神经机器翻译军事领域英译汉评测及译后编辑研究

郭望皓[1*]，胡富茂[2]

（1. 战略支援部队信息工程大学 洛阳校区，河南 洛阳，471003；2. 洛阳理工学院 外国语学院，河南 洛阳，471003）

摘要：尽管神经机器翻译技术取得巨大进步，神经机器翻译系统正在加速推进实用化和商品化，但在垂直领域上的表现还不尽如人意。本研究以国内外主流机器翻译系统军事领域英汉文本翻译为研究对象，在自主构建的 1000 个测试数据集上，谷歌、百度、腾讯、网易有道、搜狗 5 家翻译系统的 BLEU 均值仅为 20.854，较之于通用语料相差超过 130%。实验结果显示，译文在拼写、词汇、句法和语义 4 大类 15 种共 5050 处错误中，军事术语翻译错误占比最高，为 42.83%；其次为普通词语误译和层级结构错误。实验结果表明，目前现有的神经机器翻译系统还不能实现高质量的军事文本翻译，无法满足现实需求，亟需进行译后编辑研究，提升军事文本翻译的准确率。

关键词：神经机器翻译；译文评测；译后编辑；军事文本；翻译错误类型

2013 年起，以神经网络为主要特征的神经机器翻译研究发展迅速，已经取代统计机器翻译成为目前机器翻译的主流范式。谷歌、微软、百度、腾讯等国内外行业巨擘纷纷上线自己的神经机器翻译系统，推动机器翻译向实用化和商品化不断迈进，甚至微软研究团队于 2018 年 3 月宣布其研发的机器翻译系统在通用新闻报道汉英翻译上“达到了人类水平”[1]。尽管如此，神经机器翻译仍然面临诸多挑战。其中一个重要方面就是目前市面上的主流神经机器翻译系统都是面向通用领域的，针对垂直领域专业题材的文本表现并不尽如人意。本文选取的英汉机器翻译引擎分别来自谷歌、百度、腾讯、网易和搜狗 5 家公司，这些产品均采用了目前流行的神经机器翻译技术作为自己英汉机器翻译的基本架构；作为国内外知名企业研发出来的产品，技术成熟、可靠性高，实验数据可重复；此外，它们都提供了英汉翻译的 API 接口，技术实现较为简单。不过，正是由于在垂直领域仅依靠机器翻译难以取得高质量的译文，因而译后编辑环节必不可少。本文以军事领域英汉神经机器翻译译文评测为切入点，通过整理归纳军事文本汉译文本错误类型，分析影响译文质量的因素，提出针对性的译后编辑建议。

1 相关研究综述

由于军事领域翻译任务需求量大，时效性要求高，各国都高度重视军事用途机器翻译的研发工作。Hutchins J.在讲述机器翻译发展史时指出，上世纪五六十年代，美苏两国军事情报需求正是机器翻译早期发展的动力[1]。Palmer M .等利用当时现成的技术模块组合成了一个机器翻译系统，专门用于英语到法语的战场信息翻译[2]。Holland M.等研发了一个能够快速将纸质文本扫面并翻译英文的原型系统，可为获取军事情报提供便利[3]。Jones D.等指出针对低资源和受限语言的机器翻译研究有着重要的军事意义和价值，并重点介绍了林肯实验室在语言机器翻译方面所作的探索性工作[4]。卡内基梅隆大学在美国国防部高级研究计划署

* Email：guowanghao@yeah.net

1 来源于微信公众号“微软研究院 AI 头条”（2018 年 3 月 15 日）。

（DARPA）的资助下，研发了可用于战场交流的语音机器翻译系统 DIPLOMAT（外交官），其中就包含有塞尔维亚-克罗地亚语、海地克里奥尔语、韩语等美军亟需的语种[5]（Frederking, R.等，2000）。此外，DARPA 还先后投入了“战术用途的口语沟通与翻译系统（TRANSTAC）”、“全球自主语言利用计划（GALE）”、“多语言自动记录分类分析和翻译（MADCAT）”、“紧急事件低资源语言计划（LORELEI）”等多个与机器翻译相关的研究项目，部分成果已经在驻阿富汗和伊拉克的美军列装，并应用于实战之中。

国内有关军事领域机器翻译的公开资料不多，主要可以分为以下三类：一是，从理论上指出发展军事领域机器翻译技术与装备的必要性与紧迫性[6-8]；二是比较军事领域文本翻译在不同机器翻译系统中的表现[9]；三是将机器翻译技术在军事领域的应用[10, 11]。

