doi:10.3772/j.issn.2095-915x.2017.03.011

双语句对翻译众包辅助平台设计与实现

1. 中国科学技术信息研究所 北京 100038；
2. 北京市科学技术情报研究所 北京 100044；
3. 富媒体数字出版内容组织与知识服务重点实验室 北京 100038

高影繁[1,3] 李辉[2] 徐红姣[1,3] 崔笛[1,3]

摘 要 本文提出了一种采用众包工作模式的科技领域日汉机器翻译辅助平台的构建方法。在充分调研众包生产模式、质量控制等研究和实践的基础上，设计了集用户管理、团队管理、语料管理、机器辅助翻译、术语辅助翻译等功能为一体的双语句对生产平台，针对不同角色和不同技术类别分别构建出相应的功能模块。该平台在众包工作模式的基础上结合了多源信息辅助译者完成翻译，翻译效率高且翻译成本低，平台的开发和运行为科技领域实用型日汉机器翻译系统的建设提供了有力支撑。

关键词： 众包，双语语料建设，机器翻译辅助平台
中图分类号： G355

Design and Implementation of Bilingual Sentence Pairs Translation Crowdsourcing Aided Platform

1. Institute of Scientific and Technical Information of China, Beijing 100038, China;
2. Beijing Institute of Science and Technology Information, Beijing 100044, China;
3. Key Laboratory of Rich-media Knowledge Organization and Service of Diqital Publishing Content, SAPPRFT, Beijing 100038, China

GAO YingFan[1,3] LI Hui[2] XU HongJiao[1,3] CUI Di[1,3]

Abstract This paper presented a method for designing and implementing bilingual sentence pairs translation aided platform based on crowdsourcing. On basis of the research and practice of crowdsourcing production mode and quality control, we designed a bilingual sentence production platform which integrates

基金项目： 本文受中国科学技术信息研究所重点工作项目（ZD2015-6），北京市财政项目（PXM2016-178214-000006）的资助。
作者简介： 高影繁（1974-），博士，副研究员，研究方向：多语言信息处理，知识组织，Email：gaoyingf@istic.ac.cn；李辉（1975-），硕士，副研究员，研究方向：科技情报，危机管理，Email：lih@bjstinfo.com.cn；徐红姣（1985-），硕士，助理研究员，研究方向：跨语言信息检索，Email: xuhj@istic.ac.cn； 崔笛（1993-），硕士，研究方向：竞争情报。

the functions of user management, team management, corpus management, machine aided translation and terminology translation. The corresponding functional modules were constructed for different roles and different technical categories of users of this platform. Using crowdsourcing translation model, we collected multi source information to assist the translator to complete the translation. The results indicated that this method has high efficiency and low cost. The development and operation of the platform provided strong support for the construction of a practical Japanese-Chinese machine translation system.

Keywords：Crowdsourcing, bilingual corpus construction, machine translation aided platform

1 引言

在经济全球化时代，多语言和跨语言信息交流需求日益增多。随着自然语言处理技术的发展，使用计算机帮助人们克服语言障碍、实现跨语言沟通的梦想正逐步成为现实。经过半个多世纪的研究和探索，新闻领域的机器翻译和跨语言信息检索服务系统几乎已经接近实用化，很多在线服务已经面世。丰富的新闻领域语料资源在机器翻译研究中发挥了重要作用，但在科技领域，由于高质量语料缺乏，特别是作为核心语言资源的双语平行句对缺乏，使得机器翻译的质量受到了很大影响。本文研究的应用背景是科技领域的日汉机器翻译[1,2]，目前主流的日汉机器翻译系统基本是在新闻类或对话类双语句对基础上训练出来的翻译模型，在科技文献翻译效果不佳。从双语语料建设[3,4]方法上来说，现有的双语语料主要以人工构建为主，代价昂贵，难以满足迅速增长的日汉科技信息翻译处理需求。因此，有针对性地建设科技领域的双语句对语料库，突破实用型日汉科技领域机器翻译系统的技术瓶颈，无论对于自然语言处理技术的进步，还是深入全面地掌控和跟踪日本的科学技术发展动态，都具有重要的价值。

