文章编号:1005-0523(2009)06-0101-04

从自动分词角度看先秦与现代汉语词汇区别

徐紫云[1],徐雪松[2]

(华东交通大学 1.人文社会科学学院;2.电气与电子工程学院,江西 南昌 330013)

摘要:自动分词是古今汉语信息化所面临的共同课题。但古今汉语存在较大区别,现代汉语分词方法难以直接应用于古代汉语分词。为更好地借鉴现代汉语分词方法,探索符合先秦汉语的分词方法,从语言开放度、语言发展的表现及阶段、复音词的分布及构成、高频词的分布四个方面对先秦汉语与现代汉语的词汇进行了比较。发现先秦汉语既有区别于现代汉语的分词难点,又有独特的分词优势。

关 键 词:自动分词;先秦汉语;现代汉语;词汇;比较

中图分类号:H0-05 **文献标识码**:A

我国优秀传统文化多以古代汉语为载体,但随着时代发展,人们对古代语言文字越来越陌生,对浩如烟海的古籍渐渐敬而远之,实现古代汉语的计算机自动处理,对消除语言隔阂、传承中华文明具有重要意义。然而,古代汉语要实现计算机自动处理,面临的基础性问题就是自动分词。目前国内专家对现代汉语的分词技术已展开了广泛研究。但古代汉语与现代汉语无论是基本词汇还是基本语法都有较大区别,现代汉语分词方法难以直接应用于古代汉语。因此,有必要探索针对古代汉语实际的分词方法。

目前,对古代汉语现代化的研究多集中在电子文献的制作、电子语料库的建设、专家知识数据库的开发及研究专用软件的制作等方面。对古代汉语自动处理的相关研究还处于起步阶段。已有成果如《唐宋诗中词汇语义相似度的统计分析及应用》[1]《用于信息检索的古文统计分析》[2]等均未涉及分词这一基本课题,关于古代汉语分词研究的仅见《基于中文信息处理的古代汉语分词研究》等极少数论文[3]。

汉语发展有两条线,一条是由上古书面语直到白话文运动前的文言文,这条线无论是词还是句法都很大程度上保留了先秦语言的原貌;另一条是由先秦口语逐步发展直至现代汉语。由此可知,先秦是我国语言发生发展的重要时期,这一时期留下了丰富的文化遗产(尤其是先秦散文),在文学、语言及思想等方面对后世都产生了重要影响。因此,开展先秦散文自动分词研究,对实现古汉语自动处理有重要意义。而对先秦汉语与现代汉语词汇进行科学地比较,一方面可以更好地利用现代汉语分词方面所取得的成果开展先秦汉语分词研究;另一方面对现代汉语分词从语言本体方面找到突破口有一定启示意义。

1 先秦汉语与现代汉语开放度不同

现代汉语包括口语和书面语,具有鲜明的开放性特征,其词汇处于不断变化之中。其分词词表虽可以《现代汉语词典》为主体进行建设,但收词范围却不限于《现代汉语词典》,还涉及现代汉语普通话中通用的词和词组,包括专科词语、外来词语、新词语、文言词语、方言词等[4]。其分词过程中还会面临许多未登录词。这一现象也成为影响分词精度的一个重要因素。长期以来,研究人员一直把未登录词和分词歧义并列为影响分词精度的两大因素。而"由 Bakeoff 数据上的评估结果表明,未登录词造成的分词精度失落至少比分词歧义大5倍以上。"[5]因此,有不少研究人员选择绕开分词词表的方式进行分词。

先秦汉语是出现在典籍中相对封闭的语言,其词汇总量是有限的。计算机的统计优势可以将其穷尽性收入。分词词表一旦建成则相对稳定,未登录词对分词精度的影响不会像现代汉语中那么大。因而,分词词表对先秦汉语分词的意义较为重要。

收稿日期:2009-09-23

基金项目:江西省社会科学规划基金项目(06WX29);华东交通大学校立科研基金项目(06SKRW03)

作者简介:徐紫云(1973-),女,江西德兴人,硕士,副教授,研究方向为古代汉语、计算语言学。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

