A Deep Analysis of the Impact of Multiword Expressions and Named Entities on Chinese-English Machine Translations

Huacheng Song[1,2], **Hongzhi Xu**[2]
[1]The Hong Kong Polytechnic University
[2]Shanghai International Studies University
huacheng.song@connect.polyu.hk, hxu@shisu.edu.cn

Abstract

In this paper, we present a study on the impact of so-called multiword expressions (MWEs) and multiword named entities (NEs) on the performance of Chinese-English machine translation (MT) systems. Built on an extended version of the data from the WMT22 Metrics Shared Task (with extra labels of 9 types of Chinese MWEs, and 19 types of Chinese multiword NEs) which includes scores and error annotations provided by human experts, we make further extraction of MWE- and NE-related translation errors. By investigating the human evaluation scores and the error rates on each category of MWEs and NEs, we find that: 1) MT systems tend to perform significantly worse on Chinese sentences with most kinds of MWEs and NEs; 2) MWEs and NEs which make up of about twenty percent of tokens, i.e. characters in Chinese, result in one-third of translation errors; 3) for 13 categories of MWEs and NEs, the error rates exceed 50% with the highest to be 84.8%. Based on the results, we emphasize that MWEs and NEs are still a bottleneck issue for MT and special attention to MWEs and NEs should be paid to further improving the performance of MT systems.

1 Introduction

Evaluating machine translation (MT) systems on various fine-grained linguistic phenomena has become a trending practice (Manakhimova et al., 2023; Song et al., 2024, etc.). Multiword expressions (MWEs), making up approximately half of the lexicon (Jackendoff, 1995; Fellbaum, 1998), are shown to be an intractable problem to various tasks in the realm of natural language processing across different languages due to their idiosyncrasies in syntax and/or semantics (Sag et al., 2002; Rayson et al., 2010; Constant et al., 2017). Such an idiosyncratic nature of MWEs is even crucial for MT since the translation systems rely on the complete understanding of the MWEs as whole linguistic units to generate accurate translations in target languages. Although there has been a lot of work indicating that MT systems still suffer from MWEs (Han et al., 2020a; Manakhimova et al., 2023) and adding MWE-specialized components in current systems to deal with them will increase the overall performance (Rikters and Bojar, 2017; Zaninello and Birch, 2020; Garg et al., 2022), it still exists an obvious lack of systematic fine-grained investigation of how well MT systems can handle different types of MWEs and how many translation errors can be accounted for by them. Therefore, we conduct this study to give initial insights into the nuanced impacts of MWEs on MT systems in terms of both the overall performance and the correlation between different types of translation errors and various categories of MWEs, with a special focus on the Chinese-English translation.

According to Baldwin and Kim (2010), MWEs can be defined as the "lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity". In the current study, we make use of an existing categorization scheme for Chinese MWEs proposed in our previous study (Song and Xu, 2024). In this scheme, given that multiword named entities (NEs) account for a large portion of MWE items and exhibit unique linguistic behaviors and forms compared to other types of MWEs (Constant et al., 2017; Vincze et al., 2011), they are discussed separately. Specifically, the term 'NE(s)' covers 19 types of expressions according to the framework of the OntoNotes project (Weischedel et al., 2012), and an additional category 'multiword domain terminology (TER)' is further included (Constant et al., 2017). While 'MWE(s)' only refers to 9 non-NE types of expressions based on the combination of the framework of PARSEME project (Savary et al., 2023) and some related studies, e.g. Wang (2020). We adopt the same terms and denotations in the present study.

Findings of the Association for Computational Linguistics: EACL 2024, pages 6154–6165

In our previous study (Song and Xu, 2024), we extended the Chinese-English parallel corpus provided by the WMT22 Metric Shared Task (Freitag et al., 2021, 2022) by annotating MWEs and NEs in Chinese source sentences. Following this annotated data, we make further identification of MWE- and NE-related translation errors in all the MT outputs in English based on the existing error annotations according to multidimensional quality metrics (MQM) framework (Lommel et al., 2014), which also consist in the WMT22 Metric Shared Task corpus (Freitag et al., 2022)[1]. The analysis of the impacts of different kinds of Chinese MWEs and multiword NEs on MT systems is then conducted. We compare the average human scores of MT outputs on the sentence groups with and without a particular category of MWEs and NEs. The results show that most MT systems obtain lower scores on sentences with MWEs and/or NEs. Besides translation scores, we explore whether and how different MWEs and NEs cause particular translation errors. Our investigation also presents that MWEs and NEs which make up about twenty percent of Chinese characters result in one-third of translation errors. Furthermore, the error rates of 13 categories of MWEs and NEs exceed 50% with the highest being 84.8%. Our study strongly suggests that the current MT systems still struggle with most kinds of MWEs and NEs. Meanwhile, the fine-grained analysis from the perspective of translation errors can provide invaluable insights into the possible directions for the improvement of MT systems.

