学校编码：10384

学号：31520111153184

厦门大学

硕士学位论文

基于云翻译平台的协同翻译工具研究

Research on Collaborative Translation Tools based on Cloud Translation Platform

何钟豪

指导教师：史晓东

专业名称：模式识别与智能系统

答辩日期：2014年5月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　月　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文(包括纸质版和电子版)，允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于:

(　　)1.经厦门大学保密委员会审查核定的保密学位论文，于　年　月　日解密，解密后适用上述授权。

(　　)2.不保密，适用上述授权。

(请在以上相应括号内打“✓”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。)

声明人(签名):

年　　月　　日

摘　要

随着全球交流的越来越广泛以及信息科技的快速发展，不同国家不同种族之间的语言交流障碍问题日益明显，随着不同语言之间的沟通越来越重要，翻译行业也得到了迅速发展。目前的翻译市场还是以人工翻译为主，单纯的人工翻译，并且以个体为单位进行翻译工作虽然可以充分展现译员水平，但是效率低、成本高，所以现在一般采用计算机辅助人工翻译的方式。云翻译平台 是基于云平台的协同翻译平台，平台不仅融合了先进的计算技术、语言处理技术，并且能够实现协同翻译，即将多名空间上分散的翻译人员组织起来共同完成一项翻译任务，以提高整个翻译行业的效率。

本文主要研究了如何在基于云翻译平台的协同翻译工具中加入项目组管理，从而更好地协调不同译员和翻译项目之间的关系，并将辅助翻译输入法作为翻译人员和协同翻译平台之间的桥梁，在不同翻译人员之间即时共享翻译信息，有效提高译员的翻译效率。同时通过协同翻译中的术语检测与识别，以及翻译记忆的实现，在提高译员的翻译效率的同时，还可以有效解决翻译内容重复，翻译成员之间各自为战，对专业术语翻译不统一的问题。本文的创新点如下：

（1）对于协同翻译中的术语检测与识别系统，我们将识别短语看成是一个分类问题，并同时引入了集成学习的方法，充分考虑了文档中存在的专业术语和非专业术语数量不平衡的情况，通过融合多个分类器的分类结果，减小了分类误差，显著提高了术语识别的准确性，对于提高翻译小组的翻译一致性提供了有效的帮助。

（2）在协同翻译系统项目组管理的功能中，将辅助翻译输入法作为协调不同译员之间以及译员与翻译项目之间关系的工具和桥梁，通过输入法更好地达到翻译一致性的目的，同时减轻翻译人员在使用计算机进行翻译时的工作量，节省翻译时间，提高翻译效率。

关键词：协同翻译；术语识别；项目组管理

Abstract

With the increasingly popular of global communication and the rapid development of information science and technology, the language communication obstacle problem between different race and different state is increasingly obvious, as the communication between different languages become more and more important, the translation industry also got developed quickly. The current translation market is still human-based translation, although human translations can fully be up to the standard of translator, but it is inefficiency and high cost, so computer-aided translations is commonly used. Cloud translation platform is a collaborative translation platform based on the cloud platform , applies advanced computer and language processing technology to translation sector, and can realize the collaborative translation functions, organizing multiple spatially distributed users to complete one translation task, in order to improve the efficiency of the whole translation industry.

This paper mainly studied how to inject the project team management into the collaborative translation tools based on the cloud platform, so as to better coordinate the relationship between different translators and translation project, and the aided translation input method will be introduced to collaborative translation platform, as a bridge between the translator and the collaborative translation platform, to effectively improve the efficiency of different translators. At the same time we realize the terms detection and recognition system and translation memory system, to not only improve the efficiency of different translators, but also can solve the problem that translation content repetition and lack of communication between different translators. In this paper, the main work and innovation involved as following:

(1) For terms detection and recognition system, we treat terms identification problem as a classification problem, and introduce the method of ensemble

learning at the same time, give full consideration to the situation that professional terms and general phrases distribution imbalance, through integrate the classification results of the multiple classifier, to minimize the classification error, significantly improve the accuracy of the terminology recognition, and can improving the consistency among the translation group.

(2) We introduce an aided translation input method into the project team management function of the collaborative translation tools based on the cloud platform, to better coordinate the relationship between the different translators and translation projects. At the same time, to make translators work less and create more, saving time and improving the efficiency of translation.

