

人工智能、机器与人工翻译的语言特征

——以小说《活着》译本为例

秦萌*, 王宪

上海海事大学外国语学院, 上海

收稿日期: 2023年10月8日; 录用日期: 2023年11月9日; 发布日期: 2023年11月17日

摘要

在AI不断发展的时代, ChatGPT作为大型交互式语言模型, 具有高智能化、交互性特点, 被广泛用于自然语言处理, 给人工翻译和机器翻译带来了冲击和挑战。本研究以中文小说《活着》为例, 使用CLAWS7、WordSmith和AntConc语言研究工具, 对小说《活着》的AI翻译、人工翻译与机器翻译(有道翻译)这三个译本进行研究, 旨在发现三个译本的语言特征, 特别是词汇特征。研究发现, 1) ChatGPT翻译词汇量最大、变化大, 阅读难度大, 句子所承载信息最多; 2) 人工翻译用词更符合英语特点, 对文化负载词的翻译最准确; 3) 人工翻译人名前后一致, 采用拼音翻译, 有道和ChatGPT采用拼音翻译或直译, 前后翻译不一致。机器翻译和人工翻译是相辅相成的, 两者不可能被其中一方取代, 人类应借助机器来实现更高效的翻译。

关键词

机器翻译, 人工翻译, 《活着》, ChatGPT

The Linguistic Features of Artificial Intelligence Translation, Machine Translation and Human Translation

—A Case Study of the Chinese Novel *To Live*

Meng Qin*, Xian Wang

College of Foreign Languages, Shanghai Maritime University, Shanghai

Received: Oct. 8th, 2023; accepted: Nov. 9th, 2023; published: Nov. 17th, 2023

*通讯作者。

Abstract

In an era of continuous AI development, ChatGPT, as a large interactive language model, possesses high intelligence and interactivity characteristics, and it is widely used in natural language processing, posing both impact and challenges to human and machine translation. This research, using the Chinese novel *To Live* as an example, primarily explores the linguistic features of AI translation compared to human translation and machine translation (e.g., Youdao translation). The study found that: 1) ChatGPT translation has the richest vocabulary and variation, making it more challenging to read, while it carries the most information within sentences; 2) Human translation uses words more in line with English characteristics and provides the most accurate translations for culturally loaded terms; 3) Human translation maintains consistency in translating names, employing phonetic translations, while Youdao and ChatGPT translations use phonetic translation or direct translation, resulting in inconsistent translations. Machine translation and human translation complement each other, and neither can entirely replace the other. Humans should leverage machines to achieve more efficient translation processes.

Keywords

Machine Translation, Human Translation, *To Live*, ChatGPT

Copyright © 2023 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

传统的机器翻译是指使用计算机将一种语言符号换成另一种语言符号的过程[1]。而随着科学技术和人工智能的发展,从基于规则的机器翻译,到统计机器翻译,再到神经机器翻译,在这一过程中计算机的能力从将不同的符号通过建立的规则对应起来,拓展到了从语料库中获取语言风格再加以模仿[2]。目前,翻译主体已经发生了变化,机助人译逐渐走向人助机译[3]。例如,ChatGPT这类交互式自然语言处理工具的出现,使得机器翻译进一步进入智能领域,通过人机互动完成翻译、校对、编辑等任务[3],能够呈现出相当完整、甚至高质量的译文,人们难以区分译文的来源是机器翻译还是人工翻译[4]。然而,机器翻译的快速发展并不意味着翻译可以完全脱离人译。机器翻译主要考虑不同符号之间概念意义和语篇意义的对应,很少涉及情感、价值观等人际意义[1]。而人类在社会生活中所积累的情感、经验等,使得人工翻译能够更好地体现出作者的情感和言外之意。因此,为了探索机器翻译的人工翻译的差异,本文以小说《活着》为例,通过建立语料库研究机器译本和人工译本的语言特征和词汇层面的差异。

