

机器翻译资源建设与翻译引擎优化方法 研究报告

中国科学技术信息研究所

2016-12-29

目 录

1 引言	2
2. 统计机器翻译双语语料抽取方法研究.....	3
2.1 从网络或平行语料抽取双语语料的技术.....	3
2.2 从可比语料中抽取双语语料的技术.....	11
2.3 语料清洗技术.....	13
2.4 多语网站和多语资源的介绍.....	13
3. 基于二维词汇化领域知识的日汉科技术语翻译方法研究.....	15
3.1 相关工作.....	16
3.2 基于领域知识的统计机器翻译训练数据选择.....	18
3.3 实验.....	23
4. 神经网络机器翻译探究	25
4.1 神经机器翻译概述.....	25
4.2 NMT 系统 OpenNMT 测试.....	27
5. 结论	28
5.1 双语语料建设.....	28
5.2 日汉术语翻译.....	29
5.3 神经网络机器翻译.....	29

插图清单

图表 1 : 文档翻译对排名列表的过程.....	5
图表 2 : 网络文本爬取过程.....	7
图表 3 : 半自动构建领域词典的过程[17].....	20
图表 4 : 翻译系统流程图.....	21
图表 5 : 神经机器翻译“编码器-解码器”结构.....	26
图表 6 : OpenNMT 工作框架.....	27

附表清单

表格 1：语义范畴标签集及示例.....	18
表格 2：应用场景标签集及示例.....	19
表格 3：二维领域知识标注实例.....	22
表格 4：机器翻译实验数据统计量.....	23
表格 5：领域标签数量.....	24
表格 6：过滤放松规则.....	24
表格 7：翻译效果.....	25
表格 8：OpenNMT 自带的英德语料集统计量	28
表格 9：NiuTrans 自带的汉英语料集.....	28
表格 10：NMT 与 SMT 打分对比.....	28

1 引言

机器翻译利用计算机将一种自然语言(源语言)转换为另一种自然语言(目标语言)。作为计算语言学的一个分支,它涉及计算机、认知科学、语言学、信息论等多个学科,是人工智能的终极目标之一,具有重要的科学研究价值。随着经济全球化及互联网的飞速发展,机器翻译技术在促进政治、经济、文化交流等方面起到越来越重要的作用,显示了其重要的实用价值。

机器翻译技术的发展一直与计算机技术、信息论、语言学等学科的发展紧密相随。从早期的词典匹配,到结合语言学专家知识的规则翻译,从基于语料库的统计机器翻译,再到端到端的神经网络机器翻译,随着计算机计算能力的提升和多语言信息的爆发式增长,再加上深度学习算法近两年里程碑式飞跃发展,越来越机器翻译技术开始为普通用户提供实时便捷的翻译服务,翻译质量也得到了极大的提升。

本项重点工作在前期日汉机器翻译研究的基础上,进一步开展平行语料建设和科技机器翻译引擎研究和开发。本项目主要研究工作内容如下:

1) 多语平行语料建设研究

统计机器翻译亟需高质量的平行语料。扩展抽取平行语料是统计机器翻译最为实用的技术。对数年来众多学者们在平行语料抽取方面已经采用的方法和算法进行分类并对比,遴选其中较好的方法进行了经验总结。同时调研目前活跃的双语资料网站分析各自的特点并介绍了相关资源。

2) 日汉科技术语翻译方法研究

科技术语翻译要求高度的准确性和专业性,通过建立术语语料的领域知识标签,并基于待翻译术语的领域对训练语料进行筛选,可训练出针对领域的翻译模型,能极大改善科技术语翻译质量。这对于机器翻译、词典自动编纂、跨语言信息检索等自然语言处理任务都具有重要的实用价值。本项研究以日汉科技术语翻译为研究目标,通过自动获取日语术语的二维领域知识,包括语义范畴标签和应用场景标签,可以对日语术语进行领域归类。该方法应用到统计机器翻译中,可以自动标注日语术语(或句子)的二维领域知识,生成测试集和开发集的领域标签集合,进而筛选训练数据。实验表明,仅利用部分训练数据就可以获取与原始训练数据可比较的翻译结果,验证了本方法的有效性和可行性。本研究有助于解决面向科技术语的统计机器翻译的领域自适应问题。