关于机器翻译的评测问题

姚天顺，杨　莹
东北大学计算机科学与工程研究所

系统评测在学术上是一个专门的学科，国际上普遍重视。一个好的评测系统其开发难度并不亚于被测系统。系统评测对被测系统具有强劲的引导作用，推动被测系统逐步完善，是十分重要的。根据自然语言的固有特点，评测在系统的研制过程中同样也起着不可缺少的作用。著名的 L. Hirschman 和 H. S. Thompson 在 1999 年就指出："不论对系统的开发者还是技术用户而言，语音和 NLP 的评测都扮演着一个至关重要的角色。"

对评测研究赋予特殊关注的是美国政府的"自动化语言处理指导委员会"（ALPAC），60 年代他们对机器翻译的研究持过反对态度。想想这几十年，不无道理。现在他们以机器翻译和人工翻译比较的态度，从三个方面去考虑评测工作：即"速度、费用和质量"，然后给出评价。按"很清楚和理解"到"不可理解"共分成 9 个等级，得出最后的结论。非常注重市场和机器翻译的实际效益。

另一个组织就是 ARPA。他们从 1992 年开始也支持机器翻译的评测，见 O'Connell, O'Mara and White[19]的报告。由于该组织是一个提供资金的组织，关心的主要点是"前展的核心技术（Further the core technology）"，立足于全自动和批量处理，对不同语言和不同翻译技术的测试。

在美国政府、欧共体和上述类组织的推动下，比较著名的系统有：

1　EAGLES (Expert Advisory Group on Language Engineering Standards)系统[8]

按需求分析各种要求（显示的或隐含的需要），并不断细化，明晰，从而形成完整的描述，这些描述落实到设计的不同部分，按这些部分依此测试各个软件模块。该系统最重要的贡献是陈述的或隐含需求的形式化。其的目标是构造一些自动化过程。

整个系统从拼写、语法、风格检查到自然语言界面、机器翻译系统，包括：

1）书写辅助工具：

拼写检查器；语法检查器；风格检查器

2）信息管理工具：

自动索引系统；文本检索系统；自然语言的信息检索系统；具有信息管理的写作辅助工具

3）自然语言界面：

信息系统；数据库系统；计算机系统

4）翻译辅助工具：

机器翻译（translation memory）；专用工作站；术语管理数据库；电子单语和多语字典；字典访问系统；多语同义词典

5） 机器翻译系统

6） 自然语言生成系统

2 TEMAA(A Testbed Study of Evaliation Methodologies)[21]

这是一项 Sprogteknologi 中心（哥本哈根）、ISSCO(日内瓦)、Stichting Taaltechnolohgie(日本，宇都宫)、语言技术组（爱丁堡）和 Claris Ireland(都柏林)[21] 等开发的。主要用于英语、意大利语、丹麦语和荷兰语的拼音检查软件的检测。利用面向对象的程序设计，设计和利用参数化测试平台（PTB）方法进行的。

3 COBALT[8]

这是一个 LRE 项目下的子项目，利用 EAGLES 的评测方法，评测哪些能对财政类感兴趣的新闻文章的系统。

4 RENOS[8]

也是一个 LRE 项目下的子项目，利用类似 TEMAA 的方法评测哪些全文检索中减少噪声的系统。

5 TSNLP (Test Suites for Natural Language Processing)[3]

1993 年 12 月份开始，到 1995 年 10 月份结束。参加的有：艾塞克斯大学，DFKI GmbH (萨尔布鲁根), ISSCO (日内瓦) 和 Aerospatiale (巴黎)。这个项目是通用的，针对三种语言：英文、法文和德文，一个比较大的语言学处理的系统，在词法、语法、语义层进行评测。

当然还有很多评测系统，这里就不再泛述了。很多学术会议都有自动评测的专题。例如

International Conference on Human-Computer Interaction;

International Conference on Terminology and Knowledge Engineering;

International Conference of the evaluators' forum.;

COLING;

International Conference on Banff Knowledge Acquisition for Knowledge-Based Systems Workshop;

Message Understanding Conference;

Annual Conference on Online Documentation; International Conference on Research in the Consumer Interest;

International Conference on Theoretical and Methodological Issues in Machine Translation, ……

国际上关于评测研究已相当普遍，但是我们国内没有对此纳入重要日程，专们从事这方面研究的学者也较少，至于如何通过评测推动 NLP 的研究就更少。下面我们讨论三个问题。

