外语教学与测试研究丛书

主编：金 艳 吴 江

在线英译汉即时自动评分

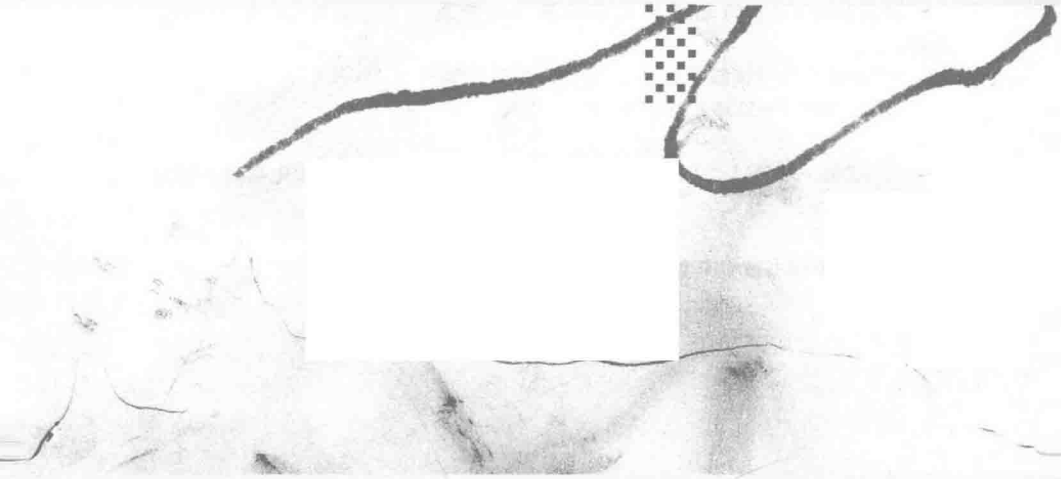
Instant Automated Scoring of Online English-Chinese Translation

田 艳 著

外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS

外语教学与测试研究丛书

主编：金 艳 吴 江



在线英译汉即时自动评分

Instant Automated Scoring of Online English-Chinese Translation

田 艳 著

外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS

图书在版编目 (CIP) 数据

在线英译汉即时自动评分 / 田艳著. — 北京 : 外语教学与研究出版社, 2013.12
(外语教学与测试研究丛书 / 金艳, 吴江主编)
ISBN 978-7-5135-3962-3

I. ①在… II. ①田… III. ①英语－翻译－评分－自动化系统－研究 IV. ①H315.9-39

中国版本图书馆 CIP 数据核字 (2013) 第 318816 号

出 版 人 蔡剑峰
责任编辑 李婉婧
封面设计 覃一彪 吕 茜
版式设计 涂 俐
出版发行 外语教学与研究出版社
社 址 北京市西三环北路 19 号 (100089)
网 址 http://www.fltrp.com
印 刷 大恒数码印刷 (北京) 有限公司
开 本 650×980 1/16
印 张 10.5
版 次 2013 年 12 月第 1 版 2013 年 12 月第 1 次印刷
书 号 ISBN 978-7-5135-3962-3
定 价 36.90 元

购书咨询：(010)88819929 电子邮箱：club@fltrp.com
如有印刷、装订质量问题，请与出版社联系
联系电话：(010)61207896 电子邮箱：zhijian@fltrp.com
制售盗版必究 举报查实奖励
版权保护举报电话：(010)88817519
物料号：239620001

We can only see a short distance ahead, but we can see plenty there that needs to be done.

—— Alan Turing

序

田艳博士的《在线英译汉即时自动评分》一书就要出版了。她请我为这本书写序言，我欣然答应了。

田艳与我是在1999年中国科技翻译学会举办的一次学术会议上认识的，这次会议的主题是讨论大中型企业中的科技翻译问题。她在这次会议上做了机器翻译系统的演示和比赛，我还做了一个关于机器翻译的发言。田艳对于机器翻译产生了浓厚的兴趣，她对我说，她决定投身于令人神往的自然语言处理（Natural Language Processing）事业。田艳的这个决定使我很受感动，当时她已经是大学英语系的骨干教师，在英语教学方面已经有了丰富的经验，可以说是功成名就了。对于很多人来说，完全可以吃老本了，而自然语言处理对于她来说完全是新的东西，对她而言有一个知识更新的再学习问题。但是，田艳不畏困难，毅然走上了研究自然语言处理的道路。她调入了上海交通大学外国语学院，接着又投考计算机系的博士研究生，进行了知识更新的再学习。现在，她基本上完成了这一再学习过程，成为一位自然语言计算机处理的专家了。田艳这种对于新知识不断探索的精神是值得我们学习的。她建立的在线英译汉即时自动评分系统既具有实用价值，又具有理论意义，是自然语言处理研究的一个很好的成果。

