·机器翻译·

机器翻译质量综合评价方法研究

孙逸群[1] 周敏康[1,2]
（1 西班牙巴塞罗那自治大学翻译系 西班牙 08193；2 厦门大学外文学院 厦门市 361005）

摘 要 基于对“词汇”、“语法”和“语篇”三个译文质量评价因素的分析，本文构建了包含12个指标、每个指标分为5个质量等级的评价指标体系，通过问卷设计与问卷调查方法获取机器翻译质量的评价数据。采用层次分析法确定评价指标的权重向量，利用模糊数学理论建立了机器翻译质量的综合评价向量模型。实例分析表明，该模型可以科学定量地评价翻译软件的译文质量及其在各项评价指标上的差异，对读者筛选优质翻译软件、开发商提高软件质量具有重要意义。

关键词 机器翻译 层次分析法 权重向量 模糊数学 综合评价模型

Abstract Based on an analysis of three factors in the assessment of translation quality, i. e. vocabulary, grammar and discourse, this paper developes an evaluation index system which includes twelve indicators and each indicator is divided into five levels. Meanwhile, a questionnaire is also designed for use. Foreign language teachers and Master's degree candidates are selected as respondents. Then, the data for machine translation evaluation is obtained from the questionnaire. The analytic hierarchy process (AHP) is applied to the data in order to determine the weighted vector of machine translation evaluation factors. On this basis, fuzzy mathematics theory is used to establish the comprehensive evaluation vector for machine translation quality. As a result, the quantity comprehensive evaluation for the translation text quality is completed. The results of our study show that the comprehensive evaluation model of translation quality based on the AHP and fuzzy mathematics method can be employed to scientifically and quantitatively analyze the translation text quality of the translation software as well as the differences of various evaluation indicators. Thus, this is significant for readers to choose high quality translation software and for software developers to find design flaws so as to improve the software design and the quality of software.

Key Words machine translation analytic hierarchy process weighted vector fuzzy mathematics comprehensive evaluation model

DOI:10.16024/j.cnki.issn1002-0489.2017.02.007

机器翻译是不同语言之间文化交流、信息检索的重要手段，是目前计算机和翻译领域的研究热点之一。翻译质量的准确评价是机器翻译系统研发的主要依据，近年来，机器翻译评价方法的研究取得了丰富成果。系统译文和参考译文的相似度评价是主要方法之一[1]，该方法需借助人工译文，采用计算机分析对比词与词、句与句、段落与段落的相似度，是局限于微观层面的评估方法，缺乏对文章宏观的把握，同时计算机也无法感知具体的语境，更无法理解文章深层次的意蕴。因此，综合评价方法受到人们的关注，成为研究热点之一。

层次分析法、模糊综合评判法是常用的综合评价方法，具有结果清晰，系统性强的特点，能较好地解决模糊的、难以量化的问题。范守义首先采用模糊数学中隶属度的概念定量评价了译文质量，考虑了九个因素，但没有考虑其权重，隶属度的数据也是专家根据经验给定的[2]。穆雷认为模糊统计是确定隶属函数的一种主要方法，这样确定的隶属度更加客观、科学[3]，并通过实例证明了模糊数学评价译文质量的可行性[4]，但其评价因素的权重亦是由专家根据经验主观人为设定的，缺乏客观性。评价因素权重的准确确定是决定评价质量的关

*本文是2016-2019年度西班牙国家科学研究项目“东亚研究”（项目编号：FFI2015-70513-P）的阶段性研究成果之一。在研究过程中获得该项目的资助。
作者电邮：13969161305@163.com，minkang.zhou@uab.cat 收稿日期：2016-07-18 修改日期：2016-12-30/20

键因素，但目前该值多人为设定。

基于上述分析，本文利用层次分析法设计递阶层次结构的机器翻译质量评价指标体系，采用问卷调查方式获取读者对译文质量评价的数据，对数据分层次进行模糊综合评判，最后得出量化评价结果。