1 评测在软件开发过程中的位置

通常，一个 NLP 的评测系统，不能简单地理解成对结果的测试。按照 EAGLES((Expert Advisory Group on Language Engineering Standards)的观点系统在软件工程生命周期中，评测是软件进入市场前的第一过程：它们按需求分析中的各种要求（显示的或隐含的需要）进行的。需求还要不断细化，明晰，从而形成完整的描述，这些描述落实到设计的不同部分，按这些部分依此测试各个软件模块。

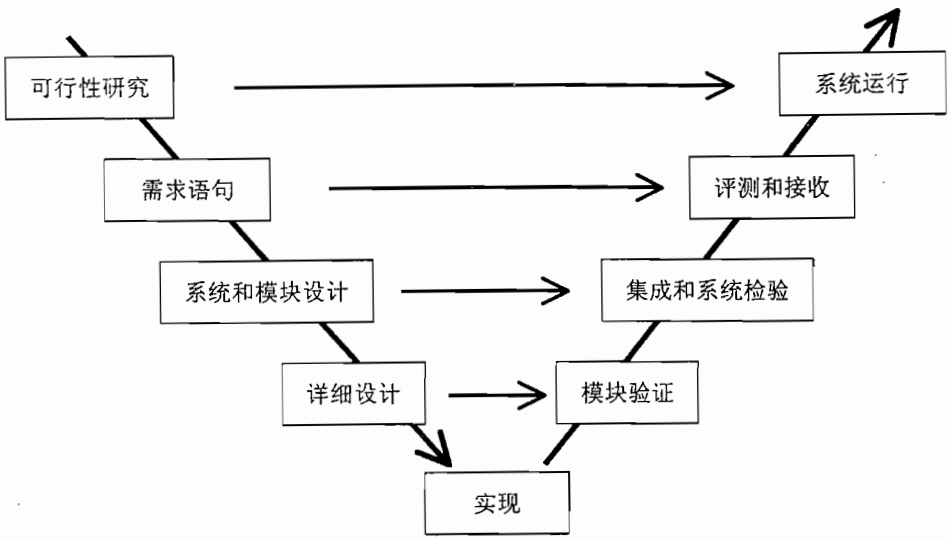

图 1　评测在软件开发中的位置

整个测试过程从可行性研究开始，提出需求语句，按需求给出系统和模块设计，再作详细设计，实现了评测软件，与此对应的是相应地逐步测试。这样一种需求分析，语言工程作为一个应用领域，总体上也遵循这样一个原则，但有很不同的，存在专门的属性，还必须要根据语言学的特点，处理需求分析过程，提炼出专用的版本。

2　ISO 9126 标准

NLP 作为一种软件，当然要符合一般 ISO 9126 软件评测标准。这个标准设计了一个评测系统的通用框架，试图设计出软件产品评价的全部质量特性。总共有六个方面的质量特性：即功能性，可靠性，可用性，效率，可维护性和可移植性。

这里一个主要观点为质量特性是属性层次组织的顶层：每个特性可分解为质量子特性，子特性本身又可以进一步分解。在同样的属性集中，特定评测或特定的软件质量观点可能更强调某些属性。ISO 提到了用户、开发者、管理者的观点。例如，ISO 的管理者观点如下：

管理者更感兴趣的是整体质量而不是特定质量性能，因此需要向个别性能赋予权值，反应商业需求。

管理者也需要权衡质量改进与管理标准，如期限延迟、价格过高等，因为他希望用有限的价格、人力资源和时间范围，获取最优的质量。

开发者比较注意环境的设置，而用户则关心最后的质量。

ISO 还建议了一个分为三个阶段的评测过程模型：每个方面的特性又都是非常广的。

1　质量需求定义：是评测的第一阶段，根据质量特征及其子特征指定的需求，以陈述的或隐含的需求为其输入，相关的技术文档，ISO 标准；输出为质量需求说明。它应该在软件开发前定义。

2　评测准备：是评测的第二阶段，由三部分构成：

a)　质量衡量标准选择。

b)　级别定义　测量得到的结果--值以级别的方式解释，即对需求满意的不同程度将

所取的值划分为范围，定义不同的级别标准。

c) 评估标准定义 综合不同特征评测的结果，给出标准定义。

3 评测过程：细化为三个步骤：

（1）测量 按规定的测量项目测量软件，得到测量值；

（2）分级 对测量值分级，得到级别水平；

（3）评估 按级别水平进行综合评估。

3 评测模型的建议

这两年，我们国家973项目专门设立了评测的研究项目。这个项目到底怎么做？要跟专家组商量。这里只是一个建议。建议参照EAGLES的思想，以ISO 9126标准为研究的起点，力图建立起一个MT软件评测体系。

