**附件6：评测组织方发布的资源列表**

如非特殊说明，评测提供的数据文件默认采用UTF-8无BOM编码

**1        汉英/英汉新闻相关资源**

1.1       训练数据

|  |  |  |
| --- | --- | --- |
| 资源名称简写及ChineseLDC 资源编号 | 资源描述 | |
| Datum2015 | 名称 | 点通汉英平行语料库（2015）（部分） |
| 提供单位 | 点通数据有限公司 |
| 语种 | 汉语—英语 |
| 领域 | 综合领域，包括：语言教材、双语图书、技术文档、双语新闻、政府白皮书、政府公文和Web上双语资源等等 |
| 规模 | 1000004个句对 |
| 说明 | 这是点通数据有限公司在863项目支持下开发的《双语/多语平行语料库》的部分内容。 |
| CASICT2011    （CLDC-2010-001）  （CLDC-2012-001） | 名称 | 计算所Web汉英平行语料库（2013） |
| 提供单位 | 中国科学院计算技术研究所 |
| 语种 | 汉语—英语 |
| 领域 | 综合领域 |
| 规模 | 1936633个句对 |
| 说明 | 该平行语料库是从互联网上自动挖掘获得的。双语平行网页的发现、确认，双语平行文本的获取，句子对齐等过程完全通过程序自动实现。语料库抽样评价的正确率在95%以上。  该研究得到国家自然科学基金项目（编号：60603095）的支持。 |
| CASICT2015 | 名称 | 计算所Web汉英平行语料库（2015） |
| 提供单位 | 中国科学院计算技术研究所 |
| 语种 | 汉语—英语 |
| 领域 | 综合领域 |
| 规模 | 2036834个句对 |
| 说明 | 该平行语料库是从互联网上自动挖掘获得的。双语平行网页的发现、确认，双语平行文本的获取，句子对齐等过程完全通过程序自动实现。计算所在此基础上进行了大致的校对，语料库抽样评价的正确率在99%以上。语料构成如下：网络语料占60%，电影字幕语料占20%，来自英汉辞书的例句语料占20%。 |
| CASIA2015 | 名称 | 中科院自动化所Web汉英平行语料库（2015） |
| 提供单位 | 中国科学院自动化研究所 |
| 语种 | 汉语—英语 |
| 领域 | 综合领域 |
| 规模 | 1050000句对 |
| 说明 | 该平行语料库是从互联网上自动挖掘获得的。双语平行网页的发现、确认，双语平行文本的获取，句子对齐等过程完全通过程序自动实现。 |
| Datum2017 | 名称 | 点通公司英汉平行语料库（2017） |
| 提供单位 | 点通数据有限公司 |
| 语种 | 汉语—英语 |
| 领域 |  |
| 规模 | 100万句对，分为20个文件 |
| 说明 |  |
| NEU2017 | 名称 | 东北大学英汉平行语料库（2017） |
| 提供单位 | 东北大学 自然语言处理实验室 |
| 语种 | 汉语—英语 |
| 领域 |  |
| 规模 | 200万句对 |
| 说明 |  |
| SSMT2007 MT Evaluation Data    （2007-863-001） | 名称 | SSMT2007机器翻译评测数据（英汉/汉英机器翻译部分） |
| 提供单位 | 中国科学院计算技术研究所 |
| 语种 | 汉语—英语 |
| 领域 | 新闻 |
| 规模 | 该机器翻译测试语料包含2个翻译方向（汉英、英汉），语料为新闻领域。其中汉英机器翻译测试语料含1,002个汉语句子。英汉机器翻译测试语料含995个英语句子。每个测试句子包括4个人工翻译的参考译文。 |
| 说明 |  |
| HTRDP(863)2005 MT Evaluation Data    （2005-863-001） | 名称 | 2005年863机器翻译评测数据（英汉/汉英机器翻译部分） |
| 提供单位 | 中国科学院计算技术研究所 |
| 语种 | 汉语—英语 |
| 领域 | 包括两种评测语料，一种是对话语料，领域为奥运相关领域，包括体育赛事、天气预报、交通住宿、旅游餐饮等；一种是篇章语料，领域为新闻领域。 |
