<<大规模考试英汉互译自动评分系统>>

图书基本信息

书名：<<大规模考试英汉互译自动评分系统的研发与应用>>

13位ISBN编号：9787040349047

10位ISBN编号：7040349043

出版时间：2012-7

出版时间：高等教育出版社

作者：秦颖，文秋芳　著

页数：121

字数：150000

内容概要

秦颖、文秋芳编著的《大规模考试英汉互译自动评分系统的研发与应用》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。

书籍目录

章节摘录

版权页： 插图： 第四章 相关语言处理技术 从语言学、翻译学和测试的角度构建翻译自动评分模型是理论研究的内容，最终如何在计算机上实现自动评分系统则是技术实现要探讨的核心：主要涉及语言信息的自动处理，包括文本特征的自动提取、语义分析方法、字符编码以及数学模型的实现、系统用户界面和操作响应等的代码编写问题。
相比理论研究，技术实现篇需要自然语言处理相关技术的支持，更关注算法实现及效率等计算机技术。
本章将对自动评分模型涉及到的有关自然语言处理相关技术予以介绍，并给出部分实现内容的程序代码。
本章编程所用的语言为Perl。
Perl为“实用报表提取语言”（Practical Extraction and Report Language）的缩写。
Perl中有强大的正则表达式，非常适合于语言信息处理。
Perl为开放源代码的免费软件，在Unix和Windows环境下均可运行。
本章所有的程序代码均在Perl v5.8.7版下通过调试。
4.1 文本特征及提取方法 英汉互译评分理论研究中从形式和语义角度分析了与译文质量相关的文本特征，表3-9和3-17分别为汉译英和英译汉模型中尝试运用的文本特征。
形式特征分为字词层面、句子层面和篇章层面三大类，又各自包含若干小类；语义特征包括n-gram，语义点、基于潜在语义分析的相似度、词或多词单位对齐等特征。
选拔性评分模型用于大规模翻译考试译文的自动评分，要求对于不同质量的译文能够较好地区分。
为提高评分速度，选拔性评分模型主要包含语义参数。
研究表明该简化模型仍然和人工评分有较高的相关性。
下面分别阐述这些特征的自动提取方法。
4.1.1 形式特征的提取 1形式参数提取前的文本预处理 预处理工作主要包括整理收集的实际语料中不规范的文本撰写内容和格式，为信息提取做必须的标注和加工等。
对于英文译文，首先去除非英文字符，如汉语标点符号；将全角的字符转换为半角；将词之间的多个空格替换为一个空格；字母全部统一为小写或大写形式等。
另外，原始的文本中没有词性信息，为获得词性分布的文本特征，就需要对所有译文（参考译文、训练译文和测试译文）做词性标注处理。
由于时间有限，我们使用了英文词性自动标注工具Gotagger进行词性标注。
英文词性标注软件较多，常用的还有tagtree、standford parser等。
但不同的软件词性标注集有差异，词性划分方法不同。
汉语译文的预处理工作更多一些。
汉语文本没有明显的词的界限，对于汉语的信息处理可基于两种语言单位——字或词进行。
很多研究表明，基于词的汉语信息提取性能优于基于字的信息提取。
因此，我们对汉语译文的处理大多以词为单位。
预处理时将所有汉语译文都预先进行了切词，并进行了词性标注。
标注的词性为北大计算语言所1997年版《现代汉语语法信息词典》中词性集。
2词汇级形式参数的提取 词汇级和译文质量相关的特征十分丰富，包括词汇多样性特征、词频广度、词汇难度、词性分布等。
词汇多样陛从类符数和类符形符比两个角度考察。
类符数指译文中不同词的数目，形符数即单词数（不包括标点符号）。
模型实际使用的参数是测试译文和参考译文的平均类符数之差和类符形符比之差，以便更合理地判断译文的词汇多样性特征。
形符类符的提取方法：预处理后的英文和汉语，词与词之间（标点与词之间）均为空格隔开，所以根据空格区分各个词。
再根据词性分隔符得到词和词性两部分，前一部分为形符或者标点符号。

对于英文形符提取可用正则表达式为/\w（＼w-＼'）*／实现，意义表示：由字母开头，后面接一个或多个字母数字及下划线和＇组成的。
汉语由于所有标点的词性标记为/w，因此凡是词性为/w的均不视为形符。
类符数就是将形符中相同的词合并后的数目。
类符形符比=类符数的平方／形符数。

编辑推荐

《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》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。

版权说明