1　机器翻译质量的评价指标体系及其权重

1.1　评价指标体系

评价译文质量考虑“词汇”、“语法”和“语篇”三个主要因素[5]，词汇从词义搭配、修辞、专门术语、方言使用四个方面考察；语法主要考察译文的语法是否正确；语篇包括衔接性、连贯性、意图性、可接受性、信息性、语境性和互文性。因此，本文将中译文质量 Y 的评价问题，分解为词义搭配 B1、修辞 B2、专门术语 B3、方言使用 B4、语法 B5、衔接性 B6、连贯性 B7、意图性 B8、可接受性 B9、信息性 B10、语境性 B11 和互文性 B12 等 12 个评价指标，12 个评价指标根据其相互关系和隶属关系形成词汇 A1、语法 A2、语篇 A3 三个评价层面。进而构建起以中译文质量 Y 作为评价目标的评价指标体系。

1.2　评价指标体系权重值的计算

评价指标体系的权重包括相对于评价目标 Y 的评价层面词汇 A1、语法 A2、语篇 A3 的权重，以及相对于三个评价层面的评价指标的权重。本文以相对于词汇 A1 的评价因素 B1、B2、B3、B4 权重值的计算方法为例，说明采用层次分析法确定权重值的方法。

B1、B2、B3、B4 两两比较，根据比较标度法赋分[6]，得到式（1）比较判断矩阵 A1，

$$A1=\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix}= \begin{pmatrix} 1 & 1 & 2 & 0.5 \\ 1 & 1 & 2 & 0.5 \\ 0.5 & 0.5 & 1 & 2 \\ 2 & 2 & 1 & 1 \end{pmatrix} \quad (1)$$

比较标度法赋分规则为[6]：两个因素相比，前者与后者相比同等重要、稍微重要、明显重要、强烈重要、极端重要分别赋分 1、3、5、7、9；2、4、6、8 则表示上述相邻判断的中间值；如果因素 i 与 j 的重要性之比为 a_{ij}，那么因素 j 与 i 的重要性之比为 $a_{ji}=1/a_{ij}$。如（1）式 A1 矩阵中元素 $a_{11}=1$，表示 B1（词义搭配）自己比较，同等重要；$a_{12}=1$ 表示 B1（词义搭配）与 B2（修辞）相比，同等重要；$a_{13}=2$ 表示 B1（词义搭配）与 B3（专门术语）相比，介于同等重要（赋值 1）和稍微重要（赋值 3）之间；$a_{13}=2$，则 $a_{31}=1/a_{13}=1/2=0.5$。

将式（1）A1 按列归一化处理，得到 A1′，

$$A1'=\begin{pmatrix} 0.2222 & 0.2222 & 0.3333 & 0.1667 \\ 0.2222 & 0.2222 & 0.3333 & 0.1667 \\ 0.1111 & 0.1111 & 0.1667 & 0.3333 \\ 0.4444 & 0.4444 & 0.1667 & 0.3333 \end{pmatrix} \quad (2)$$

将式（2）A1′ 的每行相加，然后归一化处理得到 A1 的特征向量 W1，即权重向量 W1，W1 =（0.2361，0.2361，0.1806，0.3472）^{T}

从而获得 B1、B2、B3、B4 的权重值分别为 0.2361、0.2361、0.1806 和 0.3472。

利用上述方法，计算得到 B5 的权重值为 1；B6、B7、B8、B9、B10、B11、B12 的权重值分别为 0.1256、0.1256、0.1078、0.2064、0.0994、0.1702、0.1651。A1、A2、A3 的权重值分别为 0.3119、0.1976、0.4905。经检验，计算权重值的判断矩阵一致性可以接受[6]，即判断矩阵反应了评价因素的重要程度。

2　问卷设计与测评软件和测评文本

根据中译文质量的评价指标体系设计问卷，以获得译文 12 个评价指标的质量等级。按照李克特量表的概念，每个评价指标分为五个质量等级：等级 1 为优秀、等级 2 为良好、等级 3 为中等、等级 4 为较差、等级 5 为差。词义搭配 B1 等级 1、等级 2、等级 3、等级 4、等级 5 分别要求整个测试文本词义搭配准确的占到 90% 以上、80%—90%、70%—80%、60%—70% 和 60% 以下。其他小类指标的等级分类亦按此标准界定。

