基于规则的中文零指代项识别研究

秦凯伟 孔 芳 李培峰 朱巧明
(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 提出了一个基于规则的中文零指代项识别方法,即输入一个句法分析树,根据这个句法分析树得到当前词的最小IP子树,再依据得到的IP子树提出中文零指代识别的一些规则。所用的语料是Ontonotes。从实验结果可以看到,该方法在标准的句法分析树上F值能达到82.45%,在自动句法树上其也能达到66.45%。从实验结果可以看出,该方法在中文零指代识别上具有很好的性能。

关键词 自然语言处理,中文零指代,句法分析树,基于规则,Ontonotes3.0
中图法分类号 TP391 **文献标识码** A

Rule-based Identification of Chinese Zero Anaphora

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

Abstract A rule-based approach for Chinese zero anaphor detection was proposed. Given a parse tree, the smallest IP sub-tree covering the current predicate was captured. Based on this IP sub-tree, some rules were proposed for detecting whether a Chinese zero anaphor exists. This paper also systematically evaluated the rule-based method on OntoNotes corpus. Using golden parse tree, our method achieves 82.45 in F-measure. And the F-measure is 63.84 using automatic parser. The experiment results show that our method is very effective on Chinese zero anaphor detection.

Keywords Natural language processing, Chinese zero anaphora, Parsing tree, Rule-based, Ontonote3.0

1 引言

中文零指代(Chinese Zero Anaphora)是指人们在特定的语言环境下,在不影响意思表达的前提下,为了使语言简洁明了,有时会省去某些语言成分,这部分的省略称为零指代项,中文零指代是指代的一种。

中文零指代的研究是自然语言处理的关键和热点问题之一,在自然语言篇章理解中举足轻重。通常在一段话中,为了保证文本的简洁明了,文本中往往会省略掉很多信息,人们能通过上下文获得这些信息,但是机器对于缺省的地方不能理解,这就要有一种方法来从文本中获得缺省的信息。中文零指代的研究就是为解决这样的问题而提出的。中文零指代的研究不但在信息抽取中起着重要作用,在机器翻译(Machine Translation)、文本分类(Text Categorization)和信息抽取(Information Extraction)等应用中也极为关键。

本文的主要工作是在Ontonotes3.0的语料上进行中文零指代的研究,并用规则的方法构建一个中文零指代项的识别系统,同时把构建的系统在Berkeley的自动句法分析树上进行测试研究,得到一些对中文零指代缺省项识别研究有用的信息。本文第2节讲述中文缺省项识别研究的相关工作;第3节主要介绍基于规则的中文零指代缺省项的识别研究;第4节主要分析本文所做实验的结果;最后给出本文工作的总结和未来可努力的方向。

2 相关工作

目前国外对英文的指代研究相对比较多,因为中文语法比英文语法复杂,所以在中文领域的零指代研究相对来说还是比较少。中文零指代的研究一般都分成3步进行,首先是对中文零指代项的识别,然后确定中文零指代项的候选词,最后对中文零指代项进行消解研究。对于中文零指代消解[1,2]部分,目前国内外的研究比较多,Wu等[3,4]提出了基于实例推理的零指代消解,Ryu Iida等[5,6]提出了跨语言的ILP零指代消解,而比较有代表性的是Yeh等[7,8]提出的基于中心理论的中文零指代消解。但是在第一步上的研究目前比较少,比较相似的研究有Yang等[9]提出的基于机器学习方法的缺省识别,朱勤宇[10]从语法的角度介绍了汉语零形回指的驱动力,这为后面的研究铺下了语法基础。

除了上述这些外,Kong[11]和Zhou[11]在同一个框架下同

到稿日期:2011-12-20 返修日期:2012-03-14 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),苏州市科技计划项目(SYG201112)资助。

秦凯伟(1987—),男,硕士生,主要研究方向为自然语言处理,E-mail:chinaqkw@yeah.net;孔 芳(1977—),女,副教授,主要研究方向为自然语言处理;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;朱巧明(1963—),男,教授,博士生导师,主要研究方向为自然语言处理、网格计算。

时进行了识别和消解研究，他们是目前唯一在同一框架下同时进行了零指代识别和消解研究的。该文提出了基于树核函数的零指代识别和消解的方法，该方法从结构化的信息入手，来研究中文零指代。

