军政领域社交媒体英汉
双语平行语料库构建方法

夏榕璟，张克亮*，唐亮，李铭

（信息工程大学洛阳校区，河南 洛阳，471003）

摘要：高质量的机器翻译系统的研究离不开高质量的平行语料库。随着机器翻译技术的发展，平行语料库研究的重要性也愈加凸显。构建军政领域社交媒体语料库，对于军政领域社交媒体语言特点和军政领域社交媒体机器翻译任务具有重要意义。通过分析现有机器翻译引擎对军政领域社交媒体语料的翻译中存在的用户名和标签的不译、网络非正规表达的不译等翻译问题，提取军政领域社交媒体语料特征，在利用翻译引擎的基础上，优化了流畅度，BLEU 值提升 1.37 个百分点，最终构建规模约为 20 万句对、用于机器翻译任务的军政领域社交媒体英汉双语句对齐平行语料库。

关键词：平行语料库；军政领域；社交媒体；机器翻译

中图分类号：TP391　　**文献标志码：**A

不同领域的平行语料对于构建垂直领域机器翻译系统具有重要意义。为保证某一垂直领域机器翻译的效果，语料库通常需要具有实用性和针对性。社交媒体上每天都有大量的文本数据不断涌现，这些文本数据为构建语料库提供了大量的真实语言样本。单语的社交媒体语料相对较容易获取。但是，构建服务于及军政领域社交媒体机器翻译的双语平行语料库存在较大难度，具体表现在双语句对采集难、语料质量差等问题上。

对于构建军政领域社交媒体英汉双语平行语料库难这一问题，本文提出一种解决思路：结合英语军政领域社交媒体语言特征，采用预处理手段，基于已有的翻译引擎，提升机器译文质量，服务于军政领域社交媒体机器翻译任务。

1 相关研究

语料库是具有一定规模的真实语言样本的集合[1]。双语语料一般指平行语料，是由源语言和目标语言互译数据构成的集合。平行语料用于训练机器翻译系统中的翻译模型，学习、获取翻译知识[2]。在机器翻译、双语词典编纂、对外汉语教学等学科领域中，平行语料库是重要的研究基础资源[3]

上世纪 80 年代，语料库语言学的兴起为机器翻译发展带来了转机，数据驱动的建模方法成为了机器翻译的基础。21 世纪以来，随着神经机器翻译的发展，对语料的规模和质量的需求也在不断提高。

无论是统计机器翻译还是神经机器翻译，双语平行语料必不可少。统计机器翻译系统中，平行语料被拆分成词、短语等碎片信息，以表达基本粒度的翻译知识。基于基本知识和语言模型以及其他不同的特征信息相互作用，经过组合、排序、搜索等操作后将源语言翻译为目标语言。神经机器翻译系统中，源语言作为输入，目标语言作为输出的标签数据，处于网络模型架构的两端。平行语料中的目标语言部分是源语言的参考翻译目标，真实的翻译结果与目标语言部分之间的差异最小化是神经机器翻译模型训练的优化目标。

目前开源的英汉数据集可从 OPUS 官方网站获取，可以下载包括联合国通用领域平行语料库[4]以及新闻、电影字幕、TED 演讲等特定领域的语料。但是难以获取到开源的军政主题的语料库。

* **通信作者：**kliang99@sina.com

在信息时代，丰富的网络资源为构建英汉双语平行语料库提供了许多资源，例如徐润华等人[5]提出了面向 Web 的英汉平行语料库的方法，以此构建大规模的英汉双语平行语料库。关于构建社交媒体的双语平行语料库，部分学者也已经有了一些探索和实践。首先可以依赖于社交网络平台内置的翻译引擎，但是对这样双语的网页进行数据爬取时存在两个问题——第一，通过这样方式获取的并行语料相对丰富，但是往往非常嘈杂，需要对语料进行对齐处理[6]；第二，社交网络平台内置的翻译引擎翻译的质量并不高。Wang 等人[7]发现许多人在 Twitter 和新浪微博中会根据目标受众，用不同的语言两个平台发布内容相同的帖子，以此设计了一套有效的动态规划算法，对数据进行抓取、对齐，然后构建平行语料库。相比于根据内置翻译引擎爬取双语语料的方法，后者的平行语料库的在语料质量上得到了提升，但是也存在数据量少的问题。并且，在军政领域上，难以通过这样的方法得到双语语料。

