日语文本语义接受度评价研究

杜家利,于屏方
DU Jia-li,YU Ping-fang

鲁东大学 外国语学院 汉语言文学院,山东 烟台 264025
School of Chinese Language and Literature,School of Foreign Languages,Ludong University,Yantai,Shandong 264025,China
E-mail:dujiali68@yahoo.cn

DU Jia-li,YU Ping-fang.Research on evaluation of semantic accessibility scale in Japanese text.Computer Engineering and Applications,2009,45(23):137-139.

Abstract: The study on agglutinative-language-involved Semantic Accessibility Scale(SAS) based on Japanese corpus comprises three steps.Firstly,『ゆきぐに』is extracted from corpus and divided into six groups for comparison by the systematic random sampling skill in which different equidistant extraction is included.Secondly,the definition of word height in presently-verified SAS formula reflecting inflecting language domain is adapted for agglutinative language domain.The word beyond five music beats is called the unpopular one,and the number of this kind of word every 100 words is considered word height.Finally,a conclusion is drawn that decreasing extracted-space results in increasing Sampling Ratio(SR),and that the non-relevance between SR and SAS is verified by the schema in which the contrast between increasing SR and the mean-fluctuated SAS is involved.In short, the evaluation of SAS in inflecting language text can be applicable in other fields,including agglutinative language text.

Key words: agglutinative language;information processing;corpus;Semantic Accessibility Scale(SAS);Sampling Ratio(SR)

摘 要:基于日语料库的粘着语文本语义接受度(SAS)研究分三步展开。首先提取『ゆきぐに』为分析文本,以等距离系统随机抽样方法取得6对比组。然后在屈折语SAS研究基础上提出适用于粘着语文本的词长定义,即百词所含5音拍及以上词数为超常用词量。最后得出结论:抽取间距由大变小引发抽取率(SR)由小变大的曲线变化;依次攀升的SR与围绕均值波动的SAS组图证明两者的非关联性,以实例验证了屈折语SAS评价公式对粘着语文本研究的可适用性。

关键词:粘着语;信息检索;语料库;语义接受度;抽取率
DOI:10.3778/j.issn.1002-8331.2009.23.038 **文章编号:**1002-8331(2009)23-0137-03 **文献标识码:**A **中图分类号:**TP311

1 引言

基于语料库的语义接受度(Semantic Accessibility Scale,SAS)研究是对文本可理解程度进行评价的理论研究。随着网络发展和海量文本电子化,如何借助语料库对文本SAS进行赋值和评价,已成为计算语言学研究的热点,不仅包括对先期语言假设寻求大规模真实文本量化支持的实证性语料库研究[1-4],还包括以理论建构为出发点并以大规模真实文本为支撑展开的抽象性语料库研究[5-8]。其目的都是利用公式或算法对文本进行量化研究,客观上方便了文学研究者借助计算机对文本进行形式化分析,例如通过统计数据剖析作者用词规范和特点的写作风格研究[9],通过理论构建提出和验证文本分析的通用算法[10]和评价范式[11]研究等。

以日语料库为检索源,在语义理论框架内对文本SAS进行评价研究,分析同一文本在不同抽取率下语义值有无偏差及值域变化曲线,最后从计算语言学角度对粘着语文本SAS进行总结归纳,为日文本形式化研究提供数据库理论支撑。

2 现行语义接受度的评价研究

基于语料库的文本语义研究涉及多个学科领域,如文体学[12-13]、信息检索[14-15]、自动文摘[16-18]、平行语料库对比[19-22]、计算科学[23-24]、语言应用[25-26]等。语义评价系统主要分人工和机器两类,前者以人为主,准入条件低、灵活性好,但差错率高且可验证性差;后者以机器为中心,成本昂贵、灵活性差,但差错率低且能重复验证。常见的人工评价有质量评价、问答评价和阅读性理解评价;现行的机器评价方法有召回率和精确度分析法、F-Measure测试法、Rouge分析法和F-New-Measure。该文涉及的评价方法主要是基于语料库的自动分析,属机器评价范围,强调文本评价的自动性、复现性和模式性。自动性是指这种方法以机器为中心,强调评价结果的自动生成而非人工评价。复现性是指有既定程序,而且再次评价的结果是前次评价的复现,独立于分析个体和取样文本。模式性是指自动评价系统具有可验证性的运作模式和公式,为系统的自动化评价提供理论支撑。