The remainder of the paper is organized as follows: Section 2 introduces the related work. Section 3 describes the data and error extraction procedure in detail. Section 4 presents our descriptive analysis and main experiments on the impacts of various MWEs and NEs on Chinese-English translations from the perspectives of human evaluation scores and translation errors. In Section 5, we focus on typical cases and propose possible directions for further improvement of MT systems in tackling MWEs and NEs. Section 6 concludes the study and lays out our future work.

2 Related Work

Translating MWEs has long been recognized as a challenging task, both in theory and in practice (Constant et al., 2017; Hidalgo-Ternero and Zhou-Lian, 2022). Special attention has been paid to the common issues encountered in MT concerned with MWEs (Han et al., 2020a; Esperança-Rodier and Didier, 2016). Constant et al. (2017) listed "ambiguity, discontiguity, non-compositionality, and variability" as the main challenges to MT systems when dealing with MWEs. In the study of Han et al. (2020a), they summarized six kinds of highly frequent difficulties when automatically translating Chinese MWEs into English, encompassing "common sense, super sense, abstract phrase, idioms, metaphor, and ambiguity". In addition to the qualitative analysis on MWE translations, Esperança-Rodier and Didier (2016) made a quantitative analysis by semi-automatically annotating five types of French MWEs and evaluating their English translations given by the MT systems. While they found an overall satisfactory performance of the MT system in translating French MWEs into English, they called for more detailed discussions on translation errors. Besides, tests with challenge sets conducted by Macketanz et al. (2018, 2022); Manakhimova et al. (2023) for MT evaluation tasks also rendered MWEs essential to investigate the performance of MT systems, further highlighting the contribution of the present study.

As suggested by Constant et al. (2017), improvements in MT performance by being aware of MWEs can be achieved through specific orchestration strategies tailored to different types of MT systems. Several studies have demonstrated that integrating information about MWEs can benefit both statistical MT (Lambert and Banchs, 2005; Ren et al., 2009; Okita and Way, 2011; Bouamor et al., 2012; Tan and Pal, 2014; Ebrahim et al., 2017) and neural MT (Rikters and Bojar, 2017; Han et al., 2020b; Zaninello and Birch, 2020; Garg et al., 2022) in terms of automatic evaluations.

To sum up, while the aforementioned studies have provided a broad understanding of the challenges posed by MWEs in MT and highlighted the potential benefits of incorporating MWE information into MT systems, little is known about the specific and subtle effects of different MWEs from the perspective of human scores and various translation errors caused by them, hindering the further advancements in MWE-enhanced MT systems.

3 Data and Error Annotation

3.1 Basic Data

In this study, we refer to the categorization scheme for Chinese MWEs and multiword NEs proposed

[1] https://github.com/florethsong/mte-zh-mwe

Category	Count	Type	Sentence Number	Example
NID (Noun-headed Idioms)	137	81	117	巨无霸 *ju wu ba* (lit. too huge to be bullied) (the giant)
ION (Separable Words or Ionized Words)	5	4	5	成不了像 *cheng bu liao xiang* (lit. to make no image) (fail to make an image)
IDI (Conventionally Fixed Idioms)	265	239	200	雪中送炭 *xue zhong song tan* (lit. to send charcoal in snowy weather) (to send help in one's need)
CON (Syntactically Special Constructions)	87	34	82	除_外 *chu_wai* (lit. except for)
VID (Verb-headed Idioms)	83	65	78	下决心 *xia jue xin* (lit. to set down the determination) (to make up one's mind)
VPC.semi (Semi Non-compositional Verb-particle Constructions)	397	212	343	意识到 *yi shi dao* (lit. to be aware of)
LVC.full (Light Verb Constructions with Bleached Verbs)	195	154	163	发表演讲 *fa biao yan jiang* (lit. to give a speech)
LVC.cause (Light Verb Constructions with Causative Verbs)	35	30	34	引发破坏 *yin fa po huai* (lit. to lead to damage)
MVC (Multi-verb Constructions)	226	131	198	试试看 *shi shi kan* (lit. to try and see) (to have a try)
TER	189	76	127	石英 *shi ying* (lit. Quartz)
PERSON	262	166	176	李效良 *li xiao liang* (lit. Li, Xiaoliang)
NORP	21	7	16	中华民族 *zhong hua min zu* (lit.Chinese nation)
FAC	53	29	46	天安门广场 *tian an men guang chang* (lit. Tian Anmen Square)
ORG	414	223	267	湖南日报 *hu nan ri bao* (lit. Hunan Daily)
GPE	400	125	225	青海省 *qing hai sheng* (lit. Qinghai Province)
LOC	76	33	56	青藏高原 *qing zang gao yuan* (lit. Qinghai-Tibet Plateau)
PRODUCT	48	20	38	比特币 *bi te bi* (lit. Bitcoin)
EVENT	111	79	83	民主峰会 *min zhu feng hui* (lit. Summit for Democracy)
WORK_OF_ART	122	98	67	《查理和巧克力工厂》 《*cha li he qiao ke li gong chang*》 (lit. *Charlie and the Chocolate Factory*)
LAW	6	6	6	《刑法》《*xing fa*》 (lit. *Criminal Law*)
LANGUAGE	13	4	10	日文 *ri wen* (lit. Japanese)
DATE	519	285	387	5月20日 *5 yue 20 ri* (lit. on May 20th)
TIME	101	75	83	四分钟 *si fen zhong* (lit. four minutes)
PERCENT	50	37	41	四成以上 *si cheng yi shang* (lit. more than 40%)
MONEY	68	55	49	十元 *shi yuan* (lit. ten yuan)
QUANTITY	47	42	25	几千公里 *ji qian gong li* (lit. thousands of kilometres)
ORDINAL	104	66	96	第一次 *di yi ci* (lit. at the first time)
CARDINAL	223	161	171	两匹 *liang pi* (lit. two 'horses')

