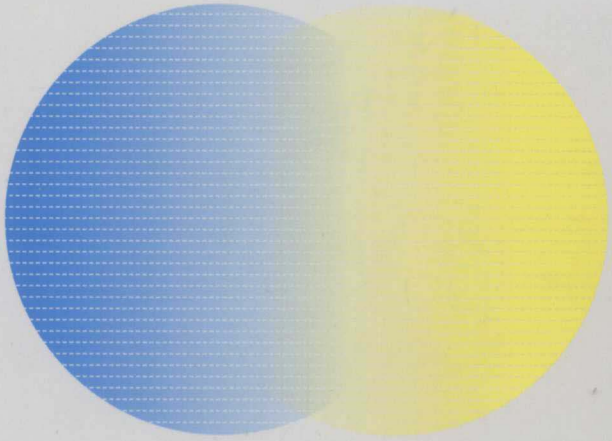
中国外语教育研究中心　外语考试自动评分研究系列丛书

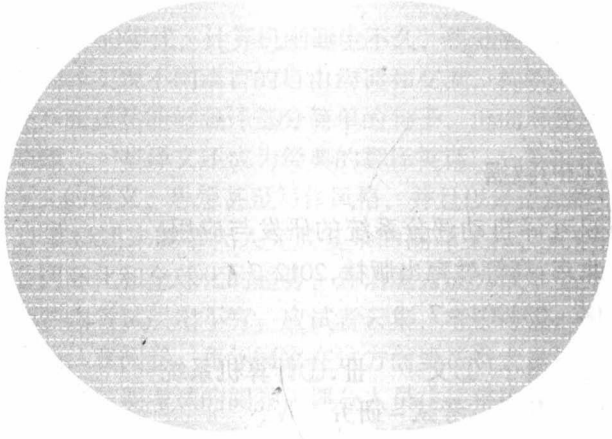


大规模考试英汉互译自动评分系统的研发与应用

秦颖　文秋芳　著

高等教育出版社
HIGHER EDUCATION PRESS

中国外语教育研究中心　外语考试自动评分研究系列丛书



大规模考试英汉互译自动评分系统的研发与应用

daguimo kaoshi yinghan huyi zidong pingfen xitong de yanfa yu yingyong

秦颖　文秋芳　著

高等教育出版社·北京
HIGHER EDUCATION PRESS　BEIJING

前 言

应用当代信息技术为外语教学与研究服务是大势所趋。计算机作为一种工具，正逐步承担起一些只有具备知识和智能的人才能从事的工作，改变着人们传统的工作思路和方式。

在20世纪50、60年代，计算机刚诞生不久，机器翻译的研究就开始了，人们希望通过机器翻译实现不同语言的自由沟通和交流，构筑人类的“巴比塔”。当时的机器翻译系统虽然能够翻译部分简单的句子，但面对复杂的语言现象，译文就变得晦涩难懂，一些译文还成为经典的翻译笑话。一篇好的译文，人们希望它能够忠实于原文的语义、思想甚至写作风格，并且以译语流畅地表达出来。

至今，计算机还未能替代人完成复杂的翻译工作，人们还是要努力学习外语，而且随着国际化和全球化的趋势，外语能力成为很多工作的必备技能。为衡量外语能力，各类考试层出不穷，应试者云集，考卷海量，更重要的是，还要组织大量评阅人员认真阅卷。每年国家在这方面要花费大量的时间和人力、物力、财力，不堪重负。重要考试的阅卷，评分人员往往集中在一起，连续数天奋战，十分疲劳，大家都迫切呼唤一种可以替代或减轻人工阅卷工作的方法。这种方法应该具有和人工阅卷可比拟的信度和准确度，同时有更高的效率和更廉价的消耗。实际上，客观题的评阅已经通过利用答题卡扫描实现了自动评阅，但像作文、翻译和简答等类型的主观题还必须人工评阅。

近年来，人工智能中自然语言处理技术的发展为自动评阅主观题答案提供了手段。英语作文自动评分已在TOFEL和GRE考试评阅中使用。在这种情形下，我们开始研究学习者翻译的自动评分，寻求解决外语考试中另一类主观题的自动评阅。

