

翻译标注说明

1.表格说明

本次实验为机器翻译的自动化测试，表格为测试工具捕获的机器翻译错误。该工具使用 Fuzzing 的思想，基于英文原句，进行删减、扩充与变形，得到一组存在包含关系的英文句子集合（一组句子中，下一行的英文句子内容包含上一行的英文句子）。根据翻译结果之间的蜕变关系，判断是否存在翻译错误，因此从 bug_id 列中可以看出，两个句子为一组，报告一个疑似 bug。从上图中可以看出，error_word 为句子中标红的部分，即工具认为可能存在错误的地方。

sent_id	error_word	bug_id	step	original_s	translation
0	寻找	0	-1 > 0	Nintendo	任天堂已经梦想为粉丝们 寻找 一条路。
				Nintendo	任天堂梦想为粉丝们找到更好的方法。
	新路	1	1 > 2	Nintendo	任天堂为马里奥卡丁车迷们想出了一条 新路 。
				Nintendo	任天堂梦想为马里奥卡丁车迷们提供一种新的游戏方式。
		2	1 > 2	Nintendo	任天堂为马里奥卡丁车迷们想出了一条 新路 。
			Nintendo	任天堂梦想为马里奥卡丁车迷们提供一条新的比赛方式。	
粉丝	3	0 > 1	Nintendo	任天堂梦想为 粉丝 们提供一种新的方式。	
			Nintendo	任天堂为马里奥卡丁车迷们想出了一条新路。	
2	消失	0	0 > 1	Twitter, a	Twitter已经受到警方和监管机构的一击，似乎完全 消失 了。
				Twitter, a	Twitter已经受到警方和监管机构的一击，周四似乎完全崩溃。
		1	0 > 1	Twitter, a	Twitter已经受到立法者和监管机构的攻击，似乎完全 消失 了。
				Twitter, a	Twitter已经受到立法者和监管机构的攻击，周四似乎完全崩溃。
	倒下	2	0 > 1	Twitter, a	Twitter已经受到来自立法者和监管机构的压力，似乎完全 倒下 了。
				Twitter, a	Twitter已经受到来自立法者和监管机构的压力，周四似乎完全下跌。
火力消失	3	0 > 1	Twitter, a	Twitter已经受到立法者和监管机构的 火力 ，似乎完全 消失 了。	
			Twitter, a	Twitter已经受到立法者和监管机构的一击，周四似乎完全崩溃。	

2.翻译错误判断

	A	B	C	D	E	F	G	H	I	J
1	sent_id	error_word	bug_id	step	original	translation				
2		为科	0	-1 -> 0	Boris Jol	鲍里斯·约翰逊为科维德统计数据的使用进行了辩护。	错误(T)/误报(F)	错误类型	误报类型	是否相关
3					Boris Jol	鲍里斯·约翰逊为在一次演示中数据分析使用科维德统计数据进行了辩护。	T	5		T
4					Boris Jol	鲍里斯·约翰逊为科维德统计数据在一次演示中解释英国被后使用方式进行了辩护。				
5			1	2 -> 3	Boris Jol	鲍里斯·约翰逊在一次演示中解释英国必须解除封锁后，为科维德的统计数据使用进行了辩护。	T		1	T

一组句子可能存在四种情况：1) 短句存在错误，长句正确，在短句后标注 T 即可；2) 短句正确，长句存在错误，在长句后标注 T 即可；3) 短句和长句均错误，短句和长句后皆标注 T；4) 短句和长句均正确，属于误报（False Positive）的情况，在短句后标注 F 即可。

标注 T 的行，需要标注翻译的错误类型和存在的错误是否与 error_word 相关。

标注 F 的行，需要标注误报的原因。

3.翻译错误类型

翻译错误类型定义及示例如下：

- 1) 译文选词错误：指原文单词中有对应的译文，但译文错误或者不准确。

原文：The soldiers took bin Laden's body with them.