吸收EAGLES 的最重要的思想是给出陈述的或隐含需求的形式化。最终的目标是构造一些自动化过程。这种过程在EAGLES中称为参数化测试平台（Parameterisable Test Bed, 简称PTB）。

1）评测中的主要概念-----形式化概述

（1）评测函数

评测就是确定某物对某人值多少。我们可以将评测描述为一个函数：

$$O \times U \rightarrow V$$

其中

- O 代表评测的对象集，表示任何对象，还包括语言知识的计算机程序，
- U 代表O的用户集，包括潜在的对O感兴趣的人或组织。U不仅包括用户环境，而且还包括对完成用户需求的整个系统中各个子系统。
- V 是值集，我们把它称为效用 (utility),

上面的基本函数可以进一步写为

$$O \rightarrow (U \rightarrow V) \cdots\cdots\cdots\cdots\cdots\cdots \quad (i)$$

$$U \rightarrow (O \rightarrow V) \cdots\cdots\cdots\cdots\cdots\cdots \quad (ii)$$

其中，（i）描述了基于对象的评测：给定某个对象，评测哪些用户喜欢它；

（ii）给出了基于用户的评测：给定某个用户，评测他们喜欢哪些对象。

（2） 特征描述

对象（及其成分）的描述在实际评测中起着中心的作用，在 EAGLES 中，评测对象是分类的，每一类都有自己的特征结构，即“属性和属性值对”的结构。

对于任何给定的O，属性指那些能被赋予值域中值的属性。属性应该与实效相关的观点选择，而且应该是可度量的，例如对于机器翻译的属性，我们可以设想有：

在词层

汉语分词的自动测试; 未登录词的自动切分

语言分析的词性标注; 概念的设定及其词汇覆盖率

词的多义或多概念的设定量

电子词典和概念词汇量量及其词法和语义的多级覆盖率（如吕叔湘常用词，现代汉语，中华大词典，...）;

……

短语层

短语库的常用短语，成语覆盖率；常用名词短语识别正确率

短语结构分析的正确率；……

句子层

目标语实词翻译的相似度自动测试；

……

语言处理环境层

语义或概念场分类体系的完整度；生语料库的分布和容量自动测试

带标语料的抽样测试；……

属性可以根据其可能的取值，区分类型，例如可以是布尔值（是/否）、列表值（无序值集）、可比较值（有序值集）、连续值、公制值（有起点和单位的实数值）、百分数等。

2）参数化测试台（PTB）

PTB (Parameterisable test bed) 是一个程序，特点如下：

- 接受评测对象和这些用户使用对象的参数；
- 可应用的测试方法库；
- 完成相测试的程序；
- 生成评估报告；

其中，

（1）对象参数：用“属性—值对”描述，不仅要构造对象集（子类，组成成分），而且要建立属性集（如功能有关，与可用性有关的属性等）。

（2）用户参数：用户实质上是对象的要求表，用户可以根据属性选项、属性的权值、属性值的说明来描述。

（3）方法库：对每一类属性的测试提供计算其值的方法。有些属性可以自动测试，例如可以开发测试拼写检查器的自动过程，但有些属性，如文字的修辞，漂亮等，自动评测就不太可行，由人工处理。不论自动测试，还是依赖人类帮助的测试，PTB 都应具备能够生成、集成并维护不同测试所用的测试材料库。

（4）测试：PTB 为每一测试或执行测试库的自动过程，如果不是自动，那由 PTB 的用户来完成测试，然后将结果输回 PTB。

（5）评估：比较测试结果和用户描述（用加权属性和说明），产生给定用户类对象的部分排列。

构造各种各样的 PTB 是评测系统的重点，也是我们的努力方向。目前我们已经实现一个非常有限的，评测分词及其词性标注的 PTB。但是要想对语言工程构造一个完整的 PTB 还是一个无限的庞大的工程。因为语言信息不断有新的对象类出现，而且不断有新的用户类出现，即使将它限制到一个特定的范围，创建这样的 PTB 工作量仍是很大的。

4 应用框架

设想了 PTB 以后，如何实现和它们的原则是什么？对于 MT 而言：

1） 属性集

评测准备阶段的核心目标是作标准，获得满足如下条件的属性集：

- 属性的值应该是可以通过观察、直接测量或由其他属性的值派生得到的；
- 给出表达所有清晰的或模糊的用户需求；
- 属性应该足够通用，可以应用到相似任务的系统中和不同的用户类中或这些系统的不同用法之中。

2） 需求

实际需求（陈述的或隐含的）起着核心的作用，因为这些需求最终产生属性。从用户抽取需求不总是直接的，我们已经从软件需求分析得到启示，需求分析应该选用基于可测量的原语（primitives）。

