国内外机器自动评分系统评述*
——兼论对中国学生翻译自动评分系统的启示

王金铨[1]　　文秋芳[2]

提要：本研究回顾了国内外机器自动评分系统的现状、内容和特点，探讨了现有机器自动评分技术对中国学生翻译自动评分系统的启示。在总结前人成功经验的基础之上，将语料库技术、自然语言处理技术和统计技术结合起来，深入挖掘能够反映译文质量的文本预测变量，建成具有中国特色的翻译自动评分系统应该为期不远。

关键词：自动评分系统；机器翻译自动评价；翻译自动评分系统

Abstract: This paper reviews the contents and characteristics of the existing automated scoring systems and provides some enlightenment for the construction of the computer-assisted scoring system of Chinese EFL learners' translation. Through the integration of corpus linguistics, natural language processing and statistics, more effective predictors can be located and the completion of the computer-assisted translation scoring system is not far away.

Key words: automated scoring system; automatic machine translation evaluation; computer-assisted translation scoring system

中图分类号：H319　**文献标识码**：B　**文章编号**：1004-5112(2010)01-0075-08

1. 引言

计算机技术在语言测试中已经得到广泛运用。客观题测试自不待言，随着技术的不断进步，主观题测试也开始实行自动评分。Chapelle和 Doughlas (2006: 2)更是明确指出，如今语言测试技术已经得到飞速发展，无论学习者置身何处，总免不了需要参加一些计算机辅助的语言测试。作文自动评分技术从 1966年问世以来，日臻成熟，美国教育考试服务中心 (Educational Testing Service, ETS)研制的 E-rater目前已经完全商业化，被正式用于 GMAT等大规模语言测试 (Marina 2005: 104)。国内，梁茂成 (2005)对中国学生英语作文自动评分进行了有益的尝试，并取得了初步成果。但是，翻译自动评分仍处于探索阶段，目前国内外尚无成熟的人工译文自动评分系统。本研究将回顾国内外机器自动评分系统的现状、内容和特点，并探讨现有自动评分技术对中国学生翻译自动评分系统的启示。

2. 机器自动评价系统概述

由于人工译文自动评分系统的缺失，本节分两部分回顾机器自动评价系统：第一部分对作文自动评分系统进行概述，第二部分对机器译文自动评价系统进行概述。这两部分的技术是构建中国学生翻译自动评分系统的基础，值得深入探讨和挖掘。

* 本研究为教育部人文社会科学重点研究基地研究项目“大规模考试主观题 (英汉互译)自动评分系统的研制”(项目编号 07JJD740070)的部分成果。

2.1 作文自动评分系统概述

作文自动评分系统 AES(Automated Essay Scoring)是机器自动评分系统的先驱。AES的定义为能够对作文进行评价、评分的计算机技术 (Shermis & Barrera 2002; Shermis & Burstein 2003; Shermis *et al* 2003)。第一个机器自动评分系统是 Ellis Batten Page研制的 PEG (Project Essay Grader)作文评分系统。该系统诞生于 20世纪 60年代,后来由于种种实际困难在 70年代至 80年代初受到冷落。Page (2003: 45)用"休眠状态"(sleep mode)来描述这段时期的 PEG。到了 80年代末,PEG系统进入了"重新觉醒期"(reawakening period) (Page 2003: 45),并获得了新的机遇。90年代以后,出现了 IEA (Intelligent Essay Assessor)、E-rater (Electronic Essay Rater)和 IntellMetric等主流作文自动评分软件。IEA的研制在 90年代末期 (Hearst 2000; Jerrams *et al* 2001),使用了科罗拉多大学 Thomas K Landauer等创立的潜在语义分析 (Latent Semantic Analysis, LSA)方法。E-rater是 Burstein等研制的作文自动评分系统 (Burstein *et al* 1998; Burstein *et al* 2001),它把语料库和自然语言处理技术 (Natural Language Processing, NLP)应用于系统中实现作文自动评分。E-rater已经商业化,美国教育考试服务中心用其对 GMAT (Graduate Management Admission Test)、AWA (Analytical Writing Assessment)中的作文题进行评分,是一个比较成熟的作文自动评分系统 (Harold & Ray 2003: 227)。IntellMetric是由 Vantage Learning开发的基于人工智能 (Artificial Intelligence, AI)和自然语言处理技术的作文评分系统 (Shermis & Barrera 2002; Elliot 2003b; Charles *et al* 2006)。在国内,梁茂成 (2005)研制了适合中国英语学习者的作文自动评分系统;该系统吸收了国外自动评分系统的长处,兼顾中国英语学习者的特点。上述系统都是比较成熟的作文评分系统,其中一些做法在构建翻译评分系统时很值得借鉴。

