自然语言文本语义接受度的在线系统评价研究

杜家利,于屏方
DU Jia-li,YU Ping-fang

鲁东大学 外国语学院 教育部汉语辞书研究中心,山东 烟台 264025
Chinese-Based Center for Dictionary Research Directly under the National Education Ministry,School of Foreign Languages,Ludong University,Yantai,Shandong 264025,China
E-mail:dujiali68@yahoo.cn

DU Jia-li,YU Ping-fang.Research on Internet-based systematic evaluation of semantic accessibility scale originating from natural language texts.Computer Engineering and Applications,2008,44(26):141-143.

Abstract: Semantic accessibility scale can be considered a fully-understood measure of natural language texts.This essay,on the basis of methods of automatic text summarization system and literary style analyses,puts forward a formula by which data-based SAS originating from English literature texts can be evaluated on line.It is by means of discussion about the corpus from *The Old Man and the Sea* written by Nobel winner that the conclusion testifies to the fact that no distinctiveness among different sampling ratios can be discovered.Despite the existence of domain limitation,domain simplicity and relativity of evaluation on line,SAS formula will benefit the literary critics who have access to internet-based English texts.

Key words: natural language text;semantic accessibility scale;systematic evaluation;automatic text summarization system;literary style

摘 要:语义接受度(SAS)是衡量自然语言文本可理解程度的标尺。通过结合自动文摘系统评价方法和文体学分析方法,提出了可用于在线分析英语文学文本SAS的系统评价公式,并通过诺贝尔文学获奖作品《老人与海》的语料分析验证了公式的可适用性:不同的抽取率不会引起评价值的显著差异。尽管存在域的有限性、域的单一性和在线评价相对性等不足,此公式为文学评论者借助网络进行英语文本SAS在线评价提供了便利。

关键词:自然语言文本;语义接受度;系统评价;自动文摘系统;文体
DOI:10.3778/j.issn.1002-8331.2008.26.043 **文章编号:**1002-8331(2008)26-0141-03 **文献标识码:**A **中图分类号:**TP301.2

1 引言

语义接受度(Semantic Accessibility Scale)是用来衡量文本可理解程度的量化标准;如何从文体学视角分析文学文本的研究已受到国内外学者的关注[1-2]。在自然语言文本分析方面,自动文摘系统(Automatic Text Summarization System)与此有异曲同工之妙。文体学的SAS强调以点代面,通过形式化分析随机抽取到的英语文本内容来总结出某作家的写作风格。自然语言的ATSS侧重以主概次,利用计算机自动提取文本主要内容并生成语义连贯的文摘以涵盖原文。虽然SAS源数据选自文学文本而ATSS选自网络文本,但两者都强调利用形式来分析概括文本,具有可比性。随着网络发展,海量文学文本已经电子化。如何借助自然语言处理领域的研究成果进行文学文本的在线分析评价,已成为计算语言学研究的新方向。本文将电子化的英语文学文本统称为自然语言文本,对其可理解程度的评价可借鉴自动文摘系统的评价模式展开。同时,尝试构建英语文本SAS的在线评价系统,并利用公式进行量化,以便于文学评论者借助计算机对英语文本进行可受度的形式化分析。

2 形式化文本分析和现行文摘评价方法

形式化英语文本分析溯源于西方文体学,无论是Bally[3]的"文本分析需要提倡科学性和系统性"的观点,还是Spitzer[4]的对文本偏离常规的语言特征进行形式分析的"语文圈"思想,都推动了文学文本的量化分析,随后的Dolezel and Bally[5],Kucera and Francis[6],Leech and Short[2]等都对此衣钵进行了传承。国内从量化角度对英语文本进行分析的学者也日益增多,曹萍[7]、吴利琴[8]、张厚振[9]等也都注重了形式化对文本分析的可验证性。

与此相对应,随着网络发展和电子文本的增多,自动文摘研究取得了长足发展,基于该系统的评价研究也在国内外学术界占有一席之地,使系统的有效性、可用性和可理解性等都得到了较好地验证。

文摘评价系统主要分人工和机器评价,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[10]、问答评价[11]和阅读性理解评价[12]。这些方法虽因惯性思维和经验思维的影响含有较多可变因素,但为