3） 方法

这里采用的方法不是某一个，而是不同的方法集，它们用来建立各种各样的属性值。以解决各种各样不同的问题，

4） 测量

（1） 测量的有效性

测量的有效性是一个至关重要的问题：如果所用测量是无效的，整个评测都是无价值的。有效性分为内部有效性和外部有效性。

内部有效性

内部有效性是指每一量度都是被测对象的适当属性的合适测量，例如，在我们这里，考虑到用户和用户的工作环境，内部有效性直接反映的是用户的利益，例如，拼写检查器的一个属性是语言，希腊语的拼写检查器不能用于处理意大利语文章。只在某个被测领域有效。

外部有效性

外部有效性是通过计算使用这个测量所得到的结果与外部标准之间相互吻合的程度确定。例如在拼写检查器的测试时，字典大小这一属性就具有外部有效性。

（2） 测量的可靠性

测量必须具有可靠性，所谓可靠性是指当测量应用到同一现象时，能够得到相同的值。可靠性是通过计算测量两次出现所得到的值之间的相互关系的程度（coefficient）确定。要保证可靠性，建议尽可能减少人的干预，但在目前的情况下，有些属性值还必须由人来决定。当人参与时很难保证可靠性，因此，研究合适的方法进行评测是十分重要的。

（3） 测试材料

通用自然语料

在 EAGLES 中称为 Test sets，指自然文本集合，著名的有：

- 布朗英语语料 BNC
- 法语语料 TLF
- 选自加拿大 Hansard 的双语（英语—法语）语料 LDC（平行文本的测试集）
- 中国 21 世纪
- 人民日报语料

随着处理和存储大量文本能力的不断提高，搜集测试集并开发应用工具的兴趣越来越重视。例如欧洲共同体 LRE 工程 MLCC，专们搜集欧洲语言的多种语言，构造平行多语语料。

语料库的选择切忌片面性。自然语料还应区分生语料和带标的语料。

人工语料

在 EAGLES 中称为 Test suites，是指人工建造的语料，用来测试系统的某些特殊现象。这是相当复杂的。即使是句法问题，定义句型就存在问题，至于义、语用上，语料的构造就更成问题。同时，语料的大小问题也很重要，随着语料快速变大，就必须解决语料管理和分析问题。

（4）翻译评测的度量

机器翻译的度量可以分为被译文本、翻译组织、和外部环境，

被译文本可用以下方面的度量：

- 被译文本的数量
- 规则系统的大小
- 处理文本的类型
- 统计语料库的特点
- 包含的语种
- 抽样检查翻译句子的质量
- ……

翻译组织用以下的度量描述：

- 翻译组织类型
- 翻译组织大小
- ……

组织中翻译的外部环境用以下的度量描述：

- 整个组织的性质
- 设备和研究环境的水准
- 国际活动的数量
- 国际活动的性质
- 组织的语言策略
- ……

（5） 评测过程

要评测的过程力求标准化。其基准的属性要有两个考虑：

— 选择的重要或核心的属性（有用性 usefulness）

— 可高效地、可靠地和有效测量（可行性 feasibility）

以下就将围绕这两点进行评测。

（1） 有用性

选择软件的最主要的考虑之一是看它是否有用。包括翻译速度和翻译质量，二者通常是缺一不可而又呈反比关系；其相对权值随翻译类型的不同而不同。

总结机器翻译中影响速度和质量的各种情况，要求满足：

- 语料库中文本类别的分布和重复度；
- 语言分析难点处理的准确率；
- 同一个概念在不同句子中分析的一致性。
- ……

（2）可测量属性

- 在线属性 包含

— 短语容量
— 翻译速度
— 对译词的命中率（实词匹配的百分比）
— ……

- 离线属性 包含
  - — 带标树库的设计和容量
  - — 切分和对齐的成功率
  - — 带标语料库规模及其自学习的能力
  - — 语义分类体系
  - — 句型体系
  - — 规则的体系结构
  - — ……

由此可见，评测系统所要评测的不仅是最后的结果，包括机译的环境，工具，效率和完备等。不仅如此，测试软件本身也必须准备良好的测试环境和工具，包括良好的带标语料，语法语义的体系，超大的字库、词库。根据不同的对象，设定不同的属性进行测试，包括、树库，被译文本、翻译组织、翻译外部环境的度量，在线和离线的属性，设计和给出汉英翻译系统的各个评测过程。可能在词层和短语层自动测试没有多大问题，进入了句子层问题就多了。可以说，评测系统的工作难度和工作量完全不亚于一个处理系统。如果这样的系统能够完成，那怕是部分完成，由于评测是带有某些强制性的，对推动 NLP 的研究，将会有很大的促进作用。

参考文献

[1] Ahmad, K. Terminology and Knowledge Acquisition: A Text-based Approach, Proceedings of Terminology and Knowledge Engineering.