从目前应用较广的谷歌翻译、必应翻译、百度翻译、有道翻译、通用翻译、灵格斯等十几种翻译软件中，选取 4 款软件分别标记为 A、B、C、D，对其翻译质量进行评估。测试文本选取了一篇英语新闻，题目为 How do you turn your tech start-up into a global giant?[7]，共

计1 100余字。

3 问卷数据及其模糊综合评判

采用网络调查方式，向外语专业的研究生和教师发出调查问卷120份，对4款翻译软件的译文质量进行问卷调查。选取其中有效问卷100份进行统计分析。以词汇A1为例，介绍问卷数据的数据处理方法。如表1所示，词汇四个评价指标评定为等级1、等级2、等级3、等级4、等级5的总人数分别为1、387、9、3和0。将1、387、9、3、4分别除以这五个数之和，即进行归一化处理得到5个等级的模糊数。同样方法可以计算出语法、语篇的模糊数。

表1 软件A译文质量问卷统计结果及数据处理

评价层面	评价指标	人数统计				
		等级1	等级2	等级3	等级4	等级5
词汇A1	词义搭配	0	96	3	1	0
	修辞	0	94	4	2	0
	专门术语	0	99	1	0	0
	方言使用	1	98	1	0	0
词汇数据处理	评定为相应等级的总人数	1	387	9	3	0
	各等级归一化模糊数	0.0025	0.9675	0.0225	0.0075	0

词汇、语法、语篇的权重值与各等级的归一化模糊数进行模糊合成运算，得到对软件A的综合评价向量，其计算见表2。表2中权重和等级1模糊数的计算为，(0.3119∧ 0.0025)∨(0.1976 (0)∨(0.4905∧0) =0.0025，其中∧、∨分别为取比较数据的最小值和取最大值，如0.3119∧0.0025的结果为0.0025，0.0025∨0∨0的结果为0.0025。同样方法计算出权重与等级2－等级5的结果列于表2。对表2的模糊运算结果归一化处理，获得软件A译文质量的综合评价向量 Y_A = (0.0040，0.7819，0.0957，0.0592，0.0592)。

表2 软件A模糊综合评价的计算

	权重	等级1	等级2	等级3	等级4	等级5
词汇	0.3119	0.0025	0.9675	0.0225	0.0075	0
语法	0.1976	0.0000	0.8900	0.0600	0.0200	0.0300
语篇	0.4905	0.0000	0.8700	0.0557	0.0371	0.0371
模糊合成运算结果		0.0025	0.4905	0.06	0.0371	0.0371
综合评价向量		0.0040	0.7819	0.0957	0.0592	0.0592

软件A、B、C、D译文质量评价目标和评价层面按上述方法进行综合模糊评判，其评价结果见表3、表4，表中L1－L5表示评定为相应等级的人数占总评价人数的百分比。

表3为译文质量评价目标的综合评价结果。可以看出，软件A、B、C、D的L1分别为0.40%，0.36%，0%和0%，认为4款软件为等级1优秀的人数非常少，因此翻译软件总体质量还需要进一步提高。L1+L2为软件质量达到等级2及以上优良的人数占总评价人数的百分比，软件A、B、C、D的L1+L2分别为78.59%、70.70%、63.25%和61.85%，译文质量优良率从好到差的软件排列顺序为A、B、C、D，软件A的L1+L2是软件D的L1+L2的1.27倍，从表3的数据亦可以看出，认为软件D差（等级5）和较差（等级4）的人数达到评价总人数的30.27%，软件D有待改进。L1+L2+L3为软件翻译质量不是较差，还能接受的评价数据，软件A、B、C、D的L1+L2+L3分别为88.16%、82.12%、74.86%和69.73%，软件A、B较好，均达到了80%以上，而软件D仅有69.73%。

设定阈值可以对译文质量评定分级。设定评价因素5个等级的阈值 λ_1 =0.15（优秀）、

$\lambda_2 = 0.45$（良好）、$\lambda_3 = 0.70$（中等）、$\lambda_4 = 0.80$（较差）、$\lambda_5 = 0.90$（差）[8]。由表 3 的数据得，$y_1^A = 0.0040 < \lambda 1$，$y_1^A + y_2^A = 0.0040 + 0.7819 = 0.7859 < \lambda 2$，因此软件 A 的译文评定为良好。同样方法软件 B、C、D 的译文亦评定为良好。

