100	31.21	47.58	17.29	55.4	26.36
Similarity	0	0	0	100	31.21	47.58	17.29	55.4	26.36
Nest in	0	0	0	100	31.21	47.58	17.29	55.4	26.36
Nest out	0	0	0	100	31.21	47.58	17.29	55.4	26.36

从表 3 可以看出，平台单独使用一个特征时，全匹配特征(String Match)、别名特征(Alias)和同位语特征(Appositive)贡献度较为明显。为了进一步探讨其它特征对平台的贡献度，文中将特征集合分为两组，一组是贡献度较为明显的全匹配特征、别名特征和同位语特征，另一组是贡献度不明显的剩下的其它特征。然后单独用这两组特征集合进行实验，在 Auto 情况下平台性能如表 4 所列。

表 4　特征组合对平台性能影响

	MUC			BCUB			CEAFE		
Auto	P	R	F	P	R	F	P	R	F
全匹配+别名+同位语特征	48.06	53.62	50.69	74.04	77.15	75.56	49.56	46.01	47.72
其它特征	42.2	9.22	15.14	91.43	42.01	57.57	22.88	51.48	31.68

从表 4 可以看出，贡献度较为明显的几个特征组合使用时，平台性能在 MUC 评测下 F 值为 50.69%。贡献度不明显的其它特征组合使用时，平台性能在 MUC 评测下 F 值为 15.14%。因此，贡献度不明显的这些特征对指代消解平台有一定的影响，这些特征对平台的贡献度较小一方面是由于这些特征对指代消解任务而言区分度没有全匹配特征，别名特征等区分度大；另一方面也是由于在文中预处理过程中存在误差，一些信息不能很好地获取，比如命名实体信息、语义角色信息等。当平台都采用标准语料库中给出的信息时，平台在 MUC 评测下 F 值可以达到 71.59%，这也说明了这些特征对指代消解任务而言是有一定的贡献度的。

结束语　本文实现了一个基于 SVM 的中文名词短语指代消解平台，其采用的是 OntoNotes 3.0 的中文新闻语料。文中详细介绍了整个平台的实现过程，并用 3 种评测工具对平台性能进行评测。同时为了探讨平台中所选特征对平台的贡献度，文中给出了系统只选用一个特征时平台的性能，并且将贡献度较为明显的特征与不明显的特征分别进行组合实验，从实验结果可以看出，这些特征对平台都是有一定的贡献度的。指代消解是自然语言处理领域的一个关键任务之一，随着篇章理解、人机对话、机器学习等领域的发展，指代消解也将受到越来越多人的关注。相对于英文而言，中文名词短语指代消解目前研究还较少，一方面是由于公开的语料库较少，缺少一些很好的中文语言的分析工具，如句法分析、语义分析等工具。另一方面，相对于英文，中文在特征选择时难度较大，比如性别特征、单复数特征等对中文而言，这些特征不容易获取，而且对指代消解的作用也不像英文那样明显。本文实现了一个基于机器学习方法的中文名词短语指代消解平台，实验结果表明，本系统是处理中文指代消解的一个不错的平台。本文中主要采取的是一些平面特征，下一步工作将继续完善该指代消解平台，针对中文语言特点，找出一些有效特征并探讨结构化特征对指代消解性能的影响。

参 考 文 献

[1] Hobbs J. Resolving pronoun reference[J]. Lingua, 1978, 44: 311-338

[2] Lappin S, Leass H J. An algorithm for pronominal coreference resolution[J]. Computational Linguistics, 1994, 20(4): 535-561

[3] Soon W M, Ng H T, Lim D C Y, et al. A machine learning approach to coreference resolution of noun phrases[J]. Computational Linguistics, 2001, 27(4): 521-544

[4] Ng V, Cardie C. Improving machine learning approaches to coreference resolution[C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002, 104-111

[5] Yang X F, Zhou G D, Su J, et al. Coreference Resolution Using Competition Learning Approach[C]// ACL. 2003: 176-183

[6] Yang X F, Zhou G D, Su J. et al. Improving Noun Phrase Coreference Resolution by Matching Strings[C]// IJCNLP'2004

[7] Kong F, Zhou G D, Zhu Q M. Employing the Centering Theory in Pronoun Resolution from the Semantic Perspective[C]// Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. 2009: 987-996

[8] 王厚峰，何婷婷. 汉语中人称代词的消解研究[J]. 计算机学报，2001, 24(2): 6-13

[9] 王厚峰，梅铮. 鲁棒性的汉语人称代词消解[J]. 软件学报，2005, 16(5): 700-707

[10] 李国臣，罗云飞. 采用优先选择策略的中文人称代词的指代消解[J]. 中文信息学报，2005, 119(14): 24-30

[11] 周俊生，黄书剑. 一种基于图划分的无监督汉语指代消解算法[J]. 中文信息学报，2007, 21(2): 77-82

[12] 孔芳，朱巧明，周国栋，等. 基于中心理论的指代消解研究[J]. 计算机科学，2009, 36(6): 219-222