基金项目:国家社会科学基金项目(No.08BYY046);教育部人文社会科学重点研究基地重大项目(No.06JJD740007);山东省社会科学规划项目(No.07CWXJ03)。

作者简介:杜家利(1971-),男,硕士,研究方向:篇章语义学和计算语言学;于屏方(1971-),女,博士,研究方向:应用语言学。

收稿日期:2008-12-16 **修回日期:**2009-02-23

召回率(*R*)和精确度(*P*)分析法依据文本句数(*T*)、自动摘要句数(*A*)和摘要文本中所含的原文本句数(*S*)三个变量进行评价,如 *T*=100,*A*=20,*S*=5 时,则 *R*=5%,*P*=25%。F-Measure 测试法侧重 *P* 和 *R* 的联动性,即 F-Measure 等于 2*PR* 与(*P*+*R*)的比值。Rouge 分析法较为复杂,主要通过机器文摘和人工文摘所重叠的单词数目来确定数值的变化。分析过程需要考虑最长公共子序列的相似程度和权重值,并需要测算出两文摘中单词共现的最大值。F-New-Measure 改进法较 F-Measure 测试法增加了一个新的参数压缩率 *C*,强调机器摘要长度 *L1* 与文本总长度 *L* 之比对评价值的影响。

由 F-New Measure 可知,压缩率是决定系统评价稳定与否的重要变量,压缩率在文本语义分析中体现为抽取率,即按照一定标准对语料库文本提取的比率。

词句抽样率与文本语义相关。句抽样率值越高,涉及的语义范畴越宽,需要正确理解文本的认知背景越高,语义理解自由度越低,语义接受度越低,即句抽样率与 SAS 成反比。词抽样率的增加带来文本理解词间语境的加大,利于文本的理解,加大语义理解的自由度,即词抽样率与 SAS 成正比例。

句长是指单位句子所含平均词数,如超过平均水平,句子中所包含的词汇承载信息就会满载,读者解码速度就会减慢,认知理解难度就会增大,最终导致文本可理解程度降低,词长可用于测定文本作者词汇通用性,口语体倾向文本超常用词较少因而易于接受,书面体倾向文本则相反。如巴金和倪海曙文本风格就能通过词长和句长进行区分:前者创作文本每句平均词数为 24.75,平均字数为 40.65,最长句含 803 个字母,最短句有 60 个字母;后者作品词数为 15.79,字数为 24.05,最长句有 363 个字母,最短句有 14 个字母。由此可知:巴金作品长句多,用词细致,符合书面体文本特征;倪海曙作品短句多,描写简洁,接近口语风格[9]。所以,词句长可作为日语 SAS 研究变量。

SAS 研究除考虑抽取率和词句长外,还要验证 SAS 是否与 SR 具有依附性。如不具有,则说明 SAS 具有恒定性,可作为文本分析者抽样调查的依据。理论情况下,文本风格不会随抽样率而变化,即抽取率不应引起 SAS 值显著波动。抽取率是计算取样比率的变量,抽样率高,覆盖文本的程度就密集,体现文本语义特征的值就丰富,描绘的文本语义特征就清晰。但有时由于示样文本较大,很难或不必要进行全样抽取,这时如果能得到一个相对独立于抽取率的评价公式,即不同的抽取率不会引起或基本不会引起原文本语义特征变化的公式的话,则会提高文本分析者进行网络文本评析的能力和效率。