| 规模 | 汉英对话句对：467句，汉英篇章句对：489句。  英汉对话句对：459句，英汉篇章句对：494句。  每个翻译方向的每个测试句子各提供4个人工翻译的参考译文。 |
| 说明 |  |
| HTRDP(863)2004 MT Evaluation Data    （2004-863-001） | 名称 | 2004年863机器翻译评测数据（英汉/汉英机器翻译部分） |
| 提供单位 | 中国科学院计算技术研究所 |
| 语种 | 汉语—英语 |
| 领域 | 两种评测语料，一种是篇章语料，一种是对话语料。领域是通用领域和奥运的相关领域，其中奥运领域包括体育赛事、天气预报、交通住宿、旅游餐饮等。 |
| 规模 | 汉英评测数据含400句对话语料，308句篇章语料。英汉评测数据含400句对话语料，310句篇章语料。每个翻译方向的每个测试句子各提供4个人工翻译的参考译文。 |
| 说明 | 2004年863机器翻译评测汉英、英汉部分测试数据。 |
| HTRDP(863)2003 MT Evaluation Data    （2003-863-004） | 名称 | 2003年863机器翻译评测数据（英汉/汉英机器翻译部分） |
| 提供单位 | 中国科学院计算技术研究所 |
| 语种 | 汉语—英语 |
| 领域 | 奥运相关领域，其中奥运领域包括体育赛事、天气预报、交通住宿、旅游餐饮等。 |
| 规模 | 汉英评测数据含437句对话语料和169句篇章语料；英汉评测数据含496句对话语料和322句篇章语料。每个翻译方向的每个测试句子各提供4个人工翻译的参考译文。 |
| 说明 | 2003年863机器翻译评测汉英、英汉部分测试数据。 |
| CWMT2008 Machine Translation Evaluation Data    （CLDC-2009-001）  （CLDC-2009-002） | 名称 | CWMT2008机器翻译评测新闻语料（英汉/汉英机器翻译部分） |
| 提供单位 | 中国科学院计算技术研究所 |
| 语种 | 汉语—英语 |
| 领域 | 新闻 |
| 规模 | 汉英评测数据含1006句对；英汉评测数据含1000句对。每个翻译方向的每个测试句子各提供4个人工翻译的参考译文。 |
| 说明 |  |
| CWMT2009 Machine Translation Evaluation Data | 名称 | CWMT2009机器翻译评测数据（英汉/汉英机器翻译部分） |
| 提供单位 | 中国科学院计算技术研究所 |
| 语种 | 汉语—英语 |
| 领域 | 新闻 |
| 规模 | 汉英评测数据含1003句对；英汉评测数据含1002句对。每个翻译方向的每个测试句子各提供4个人工翻译的参考译文。 |
| 说明 |  |
| CWMT2011 Machine  Translation Evaluation Data | 名称 | CWMT2011机器翻译评测数据（英汉机器翻译部分） |
| 提供单位 | 中国科学院计算技术研究所 |
| 语种 | 英语—汉语 |
| 领域 | 新闻 |
| 规模 | 英汉评测数据含3187句对。每个测试句子各提供4个人工翻译的参考译文。 |
| 说明 |  |
| NJU-newsdev2017-enzh    （NJU-newsdev2017-zhen） | 名称 | 南京大学CWMT2017汉英/英汉新闻语料开发集数据 |
| 提供单位 | 南京大学 |
| 语种 | 汉语—英语 |
| 领域 | 新闻 |
| 规模 | 共2,002句对 |
| 说明 | 包含1000个汉语新闻句子及其英语翻译结果，以及1002个英语新闻句子及其汉语翻译结果。 |
| NJU-newstest2017-enzh    （NJU-newstest2017-zhen） | 名称 | 南京大学CWMT2017汉英/英汉新闻语料测试集 |
| 提供单位 | 南京大学 |
| 语种 | 汉语—英语 |
| 领域 | 新闻 |
| 规模 | 共2,001句对 |
| 说明 | 包含1000个汉语新闻句子及其英语翻译结果，以及1001个英语新闻句子及其汉语翻译结果。  该数据是CWMT2018的开发集 |