总体来说，单语的社交媒体语料较容易获取，但是构建高质量的有一定规模的社交媒体双语平行语料库还十分困难。

2 军政领域社交媒体数据库构建方法

2.1 翻译引擎构建平行语料库存在的翻译问题

21 世纪以来，各大社交媒体公司纷纷跨入了“机器翻译”这一领域，让文本信息能够更好地“无障碍”交流。2013 年，Twitter 公司联合 Bing 在 Windows 上测试了翻译功能。2015 年，Twitter 与 Bing 在各个系统上正式推出了翻译功能，支持四十多种语言，包括中文。目前，Twitter 与谷歌合作，用谷歌浏览器访问可直接用谷歌翻译推文内容。2017 年，Facebook 公司将卷积神经网络应运用机器翻译任务中[8]，大幅提升了运算速度。所以，根据 Twitter 和 Facebook 现有的机器翻译功能，围绕关键词和关键军政人物，设计爬取策略，得到双语句对。但是，这样得到的双语语料存在许多问题，以 Twitter 和 Facebook 中各一条博客为例进行分析。

例 1 源文：@POTUS and I will work to deepen the U.S.-German partnership.

Google 翻译：@POTUS 和我将努力深化美德伙伴关系。

例 2 源文：ICYMI: Seabees assigned to NMCB 133's Runnin' Roos assist with road clearing operations during #HurricaneIda disaster relief efforts in Grand Isle, Louisiana.

Facebook 翻译：ICYMI：在路易斯安那州格兰德岛的#Hurricaneida 救灾工作中，分配到 NMCB 133 的 Runnin‘Roos 的海蜂协助清理道路。

从例 1 和例 2 中发现，“POTUS”、“ICYMI”、“HurricaneIda”、“NMCB 133's Runnin' Roos”都没有被翻译出，而这样的情况造成了译文句子含义不完整、可读性差等问题。通过这种方法获取的平行语料显然不足以作为平行语料训练机器翻译系统。

为全面分析和对比各机器翻译引擎对于军政领域社交媒体内容的翻译效果，从 20 万句来源于 Twitter 和 Facebook 的军政主题单语语料中抽取 3000 余句，进行人工精准校验后得到测试集，使用谷歌、百度、有道三个常用的英汉机器翻译引擎检验翻译效果，计算得到的 BLEU 值如表 1 所示。

表 1 三种机器翻译引擎的 BLEU 值表

Tab.1 Bleu value table of three machine translation engines

翻译引擎	BLEU 值
谷歌	29.01
百度	**45.00**
有道	31.89

在军政领域社交媒体文本的翻译中，百度翻译的效果最佳，并且显著高于其他两种翻译引擎，但是总体而言，三种机器翻译引擎的 BLEU 值都不高。对比翻译结果和结合上述例 1 和例 2 中翻译问题的分析，可总结得到军政领域社交媒体机器翻译存在的主要共性三个问题。

第一，用户名、标签的不译。在 Twitter 和 Facebook 中，“@”后跟用户名，与全名相区别，前者有且只有一个，不能和他人重复，后者可以视作一个昵称；“#”表示一个标签，用于标记话题和主题。由于“@”和“#”所携带内容中没有空格，所以“@”和“#”后的内容常常不会被机器翻译引擎翻译出，部分情况下这些内容或许并不重要，我们只需要了解这是一个姓名或者标签即可，但是，某些情况下会大大影响译文的可读性，甚至影响对译文主题的判断。例如，例 1 中，如果不知道“POTUS”是“President of the United States”，难以判别这句话主题是“政治外交”。