由此提出文本 SAS 评价公式,涉及语料库文本的词句长和抽取率。设单位句子含词量为句长 *L*、百词中超常用词量为词长 *H*、词句长之和的加权值为 0.4,文本取样句数 *S*1、取样词数 *W*1、文本总句数 *S*,总词数 *W*,句抽样率为 *P*1,词抽样率为 *P*2,词句综合抽取率 SR,语义接受度为 SAS[27]。

$$SR=\frac{2\times P1\times P2}{P1+P2}=\frac{2\times S1\times W1}{S1W+SW1} \tag{1}$$

$$SAS=\frac{1}{0.4(L+H)}\times\frac{P2}{P1} \tag{2}$$

$$SAS=\frac{1}{0.4(L+H)}\times\frac{S\times W1}{S1\times W} \tag{3}$$

3 粘着语文本语义接受度评价标准

粘着语是相对于屈折语和孤立语而言,通过在各种词缀中加入词根以增加意义或改变语法功能。典型的粘着语除日语外,包括芬兰语、匈牙利语、斯瓦希里语和土耳其语。东非的斯瓦希里语中“他们已付钱给我们了(wametulipa)”,构词成分分别为 “wa +me+ tu+ lipa”相对应的是“他们+(完成式符号)+我们+付钱”。这句话日语可翻译为“彼らはもう私たちが支払った”,其结构为 “彼ら+は+もう+私たち+が+支払+った”,即“他们+(助词)+已经+我们+(助词)+支付+(过去式)”。

与屈折语超常使用的三音节词不同,日语在常用词语料库统计中因其粘着构词特点,以五音拍词作为超常用词界限。超常用词承载着较多语义信息,有较浓重的书面语倾向,代表着作者独有的写作风格。由此,日文本 SAS 求证公式中词长 H 应定义为百词中所含 5 音拍及以上词总数,其他赋值标准同屈折语。

下文验证日文本 SAS 值在多抽取率时是否发生显著偏差,并以具体数据描绘偏差值变化。

4 基于日语料库的语义接受度分析

基于日语料库的 SAS 分析操作如下:(1)从日语料库中抽取文本后以系统随机抽样方法进行页码选择;(2) 起始页码按照随机量表选择,余下单位等距离抽取;(3)抽取间距由 10、5、4、3、2 和 1 页(即全文抽取)为标准组建 6 个取样对比组,抽取公式为 $A+BX\leqslant C$,*A* 为起始页码,*B* 为抽取间距,*C* 为文本总页码,*X* 为能抽取的最大页数量。

如文本总页码为 500 页,抽取间距为 10 页,共抽取 50 页。假定从随机量表中选取某单位(≤10)为起始页,如 7,则以后抽取的页码分别为 17,27,37,…,497。抽取公式 $7+10X\leqslant 500$,$X\in(0,49)$。各抽取页码组成对比组,并根据字句数、词句长等求得各组 SR 和 SAS 均值,然后对比研究两者的依附性。

4.1 日文本多抽取率对比

日文本以诺贝尔文学奖得主かわばた やすなり (川端康成)作品『ゆきぐに』《雪国》(新潮社 1969 日语版)为语料。《雪国》创作跨越整个二战,前后 13 年,是川端在了结其生命的最后阶段完成的,文本内容虽没有直面日本军国主义的侵略战争,但却通过纯洁的爱情将虚无思想深深地渗透在文本中。该文充分体现了作家的思想和风格。

该文总长 124 页,分别按照 10、5、4、3、2 和 1 页为抽样间距组成 6 个对比组。第一组抽取 12 页,间距为 10,初始页码为 6,后续为 16,26,36,…,116,即 $6+10X\leqslant 124$,$X\in(0,11)$。第二组抽取 25 页,间距为 5,初始页码为 4,后续抽取 9,14,19,…,124,即 $4+5X\leqslant 124$,$X\in(0,24)$。第三组抽取 31 页,间距为 4,初始为 3,后续抽取 7,11,15,…,123,即 $3+4X\leqslant 124$,$X\in(0,30)$。第四组抽取 41 页,间距为 3,初始为 2,后续抽取 5,8,11,…,122,即 $2+3X\leqslant 124$,$X\in(0,40)$。第五组抽取 62 页,间距为 2,初始为 1,后续为 3,5,7,…,123,即 $1+2X\leqslant 124$,$X\in(0,61)$。第六组全文统计 124 页。数据如表 1、图 1 所示。