2 众包的概念和应用

Crowdsourcing 一词由美国计算机杂志《连线》(Wired) 的记者 Howe 在题为《众包的兴起》一文中第一次使用，国内翻译成“众包”。杰夫·豪在文中这样描述众包：“从产品设计软件到数码摄像机，这些无处不在的技术进步打破了原先设在专业人士和业余人士之间的成本壁垒。原先只是兴趣爱好或临时为之或业余尝试的人，现在一下子找到了市场，…… 他们并非总是免费，但却比传统的雇员花费要少得多。这不是外包，而是众包。”众包就是“网络社会的社会化生产”。众包不同于外包就在于外包强调的是专业化分工，它的对象是专业机构和专业人士；而众包的发包对象是业余爱好者，他们社会差异大且背景多样化。在强调创新的今天，众包的优势明显：成本低、调动了潜在的生产资源、提高了生产效率、还能满足用户个性化需求[5]。

从众包的工作模式上来看，一般可将其分为四种：（1）奖励解决问题模式。基于经济奖励制度的方式让大众为你解决实际难题。（2）用户制作内容模式。将制作的工具交给用户，网站在幕后控制，同时紧盯市场，并

doi:10.3772/j.issn.2095-915x.2017.03.011

及时投放网络广告。（3）免费协作的模式。以维基百科为代表，其内容完全免费，且全部由志愿者在统一的开放性平台上完成。（4）即时的众包模式。Foodpickle 通过与即时众包模式的代表——微博共同协作，构建了即时的众包模式问答平台。从众包本身的特点来看，则具有用时短、花费少、可广纳人才、及时洞察客户需求的优势[6]，例如：在机器翻译领域，若采用人工的方式对机器翻译结果进行评估，不仅费时费力，评估代价也会非常高，而众包工作模式则能有效避免这种苦恼。正是由于众包模式的这种优势，使得它在越来越多的领域得以发展和普及，用户群体和工作模式也逐渐变得多样化。当然，这种具有较强随机性的用户群体也给众包模式带来诸多问题，很多工作者由于受利益驱使而随意给出答案，导致众包结果的质量难以控制。随着众包技术的广泛应用，众包的质量控制问题变得越来越重要，众多学者也针对这一问题展开了一系列研究。

2008 年，Sorokin 等做了一项图像标注的众包应用实验，分析赏金刺激对结果质量所产生的影响，发现奖励额度和结果质量间有很强的依赖现象[7]。同年，Kittur 等人发现众包任务的构造方式对结果有较大的影响，其中，明确可验证型问题对众包应用质量的控制起到关键作用[8]。2011 年，Eickhoff 等人通过分别对众包任务类型、用户接口形式以及工作者类型三个方面进行相关实验发现：1）一项复杂度较高且需要创造力的任务对只想简单快速完成任务的工作者缺乏吸引力；2）上下文情景改变的任务对恶意工作者吸引力不大，但对欺骗性、自动标注的答案有较好的抵抗性；3）只接受发达国家工作者承接的众包任务通常能获得较高质量的结果，同时也要付出时间的代价[9]。

针对众包任务的质量问题，目前主要通过两种途径来提高众包质量：一方面是使用黄金标准数据（Golden standard data）评估工作者完成的质量，以识别欺骗类型工作者，达到拒绝恶意工作者完成结果的目的；另一方面通过调整设计任务的方式，减少对恶意工作者的吸引力，以提高众包工作者的整体素质。

3 基于众包的语料构建方法

本研究采用基于众包平台的协同翻译方法，利用自然语言处理技术和机器学习方法自动获取日汉平行语料和日汉双语词典，将来源于日本 JST 提供的生物、化学领域论文摘要的日语文献翻译为中文语料，旨在尽量减少人工参与的构建代价，在保证语料质量的基础上，提高双语语料建设的速度。