2. 文献综述

随着技术的不断更新和进步,学者开始研究基于不同模型或规则的神经机器翻译,如融合依存句法和LSTM[5]、字符级语言建模[6]。翻译技术的进步对原本的翻译实践行为产生了影响。除了译者外,翻译机构和国家主体的作用逐渐显现出来[7],翻译主体不再以译者为核心[3]。这样高效的机器翻译翻译很快就进入了应用研究领域,有学者研究海图英语地名的机器翻译[8]。还有一些学者们开始对比研究机译和人译的语言特征差异,以探求是否存在一种更好的翻译方法。李奉栖[9]对比了英语专业学生翻译和神

经网络翻译的质量差异。Almahasees [10]对比人机翻译后总结了谷歌和必应的优缺点。虽然机器翻译已经表现出一些优势,但在法律翻译中依然面临着诸多困难[11]。此外,目前的人机翻译语言特征对比研究集中于研究神经机器翻译和人工翻译的对比,如使用谷歌机器翻译研究英语关系从句[12]、探索 SYSTRAN 系统对中医英语的翻译情况[13]、以《傲慢与偏见》为例的在线机译和人译语言特征对比[14]等。

可以看出,上述研究都从各自角度出发,对不同文体内容的语言特点进行了探索,学者们多关注于神经机器翻译研究,但是,人工智能技术下所诞生的交互式工具 ChatGPT 还未被广泛研究。因此,本研究以小说《活着》为语料库,对比研究人工翻译和两种机器翻译的语言特征。

3. 研究方法

3.1. 语料库建立

本文以《活着》的前三章为研究对象,人工翻译选取 Michael Berry (MB)的英译本,机器翻译选取 ChatGPT 和有道翻译为翻译工具,分别建立中文原文语料库、MB 英文译文语料库、ChatGPT 英文译文语料库和有道英文译文语料库。ChatGPT 和有道翻译虽同为辅助工具,但是两者本质上还是存在不小的差异。有道翻译是中国学生所熟悉并且常使用的神经机器翻译工具,所承载的语料库是主要基于中文语料库,而 ChatGPT 则是采用人工智能技术,基于国外语料库的高智能化、交互性的自然语言处理工具。总之,两种工具在语料库和功能上都有着较大差异。因此,本研究选取两种翻译工具和人工翻译进行对比。由于两种机器翻译的篇幅限制,有道翻译以每一章为单位进行翻译,ChatGPT 以最大限制内语篇意义完整的篇幅为单位进行翻译。

3.2. 研究方法

本研究使用 CLAW7 [15]、WordSmith 和 AntConc 对所选语料进行处理。首先使用 CLAW7 对三个英文译本进行在线词性标注。其次,使用 WordSmith 对英译本进行分析,得出形符、类符等数据。最后,使用 AntConc 制作三个英译本的 word list,再加以具体分析。

4. 结果与讨论

4.1. 词汇丰富程度

使用 WordSmith 和 AntConc 对三个英译本进行对比分析,得到以下计量语言特征。

如表 1 所示,在总字数上,中文原文字数为 25,333,三个英译本的字数相较于原文都相差了不少,从高到低为 20,288 字、16,912 字和 14,593 字。这是因为英语和中文属于不同语系,英语中的一些语法或句式,在汉语中需要借助词汇来实现同样的效果,这会增加汉语的字数[16]。而人工译本和两个机器译本差异也较为显著,在人工翻译时,译者为了使英语母语读者更容易理解,会增加阐释等。而机器翻译受制于原文、算法等,大多只能呈现出与原文的对应关系[14],导致译文字数比人工译文低。此外,虽然人工译本的字数最多,但是从标准化形符类符比(STTR)来看,ChatGPT 的 STTR 最高。STTR 通常用来衡量语料库的词汇变化,STTR 越大就表明语料库不同的词汇量越大,阅读的难度就越高。