1.2       单语新闻数据

|  |  |  |
| --- | --- | --- |
| XMU-CWMT2017 | 名称 | 厦门大学NLP实验室新华网新闻汉语单语语料（2017） |
| 提供单位 | 厦门大学 |
| 语种 | 汉语 |
| 领域 | 新闻 |
| 规模 | 现语料库共有662,904个文章，大约1100万词汇。 |
| 说明 | 本资源由厦门大学NLP实验室收集，包括新华网2011年不同主题频道的新闻语料，例如：国内新闻，国际新闻，财经新闻，论坛，教育等。  每篇文章包含：标题，日期，URL和内容。 |

1.3       开发集数据

|  |  |  |
| --- | --- | --- |
| WMT 2018的测试集\* | 名称 | WMT 2018汉英/英汉测试集数据 |
| 提供单位 | WMT 2018发布（<http://data.statmt.org/wmt18/translation-task/test.tgz>）\* |
| 语种 | 汉语—英语 |
| 领域 | 新闻 |
| 规模 | 共3,981句对 |
| 说明 | 包含2481个汉语句子及其英语翻译结果，以及1500个英语句子及其汉语翻译结果； |

\* CCMT2019继续与WMT2019合作，开发集设定也是使用同一个数据。区别在于：CCMT2019要求仅使用该数据的汉英翻译部分用于汉英翻译模型的开发集，仅使用英汉翻译部分用于英汉翻译模型开发集。版权上，“The data released by WMT2018 news translation task can be freely used for research purpose。”

2        蒙汉日常用语项目数据

2.1       训练数据

|  |  |  |
| --- | --- | --- |
| IMU-CWMT2013    （CLDC-2010-005） | 名称 | 内蒙古大学汉蒙平行语料库（2013） |
| 提供单位 | 内蒙古大学 |
| 语种 | 汉语—蒙古语 |
| 领域 | 政府文献和法律法规、日常对话、文学 |
| 规模 | 共104,975句对  其中：CWMT2011评测训练语料67274句对,领域包括: 日常对话、文学、政府文献和法律法规;  CWMT 2013新增训练语料:包括新闻语料17,516句对,政府文献语料10,394句对,课本语料5,052句对,蒙汉字典语料4,739句对; |
| 说明 |  |
| IMU-CWMT2015 | 名称 | 内蒙古大学汉蒙平行语料库（2015） |
| 提供单位 | 内蒙古大学 |
| 语种 | 汉语—蒙古语 |
| 领域 | 政府文献和法律法规、日常对话、文学 |
| 规模 | 共24,978句对 |
| 说明 |  |
| IIM-CWMT2015 | 名称 | 中国科学院合肥智能机械研究所蒙汉双语语料库（2015） |
| 提供单位 | 中国科学院合肥智能机械研究所 |
| 语种 | 蒙古语—汉语 |
| 领域 | 新闻 |
| 规模 | 1,682句对 |
| 说明 |  |
| ICT-MC-corpus-CWMT2017 | 名称 | 中国科学院计算技术研究所蒙汉双语语料库（2017） |
| 提供单位 | 中国科学院计算技术研究所 |
| 语种 | 蒙古语—汉语 |
| 领域 | 新闻 |
| 规模 | 30,007句对 |
| 说明 |  |
| IMU-corpus-CWMT2017 | 名称 | 内蒙古大学蒙汉双语语料库（2017） |
| 提供单位 | 内蒙古大学 |
| 语种 | 蒙古语—汉语 |
| 领域 | 综合，包括：政府文件，政府工作报告，国务院文件，法律法规等 |
| 规模 | 100,001句对 |
| 说明 |  |
| IMU-dev-mnzh -CWMT2017 | 名称 | 内蒙古大学CWMT2017蒙汉开发集数据 |
| 提供单位 | 内蒙古大学 |
| 语种 | 蒙古语—汉语 |
| 领域 | 政府文献和法律法规、日常对话、文学 |
| 规模 | 共1,000句蒙古语，每句4个汉语参考译文 |
| 说明 | CMWT2017蒙汉开发集与CWMT2011、CWMT2013、CWMT2015蒙汉开发集相同 |
| IMU-test-mnzh-CWMT2017 | 名称 | 内蒙古大学CWMT2017蒙汉测试集数据 |
| 提供单位 | 内蒙古大学 |
| 语种 | 蒙古语—汉语 |
| 领域 | 政府文献和法律法规、日常对话、文学 |
| 规模 | 共1,001句蒙古语，每句4个汉语参考译文 |
| 说明 | CWMT2017蒙汉测试集即CMWT2018蒙汉开发集数据 |

