

高频最大交集型歧义切分字段在汉语自动分词中的作用*

作者姓名 审稿时不填写

作者单位 提交审稿时不填写

E-mail 提交审稿时不填写

摘 要: 交集型歧义切分字段是影响汉语自动分词系统精度的一个重要因素。本文引入了最大交集型歧义切分字段的概念,并将之区分为真、伪两种主要类型。最大交集型歧义切分字段的高频部分表现出相当强的覆盖能力及稳定性:前 4,619 个的覆盖率为 59.20%,且覆盖率受领域变化的影响不大。而其中 4,279 个为伪歧义型,覆盖率达 53.35%。根据以上分析,我们提出了一种基于记忆的处理策略,可有效改善实用型非受限汉语自动分词系统的精度。

关键词: 中文信息处理, 汉语自动分词, 高频最大交集型歧义切分字段, 基于记忆的排歧策略

The Role of High Frequent Maximal Crossing Ambiguities in Chinese Word Segmentation

Anonymous

Address line is not needed when submitting for review

E-mail is not needed when submitting for review

Abstract: The solution of crossing ambiguities is still an open issue in the study of Chinese word segmentation. In this paper, we introduce the concept of maximal crossing ambiguity at first, divide it further into two major types, i.e., the true and the pseudo. The high frequent part of maximal crossing ambiguities is strong in coverage capacity and rather stable with regard to domain shifting. As a consequence, we propose a memory-based strategy that is expected to improve the performance of practical Chinese word segmenters significantly.

Keywords: Chinese information processing, Chinese word segmentation, maximal crossing ambiguities with high frequency, memory based disambiguation strategy.

1 前言

信息化已成为二十一世纪全球不可抗拒的选择,成为国家经济与社会发展的命脉,成为一种新的控制财富的手段,这是来自用枪炮也无法阻挡的一种新的威胁与挑战,也是一种新的机遇。发展中的国家(包括中国在内)如果不重视这种挑战不抓住这个机遇,则发达国家有可能将信息技术作为新殖民化(信息殖民化)的有力武器。中文信息处理产业是否立得起来、立得好不好,关系到我国政治、经济、社会生活的变革,关系到我国在世界上的地位,甚至关系到我国的安全生存问题。信息处理主要是语言信息的处理,因此研究汉语语言信息处理的理论、方法、工具、资源,不仅是十分必要的而且迫在眉睫。当前,语言信息处理的竞争很大程度上取决于支撑的知识资源的竞争。

目前,世界上各国学者十分重视语言信息处理的知识资源的建设,知识包括词汇学知识、句法学知识、语义学知识、语用学知识乃至常识方面的知识,核心问题是语义学知识。相比而言,句法分析理论和技术(无论是对外语还是对汉语)发展得比较成熟和完善,语义学则是难度较大、起步较晚的一个薄弱环节,空白点更多。特别是面向机器处理的语义学研究,国内外起步时间均不长。汉语缺乏屈折变化,是意合语言是语义型语言,对语义的依赖更大,句法分析对句子的贡献比英语等语言要小,语义分析对汉语机器理解尤为重要。因此研究面向机器处理的汉语语义知识表示更具有重大意义。

根据框架语义学,格关系、槽关系和情态是句义的三大语言知识工程。格关系(论旨网格)描写的是论旨角色(格角色)与动词的语义关系,槽关系是研究论旨角色内的偏词和正词之间的语义关系,即动词框架的槽内的语义关系。研究动词的格框架关系是非常重要的,在此基础上研

*为保证匿名审稿,在提交审稿时,请从文章中暂时去除有关项目信息。

究名词的槽关系也是非常重要的，对于汉语尤其重要。这是因为印欧语着眼于谓词动词和时间，汉语既着眼于动词和时间，也着眼于名物和空间，因此名词的研究对汉语具有特别重要的意义。即：与以英语为代表的西方语言“以动词为中心”不同，汉语不仅是以动词为中心也是以名词为中心的（这是汉语重要特点之一），所以仅研究动词不研究名词难以满足汉语机器理解的要求。研究槽关系，其重点是研究名词、研究名词与其前面作修饰的定语的关系。

国内外对动词框架研究比较多，例如菲尔墨 1966 年和 1968 年就提出了格语法、格框架理论，1996 年又向美国基金会申请一项语言工程项目“框架网：基于框架语义学的英语语义词库”，计划研究 5000 个英语动词的语义词库（至今未见此项工程性项目研究成果公布于世）。日本对格语法的研究比较多，例如日本大修馆书店(1989.3.1)出版了小泉堡、船城道雄等编写的人读词典《日本语基本动词用法辞典》，其中包括日语 728 个基本动词词条的词形、读音、释义、用结合价（格框架）和语义分类表示的句例、例句和一些语法信息。国内的研究有清华大学陈群秀、黄昌宁和中国人民大学林杏光等研制的“现代汉语述语动词机器词典”，该机器词典已经描写了三千多个汉语常用动词的四千五百个义项的词形、拼音、词性、动词分类、论元数、义项数、义项序号、释义、论旨网格论旨模式的基本式、变换式、句例、论旨角色的语类、句法功能、语义限制、论旨实例、否定形式、时态、语义指向动词的后状以及论旨模式的扩展式等丰富的词法、句法、语义和语用信息。