2.2 机器翻译评价系统概述

翻译领域的自动评价系统基本上以机器译文评价系统 (Machine Translation Evaluation)为主。这些系统的功能用来比对被测译文 (candidate translation)和参考译文 (reference translation,一般为人工译文)之间的相似程度,匹配的词语越多,被测译文的得分就越高。还有一些研究者 (Brew & Thompson 1994; Rajman & Hartley 2001)认为不能只看译文词语的匹配率,还需要考虑词语顺序,顺序匹配的词语得分应该高于未按顺序排列的词语。例如,"a lot of"的得分应该高于"lot of a"。机器译文评价系统的评价方法主要有以下几种。

1)完全匹配法。只有与参考译文相同的部分才算正确,如 SER (Sentence Error Rate)方法。SER将被测译文与参考译文中相异句子的比例作为衡量机器译文质量的标准,在 SER方法中参考译文只有一篇。

2)编辑距离法。计算被测译文与参考译文之间最小编辑距离 (minimum edit distance),即考查字母增删、替换、插入的次数来衡量译文质量的方法 (Levenshtein 1966; Tillmann *et al* 1997; Vidal 1997),如 WER (Word Error Rate)、mWER (Multiple Reference WER)等方法 (NieSen *et al* 2000; Leusch *et al* 2003)。WER、mWER使用的评价方法是编辑距离越小,被测译文质量越高,也就是说 WER、mWER值越小,译文的质量越高。WER和 mWER之间的不同之处在于 WER中只有一篇参考译文,而 mWER有数篇,编辑距离以被测译文与数篇参考译文之间最小的编辑距离计算 (Niessen *et al* 2000)。

3)基于 N元组模型的方法。在机器翻译评价研究中,基于 N元组的评测方法应用最为广泛,其中主要的有两种:IBM提出的 BLEU (Bilingual Evaluation Understudy)评测方法 (Papineni

& Roukos 2002)和 NIST(National Institute of Standards and Technology)评测方法(Doddington 2002)。这两个评测方法都是从文本的 N-gram(N元组)出发,比对被测译文和参照译文之间N元组的相似程度,计算出被测译文和参考译文之间的距离。BLEU的核心思想是"机器译文与专业人工译文越接近,其质量越好"(Papineni & Roukos 2002)。BLEU默认匹配 4元组以下的相似度,超出 4元组的连续词语由于频率低对自动评价的意义不大。

BLEU公式可以计算出被测译文和参考译文之间的相似程度,确定被测译文的质量。但是,BLEU公式对所有词汇采取了一视同仁的做法,即所有词汇的权重相同。根据翻译经验,译文中信息的权重应该是分等级的,有的信息属于不可或缺类型,而有的信息则可有可无。NIST是在 BLEU基础上发展而成的评测方法,基本参数与 BLEU类似,区别在于 NIST考虑了不同信息的权重系数,减少了评测方法对译文质量产生的影响。根据 NIST标准,如果一个 N元组在文本中出现的次数越少,表明它所包含的信息量越大,该 N元组就应该被赋予更高的权重。BLEU和 NIST是国际上机器自动翻译评价系统的流行指标,得分越高,表明译文质量越好。根据国内外研究,这两种评测方法与人工评分的相关度较高,能够在一定程度上反映翻译质量的优劣(Brew & Thompson 1994; Doddington 2002; Papineni & Roukos 2002; Zhang *et al* 2004)。

