基于规则的中文缺省识别研究

杨国庆 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 中文语句中广泛存在缺省现象,缺省项识别的准确与否关系到缺省消解结果,因此对缺省项的识别很重要。介绍了一种基于规则的中文缺省项识别方法,即采用CTB语料构建基准语料库,以动词驱动为核心提出规则来获得缺省项的结构化信息。实验结果显示,基于规则的中文缺省项识别方法具有可行性。

关键词 缺省识别,规则,动词

中图法分类号 TP18 **文献标识码** A

Rule-based Ellipsis Identification in Chinese

YANG Guo-qing KONG Fang ZHU Qiao-ming LI Pei-feng

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

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

Abstract The phenomenon of ellipsis is widely existed in Chinese and the results of ellipsis resolution are directly impacted by correctness of the ellipsis identification. So ellipsis identification is very important. We introduced a learning approach of rule-base ellipsis identification in Chinese. That approach constructs a corpus-base by marking all sentences in CTB manually and then proposes a verb-driven method to extract rules to get syntax structure information. Experimental results shows that our method is feasible.

Keywords Ellipsis identification, Rule-based, Verbs

1 引言

缺省是任何语言中都存在的一种语言现象。随着语言学的发展,缺省得到了越来越多的重视。人们在特定的语言环境中,在不影响意思表达的前提下,为了使语言简洁明快,会省去某些语言成分,这种现象在语法中称作缺省。

缺省是指代的一部分。缺省识别将自然语言中的模糊成分具体化,并识别缺省成分在篇章中的性质,为指代消解和零指代提供基础。缺省识别是指代消解和零指代消解的一个重要子任务,缺省成分识别的正确与否直接影响消解结果。在自然语言处理研究中,缺省识别显得尤为重要。对中文缺省的研究,可促进对中文语法、语义、语用层面的认识,提高中文语篇理解。

当前关于中文缺省识别的研究并不多,主要原因是缺省的形式多种多样,缺省的定义和识别标准不统一,给自动化识别缺省带来了困难。本文针对上述问题进行研究,首先对CTB的语料进行了相关的整合调整,构建了一个基准语料库;然后提出规则,以获得缺省项的结构化信息;最后,构建了一个基于规则的中文缺省识别系统。

本文第2节介绍国内外相关领域的研究成果和现状;第3节介绍本文提出的规则以及中文缺省识别的流程;第4节主要分析实验结果;最后是结论和今后的工作展望。

2 相关工作

国内对中文的缺省研究大都停留在理论层面,没有给出系统、有效的识别和实现方法。李旺等[9]提出通过构造语篇表述框架(DRS)来进行缺省识别。殷鸿等[10]采用语法规则进行缺省识别。贾宁等[11]采用语义块信息进行缺省识别。上述所有研究都没有给出如何在句法或语义上缺省识别的具体实现,只是提出了缺省识别的可能方法。

英语的缺省受到语法形式的严格控制,必须在一定的语法规则下缺省。而中文的缺省,灵活多样,凭借语义或句子之间的逻辑关系,只要意义表达清楚,就可以缺省。中文主要研究名词短语(NP)缺省情况,英语主要研究动词短语(VP)缺省情况[1]。因此国外对缺省识别研究仅仅适用于英语。Hardt[2]提出了通过句法分析以及句法限制,利用搜索技术和语法标注来识别动词短语缺省。Yeh等[3]将规则方法应用到中文零指代消解的零指代项识别研究中。从某种程度上说,零指代项识别和缺省识别有一定相似的地方,因此缺省识别可以近似地看作是一种零指代项识别。

3 缺省项识别

为了排除不必要的噪音干扰,本文的测试文本是CTB提

到稿日期:2011-01-12 返修日期:2011-03-24 本文受国家自然科学基金(90920004,60970056,61070123,61003153)和江苏省高校自然科学重大基础研究项目(08KJA520002)资助。

杨国庆(1986—),女,硕士,主要研究方向为自然语言处理;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;李培峰(1971—),男,副教授,硕士生导师,主要研究方向为中文信息处理。

供的标准完全句法分析树。测试语料经过预处理，获取结构化信息特征，判断语句中是否有缺省项，最后得到测试结果。

3.1 缺省识别语料库及预处理

本文采用CTB语料库中的890篇文献，其中测试样例32052个。在CTB语料中，缺省的位置标记为"-NONE-X"，其中"X"代表不同类型[4]。例如当从句中无明显主语时，缺省项用"-NONE-*PRO*"标记。"-NONE-X"的具体分类和每个分类的出现频率如表1所列。