从形式上说，自然语言处理中的操作基本上可以归纳为三个问题：分割（Segmentation）、分类（Classification）和等同（Identification），取英文第一个字母，可以把这三个问题叫做“SCI问题”（SCI problems）。自然语言处理中的所有操作，都围绕着“SCI问题”来进行。

“SCI问题”中的第一个问题是“分割”，也就是对于自然语言的字符串进行切分。

汉语书面文本是连续的汉字流，单词与单词之间没有空白，为了找出单词与单词之间的这种空白，就要进行书面文本的自动切词，这就是众所周知的汉语书面文本的自动切词问题。这种自动切词是自然语言处理中最典型的“分割”问题。

有人认为分割只是汉语书面文本的特殊问题，在其他语言中，特别

是在印欧语中不存在分割的问题。其实这是一种错误的看法。

我们知道，英语形态分析中有“接词”（cliticization）成分的分析问题。在英语中，带省略符号（'）的符号串，称为“接词”（cliticization）。在形态分析时，要把它们切分为不同的词例，省略符号前面的部分叫做“接词前段”（proclitics），省略符号后面的部分叫做“接词后段”（enclitics）。一般说来，英语中的接词有如下几种情况：

—Let's = let + us

—I'm = I + am

—{it, that, this, there, what, where}'s ={~} + is

—He's = (He + is) or (He + has)

—A've = A + have

—A'll = A + will

—A're = A + are

—A'd = (A + would) or (A + had)

—{is, was, are, were, has, have, had}n't ={~} + not

—can't = can + not

—won't = will + not

—dog's = of + dog (e.g. dog's tail = tail of dog)

接词现象既涉及动词，也涉及名词，英语动词的接词现象可以归纳如下：

完全形式	接词
Am	'm
Are	're
Is	's
Will	'll
Have	've
Has	's
Had	'd
Would	'd

注意，在这个表中，He's = (He + is) or (He + has)，A'd = (A + would) or (A + had)。这时，接词出现歧义，应当根据上下文来消除歧义。

显而易见，英语中接词成分的处理和切分就是一种分割问题，这与汉语书面文本的切词问题是非常接近的。不过这样的分割问题比较简单。

分割问题在闪米特语言中也存在，这些语言附着成分的分割也是相当困难的。例如，在阿拉伯语和希伯莱语中，定冠词（阿拉伯语为 Al，希伯莱语为 ha）附着在名词前面，为了进行词类标注、剖析或其他的自然语言处理工作，必须把这些定冠词和它们后面的名词切分开来。阿拉伯语的其他前附着成分还有介词 b（相当于英语的 by/with）和连接词 w（相当于英语的 and）；阿拉伯语还有表示代词的后附着成分，例如，单词 wbHsnAthm（相当于英语的 and by their virtues）中包含意思为 and, by, their 的三个接词成分，一个意思为 virtue 的词干以及一个表示复数的词缀。由于阿拉伯语是从右往左读的，因此这些成分在一个单词中的顺序也是从右而左的。

	前附着成分	前附着成分	词干	词缀	后附着成分
阿拉伯语	w	b	Hsn	At	hm
注释词	and	by	virtue	s	their

可见，“分割”是自然语言处理中一个最为常见的普遍问题，不仅汉语书面文本的自动处理存在分割问题，其他语言的自动处理中也存在分割问题。我们可以把它叫做 SCI 问题中的 S 问题。

SCI 问题中的第二个问题是“分类”。

自然语言中词类的划分就是一个典型的分类问题。这样的分类问题在自然语言处理中表现为“词类标注”（Part-of-Speech tagging，POS tagging），简称为“标注”（tagging），这是给语言中的每一个单词指派一个词类或者词汇类别标记的过程。标记通常也用来标注标点符号。词性标注不但是机器翻译形态分析的重要组成部分，而且它在语音识别和信息检索中都起着越来越重要的作用。