2.2       开发集数据

|  |  |  |
| --- | --- | --- |
| IMU-test-mnzh -CWMT2018 | 名称 | 内蒙古大学CWMT2018蒙汉测试集数据 |
| 提供单位 | 内蒙古大学 |
| 语种 | 蒙古语—汉语 |
| 领域 | 政府文献和法律法规、日常对话、文学 |
| 规模 | 共1,000句蒙古语，每句1个汉语参考译文 |
| 说明 | CWMT2018蒙汉测试集被指定为CCMT2019蒙汉开发集 |

3        藏汉政府文献相关资源

3.1       训练数据

|  |  |  |
| --- | --- | --- |
| QHNU-CWMT2013 | 名称 | 青海师范大学藏汉平行语料库（2013） |
| 提供单位 | 青海师范大学 |
| 语种 | 藏语—汉语 |
| 领域 | 政府文献领域 |
| 规模 | 33,145句对 |
| 说明 | 该平行语料库是通过录入、扫描、网页下载等方式获得的。双语平行的搜集、整理、确认、获取、句子对齐等过程是通过程序自动实现和人工干预实现的。语料库的正确率在99%以上。  该研究得到国家自然科学基金项目（编号：61063033）和973前期研究专项（编号：2010CB334708）的支持。 |
| QHNU-CWMT2015 | 名称 | 青海师范大学藏汉平行语料库（2015） |
| 提供单位 | 青海师范大学 |
| 语种 | 藏语—汉语 |
| 领域 | 政府文献领域 |
| 规模 | 17,194句对 |
| 说明 | 该平行语料库是通过录入、扫描、网页下载等方式获得的。双语平行的搜集、整理、确认、获取、句子对齐等过程是通过程序自动实现和人工干预实现的。语料库的正确率在99%以上。  该研究得到国家自然科学基金项目（编号：61063033）的支持。 |
| XBMU-XMU | 名称 | 央金藏汉平行语料库 |
| 提供单位 | 厦门大学人工智能研究所  西北民族大学语言（技术）研究所 |
| 语种 | 汉语—藏语 |
| 领域 | 综合领域 |
| 规模 | 52,078句对 |
| 说明 | 1)该藏汉平行语料库是用正式出版物、藏汉大词典和网络语料藏汉对照文本,经使用自主开发的“藏汉句子对齐工具”初步对齐之后,由人工逐句对齐。  2)5万句对藏汉平行语料的对齐正确率为100%。  3)该研究得到国家社科基金重点项目《藏语语料库建设研究》（批准号：05AYY001）和863重点项目《面向跨语言搜索的机器翻译关键技术研究》（批准号：2006AA010107）的支持。 |
| XBMU-XMU- UTibet | 名称 | 西北民族大学、西藏大学与厦门大学藏汉语料（2012） |
| 提供单位 | 西北民族大学语言（技术）研究所  西藏大学  厦门大学人工智能研究所 |
| 语种 | 汉语—藏语 |
| 领域 | 政论，法律 |
| 规模 | 24,159句对 |
| 说明 | 语料来源：2008年和2009年全国最新法律文件和十八大报告、2011、2012年政府工作报告等，政论类与法律类语料各占一半。系西北民族大学、西藏大学与厦门大学于2012年通过对原材料进行扫描、识别、校对并独立加工完成。 |
| ICT-TC-corpus-CWMT2017 | 名称 | 中国科学院计算技术研究所藏汉双语语料库（2017） |
| 提供单位 | 中国科学院计算技术研究所 |
| 语种 | 藏语—汉语 |
| 领域 | 新闻 |
| 规模 | 30,004句对 |
| 说明 |  |
| QHNU-dev-tizh- CWMT2017 | 名称 | 青海师范大学CWMT2017藏汉开发集数据 |
| 提供单位 | 青海师范大学 |
| 语种 | 藏语—汉语 |
| 领域 | 政府文献 |
| 规模 | 共650句藏语，每句4个汉语参考译文 |
| 说明 | CMWT2017藏汉开发集与CWMT2011、CWMT2013、CWMT2015藏汉开发集相同 |
| QHNU-test-tizh- CWMT2017 | 名称 | 青海师范大学CWMT2017藏汉测试集数据 |
| 提供单位 | 青海师范大学 |
| 语种 | 藏语—汉语 |
| 领域 | 政府文献 |
| 规模 | 共729句藏语，每句4个汉语参考译文 |
| 说明 | CWMT2017藏汉测试集即CMWT2018藏汉开发集数据 |