除 STTR 外,译本的单词、句子长度也可以体现出译本的词汇丰富程度和阅读难度(表 2、表 3)。在表 2 中,虽然 MB 译本的平均单词长度要小于 ChatGPT 译本,大于有道译本,但是结合表 3 来看,MB 对长单词的使用要比两个机器译本频繁。而有道译本,不论是在平均词长还是平均句长上,都与 ChatGPT 和 MB 译本有差异,其词汇变化小,阅读难度要低于另外两个译本。

Table 1. STTR
表 1. 标准化形符类符比

	MB	ChatGPT	有道
形符数	20,288	14,593	16,912
类符数	2772	2351	2248
类符/形符比	13.66%	16.11%	13.29%
标准化类符/形符比	40.67%	43.20%	38.08%

Table 2. The average length of words and sentences
表 2. 平均词长、句长

	MB	ChatGPT	有道
平均单词长度	4.12	4.24	3.86
句子总数	1135	817	1146
平均句子长度	17.87	17.86	14.76

Table 3. The number of letters in a word
表 3. 单词字母数

		2-letter words	3-letter words	4-letter words	5-letter words	6-letter words	7-letter words	8-letter words	9-letter words	10-letter words
ChatGPT	676	2383	3017	3036	1806	1466	1127	575	253	136
MB	1053	3781	4458	3992	2319	1722	1448	740	375	232
有道	1135	3174	4024	3489	1902	1391	1001	448	202	81

4.2. 词汇密度

为研究三个英译本的词汇密度, 本文使用 CLAWS7 的词性标注形式, 将名词 (ND1|NN|NN1|NN2|NNA|NNB|NNL1|NNL2|NN0|NN02|NNT1|NNT2|NNU|NNU1|NNU2|NP|NP1|NP2|NPD1|NPD2|NPM1|NPM2)、动词(VV0|VVD|VVG|VVGK|VVI|VVN|VVNK|VVZ)、副词 (RR|RRQ|RRQV|RRR|RRT|RT)、形容词(JJ|JJR|JJK)筛选为实词, 从而计算其词汇密度, 结果见表 4。

Table 4. Lexical density
表 4. 词汇密度

	MB	ChatGPT	有道
名词	3949	2905	3422
形容词	915	645	704
副词	1032	881	694
实义动词	3380	2520	2675
实词总数	9276	6951	7495
虚词总数	11,012	7642	9417
词汇密度	45.72%	47.63%	44.31%
实词/虚词比	84.23%	90.95%	79.59%

实词是指具有稳定词义的词语, 包括名词、动词、形容词和副词等词类[16], 词汇密度可以反映文本信息量, 词汇密度越大, 实词就越多, 句子所承载的信息就越多。从表 4 可以得出, ChatGPT 译本的词汇密度最大, 其句子所承载的信息最多, 其次是 MB 译本。

4.3. Word List 对比

本研究使用 AntConc 建立三个英译本的 Word List 进行对比。采用 lemma list 将词性还原, stop list 提出高频次、低价值的词语, 从而更准确地得出三个译文的结果, 见表 5。

Table 5. Top 20 words in frequency
表 5. 使用频次前 20 的单词

序号	MB	ChatGPT	有道
1	say 129	father 89	say 212
2	Jiazhen 113	say 80	father 125
3	see 94	mother 64	not 115
4	get 92	see 64	go 100
5	go 90	walk 59	see 90
6	old 81	go 57	mother 87
7	just 79	Jiazhen 57	come 80
8	hand 77	old 54	Jiazhen 72
9	come 75	back 50	hand 66
10	father 75	look 50	look 64
11	Long Er 74	ask 49	old 64
12	dad 73	there 49	man 62
13	there 70	time 49	time 59
14	back 68	hand 47	back 57
15	then 68	leave 45	walk 57
16	take 64	saying 43	Fengxia 54
17	time 60	then 43	ask 53
18	didn't 59	day 42	day 51
19	man 59	man 40	get 49
20	walk 59	come 39	know 49