在进行词类标注时，标注算法的输入是单词的符号串和词类标记集（tagset）。算法的输出要让每一个单词都标上一个单独的而且是最佳的标记。所以，词类标注的过程实际上是对每一个单词进行分类的过程。

下面是国家语委语料库中一个词类标注的例子。

鸟 /n 是 /v1 大 /a 自然 /n 的 /u 歌手 /n，/w 鸟语 /n 就 /d 是 /v1 大 /a 自

然 /n 的 /u 音乐 /n 和 /c 诗歌 /n 了 /u。/w

其中，n 表示名词，u 表示助词，vl 表示系动词，w 表示标点符号，d 表示副词，c 表示连接词，a 表示形容词。文本中的每一个单词都被分派给一个词类标记，也就是说，每一个单词都被正确地分类了。

我们再给出美国 ATIS 语料库（一个关于航空旅行订票对话的语料库）中的英语词类标注的例子。每一个单词都分派了一个词类标记；标点符号没有标注。

Book/ VB that/ DT flight/ NN .

Does/ VBZ that/ DT flight/ NN serve/ VB dinner/ NN ?

英语的词类标记来自宾州大学的词类标记集。

这些都是非常简单的分类实例。

这样的词类标注问题在各种语言中都普遍存在。

在汉语中，兼类词的自动标注是一个令人棘手的问题。例如，在短语“这本书的出版”中，“出版”就是一个兼类词，可以是名词，也可以是动词。由于出现在“的”后面，有人可能认为应当标注为名词，但如果这个“出版”标注为名词，那么“这本书的不出版”中的“出版”前面出现否定副词“不”，又似乎应当标注为“动词”；“这本书的迟迟不出版”中的“出版”前面除了常出现否定副词“不”，而且还出现表示状态的形容词重叠式“迟迟”，似乎更应当标注为动词了。究竟是标注为名词还是动词，使我们处于进退维谷的困难境地。

词类标注是一个最普遍的分类问题，除此之外，自然语言处理中还有语义的分类问题和句型的分类问题。这些分类问题，我们可以把它们叫做 SCI 问题中的 C 问题。

SCI 问题中的 S 问题和 C 问题，在田艳的在线英译汉即时自动评分的研究中都遇到了。她使用中科院计算所汉语词法分析系统 ICTCLAS（Institute of Computing Technology, Chinese Lexical Analysis System）对学生译文进行了自动切分和词类标注，从而为英译汉的在线自动评分奠定了基础。

SCI 问题中的另一个问题是“等同”问题，也就是 I 问题，这是判定两个语言单位是否同一的问题。

查词典是自然语言处理中经常遇到的问题，这是单词一级的等同问题。所谓“查词典”，就是判定文本中的单词是否与机器词典中存储的单词等同。对于有屈折变化的英语来说，由于文本中的单词经常是以变化形式出现的，而词典中的单词则往往以原形存储，这样，就必须把文本

中以变化形式出现的单词还原为它的原形，然后再与词典中存储的原形词相匹配，从而实现“等同”问题的判定。例如，flown 这个在文本出现的单词是 fly 的过去分词形式，需要首先还原成它的原形 fly，然后再查机器词典。

“等同”问题一般要在“分割”和“分类”问题的基础上来进行。文本在进行了分割之后，单词与单词之间的界限被泾渭分明地区分开来了，这样才有可能查词典，进行单词的等同问题的判定。

文本中的单词在进行了“分类”之后，也有利于等同问题的判定。例如，books 可能是名词 book（书）的复数形式，也可能是动词 book（预定）的单数第三人称现在时形式。在判定了 book 的词类类别之后，就可以判定 books 究竟是与名词 book（书）等同，还是与动词 book（预定）等同了。

上面所述的单词的等同判定比较简单，而单词组合（词组和句子）这样的语言片段的等同判定就比较复杂了。一般说来，如果存在语言片段 A 和语言片段 B，要判断这两个语言片段是否等同，应当具备三个必要条件。

