第 38 卷 第 22 期
Vol.38 No.22

计 算 机 工 程
Computer Engineering

2012 年 11 月
November 2012

·人工智能及识别技术· 文章编号：1000—3428(2012)22—0130—03 文献标识码：A 中图分类号：TP391

用于中文缺省识别研究的机器学习方法

秦凯伟[1,2]，孔 芳[1,2]，李培峰[1,2]，朱巧明[1,2]，徐生芹[1,2]

(1. 苏州大学计算机科学与技术学院，江苏 苏州 215006；2. 江苏省计算机信息处理技术重点实验室，江苏 苏州 215006)

摘 要：实现一个基于机器学习的中文缺省项识别系统，对语料库进行预处理，选取多个特征及其组合，通过支持向量模型(SVM)构建的缺省识别模型进行中文缺省识别。研究系统在不同句法分析树上的性能。实验结果证明，该识别系统在标准的句法分析树上 F 值能达到 84.01%，在自动句法树上能达到 68.22%。

关键词：缺省；自然语言处理；句法分析树；机器学习；语料；缺省识别

Machine Learning Approach for Chinese Ellipsis Identification Study

QIN Kai-wei[1,2], KONG Fang[1,2], LI Pei-feng[1,2], ZHU Qiao-ming[1,2], XU Sheng-qin[1,2]

(1. School of Computer Science & Technology, Soochow University, Suzhou 215006, China;
2. Key Lab of Computer Information Processing Technology of Jiangsu Province, Suzhou 215006, China)

【Abstract】 This paper presents a system for ellipsis identification in Chinese which is based on machine learning. The system can be used to select a number of features and feature combinations through preprocessing the corpus. And Chinese ellipsis identification can also be achieved by the ellipsis identification model built by Support Vector Machine(SVM). The performance of the system in different parser tree is studied as well. Experimental result shows that the system has F value of 84.01% on the standard parser tree, and 68.22% on automatic sentence parser tree.

【Key words】 ellipsis; natural language processing; sentence parse tree; machine learning; corpus; ellipsis identification

DOI: 10.3969/j.issn.1000-3428.2012.22.032

1 概述

在不影响意思表达的情况下，通常为了语言的简洁明了会省略部分语言成分，这种现象称为缺省。缺省是一种常见的语言现象，而这种现象在汉语中更加普遍。国内外对于缺省的研究起步比较早，但大多数的研究都只停留在理论层面，各个语言学派对于缺省的定义又各自不同。文献[1]认为：所谓缺省是指语法、语义作用，但不具有语音形式的语言成分。

缺省项研究的主要内容是找出句子中存在缺省的位置，并补出相应的语言成分。缺省项的研究是自然语言处理中非常重要的一部分，缺省项的识别和恢复的正确率，将直接影响到自然语言处理中其他研究的性能。

缺省项的识别是整个缺省项研究的基础，其识别正确与否直接影响后续工作。由于缺省项类别的多样性和不确定性，使得中文缺省项的识别困难较大。本文通过引入机器学习的方法，从人工智能角度解决缺省项识别的问题。

2 相关工作

早在 20 世纪 60 年代，国内就开始了对中文缺省的研究，限于当时的技术水平，中文缺省主要集中在理论研究上。到了 21 世纪，信息的高度发展使得中文缺省研究工作变得空前重要，对于中文缺省项的研究也成为了当前热点。目前国内外对于缺省项的识别研究比较多，文献[2]主要通过规则的方法识别 VPE(Verb Phrase Ellipsis)的缺省，通过对句法树的分析，专门针对 VPE 类型的缺省，提出针对性的规则，从实验结果可以看出其提出的方法非常有效，此外还提出了一种用动词来驱动识别的方法。

文献[3-4]也是通过规则的方法进行缺省项的识别，但是和文献[2]不同，文献[3-4]提出的是规则三元组 $T=\{S,P,O\}$，根据规则三元组进行缺省项的识别。从实验结果来看，该方法在其语料上具有很高的性能，但由于其语料是非公开语料，因此不具有可比性。文献[3-4]提出的规则为后来基于规则方法的研究提供了依据，文献[5]就是在其基础上对规则做了相应的改进。

目前规则的方法大都如上所述，而基于机器学习的缺省识别研究占据了重要的地位。文献[6]给出了一个使用机器学习方法进行中文零指代消解的方案，对于零指代识别

基金项目：国家自然科学基金资助项目(90920004, 60970056, 61070123, 61003153)；江苏省高校自然科学重大基础研究基金资助项目(08KJA520002)；苏州市科技计划基金资助项目(SYG201112)