Zhao 和 Ng[12] 用机器学习的方法对中文零指代项识别和消解同时做了研究，提出了用特征的方法来识别中文零指代项。不过由于他们的训练语料的不平衡性，使得提出的特征从结果来看不是很好，但是他们同时做了实验，得到了性能最好时的正负比例。

3 基于规则的中文零指代项研究

3.1 语料库

本文语料使用的是 Ontonotes3.0 中的中文新华前 100 篇语料，所以在进行缺省项的识别研究之前，先对标准语料进行一些处理。因为标准语料上已经标注好了零指代项的位置，所以要把标准语料上标注好的零指代项删掉，得到本文要进行识别的测试语料。以下识别实验都是在这个测试语料上进行的。

为了更好地对规则进行总结，本文对 Ontonotes3.0 上的中文新华 100 篇语料进行了统计，如表 1 所列。

表 1 Ontonotes3.0 语料统计

类型	承担语法		总数
-NONE- *T*	NP-SBJ	430	742
	NP-OBJ	191	
	其它	121	
-NONE- *pro*	NP-SBJ	442	446
	NP-OBJ	4	
	其它	0	
-NONE- *PRO*	NP-SBJ	397	399
	NP-OBJ	0	
	其它	2	
-NONE- *RNR*	NP-SBJ	0	44
	NP-OBJ	25	
	其它	19	
-NONE- *OP*	NP-SBJ	0	722
	NP-OBJ	0	
	其它	722	
其它	NP-SBJ	2	4
	NP-OBJ	0	
	其它	2	
总数	NP-SBJ	1271	2357
	NP-OBJ	220	
	其它	866	

从上面的数据可以看出，主语部分占了整个缺省的53.92%，宾语部分占了整个缺省的 9.33%，而其它占了 36.74%。可以看出主语和宾语两个部分占了整个缺省的 63%。所以本文所提出的规则主要针对的是主语和宾语。

从统计结果可以看出，类型“OP”占了整个缺省项的 30.63%。其在缺省中不承担主语和宾语的角色，而本文所要研究的是零指代的缺省，并为后续零指代的消解研究提供基础。所以类型“OP”对于本文没有任何的意义，本文在制定规则的时候就将其忽略掉了。这就是为什么下文提出的规则中没有考虑占整个缺省项 30.63% 的“OP”类型。去除类型“OP”后，我们发现整个主语和宾语占整个缺省项的 91.2%，而其它仅仅占整个缺省项的 9%不到。下文提出规则的时候就忽略了这一小部分，因为其所占的比例与主语和宾语相比非常得小。

3.2 规则集

本文提出的规则参考了 Yeh[13] 和 Yang[14] 提出的规则，并对其进行改进。从实验结果看，所提出的改进是有效的。

Yeh 等[13] 提出的三元组 $T=\{S,P,O\}$ 中，S 表示语法成分是子句主语的名词列表；P 表示语法成分是子句谓语的动词或介词列表；O 表示语法成分是子句宾语的名词列表。根据三元组，提出的三元规则如下：

$Triple1(S,P,O) \rightarrow np(S),vtp(P),np(O)$

$Triple2(S,P,none) \rightarrow np(S),vip(P)$

$Triple3(S,P,O) \rightarrow np(S),prep(P),np(O)$

$Triple4(S,none,none) \rightarrow np(s)$

Yeh 等[13] 的处理是在自己收集的语料上进行的，这样并不具有可比性，而且他们用的是 shallow parsing，这和本文提出的方法是不同的。

从上面对语料的统计结果可以看出，91.2%的缺省是在主语和宾语上，如果我们处理整个句法分析树，势必会影响性能。由于大部分的缺省在主语和宾语上，因此考虑把整个句法分析树切分成若干小句法分析树，并把形成的子句法分析树重新看成一个独立的句法分析树，并对这些句法分析树进行独立的处理。经过这样的处理，就可以过滤掉很多无用的信息，因为本文所要识别的大部分是主语和宾语。

基于规则的中文缺省项识别方法是非常依赖于句法分析树的。完整的句法分析树经过裁剪，得到了若干个小的句法分析树，但保留了大部分有用的信息，在裁剪过程中，有一小部分的信息可能会被丢失。但经过权衡，发现丢掉这一小部分信息可以换来很高的识别率，这样的丢失是可允许的。