该项研究工作包括三大任务：一是研究数据的收集工作：收集整理大量真实的学生在外语考试中的译文，全部电子化，然后组织专业人员制定客观、便于操作的评分标准，并为每一篇译文进行人工评分；数据收集完成后的第二项任务是确定译文质量自动评价的技术路线，选择合适的方法，从理论上探究和译文质量相关的文本特征，研究哪些因素和译文质量相关，最终建立评价的理论数学模型；第三项任务就是在计算机上编程实现自动评分模型，构建完整的评分系统，并对系统的性能进行评测。

本书的内容围绕上述任务展开，分为理论研究篇和技术实现篇两部分。理论研究篇侧重语言学分析、语言测试的有关理论，从翻译质量的人工评价方法和已有的机器译文自动评测有关算法出发，探索适合评价学习者译文质量的理论模

型。技术实现篇侧重运用自然语言处理技术构建评分系统，介绍系统实现所涉及的具体技术问题，系统的构建原则和方法，并给出部分调试过的源代码程序供读者参考。

翻译自动评分是一项跨学科的综合研究，既包括翻译语言学、外语教学和外语测试的理论，又涉及计算机编程、语言信息处理的技术问题。自然，本书的读者也需要有相关知识的背景，或者有一方面的知识又对另一方面感兴趣。

本书在内容介绍上同时考虑了跨学科研究的因素，力求明确介绍相关概念，条理清楚地介绍实现步骤，程序代码添加必要的注释等等，让读者根据内容介绍就能够逐步学习建立一个翻译自动评分系统的框架，实用性强。

最后，本书大部分内容属于探索研究的课题，个别结论也是在现有数据基础上得出的，还需要进一步验证，恳请读者批评指正，共同推进翻译自动评分研究。

目 录

第一部分 理论研究篇 …… 1

第一章 绪论 ……1

1.1 语言质量自动评价及研究的意义 ……1
1.2 相关研究回顾 ……3
1.3 本书的内容及安排 ……5

第二章 翻译质量评价 ……6

2.1 翻译质量的人工评价标准 ……6
2.2 翻译质量的自动评价方法 ……8

2.2.1 BLEU算法 ……10
2.2.2 NIST算法 ……11
2.2.3 GTM算法 ……12

2.3 小结 ……13

第三章 学习者译文质量自动评价理论模型构建 ……14

3.1 用基于n-gram算法评价学生译文 ……15

3.1.1 语料说明 ……15
3.1.2 自动评测及结果 ……16
3.1.3 算法评测的影响因素 ……19

3.2 用改进的n-gram算法评价学生译文 ……21

3.2.1 基于伪测试句的自动评测算法 ……21
3.2.2 扩展n-gram评测实验结果 ……23
3.2.3 参考译文数目对评测性能的影响 ……23
3.2.4 对机器翻译评测与学生译文评测的讨论 ……24

3.3 基于线性回归模型的学生译文评价 …… 26
3.3.1 线性回归的数学描述 …… 26
3.3.2 选拔性评分和诊断性评分 …… 28
3.3.3 汉译英评分理论模型 …… 36
3.3.4 英译汉评分理论模型 …… 44
3.4 小结 …… 51

第二部分 技术实现篇 …… 53

第四章 相关语言处理技术 …… 53
4.1 文本特征及提取方法 …… 53
4.1.1 形式特征的提取 …… 53
4.1.2 n-gram共现参数的提取 …… 55
4.1.3 语义点参数提取 …… 57
4.1.4 双语对齐参数的提取 …… 59
4.1.5 潜在语义分析LSA …… 63
4.2 逐步线性回归模型的实现和参数优化 …… 65
4.3 线性相关度的计算 …… 68
4.4 字符编码和汉语语言信息处理 …… 69

第五章 面向大规模考试的英汉翻译自动评分系统 …… 73
5.1 系统实现的原则和结构 …… 73
5.2 系统实现框架 …… 75
5.3 雷同译文检测 …… 77

第六章 翻译自动评分系统的应用 …… 82
6.1 翻译自动评分数据来源 …… 82
6.1.1 语料收集 …… 82
6.1.2 人工评分的实施和评分信度 …… 83
6.1.3 参考译文集的形成 …… 85
6.2 自动评分系统性能 …… 85
6.2.1 系统性能评估方法 …… 85
6.2.2 汉译英自动评分性能 …… 86

