数据库产品说明书

DB-ASR-107

多语种平行语料数据库

目　　录

1. 数据产品概述

1.1. 数据概述

标贝多语种平行语料数据库数据量约 1 亿句对，包含 20 个语种。文本内容涉及通用、新闻、演讲等领域。文本编码均为 UTF-8 格式。

1.2. 数据内容

数据目录树	
DB-ASR-107	
└doc	
└ 数据产品说明书.pdf	(产品说明书)
└ data	(数据文件夹)
└en-zh	(语种文件夹)
└ txt	(文本文件)
└ar-zh	(语种文件夹)
└fr-zh	(语种文件夹)
└my-zh	(语种文件夹)
└es-zh	(语种文件夹)
└pt-zh	(语种文件夹)
└th-zh	(语种文件夹)
└ru-zh	(语种文件夹)
└hindi-zh	(语种文件夹)
└vi-zh	(语种文件夹)
└de-zh	(语种文件夹)
└pt-en	(语种文件夹)
└de-en	(语种文件夹)
└en-ja	(语种文件夹)

└en-ko	(语种文件夹)
└en-ru	(语种文件夹)
└cw-zh	(语种文件夹)
└ww-zh	(语种文件夹)

质量标准	文本信息字正确率达 90%
存储方式	ftp 服务器存储
数据敏感项	无
版权所有者	标贝（青岛）科技有限公司

1.3. 应用场景

本产品适用于语音识别、机器翻译等领域，可应用于以下应用场景。

- 科研，可用于文本翻译模型训练或算法研究
- 智能科技
- 输入法、社交
- 教育、娱乐、游戏、传媒

2. 数据产品详情

2.1. 翻译语料

本数据库包含 20 个语种，主要为日常语句，涵盖了通用、旅游、金融、新闻、教育、法律等多个场景。

类型	翻译语种	数量	单位
阿拉伯语（ar）	中文（zh）	1,352,091	句对
法语（fr）	中文（zh）	3,506,709	句对

缅甸语（my）	中文（zh）	48,772	句对
西班牙语（es）	中文（zh）	3,877,545	句对
葡萄牙语（pt）	中文（zh）	700,662	句对
泰语（th）	中文（zh）	87,292	句对
俄语（ru）	中文（zh）	7,342,550	句对
印地语（hindi）	中文（zh）	1,581,312	句对
越南语（vi）	中文（zh）	1,120,879	句对
德语（de）	中文（zh）	1,403,037	句对
葡萄牙语（pt）	英文（en）	5,023,597	句对
德语（de）	英文（en）	19,545,086	句对
英文（en）	日语（ja）	126,899	句对
英文（en）	韩语（ko）	2,505,551	句对
英文（en）	俄语（ru）	5,091,644	句对
朝鲜语（cw）	中文（zh）	694,532	句对
维吾尔语（ww）	中文（zh）	138,922	句对
韩语（ko）	中文（zh）	1,187,318	句对
法语（fr）	英文（en）	29,518,828	句对
英文（en）	中文（zh）	21,096,948	句对

2.2.采集方法

2.2.1. 翻译方法

源文本内容无逻辑不通，语义混乱或者句子不完整的现象，无错别字，特殊字符，乱码或夹杂大量其他语种，目标文本可完整的表达出源文本的意思，且表达流利地道，符合目标语的表述习惯，专有名词均使用官方翻译。

2.2.2. 数据样例说明

样例均取自真实数据产品，是对完整数据产品的局部展示，仅用于参考。本产品样例涵盖每个翻译语种各 100 句对语料。