3. 现有机器评分系统评价及其对中国学生翻译自动评分系统的启示

本节分为三个部分:第一部分对有代表性的作文机器评分系统进行评价;第二部分对主流机器翻译评价系统进行评价;第三部分总结现有机器评分系统对中国学生翻译自动评分系统的启示。

3.1 作文自动评分系统

本部分主要评价国外的 PEG、IEA、E-rater三种比较有代表性的作文自动评分系统以及国内梁茂成(2005)构建的中国学生英语作文自动评分模型。

PEG是历史最悠久的作文自动评分系统,从 1966年诞生以来一直在不断更新。PEG评分需要事先在训练集中建立评分模型,然后运用评分模型为其他同题作文评分。建立模型时,PEG首先从文本中提取表面特征(surface feature),如用文本长度(essay length)表示流利度(fluency),介词数、代词数以及其他词性标记表示句子结构(sentence structure),单词长度差异表示用词(diction)等(Salvatore *et al* 2003: 320 - 321)。获得上述文本特征以后,系统以人工评分作为因变量、文本特征作为自变量进行多元回归分析得到一个回归方程,该回归方程中的变量都与作文成绩相关,能够预测作文得分。PEG系统把统计分析和文本表面特征结合起来对作文进行评分,因而在作文自动评分史上有着独特的贡献。不过,PEG系统使用的文本特征都是与语言的形式特征有关,对语义内容束手无策。

IEA是 Thomas K. Landauer等开发的作文自动评分系统,其评分原理主要依靠潜在语义分析(LSA)计算被测文本与已评分文本之间的语义相似度。潜在语义分析始于信息检索领域,是利用数据统计以及数学分析方法,通过观察词项在同一文本中出现的相对频率来计算语义的相似度(王金铨等 2007: 406)。IEA把潜在语义分析方法引入作文自动评分领域,通过矩阵分析方法分析比较文本的语义相似度来确定文本得分。

Laham (1997)、Landauer等(1998)、Turney (2005)等运用 LSA方法做过不少与语义相关的研究。他们得出的结论显示,通过 LSA分析句子间的相似度,可以有效避免自然语言中多

词同义和一词多义问题,提高语义分析的准确度,进而提高评分质量。潜在语义分析认为,"文档中出现两个或者更多词条不是偶然事件"(Manning *et al* 2005: 344)。LSA计算方法被广泛运用于信息检索领域,它甚至可以用来计算不同语言之间的相似度(Dumais *et al* 1996; Dumais *et al* 1997; Rosario 2000)。在国内语言学界,桂诗春(2003)曾运用 LSA方法对 CLEC中学生作文的失误进行了研究;梁茂成(2005, 2006)曾经采用 LSA方法分析了文本的语义内容和语篇连贯性;王金铨等(2007)运用 LSA方法计算语句间的形式和语义相似度,证明了该方法的有效性和实用性。

IEA运用潜在语义分析方法对作文进行评分,充分挖掘了文本中的语义相似度,但是对于文本的形式特征缺乏考虑。

E-rater与 PEG建模的过程相同,都需要事先对训练集进行评分,然后提取文本特征,运用统计分析获得自动评分模型。E-rater综合运用统计分析和自然语言处理技术,既关注文本的形式特征,也提取文本的语义内容。与 PEG相比,E-rater的优点在于分模块设计。E-rater具有独立的句法模块、语篇模块和主题分析模块(Marina 2005: 104)。根据 Burstein (2003: 116 - 118)的观点,句法模块是通过词性赋码的方法实现文本句法分析;语篇模块是通过识别语篇连接词语和句法结构的方法分析文本中的语篇关系;主题分析模块是借助向量空间模型(Vector Space Model, VSM)分析文本的词汇使用来判定内容相似程度。不过,Dodigovic (2005: 104)指出,与潜在语义分析相比,E-rater对于近义词的识别能力较差,影响评分结果。与 PEG和 IEA相比,E-rater考查了作文的三个主要方面:句法、语篇和主题,因而它的结构效度较好。