第一，两个语言片段中的单词是否相同：例如，语言片段 A=“看 打 球 的 同学”，语言片段 B=“看 打 球 的 同学”，其中的单词完全相同，这是判定两个语言片段等同性的第一个必要条件。显而易见，判断单词的等同性首先要对书面文本进行切分，这样的等同性判定是在分割的基础上进行的。

第二，两个语言片段中单词的顺序是否相同：如果两个语言片段中的单词相同而单词的顺序不同，这样语言片段就是不等同的。例如，如果语言片段 A=“看 打 球 的 同学”，其中单词的顺序是“看→打→球→的→同学”，而语言片段 B=“同学 的 看 球 打”，其中单词的顺序是“同学→的→看→球→打”，这两个语言片段中的单词虽然相同，但是单词的顺序不同，它们显然不能是等同的语言片段。可见，单词的顺序相同是判定两个语言片段是否等同的第二个必要条件。

第三，两个语言片段中各个单词所在的层次是否相同：如果两个语言片段中的单词相同，单词的顺序也相同，但其中各个单词所在的层次不完全相同，也不能判定这两个语言片段是等同的。例如，如果语言片段 A 的层次是“[看 打 球 的][同学]”，其中 [看 打 球 的] 作为 [同学] 的定语，是一个偏正结构；而语言片段 B 的层次是“[看] [打 球 的 同学]”，其中 [看] 作为谓语，而 [打球的同学] 作为 [看] 的直接宾语，是一个述

宾结构。在这种情况下，我们仍然不能判定它们是等同的。可见，除了单词相同和单词的顺序相同这两个必要条件之外，语言片段的层次相同是判定它们等同性的第三个必要条件。

事实上，即使两个语言片段具备了上面的三个必要条件，也不一定就能判定它们是等同的。例如，语言片段 A=“[张三][理发]”，语言片段 B=“[张三][理发]”，其中的单词相同，单词的顺序相同，整个语言片段的层次也相同。但如果在语言片段 A 中，张三是一个理发师，那么他是在给别人理发，是“理发”这个行为的施事者；而在语言片段 B 中，张三不是理发师，他是被理发师理发，是“理发”这个行为的受事者，那么，这两个语言片段的含义显然是不等同的。要判断这两个语言片段的等同性，需要关于现实世界的常识性知识，是一件非常困难的工作。在特殊情况下，如果“张三”是一个理发师，而且他也需要别人给他理发，那么，要判断“[张三][理发]”这个语言片段的等同性显然就更加困难了。

可见，尽管我们可以研究出判定两个语言片段等同性的某些必要条件，但是要最终找到判定语言片段等同性的必要而充分的条件，至今仍然是一个非常困难的问题。因此我们可以说，SCI 问题中的 I 问题的判定，是自然语言处理中在理论上尚未完全解决的一个难题。

田艳在这本书中要研究在线英译汉即时自动评分，就需要判定学生译文的句子和标准译文中的句子之间的等同性问题，这样，田艳就必须面对 SCI 问题中的 I 问题，而她所进行的自动分词和词类标注则为研究 SCI 问题中的 I 问题做了必要准备。根据我们前面的分析，这个问题显然是一个非常困难的问题。

为了解决这一问题，田艳首先尝试了字符串匹配方法。她将学生译文与标准译文中的一串字符进行完全匹配，判定学生译文中的字符串是否与标准译文中的某一字符串相同。这种方法接近于我们上面所说的判定两个语言片段是否等同的第一个条件，也就是判定两个语言片段中的单词是否相同，字符串往往就是单词。实验结果显示，采用这种方法自动评分，结果与人工评分结果相距甚远，分数普遍过低，不符合实际情况。

由于 SCI 问题中等同性问题的困难，田艳没有再继续使用形式对比的方法进行等同性问题的研究，她转而借助于语义，采用了基于语义的自动评分方法，试图把“语义”引入到 SCI 问题的等同性判定这个困难的问题中。自然语言处理的经验说明，每当我们遇到了困难而几乎到了

“山穷水尽疑无路”的境地时，采用语义往往使我们摆脱困境，进入“柳暗花明又一村”的福地。采用语义确实是自然语言处理中一条又方便又省力的捷径。我认为她的这个“语义”转向的决定是明智的。