先秦汉语分词词表可参照现代汉语分词词表建设。关于词,至今没有令人信服的定义。对词的判定标准不同,结果自然存在差别。但从语言使用实际来看,除少数语言学家外,绝大多数人并不能准确区分语素、词与短语,却不影响其对语言的理解。同时,词汇复音化过程是个动态过程,部分短语与词的界限必然不会太明晰。因此,《信息处理用现代汉语分词词表》针对语言信息处理的需求所提出的是:"这个词表既要向根据语言学理念建立起来的词表尽量靠拢,同时又要与老百姓心目中'朦朦胧胧'但又确乎存在的'词表'尽量兼容。"[6]事实上,这一规范中分词单位既包括语言学规则认可的词,还包括部分短语、俗语及一些不可单说但可单用的"词素"。先秦汉语分词词表可借鉴现代汉语规范中对分词单位的收录原则。

但先秦汉语作为已过时的书面语言,我们是从古为今用的角度对其进行研究。语言虽是古人所说,却应为今人所理解。建立先秦汉语分词词典时,要充分考虑与现代汉语分词词表的对接。尤其对古今形式相同,但性质或意义存在差别的结构的判定,必须兼顾先秦语言运用实际和词汇发展情况,作具体分析。如"衣裳"、"东方未明、颠倒衣裳"(《诗经·齐风·东方未明》)中,"衣",为"上衣","裳"为"下衣",上衣与下衣穿颠倒了,"衣裳"显然为短语。但"衣裳"后来常连用泛指衣服,形成复音词,且意义与先秦短语有紧密的相承关系,因而,可以为一个分词单位收入。但"和平"却不能如此处理,在先秦汉语中,"和平"既可连用,如"血气和平"(《荀子》),亦可分用,如"既和且平"(《诗经》),应为短语,但不管连用还是分用,其意义与现代汉语"和平"一词均不同,因此,不能作分词单位收录。

2 先秦汉语与现代汉语发展的表现及阶段不同

任何时代的语言都处于发展变化之中,但发展的表现及阶段各不相同。

现代汉语发展的表现是出现众多难以预测的未登录词,先秦汉语则是短语的词汇化,出现众多词与短语纠缠不清的同形结构。考察《说文解字》会发现,其中任何一个汉字,都有其独立的意义,都不是作为某个词的一部分被解释的。随着语言的发展演变,一些概念逐渐由单音转变为复音。为顺应这变化,部分汉字结合成复音词,联绵词甚至不考虑构成汉字的意义,而单纯借其语音来表示一个相对完整的概念。在这一过程中,部分汉字失落了原有意义,失去了独立运用的能力,成为纯粹的构词语素。但先秦汉语中,这类语素很少,绝大多数单音语素可以独立运用,如"祖",既可同其他词合成"先祖""太祖"等,也可以独立运用,如"诬其祖矣"。因此,偏义或有了转义的复音词,多存在歧义。如:"人有恒言,皆曰天下、国家。天下之本在国,国之本在家,家之本在身。"(《孟子·离娄上》)此处,"国""家"分别表示不同概念,"国家"为短语,在"小人少而君子多,故社稷常立,国家久安"中,"国家"与"社稷"为相对概念,是作为一个偏义复词来用的。这种歧义情况在先秦汉语中远远大于现代汉语,显然不能通过词表匹配或概率模型等方式进行区别,而要考虑对语义的理解。这是先秦汉语区别于现代汉语的分词难点之一。

先秦汉语处于特定的发生发展阶段,词汇相对贫乏,句法结构也相对简单,有着大量名、动、形同词及词类活用现象,这一现象对词性的自动判别必然造成很大障碍。无论从语言本身还是信息处理角度来看,先秦汉语中兼类与词类活用均可看作同一类现象——一词兼多种词性。赵克勤指出:"在先秦,词类活用虽然不能完全排斥修辞方面的因素,但大多数还是受当时的语法规律所制约,是一种正常的语言现象。""与其说是运用修辞手段的结果,不如说是当时名、动、形三类词时常通用的表现"[7]。现代汉语中也存在部分兼类词及词类活用现象。但现代汉语处于成熟阶段,语言形式组合规则更为精确,有较明显的语法标志,兼类词数量少,使用也相对规范,因而,词性多可根据语法条件加以判别。如:"他正在工作。""他正在找工作。"两个"工作"的词性很容易判别;现代汉语的活用与先秦汉语的活用也有本质区别。现代汉语的活用大部分为临时的修辞[8],如"他的相声最近很火",本为名词的"火"受"很"修饰,活用为形容词。但处于初级阶段的先秦汉语,缺乏必要的语法形式标志,加之单音语素过多,缺少理解词汇义的标志性特征,活用词词性的判断要复杂得多。如"请勾践女女于王",第一个用如动词的"女",就很难根据语言形式标志加以判别,而必须充分考虑上下文的语境。这是区别于现代汉语的又一分词难点。

