英日语料库语义接受度对比研究

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 Foreign Languages,School of Chinese Language and Literature,Ludong University,Yantai,Shandong 264025,China
E-mail:dujiali68@yahoo.cn

DU Jia-li,YU Ping-fang.Comparative study on semantic accessibility scale originating from English and Japanese corpora.Computer Engineering and Applications,2009,45(24):146-149.

Abstract: The corpus-based study on Semantic Accessibility Scale (SAS) is a useful method to evaluate the acceptance of electronic texts.On the basis of large-scale natural language texts,this paper compares *The Old Man and the Sea* and 『ゆきぐに』 from English and Japanese corpora by means of the information retrieval and semantic assignment.A conclusion is drawn that *SAS* is related to vocabulary density(*P*1,*P*2),vocabulary length(*H*) and sentence length(*L*),namely $SAS=P2/[P1\times0.4\times(L+H)]$.Correspondingly,different sampling ratios will not result in fundamental difference of *SAS*.This study provides the theoretical support for the literary critics to analyze the acceptance of internet-based texts.

Key words: text;corpus;natural language;semantic accessibility scale;information retrieval

摘 要: 基于语料库的语义接受度(SAS)研究是在线衡量文本理解程度的可行性方法。在大规模真实文本语料的基础上,利用赋值限域方法进行英日文本对照研究。并通过分析不同赋值区间对英日小说文本语义接受度进行解读。经过验证的语义接受度公式证明了文本理解与词汇密度(*P*1,*P*2)、词长(*H*)和句长(*L*)相关,即 $SAS=P2/[P1\times0.4\times(L+H)]$,而且不同的抽取率不会引起评价值的显著差异。此公式为文学研究者借助网络对电子文本进行理解度评价提供了理论支持。

关键词: 文本;语料库;自然语言;语义接受度;信息检索
DOI: 10.3778/j.issn.1002-8331.2009.24.043 **文章编号:** 1002-8331(2009)24-0146-04 **文献标识码:** A **中图分类号:** TP391

1 引言

语义接受度(Semantic Accessibility Scale)是系统从语义角度评价语料库文本可理解程度的标准之一。随着网络发展,多语种电子文本得到广泛应用。这为语料库研究提供了充足的语料。但如何借助语料库对多语种文学文本的语义接受度进行赋值和评价,已成为国内外学者研究的热点。

2 基于多语种的平行语料库研究

基于多语种的平行语料库研究,主要分为实践性和理论性两类。

实践性语料库研究是指以语料库为检索源进行的实证性研究,对先期的语言假设寻求大规模真实文本的量化支持,如利用中日平行语料库进行面部词语的辨析[1],基于大规模依存关系语料库进行的支持向量机 SVM(Support Vector Machines)的使用[2],依托日本奈良 ChaSen2.1 和北大 SLEX 分词系统建立的日汉译词语料库展开的一对一自动抽取研究[3],自建语料库对日语复合词辨析的实例运用[4],采用英日汉口语对比语料库进行的 RT(Reactive Token)研究[5],基于英语和阿拉伯语平行语料库提出的可适用于 P-NNT(Probabilistic Neural Network)和 GMM(Gaussian Mixture Model)的句子对齐实例模式研究[6],借助韩日英跨语言检索系统 CLIR(Cross-Language Information Retrieval)使用动态增量聚类(Dynamic Incremental Clustering)方法所进行的解歧处理[7],利用中日英语料库开发的具有实时便携功能的口译系统的应用[8],基于日意 8 月龄童语料库展开并得出"频率分布和句位(frequency distributions and sentential positions)"是影响幼儿词序习得的实践研究[9],采用 CSJ(Corpus of Spontaneous Japanese)语料库进行的言语实时分类和任务指向自动语音识别研究[10],基于语义的 Web 信息检索[11]等。

理论性语料库研究是指以语料库的理论建构为出发点,以大规模真实文本为支撑展开的抽象性研究,如利用义素本体研究成果[12]进行日语机语义构建的尝试[13],对日语语料库构建的理论探索[14],从神经认知机制角度探索的日语非语序线性句子理解系统的建立[15],从英日对比语料库入手提出的对"光泽翻译策略(gloss translation strategy)"敏感的"基于询问的信息检索系统(query-based information retrieval systems)"的理论构建[16],可适用于汉语、西班牙语、日语和阿拉伯语互为检索的 SpidersRUs 搜索引擎系统的确立和验证[17],利用关联规则进行的检索模型研究,利用"询问翻译法(query-translation approach)"