第二，网络非正规语言表达的不译。ICYMI 是“In Case You Missed It”的首字母缩略形式；“POTUS”是“President of the United States”的缩略形式；网络中类似的表达，例如“POTUS（President of the United States）”、“BTW（By the Way）”、“B4（Before）”。以上这些表达统称为网络非正规语言表达（Network Informal Language Expression, NILE）。网络非正规表达指应用于网络中的特殊语言形式，可以体现为单词、词组、短语、句子多种等多种表现形式，但其主要表现形式为单词和词组两种类型。

第三，命名实体的不译或错译。例 2 中的“NMCB”是“美国海军流动建筑营”，“133's Runnin' Roos”是其下属单位，在这一例句中都没有被翻译出。但是可以通过译文能猜测到这是一个命名实体，表示某一机构。这种不译的情况对译文的可读性影响不大。但是，例 2 中还有一个命名实体错译的问题，“Seabees”在此句中表示“海军工程兵”，但是被译成了“海蜂”。

2.2 基于军政领域社交媒体语料特征构建平行语料库

2.1 中所提到的三个问题中，“用户名、标签的不译”、“网络非正规表达的不译”可以通过预处理的方法基本解决，以提高译文质量。针对这两种问题，分别有以下解决方案。

第一，针对“用户名、标签的不译”问题，我们选择将用户名和标签切分为 Piece。

“@”和“#”的内容常见的有 5 种格式，如表 2 所示。

表 2 用户名和标签格式表

Tab.2 User name and label formats table

格式	举例
全大写	#MAGA、#HKIA、@USMC
全小写	@usairforce、@leezeldin
以首字母大写来分割	#WeeklyClimateShow、#ArmyValues 、@AlinejadMasih
大小写混合	@ USArmy、@TeamUSA、@USNavy
大小写混合且带缩略	@ SecDef、@PentagonPresSec

全大写格式通常是某一表达的缩略形式，机器翻译引擎能翻译出部分这一格式的标签或用户名。此外的四种格式，都是英语单词或英语单词的缩略形式的组合。对于人来说，可以很快地识别出这个用户名或标签是如何构成的，但是在机器翻译引擎中则选择忽略这一部分内容，选择不译，这也就导致译文不顺畅或无法理解。但是一旦把这些表达切割开，翻译引擎就会将其翻译。因为被切分后得到的既包括单词，也包括“Sec”、“Def”这样的缩略形式，我们统一称这种语言单元为 Piece。

所以，针对用户名、标签的不译问题，可以进行一个分词任务，不同于常见英文分词任

务的是，这个分词没有空格。这一分词任务类似于中文分词，所以我们构建一个常用词词表，将所有用户名和标签小写，然后匹配词表，得到最佳匹配后进行分割。

第二，在针对英语非正规语言表达上，采取自动识别然后替换其英文解释的方法，让翻译引擎能够准确翻译其含义。

首先，采用统计和规则融合的方法[9]（即互信息和信息熵相结合的统计方法，以及字母与数字混合出现、全大写字母等规则），识别出英语源文中的网络非正规语言表达，结合网络中所爬取的常用网络非正规语言表达，构建网络非正规语言表达库，将其表现形式和英文释义，存储到本地。然后，在调用翻译引擎的 API 前，对初始译文中存在的网络非正规语言表达替换为英语释义，例如“ICYMI”被替换为“In Case You Missed It”。表 3 中以一个例句展示了在进行以上两步的预处理前后调用百度翻译 API 得到的翻译结果。