基金项目:教育部项目(No.06JJD740007);山东省项目(No.07CWXJ03);鲁东大学项目(No.W20072602)。
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:应用语言学和计算语言学。
收稿日期:2007-12-28 **修回日期:**2008-03-04

机器的自动评价提供了平台。

常见的机器评价方法有:(1)句子召回率和精确度分析法;(2)F-Measure 分析法;(3)Rouge 分析法;(4)F-New-Measure 分析法。

法(1)中设定三个参数:文本句数 A,机器摘要句数 B,机器摘要中所含有的原文本句数 C,并通过三者的比值确定精确度和召回率。

公式 1:精确度(Precision):$P=\frac{C}{B}$

公式 2:召回率(Recall):$R=\frac{C}{A}$

这种方法将两个评价标准分开处理,割裂了两者的关联性,法(2)则推进一步,综合考虑 P 和 R 的联动性,但对文本不同压缩率对评价结果的影响没有在公式中体现。

公式 3:F-Measure=$\frac{2}{\frac{1}{P}+\frac{1}{R}}$

法(3)是更为复杂的系统评价方法,涉及 Rouge-L,Rouge-N 和 Rouge-W 三种 Rouge 评价,即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,既考虑了最长公共子序列所带来的相似程度,也计算了带权重的最长公共子序列,并推导出机器文摘和人工文摘中同时出现的最大单词数目。此方法虽具有很高的可信度但参数过多,操作困难。

法(4)兼顾了 F-Measure 易操作的优势,并参考 Rouge 分析法参数翔实的特点提出新的测算方法,即在 F-Measure 公式基础上添加了一个新的参数-压缩率 C(compression),它是机器摘要长度 $L1$ 与文本总长度 L 之比,并将其融合到 F-New-Measure 公式中。

公式 4:$C=\frac{L1}{L}$

公式 5:F-New-Measure=$\frac{2(1-C)}{\frac{1}{P}+\frac{1}{R}}$

此法在非受限领域的自动文摘系统中进行了验证,并对照了 4 种评价方法的数值差异,最后得出在压缩率不同的情况下,操作较为简便的 F-New-Measure 对文摘的评价具有稳定性[13]。

现行的形式化英语文本分析落后于自动文摘评价系统研究的发展,主要体现在 6 个方面:(1)目前尚未形成统一的、可验证的量化公式;(2)对文本的量化尚停留在人工计算数值的低级层面;(3)对数值分析缺乏域的概念,往往是简单枚举为主;(4)对同一文本的量化因抽取范围不同而出现较大偏差;(5)对不同文本的形式讨论因文本长度不同易产生数值虚化,不能真正代表原文本语言特点的全貌;(6)对文本语义接受度的在线评价研究尚属于拓荒之地,文本分析者对信息检索、提取等领域知识的缺乏使其形式化分析成为无源之水。

鉴于此,融合文本形式分析定性之长和自动文摘评价定量之优,尝试提出对英语文本语义接受度的在线评价方法。

3 语义接受度在线评价方法

在英语文本语义接受度研究方面,公式 Fog Index=0.4(L+H)[1]可用来分析可理解程度(与 SAS 成反比例)。其中 L 是平均每句所含单词的数量,H 是平均每百个单词所含的三音节或以上单词数量(不含屈折变化所致的多音节)。此公式考虑了句长和词长对文本的影响,而且数值能通过系统获得,所以可应用于文本 SAS 的系统在线评价。

通过分析 F-New-Measure 的测算思路可知,抽取率是决定系统评价稳定与否的重要变量。文本抽取涉及句数 $S1$ 和词数 $W1$ 两个变量,它们与总句数 S 和总词数 W 之比可作为系统评价的参数(Parameter),设为 $P1$ 和 $P2$,并设定文本的抽样率(Sampling Ratio)为 SR。

公式 6:$P1=\frac{S1}{S}$

公式 7:$P2=\frac{W1}{W}$

公式 8:$SR=\frac{2}{\frac{1}{P1}+\frac{1}{P2}}=\frac{2\times P1\times P2}{P1+P2}$

系统评价除需要考虑文本 SAS 与 Fog Index 成反比例之外,还需要考虑抽取率 SR 不同对系统评价的影响。借鉴 F-New-Measure 的公式推测方法,将抽取率中的两个变量 $P1$ 和 $P2$ 融合到公式中,并推导出 SAS 在线评价的理论公式:

公式 9:SAS=$\frac{1}{\text{FogIndex}}\times\frac{P2}{P1}=\frac{1}{0.4(L+H)}\times\frac{P2}{P1}$

其中 L 代表文本的平均句长;H 代表文本每百词中三音节或以上单词数量;$P1$ 代表文本句数的抽取率;$P2$ 代表文本词数的抽取率。此公式只具有理论上的可行性,即当单位句长和多音词数值较高时,理解文本的难度加大,SAS 数值降低,而其数值的变化还受到文本句数和词数抽取率的影响。如何将理论实践化是公式运用的前提,对 SAS 的验证将决定其可推广程度。

4 语义接受度在线评价公式的验证

从英语语料库中抽取海明威 1954 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)为语料,分别抽取总数为 20 页、30 页、40 页、50 页、65 页和全文 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,则公式 9 具有可推广性。

取样按照随机数表[14]方式进行,数值划定示例如下:

81833	93449	57781	94621	90998
37561	59688	93299	27726	82167
63789	54958	33167	10909	40343
81023	61590	44474	39810	10305
61640	81740	60986	12498	71546
42249	13812	59902	27864	21809
42243	10153	20891	90883	15782
98167	86837	99166	92143	82441
45236	09129	53031	12260	01278
14404	40969	33419	14188	69557

因被试文本总体规模是含有三位数的 124 页,所以横向抽取上面数值每三位为实际研究页码 $N(1\leqslant N\leqslant 124)$,并依次将不重复的 N 值设定为 20、30、40、50 和 65 五个组作为源页码(不足部分可按照随机数表顺次查找),并且和全文 124 页进行对比研究[1],见表 1~表 7。

由表 7 数值可知,当抽取率 SR 分别为 16%,24%,32%,39%和 52%时,其各自 SAS 值与全文 SAS 值没有显著差异,即对同文本的系统评价不会因抽取率不同发生评价性差异,公式 9 符合系统评价的要求,具有理论可行性和实践可验证性。

[1] 各样组的 Fog Index 取均值。

表 1 20 页组数据统计

抽取页	句总数	单词数	*L*	多音数	*H*	Fog Index
12	19	174	9.16	4	2.30	4.58
22	8	230	28.75	8	3.48	12.89
30	11	215	19.55	11	5.12	9.86
44	12	210	17.50	8	3.81	8.52
…	…	…	…	…	…	…
99	19	209	11.00	6	2.87	5.55
101	21	240	11.43	8	3.33	5.90
103	10	210	21.00	6	2.86	9.54
112	24	210	8.75	4	1.90	4.26

表 2 30 页组数据统计

抽取页	句总数	单词数	*L*	多音数	*H*	Fog Index
6	18	150	8.33	4	2.67	4.40
9	20	182	9.10	7	3.85	5.18
12	19	174	9.16	4	2.30	4.58
13	21	180	8.57	6	3.33	4.76
…	…	…	…	…	…	…
119	14	209	14.93	6	2.87	7.12
121	25	169	6.76	5	2.96	3.89
123	15	187	12.47	5	2.67	6.06
124	11	98	8.91	3	3.06	4.79

表 3 40 页组数据统计

抽取页	句总数	单词数	*L*	多音数	*H*	Fog Index
1	7	192	27.43	5	2.60	12.01
3	12	213	17.75	9	4.23	8.79
5	17	177	10.41	7	3.95	5.75
6	18	150	8.33	4	2.67	4.40
…	…	…	…	…	…	…
117	21	213	10.14	6	2.82	5.18
120	20	170	8.50	7	4.12	5.05
122	27	171	6.33	5	2.92	3.70
124	11	98	8.91	3	3.06	4.79

表 4 50 页组数据统计

抽取页	句总数	词总数	*L*	多音数	*H*	Fog Index
2	18	207	11.50	4	1.93	5.37
3	12	213	17.75	9	4.23	8.79
4	16	187	11.69	6	3.21	5.96
6	18	150	8.33	4	2.67	4.40
…	…	…	…	…	…	…
112	24	210	8.75	4	1.90	4.26
113	22	218	9.91	3	1.38	4.51
118	16	220	13.75	5	2.27	6.41
124	11	98	8.91	3	3.06	4.79