田艳先利用《知网》资源，把学生译文与标准译文中的主要实词（动词、形容词和副词）进行语义相似度计算，其他词类仍然采用关键词匹配评分。实验结果显示，此方法比前一种基于形式匹配的方法有了明显的改善，所得到的学生分数接近于人工评分结果。

田艳又引入了《同义词词林扩展版》，对标准译文中的名词进行语义扩展，生成标准译文中名词的同义词集合，然后将学生译文中的名词与这个同义词集合中的名词进行匹配，从而实现了实词中名词的语义评分，除名词、动词、形容词和副词以外的其他词类仍然采用关键词匹配方法评分。实验结果显示，此方法比只采用《知网》的语义评分结果又有所改善。

最后，为了弥补上述只考虑词一级的测评而忽略全句整体意义的评分方法，田艳还尝试加入了句式模板匹配的方法，把基于词一级的评分方法和句式模板匹配整合在一起，使自动评分既做到了对语言点的测评，又实现了对整句的测评。实验结果表明，这种评分方法所得到的结果优于上述所有方法，自动测评的分数更加接近于人工评分的分数。

田艳采用这些方法的效果，说明了在解决 SCI 问题中的等同性问题时，引入“语义”和“句式模板”是很有必要的。但我认为，尽管田艳使用上述方法取得了一些进展，SCI 问题中的等同性问题还远远没有解决。

由田艳的研究可以看出 SCI 问题中等同性问题的困难程度。目前在自然语言处理的研究中，不论是 SCI 问题中的分割问题、分类问题还是等同问题，都还没有很好地得到解决，特别是没有从理论的深度进行抽丝剥茧那样仔细的探讨。SCI 问题的进一步研究是自然语言处理取得突破性进展的关键问题，值得我们进一步深入探讨。

对于英汉翻译的自动评分来说，译文的等同性实质上是以原文与译文的翻译等同性为基础的。机器翻译的奠基人之一 Weaver 早就指出，在机器翻译中，原文与译文“说的是同样的事情”，因此，当把语言 A 翻译为语言 B 时，就意味着从语言 A 出发，经过某一“通用语言”（Universal Language）或“中间语言”（Interlingua），然后转换为语言 B，这种“通用语言”或“中间语言”，可以假定是全人类共同的。所以从本质上说，英汉翻译中译文的等同性问题归根结底是译文是否与

“通用语言”或“中间语言”等同的问题。然而，这样的等同性问题怎样求解，怎样通过计算机程序来进行计算，显然还是一个非常值得深入探讨的问题。

英国朴茨茅斯大学（University of Portsmouth）的自动文本评分系统（The Automated Text Marker，简称 ATM）利用“概念依存关系”（conceptual dependency），充分挖掘了句子深层的语义，一旦正确建立了概念之间的依存关系，那么无论被试者的答案形式怎样变化，只要与这种依存关系等同，就能判断其正确，实现了在句子层面上的概念和概念关系等同性的评测。ATM 自动文本评分系统的这种概念依存关系是在概念的基础上建立起来的，已经比较接近于 Weaver 所说的“通用语言”或“中间语言”。他们的工作是 SCI 问题中的 I 问题判定研究中一个很鼓舞人心的进展，我们应当密切关注这样的进展。

田艳的英译汉评分是在 Web 上进行的，因此，她的工作与 Web 有着密切的关系。我们现在所指 Web 就是 WWW，是基于 Internet 的计算机网络。用户使用 WWW，可以通过互联网（Internet）访问存储在世界范围内的 Internet 上的海量信息。WWW 是根据“客户端—服务器”（Client-Server）的模式来进行工作的。客户通过叫做“客户端”（Client）的程序与远程存储着数据的“服务器”（Server）连接，Web 的浏览通过叫做“浏览器”（browser）的 Client 程序来进行（例如 Navigator, Internet Explorer 等）。Web 浏览器把用户的提问传送给远程的服务器搜索有关的信息，然后返回搜索到的文件，这些文件使用“超文本标记语言”（Hyper Text Makeup Language，简称 HTML）书写，最后在客户端用户的计算机屏幕上显示出来。

Web 的操作依赖于超文本文件的结构。超文本可以让网页的作者把他们的文件与 Web 的其他文件进行“超链接”（Hyperlink），从而看到 Web 上的有关文件。

