

基于树核函数的中文零指代项识别研究

黄李伟 孔 芳 朱巧明 周国栋

(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 介绍了一种基于树核函数的零指代项识别方法,并给出了具体的实现系统。为了能够裁剪出包含零指代项的句法结构树,首先对 CTB 的语料进行了相关的人工标注,构建了一个基准语料库;然后,提出了 3 种不同的裁剪策略,获得了零指代项的结构化信息,生成了零指代项分类器;最后,构建了一个基于规则的原型系统,作为研究内容的实验对象。实验结果显示,提出的基于树核函数的零指代项识别方法获得的识别率明显优于基于规则的原型系统。

关键词 零指代,零指代项识别,树核函数,机器学习

中图法分类号 TP18

文献标识码 A

Tree Kernel-based Zero Pronouns Recognition

HUANG Li-wei KONG Fang ZHU Qiao-ming ZHOU Guo-dong

(School of Computer Science and Technology of Soochow University, Suzhou 215006, China)

(Key Lab of Computer Information Processing Technology of Jiangsu Province, Suzhou 215006, China)

Abstract This paper introduced a learning approach of Tree Kernel-based Zero Pronouns Recognition, and presented a system based-upon this approach. In order to get the syntax structure tree, which includes the zero pronoun, it marks the sentences of the CTB manually, and constructs a base-corpus. Then, it proposed three different policies to get syntax structure tree, which can be trained by SVM as a feature. At last, it also presented a baseline system based on rules, with which we can compares. Experimental results show that the recognition rate of the approach based on tree kernel is better than the one of the System based on rules.

Keywords Zero anaphora, Zero pronoun recognition, Tree kernel, Machine learning

1 引言

指代是指在篇章中用一个指代词回指某个以前说过的语言单位。在语言学中,指代词称为照应语,所指的对象或内容称为先行语。指代消解就是确定照应语和先行语之间相互关系的过程,是自然语言处理的关键问题之一。指代消解在文本段落的理解中起到了相当重要的作用,如信息抽取、机器翻译等。

在自然语言中,读者能够根据上下文的关系推断出来的部分经常会被省略,被省略的部分在句子中承担句子的句法成分,并且回指前文中说过的语言单位,这个现象称为零指代。Zhao 和 Ng(2007)^[1]指出,英文中句法结构比较严谨,代词和连词用得比较多,对于前文中出现过的名词短语、主题等等,在后文中一般都代词或名词短语表示出来。省略句法结构的某个部分的现象在英文中很少见,大概不超过 4%。而对于中文等亚洲文字而言,省略句法结构中某个部分的现象高达 36%^[1]。这项研究表明了中文中出现零指代的现象是很普遍的。由于零指代现象的普遍存在性,使得中文在很

多领域的研究都困难重重。比如,在机器翻译领域中,在无法知道省略部分所代表意思的时候,无法将中文翻译成目标语言等等。因此,中文零指代的研究是中文指代消解研究的一个重要的子课题,具有非常重要的意义。

零指代消解工作,一般可以分为 2 个阶段:零指代项的识别和零指代项的消解。本文实现了零指代第一阶段的主要工作:首先对 CTB 的语料进行相关的人工标注,构建了一个基准语料库;然后,提出了 3 种不同的裁剪策略,获得了零指代项的结构化信息,生成了零指代项分类器;最后,构建了一个基于规则的原型系统,作为本文研究内容的实验对象。本文第 2 节讲述了零指代消解研究的相关工作;第 3 节主要介绍了本文所研究零指代项识别的框架结构以及详细介绍;第 4 节主要分析了本文实验的结果;最后总结全文。

2 相关工作

中文零指代消解工作近年来的代表性工作如下:Ferrandez(2000)等^[2]提出了基于手工标注规则的方法完成了西班牙语主语位置的零指代项的识别工作,这种方法在中文零指

到稿日期:2010-03-03 返修日期:2010-05-20 本文受国家自然科学基金(60873150,60970056),江苏省自然科学基金(BK2008160)和江苏省高校自然科学重大基础研究项目(08KJA520002)资助。

黄李伟(1985—),男,硕士,主要研究方向为自然语言处理,E-mail:20074227064005@suda.edu.cn;孔 芳(1977—),女,博士,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、网络挖掘;周国栋(1967—),男,教授,博士生导师,主要研究方向为自然语言处理。