6.2.3 英译汉自动评分性能 ……91
6.2.4 雷同译文检查性能 ……92

第七章 翻译自动评价的总结和展望 ……94
7.1 研究结论总结 ……94
7.2 翻译自动评价应用展望 ……96

参考文献 …… **97**

英文参考文献 ……97
中文参考文献 ……105

附 录 …… **108**

附录1 机器翻译自动评测程序的格式要求（XML）和转换程序 …108
附录2 英文停用词表……111
附录3 汉语停用词表……114
附录4 面向考试的自动评分系统的用户文档……115
附录5 诊断性翻译评分系统的界面……120

第一部分 理论研究篇

第一章 绪 论

1.1 语言质量自动评价及研究的意义

语言是人们信息交流最重要的媒体和手段。全球化和国际化的趋势要求人们能够超越语言障碍，实现不同文化和信息的互通。当今，有两种方法来实现跨语言信息交流：一种是从人自身的角度，学习、掌握和使用外语，尤其是作为国际语的英语；另一种则是借助外力，如机器自动翻译等。

在我国，英语是学生必修的课目，甚至从幼儿园就开始教授。众多的外语学习者还要参加各种外语考试以证明他们的外语能力所能达到的水平：

- 根据美国教育考试服务中心（ETS）的数据，截止到2010年底，全球已有2500多万人参加过托福考试，2010年参加GRE考试的人数达63.3万；
- 我国2011年上半年的英语四六级考试中，单次报名人数达909万。

根据外语测试的有关研究结论，目前的考试题型设计一般都包括客观题和主观题两大类。客观题答案唯一而且确定，填涂答题卡的答题方式加上光电扫描技术，使得客观题阅卷基本实现自动化，凡是和标准答案填涂位置一致的就是正确的，不一致的均判错。但是对于主观题的答案往往是以语言叙述形式，短可为简单的句子，长可至段落或篇章，没有确定的标准答案，所以仍然依赖人工评阅。有研究表明，和客观题相比，主观题更能够反映出应试者的语言产出能力，更有人主张在考试中增大主观题的比例。但是这个主张实施起来却不容易。每次大规模的外语考试后，都要组织大量评阅人员对这些没有确定答案的题目进行人工评判，而这方面投入的人力、物力、财力已经相当大了。

语言学习并非易事，有人历时十多年的学习外语，仍然面临交流的困难，人们就幻想着是否能有一种工具可以自动完成不同语言的翻译而免去学习之苦。机器翻译（machine translation）的研究就在计算机诞生的同时也开始了。经过20世纪70和80年代的一段沉寂，现在机器翻译研究在广阔的市场驱动下如火如荼，国内外都有不少著名的机器翻译系统在运作，如：

- 美国乔治敦大学的商用机器翻译系统SYSTRAN，每小时能翻译几十万词，能够实现多语种之间的互译；

- ◆ 我国中科院直属的华建机器翻译也已有十多年的历史，能支持8个语种，14个语对的翻译；
- ◆ 更有不计其数的人每天在使用Google和yahoo等公司的在线机器翻译系统帮助解决翻译中的困难（网址：translate.google.com；babelfish.yahoo.com）。

但是目前机器翻译系统的性能不尽如人意，很多译文晦涩难懂。很多人一边使用机器翻译，一边心存疑虑：机器翻译的结果可信度到底有多高？

上述两种情形面临的是同样的问题，即对翻译质量的评价问题。可以看出，无论在语言教学，还是在技术研究领域，语言质量评价都是十分重要和值得研究的课题。目前，大多的语言质量评价工作还必须依赖人力完成。随着各类外语考试应试者的增加，人工评价翻译质量所需要的时间、人力、物力、财力消耗已经让人们不堪重负，迫切呼唤一种新的、可信度高、快捷的自动语言质量评价方式。而现在正是契机：

- ◆ 托福于2004年开始采用上机考试；
- ◆ GRE宣布自2011年8月开始以机考代替笔考；
- ◆ 据《北京考试报》2008年12月13日的消息："大学英语四六级考试将首次试行机考，并在全国50所高校试点"。