3.1 众包翻译项目管理的一般流程及关键问题

众包翻译项目管理和一般的翻译项目管理具有基本的共性，但又有其自身的特性。两者都要经历基本的五大阶段：项目启动——项目计划——项目执行——项目监控——项目收尾。简单来说就是都有“译前、译中、译后”这样一个过程。但是，在众包模式下，翻译项目管理在各个阶段又有其自身的特点。

（1）在项目启动和计划阶段，除了进行一

般的分析、评估和获取之外，管理者还要分析项目进行众包的可行性、项目的保密性、译员队伍的组建、项目众包的成本、原文的分割、译文质量的控制等等问题。

（2）在项目执行阶段，要制定更为详细的计划和更为规范的统一执行标准。

（3）在项目监控阶段，要做的工作更为繁杂，要求也更为严格。

众包翻译模式作为一种新的翻译模式，具有一些独特的特点。爱尔兰学者 Dimitra 认为众包模式有三个关键成功要素，即 Quality(质量)、Control(控制)、Motivation(激励)[10]，质量主要指翻译的质量水平，控制反映翻译的安全和译者管理，激励指如何激励众包参与者的问题。对于众包模式的核心要素可以从平台建设、译者和质量控制等三个方面构建。通过平台建设完善流程控制，通过译者管理有效激发参与度，通过翻译质量控制保证众包的翻译水平。比如，在项目执行前，建立供管理者与译员、译员与译员间进行沟通的平台，以制定统一的执行标准，也是保证翻译质量的关键；同时，由于平台系统的开放性，译者的翻译时限相对灵活，因此，建立完善的翻译进度可视化平台，方便译者把握和管理翻译任务的进度，实现流程的可视化，也是十分必要的。

3.2 基于众包的双语句对翻译辅助平台设计

3.2.1 平台设计架构

本平台的架构设计图如图 1。一共包含 3 个模块：译员集合模块、译员交流模块和翻译辅助信息模块。译员集合模块负责通过众包网络发布翻译需求信息，经过译员过滤，确定众包译员，在网络交流平台上进行联络，包括 qq、微信、翻译论坛等。译员交流模块负责任务的分发，可视化进度展示，并进行翻译结果的回溯，以及翻译质量的控制。输入待翻译的句子之后，翻译辅助信息模块提供四种翻译辅助信息，术语识别，句子匹配，机器翻译和互联网翻译。

众包翻译平台的基本目标是：允许任何组织和个人通过本平台在任意时间、任意地点完成“日汉句对”翻译任务。本平台的功能依据角色进行划分，包括平台管理员、译员和审核专家三种，平台的工作流程如图 2、图 3 所示：

3.2.2 按角色划分的平台功能描述

众包辅助翻译平台中按照三种角色划分的平台功能细节如下：

（1）平台管理员具有用户管理功能、评价统计功能、数据管理功能、任务管理功能。

用户管理功能包括：

1）平台用户的增加、修改、删除功能：平台的用户按照所属单位进行管理，每个所属单位包括若干普通译员和审核专家，平台管理员具有对这些用户的增添、修改和删除权限。对于没有所属单位的译员，系统也要自动生成一个唯一的所属单位流水号。

2）基于系统评价的用户处理结果发布：在系统评价和审核专家评价基础上，生成对各普通译员的工作能力评价，初步分为“审核通过”和“评审不通过”两个级别，将此评价结果反馈给普通译员。