近些年，“机器翻译+译前译后编辑”已经成为翻译工作的一个重要模式，有关译后编辑的研究也呈井喷态势，其中魏长宏、张春柏、冯全功、崔启亮、刘明等人分别从译后编辑的概念、发展现状、人才培养、能力建设及未来趋势等方面进行理论探讨[12-15]，也有从某一个译后编译实践的出发，探讨译后编辑的方法[16]、作用[17, 18]、策略[19, 20]、适用性[21]以及与人工翻译的认知差异[22]等实证研究。文献整理发现，以机器翻译译后编辑作为选题的 MTI 硕士论文呈逐年增多的趋势，这也在另一个侧面反映出学界的关注。但是有关军事领域文本译后编辑的论文数量较少，许杰、程露蓝以 The Defense Industrial Base: Strategies for a Changing World 翻译实践为例，归纳总结了国防科技本文特征，提出了译后编辑的策略[23-24]。

从前人研究中可以看出，目前学界对于军事领域文本译后编辑关注程度不够高，此外，大部分研究对机器翻译背后的机制了解不够清楚，未能提出针对性强的译后编辑策略。本文将在错误分析统计的基础上，指出哪些错误是目前机器翻译难以克服的，哪些错误可以通过技术手段迅速加以解决，以期切实提高译后编辑的效率与质量。

2 军事领域英汉神经机器翻译译文评测实验设计

本研究通过提出研究假设，针对性地选取了谷歌等国内外 5 个主流的神经机器翻译系统作为实验对象，选取 1000 句军事领域文本作为实验材料和 1000 句通用领域文本作为对照材料，以国际通用 BLEU 值作为译文质量的评价标准，客观反映目前神经机器翻译系统在军事领域英汉翻译上的表现。在 BLEU 值测定的基础上，再通过人工手段针对机器翻译的错误进行分类统计。

2.1 研究问题

本研究要探求的两个问题是：（1）主流机器翻译系统军事文本的译文质量与通用文本之间有无显著差异；（2）军事文本译文的错误类型有哪些，呈现何种特点。

2.2 研究路径

本研究的采用的技术路线为：（1）构建实验数据集和对照数据集；（2）运用国内外主流的神经机器翻译系统对实验数据进行英汉机器翻译；（3）计算 BLEU 值；（4）对机器翻译错误进行分类统计。

2.3 研究数据集

为研究目前国内外主流机器翻译系统在军事领域英汉翻译上于通用领域有无显著差异问题，我们构建了两个数据集。实验数据集由军事领域英汉双语文本构成，对照数据集以新闻领域英汉双语文本为主。两个数据集的原始语料均来自于互联网，都属于开放型的数据，目的是最大限度保证实验的可比性。

2.3.1 中英双语语料的选择

互联网上有关军事领域的文本汗牛充栋，但高质量的英汉对照文本并不容易取得。经过仔细甄别筛选，实验数据集选取 *AirSea Battle: A Point-of-Departure Operational Concept*（汉译《作战概念起点》）、*The United States Army's Cyberspace Operations Concept Capability Plan 2016-2028*（汉译《美国陆军网络作战概念能力构想 2016-2028》）、*2014 Quadrennial Defense Review*（汉译《四年防务评估报告（2014 版）》）作为原始数据来源。这三篇报告的原始汉译译文来自于知远战略与防务研究所。

对照数据集我们选取了 VOA 双语新闻、中国日报双语新闻和经济学人双语对照文本中的部分内容。这些语料都是通过人工校对后才发到互联网上供人们阅读使用的，因此翻译质量是能够得到保证的。

2.3.2 语料预处理

由于互联网上下载的语料格式不统一，存在多余的空格、与研究内容无关的图表等，需要进行数据清洗。因此我们通过以下三个步骤对语料进行了预处理：（1）将所有文本转化为 TXT 格式；（2）去除文本中多余的空格、空行以及与研究任务无关的图表、符号等；（3）删除了原始文档中的目录及各种类型的注释。

双语语料句对齐

在完成原始语料预处理之后，我们对英汉双语文本进行了句对齐处理。本次研究选用了成都优译信息技术有限公司研发的单机版工具 Transmate 作为我们的对齐工具。这款工具对中文支持好，对齐准确率较高，个人版供免费下载使用。

2.3.3 构建研究数据集

由于原始数据较多，译文质量参差不齐，为了在有限的时间和精力内保证实验的效度与信度，我们对实验数据集和对照数据集进行了进一步的处理。一方面，采用随机采样的方法，利用计算机在军事文本实验数据集中抽取了 1000 个句长大于 10 个单词的英文句子，并请从事军事翻译工作的 2 名研究人员对这 1000 句汉译文本进行了校对，以保证翻译的准确专业。对于通用文本对照数据集也采用相同的方法提取了 1000 句，但未对译文再作修改。最后，对于已选取的双语对齐数据的对齐质量和文本格式又进行了人工核对，最终完成了研究数据集的构建工作。两个数据集的基本情况如下图所示。