表 3 预处理前后翻译对比表

Tab.3 Comparison table of translation before and after pretreatment

源文	#ICYMI: Twenty service members recently became U.S. citizens in a naturalization ceremony aboard @USSMidwayMuseum.
预处理前的翻译	#ICYMI：最近，在@USSMidwayMuseum 的入籍仪式上，20 名服役人员成为美国公民。
预处理后的翻译	#万一你错过了：20 名服役人员最近在@中途岛美国海军博物馆（USS Midway Museum）的入籍仪式上成为美国公民。

可以看出预处理后的翻译译文质量明显得到了提升。

针对以上这两种情况，设计基于翻译引擎构建平行语料库的方法，提升机器翻译引擎对军政领域社交媒体英语源文的翻译质量，并进行实验验证。

3 实验验证

译文质量评测（Quality Evaluation of Translation）主要从流畅度（Fluency）和忠诚度（Fidelity）[10]两个方面对机器翻译系统的得到的译文进行评价。其中，BLEU 值[11]是现在机器翻译系统评测的主流方法，主要用于考察译文的忠诚度。随着神经网络应用于机器翻译评测，提出了分布式表示评价度量方法，常用于考察译文的流畅度。

我们从忠诚度和流畅度上对以上方法的效果进行考察。首先采用 BLEU 值作为评判忠诚度的指标，然后利用百度开源的基于深度神经网络的语句流畅度计算方法作为评判流畅度的指标，对上述方法的有效性进行验证。并以此构建大规模军政领域社交媒体双语平行语料库。语料库构建流畅图如图 1 所示。

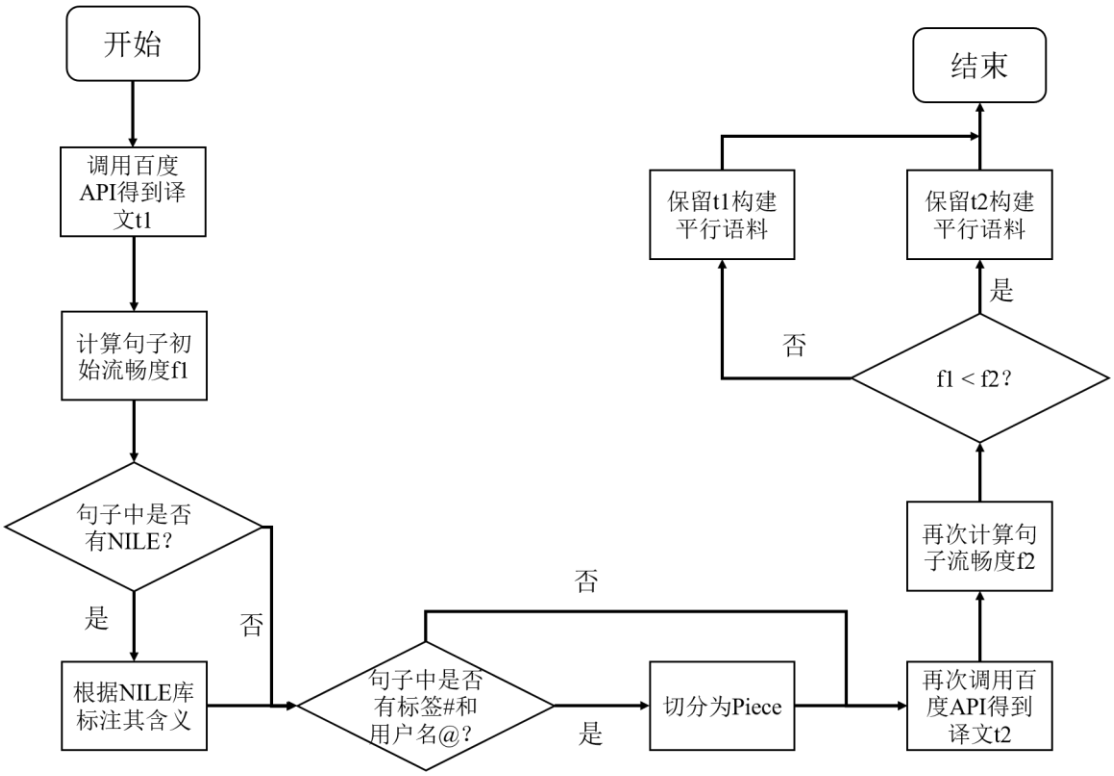

图 1 军政领域社交媒体英汉双语句对齐语料库构建方法流程图

Fig.1 Flow chart of English-Chinese sentence alignment corpus construction method for social media in military and political fields