表 3　译文质量评价目标的综合模糊评价结果

软件	L1	L2	L3	L4	L5
A	0.0040	0.7819	0.0957	0.0592	0.0592
B	0.0036	0.7034	0.1147	0.1004	0.0779
C	0.0000	0.6325	0.1161	0.1289	0.1225
D	0.0000	0.6185	0.0788	0.1639	0.1388

表 4 为译文质量三个评价层面的综合模糊评判结果。在词汇层面，软件 A、B、C、D 的 L1 分别为 2.4%、2.29%、0% 和 0%，4 款软件在词汇方面也表现欠佳，评定为优秀的数据极小。软件 A、B、C、D 词汇的 L1 + L2 分别为 85.62%、81.7%、52.04% 和 53.65%，因此，词汇翻译最好的是 A，其次是 B，排在第三位的是 D，最差的是 C。软件 A、B 在词汇方面比较优秀，软件 A 的 L1 + L2 为软件 D 的 L1 + L2 的 1.60 倍，为软件 C 的 1.65 倍，软件 C、D 在词汇方面有待改进。软件 D 虽然在总体评价方面（见表 3）表现最差，但在词汇方面好于软件 C。

在语篇层面，软件 A、B、C、D 的 L1 均为 0%，4 款软件在语篇方面表现较差。软件 A、B、C、D 语篇的 L1 + L2 分别为 43.33%、43.33%、58.56% 和 54.84%。因此，语篇翻译最好的是 C，其次是 D，A 和 B 并列排在第三位。因此，软件 A 虽然在总体评价方面（见表 3）最好，但其在语篇方面排在最后，还有改进提升的空间，而总体评价排名第二的软件 B，在语篇方面表现亦最差，其语篇也急待改进，以提升软件翻译质量。软件 C 虽在总体评价方面（见表 3）位于第 3，但其语篇表现最好。

在语法层面，软件 A、B、C、D 语法的 L1 均为 0%，4 款软件在语法方面没有达到优秀的。从表 4 软件 A、B、C、D 语法的 L1 + L2、L1 + L2 + L3 可以看出，语法翻译最好的是 A，B、C 的表现也很好，D 表现最差。B 总体评价上优于 C（见表 3），但在语法方面，两者比较接近。软件 D 总体评价最差（见表 3），语法上也是表现最差的。

表 4　译文质量评价层面的综合模糊评判结果

评价层面	软件	L1	L2	L3	L4	L5
词汇	A	0.0240	0.8322	0.0959	0.0479	0.0000
	B	0.0229	0.7941	0.0916	0.0457	0.0457
	C	0.0000	0.5204	0.1349	0.1649	0.1798
	D	0.0000	0.5365	0.1236	0.1700	0.1700
语篇	A	0.0000	0.4333	0.1889	0.2099	0.1679
	B	0.0000	0.4332	0.1470	0.2099	0.2099
	C	0.0000	0.5856	0.1703	0.1873	0.0568
	D	0.0000	0.5484	0.2391	0.1328	0.0797
语法	A	0.0000	0.8900	0.0600	0.0200	0.0300
	B	0.0000	0.8000	0.0800	0.0700	0.0500
	C	0.0000	0.8000	0.0900	0.0700	0.0400
	D	0.0000	0.4905	0.0625	0.1300	0.1100

4 结 论

（1）采用层次分析法，通过构建两两比较判断矩阵，定量确定评价机器翻译相对于评价目标和相对于评价层面的权重，克服了人为设定权重数值的弊端。

（2）以问卷调查方法获得的评价译文质量的数据为基础，采用模糊数学方法建立机器翻译质量的模糊评价矩阵，结合基于层析分析法的权重模糊向量，可以获得评价译文质量的综合评价向量，实现了对译文质量的定量评价。

（3）采用层次分析法和模糊综合评判法，可以科学定量地分析翻译软件的译文质量和翻译软件在各类评价指标上的差异，以帮助读者筛选翻译质量好的软件，并为软件开发商指出其软件在某些特定方面的缺陷，指导机器翻译软件开发商改进其软件设计和程序，提高软件的运用质量与效率。