表 1 实验数据集和对照数据集的基本情况

	原文（英语）		译文（汉语）	
	句子数	词数	句子数	字数
军事文本	1000	26799	1000	47643
通用文本	1000	25376	1000	44172

2.4 神经机器翻译系统进行翻译

我们将准备好的数据集导入 EXCEL 中，通过自编的 Python 小程序调用谷歌、百度、腾讯、网易有道和搜狗 5 家公司的在线机器翻译引擎对数据集中的英语原文进行英汉翻译，并将翻译结果返回显示在 EXCEL 中，以便后期进行数据分析。翻译结果如图所示。

	A	B	C	D	E	F	G	H
1		reference	source	google	baidu	tencent	youdao	sougou
53	56	同时，美军的通信	And US communic	美国通信，ISR和	美国通信、ISR和	而美国通信、ISR	和我们通信、ISR	美国通信、情报矿
54	57	这种连接能力则依	This connectivi	这种连接高度依赖	这种连接性高度依	这种连通性高度依	这个连接是高度依	这种连接高度依赖
55	58	网络空间也是如此	The same can be	关于网络空间也是	对于网络空间也可	对于网络空间可以	也是说，对网络空	网络空间也是如此
56	59	随着中国“反介入	The growing Chi	不断增长的中国A2	中国日益增强的A2	中国日益增强的A2	中国不断增长的A2	日益增长的中国A2
57	60	失去现实及虚拟领	Loss of forward	物理域和虚拟域（	物理域和虚拟域（	在物理域和虚拟域	损失在物理域和虚	物理领域和虚拟领
58	63	过去20年来，尽管	While the favor	虽然过去二十年来	尽管过去20年来西	尽管过去二十年来	虽然良好，稳定的	尽管过去20年西大
59	64	国防部继续侧重于	Thus DoD contin	因此，国防部继续	因此，国防部继续	因此，国防部继续	因此国防部继续强	因此，国防部继续

图 1 谷歌等 5 家机器翻译系统的翻译结果

2.5 BLEU 值计算

较之于人工打分，对译文质量进行自动评测具有速度高、成本低、不依赖于人的主观经验等优势。 BLEU （Bilingual Evaluation Understudy）是 IBM 公司提出的一种文本评估算法，一般用于计算机器翻译与专业人工翻译之间的对应关系，是目前被普遍采用的一种机器翻译自动评测指标。其核心思想是，利用 N-gram 和惩罚因子，对机翻译文和人工译文进行相似度计算，二者越相似，说明机翻译文的质量越高，BLEU 值越接近 1；反之，BLEU 接近于 0。在实际操作中，可以在 Python 中直接调用 NLTK 的 nltk.translate.bleu_score 工具包进行计算。本实验中，各项参数均选用默认值，数据平滑算法采用 Chen B.和 Cherry C.推荐的 method4[25]，通过计算最终得到的 BLEU 值数据。

本研究用于进行数据统计分析的工具是 IBM SPSS Statistics 22。

3 实验结果及讨论

3.1 总体情况

在基于 BLEU 值的自动评测中，BLEU 值越高，说明机器翻译的结果越接近人工译文。实验表明，BLEU 评测结果与人工评测结果具有高相关性[26]，因此能够在一定程度上反映机器翻译系统质量的优劣。两组数据的描述性统计结果显示，谷歌、百度、腾讯、有道、搜狗五个神经机器翻译系统针对实验数据集 1000 句军事文本的译文 BLEU 均值分别为 20.59、22.03、21.50、18.65、21.50，系统平均值为 20.85。同样的五个神经机器翻译系统针对对照数据集 1000 句普通文本的译文 BLEU 均值分别为 27.90、27.52、27.69、26.47、27.75，系统平均值为 27.47。实现结果证实，以目前市场上主流的机器翻译系统为代表的神经机器翻译在英汉通用领域文本翻译方面显著优于军事领域文本翻译，前者平均高出 6.62 个 BLEU 值，近 30 个百分点。

表 2 谷歌等五个机器翻译系统在军事领域文本上的表现（BLEU 值）

	N	Minimum	Maximum	Mean	Std. Deviation
谷歌	1000	5.2547	80.9107	*20.585884*	9.0886753
百度	1000	5.8297	90.2471	*22.025858*	10.2746282
腾讯	1000	4.0671	89.3962	*21.496574*	9.6914799
有道	1000	4.5323	80.9107	*18.649330*	7.3446064
搜狗	1000	5.4114	92.0530	*21.504140*	9.8810298