Web 的概念最早是 Tim Berners-Lee 于 1989 年提出的。当时他在瑞士的欧洲核研究中心（Centre European pour la Recherche Nucleaire, 简称 CERN）工作，写了第一个 WWW 的 Server 和 Client 程序，并且把它们叫做 World Wide Web。1989 年 3 月，Tim Berners-Lee 给 CERN 的高层领导提交了一个建议。在这个建议中，他分析了当时使用的层级式信息组织方法（hierarchical organization of information）的缺点，同时又指出了基于超文本系统（Hypertext System）的优点，初步提出了建立“分布式超文本系统”（Distribution Hypertext System）的基本方法。可惜他的这

个建议没有得到 CERN 高层必要的支持。

1990 年，Berners-Lee 又再次向 CERN 提出他的建议，这一次他的建议得到了支持。于是，Berners-Lee 和他在 CERN 的同事们立即采用分布式超文本系统的思想来研究 Web，为 Web 后来的发展做了奠基性的工作。他们为此研制了 Web 的服务器和浏览器，并研制了客户端和服务器之间的通信模型、超文本传输协议（HyperText Transfer Protocol，简称 HTTP）、超文本标记语言（HyperText Makeup Language，简称 HTML）、通用资源定位器（Universal Resources Locator，简称 URL，也就是网址）等等。

1993 年 2 月，美国伊利诺斯大学（University of Illinois）国家超级计算机应用中心（National Center of Supercomputer Application）的 Marc Andereeson 和他的研究小组设计了使用 Mosaic 技术的用户图形界面，并把它用来作为 Unix 的 Web 浏览器。短短的几个月之内，Macintosh 和 Windows 的操作系统都先后使用了 Mosaic 的用户图形界面技术。用户只要点击计算机屏幕上的图形，就可以对计算机进行各种操作。1994 年，Jim Clarc 与 Marc Andereeson 合作，成立了 Mosaic Communication 公司，后来改名为 Netscape Communication 公司。在几个月之内，他们就研制出了 Netscape 的浏览器，并在 Web 用户中普及。1995 年 8 月，微软公司公布了他们的 Web 浏览器 Internet Explorer，并向 Netscape 挑战。从此，用户就可以通过浏览器在 Web 上随心所欲地漫游了。

Tim Berners-Lee 创立的 World Wide Web 以及 Mosaic 浏览器的出现，是 Web 发展历史上两个最重要的事件，它们使得 Web 能够迅速地在用户中得到推广和普及。

Internet 是 Web 的通信网络。没有 Internet，Web 是不可能发挥其功能的。Internet 的前身是计算机网络 ARPANET，这个计算机网络是在美国国防部高等研究计划处（Advanced Research Project Agency 简称 ARPA）的支持下研制的。早在 1969 年 ARPANET 就建成了。1972 年，ARPANET 在计算机与通信第一次国际会议上表演，ARPA 的科学家们出色地利用 ARPANET 把处于四十多个不同地方的计算机连接在一起。后来，这个 ARPANET 进一步发展成为当今的 Internet。

在 1973 年，Vinton Cerf 和 Bob Kahn 就开始研究“网络协议”（Internet Protocol）。1974 年，他们发表了《传输控制协议》（Transmission Control Protocol）的文章，正式把他们提出的协议叫做 TCP/IP 协议（Transmission Control Protocol / Internet Protocol）。TCP/IP 协议可以使得计算机网络彼

此连接起来，彼此进行通信。但是直到 1982 年，TCP/IP 协议才正式得到采用，Internet 使用 TCP/IP 协议把不同网络联系起来了。

为了有效地获取分布在全世界网络上的信息，需要研制“搜索引擎”(search engine)。1993 年，美国斯坦福大学（Stanford University）的 6 名学生研制了搜索系统 Excite；1994 年，美国得克萨斯大学（Texas University）研制了 EINet Calaxy；同年，著名的搜索引擎 YAHOO 问世。1998 年，斯坦福大学的 Sergey Brim 和 Larry Page 推出了搜索引擎 Google。2005 年，微软推出了搜索引擎 MSN。