表1 "-NONE-"分类及频率

标记	描述	频率
-NONE-*T*	缺省的为主题或从句实施者	39.8%
-NONE-*	缺省在"把"字句、"被"字句	1.0%
-NONE-*PRO*	从句中无明显主语	34.2%
-NONE-*pro*	缺省的为主语或宾语	23.3%
-NONE-*RNR*	缺省成分为宾语	1.5%
-NONE-*?*	其他类型	0.2%

预处理部分包括中文分词、词性标注和句法分析等，以获得词语本身、词性以及词语与词语之间的句法关系。在CTB语料中，"WHNP(-NONE-*OP*)"代表的是关系子句中的一个实施者[4]。在宾州中文树库中，关系子句用来修饰中心名词短语，即在CP的下一层有一个假设的操作者，用"WHNP(-NONE-*OP*)"标记。但是在中文中，一般操作者本身在文中并没有相关的语言单位与缺省的部分有联系，所以在处理CTB语料时，"WHNP(-NONE-*OP*)"不在本文的研究范围内。

3.2 句法树

经过预处理后，用"NONE"来表示缺省项。如果"NONE"前面标注为"NP-SBJ"，则表示主语缺省；如果"NONE"前面标注为"NP-OBJ"，则表示宾语缺省。可以得到含有缺省标记的句法树，如图1所示。

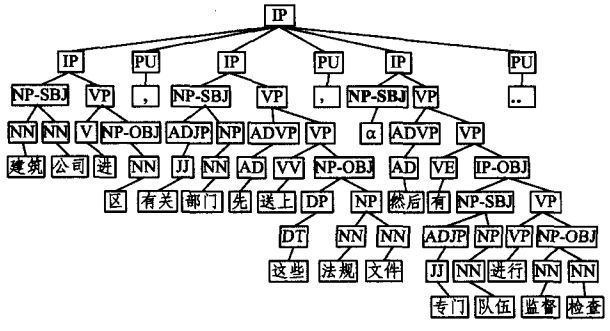

图1 含有缺省标记的句法树

其中，"α"表示此处有缺省成分，用"NONE"来表示。"α"结点的父结点"NP-SBJ"表示的是缺省成分在主语中的位置；同样地，当宾语缺省时，句法树中是用"NP-OBJ"来表示的。分析含有缺省成分的句法树，可以更好地提出相应的规则来进行缺省识别。

3.3 动词

在CTB语料中，根据缺省位置不同，统计出主语缺省频率是81.73%，宾语缺省频率为11.48%，其他情况的缺省情况为6.79%。根据统计结果可知，主语缺省和宾语缺省占缺省情况的93.21%，因此本文主要考虑主语缺省和宾语缺省。

英语是以主语为主导的语言。英语的主语是重要的语法结构，一般情况是必不可少的。识别英语主语缺省主要是以动词短语为驱动。同样地，中文主语或宾语缺省时，主要由动词短语为驱动[6]。因此，针对这一特点，本文主要讨论以动词驱动为核心的中文缺省项识别。

在CTB语料中，动词的分类[4]以及各种类型动词的频率如表2所列。

表2 动词分类以及频率

标记	描述	例子	频率
VA	表语形容词	好、勇敢	8.9%
VC	系动词	是、为	6.2%
VE	存在动词	有、没有	3.4%
BA	主控动词	把、将	0.9%
SB	被控动词	被、叫	0.5%
VV	其他动词	实行、喜欢	80.1%

由表2可知，在所有动词中，"VA"、"VC"、"VV"出现的总频率是95.2%，因此本文主要考虑的是"VA"、"VC"、"VV"。

在CTB语料中，连续两个或两个以上的动词出现时，连续的动词组合成复合动词[4]。出现复合动词时，把复合动词中的所有动词看作一个整体处理。在CTB语料中，复合动词的类型以及描述如表3所列。

表3 复合动词类型

标记	描述	例子
VCD	同步复合动词	回国就职、登记注册
VCP	VV+VC	看作是、想成是
VNV	V+不+V/V+一+V	能不能，看一看
VPT	V+的+V	打的赢
VRD	结果动词复合	降下来、表达出
VSB	修饰语+关系动词	仰头望去、驱车行程