本文所用的规则如下：

1)输入整个句法分析树，将整个句法分析树分成若干个不相等的 IP 子树。

2)输入 IP 树，对输入的句法分析树进行层次遍历，找到第一个“VP”节点，然后找到“VP”节点的左兄弟节点，如果存在左兄弟节点且左兄弟节点承担的语法角色是“-SBJ”，那么该“VP”节点不存在缺省。如果“VP”节点不存在左兄弟，那么“VP”节点前面存在缺省成分，并且是主语缺省。

3)输入 IP 树，对输入的句法分析树进行层次遍历，找到最近的“VP”节点，判断此“VP”的动词是否是不及物动词，若动词是及物动词，遍历这个“VP”子树，找到其右孩子，如果右孩子承担的语法角色是“-OBJ”，那么不存在缺省，否则，次动词后面存在缺省，并且是宾语缺省。若动词是不及物动词，则不存在宾语缺省。

4)判断句法分析树的若干小 IP 树是否全部完成，如果没有就跳到规则 2)开始执行。

从以上规则可以看出，本文处理的对象是最小 IP 树，核心是动词短语“VP”，用“VP”动词短语进行驱动，来识别缺省项。本文不在整个句法分析树上进行识别是因为整个句法分析的噪音太大，对性能的影响也很大，所以本文提出了在 IP 树上进行处理。

3.3 规则的使用方法

根据本文提出的规则，对中文零指代项识别的流程大致如图 1 所示。

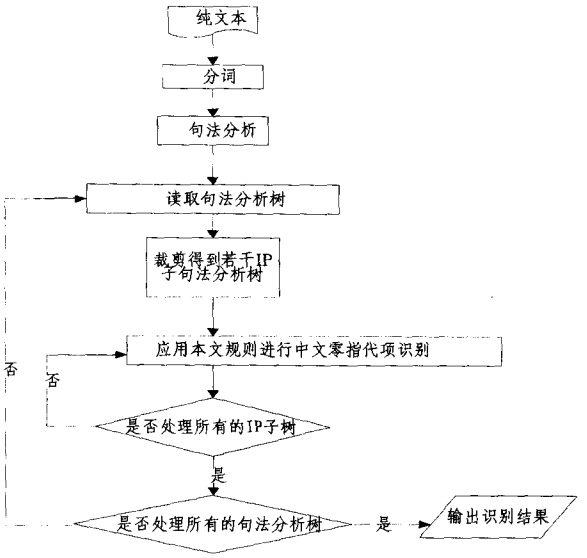

图 1　基于本文规则的识别流程图

下面详细介绍基于本文所提出的规则进行缺省项识别的例子。

建筑是开发浦东的一项主要经济活动，这些年有数百家建筑公司、四千余个建筑工地遍布在这片热土上。

【【(建筑)(是)【(NP-SBJ(NONE))(开发)(浦东)】(的)(一项)(主要)(经济)(活动)】，【(NP-SBJ(NONE))(这些)(年)(有)【(数百)(家)(建筑)(公司)、(四千余)(个)(建筑)(工地)(遍布)(在)(这)(片)(热土)(上)】】。】

1)先对这一输入串进行处理，得到若干个不相等的子句，也就是"【】"里面的内容。总共得到 4 个子句。

2)处理第一个子句，即"建筑是开发浦东的一项主要经济活动"，其中因为"开发浦东"也是一个子句，把"开发浦东"看作一个整体进行处理。应用本文提出的规则进行识别研究，先找到"是"作为 VP 驱动词，并发现 VP 前面存在主语"建筑"，则 VP 前面不存在主语缺省，又因为"是"是不及物动词，所以也不存在宾语缺省。

3)处理第二个子句"开发普东"，应用规则找到动词"开发"作为驱动词，并发现 VP 前面不存在主语，所以"开发"前面存在主语缺省。跳出处理，本文假设每个子句中只存在一个缺省。

4)处理第三个句子"这些年有数百家建筑公司、四千余个建筑工地遍布在这片热土上"，找到"这些年"作为动词短语作为驱动词，并发现前面不存在主语，所以"这些年"前面存在主语缺省，跳出处理。

5)处理最后一个子句"数百家建筑公司、四千余个建筑工地遍布在这片热土上"，找到动词驱动"遍布"，发现前面存在主语"数百家建筑公司、四千余个建筑工地"，则"遍布"前面不存在缺省，又因为"遍布"为不及物动词，所以不存在宾语缺省。