表 3 谷歌等五个机器翻译系统在通用领域文本上的表现（BLEU 值）

	N	Minimum	Maximum	Mean	Std. Deviation
谷歌	1000	10.3745	90.5521	*27.896913*	8.9847165
百度	1000	10.2752	92.3786	*27.521496*	9.2347316
腾讯	1000	11.6328	88.9214	*27.697215*	8.5932702
有道	1000	10.0211	89.7705	*26.465237*	8.9537496
搜狗	1000	10.7739	88.5076	*27.745608*	9.1370967

3.2 错误类型分析

尽管 BLEU 值为译文质量的优劣提供了良好的参考价值，但是仅依靠 BLEU 值并不能

给译后编辑带来更多的帮助，因此我们要对错误类型进行更加深入细致地分析。李梅、朱锡明等人的研究表明，深入研究分析机器翻译的错误类型，能够有效地提升译后编辑的效率[27]。

我们对所有机器翻译的结果进行了统计分析。首先，我们在借鉴 Vilar 等[28]，Farrus 等[29]，Kirchhoff 等[30]，Stymne 和 Ahrenberg[31]，Comelles 等[32]以及罗季美和李梅[33]，李梅和朱锡明[27]，罗季美[34]，刘艳丽[21]等人的研究基础上，从语言学角度对翻译错误类型进行了划分。具体而言，就是将错误类型分为拼写、词汇短语、句子句法和语义 4 个大类，每个大类下又分成若干个小类，一共 15 种错误类型。

表 4 机器翻译错误类型统计分析

错误大类	错误小类	数量	比例（%）
拼写错误	错别字	41	0.81
	标点符号	69	1.37
词汇短语错误	词性误译	178	3.52
	术语误译	2163	42.83
	缩略语误译	251	4.97
	俗语误译	0	0
	漏词	247	4.89
	多词	165	3.27
	普通词语误译	738	14.61
句子句法错误	断句错误	95	1.88
	层次结构错误	313	6.20
	语序错误	222	4.40
	时态错误	191	3.78
语义错误	逻辑关系混乱	248	4.91
	指代不清	129	2.55
（汇总）4	15	5050	100%

4.2.1 拼写错误

（1）错别字

由于构建神经机器翻译模型需要大规模的语料，使得译文的流利度较之于以前的统计机器翻译系统有了较大的提高，加之成熟的系统往往在译文生成后会利用语言模型等技术对译文进行纠错处理，所以一般来说，单纯的错别字错误还是比较少的。主要集中在“的”和“地”的错用方面，即在该用“地”的时候用了“的”。究其原因很可能是原始语料种就大量存在此类错误，在模型训练过程中，这类错误也被学习到并保留了下来。如：“unprovoked and unwarranted military buildup”译为“无端又毫无根据的提升军力（谷歌）”，应为“无端又毫无根据地提升军力”。

（2）标点符号错误

在机器译文中，有些标点符号的用法不符合中文表达的习惯。其中大部分是格式错误，如使用了半角标点以及单破折号。考虑到这类错误并不影响译文理解，尽管数量不少，但是在后期编辑过程中极易统一处理，故未纳入错误统计之中。纳入统计的错误主要是顿号和逗号之间的误用，以及引用他人说法时未使用冒号、引号，连词“和”前使用逗号的情况。总的来说这类错误也不算太多，属于行文不规范类，对语义理解没有造成太大障碍。

4.2.2 词汇短语错误

词汇短语误译是翻译中最为常见的错误，无论从数量上还是类型上，都占据着绝对多数。

（1）词性误译

词性误译是指将原本译为A词类的词译成了B词类，如将动词译为了名词。例如：“ability to project”指“（兵力的）投送能力”而非“项目能力（有道）”。这类错误时常出现，并且没有呈现出任何规律性，即使是同一机器翻译系统在不同句子中翻译同一短语结构也可能出现不同的结果。

（2）术语误译

军事文本与其他专业文本一样，都有大量的专业术语存在。因为与普通词语相比，术语的最大特点就是无歧义性，能够满足表义准确的要求，符合军事本文文体的要求。美国翻译协会（ATA）秘书长 Alan K Melby（转引自 Cheng J. 和 Min W.）将“术语”视为翻译“三脚架”（Tripod）之一[35]，由此可见，术语在翻译，尤其是专业领域翻译中的重要地位和作用。在我们的实验中，术语误译在 15 种错误类型中数量最多、占比最高。具体而言，可分为以下几个方面：

一是把军事术语译为普通词语，例如把“theater 战区”误译为“剧院（谷歌、百度、搜狗）”，把“logistics 后勤”误译为“物流（谷歌、百度、腾讯、有道、搜狗）”等。这样的例子不胜枚举。