doi:10.3772/j.issn.2095-915x.2017.03.011

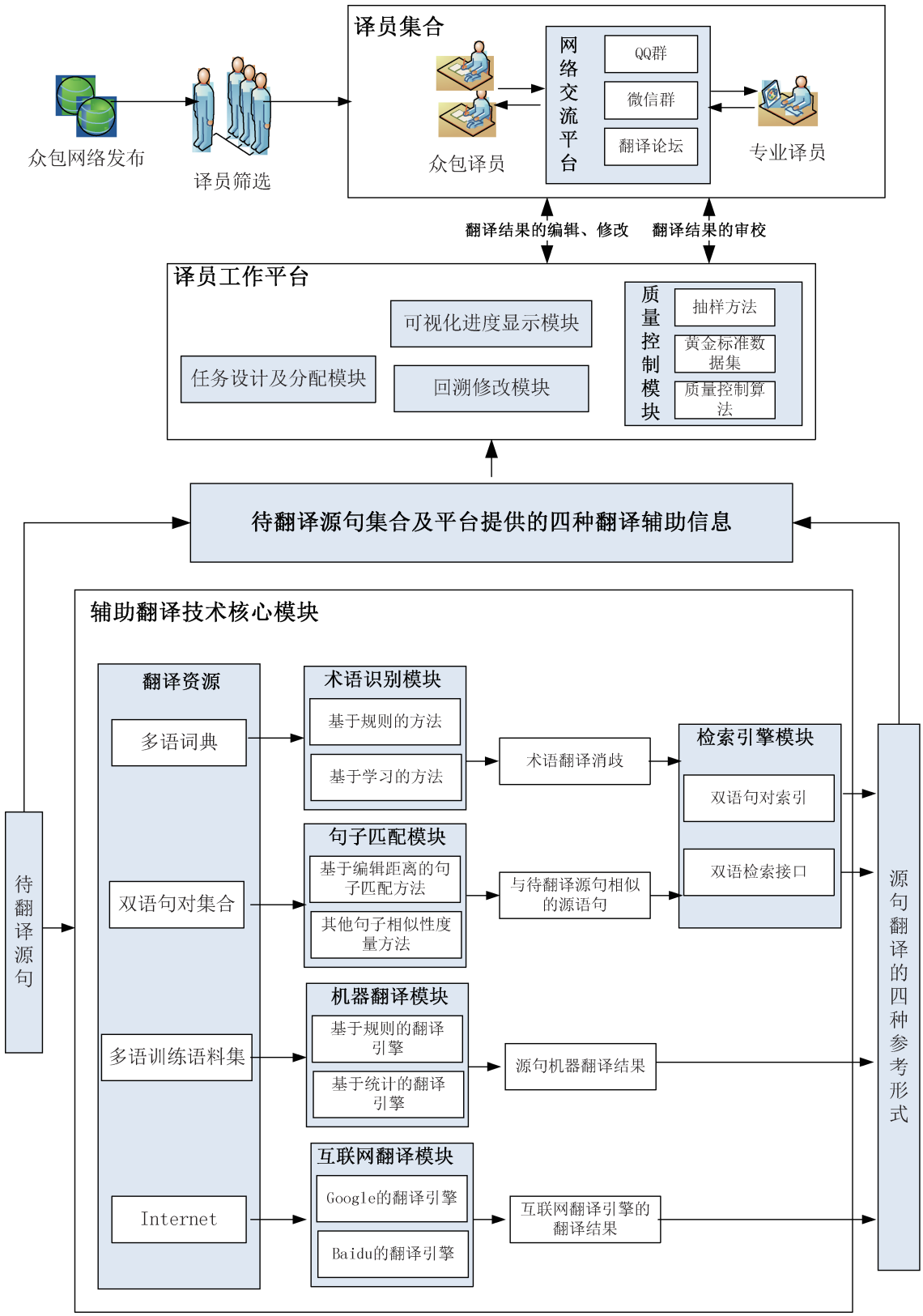

图1 平台架构设计图

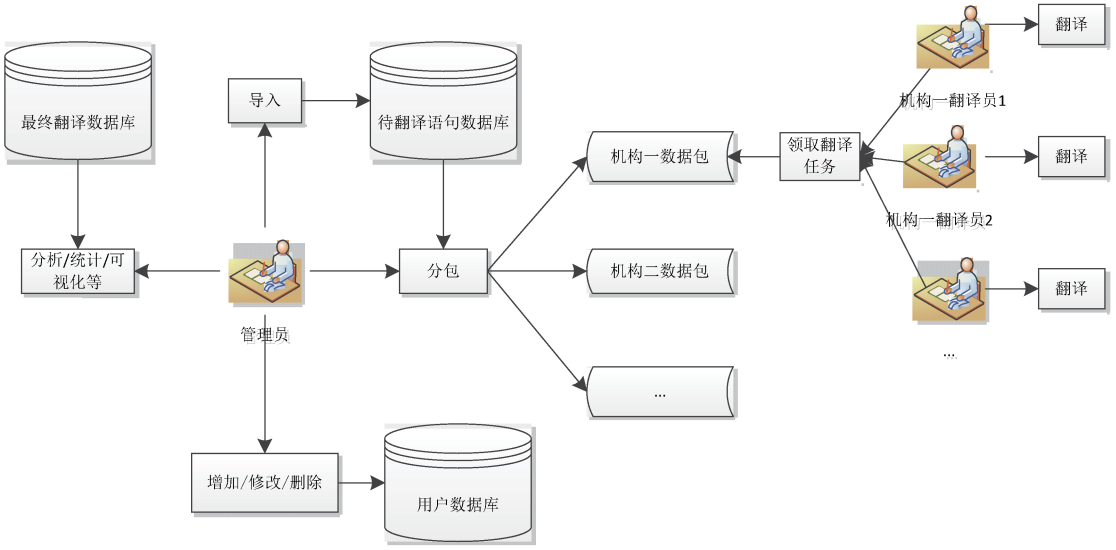

图2 平台管理员和翻译员的工作流程图

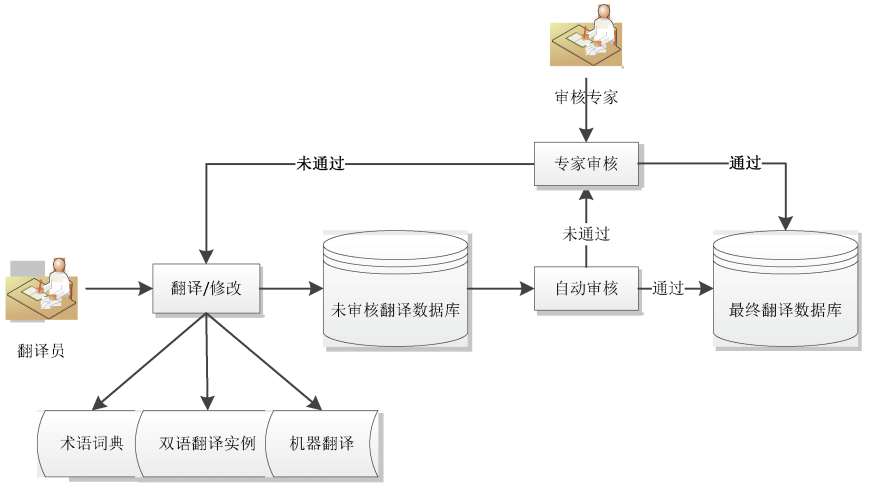

图3 按照角色划分的平台工作流程图

3）用户申诉受理和申诉结果反馈：针对普通译员在收到工作能力评价结果后，对有异议的评价结果具有申诉权。用户管理模块收到申述后，可通过审核专家信息发布平台进行申诉提交，审核专家组对申诉进行重新审核后给出申诉反馈意见，将此反馈结果返回申诉译员。

评价统计汇总功能包括：

1）按照所属单位进行工作结果显示：包括工作进度和工作质量评价两个方面。工作进度显示承接单位全部译员的工作数量，有普通译员的翻译数量和专职译员的审核数量两种；工作质量评价按照普通译员和审核专家来分，所有普通译员的质量评价指标主要有三个：审核通过量，通过率，历史评测指标的均值；所有

doi:10.3772/j.issn.2095-915x.2017.03.011

审核专家的质量评价指标主要是被申诉的数量。

2）按照译员进行工作结果展示，译员包括普通译员和审核专家两种，因此展示方式也分两种：第一种显示普通译员的翻译工作进度和工作结果的质量评价。工作进度显示该部译员的工作数量；工作质量评价的质量评价包括：审核通过量，通过率，历史评测指标的均值等三个指标。第二种显示审核专家翻译审核情况，主要包括：审核结果总量和审核结果的被申诉次数。