答卷形式的转变本身没有多大的意义，却为自动评阅和结果的自动处理提供了必要的前提条件，因为电子化的文档才方便计算机进行数据处理。真实的语言数据为研究主观题自动评分提供了研究内容和评测数据。

自动评测的内容可以是学习者产出的语言，也可以是机器生成的语言。自动评测的基本目标是要快速、廉价、准确地区分不同质量的文字，并给出一个具体的数值得分。当然，还可以更进一步指出文字中的缺陷和不足用于改进和提高。

机器翻译的译文有自动评测算法（Papineni 等，2002；Doddington，2002），应试者的作文可以实现自动评分（Burstein & Chodorow，1999；梁茂成，2005）。研究形势表明，语言质量自动评价是大势所趋，尤其在外语教学和测试中，更是对传统教学和测试观念和方法的变革。语言质量自动评价研究对外语教学和测试的意义重大：

首先，在大规模考试中，可以将评分人员从疲惫不堪的重复性阅卷工作中解放出来，节省大量的时间和消耗。

第二，机器自动评价语言质量将实现主观题评价方式的根本转变，由人的主观评价变为机器的客观评价，克服和避免人工评价一致性差，可信度低，并且还受时间、评阅人不同、评阅人心情和疲劳程度等因素的影响而变化的弊端。

第三，通过运行软件，评价工作可随时随地进行，因此，大大方便语言教学和学生的学习。教师可以及时掌控学生的学习状况，学习者可以及时得到学习效果的反馈。

第四，自动语言质量评价涉及到语言学、语言教学和测试、计算机技术、自然语言处理、人工智能等多个学科，理论研究意义重大。该研究课题将极大地促进各个学科的发展和学科之间的融合，改进教学方式，促进技术进步。

语言质量自动评价研究是一项包括众多因素的综合研究，难度很大，目前尚在起步探索阶段，整体研究水平不高，本书也是尝试性研究成果的汇报，很多结论亦是在一定数据基础上、在现有的研究方法下得出的。

1.2 相关研究回顾

研究语言质量评价，主要涉及的内容有翻译质量和写作质量等以自然语言形式表述的语言质量问题；而需要评价语言质量的主要是机器产出的语言和学习者产出的语言。将这两个方面排列组合一下，也就得到了语言质量自动评价相关研究的视角。

一、自动评价学习者写作水平的研究

影响写作质量的因素众多，作文也分为命题作文和情景作文等多种情况。如何判断一篇习作的优劣，标准不一。对评价人来讲，可以依据的评判作文质量的标准通常是抽象和粗略的，常见的有内容切题、主题明确、表达清楚、文字连贯、句式有变化、用词正确等等。对其中每一项标准的认识和理解，不同的人并不完全相同。这些写给人看的评价标准都无法让计算机理解并执行，这正是自动评价研究的困难之处。

国外对于作文自动评分的研究起步于20世纪60年代。到90年代末，已有商用的系统，如BETSY，IEA，IntelliMetric，E-Rater等投入实际应用。BETSY（Bayesian Essay Test Scoring sYstem）（Runder & Liang，2002）是一个基于Bayesian文本分类的作文评分系统，根据作文质量分类，不同的类对应不同的得分等级。美国教育考试服务中心用于TOFEL、GRE等考试中的作文评分软件称为E-Rater，目前和人工评分一起来评价应试者的英语语言能力[1]。E-Rater运用自然语言处理技术（Natural Language Processing，NLP）从测试作文中提取与写作能力有关的语言特征，如词汇的使用、文章结构、组织、展开、写作风格等（Burstein & Chodorow，1999），根据测试作文这些特征的多少来评价它的质量高低。语言特征以形式的为主，很少在语义层上展开。而IEA（Intelligent Essay Assessor）因为采用了潜在语义分析法（Latent Semantic Analysis）而号称可以从语义上分析文章质量；IntelliMetric（Elliot，2001）也将人工智能的成果用于评分系统设计中。

国内梁茂成（2005）、葛诗利和陈潇潇（2007）也对我国学生作文自动评分

1 http:// www.ets.org/understanding_testing/scoring

进行了富有成效的探索。

正如美国教育考试服务中心所述，对于考生作文质量的自动评分目前还只是作为“第二评分员”，如果人工评分和机器自动评分分歧较大时，则要对作文重新评分，以此来进一步提高人工评分的信度和准确率。