在CTB语料中，复合动词也可以与复合动词组合，即为复合动词组合。出现复合动词组合时，把复合动词组合中的所有动词看作一个整体处理。但是，并不是所有的复合动词都可以与其他复合动词组合，只有满足彼此相容性，复合动词之间才能进行组合。复合动词彼此相容性[4]如表4所列。

表4 复合动词彼此相容性

	VCD	VRD	V-ASP	V+不+V	V+一+V
VCD	√	√	√	?	×
VRD	√	×	√	?	×
V-ASP	√	×	√	×	×
V+不+V	√	?	×	×	×
V+一+V	√	×	×	×	×

注："√"代表可以组合；"×"代表不能组合；"?"表示不确定；V-ASP表示动词+助词(比如了、着、过等)。

3.4 规则

Yeh等[3]提出三元组$T=\{S,P,O\}$，其中S表示语法成分是从句主语的名词列表；P表示语法成分是从句谓语的动词或介词列表；O表示语法成分是从句宾语的名词列表。根据三元组，提出三元规则如下：

Triple1(S,P,O) → np(S),vtp(P),np(O)

Triple2(S,P,N) → np(S),vip(P)

Triple3(S,P,O) → np(S),prep(P),np(O)

Triple4(S,N,N) → np(S)

其中，np(S)表示主语是名词短语，np(O)表示宾语是名词短语，vtp(P)表示谓语是及物动词，vip(P)表示谓语是不及物动词，prep(P)表示谓语是介词，N代表此处缺失某成分，Tri-

ple4 表示句子中除了一个名词短语没有其他的成分。

英汉句子基本结构均是主语+谓语+宾语[8]。对每个句子进行抽取后可以得到句子的基本结构，即主谓宾结构。得到句子基本结构后，可以看出主语或宾语是否缺省，进而找到缺省位置。综上所述，Yeh 等提出的三元规则基于句子结构，适用于中文缺省识别。本文参考 Yeh 等提出的方法，以动词为核心，提出缺省三元规则，具体如下：

$Triple1n1(none,P,O) \rightarrow vtp(P),np(O)$

$Triple1n2(S,P,none) \rightarrow np(S),vtp(P)$

$Triple1n3(none,P,none) \rightarrow vtp(P)$

$Triple2n1(none,P,none) \rightarrow vip(P)$

其中，$np(S)$表示主语是名词短语，$np(O)$表示宾语是名词短语，$vtp(P)$表示谓语是及物动词，$vip(P)$表示谓语是不及物动词，*none* 代表缺省项。

本文研究的缺省成分主要是在主语和宾语位置；规则 *Triple1n1* 表示缺省成分在主语位置；规则 *Triple1n2* 表示缺省成分在宾语位置。*Triple1n3* 和 *Triple2n1* 表示缺省成分在主语位置和宾语位置。例如在表 5 中，“NP-SBJ NONE”是根据缺省三元规则中的 *Triple1n1* 规则来表示主语位置有缺省。

表 5　缺省三元规则样例

建筑 公司 进 区，有关 部门 先 送上 这些 法规性 文件，然后 有 专门 队伍 进行 监督 检查 。
→[[[建筑],[公司],[进],[区]],,[[有关],[部门],[先],[送上],[这些],[法规性],[文件],,[[NP-SBJ],[NONE],[然后],[有],[专门],[队伍],[进行],[监督],[检查]]] 。

根据缺省三元规则，缺省识别具体流程如图 2 所示。

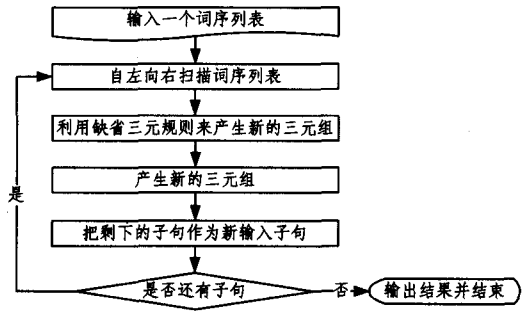

图 2　缺省三元规则流程图

4　实验及结果分析

本系统采用了国际上通用的 MUC 评测系统进行评测。MUC 有 3 个评价指标：召回率 R(Recall)、准确率 P(Precision)和 F 值(F-measure)。召回率是指识别出来的正确缺省项的数目占实际上缺省项数目的百分比，反映的是缺省项识别系统的完备性。准确率是指识别出来的正确缺省项的数目占实际识别的缺省项数目的百分比，反映的是缺省项识别系统的准确程度。当比较两个不同系统的性能时，一般使用这两个指标的综合值，即 F 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，权重取 1。