梁茂成(2005)构建的中国学生英语作文自动评分模型综合了上述三个主流英语作文自动评分系统的优点,从语言、内容和结构三个方面衡量作文质量。语言模块又分为三个分模块:流利度、复杂度以及地道性;内容模块的考查借助潜在语义分析等变量;结构模块的考查则借助作文中的话语连接词等变量。梁茂成(2005)的评分模型从多维度考查了英语作文的质量,实践证明该模型能够有效地对中国英语学习者的作文进行评分。

3.2 机器翻译自动评价系统

本部分主要讨论机器翻译自动评价系统的两个代表性评测方法:BLEU和 NIST。

BLEU和 NIST这两种评测方法实质上是对机器被测译文和参考译文在 N元组(1—4元组)层次上进行匹配,如出现同样的 N元组,则默认为一致词项,赋值为 1,否则赋值为 0。Papinenl和 Roukos (2002)、Doddington (2002)证实了这种方法计算的相似度结果与人工判断之间的相关系数较高,表明该方法确实能够在某种程度上检验被测译文与参考译文之间的相似度,但是这种相似度检验方法存在着一个致命弱点。自然语言中存在大量的多词同义(synonymy)和一词多义(polysemy)现象,单纯的词形匹配无法计算译文之间的相似度;自然语言之间的相似度不可能只是一个 0和 1的真假二值的判断,人工译文的复杂性和灵活性要远远超过机器译文。更何况,BLEU和 NIST未提及对译文中的单词进行词形还原(lemmatisation)处理,使得同一个单词的不同形式被归于不同单词之列,进一步影响了算法的信度和效度。因此,该方法不能单独用来检验被测译文的质量。在自动评价系统中使用该方法必须具备两个条件:(1)参考译文要足够多,能够覆盖可能出现的 N元组;(2)同时使用其他相似度测量工具,互为补充。

3.3 现有机器评分系统对中国学生翻译自动评分系统的启示

现有机器评分系统分别从语言、内容、结构三方面对文本质量进行测量。语言测量主要挖掘形式特征,如 PEG系统中的文本长度、介词数、代词数以及句子结构等;内容测量主要研究语义相似度,如 IEA中的潜在语义分析、E-rater中的向量空间模型(VSM);结构测量则是通过作文中的话语连接词等变量考查文本的语篇特征,如 E-rater中的语篇模块和梁茂成(2005)评分模型中的结构模块。机器翻译自动评价系统中的一些评价方法也很值得借鉴,如 N元组,该变量实际上更偏重于考查译文的语义内容,可以纳入内容模块。表 1对各个自动评分系统的测量维度进行了综合比较。

表 1 各评分系统测量维度比较

		PEG	IEA	E-rater	梁茂成(2005)评分模型	BLEU, NIST
测量方法	训练集	有	有	有	有	无
	多元回归	有	无	有	有	无
	模块化设计	无	无	有	有	无
测量内容	语言	有	无	有	有	有
	内容	无	有	有	有	有
	结构	无	无	有	有	无

3.3.1 测量方法对本研究的借鉴作用

首先,从测量方法来看,PEG系统率先在评分系统中使用事先经过评分的文本作为训练评分系统的材料,使得构建系统成为一个有指导的训练,能够提高评分系统的信度和稳定性。其他作文评分系统也使用了训练集作为构建模型的基础。严格来说,在机器翻译自动评价系统中使用的参考译文不能等同于作文评分系统中的训练文本,因为参考译文只是作为一个参照物与被测译文进行比较,并未纳入模型构建过程。本研究构建的翻译自动评分模型将借鉴作文自动评分系统的做法,创建训练集作为模型构建的基础。