对于输入的一个英语军政领域社交媒体源文，首先调用百度 API 对源文进行翻译，得到目标语句 t1，计算句子初始流畅度得到句子初始流畅度 f1，判断句子中是否有网络非正规语言表达，如果有，标注其英文解释；如果没有，继续判断句子中有无需要处理的“#”和“@”的内容，如果有，切分为 Piece，如果没有，则再次调用百度 API 对处理后的译文进行翻译，得到译文 t2，再次计算句子流畅度得到 f2。最后比较 f1 和 f2，保留流畅度值低的那一句目标语言作为源文的平行语料。

英文单语语料来源于 Twitter 和 Facebook 中重要军政人物和军政机构所发布的博客以及带有军政主题标签的群众博客。从约二十万句军政领域社交媒体源文中随机抽取 3000 余句，对源文进行预处理。因百度翻译计算句子流畅度的 DNN 模型对于语句长度有限制，所有筛选了部分长句，剩余 2923 句，并使用百度翻译引擎进行翻译。这 2923 句经过多个机器翻译引擎的初步翻译和两轮人工核验校对，得到了参考译文，以计算 BLEU 值。对未经过处理的目标语句、经过处理的目标语句和参考译文进行流畅度计算并将实验结果进行可视化展示。

经过实验，468 句源文经过处理后再次调用百度 API 进行翻译后的句子流畅度小于等于初始流畅度，即 f2 小于等于 f1，占样本总数的 16.01%。其中，有 53 句经过处理后，流畅度减幅超过 1000，占样本总数的 1.81%，占流畅度减少样本数的 11.32%。可以证明上述处理方法对获取平行语料具有一定效用。直方图和扇形图可以清晰地展现处理前后流畅度数据的分布情况。