根据本文提出的规则进行实验，实验结果如表 6 所列。

表 6　基于规则缺省识别实验结果

P(%)	R(%)	F(%)
70.2	68.5	69.3

根据表 6 中实验结果可以看出，得到的准确率并不是很高，是由以下方面造成的：一是本文主要以动词驱动为核心进行缺省项识别，但是当缺省项充当子句的主语时，由动词短语和介词短语作为驱动，本文并未考虑以介词短语为驱动；二是当以动词驱动为核心进行缺省项识别时，语料库中的动词形式千变万化，并不能考虑周全。比如当动词作为名词使用时，语料库中仍是以动词的形式出现，这样的动词也作为本文待处理的动词形式，但实际上动词已作为名词，不应以动词形式来进行处理。这部分原因占主要方面。针对上述原因，在后续工作中应考虑以机器学习方法进行缺省项识别，并对其进行改进。

根据“NONE”类型的不同，统计每种类型“NONE”的召回率，如表 7 所列。

表 7　不同类型“NONE”召回率

标记	总数	正确识别数	R(%)
-NONE-*T*	12759	7272	57.0
-NONE-*	324	98	30.2
-NONE-*PRO*	10948	8860	80.9
-NONE-*pro*	7479	4910	65.7
-NONE-*RNR*	478	200	41.8

根据表 7 可以看出，“-NONE-*PRO*”的召回率最高，达到 80.9%，主要因为当语句从句中无明显主语时，用“-NONE-*PRO*”标记，而主语缺省占整个缺省识别工作的 81.73%，所以缺省识别的 Recall 很高。“-NONE-*pro*”的召回率次之，主要因为本文主要研究的是主语缺省和宾语缺省，缺省为主语或宾语时，用“-NONE-*pro*”标记。“-NONE-*PRO*”和“-NONE-*pro*”的召回率均不低，证明中文语句缺省中主语缺省和宾语缺省为主要缺省情况。

结束语　目前，中文缺省识别是自然语言处理领域中的一个难点，作为缺省消解的第一阶段的工作，缺省识别是必不可少的。本文参考 Yeh 等介绍的三元规则方法，提出了缺省三元规则来进行缺省识别。实验结果显示，基于规则的缺省识别方法具有可行性。在后续工作中，我们会寻求更加合理的方法来改进本文所提出的缺省识别模型，同时提出相应的特征，运用机器学习方法进行缺省识别。本文提出的基于规则的缺省识别工作可以作为一个基准平台，为后续工作提供比较对象。

参 考 文 献

[1] Nielsen L A. A corpus-based study of verb phrase ellipsis[C]// Proceedings of the Annual CLUK Research Colloquium. London, British, 2003: 109-115

[2] Hardt D. VP Ellipsis: Form, Meaning, and Processing[D]. USA: University of Pennsylvania, 1993

[3] Long Y-C, Chen Yi-chun. Zero anaphora resolution in Chinese with shallow parsing[J]. Journal of Chinese Language and Computing, 2004, 17(1): 41-56

[4] Xue Nian-wen, Xia Fei, Huang Shi-zhe, et al. The bracketing guidelines for the Penn Chinese Treebank(3.0)[R]. IRCS Report 00-08. USA: University of Pennsylvania, 2000

(下转第 273 页)

for Rotocraft Low altitude Flight[J]. IEEE Transaction on System, Man, and Cybernetics, 1992, 22(2): 290-299

[2] Yakimenko O A, Kaminer I I, et al. Unmanned aircraft navigation for shipboard landing using infrared vision[J]. IEEE Transactions on Aerospace and Electronic Systems, 2001, 38(4): 1181-1200

[3] Johnson A E, Matties L. Precise Image-based Motion Estimation for Autonomous Small Body Exploration[C]// Proc. 5th Int' l Symp. Artificial Intelligence, Robotic, and Automation in Space (iSAIRA 99). JAXA, 1999: 627-634

[4] Kawaguchi J, Uesugi T, Fujiwara A, et al. The Muses-C, World's first sample and return mission from a near earth asteroid Nereus[C]// 2nd Int'l Conf. Low-cost Planetary Missions. Space Age, 1996: 15-23