3.2       开发集数据

|  |  |  |
| --- | --- | --- |
| QHNU-test-tizh- CWMT2018 | 名称 | 青海师范大学CWMT2018藏汉测试集数据 |
| 提供单位 | 青海师范大学 |
| 语种 | 藏语—汉语 |
| 领域 | 政府文献 |
| 规模 | 共1000句藏语，每句1个汉语参考译文 |
| 说明 | CCMT2019藏汉开发集使用CWMT2018的藏汉测试集 |

4        维汉新闻相关资源

4.1       训练数据

|  |  |  |  |
| --- | --- | --- | --- |
| XJU-CWMT2013   XJU-corpus-CWMT2017  XJU-dev-uyzh-CWMT2017  XJU-dev-uyzh-CWMT2018 | | | *这几份数据2019年暂未进入训练语料；*  (注CMWT2017维汉开发集与CWMT2011、CWMT2013、CWMT2015维汉开发集相同) |
| XJIPC-CWMT2015 | 名称 | 中国科学院新疆理化技术研究所维汉双语语料库（2015） | |
| 提供  单位 | 中国科学院新疆理化技术研究所 | |
| 语种 | 汉语—维吾尔语 | |
| 领域 | 新闻 | |
| 规模 | 59,990句对 | |
| 说明 | 此语料在CWMT 2013年的基础上新增加约3万句对  语料比例：2007年-2014年媒体新闻类语料比例约占95%，2012-2014年政府报告类语料和法律法规类语料合计比例约占5%。  语料来源：系中国科学院新疆理化技术研究所在2012年-2014年间采集、标注、校对完成。 | |
| ICT-UC-corpus-CWMT2017 | 名称 | 中国科学院计算技术研究所维汉双语语料库（2017） | |
| 提供单位 | 中国科学院计算技术研究所 | |
| 语种 | 维吾尔语—汉语 | |
| 领域 | 新闻 | |
| 规模 | 30,071句对 | |
| 说明 |  | |
| XJIPC-corpus-CWMT2017 | 名称 | 中国科学院新疆理化技术研究所维汉双语语料库（2017） | |
| 提供单位 | 中国科学院新疆理化技术研究所 | |
| 语种 | 维吾尔语—汉语 | |
| 领域 | 新闻 | |
| 规模 | 30,000句对 | |
| 说明 |  | |
| XJIPC-corpus-CWMT2018 | 名称 | 中国科学院新疆理化技术研究所维汉双语语料库（2018） | |
| 提供单位 | 中国科学院新疆理化技术研究所 | |
| 语种 | 维吾尔语—汉语 | |
| 领域 | 新闻 | |
| 规模 | 50,000句对 | |
| 说明 |  | |

4.2       开发集数据

|  |  |  |
| --- | --- | --- |
| XJIPC-test-uyzh-CWMT2018 | 名称 | 中国科学院新疆理化技术研究所CWMT2018维汉测试集数据 |
| 提供单位 | 中国科学院新疆理化技术研究所 |
| 语种 | 维吾尔语—汉语 |
| 领域 | 新闻 |
| 规模 | 共1000句维语，每句1个汉语参考译文 |
| 说明 | CCMT2019维汉开发集数据为CWMT2018维汉测试集数据 |

5       多语言机器翻译（日英专利领域）相关资源

5.1       训练数据

|  |  |  |
| --- | --- | --- |
| Lingosail-train-zhjp-CWMT2018 | 名称 | 北京语智云帆科技有限公司日汉专利平行语料库（2018） |
| 提供单位 | 北京语智云帆科技有限公司 |
| 语种 | 日语—汉语 |
| 领域 | 综合 |
| 规模 | 3,000,000句对 |
| 说明 | 该数据更新了2017版的日汉专利平行数据 |
| Lingosail-train-enzh-CWMT2018 | 名称 | 北京语智云帆科技有限公司英汉专利平行语料库（2018） |
| 提供单位 | 北京语智云帆科技有限公司 |
| 语种 | 英语—汉语 |
| 领域 | 综合 |
| 规模 | 3,000,000句对 |
| 说明 |  |