二是有的军事术语未采用通用译法。例如：“Western Pacific Theater of Operations”，通用译法为“西太平洋战区”，有的系统翻译为“太平洋战场西部（搜狗）”；“AirSea Battle”，通用译法为“空海一体战”，有的系统翻译为“空海战役（百度、腾讯）”；这种情况不仅存在于军事术语上，也存在于包括组织机构名称在内的其他领域术语之中。例如将日本的“小笠原群岛”直接音译为“Bonin Islands 博宁群岛（谷歌、百度、腾讯）”。

（3）缩略语误译

在一些军事文本中，一些常用的缩略语往往默认读者了解其涵义而不写全称，这就对机器翻译造成了很大的挑战。如果模型训练过程中根本就没有接触过这种缩略形式，机器不可能给出正确的翻译，因为目前的机器翻译系统的推断能力非常弱，甚至可以说根本就不具备这种能力。例如有关美军的文本中经常出现的“C4ISR”，是“Command Control Communication Computer Intelligence Surveillance Reconnaissance”七个单词的缩写，即“指挥、控制、通信、计算机、情报及监视与侦察”，是美军开发的一个自动化指挥控制系统。机器翻译系统通常的做法是保留原缩略语形式保持不变，但有时也会出现丢失或者误译。如将“A2/AD（anti-access/area-denial）反介入/区域拒止”中的“AD”误译为“广告（有道）”，令人啼笑皆非。这就需要在译后编辑过程中多加注意。

（4）俗语误译

军事文本中时常会出现对名人名言、兵书著作的引用，如果不借助于已经建立好的双语对齐的俗语库，单纯依靠“不求甚解”的直接翻译，对于包括谚语、名言、歇后语等具有强烈历史文化内涵的语句，机器自动翻译可以说是“束手无策”。由于在本次实验中并未出现此类语句，故而这类错误为 0。

（5）漏词

由于神经机器翻译模型中缺乏“硬对齐”模块，所以导致译文中经常会出现漏译，即“该翻译的没翻译”，最后造成意义的不完整。例如：“a global power with global interests”，意思是“拥有全球利益的全球大国”，但是有系统将其翻译为“全球大国全球利益（有道）”，漏了表示修饰关系的引导词“with”，从而产生了错误。值得注意的是，有时系统在单独翻译某一子句或者某一短语时，漏译现象并不会发生，但当将这一子句或者短语放到句子之中再进行翻译时，就会有漏译现象出现。

（6）多译

这里的所谓多译，有两种情况。第一是指，把“不该翻译的给翻译了”。例如，“the Chinese PLA”翻译成“中国人民解放军”就可以了，即只反映“PLA”即可，前面的“the Chinese”

不需要再翻译一次。有的系统就会误译为“中国中国人民解放军（搜狗）”。第二种情况是指，将句子中的某个词语或者短语翻译了至少两次，例如：“…will almost certainly create downward pressure on both countries' defense budgets”的意思为“几乎肯定会对两国的国防预算造成下行压力”，有系统将“下行”翻译了两次，译为“几乎肯定会对两国的国防预算造成下行下行压力（百度）”。

（7）普通词语误译

普通词语误译也是一大类错误，这里的普通词语就是军事术语以外的词语。例如：“play important enabling roles”中的“enabling”可以翻译为“推动”或者“扶持”，译为“授权（腾讯）”显然是不合适的。

4.2.3 句子句法错误

句子句法类错误主要表现在以下五个方面。

（1）断句错误

对于超过某一定长的句子，机器翻译系统可能会从中进行截断，从而造成断句错误。例如：

原文：“The PLA's efforts are made all the more effective as the US defense program finds the bulk of the most stealthy US strike aircraft will be relatively short-ranged late-generation strike fighters carrying very small payloads of guided munitions, while US bombers, with their much greater payloads, are unlikely to be able to penetrate the PLA's robust IADS systems without considerable risk of loss.”

译文：“随着美国国防计划的实施，中国人民解放军的努力变得更加有效，因为美国最隐秘的攻击机大部分将是相对短程、携带非常小有效载荷制导弹药的晚一代攻击机，而美国轰炸机的有效载荷更大，不太可能穿透中国人民解放军。强大的 IADS 系统，不会有相当大的损失风险。（腾讯）”

译文在“PLA's”和“robust”之间进行了切分，把一个完整的子句分成了两个句子，导致了断句错误。

（2）层次结构错误

在未嵌入句法分析模块的神经机器翻译系统中，层级结构错误是一类常见的错误。例如：

原文：“Their main operating bases, ports and facilities have been largely invulnerable to serious conventional attack since World War II.”