策略,对人工标注有 ZP 和 NZP 的 CTB 语料进行句法树的裁剪,得到了包含有零指代项的句法树结构,将其标注为正例。而没有指代关系的、包含有 NZP 项的句法树结构标注为负例。其中负例包含 2 种情况:一种是句子中并没有省略句法成分;另外一种是在句子中省略了句法成分,但是其前面并没有先行词。

构建测试样例:要通过简单的规则,将零形代名词的候选集找出来,规则如下:如果一个节点是 VP 短语,那么 VP 短语最左边的叶子节点的左边一个位置便作为零指代项的候选项,用 E 将其标识出来。然后将包含有 E 的子句法树裁剪下来,构成一个测试实例,交由分类器进行分类,这样会得到一个值。如果值大于 0,表示这个测试实例中的零形代名词候选项为零指代项,即其照应前面的一个或多个名词短语。

3.5 卷积核函数

本文所用的树核函数是卷积核函数。卷积核函数是通过计算两棵解析树之间的相同子树的数量来比较解析树之间的相似度。例如有两棵解析树 T_1 和 T_2 ,要计算相似度 $K_c(T_1, T_2)$:

$$K_c(T_1, T_2) = \sum_{n_1 \in N_1, n_2 \in N_2} \Delta(n_1, n_2)$$

式中, N_j 是 T_j 的节点集合, $\Delta(n_1, n_2)$ 计算以 n_1 和 n_2 为根的共同子树的个数,可以按照下面这种递归的计算方法:

(1) 如果 n_1 和 n_2 节点处的产生式不同,则 $\Delta(n_1, n_2) = 0$, 否则转向(2);

(2) 如果 n_1 和 n_2 都是叶子前的一个节点,则 $\Delta(n_1, n_2) = 1 \times \lambda$, 否则转向(3);

(3) 递归地计算 $\Delta(n_1, n_2)$:

$$\Delta(n_1, n_2) = \lambda \prod_{k=1}^{\#ch(n_1)} (1 + \Delta(ch(n_1, k), ch(n_2, k)))$$

式中, $\#ch(n_1)$ 是节点 n_1 的孩子节点数目, $ch(n, k)$ 是节点 n 的第 k 个孩子节点, $\lambda (0 < \lambda < 1)$ 是衰退因子。

实验中,采用了如表 1 所列的训练、测试实例。

表 1 树核训练、测试的实例

正例	1 BT (NP(NR(NN 工程)))(PU,)(E ZP)(VP(PP(P 由)(NP(NN 港商)))(VP(VP(VV 投资)))) ET
负例	-1 BT (IP(VP(E NZP)(NP(NT 去年)))(VP(VV 新增)))) ET

|BT| 与 |ET| 之间的内容是按照剪裁策略得到的包含零指代项自身及其周围相关的结构化信息。

4 实验及结果分析

本系统采用了国际上通用的 MUC 的评测系统进行评测。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零指代项的数目占实际上零指代项的数目,反映的是零指代项识别的完备性。准确率 P 是指识别出来正确的零指代项的数目占实际识别的零形代名词数目的百分比,反映的是指代消解系统的准确程度。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,定义如下:

$$F = \frac{(\beta + 1)P \times R}{(\beta \times P) + R}$$

式中, P 为准确率, R 为召回率, β 为召回率和准确率的相对权重,取 1。

为了验证本文的研究,借鉴了 Zhao 和 Ng (2007) 的做法,

即构建一个基于规则的原型系统作为比较对象,其 F 达到了 26.1。

本文参考了 Zhao 和 Ng^[1] 以及 Wu 和 Liang^[6] 基于规则的零指代消解系统的研究,借鉴了他们文献中用作比较的原型系统的规则,通过这些规则构建了一个基于规则的零指代项识别系统,以供本文研究比较。

规则 1 一个节点是 VP 节点,并且当它的父亲节点不是一个 VP 节点时,如果其兄弟左节点不是一个 NP 节点,或者其没有兄弟左节点,那么该节点的左边位置便是零指代项。

规则 2 一个节点是 VP 节点,并且当它的父亲节点不是一个 VP 节点的时候,如果其所扩展的节点中包含一个及物动词,且其兄弟右节点不存在,那么 VP 节点右边位置便是一个零指代项;如果其所扩展的节点中包含的是一个不及物动词,且其兄弟右节点不存在,那么该 VP 节点右边位置不是一个零指代项。