3）按照任务的工作结果展示，由于众包翻译是以任务拆分的方式发布的，因此不同的任务会被划分到不同的承包单位，各任务的完成情况指标主要有：任务总量，任务通过量，任务通过率三个。

数据管理功能包括：

1）原始数据入库：即完成双语据对的数据库导入。

2）中间数据结果保存：普通译员在翻译时点击“保存”之后的所有结果会被写入数据库；审核专家的所有审核结果在点击“保存”之后被写入数据库。

3）修改记录登记：普通译员在完成一次翻译后进行的回溯修改操作也被写入数据库，该数据的保存位置要与第一次翻译结果的保存位置有所不同，以便随时找回原始操作记录。

4）数据导出：数据库中的全部数据能够实现按照要求的导出功能，这里还要涉及导出形式、导出编码等选择。

任务管理功能主要指众包翻译的对象在进行任务拆解后，如何实现任务的有效管理。主要包括以下功能模块：

1）新建任务

2）任务数据管理

3）任务分配管理

4）主要包括普通译员的任务分配和审核专家的任务分配两个方面。

5）任务完成进度的可视化

6）任务质量评价

7）任务完成情况汇总表

该汇总表作为平台管理员与承接单位责任人进行薪酬支付的依据。

（2）审核专家

作为使用翻译平台的审核专家，平台主要为其提供的功能包括：任务审核、任务打分和任务评价说明。

（3）普通译员

作为使用翻译平台的平台译员，平台主要为其提供的功能包括：领取任务包、完成任务包、任务完成进度可视化和任务完成质量可视化。

3.2.3 按技术模块划分的平台功能描述

本众包翻译平台的特色之一，就是为译员提供尽可能多的翻译辅助信息，以帮助他们高效地进行双语句对的翻译。

（1）机器翻译结果导入模块

本平台的全部待翻译句子，都会事先经过机器翻译系统的翻译，平台在导入待翻译句子集合时，可以同时导入机器翻译系统翻译后的句子，即机器翻译模块不是平台的一个部分，只是留出与机器翻译系统翻译结果的导入接口即可。

（2）查词典模块

本平台需要留出翻译词典导入接口，同时，

对于待翻译句子中包含的词典中的词，要能够给出翻译结果。这里的要求是，要查询出词典中包含的最长的词的翻译结果，而不是仅能提供短词的翻译。比如：“这是一个信息检索系统”这句话，翻译词典中有“信息”、“检索”、“系统”、“信息检索”、“检索系统”、“信息检索系统”等词及其翻译，则我们要的结果一定是最长的那个，即“信息检索系统”。

（3）翻译结果的网络抓取模块

百度、Google 等均能提供面向公众的翻译引擎，这些翻译引擎通常都有 API 接口，本平台能够通过这些开放 API 实现待翻译句子翻译结果的网络抓取。

（4）翻译记忆模块

翻译记忆功能主要是实现双语句对的检索，这里又有两种情况：一是看现有双语句对能否与待翻译句子完全匹配，如果是，则直接将翻译结果取出；二是按照句子相似性，将双语句对中与待翻译句子最相似的句子按照相似度排序，最后取出排在前面的 5 个句对即可。