其中的“since World War II”用于修饰全句，通常译作“自二战以来”并放在句首。有系统误将其作为“attack”的修饰语，译作“……严重的自二战以来传统的攻击（有道）”，显得混乱，令人感觉不知所云。此外，本句中的“conventional attack”应译为“常规攻击”而非“传统攻击”。

（3）语序错误

尽管英汉两种语言都是“SVO”语言，语序大体相同，但是二者在某些方面也存在差异。比如，汉语倾向于将修饰、补充部分前置，而英语修饰、补充部分后置的情况则更为常见。如果不改变语序，译文读起来就显得特别别扭。例如：“high-value Navy surface units, including carriers”通常译作“包括航母在内的高价值水面部队（或目标）”，但有系统译作“高价值水面部队，包括航母”（搜狗）。

（4）时态错误

英汉对于时态采用了不同的表现形式，纯粹的神经机器翻译系统中并没有时态分析生成模块，这就给时态错误的产生带来了可能。例如：

原文：“This state of affairs is almost certainly ending, with significant consequences for US security.”

译文：“这种状况几乎肯定是结局，对于美国的安全产生了重要影响。” （有道）

机器翻译的结果把将来时译成了过去时。

4.2.4 语义错误

语义错误多集中在以下两个方面：逻辑关系混乱和指代不清。

（1）逻辑关系混乱

以下面这句为例：

原文：“This is the key to maintaining the stable military balance that has preserved the peace in the Western Pacific.”

这句话的意思是：“这是保持军事态势平衡的关键，正是这一态势维持了西太平洋地区的和平。”不少系统将“maintaining the stable military balance 保持军事态势平衡”与“preserved the peace in the Western Pacific 维持西太平洋地区和平”相并列（百度、腾讯、有道、搜狗），造成整句逻辑关系错误。

（2）指代不清

指代不清是指在译文中的指示代词所指对象不明，导致意义含混。例如：

原文：“Consequently, the United States confronts a strategic choice: either accept this ongoing negative shift in the military balance, or explore options for offsetting it.”

这句话的最后一个单词“it”指代“ongoing negative shift”，如果简单翻译为“它”，就会造成指代不清，谷歌的译文“因此，美国面临一个战略选择：要么接受这种军事平衡持续地向负面转变，要么探索抵消它的方法”就存在这种错误。

4 提升军事领域文本英汉译后编辑效率的建议

通过上文的统计分析，针对军事文本英汉神经机器翻译的结果，我们提出如下四项建议，以期提升该领域英汉以后编辑的效率。

4.1 通过构建术语库增加术语翻译的准确性和一致性

上文统计显示，术语误译占全部错误的四成以上。经过分析不难看出，大部分出现误译的词语都是一词多译的，在一般的文本中多数时间使用的是普通义，而在军事文本中则多数作为术语使用。由于我们选用的机器翻译模型都利用大规模通用语料训练进行训练，自然就难以翻译出术语义，这也在一定程度上造成了神经机器翻译系统可移植性差的问题。军事文本专业性强、术语众多，如果能较好地解决术语翻译问题，势必能够大幅提升译文质量。

目前“机器翻译+译前译后编辑”的基本工作流程大致可分为以下几个阶段：原文预处理、构建术语库、机器翻译、译后编辑、校对定稿。其中构建术语库通常采用术语自动提取和人工提取相结合、提取后再进行人工翻译的方式进行。构建好的术语库既可以用于机器翻译，也可以用于译后编辑。这样看来，构建和维护一个较大规模的军事术语库显得尤为必要了。尤其在译后编辑阶段，利用军事术语库能够对译文进行批量修改，不仅提升了编辑效率，而且还能保证译文的统一性。

4.2 充分利用语法检查工具快速定位漏译、多译及断句错误问题

断句错误在各种类型的机器翻译系统中都会或多或少地存在，其主要原因是，目前机器翻译系统处理过长的句子还比较困难，一旦句子长度超过 50~60 个单词，系统性能急剧下降。因此，在训练模型时，通常会将训练语料文本句长限定在一定范围之内。系统使用过程中，面对过长的、超出系统处理范围之外的语句，就会依据某一条件进行断句处理，这时就有可能会触发断句错误。而漏译、多译问题通常会出现在神经机器翻译系统之中，这是神经机器翻译架构自身缺少“硬对齐”造成的。这三类错误的出现一般缺乏规律，难以预测。但是，

这三类错误一旦出现，往往导致整个句子不符合语法规则。这时，若能有效利用语法检查工具，如 Microsoft Word 中就有“拼写与语法”模块，就可以对此类错误进行快速定位，再通过译后人工编辑进行修订。

不过神经机器翻译译文的流利度好，有时即使多了一个词或者少了一个词都不会影响整句的可读性，尤其是漏译更为显著，对于语法检查工具未能发现的错误，就需要与原文进行仔细比对，切不可遗失关键信息。