（4）在国内外高校翻译教学中，机器翻译质量综合评价方法可望成为学生进行翻译实践的有效手段之一，为高校翻译教学的信息化和现代化提供新的途径与方法。

5 参考文献

1 王博. 机器翻译系统的自动评价及诊断方法研究. 工学博士论文，哈尔滨：哈尔滨工业大学，2010，8-10
2 范守义. 模糊数学与译文评价. 中国翻译，1987（4）：2-9
3 穆雷. 用模糊数学评价译文的进一步探讨. 外国语，1991（2）：66-69
4 穆雷. 模糊数学评价译文的再探讨. 中国科技翻译，1992（18）：39-43
5 张化丽. 对译文质量评估的参数分析. 译林，2011（8）：70-76
6 Thomas L. Saaty, Luis G. Vargas: Models, Methods, Concepts & Applications of the Analytic Hierarchy Process, Springer Science + Business Media New York, 2012, 23-40, 100-102, 149-158, 161-165, 203-247.
7 Daniel Thomas. How do you turn your tech start-up into a global giant? BBC NEWS, the section Business. 6 November 2015, http://www.bbc.com/news/business-34731456
8 马丽，李平，王秀英，赵丽华. 机器翻译系统的模糊评价方法. 微计算机信息，2008，24（1-1）：299-300

（上接第10页）

④interstellar extinction 定义来源 https://en.wikipedia.org/wiki/Extinction_（astronomy）

⑤本文表格内英汉对照词语均摘自1958年科学出版社《天文学名词》一书。

3 参考文献

1 Tyson. N. *The Pluto Files*. New York: W. W. Norton & Company, 2014, 55-91
2 Kasting. J. *How to Find a Habitable Planet*. Princeton: Princeton UP, 2010, 8-35
3 刘增羽. 中国古代的天文翻译. 中国翻译，1996（4）：54-57
4 欧阳修. 新唐书. 北京：中华书局，1975，36
5 房玄龄. 晋书. 北京：中华书局，2008，515
6 Hornby, A. 牛津高阶英语词典. 北京：商务印书馆，牛津大学出版社（中国）有限公司，2004，389
7 清代学者. 康熙字典. 汉语大词典编纂处整理. 北京：汉语大词典出版社，2002，834
8 范晔. 后汉书. 北京：中华书局，2009，2999-3100
9 司马迁. 史记. 上海：上海古籍出版社，2015
10 王充. 论衡. 长沙：岳麓书社，2015，365
11 脱脱. 宋史. 北京：中华书局，1985
12 刘向. 战国策. 上海：上海古籍出版社，1985，992
13 刘安. 淮南子. 哈尔滨：北方文艺出版社，2013，118
14 金兆梓. 尚书注释. 北京：中华书局，2016
15 李渔. 闲情偶寄. 杭州：浙江古籍出版社，2008，305
16 许慎. 说文解字. 北京：中华书局，1978，248
17 Lewin. W, M. Klis. *Compact Stellar X-Ray Sources*. Cambridge: Cambridge UP, 2006, 461
18 Firth. J. Papers in Linguistics. 1934-1951. London: Oxford UP, 1957
19 Sterken. C, J. Manfroid. *Interstellar Extinction*. Berlin: Springer Netherlands, 1992, 1

四种“新”元素有了中文名

中国科学院、国家语言文字工作委员会和全国科学技术名词审定委员会2017年5月9日在北京召开新闻发布会，正式发布第113号、115号、117号、118号化学元素中文名称。它们的汉语发音分别为"nǐ"、"mò"、"tián"、"ào"。

这4种元素大多于2002-2003年后被科学家发现，其英文名、符号和汉译名写法分别为：1 nihonium, Nh,"nǐ" 字左右搭配“钅”+“尔”；2 moscovium, Mc, "mò" 字作“镆”；3 tennessine, Ts,"tián" 字左右搭配“石”+“田”；4 oganesson, Og, "ào" 字为“气”字头下加“奥”构成。据悉，表示第113、117和118号元素的汉字是新造的，名词委正在按照国际标准提案的要求履行手续，争取尽快取得其在ISO/IEC10646的区位码和字符集，实现电脑输入。

（宜司南）

DOI:10.16024/j.cnki.issn1002-0489.2017.02.020