5.2       开发集数据

|  |  |  |
| --- | --- | --- |
| Lingosail-dev-jpzh-CWMT2017 | 名称 | 北京语智云帆科技有限公司日汉双语开发集数据（2017） |
| 提供单位 | 北京语智云帆科技有限公司 |
| 语种 | 日语—汉语 |
| 领域 | 综合 |
| 规模 | 3000句日语，每句含一个汉语参考译文 |
| 说明 | CWMT2017日汉专利领域翻译开发集数据 |
| Lingosail-dev-enzh-CWMT2018 | 名称 | 北京语智云帆科技有限公司英汉双语开发集数据（2018） |
| 提供单位 | 北京语智云帆科技有限公司 |
| 语种 | 英语—汉语 |
| 领域 | 综合 |
| 规模 | 3000句英语，每句含一个汉语参考译文 |
| 说明 |  |
| Lingosail-dev-jpzh-CWMT2018 | 名称 | 北京语智云帆科技有限公司日汉双语开发集数据（2018） |
| 提供单位 | 北京语智云帆科技有限公司 |
| 语种 | 日语—汉语 |
| 领域 | 综合 |
| 规模 | 3000句日语，每句含一个汉语参考译文 |
| 说明 |  |
| Lingosail-dev-enjp-CWMT2018 | 名称 | 北京语智云帆科技有限公司英日双语开发集数据（2018） |
| 提供单位 | 北京语智云帆科技有限公司 |
| 语种 | 英语—日语 |
| 领域 | 综合 |
| 规模 | 3000句英语，每句含一个日语参考译文 |
| 说明 |  |

5.3       汉语专利数据

|  |  |  |
| --- | --- | --- |
| Lingosail-cn\_for\_lm-CWMT2017 | 名称 | 北京语智云帆科技有限公司汉语专利语料（2017） |
| 提供单位 | 北京语智云帆科技有限公司 |
| 语种 | 汉语 |
| 领域 | 综合 |
| 规模 | 7,114,700句对 |
| 说明 | CWMT2017日汉专利领域翻译汉语单语数据 |

6        语音机器翻译评测相关资源

6.1训练数据

|  |  |  |
| --- | --- | --- |
| BSTC-zhen-train | 名称 | 百度语音翻译语料库-训练集 |
| 提供单位 | 百度 |
| 语种 | 汉-英 |
| 领域 | 综合 |
| 规模 | 50小时以上 |
| 说明 | 包含中文语音、中文文本、英文文本 |

6.2开发集数据

|  |  |  |
| --- | --- | --- |
| BSTC-zhen-dev | 名称 | 百度语音翻译语料库-开发集 |
| 提供单位 | 百度 |
| 语种 | 汉-英 |
| 领域 | 综合 |
| 规模 | 1小时 |
| 说明 | 包含中文语音、中文文本、英文文本 |

7        汉英/英汉多领域机器翻译质量评估相关资源

7.1      词汇级翻译质量估计

7.1.1训练数据

|  |  |  |
| --- | --- | --- |
| Alibaba-train-enzh-qe-word-level-CCMT2019 | 名称 | 阿里巴巴英汉多领域机器翻译质量评估语料库（2019） |
| 提供单位 | 阿里巴巴（中国）有限公司 |
| 语种 | 英语—汉语 |
| 领域 | 经济，政治，科技，口语，电商，IT等领域 |
| 规模 | 10878句对 |
| 说明 | 训练集由四个文件组成，包括由逐行对应的源语言文件train.source、译文文件train.target、对译文进行人工译后编辑的文件train.pe，每行为一个句子；以及待评估译文的Tags文件train.tags。 |
| Alibaba-train-zhen-qe-word-level-CCMT2019 | 名称 | 阿里巴巴汉英多领域机器翻译质量评估语料库（2019） |
| 提供单位 | 阿里巴巴（中国）有限公司 |
| 语种 | 汉语—英语 |
| 领域 | 经济，政治，科技，口语，电商，IT等领域 |
| 规模 | 11039句对 |
| 说明 | 训练集由四个文件组成，包括由逐行对应的源语言文件train.source、译文文件train.target、对译文进行人工译后编辑的文件train.pe，每行为一个句子；以及待评估译文的Tags文件train.tags。 |