[5] Cheng Yang, Johnson A, Matthies L. MER-DIMES: a planetary landing application of computer vision[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2005, 1: 806-813

[6] 石德乐，叶培建. 月球着陆器精确定点及安全着陆技术研究[J]. 航天器工程，2007，16(3)：9-16

[7] 吴伟，周金鹏，秦石乔，等. 光学相关识别在飞行器精确定点着陆中的应用[J]. 光电子·激光，2008，19(12)：1653-1655

[8] 田阳，崔平远，崔祜涛. 基于图像的着陆点评估及探测器运动估计方法[J]. 宇航学报，2010，31(1)：98-102

[9] 翟冬丽. 基于图像序列的小行星软着陆导航方法研究[D]. 哈尔滨：哈尔滨工业大学，2006

[10] Ming Jie, Huang Xian-lin. The Vision Navigation Based on Lunar Surface Control Point Registration[C]// Proceedings of the 25th Chinese Control Conference. Beihang University Press, 2006: 2236-2239

[11] Zitova B, Flusser G. Image registration methods: a survey[J]. Image and vision Computing, 2003, 21(11): 997-1000

[12] Harris C, Stevens M. A Combined Corner and Edge Detector[C]// Proc 4th Alvey Vision Conf. BMVA. 1988: 147-151

[13] Lowe D G. Object recognition from local scale-invariant features [C]// Proceedings of International Conference on Computer Vision. IEEE, 1999, 2: 1150-1157

[14] Cheng S, Stankovic V, Stankovic L. Improved sift-based image registration using belief propagation[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2009: 2909-2912

[15] Allaire S, Kim J J, Breen S L, et al. Full orientation invariance and improved feature selectivity of 3D SIFT with application to medical image analysis[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 2008

[16] Guo Xiao-jie, Cao Xiao-chun. Triangle-constraint for Finding More Good Features[C]// 20th International Conference on Pattern Recognition(ICPR). IEEE, 2010: 1393-1396

[17] Guo Xiao-jie, Cao Xiao-chun. FIND: A Neat Flip Invariant Descriptor[C]// 20th International Conference on Pattern Recognition(ICPR). IEEE, 2010: 515-518

[18] 袁修国，彭国华，王琳. 基于 GPU 的变型 SIFT 算子实时图像配准[J]. 计算机科学，2011，38(3)：300-303

[19] Bay H, Tuytelaars T, Gool L V. SURF: Speeded Up Robust Features[C] // European Conference on Computer Vision. Springer, 2006: 404-417

[20] Fischler M A, Bolles R C. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography[J]. Comm. of the ACM, 1981, 24(6): 381-395

[21] 殷飞，桑农，王洗. 一种新的序列图像匹配定位算法[J]. 红外与激光工程，2001，30(6)：422-425

[22] Tang Chun-ming, Dong Yan, Su Xiao-hong. Automatic Registration Based on Improved SIFT for Medical Microscopic Sequence Image[C]// Second International Symposium on Intelligent Information Technology Application. IEEE, 2008: 580-583

[23] Li Ming, Jiang Yu-gang, Liu Ya-dong, et al. Registration of intra-operative optical image sequence, Fifth International Conference on Photonics and Imaging in Biology and Medicine. SPIE, 2007, 6534: 65340L-1-65340L-7

[24] Lin Yu-ping, Medioni G. Map-enhanced UAV Image Sequence Registration and Synchronization of Multiple Image Sequences [C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2007: 1-7

(上接第 257 页)

[5] Zhao Shan-heng, Hwee T N. 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

[6] Li Liu. A contrastive analysis of the ellipsis rules of chinese and english language[J]. US-China Foreign Language, 2004, 2(2): 57-63

[7] 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

[8] 单文波. 试论英汉句子结构的差异[J]. 江汉大学学报，2001，18(1)：74-77

[9] 李旺，李绍滋. 基于 DRT 理论的汉语省略恢复研究[J]. 计算机工程，2004，30(17)：39-41

[10] 殷鸿，许威，赵克，等. 基于概念模型的省略恢复研究[J]. 计算机工程，2007，33(22)：229-231

[11] 贾宁，张全. 基于句间关系的汉语语义块省略恢复[J]. 中文信息学报，2008，22(6)：33-37