二、对自动评价机器翻译质量的研究

机器翻译质量评价问题是和机器翻译研究一同起步的。开始以人工评价为主，但随着机器翻译研究的深入，早期人工对翻译结果评价的方法因为费时费力而且没有重用性，已经不再能满足需求。系统开发人员对系统的每一次改进都希望得到廉价、准确、迅速的反馈，以把握系统性能的变化。人工评价的反馈时间过长，延长了系统开发周期。为促进机器翻译研究而举办的国际机器翻译竞赛，则需要对很多参赛的翻译系统给予评价，对其性能进行排序，人工评测的任务十分繁重。另一方面，人力评测译文质量，也会融入主观因素，影响评测的信度。用户购买翻译软件，在不同的系统之间选择时，也需要快速准确地了解不同系统的性能。因此，机器翻译自动评测的研究受到重视，产生了多种评测方法。同时，评测方法的改进也极大地促进了机器翻译的研究，为开发者和用户提供了客观的比较和交流的平台。

为评测机器译文，通常要准备一篇或数篇人工的译文作为参考。一般认为，一个翻译系统的译文和人工译文越相似，译文质量越高。基于这种思想，Papineni等（2002）提出了基于n-gram共现的BLEU（bilingual evaluation understudy）算法，之后Doddington（2002）对BLEU在n-gram的权重、平均算法等方面进行了改进，形成了目前美国国家标准和技术所（NIST）用于国际机器翻译自动评测的NIST算法。这两种算法都是通过统计测试译文和参考译文共有的n-gram数目（n通常小于4）来判断译文的质量。实验表明，在评测机器译文质量的性能上，算法和人工评测有很高的相关度。为更多地查找到测试译文和参考译文相同或相似的n-gram，很多研究者从各个方面对n-gram算法进行了扩展。Kauchak 和 Barzilay（2006）利用重述（paraphrasing）方法合成了部分参考译文，如果测试译文和合成译文匹配，就认为和参考译文也匹配。重述词是根据外部资源WordNet及上下文环境得到的。为克服严格n-gram匹配带来的问题，Lin & Och（2004）提出了skip-gram的方法，将n-gram匹配和松弛最长公共子串匹配的优势结合起来评测译文质量。还有部分研究者借鉴文档相似、句子相似及机器学习的思想改进自动评测方法。

经过十年来的研究，机器翻译自动评测方法已经成功用于系统译文的评测。Zhou 等（2008）提出了一种诊断性机器翻译评测方法，从词、短语和句子级分别给出翻译系统在若干语言现象上的翻译能力的评测。这个评测系统

（Woodpecker）已被第四届全国机器翻译研讨会采纳为一个评测指标[2]。

近年来，随着机器译文质量的不断改进，有研究者发现以n-gram匹配为主的语言形式上的相似已经不能反映系统译文质量在细节层面的差别了（黄瑾 等，2007）。

三、对自动评价机器文摘质量的研究

面对海量文件，人们希望能够快速、简洁、准确地获得核心内容，自动文摘（Automatic Summary）技术也应运而生。自动文摘可以看做是机器自动生成的文章，但这种文摘的流利度、要点是否准确等指标同样需要评测。机器自动文摘研究的同时，其评价方法也十分关键。

Lin and Hovy(2003) 借鉴机器译文评价方法，基于n-gram共现的统计数据实现了对文摘的质量评价。在2005年ACL会议中，举办了一个关于评价机器翻译和自动文摘质量的研讨（workshop），Amig'o（2005）等人还提出了一种基于相似模型的QARLA框架来评价文摘质量。

总的看来，对于语言质量自动评价的研究，开展得较多的是前两个方向，即对机器译文质量的评价和对学习者作文质量的评价，已经在国际机器翻译比赛和有关外语考试中应用。但是对于学习者翻译质量的评价研究较少。这正是本书要讨论的内容。

1.3 本书的内容及安排

本书对于学习者译文自动评分研究分为理论构建和技术实现两部分，理论构建部分包括第二和第三章，将论述对学习者译文质量评价数学模型的对比、选择和构建；技术实现部分包括第四章至第六章，重点是在理论论证的基础上如何借助NLP技术实现自动评分系统，并考虑实际考试中的需求，加入一些实用功能，如雷同译文检查。