规则 3 一个节点是介词短语,并且当它的父亲节点不是一个介词短语的时候,如果该节点的兄弟左节点不存在,那么该节点的左边位置便是零指代项。

所标注的语料库的具体数据如表 2 所列。

表 2 语料库数据

训练样例		测试样例	
536		278	
正例	负例	正例	负例
320	216	161	117

根据以上 3 条规则分别对其进行实验,将规则 1 和规则 2 归成一类进行实验,然后将规则 1, 2, 3 放在一类进行实验,数据结果如表 3 所列。

表 3 规则 base-line 实验结果

规则	基于规则的方法				
	Zp_true	Zp_false	P	R	F
1, 2	32	108	22.86	27.83	25.10
1, 2, 3	35	130	19.75	30.43	23.95

表 3 中, Zp_true 是指正确识别出来零指代项的个数, Zp_false 是指错误识别出来零指代项的个数。本文将上述 3 条规则以表达式的形式展现出来:

1) $\langle \text{Zero}, P, O \rangle: \text{vtp}(P), \text{np}(0)$;

2) $\langle S, P, \text{Zero} \rangle: \text{np}(s), \text{vtp}(P)$;

3) $\langle \text{Zero}, P, O \rangle: \text{prep}(p), \text{np}(0)$ 。

其中, $\text{vtp}(P)$ 是动词短语做谓语, $\text{prep}(p)$ 是介词短语做谓语, np 名词短语既可做主语又可做宾语。从表 3 的实验数据可以看出,就规则 1, 2 而言,判断零指代项位置是由 vtp 驱动的。而规则 3 中是由 prep 驱动的,比较两次实验数据,在正确识别个数提高 3 个的情况下,错误识别的个数提高了 22 个。可以看出,在本基准语料库中,零指代项由 vtp 驱动出现的可能性比由 prep 驱动出现的可能性要大,所以本文基于树核函数的方法主要针对的是有 vtp 驱动的零指代项出现的情况。

所研究的基于树核函数的零指代项识别系统的结果,以及基于规则 1, 2, 3 的 base-line 的实验结果如表 4 所列。

表 4 比较实验结果

	P(%)	R(%)	F
规则方法	19.75	30.43	23.95
树核方法	57.3	90.1	70.05

(下转第 250 页)

β -三枝决策、 $(1,0)$ -三枝决策和 γ -二枝决策各自犯错的详细情况,给出 (α,β) -三枝决策优于其他两者的成立条件。最后,利用贝叶斯理论来诠释决策粗集三枝决策方法在实际问题中的应用过程。相对于其他概率粗集理论,决策粗集理论具有其深刻的研究背景和独有的语义解释。可以预见,决策粗集理论不仅能够给予决策问题以理论指导,而且能在实际应用中发挥越来越重要的作用。

参 考 文 献

- [1] Pawlak Z. Rough sets [J]. International Journal of Computer and Information Sciences, 1982, 11: 341-356
- [2] Yao Y Y. Three-way decision: an interpretation of rules in rough set theory [J]. LNAI, 2009(5589): 642-649
- [3] Yao Y Y. Three-way decisions with probabilistic rough sets [J]. Information Sciences, 2010, 180: 341-353
- [4] Yao Y Y. Two semantic issues in a probabilistic rough set model [J]. Fundamenta Informaticae, Manuscript, 2009
- [5] Pawlak Z, Wong S K M, Ziarko W. Rough sets: probabilistic versus deterministic approach [J]. Inter. Journal of Man-Machine Studies, 1988, 29: 81-95
- [6] Yao Y Y, Wong S K M. A decision theoretic framework for approximating concepts [J]. Inter. Journal of Man-machine Studies, 1992, 37: 793-809
- [7] Yao Y Y. Decision-theoretic rough set models [J]. Lecture Notes in Artificial Intelligence, 2007(4481): 1-12
- [8] Ziarko W. Variable precision rough set model [J]. Journal of Computer and System Sciences, 1993, 46: 39-59
- [9] Slezak D. Rough sets and Bayes factor [J]. LNCS Transactions on Rough Sets III, 2005: 202-229
- [10] Slezak D, Ziarko W. The investigation of the Bayesian rough set