机器译文：(这些)士兵带上本·拉登的身体(遗体)。

- 2) 实词译文缺失：指缺失应有的实词译文，主要包括以下几种情形：

译文消失：指原文缺少必须翻译出来的相应译文。

原文：The decision is made in a short story in "Action Comics" No. 900.

机器译文：这一决定是在很短的故事在“漫画行动”900号。

原文照搬：实词译文为不符合要求的原文照搬。

原文：And there was no lack of self-lovin' in the '80s.

机器译文：在80年代，就不缺少 self-lovin'。

省略未译：没有根据翻译需要，翻译出原文中省略的实词。

原文: Tornadoes devastate South, killing at least 290.

机器译文: 龙卷风摧毁南方, 杀害至少 290 个。

3) **逻辑错误**: 原文中的单词或短语虽然被正确翻译, 但在译文中出现的位置不正确, 导致译文逻辑混乱, 令人费解。

原文: who soon became separated from her team and bodyguard as the crowd swept her up.

机器译文: (其[马上;不久]与她的团队和保镖失散)随着人群席卷她。

4) **意思相反**: 指译文中有的词语或短语与原文意思相反, 导致译文含义出现严重偏差。

原文: He denies her kidnap and murder.

机器译文: 他否认她的绑架和谋杀。

5) **命名实体问题**: 指原文中的命名实体 (包括人名、地名、机构名等) 翻译错误或者没有翻译

原文: Software company Adobe, which develops software such as Photoshop and Acrobat, will repurchase stock worth \$2 billion.

机器译文: 开发 Photoshop 和杂技等软件的软件公司 Adobe 将回购价值 20 亿美元的股票。

6) **数量词/时间词问题**: 原文中的数量词或时间词翻译错误或缺失。

7) **其他错误**: 不属于上述分类的错误翻译。

在标注过程中填写数字即可。若某组句子属于误报, 则该列为空。

4. 误报类型

误报类型定义如下:

- 1) **句子结构变化**: 不同的句式表达相同的句义, 即句型变换。
- 2) **同义词**: 长句中存在与 error_word 相同意思的词语
- 3) **缩写**: 长句中使用一个词替换了短句两个或以上的词语, 或短句中使用一个词语替换了长句中两个或以上的词语

例子:

房屋所有者的抵押贷款和假期也延长了。

受疫情影响的房主的抵押贷款和假期也延长了。

4) **分词错误**: 如下图所示

一大	0	0 -> 1	The pande 这 一大 流行病严重扰乱了教育和卫生系统。	
			The pande 这一流行病严重破坏了存在的教育和卫生系统。	
	1	2 -> 3	The pande 这 一大 流行病严重扰乱了数百万父母的教育和卫生系统。	
			The pande 这一流行病严重扰乱了数百万父母赖以生存的教育和卫生系统。	
	2	2 -> 3	The pande 这 一大 流行病严重扰乱了父母存在的教育和卫生系统。	
			The pande 这一流行病严重扰乱了父母依赖的教育和卫生系统。	
	3	0 -> 1	The pande 这 一大 流行病严重扰乱了教育和卫生系统。	
			The pande 这一流行病严重扰乱了数百万人的教育和卫生系统。	
	4	2 -> 3	The pande 这 一大 流行病严重扰乱了数百万父母的教育和医疗系统。	
			The pande 这一流行病严重扰乱了数百万父母赖以生存的教育和医疗系统。	
	5	1 -> 2	The pande 这 一大 流行病严重扰乱了数百万人的教育和医疗系统。	
			The pande 这一流行病严重扰乱了数百万人的教育和医疗系统。	

5) **其他**

填写数字即可。

若该组句子中的某一句存在翻译错误, 则该组的两行中, 此列皆为空。

5. 是否与红字相关

某句存在翻译错误, 需要使用 **T** 或 **F** 标注存在的错误是否与红字标注的相关, 如词序颠倒、词不达意等等。若该组属于误报, 则该组的两行中, 此列皆为空。