理论部分的第二章探讨人工对译文质量的评价方法，或者说从翻译学的角度看，如何界定一篇译文质量的高低。基于这些评价观点，如何将这些方法形式化，量化为具体的评价指标，才能让计算机实现自动评价。本章将对一些著名的机器译文自动评价算法进行分析和评价。第三章是构建学习者译文评分的理论框架，尝试和对比了n-gram模型，扩展n-gram模型和线性回归模型。

技术实现部分的第四章介绍常用的自然语言处理技术，第五章是英汉双向自动互译评分系统的构建，在实际数据上对评分系统性能的测试结果和分析安排在第六章。第七章是有关研究结论和对未来自动评分系统的展望。

---

2 http:// www.chineseldc.org/doc/Guidelines for CWMT2008 Machine Translation Evaluation-Chinese.pdf

第二章　翻译质量评价

2.1 翻译质量的人工评价标准

在翻译界，为大多人熟知的评价译文质量的三大原则是："信、达、雅"。"信"是指要忠实于原文，这是最基本的原则。翻译不同于无蓝本的新的创作，与原著的思想、内涵和风格大相径庭的不能称为翻译作品；"达"是指译文要通顺，符合译语的语言表达习惯，读起来不晦涩难懂，"达"的要求比"信"更难以做到；"雅"则是要求译文用词确切，起到对原文思想精髓的传神表述，在这三原则中，"雅"的要求最高。尽管现在有不少人对严复先生提出的这三大翻译原则有颇多质疑和修改，但"信""达""雅"仍不失为简洁而又精辟的评价标准。

这里我们关注的不是对专业人士译作的评价，而是对两类不完美译文的质量评价问题。一类在外语教学中，通过外语学习后，学习者译文质量的评价；一类是机器翻译中，人们如何评判系统的性能，译文质量无疑是机器翻译系统最关键的性能指标。这些译文还未达到"雅"的标准，所以，我们基本不考虑如何认定什么是雅文。

如何具体运用评价标准评测翻译的质量？我们先看看工业界的做法。在20世纪60年代，机器翻译刚刚兴起，吸引了很多的研究者，人们对机器翻译给予了极大的热情和期待。为了评价机器翻译的性能，美国国家科学院语言自动处理咨询委员会(Automatic Language Processing Advisory Committee ,即ALPAC)就提出了一套人工评价机器译文质量的两个指标：忠实度（fidelity）和可理解度（intelligibility）。忠实度用于衡量译文是否正确反映原文所表达的含义，可理解度用来评价译文的表达是否通顺、可理解和符合目标语的表达习惯等。可理解度指标被分为9级，涉及的内容主要是能否准确地表达思想（idea），是否存在未译的词（untranslated words），是否有文法或语法问题（stylistic or grammatical infelicities），用词是否得当（unusual word usage），句法排列次序上是否有问题（syntactic arrangement）等。忠实度采用信息度（informative）方式间接评测，这种方法是，首先让同时熟悉原语和译语的评价人员阅读译文，获取一定的信息，然后再去阅读原文，根据他的理解来判断原文提供的信息和译文提供的信息的差别。Informative的级别分10级。根据这些指标对当时机器翻译系统的评价报告认为机器译文完全不可用，机器翻译研究一度被打入冷宫。

我国863机器翻译评测中，也采用了类似的人工评价指标。表2-1和2-2列出了详细的忠实度和流利度评判等级和得分标准。

表2-1 人工评测译文的忠实度得分标准

等级分	得分标准
0	完全没有译出来
1	译文只有个别词符合原文
2	译文有少数内容符合原文
3	译文基本表达了原文的意思
4	译文表达了原文的绝大部分信息
5	译文准确完整地表达了原文信息

表2-2 人工评测译文的流利度得分标准

等级分	得分标准
0	完全不可理解
1	译文晦涩难懂
2	译文很不流畅
3	译文基本流畅
4	译文流畅，但是在地道性方面有所不足
5	译文是流畅而且地道的句子

在具体实施时，人工对系统的打分可以带一位小数，这也就意味着，评判等级不只是6个级别。上述标准执行中，不同的人对于完整、准确、流畅、地道这些词语的理解不同，评分过程必然会融入主观成分，导致不同人给同一篇译文的评分不同。最终一般是取多个评判结果的平均值作为译文的得分。