但是国内外对名词的槽关系都很少有研究，至于工程性的描写更为罕见。因此对名词槽关系的系统性研究和工程性描写是一项开拓性工作。

2 名词槽关系研究的概况

名词槽关系研究在国内外都比较少见，汉语名词槽关系研究尤其是作为机器处理使用的槽关系研究尚未见到。汉语的句子中，名词短语（以名词为中心的名词词组）的分析和理解对句子的整体结构和语义理解有很大的影响，但汉语名词短语常常有着复杂的内部结构，即名词词组中修饰中心词名词的定语可能有多项，而且定语的语类以及定语与中心词名词的语义关系也非常复杂，因为汉语名词词组的定语的多样化以及定语与名词的语义关系的复杂性正是汉语的一个特点和机器处理的难点。例如：汉语名词修饰名词十分自由，有时加“的”，有时不加“的”，只要意义配搭得上，就可以直接粘合在一起组成偏正结构，甚至把一连串名词叠加在一起造成复杂的偏正结构。因此研究汉语名词槽关系和建立汉语名词槽关系系统的意义很大而且难度也很大，何况还没有研究先例和可以参照的外语的相应工作。汉语语言界研究名词性词组的定语多从定语出现的位置、定语的语类入手，对定语的划分也各不相同（例如，有的把定语划分为修饰性定语和限制性定语，也有的划分为描写性定语和限制性定语，还有的把定语划分为限定性定语、区别性定语和描写性定语），特别是无法揭示定语与中心语的语义关系是什么，而且一般也没有做到系统化，更无法形式化。本文作者研究名词槽关系在继承语言学的研究成果基础上又有新的思路，既从定语出现的位置、定语的语类出发研究定语，又采用定义 70 个槽类型新颖方法，从研究定语与中心词的语义关系的新的角度来研究汉语定语与名词中心词的关系，是寻求研究汉语名词定语的新突破口。论文作者认为，研究名词槽关系，关系在于研究和揭示定语与中心词的语义关系。在对定语与中心词的语义关系研究基础上，还需对大量名词进行工程性描写。目的在于研制一个

6 结语

通过几年来对名词槽关系的研究和槽类型的设计、应用，有以下的体会和问题。

第一，我们试图以计算词典学和传统词典学相结合的方法使名词槽关系的研究建立在丰富翔实的大量语言事实上，并试图以槽类型（语义关系）为主、语类（句法关系）为辅的方法来对名词作工程性的描写，目的在于为语言学工作者、计算语言学工作者的汉语研究、汉语机器理解研究提供丰富的语义信息；

第二，目前描述的有 3000 个名词，但作用远不止于这 3000 个名词。通过这 3000 个名词我们作了少量的统计和推理试验，发现可以用来推测新的尚未描述的名词的槽关系表达式联想。例如，已经描述了新娘、小偷、校长、总统，可以推测出部长的槽关系表达式联想，同时，若已描述了羊、牛、马，可以推测出驴的槽关系表达式联想；

第三，70 个槽类型的设立在使用过程中虽然也经过多次讨论、多次修改和填写验证，目前可能也还存在有不周全或填写时不好把握的问题，有待我们继续去修改验证；

第四, 名词槽关系系统工作单中尚有“论元数目”一项空缺, 这“论元数目”实际上是名词的配价, 有待于今后去研究和填补;

第五, 填写的 3000 个名词的分布性和典型性还有待于添加、补缺和平衡;

第六, 已录入名词槽关系系统的信息有待于进一步校对修改;

第七, 名词槽关系系统与动词机器词典、信息处理用现代汉语语义分类词典现在是互相独立的三个系统, 有待于将其整合成一个平台, 以利于互相支持、相互调用。

下面我们将进一步进行研究, 一是扩大描写名词数量及调整分布和补缺查漏, 二是对原有信息进行进一步校对修改, 三是将名词槽关系系统、动词机器词典、语义分类词典整合在同一平台上, 并进行机器学习的研究, 四是进一步验证和修改槽类型, 完善填写规范。

参 考 文 献

- [1] Chor B, Rivest RL. A knapsack-type public key cryptosystem based on arithmetic in finite fields. IEEE Transactions on Information Theory, 1988,34(5):901~909.
- [2] Dantchev S. Improved sorting-based procedure for integer programming. Mathematical Programming, Serial A, 2002,92:297~300.
- [3] Garlan D, Monroe R, Wile D. Acme: an architecture description interchange language. In: Johnson JH, ed. Proceedings of the CASCON'97: The 7th Annual IBM Centre for Advanced Studies Conference. 1997. 169~183.
- [4] <http://www-2.cs.cmu.edu/afs/cs/project/compose/ftp/pdf/acme-cascon97.pdf>
- [5] Luckham DC, Vera J. An event-based architecture definition language. IEEE Transactions on Software Engineering, 1995,21(9): 717~734.
- [6] 唐稚松,等.时序逻辑程序设计与软件工程.北京:科学出版社,2002.
- [7] 周莹新,艾波.软件体系结构建模研究.软件学报,1998,9(11):866~872.