基金项目: 国家社会科学基金项目(No.08BYY046);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: 2009-03-09 **修回日期:** 2009-05-04

和“迭代解歧图式(an iterative disambiguating scheme)”构建的日汉交叉语言信息检索系统[18]，基于日语口语语料库进行的LVCSR(Large Vocabulary Continuous Speech Recognition)系统的扩展性研究[19]，建立在语料库统计基础上利用日语“共现可能词对(co-occurrence probabilities for word pairs)”提出的基于类别的三类神经网络模式[20]，从日汉词汇相似性提出的自组织单语语义图谱(self-organizing monolingual semantic maps)和基于日汉平行语料库词汇对齐的语义模式的构建[21]等。

该文以英日文学文本语料库为检索源，在语义理论[22-23]基础上对多语种自然语言信息检索进行对照研究，分析英语(屈折语)和日语(黏着语)文本在系统检索过程中的语义接受度的对比性，并进行数据测算，分析评价语义值有无偏差及出现偏差的原因，最后从信息检索角度对英日语料库文本的语义接受度进行总结归纳，为文学文本的形式化研究提供数据库理论支撑。

3　语义接受度在线评价方法及理论

随网络发展而产生的语义评价系统的建立(如基于电子文本的质量评价、问答评价和阅读性理解评价)，标志着现代文本形式研究的开始，其主要特点体现在文本评价的自动性、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。复现性是指按照既定程序，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。模式性是指自动文摘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。召回率和精确度分析法、F-Measure 测试法、Rouge 分析法和 F-New-Measure 改进法是目前广泛使用的语义评价方法。

召回率(R)和精确度(P)分析法相对简单，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)三个变量进行文摘的自动评价，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 $F\text{-}Measure=2PR/(P+R)$。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(机器摘要长度 $L1$ 与文本总长度 L 之比)，并在非受限领域的评价系统中对公式 $F\text{-}New\text{-}Measure=2PR[1-(L1/L)]/(P+R)$ 进行了可行性验证[24]。

基于语料库文本的语义评价通常与文本结构相关，如词长、句长、段落值、章节值、抽取率等。

词长是指作者惯于使用的单词长度，如英语(屈折语)文本侧重的超常使用的三音节或以上(非屈折变化)单词数量，汉语(孤立语)关注的超常使用的双字组合以上的词的总量[1]，日语(黏着语)文本强调的超常使用的五音拍及以上的字词数量。词长参量通常可测定文本作者的词汇通用性，口语体倾向的文本超常量词较少因而易于接受；书面体倾向的文本则相反。

句长是指文本句子中所含有的平均词数，句长与文本可理解程度成反比。如句长超过平均水平，句子中所包含的词汇承载信息就会满载，读者进行解码的速度就会减慢，认知理解难度就会增大，最终导致文本可理解程度降低。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[25]。

段落值和章节值通常指示样文本的自然段落和章节的数量值，文本段落值和章节值提高，文本长度加长，可供理解的背景知识增多，进行文本语义评价的素材也相应增多，有利于文本评价的展开。

抽取率是用来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。

基于上述变量的讨论，提出可用于对电子文本进行在线评价的语义接受度公式，涉及语料库文本的词长、句长、抽取率和词汇密度。段落值和章节值具有文本研究的特定性，不适合在多语种文本中进行横向比较，所以，在英日语料库基础上展开的文本语义接受度的讨论不涉及这两个变量。

设定句长 L(平均每句所含单词的数量)、词长 H(平均每百个单词所含的非屈折变化所致三音节或以上单词数量)、文本取样句数 $S1$、取样词数 $W1$、文本总句数 S，总词数 W，基于句数的词汇密度为 $P1$，基于词数的词汇密度为 $P2$，词句综合抽取率 SR，语义接受度为 SAS。