PEG、E-rater和梁茂成(2005)的作文自动评分模型都使用了多元回归分析的统计方法。多元回归分析是研究因变量(Y)和多个自变量(IVs)之间依存关系的统计方法。在构建评分系统时,多元回归分析能够考查自变量(预测因子)对因变量(人工评分)的影响作用,从而得到能够预测分值的回归方程。机器翻译自动评价系统没有使用多元回归分析,也没有使用其他统计手段。实际上,多元回归分析在模型构建过程中的作用是不可替代的,它可以进行多种分析,除了能够分析因变量与自变量之间的相互影响(多元回归),还可以分析因变量和自变量之间的关系(相关关系)、自变量之间的关系(偏相关关系),确定进入回归方程的自变量,运用回归方程进行预测等。本研究的译文自动评分系统构建将借鉴作文自动评分的做法,采用多元回归分析方法构建模型。

E-rater和梁茂成(2005)采用了模块化设计。E-rater中共有三个模块:句法模块、语篇模块和主题分析模块。梁茂成(2005)的评分模型中也有三个模块:语言模块、内容模块和结构模块;语言模块对应 E-rater中的句法模块,内容模块对应主题分析模块,结构模块对应语篇模

块。这些模块涵盖了作文质量评判的各个方面:语言模块主要考查作文语言形式的准确性;内容模块考查作文是否紧扣主题;结构模块考查作文的语篇要素,决定其能否成为一个独立语篇。这三个方面都是一篇作文应该具备的,通过提取相应的文本特征对评分模型进行有指导的训练,可以使机器获得评判文本优劣的能力。模块化设计考虑了作文质量评判的方方面面,结构效度较好,因而训练出的评分模型信度更高、稳定性更好。

3.3.2 测量内容对本研究的借鉴作用

从测量内容来看,E-rater和梁茂成(2005)的评分模型对文本的语言、内容和结构都进行了测量。机器翻译评价系统只是通过文本中的 1—4元组来比较被测译文和参考译文。从严格意义上来说,机器翻译评价系统的评测方法更偏重于评测意义,因为 1—4元组的匹配率主要反映意义。不过在任何语言中,意义和形式是不可分割的。在表现意义的同时,1—4元组也能够反映一部分形式,不过形式的成分要小于意义。即使在表现意义方面,N元组也是不够的,因为在语言中有很多近义词,不能用非此即彼的方式进行排除。

综上所述,中国学生翻译自动评分系统为了保证模型的信度和效度,首先需要采用模块化设计,对译文质量进行全方位的测量;其次,应该借鉴作文自动评分系统和机器译文评价系统的成熟做法,在模型构建过程中不断扬弃,保证评分模型的性能;最后,多元回归方法是模型构建的有效统计手段,值得在翻译自动评分系统中尝试。

4. 结语

本文分析回顾了国内外现有机器自动评分系统的优缺点,探讨了现有机器自动评分系统对中国学生翻译自动评分系统的启示以及可以借鉴的方方面面。在总结以往成功经验的基础之上,将语料库技术、自然语言处理技术和统计技术结合起来,深入挖掘能够反映译文质量的文本预测变量,建成具有中国特色的翻译自动评分系统应该为期不远。

参 考 书 目

[1] Brew C & Thompson H. *Automatic Evaluation of Computer Generated Text: A Progress Report on the TextEval Project* [R]. Plainsboro, NJ: The Workshop on Human Language Technology Workshop, 1994.

[2] Burstein J. The e-rater scoring engine: Automated essay scoring with natural language processing [A]. In Shermis M D & Burstein J (eds). *Automated Essay Scoring: A Cross-Disciplinary Perspective* [C]. Mahwah, NJ: Lawrence Erlbaum Associates, 2003.

[3] Burstein J, Kukich K, Wolff S, Lu C & Chodorow M. *Enriching Automated Essay Scoring Using Discourse Marking* [R]. Montreal, Canada: The Workshop on Discourse Relations and Discourse Marking, Annual Meeting of the Association of Computational Linguistics, 1998.