Keywords: Collaborative Translation; Terms Recognition; Project Team Management

参考资料

[1] 叶娜,张桂平,韩亚冬,蔡东风. 从计算机辅助翻译到协同翻译[J]. 中文信息学报,2012,26（6）：1-10.
[2] 韩亚冬. 协同环境下基于模板的机器翻译技术的研究[D]. 沈阳：沈阳航空航天大学,2011.
[3] 何鸿君,吴泉源,罗莉. 协同编辑中维护操作意愿的文档标注方法[J]. 软件学报,1999,10（02）：160-164.
[4] 周明骏,徐礼爽,田丰,戴国忠. 协作笔式用户界面开发工具研究[J]. 软件学报,2008,19（10）：2780-2788.
[5] 王潜平,林宗楷,郭玉钗. 支持协同设计的工程数据库版本管理[J]. 软件学报, 1996,7（11）：691-697.
[6] Pax Humana: Translation of Various Humanitarian Reports in French, English, German, Spanish[EB/OL]. http://paxhumana.info.
[7] Sayori Shimohata,Mihoko Kitamura,Tatsuya Sukehiro. Collaborative Translation Environment on the Web[C]//Proceedings of the MT Summit VIII,2001：331-334.
[8] 王建德,陈肇雄,黄河燕. 基于协同机制的多用户交互翻译系统的设计与实现[J]. 中文信息学报,2007,4（10）：34-43.
[9] Bey Y,Kageura K,Boitet C. A Framework for Data Management for the Online Volunteer Translators"Aid System QRLex[C]//Proceedings of the Pacific Asia Conference on Language,Information and computation：51-60.
[10] Youce Bey,Christian Boiter,Kyo Kageura. The TRANSBey Prototype: An Online Collaborative Wiki-Based CAT Environment for Volunteer Translation[EB/OL]. http://paxhumana.info.
[11] 张桂平,蔡东风. 基于知识管理和智能控制的协同翻译平台——知识管理和机器翻译的融合[J]. 中文信息学报,2008,22（5）：3-11.
[12] Daisuke Morita,Toru Ishida. Designing Protocols for Collaborative Translation [C]//Proceedings of the International Conference on Principles of Practice in Multi-Agent Systems,Berlin,Heidelberg,Springer：61-70.
[13] Toshiki Murata,Mihoko Kitamura,Tsuyoshi Fukui,Tatsuya Sukehiro. Implementation of Collaborative Translation Environment‘ Yakushite Net ’[EB/OL]. http://paxhumana.info.
[14] 方瑞玉. 基于输入法的辅助翻译工具研究和实现[D]. 厦门：厦门大学,2013.
[15] 冯志伟. 现代术语学引论[M]. 北京：语文出版社,1997.
[16] Chen Wenliang,Zhu Jingbo,Yao Tianshun. Automatic Learning Field Words by Bootstrapping [C]//Proceedings of the JSCL 2003.
[17] Luhn,H. P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information [J]. IBM Journal of Research and Development,1957,1(4)：159-165.
[18] John S. Justeson,Slava M. Katz. Technical terminology: Some linguists properties and an algorithms for identification in text[J]. Natural Language Engineering,1995 (1)：9-27.
[19] Frantzi,K. T.,Ananiadou. Statistical Measures for Terminological Extraction[C]//Proceedings of the International Conference on Statistical Analysis of Textual Data,1995：297-308.
[20] Damerau,F. J. Evaluating Domain-Oriented Multi-Word Terms from Texts[J]. Information Processsing and Management,1993,29(4)：433-447.
[21] Cohen. J. D. Highlights: Language and Domain-Independent Automatic Indexing Terms for Abstracting[J]. Journal of the American Society for Information Science,1995,46(3)：162-174.
[22] Patrick Pantel,Dekang Lin. A statistical Corpus-Based Term Extractor[C] //Proceedings of Canadian Conference on AI,2001：36-46.
[23] Luo Zhiyong. An Integrated Method for Chinese Unknown Word Extraction[J]. Proceedings of the ACL2004.
[24] 叶志飞,文益民,吕宝粮. 不平衡分类问题研究综述[J]. 智能系统学报,2009,4（2）：148-156.
[25] 杨明,尹军梅,吉根林. 不平衡数据分类方法综述[J]. 南京师范大学学报,2008,8（4）：7-12.