从整体上来看, 三个译本使用频率最高的前 20 个单词大多是相同的, 主要差异集中于动词的使用上。以下将主要举例对三个译本的动词使用情况进行说明(表 6)。

Table 6. Top 3 verbs in frequency
表 6. 三个译本使用频率前 3 的动词

	1	2	3
MB	say, said, says (129)	see, saw, seeing, seen (94)	get, gets, got, getting, gotten (92)
ChatGPT	say, said (80)	see, saw, seeing, seen (64)	walk, walked, walking (59)
有道	say, said, says (212)	go, goes, going, gone, went (100)	see, saw, seeing, seen (90)

在小说《活着》中包含着大量的对话, 常出现“说”、“说道”等动词, 因此在英译本中英文动词“say”出现频率是最高的。其中, 有道译文中动词“say”的使用频次是三个版本中最高的, 这和不同译本的翻译准确度有关。在一些中文语境下, “说”并不能直接对应为“say”, 而是可以表示一种状态。如以下例子:

我听他们不是说死了, 就是说前两天还见过。

It seemed as if everyone they asked about was either dead or someone had just seen them within the last few days. (MB)

From what I overheard, they either talked about those who had died or mentioned having seen them just a couple of days ago. (ChatGPT)

I heard them not say dead, that is to say two days ago also met. (有道)

有道翻译直接将原文中的“说”翻译为“say”, 显然并没有另外两个译文更加符合原文的真实意思。原文可以理解为“我听他们说不是死了, 就是前两天还见过。”, 这里的“说”是“听说”的意思, 而“say”更倾向于向某人传达某种信息的动作。此外, 就整句话的翻译而言, 有道是字对字的直译, 并不准确。

在前 20 名高频词表中, 三个译本中另一差异较大的动词是“get”。MB 译本中使用了 92 次, ChatGPT 译本中使用了 38 次, 有道译本中使用了 49 次。在英语中, “get”是一个非常常见的动词, 特别是在非正式口语和写作中十分常见[17]。尽管“get”看起来是一个很简单的单词, 但是它的用法却非常复杂, 可以和形容词、名词/代词、宾语、副词、谓语介词和描述地点/时间/事物的词语一起使用, 或者可以用在习语中[18]。MB 译本中, “get”的搭配有“get to v.”、“get adv. (get up, get out)”、“get adj.”等。

“再说我爹年轻时也和我□样, 我家祖上有两百多亩地, 到他手上□折腾就剩□百多亩了。”

Our family used to have over two hundred mu of land, but once my father got his hands on it he managed to lose over half. (MB)

My family used to have over 200 acres of land, but he managed to reduce it to just over 100 acres. (ChatGPT)

My family had more than 200 mu of land in my ancestors, which would be left over 100 mu after he had done something about it. (有道)

对例句中“到他手上一折腾”的翻译, 三个译本各不相同。结合上下文“到他手上一折腾”的意思是由于“我爹”管理不当(变卖家产用于玩乐), 家里的地少了一半。MB 翻译使用短语“get one’s hands on”、“manage”和“lose”来表达原文意思。短语“get one’s hands on”将“手”这一概念翻译了出来, 和“manage”都有管理的意思, 而“lose”则是解释了家里的地是赌博输掉了的原因。相比之下, ChatGPT 和有道都没有将“手”翻译出来, 只把“折腾”翻译了出来, 且使用的动词贬义程度不如 MB 翻译。

4.4. 文化负载词

语言是文化的一部分, 是文化的载体, 人类社会独特的文化传统在语言中得以体现[19]。翻译作为一种跨文化交流活动, 涉及到使用一种语言准确表达另一种语言意义的过程。而东西方较大的文化差异给文化负载词的翻译带来了不少困难。本文所研究的语料库中也有不少关于文化负载词的翻译。