7.1.2开发集数据

|  |  |  |
| --- | --- | --- |
| Alibaba-dev-enzh-qe-word-level-CCMT2019 | 名称 | 阿里巴巴英汉多领域机器翻译质量评估开发集数据（2019） |
| 提供单位 | 阿里巴巴（中国）有限公司 |
| 语种 | 英语—汉语 |
| 领域 | 经济，政治，科技，口语，电商，IT等领域 |
| 规模 | 1129句对 |
| 说明 | 开发集由四个文件组成，包括由逐行对应的源语言文件dev.source、译文文件dev.target、对译文进行人工译后编辑的文件dev.pe，每行为一个句子；以及待评估译文的Tags文件dev.tags。 |
| Alibaba-dev-zhen-qe-word-level-CCMT2019 | 名称 | 阿里巴巴汉英多领域机器翻译质量评估开发集数据（2019） |
| 提供单位 | 阿里巴巴（中国）有限公司 |
| 语种 | 汉语—英语 |
| 领域 | 经济，政治，科技，口语，电商，IT等领域 |
| 规模 | 1050句对 |
| 说明 | 开发集由四个文件组成，包括由逐行对应的源语言文件dev.source、译文文件dev.target、对译文进行人工译后编辑的文件dev.pe，每行为一个句子；以及待评估译文的Tags文件dev.tags。 |

7.2句子级翻译质量估计

7.2.1训练数据

|  |  |  |
| --- | --- | --- |
| Lingosail-train-enzh-qe-CCMT2019 | 名称 | 北京语智云帆科技有限公司英汉多领域机器翻译质量评估语料库（2019） |
| 提供单位 | 北京语智云帆科技有限公司 |
| 语种 | 英语—汉语 |
| 领域 | 时政、经济、文化、科技等领域 |
| 规模 | 14789句对 |
| 说明 | 子训练集由四个文件组成，包括由逐行对应的源语言文件train. source、译文文件train. target、对译文进行人工译后编辑的文件train.pe，每行为一个句子；以及待评估译文的HTER值文件train.hter，每行为一个区间[0,1]的数值。 |
| Lingosail-train-zhen-qe-CCMT2019 | 名称 | 北京语智云帆科技有限公司汉英多领域机器翻译质量评估语料库（2019） |
| 提供单位 | 北京语智云帆科技有限公司 |
| 语种 | 汉语—英语 |
| 领域 | 时政、经济、文化、科技等领域 |
| 规模 | 10070句对 |
| 说明 | 子训练集由四个文件组成，包括由逐行对应的源语言文件train.source、译文文件train.target、对译文进行人工译后编辑的文件train.pe，每行为一个句子；以及待评估译文的HTER值文件train.hter，每行为一个区间[0,1]的数值。 |

7.2.2开发集数据

|  |  |  |
| --- | --- | --- |
| Lingosail-dev-enzh-qe-CCMT2019 | 名称 | 北京语智云帆科技有限公司英汉多领域机器翻译质量评估开发集数据（2019） |
| 提供单位 | 北京语智云帆科技有限公司 |
| 语种 | 英语—汉语 |
| 领域 | 时政、经济、文化、科技等领域 |
| 规模 | 300句英语，每句含多个汉语参考译文，共计1381句译文 |
| 说明 | 开发集由四个文件组成，包括由逐行对应的源语言文件dev. source、译文文件dev. target、对译文进行人工译后编辑的文件dev.pe，每行为一个句子；以及待评估译文的HTER值文件dev.hter，每行为一个区间[0,1]的数值。 |
| Lingosail-dev-zhen-qe-CCMT2019 | 名称 | 北京语智云帆科技有限公司汉英多领域机器翻译质量评估开发集数据（2019） |
| 提供单位 | 北京语智云帆科技有限公司 |
| 语种 | 汉语—英语 |
| 领域 | 时政、经济、文化、科技等领域 |
| 规模 | 300句汉语，每句含多个英语参考译文，共计1143 |
| 说明 | 开发集由四个文件组成，包括由逐行对应的源语言文件dev. source、译文文件dev. target、对译文进行人工译后编辑的文件dev.pe，每行为一个句子；以及待评估译文的HTER值文件dev.hter，每行为一个区间[0,1]的数值。 |