对传统英语文本可理解程度的研究，通常引入 $Fog\ Index=0.4(L+H)$，此公式考虑了句长和词长对文本理解的影响，即词长和句长值越高，读者理解文本所需的认知值越高，通过文本结构体现的 $Fog\ Index$ 越高，文本的可理解程度越低，因此，语义接受度 SAS 与 $Fog\ Index$ 成反比例关系。同时，两类词汇密度(取样句数 $S1$ 与总句数 S 之比，即 $P1=S1/S$；或取样词数 $W1$ 与总词数 W 之比，即 $P2=W1/W$)均与文本结构相关。基于句数的词汇密度 $P1$ 语义跨度较大，涉及抽样文本句数和文本总句数之比，其值越高，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 $P1$ 与 SAS 成反比。基于词数的词汇密度 $P2$ 语义跨度较小，限定在词层，其单位词密度的增加，为取样文本的理解提供必须的词间语境，利于文本的理解，即 $P2$ 与 SAS 成正比。语义接受度公式可设定为 $SAS=P2/[P1\times0.4\times(L+H)]$。

通过分析 F-New-Measure 的测算思路可知，抽取率 SR 是决定系统语义评价稳定与否的重要变量。单位数量的抽取率与词汇密度具有关联性。设定两种词汇密度的表示方法与抽取率具有 F-New-Measure 中召回率和精确度类似的语义关系，可获得公式 $SR=(2\times P1\times P2)/(P1+P2)$。

综上所述，语料库文本的语义接受度研究除需要考虑基于词句的词汇密度、词长和句长外，还需要关注一定的抽取率，并验证文本语义是否与抽取率具有依附性。

4　语义接受度与文本抽取率

文本语义接受度通常情况下代表了作者独有的写作风格，

[1] 双字组合占词典收条总数的 67.625%，在现代汉语中占有绝对的优势，而其他组合所占比例偏低。

理论情况下，不同的文本抽取率不应该引起文本 *SAS* 值的显著波动。也就是说，按照不同比率从语料库文本中取出的段落词句应该具有基本一致的文本语义接受度，而且能代表作者的特有风格。如能通过数值证明 *SAS* 与 *SR* 变化的非一致性，则可断定两者的独立关系。

下文分别从英日语料库中抽取具有同等页码长度的海明威 1954 年、川端康成 1968 年诺贝尔文学获奖作品 The Old Man and the Sea《老人与海》(译林出版社 2001 英文版)、『ゆきぐに』《雪国》(新潮社 1969 日语版)为语料，并分别抽取 20、30、40、50、65 和 124 页的 6 个组进行评价结果的对比测试。如果证明 *SR* 差异不能带来 *SAS* 的显著变化，则可证明 *SAS* 独立于 *SR*，因而具有文本抽样调查的可推广性。

统计测试中，词长 *H* 取值略有不同：在英文本中代表为平均每百个单词所含的非屈折变化所致三音节或以上单词数量，在日文本中表示为平均每百个单词所含的五音拍及以上的字数量。其他参量均相同。

4.1 基于英文本的语义接受度与文本抽取率对比

《老人与海》全文共 124 页，按照抽取的 6 组进行对照，具体页码选择按照统计学中随机统计量表从语料库中进行抽取，并计算出词汇密度 *P*1 和 *P*2 的值，以公式 $SR=(2\times P1\times P2)/(P1+P2)$ 得到具体每组的基于词句的文本抽取率，并以该值为第一纵轴线。文本句长 *L* 和词长 *H* 可通过数据统计获得，经公式 $Fog\ Index=0.4(L+H)$ 和 $SAS=P2/[P1\times 0.4\times(L+H)]$ 计算出各对照组的语义接受度，并以该值为第二条纵轴线，横轴以 6 对照组为自然轴线，见图 1。

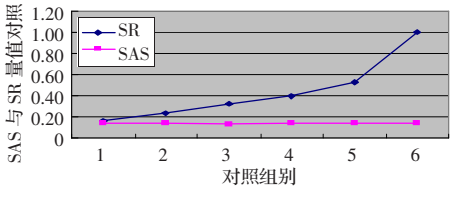

图 1 《老人与海》语义接受度与抽取率对照图

如图 1 所示，各对照组 *SR* 分别为 16%、24%、32%、39%、52%、100%时，各组所对应的 *SAS* 值除 40 页组为 0.13 外，其余均为 0.14，也就是说，单一文本 *SAS* 值基本保持一致，没有随抽取率的变化而显著变化，见表 1：

表 1 《老人与海》语义接受度与抽取率量值对照

Group	Fog Index	*P*1	*P*2	*SR*	*SAS*
P20	7.29	0.16	0.17	0.16	0.14
P30	7.04	0.24	0.23	0.24	0.14
P40	7.12	0.33	0.31	0.32	0.13
P50	7.63	0.38	0.40	0.39	0.14
P65	7.16	0.52	0.53	0.52	0.14
P124	7.13	1.00	1.00	1.00	0.14