[26] Shoushan Li,Guodong Zhou,Zhongqing Wang. Imbalanced Sentiment Classification[C]//Proceeding of the 20th ACM international conference on information and knowledge mangement,UK,2011：2469-2472.
[27] Holte R C,AckerL E,PorterBW. Concept learning and the problem of small disjuncts[C]//Proceedings of the 11th International Joint Conference on Artificial Intelligence,Austin: Morgan Kaufmann, 1989：813-818.
[28] WEISS GM. Mining with rarity: a unifying framework[J]. Sigkdd Explorations,2004,6 (1) ：7219.
[29] Sun YM,Kamel M S,Wong A K C. Cost-sensitive boosting for classification of imbalance data[J]. Pattern Recognition,2007,40：3358-3378.
[30] 钱洪波,贺广南. 非平衡类数据分类概述[J]. 计算机工程与科学,2010,32（5）：85-88.
[31] Provost F,Fawcett T. Robust classification for imprecise environments[J]. Machine Learning,2001,42(3)：203-231.
[32] Drummond C,Holte R C. class imbalance, and cost sensitivity: why under-sampling beats over-sampling[C]//Proceedings of International Conference on Machine Learning,Washington DC,2003：152-154.
[33] Chawla.N.V,Bowyer.K.W,Hall.L.O. SMOTE: synthetic minority over-samp ling technique[J]. Journal of Artificial Intelligence Research,2002,16：321-357.
[34] Kubatm,Matw IN S. Addressing the curse of imbalanced training sets: one-sided selection[C]//Proceedings of the International Conference on Machine Learning,San Francisco: Morgan Kaufmann,1997： 179-186.
[35] Pazzan IM,Merz C,Murphy P.Reducing misclassification costs[C] //Proceedings of the International Conference on Machine Learning. San Francisco,CA,USA,1994：217-225.
[36] Tom M. Mitchell: Machine Learning[M],McGraw Hill,1997.
[37] Sollich P,Krogh A. Learning with ensembles: how over-fitting can be useful[C] //Proceedings of Advances in Neural Information Processing Systems 8,Cambridge,MA: MIT Press,1996：190-196.
[38] Opitz D,Maclin R. Popular ensemble methods: an empirical study[C] //Proceedings of Journal of Artificial Intelligence Research,1999,11：169-198.
[39] Feature Selection and Classifier Ensembles: A Study on Hyperspectral Remote Sensing Data[EB/OL],http://143.129.203.3/visielab/theses/shixin/ thesis_yu.pdf,2003.
[40] Robert E,Schapire. The Strength of Weak Learnability[C]. Machine Learning,1990,5(2)：197-227.
[41] Breiman L. Bagging predicators. Machine Learning[J],1996,24(2)：123-140.
[42] Schapire R E,Singer Y. Improved Boosting Algorithms Using Confidence-Rated Predictions[J]. Machine Learning,1999,37(3)： 297-336.
[43] Friedman J H,Hastie T,Tibshirani R. Additive Logistic Regression: A Statistical View of Boosting[J]. Annals of Statistics,2000,28(2)：337-374.
[44] Breiman L. Arcing Classifiers[J]. Annals of Statisties,1998,26(3)：801-849.
[45] Freund Y. An Adaptive Version of the Boost by Majority Algorithm[J]. Machine Learning,2001, 43(3)：293-318.
[46] Alexey T.,Mykola P.,Padraig C. Sequential Genetic Search for Ensemble Feature Selection[C]//Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence,2005.
[47] T.G. Dietterich. Ensemble Methods in Machine Learning[J]. In Multiple Classier Systems,Cagliari,Italy,2000.
[48] Hiroshi Mamitsuka. "Empirical Evaluation of Ensemble Feature Subset Selection Methods for Learning from a High-Dimensional Database in Drug Design"[C] //Proceedings of IEEE Symposium on BioInformatics and BioEngineering,2003.
[49] Thomas G. Dietterich,Ghulum Bakiri. Solving Multiclass Learning Problems via Error-Correcting Output Codes[C]//Proceedings of Journal of Artificial Intelligence Research,1995：263-286.
[50] Dietterich, T. G. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization[J]. Machine Learning,40,139-157.
[51] Xu, L., et al, Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition[C]//Proceedings of IEEE Transactions on Systems,Man and Cybernetics 1992.
[52] Verlinde P,Chillet G. Comparing decision trees fusion paradigms using K-NN based classifiers, decision

trees and logistic regression in a multi- modal identity verification application[C]//Proceedings of International Conference on Audio and Video-Based Biometric Person Authentication,1999,188-193.
[53] 熊德意. 基于括号转录语法和依存语法的统计机器翻译研究[D]. 北京：中国科学院计算技术研究所,2007.
[54] 最大熵工具包[EB/OJ]：http://ir.hit.edu.cn/~taozi/ME.htm.
[55] 王传英,闫栗丽,张颖丽. 翻译项目管理与职业译员训练[J]. 译业论文,2011,（1）：55-59.
[56] 邵敏. 专业翻译项目团队建设初探[D]. 上海：上海外国语大学,2011.
[57] 陈懿. 多译者参与翻译项目中译文风格统一性的探讨[J]. 文化与教育技术,2010,（22）：242.
[58] 吉胜军. 基于Levenshtein distance算法的句子相似度计算[J]. 电脑知识与技术,2009,5（9）：2177-2178.
[59] 殷耀明,张东站. 基于关系向量模型的句子相似度计算[J]. 计算机工程与应用,2014,50（2）：198-203.
[60] 王惠敏. 浅析翻译项目中的质量管理[J]. 长江大学学报,2012,35（10）:101-102.
[61] Resnick P,Varian H R.Recommender systems[J]. Communications of the ACM, 1997,40(3)：56-58.