但是,综合考察先秦汉语中的兼类及活用现象,我们会发现,这两种现象都是人类思维联想的结果。如"禽",本义为动词"捕获",后由"捕获"联想到"捕获的对象","禽"就有了"走兽总名"这一名词意义。又如:"从左右,皆肘之","肘"的本义为名词"手肘",由此联想到手肘的动作,因而,文中用为动词"用手肘去

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

碰”。由此可见,处于发生阶段的先秦汉语更少形式上的标志,而更贴近人类的朴素思维模式。

3　复音词分布及构成情况的不同

从复音词所占比例来看,现代汉语中复音词占70%以上,而先秦汉语中复音词处于萌芽阶段,根据程湘清先生对《尚书》《论语》《韩非子》《孟子》等几部代表性作品的统计,除专名外,复音词大概只占13%左右[9]87。复音词所占比例的显著区别使得两者分词思路也不尽相同。

现代汉语词汇异常丰富,且以复音词为主,目前常用分词方法,或基于大规模分词词典(词典规模对分词的精确度有很大影响),或基于对大规模真实文本的统计,对语料中相邻共现汉字的组合频度进行统计,计算他们的统计信息并作为分词的依据。这两种主要的分词方法都是从“词”入手切分词(目前不少研究者利用基于人工智能的专家系统或神经网络分词方法。这类方法不同于前两者的思路,但目前还不成熟)。

先秦汉语词以单音词为主,词的数量较少,根据统计来看,常用词也不多。以《论语》为例,汉字1 335个,只出现1次的就有461,占34.5%,10以内(含10次)的有1 098个,占82.2%,10次以上的只有237个,占17.8%。同时,从语言发生发展规律来看,复音词是通过单音词组合而成并逐步取代复音词的,大部分词在由字成词的过程中遵循一定规律。程湘清先生统计认为,先秦词汇双音化过程中除部分语音成词(如叠音词、联绵词)外,绝大多数是运用语法手段构成的,且以并列式与偏正式为主,占了总数的94.8%[9]34。因此,结合上文所提到的先秦汉语的两个分词难点,先秦汉语的分词可考虑采取不同于现代汉语的分词方式,在特殊分词标志及分词词典(以专名、联绵词为主)辅助下,借鉴人的联想思维模式,利用人工智能技术,分析上下文语境,用构词规则来对单字进行词或固定短语的组合。

4　高频字的分布情况不同

不同类型现代汉语作品中字频分布有较大不同。以余秋雨散文《一个王朝的背影》与《流放者的土地》为例,两篇总字次为25 000,其中共出现单字2 020个,字的出现频率较为平均,100次以上的只有29个,10次以上的共463个,“的、一、是、不、人、这、在、他、有、来、个”等频率居于前列。而池莉的小说《来来往往》,总字次为62 593,其中单字2 561个,高频字分布与余秋雨的散文出现不同,除“的、一、是、不”外,“康、业、伟、说、段、好、时、里”等实词频率居于前列。这些高频字大部分能与其他字组成复音词(只有“的、不、在、有”等极少数高频字构词能力较弱),有些甚至是不能单独成词的语素,难以成为独立的分词标志。

根据本课题组统计,在先秦散文中,部分高频字在不同作品中出现频率非常稳定,具体分布见表1。

表1　部分高频字频次分布表

高频字	书名				
	论语	孟子	庄子	荀子	韩非子
之	611(3)	1 925(1)	2 944(1)	3 947(1)	3 315(1)
也	531(5)	1 245(2)	1 636(4)	2 694(2)	1 677(4)
不	583(4)	1 089(3)	1 897(3)	2 441(3)	1 950(3)
曰	756(2)	958(4)	995(8)	523(22)	1 163(6)
子	971(1)	948(5)	963(10)	746(12)	678(13)
而	343(6)	778(6)	2 075(2)	2 422(4)	1 992(2)
以	211(9)	641(7)	1 172(6)	1 558(6)	1 435(5)
者	219(8)	640(8)	1 133(7)	1 640(5)	1 134(9)
人	219(8)	626(9)	971(9)	1 244(8)	1 156(7)
其	270(7)	588(10)	1 203(5)	1 194(9)	1 146(8)
于	183(11)	562(11)	808(14)	586(20)	792(12)
为	170(13)	519(12)	941(11)	820(11)	818(11)
有	199(10)	465(13)	770(15)	741(13)	672(14)
则	124(21)	428(14)	485(19)	1 289(7)	840(10)