年过花甲的私塾先生对我爹说: “你家少爷大了准能当个二流子。”

The next time he saw my father, my teacher, who was really getting on in years, told him, “I guarantee you that when

that son of yours grows up, he'll be nothing but trouble.” (MB)

The elderly private school teacher said to my father, “Your young master will probably end up as a mediocre person.” (ChatGPT)

The elderly private school teacher said to my father, “Your young master will probably end up as a mediocre person.” (有道)

“二流子”指整天无所事事，到处惹是生非的人，而 ChatGPT 和有道翻译中的 “a mediocre person” 更倾向于指普通、没有成就或者出色的表现的人，显然没有 MB 译本更加准确。

我是徐家仅有的一根香火，我要是灭了，徐家就得断子绝孙。

I'm the only flame the Xu family still has burning. If I'm extinguished, the Xu family will be finished.

I was the only hope of the Xu family. If I failed, the Xu family would be finished.

I am the only incense in the Xu family. If I go out, the Xu family will have no children.

MB 译文将香火燃烧的状态描述了出来，有道译文更倾向于直译，而 ChatGPT 的翻译并不能准确地表达原文意义，“香火”指家中的男孩，并不一定能有所成就，成为家里的希望。对于“断子绝孙”，有道翻译更加直白一些。

4.5. 人名翻译

中国人名、地名的英译策略经历了很长时间的发展，从罗马音译到拼音翻译，中国人名、地名不统一的翻译策略，增加了翻译成本和难度，不利于中国与别国交流[20]。随着中国国际地位、经济实力的提升，中国文化的不断推广，中国人名、地名的拼音翻译逐渐成为了主流。

小说《活着》的前三章共出现了不少人名，福贵、二喜、有庆、家珍、凤霞、长根、龙二、老全、春生等。MB 译本将所有人名按拼音翻译处理，如福贵(Fugui)、家珍(Jiazhen)、长根(Changgen)、龙二(Long Er)等，而 ChatGPT 和有道对于相同人名的翻译会出现拼音翻译或直译现象，导致上下文译名不一致。例如，长根(Long root)、龙二(Long two, Dragon two)、春生(Spring-born)等，这就会提升阅读难度。

人名翻译没有一定的标准，并非都要翻译成拼音。外国人对拼音系统的不熟悉，也可能会增加阅读难度。相反，一些人名通过直译也能更好地传达名字本身的含义，如春生被译为“Spring-born”可以比翻译成拼音更好地传达含义。

5. 结语

本文以《活着》前三章的三个英译本建立语料库，通过语料库检索工具 AntConc 和 Wordsmith 对其进行对比研究，旨在探究其语言特征和差异。研究发现：1) 人工翻译篇幅最长，但 ChatGPT 翻译词汇量最大、变化大，阅读难度大，句子所承载信息最多，有道翻译用词单一；2) 人工翻译用词更符合英语特点，对文化负载词的翻译最准确 3) 人工翻译人名前后一致，采用拼音翻译，而有道和 ChatGPT 采用拼音翻译或直译，前后翻译不一致。

尽管机器翻译目前无法和人工翻译相媲美，其快速发展已经让人们担忧未来替代人工翻译的可能性。然而，即使是像 ChatGPT 这样的工具的出现，机器翻译也不会代替人工翻译，其功能的实现来自于对人类的模仿，想让它真正理解人类语言还存在很大的困难。因此，我们应该抱有积极的态度，加以利用，从而为人类提供更多的便利。“人工翻译 + 机器翻译(人工智能)”是未来翻译发展的方向，有利于打破语言壁垒，高效且准确地为跨文化交流提供便捷。