作者简介：秦凯伟(1987－)，男，硕士研究生，主研方向：自然语言处理；孔 芳、李培峰，副教授；朱巧明，教授、博士生导师；徐生芹，硕士研究生

收稿日期：2012-02-29 **修回日期**：2012-03-20 **E-mail**：chinaqkw@yeah.net

问题只给出了简单的基于规则的处理，该文的研究表明，零元素识别性能对零指代消解的性能至关重要。文献[7]给出了使用机器学习方法进行空语类识别的一个完整方案，给出了一组语法和词性相关的特征集合，将空语类的识别过程看作一个类似于 POS(Part-of-Speech)标注的过程。该方案在完全正确的句法树上获得了很好的性能，但在自动句法树上得到的性能较差，并没有对空语类进行消解研究。

3　缺省识别

3.1　语料预处理

本文使用的语料是 OntoNote3.0 中的 120 篇文章，其中 100 篇是训练语料(0001~0099)，20 篇为测试语料(0100~0103,0109~0112,0118~0121,0127~0130,0136~0139)。首先对 100 篇训练语料进行统计，统计出各个分类在语料中的频率，如表 1 所示。

表 1　缺省类别在语料中的占比

类别	总数	频率/(%)
-NONE-*T*	742	31.48
-NONE-*pro*	446	18.92
-NONE-*PRO*	399	16.93
-NONE-*RNR*	44	1.87
-NONE-*OP*	722	30.63
其他	4	0.17

从表 1 可以看出类别“*T*”和“*OP*”所占的比重比较接近，再次观察语料发现类型“*T*”和类型“*OP*”大部分都是同时出现，如图 1 所示。所以在下文提出识别特征时，可以将识别类型“*T*”和“*OP*”合用。

(NP(CP(WHNP-1**(-NONE-*OP*)**)(CP(IP(NP-TPC**(-NONE-*T*-1)**)(IP(IP(NP-SBJ(NN 投资额)(VP(ADVP(AD 较)(VP(VA 大))))(PU 、)(IP(NP-SBJ(NN 技术))(VP(ADVP(AD 较))(VP(VA 高))))))(DEC 的)))(NP(NN 台资)(NN 企业))))

图 1　类型“*T*”和“*OP*”例子

3.2　句法分析树

OntoNotes3.0 语料库中提供了标准的句法分析树，但本文系统所采用的是最小 IP 子树，如图 2 所示。所谓的最小 IP 子树是标准句法分析树中各个叶子节点的最小 IP 父子树，通过这样的处理，就把整个句法分析树分成了若干小的 IP 子树，这样处理的好处是在特征抽取的时候能尽量减少噪音，提高系统的性能。

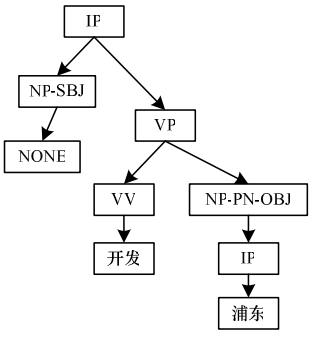

图 2　处理得到的最小 IP 树

3.3　特征介绍

基于机器学习方法的中文缺省项识别，最主要的是特征的抽取，特征选择的优劣将直接影响模型的识别性能。本文共选取了 5 个特征作为组合，考虑到 SVM 分类器对特征值表示方法的要求，本文全部将特征数字化。

本文的特征获取过程如下：首先输入整个句子的句法分析树，然后基于整个句子的句法分析树得到每个词的最小 IP 子树，最后在最小 IP 子树上进行特征抽取。标准的句法分析树由 OntoNotes3.0 语料库提供；自动的句法分析树，是由 Berkeley 句法分析工具对同样的句子进行处理后得到。

对于特征的正负例判断，本文通过标准的句法分析树进行匹配得到，如当前词在标准句法分析中存在缺省，则把当前词的特征标注为正例，即“+1”，否则为负例，标为“–1”。具体各个特征所代表的意义如下：

(1)1st-word-in-IP：判断当前词是否为 IP 子树的第 1 个左孩子节点，是为 1 否则为 0。

(2)In-NP：特征 1 为 1 的情况下，如果当前词的词性是 NP，“NP-SBJ”等以“NP”为开头的，则特征 2 为 0，否则为 1。

(3)In-V：特征 1 为 1 的情况下，如果当前词的词性是以“V”为开头的，则特征 3 为 1，否则为 0。

(4)1st-word-In-Verb：判断首单词是否为动词，如果为动词，则特征 4 为 1，否则特征 4 为 0。

(5)Has-Object：判断当前词是否为及物动词，如果是及物动词，判断当前词的右子树词性中是否包含“NP-OBJ”，如果含有则为 0，否则为 1，如果是不及物动词为 0。