[4] Burstein J, Leacock C & Swartz R. Automated evaluation of essay and short answers [A]. In Danson M (ed.). *Proceedings of the Sixth International Computer Assisted Assessment Conference* [C]. Loughborough, UK: Loughborough University, 2001.

[5] Chapelle C A & Doughlas D. *Assessing Language through Computer Technology* [M]. Cambridge: Cambridge University Press, 2006.

[6] Charles A M, Steve G & Jill F. *Handbook of Writing Research* [M]. New York: Guilford Press, 2006. 409-410.

[7] Doddington G. *Automatic Evaluation of Machine Translation Quality Using N-Gram Co-Occurrence Statistics*

[R]. San Diego: The Human Language Technology Conference, 2002.

[8] Dumais S, Landauer T & Littman M. *Automatic Cross-Linguistic Information Retrieval Using Latent Semantic Indexing* [R]. Zurich: SIGIR96 Workshop on Cross-Linguistic Information Retrieval, 1996.

[9] Dumais S T *et al*. *Automatic Cross-Language Retrieval Using Latent Semantic Indexing* [R]. Palo Alto, CA: AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, 1997.

[10] Elliot S. Intellimetric: From here to validity [A]. In Shermis M D & Burstein J (eds). *Automated Essay Scoring: A Cross-Disciplinary Approach* [C]. Mahwah, NJ: Lawrence Erlbaum Associates, 2003.

[11] Harold F & Ray S. *Technology Applications in Education: A Learning View* [C]. Mahwah, NJ: Lawrence Erlbaum Associates, 2003. 227.

[12] Hearst M. The debate on automated essay grading [J]. *IEEE Intelligent Systems*, 2000, 15(5): 22 - 37.

[13] Jerrams S, Soh V & Callear D. *Bridging Gaps in Computerized Assessment of Texts* [R]. Madison, USA: The 2nd IEEE International Conference on Advanced Learning Technologies, 2001.

[14] Laham D. Latent semantic analysis approaches to categorization [A]. In Shafto M G & Langley P (eds). *Proceedings of the 19th Annual Conference of the Cognitive Science Society* [C]. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc, 1997.

[15] Landauer T K, Foltz P W & Laham D. Introduction to latent semantic analysis [J]. *Discourse Processes*, 1998, (25).

[16] Leusch G, Ueffing N & Ney H. *A Novel String-to-String Distance Measure with Applications to Machine Translation Evaluation* [R]. New Orleans, USA: MT Summit IX, 2003.

[17] Levenshtein V I. Binary codes capable of correcting deletions, insertions and reversals [J]. *Soviet Physics Doklady*, 1966, 10(8): 707 - 710.

[18] Manning C D & Schutze H. **苑春法等译. 统计自然语言处理基础** [M]. **北京: 电子工业出版社**, 2005.

[19] Marina D. *Artificial Intelligence in Second Language Learning: Raising Error Awareness* [M]. Buffalo, NY: Multilingual Matters, 2005.

[20] Niesen S *et al*. *An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research* [R]. Athens, Greece: The 2nd International Conference on Language Resources and Evaluation, 2000.

[21] Page E B. Project Essay Grade: Peg [A]. In Shermis M D, Burstein J (eds). *Automated Essay Scoring: A Cross-Disciplinary Perspective* [C]. Mahwah, NJ: Lawrence Erlbaum Associates, 2003.

[22] Papineni K & Roukos S. *BLEU: A Method for Automatic Evaluation of Machine Translation* [R]. Philadelphia: The 40th Annual Meeting of the Association for Computational Linguistics, 2002.

[23] Rajman M & Hartley T. *Automatically Predicting MT Systems Rankings Compatible with Fluency, Adequacy or Informativeness Scores* [R]. Santiago de Compostela, Spain: The Workshop on Machine Translation Evaluation: "Who Did What to Whom", 2001.

[24] Rosario B. *Latent Semantic Indexing: An Overview* [R]. *INFOSYS* 240, 2000.

[25] Salvatore V, Francesca N & Alessandro C. An overview of current research on automated essay grading [J]. *Journal of Information Technology Education*, 2003, (2): 319 - 330.