6)所有子句处理完成，这个识别过程完成。

以上就是本文基于规则识别方法的整个过程，基于规则的缺省项识别结果在下面章节进行详细分析。

4　实验结果与分析

本文采用了国际上通用的 MUC 的评测方法进行评测。MUC 对指代消解结果的技术评估有 3 个重要标准：召回率 R(Recall)、准确率 P(Precision)和 F 值。召回率 R，是指识别出来正确的缺省项的数目占实际上缺省项的数目，它反映的是缺省项识别的完备性，即式(1)。准确率 P，是指识别出来正确的缺省项的数目占实际识别的缺省代名词数目的百分比，它反映的是指代消解系统的准确程度，即式(2)。当比较两个不同指代系统的性能时，一般使用这两个指标的综合值：F 值，即式(3)。

$$R=\frac{\text{识别出来正确的缺省项数目}}{\text{实际缺省项的数目}} \tag{1}$$

$$P=\frac{\text{识别出来正确的缺省项数目}}{\text{实际识别出来缺省项的数目}} \tag{2}$$

$$F=\frac{2*P*R}{P+R} \tag{3}$$

4.1　基于标准的句法分析树

将本文提出的规则在 Ontonotes3.0 语料的新华 100 篇文章下进行了测试，得到了表 2、表 3 所列的测试结果。

表 2　基于标准句法分析树的识别数目

总数	识别出数目	正确识别出数目
1635	1344	1228

表 3　基于标准句法分析树的识别结果

方法	召回率	准确率	F 值
规则	75.11	91.37	82.45

从表 2 和表 3 可以看出，在 Ontonotes3.0 语料上，本文提出的规则具有很好的效果，F 值能达到 82.45%。为了更好地为中文零指代消解做贡献，本文还对中文零指代缺省项的类别做了统计，实验数据见表 4。

表 4　各个零指代类型的识别数目

类别	总数	正确识别数目	召回率(%)
-NONE- *T*	742	422	56.87
-NONE- *pro*	446	434	97.31
-NONE- *PRO*	399	370	92.73
-NONE- *RNR*	44	0	0
其它	4	2	50

从表 4 的数据可以很明显地看出，本文提出的规则对于类型"pro"和"PRO"具有很高的召回率，分别达到了 97.31% 和 92.73%。对于中文零指代的研究，大部分在这两个类型上。

4.2　基于自动的句法分析树

为了更好地研究本文提出的规则，将其在自动的句法分析树上进行了实验。本文用的是 Berkeley 的句法分析工具，同时在标准的分词下面进行实验，结果见表 5。

表 5　基于自动句法分析树的识别结果

方法	召回率	准确率	F 值
规则	56.94	79.78	66.45

从表 5 中可以看出，本文提出的规则在自动的句法树上的性能 F 值能达到 66.45%，比标准的低了将近 16%。为了更好地研究和标准句法树的区别，本文对自动句法分析树也做了类别统计，实验结果见表 6。

表 6　基于自动句法分析树的各个零指代类型识别结果

类别	总数	正确识别数目	召回率(%)
-NONE- *T*	742	361	48.65
-NONE- *pro*	446	267	59.87
-NONE- *PRO*	399	303	75.94
-NONE- *RNR*	44	0	0
其它	4	0	0

从表5和表6的数据可以看出，本文提出的规则对句法树依赖比较大，例如类型"pro"在标准的句法分析树上本文的规则对其的召回率是97.31%，在自动的句法分析树上召回率为59.87%，可见基于规则的中文零指代项识别对于句法树的依赖是比较大的。正是因为规则的方法对句法分析树的依赖大，所以本文提出的规则中最重要的一点就是不针对整个句法分析树进行处理，而是对叶子节点的第一个IP节点的子树进行处理。这样，本文提出的规则无论在标准的句法分析树上还是自动句法分析树上，都比别人的方法性能要好，见表7。

表7 性能比较

作者	Parser类型	召回率(%)	正确率(%)	F值(%)
Yang,Xue[9]	标准	70.5	75.3	72.8
Yang,Kong[14]	标准	70.2	68.5	69.3
本文	标准	75.11	91.37	82.45
Yang,Xue[9]	自动	50.2	57.9	53.8
本文	自动	56.94	79.78	66.45