再看外语教学中，对学习者翻译质量的评价准则。在英语专业八级考试中，英译汉和汉译英各有10分的题，评分都分为5档，标准如表2-3：

表2-3 英语专业八级考试翻译题评分标准

得　分	标　准
9-10	忠实原文，只有1-2个词汇、拼写、标点、句法上的小错。译文优美（词汇运用恰当，句式有变化）。
7-8	近乎忠实原文，有相对较少的有意义的词汇、拼写、标点和句法错。译文可读（整体清晰、流畅、连贯）。
5-6	能够反映出大多数原文的意思，偶尔出现词汇、拼写、标点和句法错。译文大部分可读。

得　分	标　准
3-4	译文只能反映原文一半左右的意义，词汇、拼写、标点和句法错频率较高。译文部分不可读。
1-2	译文对原文意义的反映不足一半，几乎所有的句子都有词汇、拼写、标点和句法错。译文大部分不可读。

一般外语考试的翻译阅卷标准中，除了上述较笼统的评分标准外，还以具体的题目划分出具体的得分点，规定得分的细则，同时也会给出一个参考的译文，或者提供在某些关键点上的多种可接受的译法以供参考。

评分点的划分主要是要考查的语言知识点，一般根据句子的语义内容进行划分，将那些与句义密切相关的语言片段作为评分点。如果应试者能够将评分点译出，句子的基本语义便可反映出来。评分时，评分员依据参考答案和评分点是否有语法错误，为每一个评分点给定一个分值。所有句子得分的累加就是整篇译文的得分。当然，也有采取扣分法的。

总之，公认的好的译文质量评分标准应该有较强的可操作性，得分能够反映译者的语言水平，而且不同评分员对同一篇译文的评分差异性小，同一个评分员对同一篇译文在不同时间的评分结果也不应有明显的变化。换句话说，也就是好的评分标准应该是客观的、可操作性强，尽量不受评分员主观因素的影响。

对人工评价译文质量的思路和方法的研究有助于我们研究自动译文质量评分方法。自动评分的目标也就是实现对人工评分的充分模拟。

2.2 翻译质量的自动评价方法

实现机器自动评价译文质量，人们首先会直觉地考虑能否将人工评分标准转换为计算机能实现的方法，也就是使评分标准形式化和量化。

第一种思路是，效仿人工评分中的扣分法，基于译文中的错误数目和类型判定来评分。采用这种方法的研究者持有的观点是，既然什么是好的译文涉及太多的因素，难以穷尽地列举高质量译文“好”在哪里，就考虑问题的反面，因为什么是译文中的错误相对好定义，所以就统计译文的错误数目和类型，错误数目越少，错误情节越轻，就可以认为译文的质量越高。毕竟高质量的译文中不应该有太多的错误。这是一种减分法。如果不能发现译文中的错，就认为是满分的译文。他们认为，语言中的错误类型和错误数目是比较容易量化的。

2006年，Snover 等人提出了基于翻译错误率（translation error rate，TER）自动评测译文质量的方法。首先准备一篇或者数篇参考答案，对测试译文进行编辑修改使之和某一篇参考译文完全一致，并统计所需的最少改动次数，最后将修

改次数除以参考译文的平均长度，得到TER指标。用公式表达即：

$$\text{TER} = \frac{\#\text{ of edits}}{\text{average }\#\text{ of reference words}}$$

编辑操作具体包括插入、删除、替换和语序的调整。例如：

参考译文：*SAUDI ARABIA* denied *THIS WEEK* information published in the *AMERICAN* NEW YORK TIMES.

测试译文：*THIS WEEK THE SAUDIS* denied information published in the NEW YORK TIMES.

