

基于 Elasticsearch 的机器翻译研究

中国科学技术信息研究所

2020 年 6 月 28 日

目 录

1 引言.....	5
2. 相关工作.....	5
3 基于 Elasticsearch 的机器翻译系统.....	7
3.1 模型结构.....	7
3.2 语料预处理.....	9
3.2 Elasticsearch 相似语料过滤.....	9
3.3 译码策略.....	10
3.3.1 模型平均.....	10
3.3.2 候选译文合并.....	10
3.4 实验.....	10
3.4.1 参数设置.....	11
3.4.2 使用数据.....	11
3.4.3 实验结果.....	11
4 翻译系统部署.....	15
4.1 组织结构.....	15
4.2 功能布局.....	16
4.3 技术架构.....	17
4.4 系统功能.....	18
4.4.1 系统首页（改进完善）.....	18
4.4.2 关于我们（新增页面）.....	19
4.4.3 会员登录（改进完善）.....	20
4.4.4 联系我们（新增页面）.....	22
4.4.5 文本翻译（改进完善）.....	23
4.4.6 词典翻译（改进完善）.....	24
4.4.7 文档翻译（改进完善）.....	26
4.4.8 排版翻译（新增功能）.....	29
4.4.9 翻译设置（新增功能）.....	30
4.4.10 翻译统计（新增功能）.....	32
4.4.11 用户管理（新增功能）.....	33
4.4.12 系统设置（新增功能）.....	35
4.4.13 个人中心（改进完善）.....	36
4.5 性能测试.....	37
4.5.1 测试内容.....	37
4.5.2 测试结果分析.....	37
5 总结.....	40

插图清单

图 1	评测整体流程图.....	7
图 2	基于完全自注意力机制的 transformer 模型架构.....	8
图 3	中日方向分词对比.....	12
图 4	日中方向分词对比.....	12
图 5	科信智译 V2.0 系统组织结构.....	16
图 6	科信智译 V2.0 Web 界面层级结构.....	17
图 7	科信智译 V2.0 Web 界面开发组件依赖关系.....	18
图 8	科信智译 V1.0 首页.....	19
图 9	科信智译 V2.0 首页.....	19
图 10	科信智译 V2.0 “关于我们” 页面.....	20
图 11	科信智译 V2.0 “App 下载” 页面.....	21
图 12	科信智译 V1.0 “用户登录” 页面.....	21
图 13	科信智译 V2.0 “用户登录” 页面.....	22
图 14	科信智译 V2.0 “联系我们” 页面.....	23
图 15	科信智译 V1.0 “文本翻译” 页面.....	23
图 16	科信智译 V2.0 “文本翻译” 页面.....	24
图 17	科信智译 V1.0 “词典翻译” 页面.....	25
图 18	科信智译 V2.0 “词典翻译” 页面.....	26
图 19	科信智译 V1.0 “文档翻译” 页面.....	27
图 20	科信智译 V2.0 “文档翻译” 页面.....	28
图 21	科信智译 V2.0 “排版翻译” 功能.....	30
图 22	科信智译 V1.0 “翻译语种选择” 功能.....	30
图 23	科信智译 V2.0 “翻译语种设置” 功能.....	31
图 24	科信智译 V2.0 “翻译统计” 页面.....	33
图 25	科信智译 V2.0 “用户管理” 界面.....	35
图 26	科信智译 V2.0 “系统设置” 界面.....	36
图 27	科信智译 V1.0 “个人中心” 界面.....	36
图 28	科信智译 V2.0 “个人中心” 界面.....	37
图 29	并发量 50 测试界面.....	38
图 30	并发量 100 测试界面.....	38
图 31	并发量 200 测试界面.....	39
图 32	并发量 400 测试界面.....	39
图 33	并发量 580 测试界面.....	39
图 34	并发量 680 测试界面.....	40

附表清单

表 1	评测方提供数据.....	11
表 2	数据过滤结果.....	12
表 3	中-日 基线系统结果.....	13
表 4	日-中 基线系统结果.....	13
表 5	中-日 开发集对比.....	13
表 6	中-日 开发集对比.....	13
表 7	系统使用数据统计量.....	14
表 8	中-日 实验结果.....	14
表 9	日-中 实验结果.....	14
表 10	中-日 翻译结果.....	15
表 11	语言配置表.....	31
表 12	测试参数.....	37
表 13	并发测试结果.....	37