注:表中数字分别表示使用次数和位数

根据表1可以看出,这些高频字虽在各部作品中的位次稍有变化,但大多高居前列。其中包括大部分

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

常用虚词及少部分常用实词,除极少数构词能力较强的,如"子、人"外,一般很少与其他字构成复音词,如"之、也、不、而、者、则、以、曰、为"等。这些字与其他字共现频率非常高,却很少与其他字构成复音词,如果以之为分词标志,将大大提高分词效率。

以《论语》为例,共现2次以上2~4字的组合共有1 898条,而加上"之、不、以、而、也、其、则、为、者、于、矣、必、与、欲、非、皆、犹、乎、岂、盍、如、焉、故、夫、所、请、虽、若、诸、哉、亦、乃、未、甚、及、自、弗"等虚词作为断开标志后,共现组合只有608条,减少近68%;以"此、是、或、莫、何、孰、无、曰、有、得、可、谓、问、能、今"等部分高频常用实词作为断开标志后,也只有1 477条,减少近22.2%。如果将前面所列常用实词虚词共同用作断词标志,则所出现的组合只有412条,较之1 898条的总数减少了78.1%。而在加入分词标志后被删除的组合中,除一部分人名外,只有22条能被判断为词或常用固定短语。因此,在以构词代替分词的思路下,可以以部分高频字作为独立分词标志,以减少相邻字的共现频率,提高构词的效率。

5 结语

先秦汉语处于汉语发展的源头,与现代汉语有本质联系,又有各种区别。对两者从工程角度进行比较,我们发现,先秦汉语既有区别于现代汉语的分词难点,又有其独特的分词优势。只有针对先秦汉语的特征,充分利用其自身优势,同时吸收现代汉语分词方法,才能探索出符合先秦汉语实际的分词之路。

参考文献:

[1] 胡俊峰,俞士汶.唐宋诗中词汇语义相似度的统计分析及应用[J].中文信息学报,2002,16(4):39-44.
[2] 张 敏,马少平.用于信息检索的古文统计分析[J].中文信息学报,2001,15(6):40-46.
[3] 邱 冰,皇甫娟.基于中文信息处理的古代汉语分词研究[J].微计算机信息,2008,24(8-3):100-102.
[4] 刘开瑛.中文文本自动分词和标注[M].北京:商务印书馆,2000.45.
[5] 黄昌宁,赵 海.中文分词十年回顾[J].中文信息学报,2007,21(3):14.
[6] 孙茂松,王洪君,李行健,等.信息处理用现代汉语分词词表[J].语言文字应用,2001,17(4):85.
[7] 赵克勤.古汉语修辞简论[M].北京:商务印书馆,1983.107.
[8] 邹立志,白 聪.论古今汉语词类活用的不同本质[J].语言研究,2009,29(2):37.
[9] 程湘清.汉语史专书复音词研究[M].北京:商务印书馆,2003.87、34.

Differences between the Pre-Qin Chinese Words and the Contemporary Chinese Words from the Perspective of Automatic Segmentation

XU Zi-yun[1], XU Xue-song[2]

(1. School of Humanities and Social Sciences; 2. School of Electrical and Electronic Engineering, East China Jiaotong University Nanchang 330013, China)

Abstract: Word segmentation is the same problem of the pre-Qin Chinese and the contemporary Chinese information disposal. The methods of contemporary Chinese automatic segmentation can't be applied to pre-Qin Chinese automatic segmentation because of the difference of the two languages. The paper compares the pre-Qin Chinese words with the contemporary Chinese words from four angles: degree of language opening, manifestation and stage of the language development, distribution and constitution of the polysyllabic words, and distribution of the high-frequency words. It is found that there are not only distinctive difficult points but also distinctive merits for pre-Qin Chinese automatic segmentation.

Key words: word automatic segmentation; pre-Qin Chinese; contemporary Chinese; words; comparing

(责任编辑:刘棉玲 李 萍)

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net