在标准句法分析树上应用本文提出的规则，从表7可以看出，本文提出的规则的性能要比Yang,Xue[9]的高10%左右，比Yang,Kong[14]的高13%。所以可以证明，其对IP子树的处理是非常有效的。对自动句法分析的结果也可以看出，本文提出的规则性能要比Yang,Xue[9]的系统性能高将近13%，这样更加证明了在自动句法分析树进行时，对IP子树进行处理时，其可以有效地避免句法分析树上的错误信息。

结束语 在中文零指代研究中，对中文零指代项的识别一直是一个比较难的研究方向，不过近年来，对于缺省识别的研究的重视，使得中文零指代识别研究有了不错的进展。本文是在基于规则的情况下对中文零指代项进行识别的。从实验结果来看，本文提出的规则是可行的；并且在自动的句法树情况下，本文提出的识别方法具有很高的效率，最好的F值能达到66.45%。

本文制定的规则中也漏掉了很多缺省项，在接下来的工作中，我们可以将机器学习的方法和规则相结合，以尽可能地提高识别系统的性能。

参考文献

[1] Soon W M, Ng H T, Lim. A machine learning approach to coreference resolution of noun phrase[J]. Computational Linguistics, 2001,27(4):521-544

[2] 王厚峰. 指代消解的基本方法和实现技术[J]. 中文信息学报, 2002(6):9-17

[3] Wu Dian-song, Liang T. A Case-Based Reasoning Approach to Zero Anaphora Resolution in Chinese Texts[C]// ICCPOL 2006. 2006:520-531

[4] Wu Dian-song, Liang T. Zero anaphora resolution by case-based reasoning and pattern conceptualization[J]. ExperSyst. With Appl.: An International Journal, 2009, 36(4):7544-7551

[5] Iida R, Poesio M. A Cross-Lingual ILP Solution to Zero Anaphora Resolution[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. 2011:804-813

[6] Iida R, Inui K, Matsumoto Y. Exploiting syntactic patterns as clues in zero-anaphora resolution[C]// Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics COLING-ACL2006. 2006:625-632

[7] Yeh C-L, Chen Yi-chun. Zero anaphora resolution in Chinese with shallow parsing[J]. Journal of Chinese Language and Computing, 2004

[8] Yeh C-L, Chen Yi-jun. An Empirical study of zero Anaphora Resolution in chinese Based on Centering Model[Z]. 2010

[9] Yang Ya-qin, Xue Nian-wen. Chasing the ghost: recovering empty categories in the Chinese Treebank[C]// Coling2010. 2010: 1382-1390

[10] 朱勤宇. 汉语零形回指的句法驱动力[J]. 汉语学习, 2002(4): 73-80

[11] Kong Fang, Zhou Guo-dong. A Tree Kernel-based Unified Framework for Chinese Zero Anaphora Resolution[C]// Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. 2010:882-891

[12] Zhao Shan-heng, Ng H T. Identification and Resolution of Chinese Zero Pronouns: A Machine Learning Approach[C]// Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2007:541-550

[13] Yeh C-L, Mellish C. An Empirical Study on the Generation of Anaphora in Chinese[J]. Association for Computational Linguistics, 1997, 23(1):171-190

[14] 杨国庆，孔芳，朱巧明，等. 基于规则的中文缺省识别研究[J]. 计算机科学，2011(12):255-257

(上接第253页)

[4] Cheng Jian, Druzdzel M J. Ais-bn: An adaptive importance sampling algorithm for evidential reasoning in large bayesian networks [J]. Journal on Artificial Intelligence, 2000, 13:155-188

[5] 樊兴华，张勤，孙茂松，等. 多值因果图的推理算法研究[J]. 计算机学报，2003,26(3):310-322

[6] 张勤. DUCG：一种新的动态不确定因果知识的表达和推理方法(I)：离散、静态、证据确定和有向无环图情况[J]. 计算机学报, 2010,33(4):625-651

[7] Zhang Qin, An Xue-gao, Jin Gu, et al. Application of FBOLES-a prototype expert system for fault diagnosis in nuclear power plants[J]. Reliability Engineering and System Safety, 1991, 34: 225-235

[8] Zhang Qin. Dynamic Uncertain Causality Graph for Knowledge Representation and Reasoning: Discrete DAG Cases[J]. Journal of Computer Science and Technology, 2012, 27(1):1-23