model [J]. International Journal of Approximate Reasoning, 2005, 40: 81-91

- [11] Yao Y Y. Probabilistic approaches to rough sets [J]. Expert Systems, 2003, 20: 287-29
- [12] Yao Y Y. Probabilistic rough set approximations [J]. International Journal of Approximate Reasoning, 2008, 49: 255-271
- [13] Ziarko W. Probabilistic approach to rough set [J]. International Journal of Approximate Reasoning, 2008, 49: 272-284
- [14] Duda R, Hart P. Pattern Classification and Scene Analysis [M]. New York: Wiley Press, 1973
- [15] Yao Y Y. The superiority of three-way decision in probabilistic rough set models [J]. Information Sciences, Manuscript, 2010
- [16] Pauker S, Kassirer J. The threshold approach to clinical decision making [J]. The New England Journal of Medicine, 1980, 302: 1109-1117
- [17] Woodward P, Naylor J. An application of Bayesian methods in SPC [J]. The Statistician, 1993, 42: 461-469
- [18] Li Y, Zhang C, Swan J. An information filtering model on the Web and its application in JobAgent [J]. Knowledge-Based Systems, 2000, 13: 285-296
- [19] 赵文清, 朱永利, 高伟. 一个基于决策粗糙集理论的信息过滤模型 [J]. 计算机工程与应用, 2007, 43(7): 185-187
- [20] Zhou B, Yao Y Y, Luo J. A Three-way decision approach to email spam filtering [C]// The 23th Canadian AI, 2010: 28-39
- [21] Macmillan F. Risk, Uncertainty and Investment Decision-making in the Upstream Oil and Gas Industry [D]. UK: University of Aberdeen, 2000
- [22] Yusgiantoro P, Hsiao F. Production-sharing contracts and decision making in oil production [J]. Energy Economics, 1993, 10: 245-256

(上接第 216 页)

从表 4 可以看出,本系统的召回率较高,说明本系统对于应该识别的零指代项的识别率较高,且完备性较好,但准确率只有 57.3%,说明本系统对于不具有零指代关系的零形代名词的识别率也较高。从表 2 中可以看出,训练和测试的时候正负例的选取比较平衡。从实验数据可以看出,本文所研究的基于树核函数零指代项识别的正确率较基于规则系统的正确率要高,说明本文研究的结构化信息在零指代项识别这个阶段具有一定的研究价值。

结束语 目前,中文零指代消解是中文指代消解研究领域一个热点课题,作为零指代消解第一阶段的工作——零指代项识别是必不可少的。本文参考 Zhao 和 Ng(2007)的基于机器学习的零指代消解方法,实现了一个基于树核函数的中文零指代项识别系统。为了能够获得包含零形代名词的结构化信息,进行了语料人工标注,构建了一个基准语料库,并通过各种裁剪策略得到包含零形代名词的句法分析树,将其作为结构化特征交由 SVM 进行训练。同时本文实现了一个基于规则的零形指代项识别的系统。通过实验结果比较,本文所研究的基于树核函数的零指代消解具有一定的价值。

参 考 文 献

- [1] Zhao Shanheng, Ng H T. Identification and Resolution of Chinese Zero Pronouns: A Machine Learning Approach [C]// ACL, 2007: 541-550
- [2] Ferrández A, Peral J. A computational approach to zero-pronouns in Spanish [C]// ACL, 2000: 166-172

- [3] Yeh Ching-long, Chen Yi-chun. Zero anaphora resolution in Chinese with shallow parsing [J]. Journal of Chinese Language and Computing, 2004
- [4] Converse S. Pronominal Anaphora Resolution in Chinese [D]. Department of Computer and Information Science, University of Pennsylvania, 2006
- [5] 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
- [6] Wu Dian-song, Liang T. Zero anaphora resolution by case-based reasoning and pattern conceptualization [J]. Expert Systems with Applications, Expert Syst, 2008, 36(4): 7544-7551
- [7] Zhou G D, Kong F. Global Learning of Noun Phrase Anaphoricity in Coreference Resolution via Label Propagation [C]// EMNLP, 2009: 978-986
- [8] Yang X F, Su J, Tan C L. Kernel-based Pronoun Resolution with Structured Syntactic Knowledge [C]// ACL, 2006: 41-48
- [9] Yeh Ching-long, Chen Yi-chun. Using Zero Anaphora Resolution to Improve Text Categorization [C]// Proceeding of PA-CLIC17, Sentosa, Singapore
- [10] Kong F, Zhou G D, Zhu Q M. Employing the Centering Theory in Pronoun Resolution from the Semantic Perspective [C]// EMNLP, 2009: 987-996
- [11] 张威, 周昌乐. 汉语语篇理解中元指代消解初步 [J]. 软件学报, 2002, 1(13): 732-738
- [12] 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