[26] Shermis M D & Barrera F. *Exit Assessments: Evaluating Writing Ability through Automated Essay Scoring* [R]. New Orleans: The Annual Meeting of the American Educational Research Association, 2002.

[27] Shermis M D & Burstein J. *Automated Essay Scoring: A Cross-Disciplinary Perspective* [M] Mahwah, NJ: Lawrence Erlbaum Associates, 2003.

[28] Shermis M D, Raymat M V & Barrera F. *Assessing Writing through the Curriculum with Automated Essay Scoring* [R]. Chicago: The Annual Meeting of the American Educational Research Association, 2003.

(下转第 91页)

[32] 亓鲁霞. 意愿与现实:中国高等院校统一招生英语考试的反拨作用研究 [M]. 北京:外语教学与研究出版社, 2004.
[33] 燕娓琴,谢小庆 (译). 教育与心理测试标准 [M]. 沈阳:沈阳出版社, 2003.
[34] 杨惠中. 语言测试和语言教学 [J]. 外语界, 1999, (1).
[35] 杨惠中. 导读 [A]. Alderson J C, Clapham C M, Wall D M著. 语言测试的设计与评估 [M]. 北京:外语教学与研究出版社, 2000.
[36] 杨惠中,桂诗春. 语言测试的社会学思考 [J]. 现代外语, 2007, (4).

作者单位: 上海交通大学外国语学院,上海 200240

(上接第 81页)

[29] Tillmann C *et al*. *Accelerated DP Based Search for Statistical Translation* [R]. Rhodes, Greece: The 5th European Conference on Speech Communication, 1997.
[30] Turney P D. *Measuring Semantic Similarity by Latent Relational Analysis* [R]. Edinburgh, Scotland: The 19th International Joint Conference on Artificial Intelligence, 2005.
[31] Vidal E. *Finite-State Speech-to-Speech Translation* [R]. Munich, Germany: The International Conference on Acoustics, Speech and Signal Processing, 1997.
[32] Zhang Y, Vogel S & Waibel A. *Interpreting Bleu/Nist Scores: How Much Improvement? Do We Need to Have a Better System?* [R]. Lisbon: The International Conference on Language Resources and Evaluation (LREC), 2004.
[33] 桂诗春. 潜伏语义分析的理论及其应用 [J]. 现代外语, 2003, (1): 76 - 84.
[34] 梁茂成. 中国学生英语作文自动评分模型的构建 [D]. 南京:南京大学, 2005.
[35] 梁茂成. 学习者书面语语篇连贯性的研究 [J]. 现代外语, 2006, (3): 284 - 292.
[36] 王金铨,梁茂成,俞洪亮. 基于 N-Gram和向量空间模型的语句相似度研究 [J]. 现代外语, 2007, (4): 405 - 413.

作者单位: 1. 扬州大学外国语学院,江苏 扬州 225009
2. 北京外国语大学中国外语教育研究中心,北京 100089

外教社引进出版《牛津英国文学百科全书》

《牛津英国文学百科全书》是牛津大学出版社于 2006年推出的重量级产品。为满足国内读者研习的需要,上海外语教育出版社最近原版引进出版了这套世界文库中的辉煌巨作。《牛津英国文学百科全书》共 5卷,收录了 500篇高质量的学术论文,涵盖了英国文学的历史与精华。《牛津英国文学百科全书》以新颖的方式呈现英国文学 1400年的发展全貌及其衍变特征,对英国文学史上的重要作家进行深度评析,阐述其生平、作品、学术和政治观点等,同时对英国文学的流派、作品体裁、当时的社会运动、对文学产生重大影响的事件以及其他重要主题也做了深入记述和探讨。全书内容翔实,条理清晰,对于英语专业教师和学生、英国文学的研究人员以及对英国文学感兴趣的读者来说,是一部不可多得的工具书。《牛津英国文学百科全书》的引进出版必将有效促进我国英语文学的教学与研究。