（5）质量控制模块

这个模块已经有代码（c 语言代码），要求嵌入平台，调用即可。

3.3 基于众包的双语句对翻译辅助平台实现

本文截取了“基于众包的双语句对翻译辅助平台”的部分功能进行展示说明。

（1）平台登陆界面

在浏览器中输入网址：“http:// 服务器 IP 地址 /znxq/”，回车进入平台登录页面。输入管理员用户名和密码，点击登录进入系统。

管理员账号：admin

图4 平台登陆界面

管理员登陆后可以进入系统设置界面修改密码，如图 5 所示：

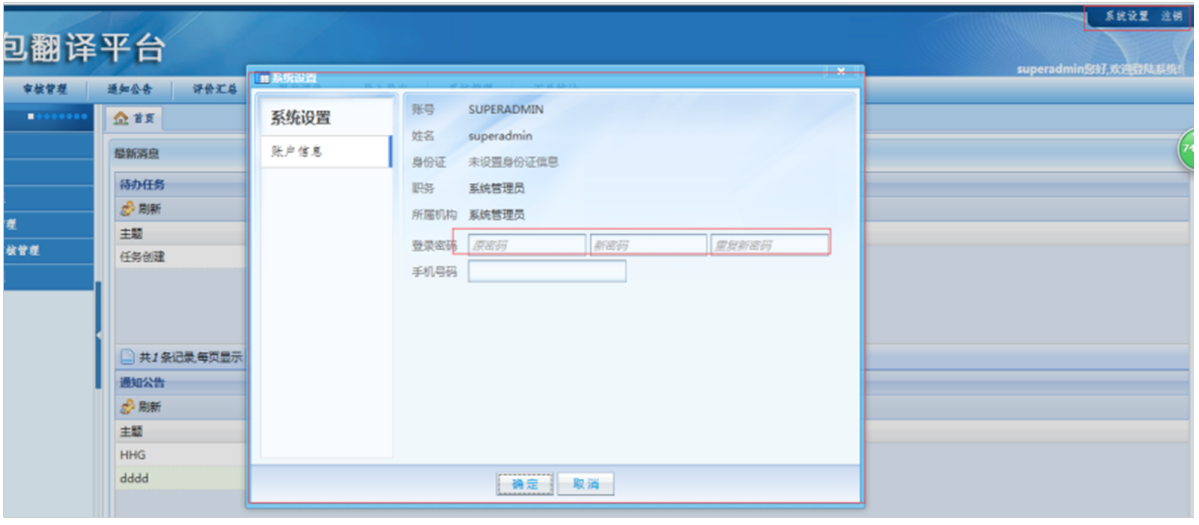

图5 系统设置界面

doi:10.3772/j.issn.2095-915x.2017.03.011

（2）用户管理界面

机构添加完成之后，点击“用户管理”，可以给各级部门添加人员，如图 6 所示。机构 A 翻译部门、机构 B 审核部门、机构 A 办公室添加人员的方式相同，下面我们以机构 A 翻译部门下面添加翻译人员为例。点击“机构 A 翻译部门”：右侧显示当前机构 A 翻译部门下的所有人员，可以继续完成新建用户、查看明细、删除等操作。

图6 用户管理界面

（3）导入导出界面

进入“导入导出”模块，左侧选择“导入数据”，可以完成三种数据的导入，如图 7 所示。

1）导入数据：包含待翻译的数据，系统翻译数据。

2）导入词典：导入词典数据。

3）到入双语句对：到入双语句对数据。

以“导入数据”为例：左侧点击导入数据，右侧为导入页面。首先上传数据，上传数据和任务创建上传附件方式相同，选择你要导入的批次（注：必须按顺序进行添加），选择完成之后点击“导入数据”按钮进行导入，如果数据量比较大，导入时间就会较长，请耐心等待，直至出现“导入成功”。