4.3 开发符合实际需求的译后编辑工具

工具给人类带来效率的提升，翻译工作也概莫能外。一个优质的译后编辑工具能够实现快速句对齐、术语智能提示、语法错误自动标记等功能，还能够方便地进行保存和导出，并且具备可扩展性。还可以在此基础上开发具备自主学习功能的模块，当用户以特定的形式将某一规则输入编辑工具后，学习模块在遇到相同条件的问题时能够实现自动处理。目前军事翻译目前需求量巨大，如果能够联合专业译员一起合作开发符合军事翻译特点和需求的译后编辑工具，必然能够助力军事翻译工作效率与质量的提升。

4.4 设计研发军事领域英汉智能机器翻译系统

机器翻译模型的优劣直接决定译后编辑的难易程度，高质量的机器翻译势必能够减少译后编辑中无谓的工作，提高译后编辑的效率。尽管神经机器翻译已经给人类带来了足够多的惊喜，然而通过上文的实验不难看出，目前市面上主流的机器翻译软件在处理军事领域文本时还不能令人满意，亟需设计开发专门用于军事领域的神经机器翻译系统。从目前机器翻译发展的现实情况来看，人机结合的翻译策略应该是目前阶段的主流翻译方式，单纯依靠机器是不可行的，机器完全取代人类也不会是一时之功。因而，在设计开发军事领域英汉机器翻译系统时要突出“智能化”：设想中的这一翻译系统应是“MT（机器翻译）+CAT（计算机辅助翻译）”的综合平台：其中的 MT 部分是专为军事领域研发的机器翻译引擎，甚至可以是多引擎融合式的，能够输出若干翻译结果供使用者选择；而 CAT 部分则涵盖专门的军语辞典、军事术语库、双语实例库、翻译记忆库等模块，具备增、删、查、改等常用功能，满足用户的通用性及个性化需求。

5 结语

从本文的研究结果来看，目前国内外主流机器翻译系统在军事领域英汉文本翻译方面，与通用语料相比，仍有较大差距。以 BLEU 值计算，机器翻译系统在军事文本方面的表现低于通用语料 30 个百分点；从错误类型上来看，军事术语误译、普通词语误译和层级结构错误占前三位，超过错误总数的六成。据此，我们认为目前现有的神经机器翻译系统还不能实现高质量的军事文本翻译，无法满足现实需求，“机器翻译+译后编辑”的人机协同应该成为军事翻译的主要工作模式。在此基础上，我们建议应重视术语库在译后编辑中的作用，充分利用语法检查、双语对齐、术语提示等工具提高译后编辑的效率，开发能够满足实际需要的译后编辑工具，设计研发军事领域英汉智能机器翻译系统等，以期不断提高军事翻译工作的效率与质量，满足日益增长的军事翻译实际需要。

参考文献

[1] Hutchins J. Machine translation: A concise history[J]. *Computer aided translation: Theory and*

practice, 2007, 13(29-70): 11.
[2] Palmer M, Rambow O, Nasr A. Rapid prototyping of domain-specific Machine Translation systems[C]//*Conference of the Association for Machine Translation in the Americas*. Springer, Berlin, Heidelberg, 1998: 95-102.
[3] Holland M, Schlesiger C, Tate C. Evaluating embedded machine translation in military field exercises[C]//*Conference of the Association for Machine Translation in the Americas*. Springer, Berlin, Heidelberg, 2000: 239-247.
[4] Jones D, Shen W, Herzog M. Machine translation for government applications[J]. *Lincoln Laboratory Journal*, 2009, 18(1): 41-53.
[5] Frederking, R., Rudnicky, A., Hogan, C. et al. Interactive Speech Translation in the Diplomat Project[J]. *Machine Translation*, 2000,15: 27–42.
[6] 解国栋，易琼，朱斌. 现代军事情报处理方法中的语言、语音技术. 装甲兵工程学院学报[J]. 2006（03）：19-23.
[7] 刘明，彭天笑. 军事安全视角下的军队翻译能力建设. 国防科技[J]. 2018a（03）：32-36.
[8] 刘明，彭天笑. 军事翻译语言资源平台建设构想. 云梦学刊[J]. 2018b（02）：12-17.
[9] 张卉媛，杨士超. 谷歌和百度机器翻译系统对军事英语文本中句子翻译之对比研究. 科教文汇(上旬刊)[J]. 2019（12）：184-185.
[10] 鲍广宇，杨飞，刘晓明. 军事文本标图系统的设计与原型实现. 解放军理工大学学报(自然科学版)[J]. 2003（03）：30-34.
[11] 黄金柱，樊信展，李峰等. 基于军事平行语料库的人机结合翻译策略. 洛阳师范学院学报[J]. 2016（08）：56-61+67.
[12] 魏长宏，张春柏. 机器翻译的译后编辑. 中国科技翻译[J]. 2007（03）：22-24+29.
[13] 崔启亮. 论机器翻译的译后编辑. 中国翻译[J]. 2014（06）：68-73.
[14] 冯全功，崔启亮. 译后编辑研究:焦点透析与发展趋势. 上海翻译[J]. 2016（06）：67-74+89+94.
[15] 冯全功，刘明. 译后编辑能力三维模型构建. 外语界[J]. 2018（03）：55-61.
[16] 陈齐祖. 机器翻译结合译后编辑模式的科技类英译汉翻译实践报告[D]. 重庆大学.2014.
[17] 王萍. 机器翻译下预编辑和译后编辑在文史翻译中的作用[D]. 山东师范大学.2016.
[18] 郭高攀，王宗英. 机器翻译的译前与译后编辑在科技文本翻译中的探究. 浙江外国语学院学报[J]. 2017（03）：76-83.
[19] 张瑞雪. 机器翻译+译后编辑在英汉翻译中的使用[D]. 上海外国语大学.2018.
[20] 褚闽闽. 英汉机器翻译的译前和译后编辑策略[D]. 上海外国语大学.2019.
[21] 刘艳丽. “机器翻译+译后编辑”在不同文本类型中的适用性分析[D]. 上海外国语大学.2020.
[22] 周博. 译后编辑与人工翻译过程中认知努力的对比实证研究[D]. 广东外语外贸大学.2017.
[23] 许杰. 国防科技文本译后编辑实践报告[D]. 湖南大学.2018.
[24] 程露蓝. The Defense Industrial Base（第二章）汉译实践报告[D]. 湖南大学.2018.
[25] Chen B, Cherry C. A systematic comparison of smoothing techniques for sentence-level bleu[C]//*Proceedings of the Ninth Workshop on Statistical Machine Translation*, 2014.
[26] Doddington G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics[C]//*Proceedings of the second international conference on Human Language Technology Research*, 2002: 138-145.