表 1 『ゆきぐに』对照组抽取率数据列表

类别	*P*12	*P*25	*P*31	*P*41	*P*62	*P*124
*W*1	5 925	12 810	17 254	21 368	33 383	65 200
*S*1	248	515	661	897	1 379	2 718
SR	0.091 1	0.192 9	0.253 5	0.328 9	0.509 7	1

如表 1 和图 1 所示,抽取间距为 10、5、4、3、2 和 1 页时,抽取率依次为 9.11%、19.29%、25.35%、32.89%、50.97%和 100%。

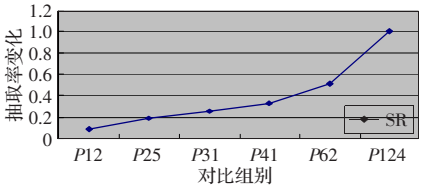

图 1　『ゆきぐに』对照组抽取率变化图

抽取率为依次攀升的曲线，抽取间距越大，抽取率越小，越位于曲线的底端。抽取间距引发抽取率规则性变化。

4.2　日文本多语义接受度对比

各抽取率获值后，可计算出单位句子含词量（句长 L）和百词中 5 音拍及以上超常用词量（词长 H），再利用公式求得各 SAS。

由表 2 和图 2 所示，各 SAS 值为 0.085 2，0.089 3，0.089 1，0.089 3，0.088 4 和 0.088 1。值域 ∈（0.085 2，0.089 3）。当抽取间距为 10 时，抽样组与 SAS 均值有较大差距，但当进入较密集抽取间距的 5、4、3、2 页和全文抽取时，SAS 呈现不规则变化。即抽取间距最大的 P12 组 SAS 值最小但偏离度最大，其他 SAS 围绕均值波动。规则变化的抽取间距与非规则变化的 SAS 组图说明 SAS 与抽样间距不具有关联性。

表 2　『ゆきぐに』对照组语义接受度数据列表

类别	*P*12	*P*25	*P*31	*P*41	*P*62	*P*124
*W*1	5 925	12 810	17 254	21 368	33 383	65 200
*S*1	248	515	661	897	1 379	2 718
L	26.27	26.48	27.83	25.05	25.86	25.65
H	2.97	2.55	2.69	2.75	2.68	2.72
SAS	0.085 2	0.089 3	0.089 1	0.089 3	0.088 4	0.088 1

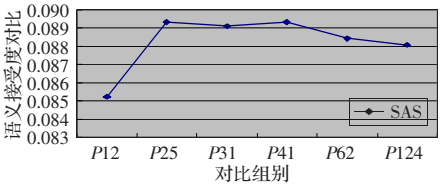

图 2　『ゆきぐに』对照组语义接受度变化图

4.3　日文本语义接受度与抽取率对比

由图 1 可知，抽样间距大小决定抽样率的变化，间距越大，抽样率越小。由图 2 可知，规则的抽样间距未带来 SAS 渐进性变化，两者不成规律曲线。抽样率和语义接受度数据和变化曲线如表 3 和图 3 所示。

表 3　『ゆきぐに』对照组抽取率与语义接受度数据列表

类别	*P*12	*P*25	*P*31	*P*41	*P*62	*P*124
SR	0.091 1	0.192 9	0.253 5	0.328 9	0.509 7	1
SAS	0.085 2	0.089 3	0.089 1	0.089 3	0.088 4	0.088 1