4.2 基于日文本的语义接受度与抽取率对比

与《老人与海》的总页码数一样，《雪国》全文 124 页，分别抽取 20、30、40、50、65 和 124 页的 6 个组进行对照。通过公式 $P1=S1/S$ 和 $P2=W1/W$ 计算出词汇密度，以公式 $SR=(2\times P1\times P2)/(P1+P2)$ 计算出文本抽取率，再经公式 $SAS=P2/[P1\times 0.4\times(L+H)]$ 计算出语义接受度。最后以 6 个对照组为横轴，以 *SR* 和 *SAS* 值为两条纵轴进行对比，见图 2。

如图 2 所示，虽然《雪国》与《老人与海》的总页码数和抽取

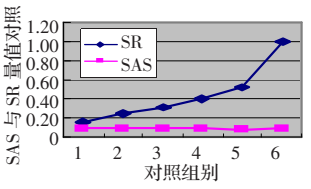

图 2 《雪国》语义接受度与抽取率对照图

的页数都一样，但两者 *SR* 值不尽相同，《雪国》抽取率分别为 16%、25%、31%、40%、52%、100%，各组所对应的 *SAS* 值除 65 页组为 0.08 外，其余均为 0.09，较显著的文本抽取率变化没有带来较明显语义接受度的变化。

表 2 《雪国》语义接受度与抽取率量值对照

Group	Fog Index	*P*1	*P*2	*SR*	*SAS*
P20	10.82	0.16	0.15	0.16	0.09
P30	11.22	0.25	0.24	0.25	0.09
P40	12.29	0.30	0.33	0.31	0.09
P50	11.65	0.40	0.40	0.40	0.09
P65	12.79	0.51	0.53	0.52	0.08
P124	11.35	1.00	1.00	1.00	0.09

4.3 结果分析

为保持抽样文本的可对照性，从英日语料库中抽取等长的诺贝尔获奖文本《老人与海》与《雪国》进行语义接受度对比研究。由表 1 和表 2 可知如下不同：(1)抽取率的不同。尽管两文本具有相同的抽取页数，但涉及不同的抽取词句量，因而导致抽取率不尽相同，也就是说相同的抽取页数不是抽取率取值的恒定条件；(2)语义接受度的不同。两文本涉及不同的抽取率、词长、句长和词汇密度，因而语义接受度也不同；《老人与海》较《雪国》有较高的语义接受度，说明素以简洁文体风格著称的海明威文本比素以描写细腻的川端康成文本有着更易理解的特性，也从一个侧面表明具有屈折特性的英语文本比依靠黏着成分组词的同类日语文本有着较高的文本可接受性。

5 结论

英语是屈折语系的典型，对其语义接受度的研究必须考虑屈折变化所造成的词长、句长和词汇密度。日语是黏着语系的代表，其主词附加黏着成分而产生新词的机制不同于英语屈折生词的造词机制，对其研究需要考虑音拍语的黏着特性。基于英日语料库语义接受度的研究，对英日抽样文本的词长、句长、词汇密度给予了重视，并提出语义接受度计算公式：$SAS=P2/[P1\times 0.4\times(L+H)]$，其中，词汇密度涉及词和句的双重抽取，英文本词长以非屈折变化所致的三音节或以上词量为准，日文本以附带黏着成分的五音拍及以上的字数为准。并通过对等长的《老人与海》与《雪国》原著的统计计算，证明了语义接受度公式具有独立于抽取率的可行性。并得出《老人与海》具有简洁易懂、倾向口语体的风格，而《雪国》具有描写细腻、倾向书面语的特征。基于抽样文本的结论是否适用于其他屈折语(英语)和黏着语(日语)文本值得进一步探索。

参考文献：

[1] 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.