为了促进 Web 在全世界范围内的推广和使用，美国麻省理工学院（MIT）和瑞士的 CERN 在 1994 年成立了万维网协会（The World Wide Web Consortium，简称 W3C），W3C 是万维网的国际性组织。W3C 的成立使得 Web 在国际范围内迅速地得到普及，几乎每一个现代人的生活和工作都与 Web 息息相关。自 1994 年第一次 W3C 会议召开以来，每年都召开一次 W3C 的国际会议。

通过这一系列的努力，Web 在全世界范围内得到了普及，我们也随之进入了一个多语言信息网络的新时代。在这样的背景下，国内外开始了利用 Web 进行语言测试的研究，并且取得了明显的进展。

利用 Web 进行语言测试一般使用 HTML 语言来编写测试工具。测试文件由 HTML 文件组成，存放在考试设计者的服务器上，然后被下载到考生的计算机上进行。考生使用 Netscape Navigator 或 Microsoft Internet Explorer 等 Web 浏览器来解读和展现下载的 HTML 文件，在自己的计算机上答题，然后把答案发送到服务器上，或使用已下载的评分功能，得到考试结果。

Web 为语言测试的网络化创造了很好的条件。目前，多项选择(multiple choice)、完型填空（cloze test)、完成语篇（discourse completion)、论文写作（essays)、阅读理解（reading comprehension）的短文回答问题(brief-response questions）等已实现了基于网络的自动测评。除文本形式的网上测试题目外，还出现了音频和视频的网上测试题目。

目前国内外还没有学生在线英译汉即时自动评分系统研究，所以田艳的研究没有可以直接借鉴的成功经验；而国外其他语言的自动测评的成功经验则由于汉语本身的特殊性，也不能直接照搬。田艳出色地克服了这些困难，建立了在线英译汉即时自动评分系统。这个系统是一个基于 Web 的 client/server 结构系统，服务器端采用 ASP.NET Framework 2.0 的 C# 作为脚本语言。客户可使用 Internet Explorer 等一般的浏览器来浏

览这个网站的网页，进行自动测评。田艳还采用机器翻译技术，研究了部分标准译文（来自英语名词短语的译文）的自动生成和自动筛选。这样的探索也是很有实用价值的。田艳对于在线英译汉即时自动评分的这些研究是 Web 上自然语言处理的一个很好的成果。

目前，除了 Web 上的翻译文本自动评测之外，Web 上多语言的机器翻译（machine translation）、信息检索（information retrieval）、信息抽取（information extraction）也正在迅猛地发展。语言辨别（language identification）、跨语言信息检索（cross-language information retrieval）、双语言术语对齐（bilingual terminology alignment）和语言理解助手（comprehension aids）等自然语言处理的多语言在线处理技术（multilingual on-line processing）已经成为了 Web 技术的重要支柱。

在多语言信息网络时代，科学技术的发展日新月异，新的信息、新的知识如雨后春笋地不断增加，出现了“信息爆炸”（information explosion）的局面。现在，世界上出版的科技刊物达 165,000 种，平均每天有大约两万篇科技论文发表。专家估计，我们目前每天在因特网上传输的数据量之大，已经超过了整个 19 世纪的全部数据的总和；我们在 21 世纪所要处理的知识总量将要大大地超过我们在过去 2500 年历史长河中累积起来的全部知识总量。Web 上 90% 以上的信息都是文本信息，它们都是以语言文字为载体的信息，也就是说，Web 世界主要是由语言文字构成的。为了说明 Web 上自然语言处理的重要性，我们可以把它与物理学做出如下的类比：“我们说物理学之所以重要，是因为物质世界是由物质构成的，而物理学恰恰是研究物质运动的学科；我们说自然语言处理之所以重要，是因为 Web 世界主要是由语言文字构成的，而自然语言处理恰恰就是研究语言文字自动处理的学科。”因此，我们可以预见，在不久的将来，自然语言处理将成为一门像物理学一样重要的学科，物理学研究物质世界的规律，自然语言处理研究 Web 这个虚拟世界的规律，一实一虚，各有分工。

我们是普通的凡人，我们的眼光当然是比较短浅的，这样的预见也许有很大的局限性，一些专家可能不以为然。不过，当代计算机科学的奠基人 Alan Turing 曾经说过：“尽管我们只能往前看很短的距离，但是我们能看清楚什么是需要做的事情（“We can only see a short distance ahead, but we can see plenty there that needs to be done.”）。”因此，尽管我们这些凡人的眼光比较短浅，但也许还可能看清楚我们现在需要做的事情究竟是什么。