[27] 李梅，朱锡明. 英汉机译错误分类及数据统计分析. 上海理工大学学报（社会科学版）[J]. 2013（03）：201-207.
[28] Vilar D, Xu J, D'Haro LF, Ney H. Error analysis of statistical machine translation output[C]// *Proceedings of 5th international conference on Language Resources and Evaluation (LREC 2006)*, 2006: 697-702.
[29] Farrús M, Costa-Jussà MR, Mariño JB, et al. Linguistic-based evaluation criteria to identify statistical machine translation errors[C]//*Proceedings of the 14th annual conference of the European Association for Machine Translation (EAMT 2010)*, 2010: 167-173.
[30] Kirchhoff K, Capurro D, Turner A. Evaluating user preferences in machine translation using conjoint analysis[C]//*Proceedings of the 6th conference of European Association for Machine Translation (EAMT-12)*, 2012: 119-126.
[31] Stymne S, Ahrenberg L. On the practice of error analysis for machine translation evaluation[C]//*Proceedings of the 8th international conference on Language Resources and Evaluation (LREC 2012)*, 2012: 1785-1790.
[32] Comelles, E., Arranz, V., &Castellón, I. Guiding automatic MT evaluation by means of linguistic features[J]. *Digital Scholarship in the Humanities*, 2017, 32(4): 761-788.
[33] 罗季美，李梅. 机器翻译译文错误分析. 中国翻译[J]. 2012（05）：84-89.
[34] 罗季美. 机器翻译句法错误分析. 同济大学学报(社会科学版)[J]. 2014（01）：111-118+124.
[35] Cheng J, Min W. A comparative study of term extraction methods in translation[C]// Chan, S (Ed.) . *The Human Factor in Machine Translation*. London; New York: Routledge Taylor & Francis Group, 2018: 64-82.

A Study on Assessment of Translations and Post-Translation Editing in Neural Machine Translation

Abstract: Although great progress has been made in neural machine translation technology and neural machine translation system is accelerating its practicality and commercialization, its performance in the vertical field is still not satisfactory. In this study, English-Chinese text translation in the military field of the mainstream machine translation system at home and abroad is taken as the research object. On the 1000 test datasets independently constructed, the BLEU average value of the five translation systems—Google, Baidu, Tencent, NetEase Youdao and Sogou—is only 20.854 with a difference of more than 130% compared with the general corpus. The experimental results indicate that among the 5050 errors in 15 categories of spelling, vocabulary, syntax and semantics, the military term translation errors account for the highest proportion (42.83%), followed by common word translation errors and hierarchy errors. The experimental results show that the existing neural machine translation system can not achieve high-quality military text translation and it cannot meet the actual needs, so it is urgent to conduct post-editing research to improve the accuracy of military text translation.

Keywords: Neural Machine Translation; Assessment of Translations; Post-Editing; Military Texts; Translation Error Types