[2] Athappily, K. and Galbreath, R. Oractical Methodology Simplifies DSS Software Evaluation Process, Data Management 24(2): 10-28. 1986

[3] Balkan, L., Netter, K., Arnold, D. and Meijer, S. TSNLP – test suites for natural language processing, Proc. of the Language Engineering Convention, ELSNET, Paris, pp.17-22., 1994

[4] Boisen, S. and Bates, M. A practical methodology for the evaluation of spoken language system, Proceedings of the Third Conference on Applied Natural Language Processing, Trento, pp.162-169.

[5] Bukowski, J. Evaluating software test results: A new approach, Proceedings Annual Reliability and Maintainablity Symposium, Philadephia, USA, 27-29. Jan, pp.369-375. 1987

[6] Cary, R. and Sproles, G. Evaluating product testing methods: A theoretical framework, Home Economics Research Journal 7: 66-75. 1978

[7] Chinchor, N. MUC-3 evaluations metrics, Proceedings of the Third Message Understanding Conference (MUC-3), Morgan Kaufmann, San Mateo, CA, pp.17-24. 1991

[8] EAGLES, Evaluation of Natural Language Processing Systems, EAGLES DOCUMENT

EAG-EWG-PR.2, Version of Septmber, 1995
[9] Falkedal, K. Evaluation methods for machine translation systems: An historical overview and critical account, Issco draft report, University of Geneva, Geneva.
[10] Flickinger, D., Nerbonne, J., Sag, I. And Wasow, T. Toward evaluation of NLP systems, Hewlett Packard Lab., Palo Alto, CA.
[11] Gallier, J. and Jones, K.S. Evaluating natural language processing system, Technical report no.291, University of Cambridge Computer Lab., Cambridge.
[12] Hayward, S., Breuker, J.A. and Wielinga, B.J., The KADS methodology: Analysis and design for knowledge based systems, ESPRIT P1098 Deliverable Y1, STC Technology Ltd., Alborg., 1987
[13] Hoge, M., Hohmann, A. and Mayer, R. Evaluation of TWB – operationalization and test results, Final report of the ESPRIT II P2315 Translator's Workbench (TWB), Fraunhofer Society IAO and Mercedes – Benz AG Stuttgart. 1992
[14] JEIDA., JEIDA methodology and criteria on machine translation evaluation, JEIDA, Tokyo, 1992
[15] King, M. and Falkedal, K., Using test suites in the evaluation of machine translation systems, Proceedings of COLING-90, ACL, Helsinki, pp.211-219. 1990
[16] Mikheev, A. and Moens, M., KADS methodology for knowledge-based language processing systems, in B.R. Gain and M. Musen (eds) Proceedings of the 6th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop, pp.5-1-5-17, 1994
[17] MT. On evaluation, Machine Translation 8(1.2). Edited by D.Arnold, R.L.Humphreys and L.Sadler. 1994
[18] MUC-3. Proceedings of the Thrid Message Understanding Conference (MUC-3), Morgan Kaufmann, San Matoo, CA. 1991
[19] O'Connell, T., O'Mara,F. and White, J. The ARPA MT evaluation methodologies: Evolution, lessons and further approaches, Proceedings of the First Conference of the Association of Machine Translation in the Americas, Colunbia, USA 1994
[20] Sundheim, B. Overview of the third message understanding evaluation and conference, Proceedings of the Third Message Understanding Conference (MUC-3), Morgan Kaufmann, San Mateo, CA, pp.3-24, 1991
[21] Thompson, H., TEMAA: A testbed study of evaluation methodologies: Authoring aids, Proceedings of the Language Engineering Convention, ELSNET, Paris, pp.147-149
[22] Vainio-Larsson, A. Evaluating the usability of user interfaces: Research in practice, in D.Diaper, D.Gilmore, G.Cockton and B.Shackel (eds), Human Computer Interaction – INTERACT'90, Elsevier, Amsterdam, pp.323-328. 1990
[23] 俞士汶等，关于汉英机器翻译测试大纲的思考，ICMI99 (HONG KONG)，IV38-40.
[24] 刘群等，汉英机器翻译的难点分析，1998 中文信息处理国际会议论文集，清华大学出版社，1998，11
[25] http://issco-www.unige.ch/projects