[2] 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.
[3] 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.
[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] Clancy P M.The conversational use of reactive tokens in English, Japanese, and Mandarin[J].Journal of Pragmatics,1996,26:355-387.
[6] Fattah M A.Sentence alignment using P-NNT and GMM[J].Computer Speech and Language,2007,21:594-608.
[7] Lee K S.Implicit ambiguity resolution using incremental clustering in cross-language information retrieval[J].Information Processing and Management,2004,40:145-159.
[8] Shimizu T.NICT/ATR Chinese-Japanese-English speech-to-speech translation system[J].Tsinghua Science and Technology,2008,13(4):540-544.
[9] Gervain J.Bootstrapping word order in prelexical infants:A Japanese Italian cross-linguistic study[J].Cognitive Psychology,2008,57:56-74.
[10] Furui S.Analysis and recognition of spontaneous speech using corpus of spontaneous Japanese[J].Speech Communication,2005,47:208-219.
[11] 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.
[12] 池上嘉彦.意味论[M].东京:大修馆书店,1975.
[13] 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.
[14] 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.
[15] Wolff S.The neural mechanisms of word order processing revisited:Electrophysiological evidence from Japanese[J].Brain and Language,2008(6).
[16] Oard D W,Resnik P.Support for interactive document selection in cross-language information retrieval[J].Information Processing and Management,1999,35:363-379.
[17] Chau M.SpidersRUs:Creating specialized search engines in multiple languages[J].Decision Support Systems,2008,45:621-640.
[18] Lin C C.Learning weights for translation candidates in Japanese-Chinese information retrieval[J].Expert Systems with Applications,2008(9).
[19] Ohtsuki K.Japanese large-vocabulary continuous-speech recognition using a newspaper corpus and broadcast news[J].Speech Communication,1999,28:155-166.
[20] 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.
[21] Ma Q.Self-organizing semantic maps and its application to word alignment in Japanese-Chinese parallel corpora[J].Neural Networks,2004,17:1241-1253.
[22] 杜家利,于屏方.自然语言文本语义接受度的在线系统评价研究[J].计算机工程与应用,2008,44(26):141-143.
[23] 吴开贵,万红波,朱郑州.一种基于语义的本体概念相似度的计算方法[J].计算机科学,2008,35(5):123-124.
[24] 傅间莲,陈群秀.一种新的自动文摘系统评价方法[J].计算机工程与应用,2006,42(18):176-177.
[25] 钱锋,陈光磊.关于发展汉语计算风格学的献议[C]//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.

(上接 135 页)

[3] 武薇,范影乐,庞全.基于广义维数距离的语音端点检测方法[J].电子与信息学报,2007,29(2):465-468.
[4] 闫润强,朱贻盛.基于信息递归度分析的语音端点检测方法[J].通信学报,2007,28(1):35-39.
[5] 马龙华,臧义华,刘利强.车内噪声环境下的语音端点检测和增强技术[J].计算机工程与应用,2007,43(36):217-219.
[6] 刘华平,李昕,郑宇,等.一种改进的自适应子带谱熵语音端点检测方法[J].系统仿真学报,2008,20(5):1366-1371.
[7] Shen J L,Hung J W,Lee L S.Robust entropy-based endpoint detection for speech recognition in noisy environments[C]//ICSLP'98,1998:232-235.
[8] Wu Bing-fei,Wang Kun-ching.Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments[J].IEEE Transactions on Speech and Audio Processing,2005,13(5):762-775.
[9] 陈振标,徐波.基于子带能量特征的最优化语音端点检测算法研究[J].声学学报,2005,30(2):171-176.

(上接 142 页)

实验结果表明,当 AXML 模式定义中包含较多函数节点数目时,算法依然能够有效地完成对模式无环性进行判定。

4 结语

研究了 AXML 文档物化终止性判定问题,提出了约束条件下的多项式时间判定算法。所考虑的 AXML 文档包含的 Web 服务调用之间是相对独立的,当服务调用之间相互依赖时,其终止性问题如何判定是未来工作的重点。

参考文献:

[1] Abiteboul S,Benjelloun O,Manolescu I,et al.Active XML:Peer-to-Peer data and Web services integration[C]//Proceedings of 28th International Conference on Very Large Data Bases,Hong Kong,2002.New York,NY,USA:ACM Press,2002:1087-1090.
[2] Milo T,Abiteboul S,Amann B,et al.Exchanging intensional XML data[C]//Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data,2003.New York,NY,USA:ACM Press,2003:289-300.
[3] Abiteboul S,Benjelloun O,Milo T.Positive active XML[C]//Proceedings of the 23rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems,2004.New York,NY,USA:ACM Press,2004:35-45.
[4] Fagin R,Kolaitis P G,Miller R J,et al.Data exchange:Semantics and query answering[C]//Proceedings of the 9th International Conference on Database Theory,2002.London,UK:Springer-Verlag,2002:207-224.