目前自然语言处理正处于激动人心的时刻。普通计算机用户可以使用的计算资源正以惊人的速度迅速增长，Web 已经成为了无比丰富的信息资源，无线移动通信日益普及和发展，这些都使得自然语言处理的应用成为当前科学技术的热门话题。我们相信，知识日新月异的增长和 Web 技术突飞猛进的进步一定会把自然语言处理的研究推向一个崭新的阶段。自然语言处理有可能成为当代语言学中最有发展潜力的学科，它已经给有着悠久传统的古老的语言学注入了新的生命力，在它的推动下，语言学有可能真正成为当代科学百花园中的一门名副其实的领先学科。

我研究自然语言处理已经五十多年的时间了。五十多年前，我还是一个不谙世事的十九岁的小青年，现在，我已经是白发苍苍的老人了。我们这一代人正在一天天地变老，然而，我们如痴如醉地钟爱着的自然语言处理事业却是一个新兴的学科，她还非常年轻，充满了青春的活力，尽管她还很不成熟，但是她无疑有着光辉的发展前景。我们个人的生命是有限的，而科学知识的探讨和研究却是无限的。我们个人渺小的生命与科学事业这棵常青的参天大树相比较，显得是那么地微不足道，有如沧海之一粟。想到这些，怎不感慨万千！

“路漫漫其修远兮，吾将上下而求索”，自然语言处理的探索者任重道远，不论在理论方面还是在应用方面，我们都需要加倍地努力。当前自然语言处理的研究和应用仍然面临诸多困难，我们还要继续奋战，才能度过难关，走向一马平川的坦途。

田艳原来是从事英语教学的语言学者，她刚加入自然语言处理的队伍，就在英译汉的在线自动评分方面做了如此出色的工作。我抽空为她的专著写了这篇序言，作为对她的热烈祝贺。

冯志伟

2013 年春于杭州下沙

前　言

随着现代信息技术在教育领域的广泛应用，外语学习方式和测试方式正在发生着根本性的变化。以网络课程和网上测试为主要形式的人机交互学习和测试就是其中之一。目前，不论国内还是国外，利用网络和计算机技术对客观题进行网上自动评分已经非常普遍，然而利用网络和计算机技术对主观题进行自动评分，因涉及自然语言的计算机处理技术，具有很大的挑战性，所以仍处在不断的探索之中。

英译汉题是主观题的一种，对其进行在线自动评分的探索，不但可以提高学生网上自主练习英译汉的兴趣，还可以通过提供及时的反馈，帮助其提高英译汉技能和英汉语言水平。

英译汉技能不但是我国英语学习者应该掌握的基本语言技能之一，也是我国翻译专业人员必须掌握的翻译技能之一。我国高等学校英语专业和非英语专业英语教学大纲对英译汉技能都有明确的量化要求，各类英语教材也都有对英译汉技能的讲解和练习。国内各类英语考试中广泛采用英译汉题作为对学生英语准确理解能力考核的有效手段之一。

利用网络提供英译汉练习是非常有效的帮助学生通过英译汉实践提高翻译水平的途径，因为学生可以不受时间和地点的限制，根据个人的水平和目的，随时随地进行翻译练习。但目前，不论是英语网络课程里包含的网上英译汉练习，还是网上专门提供的英译汉训练栏目，大多不能针对学生的译文提供及时的反馈，而只提供参考译文，这大大降低了学生网上自主练习英译汉的积极性和兴趣，从而降低了网上提高英译汉水平的效果。因而，从理论和实践上探索如何利用网络和计算机技术对英译汉题进行即时自动评分，可为学生网上练习英译汉提供有力的技术支持，有效拓展其英译汉技能的提高途径，最终实现英译汉技能培训和自测的网络化。

在线英译汉即时自动评分的对象是学生从英语原文翻译成汉语的句子，因此，本书不仅涉及自然语言处理（Natural Language Processing，简称 NLP），尤其是中文信息处理（Chinese Information Processing，简称 CIP），而且还涉及语言学、测试学、翻译学和网络信息技术等。

本书在反复实践的基础上，从理论上探讨了在线英译汉即时自动评