以上就是本文识别系统用的 5 个特征。根据上述的特征提取方法，对图 2 中的词“开发”进行特征提取，需要说明的是在特征提取的过程中，标准语料中的缺省信息是被忽略的，所以根据提取特征的判断规则，可以得到“开发”的特征为“+1 1:1 2:1 3:1 4:1 5:0”。

4　实验结果与分析

本文采用了国际上通用的 MUC 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 2 个不同指代系统的性能时，一般使用这 2 个指标的综合值：F 值，即式(3)。

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

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

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

应用本文提出的特征提取规则，在标准的句法分析树上构建出训练样例，由 SVM 生成分类模型，并在测试集上进行测试，最终实验结果如表 2、表 3 所示。

表 2 标准句法分析树上的结果 (%)

方法	召回率	准确率	*F* 值
本文方法	73.48	98.05	84.01
文献[7]方法	83.00	95.90	89.00

表 3 自动句法分析树上的结果 (%)

方法	召回率	准确率	*F* 值
本文方法	56.66	85.71	68.22
文献[7]方法	52.10	80.30	63.20

从表 2 可以看出，本文选取的特征集在标准的句法分析树上具有很好的性能，*F* 值能达到 84.01，本文在标准句法树上的性能没有文献[7]方法的高，但是从表 3 可以看出，本文提出的系统在自动句法树上的性能要比文献[7]的方法要好，*F* 值比文献[7]的方法高出 5%。在实际应用中一般都使用自动句法分析树。表 2 和表 3 只能从整体性能上说明本文提出的特征具有比较好的性能。对于针对具体的缺省类别，也做了实验，从表 4 和表 5 可以看出各个类型的缺省项具体的识别率。

表 4 标准句法分析树各缺省类别的识别结果

类别	总数	正确识别数	召回率/(%)
-NONE-*T*	165	103	62.42
-NONE-*pro*	79	78	98.73
-NONE-*PRO*	95	93	97.89
-NONE-*RNR*	8	2	25.00
-NONE-*OP*	159	96	60.38

表 5 自动句法分析树各缺省类别的识别结果

类别	总数	正确识别数	召回率/(%)
-NONE-*T*	165	85	51.50
-NONE-*pro*	79	60	75.95
-NONE-*PRO*	95	75	78.95
-NONE-*RNR*	8	1	12.50
-NONE-*OP*	159	66	41.51

从表 4 和表 5 可以看出，本文的识别系统对于类型“pro”和类型“PRO”具有非常高的识别率，尤其在标准的句法分析树中，更是达到了 98.73%和 97.89%。在自动句法树下也有将近 76%和 79%。系统对于缺省类型是不进行识别的，也就是说系统只是逐个词逐个词地进行判断，如果当前词前面有缺省，识别系统就进行标记，而具体的类型并不进行识别。但从表 4 和表 5 可以发现，缺省类别“pro”和“PRO”比较容易识别，在不同的句法树下，这 2 个类型都具有很高的性能，原因是这 2 个缺省类型主要承担主语和宾语类型。而在句法分析树中，这个是基本的标注，所以识别的准确率比较高。

5 结束语

中文缺省项的识别研究，一直是一个相对较难的研究方向，不过由于近年来对中文缺省项研究的重视，使得中文缺省项识别研究得到了很大的发展。本文提出了基于机器学习的方法来识别缺省项，从实验结果可以看出，无论是在标准的句法分析还是自动句法分析树上，都具有很高的性能。但是，本文提出的方法在自动句法分析树的性能和标准方面还是有相当大的差距，在以后的研究中可以重点考虑解决这方面的问题。

参考文献

[1] 徐烈炯. 与空语类有关的一些汉语语法现象[J]. 中国语文, 1994, (5): 321-329.

[2] Nielsen L A. Verb Phrase Ellipsis Detection Using Automatically Parsed Text[C]//Proc. of the 20th International Conference on Computational Linguistics. Geneva, Switzerland: [s. n.], 2004.

[3] Yeh Ching-Long, Chen Yi-Chun. Zero Anaphora Resolution in Chinese with Shallow Parsing[J]. Journal of Chinese Language and Computing, 2004, 17(1): 41-56.

[4] Yeh Ching-Long, Chen Yi-Jun. An Empirical Study of Zero Anaphora Resolution in Chinese Based on Centering Model[C]// Proc. of Rocling'01. Taiwan, China: [s. n.], 2001.

[5] 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, 38(12): 255-257.

[6] Zhao Shanheng, Ng H T. Identification and Resolution of Chinese Zero Pronouns: A Machine Learning Approach[C]//Proc. of ACL'07. Prague, Czech: [s. n.], 2007.

[7] Yang Yaqin, Xue Nianwen. Chasing The Ghost: Recovering Empty Categories in The Chinese Treebank[C]//Proc. of Coling'10. Beijing, China: [s. n.], 2010.

编辑 顾逸斐