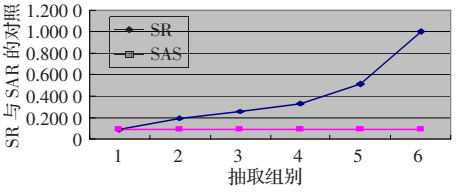

图 3　『ゆきぐに』抽取率与语义接受度对比图

由表 3 和图 3 所示，抽取率为 9.11%、19.29%、25.35%、32.89%、50.97%和 100%时，语义接受度为 0.085 2，0.089 3，0.089 1，0.089 3，0.088 4 和 0.088 1。各语义接受度围绕均值波动而不随抽取率变化。这说明多样抽取率不会或基本不会带来特定文本语义值的偏差，日文本 SAS 公式利于文学分析者采用抽样方法进行语料库文本语义分析。

5　结论

基于自然语言的文本语义接受度（SAS）研究是计算语言学研究的一个方向。以粘着语代表的日语语料库文本为语料，进行了适用于文学文本可理解程度的量化探索，验证了语义接受度公式并得出以下结论：（1）抽样间距带来抽取率依次攀升的曲线变化，间距越大，抽样率越小，越位于曲线底端；（2）抽样率规则变化未引起代表文本风格特点的语义接受度的规则变化，除抽样间距 10 页组表现出偏离 SAS 均值的特点外，其他组别围绕 SAS 均值波动；（3）日文本 SAS 公式独立于抽样率，证明其对粘着语文本语义评价具有敏感性，便于文学研究者抽样调查。尽管 SAS 公式独立于抽样率已经得到验证，但较大抽样间距的 SAS 值偏离性的临界点是否可测以及语义值失真是否可控值得进一步探讨。

参考文献：

[1] Furui S，Nakamura M，Ichiba T，et al.Analysis and recognition of spontaneous speech using corpus of spontaneous Japanese[J].Speech Communication，2005，47：208-219.

[2] 施建军，徐一平.日语词汇单一汉译词自动获取研究[J].解放军外国语学院学报，2003，26(5)：65-68.

[3] 陈治平，尤文虎.义素分析法在日语计算机处理中的基础性应用[J].解放军外国语学院学报，2001，24(3)：32-35.

[4] 毛文伟.试析复合辞“~ テナラナイ”、“~テショウガナイ”、“~テタマラナイ”的异同——语料库统计法在语法研究中的应用一例[J].解放军外国语学院学报，2002，25(3)：62-66.

[5] 周惠巍，黄德根，李巍.基于支持向量机的日语并列关系解析[J].大连理工大学学报，2007(6)：904-908.

[6] 施建军，徐一平.语料库与日语研究[J].日语学习与研究，2003(4)：7-11.

[7] 王冠华.关于面部表情描写的中日对比研究—基于语料库所进行的调查[J].日语学习与研究，2007(4)：10-17.

[8] Sakamoto K，Terai A，Nakagawa M.Computational models of inductive reasoning using a statistical analysis of a Japanese corpus[J].Cognitive Systems Research，2007，8：282-299.

[9] 钱锋，陈光磊.关于发展汉语计算风格学的献议[M]//胡裕树，宗廷虎.修辞学发凡与中国修辞学.上海：复旦大学出版社，1983.

[10] 冯志伟.花园幽径句的自动分析算法[J].当代语言学，2003(4)：339-349.

[11] Yu P F，Du J L.Automatic analysis of textual garden path phenomenon：A computational perspective[J].Journal of Communication and Computer，2008，5(10)：58-65.

[12] 杜家利，于屏方.迷失与折返-海明威文本“花园路径现象”研究[M].北京：中国社会科学出版社，2008.

[13] Clancy P M，Thompson S A，Suzuki R，et al.The conversational use of reactive tokens in English，Japanese，and Mandarin[J].Journal of Pragmatics，1996，26：355-387.

[14] 吴云芳，俞士汶.信息处理用词语义项区分的原则和方法[J].语言文字应用，2006(2).

[15] 王诚，张璟.基于语义的 Web 信息检索[J].计算机应用研究，2005(8)：111-112.

（下转 148 页）