参考文献

- [1] 胡开宝, 李翼. 机器翻译特征及其与人工翻译关系的研究[J]. 中国翻译, 2016, 37(5): 10-14.
- [2] 冯志伟, 张灯柯, 饶高琦. 从图灵测试到 ChatGPT——人机对话的里程碑及启示[J]. 语言战略研究, 2023, 8(2): 20-24.
- [3] 胡健, 范梓锐. 机器翻译视角下的翻译本质[J]. 当代外语研究, 2023(2): 90-96.
- [4] Popel, M., Tomkova, M., Tomek, J., Kaiser, L., Uszkoreit, J., Bojar, O. and Žabokrtský, Z. (2020) Transforming Machine Translation: A Deep Learning System Reaches News Translation Quality Comparable to Human Professionals. *Nature Communications*, **11**, Article No. 4381. <https://doi.org/10.1038/s41467-020-18073-9>
- [5] 郑鑫, 陈海龙, 马玉群, 等. 融合依存句法和 LSTM 的神经机器翻译模型[J]. 哈尔滨理工大学学报, 2023, 28(3): 20-27.
- [6] 胡泽林, 高翊, 李淼, 等. 基于字符级语言建模的汉蒙神经机器翻译方法研究[J]. 昆明理工大学学报(自然科学版), 2023, 48(3): 85-92.
- [7] 仝亚辉. 翻译技术时代下的翻译主体性研究[J]. 解放军外国语学院学报, 2022, 45(6): 132-140.
- [8] 乔晶, 李鹤元. 一种海图英语地名机器翻译方法[J]. 海洋测绘, 2022, 42(5): 73-77, 82.
- [9] 李奉栖. 人工智能时代人机英汉翻译质量对比研究[J]. 外语界, 2022(4): 72-79.
- [10] Almahasees, Z.M. (2017) Assessing the Translation of Google and Microsoft Bing in Translating Political Texts from Arabic into English. *International Journal of Languages, Literature and Linguistics*, **3**, 1-4. <https://doi.org/10.18178/IJLL.2017.3.1.100>
- [11] 张法连. 法律翻译中的机器翻译技术刍议[J]. 外语电化教学, 2020(1): 53-58+8.
- [12] 贺文照, 李德凤. 英语关系从句机译汉语评价——以谷歌机器翻译为例[J]. 中国科技翻译, 2019, 32(3): 30-34.
- [13] 屈亚媛, 周玉梅. 机器翻译还是人工翻译?——浅析《黄帝内经·素问》双字格养生术语机译错误人工评测[J]. 医学争鸣, 2016, 7(4): 50-53.
- [14] 蒋跃, 张英贤, 韩纪建. 英语被动句人机翻译语言计量特征对比——以《傲慢与偏见》译本为例[J]. 外语电化教学, 2016(3), 46-51, 63.
- [15] University Centre for Computer Corpus Research on Language of Lancaster University (n.d.) Free CLAWS Web Tagger. <https://ucrel-api.lancaster.ac.uk/claws/free.html>
- [16] 王克非. 语料库翻译学探索[M]. 上海: 上海交通大学出版社, 2012.
- [17] Cambridge University Press & Assessment (n.d.) Get. English Grammar Today. Cambridge Dictionary. <https://dictionary.cambridge.org/us/grammar/british-grammar/get>
- [18] Pusparini, N.M.D.U., Fitriari, D.A.C., Kasni, N.W. and Susanthi, I.G.A.A.D. (2022) The Use of the Word “Get” in English. *KnE Social Sciences*, **2022**, 383-391. <https://doi.org/10.18502/kss.v7i10.11307>
- [19] Zhou, X. and Hua, Y. (2021) Culture-Loaded Words and Translation Equivalence. *Theory and Practice in Language Studies*, **11**, 210-215. <https://doi.org/10.17507/tpls.1102.14>
- [20] Qu, W. and Li, R. (2015) Translation of Personal and Place Names from and into Chinese in Modern China: A Lexicographical History Perspective. *International Journal for the Semiotics of Law*, **28**, 525-557. <https://doi.org/10.1007/s11196-015-9414-0>