图7 数据导入界面

展开导出列表，可以进行词典导出、导出待翻译数据和系统翻译、导出待翻译数据和译员翻译（译员翻译后的数据）。

以词典为例：点击词典页面，显示词典数据，点击“导出 EXCEL”直接导出，如图 8 所示。

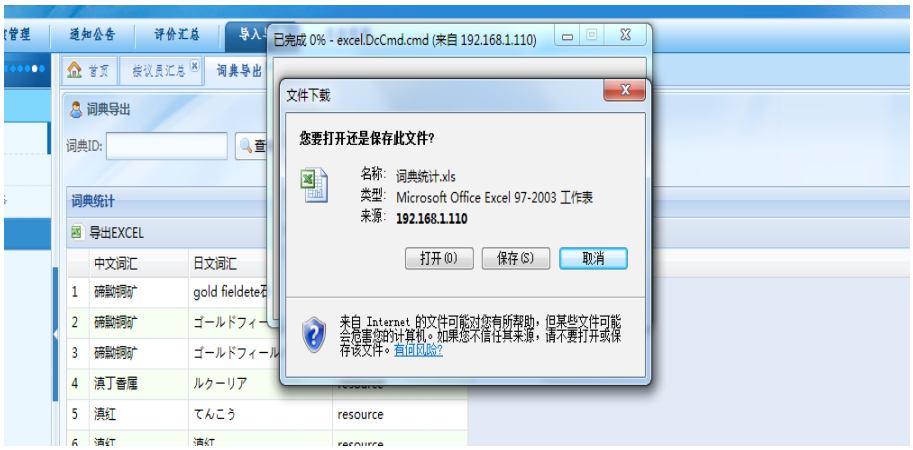

图8 数据导出界面

（5）汇总评价界面

汇总评价包含：按任务结果汇总、按所属单位汇总、按译员汇总。以译员为例，每个统计页面都分为进度和质量统计，有任务数据、加工完成等，选择一条数据可以查看数据统计图图 9（饼状图）。

图9 汇总评价界面

4 结论

本研究采用众包工作模式构建了日汉机器翻译系统双语语料翻译平台。从功能上来看，该平台能够为平台管理员、审核专家和专业译员三种角色分配进行职能划分，并通过机器翻译结果导入模块、查词典模块、翻译结果的网络抓取模块、翻译记忆模块和质量控制模块这五大技术模块的设置，为译员尽可能多地提供翻译辅助信息，以帮助其高效地完成双语句对的翻译工作。将众包工作模式与多源辅助翻译信息的结合，充分发挥了两种模式的优势，以本研究为基础构建出的日汉双语语料生产平台成本低、效率高，为面向科技领域的实用型日汉机器翻译系统的建设提供了重要支撑。在未来研究工作，需要对本研究所构建的平台性能

doi:10.3772/j.issn.2095-915x.2017.03.011

进行反复测试和评估，以期进一步的优化和完善。

参考文献

[1] 张均胜，何彦青，李颖，等 . 中日两国机器翻译研究进展及比较 [J]. 数字图书馆论坛，2011(12): 20-31.

[2] 李颖，朱礼军，张钧胜，等 . “第四届中日韩科技信息机构联合研讨会”概述 —— 面向开放获取、数字标识及实用型汉日双向机器翻译系统 [J]. 数字图书馆论坛，2013(11): 33-45.

[3] 常宝宝，詹卫东，张华瑞 . 面向汉英机器翻译的双语语料库的建设及其管理 [J]. 产品安全与召回，2003(1): 28-31.

[4] 姚树杰 . 面向统计机器翻译的语料处理与评价技术研究 [D]. 沈阳 : 东北大学，2011.

[5] 陆艳 . 网络众包翻译模式研究 [M]. 广东 : 世界图书出版广东有限公司，2014.

[6] Booth T L, Thompson R A. Applying Probability Measures to Abstract Languages[J]. IEEE Transactions on Computers, 1973, C-22(5): 442-450.

[7] Sorokin A, Forsyth D. Utility data annotation with Amazon Mechanical Turk[C]// Computer Vision and Pattern Recognition Workshops, 2008. CVPRW ‘08. IEEE Computer Society Conference. IEEE Xplore, 2008: 1-8.

[8] Kittur A, Chi E H, Suh B. Crowdsourcing User Studies with Mechanical Turk[C]// Conference on Human Factors in Computing Systems, CHI 2008, 2008, Florence, Italy, April. DBLP, 2008: 453-456.

[9] Eickhoff C, Vries A P D. How Crowdsourcable is Your Task?[J]. In Proceedings of the Workshop on Crowdsourcing for Search and Data Mining, CSDM, 2011.

[10] Desilets A, Meer J V D. Co-creating a Repository of Best-practices for Collaborative Translation[J]. Linguistica Antverpiensia, 2011, 10(2): 27-45.