表 5 65 页组数据统计

抽取页	句总数	词总数	*L*	多音数	*H*	Fog Index
2	18	207	11.50	4	1.93	5.37
6	18	150	8.33	4	2.67	4.40
7	12	220	18.33	4	1.82	8.06
8	18	209	11.61	2	0.96	5.03
…	…	…	…	…	…	…
119	14	209	14.93	6	2.87	7.12
120	20	170	8.50	7	4.12	5.05
122	27	171	6.33	5	2.92	3.70
124	11	98	8.91	3	3.06	4.79

表 6 124 页组数据统计

抽取页	句总数	词总数	*L*	多音数	*H*	Fog Index
1	7	192	27.43	5	2.60	12.01
2	18	207	11.50	4	1.93	5.37
3	12	213	17.75	9	4.23	8.79
4	16	187	11.69	6	3.21	5.96
…	…	…	…	…	…	…
121	25	169	6.76	5	2.96	3.89
122	27	171	6.33	5	2.92	3.70
123	15	187	12.47	5	2.67	6.06
124	11	98	8.91	3	3.06	4.79

表 7 不同 SR 对应 SAS 值统计

表号	*S*1	*W*1	*P*1	*P*2	*SR*	Fog Index	*SAS*
1	306	4 433	0.16	0.17	0.16	7.29	0.14
2	451	6 236	0.24	0.23	0.24	7.04	0.14
3	622	8 298	0.33	0.31	0.32	7.12	0.13
4	730	10 714	0.38	0.40	0.39	7.63	0.14
5	996	13 967	0.52	0.53	0.52	7.16	0.14
6	1 900	26 587	1.00	1.00	1.00	7.13	0.14

5 结论

自然语言文本语义接受度(SAS)研究是计算语言学研究的一个方向,本文以英语文本为语料,结合自动文摘系统评价方法和文体学分析方法进行文本可理解程度的量化探索,并提出了具有理论可行性的 SAS 公式。同时,通过《老人与海》的语料分析使公式实践性得到了验证:抽取率 SR 与 SAS 不具有显著相关性;相关参数可由系统获得且具有在线评价潜能。这为文学评论者通过抽样对文本进行在线可受度分析提供了便利。

本文虽基于英语文本提出并验证了 SAS 公式, 但其推广尚有一定难度,体现在:(1)域的有限性:网络语不局限于屈折语,作为孤立语(如汉语)和粘着语(如日语)的语言是否适用于此公式未得到验证;(2)域的单一性:同属屈折语的法、德等语言也未为此公式提供支撑;(3)在线评价相对性:虽然字处理系统工具栏中具有自动计词功能,但尚无计句和计算多音词的功能,这使 SAS 公式的在线自动评价不具有绝对性。

参考文献:

[1] 孙华祥.从《乞力马扎罗的雪》看海明威的文体风格[J].外国文学研究,1999,1:104-108.

[2] Leech G N,Short M H.Style in fiction:a linguistic introduction to English fictional prose[M].Beijing:Pearson Education Limited/Foreign Language Teaching and Research Press,1981/2001:113-117.

[3] Bally C.Traite de stylistique francaise[M].3rd ed.Paris:Klincksieck,1951.

[4] Spitzer L.Linguistics and literary history[M].Princeton:Princeton University Press,1948.

[5] Dolezel L,Bally R W.Statistics and style[M].New York:American Elsevier Publishing Co,1969.

[6] Kucera H,Francis W N.Computational analysis of present-day american english[M].Providence:Brown University Press,1967.

[7] 曹萍.海明威短篇小说《杀人者》的语言风格[J].外国文学研究,1997(4).

[8] 吴利琴.从短篇小说 *The Killers* 的句型统计看海明威的"冰山理论"[J].山西教育学院学报,2002(2).

[9] 张厚振.基于语料库的海明威作品《一个干净明亮的地方》分析[J].新乡教育学院学报,2004(2).

[10] Minel J L,Nugier S,Piat G.How to appreciate the quality of automatic text summarization[C]//Proc of the ACL/EACL'97,1997:25-30.

(下转 157 页)