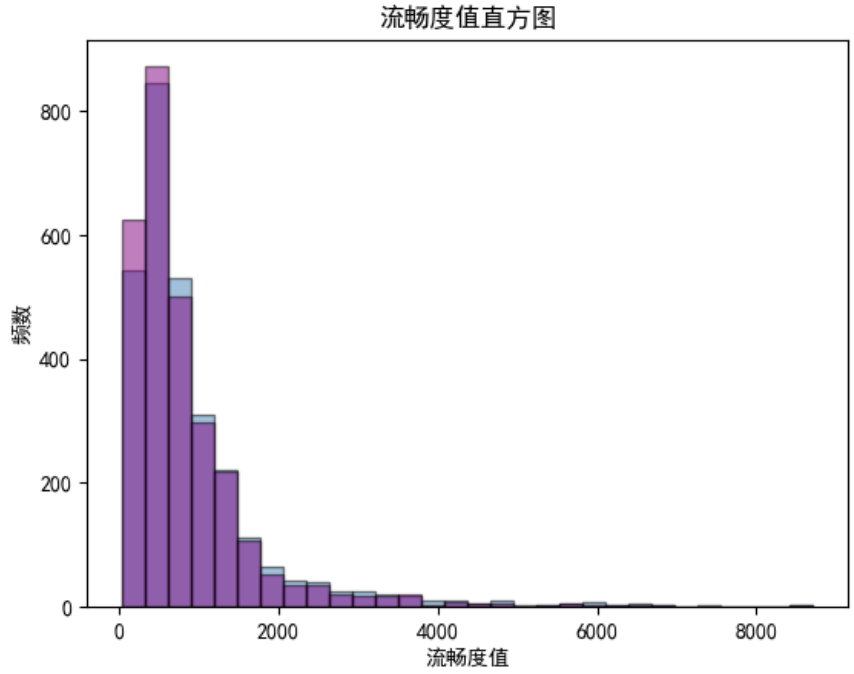

图 2 源文处理前处理后译文流畅度值数据直方图

Fig.2 Histogram of fluency value of translation after source language processing

图 2 中，蓝色表示处理前目标语句流畅度值直方图，紫色表示处理后目标语言流畅度值直方图。从图中可以看出，源文处理后得到的译文流畅度值的分布相较于处理前，值更集中于 2000 以下。可以用扇形图对源文处理前后以及参考译文流畅度进行对比。

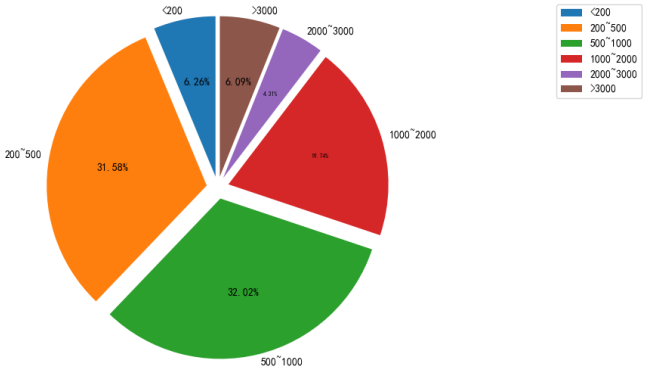

图 3 源文处理前译文流畅度值扇形图

Fig.3 Fan chart of translation fluency value before source language processing

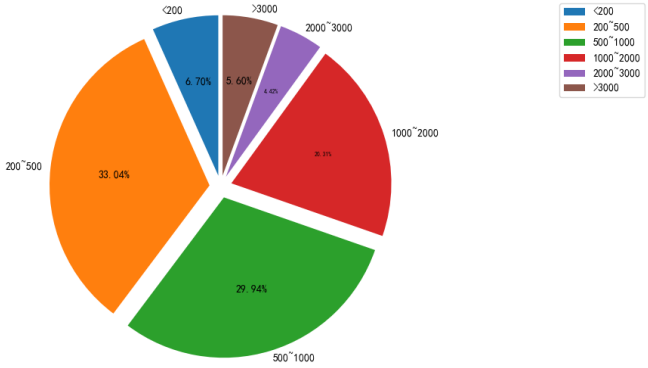

图 4 源文处理后译文流畅度值扇形图

Fig.4 Fan chart of translation fluency value after source language processing

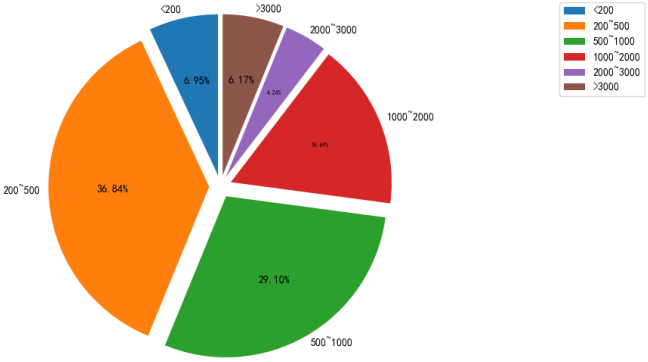

图 5 参考译文流畅度值扇形图

Fig.5 Reference translation fluency value fan chart

从扇形图中可以看出，源文处理前和处理后流畅度值低于 200 的占比提升了 0.44%，流畅度介于 200 至 500 的占比提升了 1.46%，证明在低流畅度值中，处理后的数据分布更加集中。并且，源文处理后的流畅度分布更接近于参考译文的流畅度分布。