测试译文要和参考译文一致，需要修改4次（2次替换，1次插入，1次语序调整）。其TER值为4/13=0.31。

明显地，TER值越小，译文质量被认为越高。既考虑了测试译文和参考译文用词的一致，又考虑了语序一致。

第二种思路是让计算机依据测试译文和参考译文的相似程度进行评价。以参考译文为标准，测试译文和参考译文越相似译文质量就越高。基于这种思想，评价译文质量的问题转换为求测试译文和参考译文相似度的问题。相似度计算成为这种质量评价方法的核心。语义的相似在一定程度上表现为语言形式上的相似，因此就有了基于n-gram共现的评分方法。这也正是目前国际机器翻译比赛评测所采纳的主要方法之一。

2001年在IBM工作的Papineni提出了基于n-gram共现的BLEU（bilingual evaluation understudy）算法，之后Doddington （2002）对BLEU在n-gram的权重、平均算法等方面进行了改进，这就是著名的国际机器翻译自动评测的NIST算法。这两种算法都是通过统计测试译文和参考译文共现的n-gram数目（n通常小于4）来判断译文的质量。虽然只是形式上的比对，实验却表明，在评测机器翻译的性能上，该算法和人工评测有很高的相关度。

n-gram又称为n元语法，是一种以计算的观点处理自然语言而建立的数学模型。自然语句视为一个随机序列，根据马尔科夫（Markov）假设，一个句子中前面连续n-1个词是影响后续词的上文信息（context），根据这n-1个词可以预测第n个词的出现概率。n=1时称为一元语法（unigram），n=2时，称为二元语法（bigram），n=3则称为三元语法（trigram）。

基于n-gram共现算法求测试译文和参考译文的相似度的主要思想是，如果测试译文的词串同时也出现在参考译文中，则进行统计计数，共现的次数越多，测试译文的质量越高。可以看出，当n取值较小时（n=1或2），也就是单个词匹配或连续2个词匹配的情况。因为匹配时没有次序的要求，所以n-gram共现更多反映的是测试译文用词和参考译文词的一致程度。参考译文是忠实于原文的，因

此unigram和bigram共现情况主要反映了测试译文的忠实度指标。n较大时，意味着测试译文和参考译文在较大的语块上相同，而参考译文一般都是专家给出的译文，所以较大n-gram的匹配在一定程度上又反映出译文的流利程度。因此，基于n-gram共现的算法是一个综合了忠实度和流利度的指标。由于语言灵活多变，当n较大时，n-gram相同的几率大大减小，因此，实际运用这种方法时，通常n最大取4或5，更大的n时，匹配率都为0。

下面简要回顾这几种基于n-gram的评测译文质量的算法。这些算法的共性是要有参考译文，有的只要求有一篇参考译文，有的则允许多篇。通过对比测试译文和参考译文的形式上的相似之处来判断译文质量得分。

2.2.1 BLEU算法

BLEU算法在查找测试译文和参考译文相同的n-gram前，要将译文进行分段。段（segment）是评测的单位，一段是一块连续的文本，通常为一个句子，有时也可以是多句。算法中参考译文允许为多篇，构成参考译文集。只要是和参考译文集中共现的n-gram都统计，没有区分是否出现在不同的参考译文中。参考译文和测试译文都要分割为段，而且段要相对应。然后提取n-gram。所有n-gram均不跨段提取。下面通过一个汉译英译文BLEU评分的例子详细说明这种算法。

测 试 句：The best advantage of writing letter is let me felt the humor of my parents.

参考句1：The best thing of writing letters home was that I became conscious of parents' hidden humor suddenly.

参考句2：The greatest benefit of writing home letters is that it made me realize my parents' hidden humor.

测试句和2个参考句共现的unigram有9个：the, best, of, writing, is, me, humor, my, parents，共10次共现[3]。测试句共15个词（不包括标点），因此unigram的共现率为0.667。但是如果有这样的句子：The best of writing is me humor my parents. Unigram共现次数相同，但由于测试句的长度也为10个词，导致unigram共现率将为1.00。可见，句子的长度对得分有影响。所以，算法最后根据测试句的长度加入了一个惩罚因子，用于惩罚过短的测试句。句子的标准长度来自参考译文的平均长度。例如，上面测试句的长度为10，参考译文平均长度为18，加入惩罚因子后，10个词的测试句的unigram BLEU得分为0.449，相比15个词的测试句，得分要小，因此较合理。

同样可以得到2-gram，3-gram及4-gram的得分值，最后将这些值加权平均，乘以惩罚因子就是最终的BLEU得分。BLEU评测译文得分的计算公式为

3 the因为在一个参考译句只出现一次，因此最多记1