对源文处理前的翻译结果和源文处理后的翻译结果，利用 Python 调用 sacrebleu 模块进行 BLEU 值的计算，在 3000 余句的测试数据上 BLEU 值提升了 1.37 个百分点，如表 4 所示。证明了以上该方法在流畅度和忠诚度上能够提升译文质量。并通过以上方法，利用单语语料构建了规模约为 20 万句对的军政领域社交媒体双语句对齐语料库。

表 4 预处理前后 BLEU 值对比表

Tab.4 Comparison table of BLEU value before and after pretreatment

源文是否进行预处理	BLEU 值
否	45.00
是	**46.37**

4 总结

本文首先总结平行语料库构建的研究情况，通过对翻译引擎翻译军政领域社交媒体所存在的问题，结合军政领域社交媒体语言特征，设计了基于单语语料的语料库构建方法，并通过实验验证了该方法的有效性，对军政领域社交媒体双语平行语料库的构建提供了一种新思路和新方法。

参考文献：

[1] 常宝宝,俞士汶.语料库技术及其应用[J].外语研究,2009(05):43-51.

[2] 李沐, 刘树杰, 张冬冬, 周明.机器翻译[M]. 高等教育出版社, 北京, 2018.

[3] 赵小曼. 英汉平行语料库句子级对齐研究及其在机器翻译中的应用[D].安徽大学,2010.

[4] Ziemski M, Junczys-Dowmunt M, Pouliquen B. The United Nations Parallel Corpus v1.0[C]// LREC 2016. 2016.

[5] 徐润华,王东波.一种面向 Web 的英汉平行语料库的构建方法[J].金陵科技学院学报(社会科学版),2021,35(04):51-56.DOI:10.16515/j.cnki.32-1745/c.2021.04.008.

[6] Ling W, Marujo L, Dyer C, et al. Crowdsourcing High-Quality Parallel Data Extraction from Twitter[C]// Proceedings of the Ninth Workshop on Statistical Machine Translation. 2014.

[7] Ling W, Marujo, Luís, Dyer C, et al. Mining Parallel Corpora From Sina Weibo and Twitter[J]. Computational Linguistics, 2016, 42(2):307-343.

[8] Gehring J, Auli M, Grangier D, et al. Convolutional Sequence to Sequence Learning[J]. 2017.

[9] 夏榕璟,张克亮.英语网络非正规语言表达的自动识别与术语库构建[J].中国科技术语,2022,24(01):36-44.

[10] Papineni K. BLEU: a method for automatic evaluation of MT[J]. Research Report, Computer Science RC22176 (W0109-022), 2001.

[11] Chen B, Guo H. Representation Based Translation Evaluation Metrics[C]//. Annual Meeting of the Association for Computational Linguistics, 2015:150-155.

Construction Method of English-Chinese Bilingual Parallel Corpus of Social Media in Military and Political Field

Rongjing Xia, Keliang Zhang*, Liang Tang, Ming Li

(Luoyang Campus, Information Engineering University, Luoyang 471003, China)

Abstract: The research of high-quality machine translation system is inseparable from high-quality parallel corpus. With the development of machine translation technology, the importance of parallel corpus research is becoming more and more prominent. The construction of social media corpus in military and political fields is of great significance for the language characteristics research and the task of social media machine translation in military and political fields. By analyzing the translation problems such as non-translation of user names and labels, non-translation of informal expressions on the Internet and other translation problems in the translation of social media corpus in the military and political field by existing machine translation engines, the characteristics of social media corpus

in the military and political field are extracted. The fluency was optimized, and the BLEU value was increased by 1.37 percentage points. Finally, a parallel corpus of English-Chinese sentence alignment for social media in the military and political fields with a scale of about 200,000 sentence pairs was constructed for machine translation tasks.
Keywords: parallel corpus; military and political fields; social media; machine translation