Table 1: Detailed information of 9 types of **MWEs** and 19 types of NEs in Chinese (Song and Xu, 2024). 'Count', 'Type', and 'Sentence Number' represent the number, the deduplicated number, and the number of sentences regarding a particular category of expressions, respectively.

by Song and Xu (2024) and base our extraction of translation errors on a well-annotated dataset constructed accordingly. The scheme combines the general MWEs categories proposed by the PARSEME project (Savary et al., 2023), the NEs categories by the OntoNotes project (Weischedel et al., 2012), and some other categories proposed by several existing studies, e.g. Constant

et al. (2017); Wang (2020). The MWE- and NE-extended dataset in our previous study (Song and Xu, 2024) therefore covers the labels of 9 types of Chinese MWEs and 19 types of Chinese multiword NEs by taking the Chinese-English parallel corpus from the WMT22 Metrics Shared Task (Freitag et al., 2022) as the basis. Finally, in our dataset, 1,359 Chinese source sentences containing 1,430 MWEs and 2,827 NEs (accounting for 15,585 tokens/characters) are marked out. More details are shown in Table 1.

The original parallel corpus of WMT22 consists of 1,875 Chinese sentences (74,616 tokens/characters) along with their corresponding English translations. These translations were generated by 14 state-of-the-art MT systems that participated in the shared translation task of WMT22 (Kocmi et al., 2022), standing at the forefront in the field of MT. Therefore, a total of 26,250 translation items are evaluated in the current study.

3.2 Error Typology

In accordance with a modified framework of MQM (Lommel et al., 2014; Freitag et al., 2021, 2022), the manual evaluation scores and the translation error labels on English texts given by human experts for further assessing the English translation quality are also included in the WMT22 dataset, serving as the foundation for our fine-grained error analysis. Specifically, it includes three aspects for translation error description: 1) error severity (*Major*, *Minor*, or *Neutral/No*), 2) error type (*Accuracy*, *Fluency*, *Terminology*, *Style*, and *Locale*), and 3) error subtype (for each type of error, there are some different subtypes, for example, *Addition* in *Accuracy* and *Spelling* in *Fluency*). In total, 20 kinds of *Major* errors, 21 kinds of *Minor* errors, and *No-error* are observed. More details are exhibited in Figure 6. Each sentence is limited to a maximum of five errors. Based on the identified errors, the final score for a translation ranges from 0 (the best) to -25 (the worst), with a *Major* error weighted at -5, a *Minor* error weighted at -1, and no deduction of scores for a *No-error* in general.

3.3 Error Annotation

In order to make an analysis of the impact of Chinese MWEs and NEs in terms of causing translation errors, we further link the translation errors in the target language (English) with their Chinese counterparts by manually aligning the error spans in the WMT22 data with the MWE and NE annotations by Song and Xu (2024). Then, an error is considered to be caused by the MWE or the NE if the range of its Chinese counterpart overlaps with the given span of an English error instance. After annotation, we find that among the 40,915 MQM-based labels in the overall data, 16,652 items are directly related to Chinese MWEs and NEs.

4 Impacts of MWEs and NEs on General Performance of MT

All these categories we include here due to their idiosyncrasy nature might cause problems for MT systems. MT systems need to fully understand the meaning of such linguistic units as a whole to generate the appropriate translations in the target language. On the other hand, although they all show semantic indecompositionality, their degrees vary across categories, which may affect MT systems differently. Roughly, all the categories can be further divided into two groups, namely fixed and non-fixed.

NID (noun-headed idiom) is usually manifested as a nominal unit possibly with modifiers of the head noun, e.g. 铁公鸡 *tie gong ji* 'iron cock' (miser). Such nominal idioms are quite flexible regarding their key components. For example, different modifiers can be added to the head noun or the order of the components can be changed like 铁打的 (*da de*, 'made of') 公鸡. VIDs (verb-headed idioms) differ from NIDs in that they are usually verbal units, typically verb-object constructions, such as 下决心 *xia jue xin* (make up one's mind). VIDs also show high flexibility in that a variety of constituents can occur in the middle, like 下了 (*le*, 'perfect aspect') 决心, 下定 (*ding*, 'firmly') 决心 to emphasize the status, and so on. Besides, some NID and VID expressions are ambiguous between their literal meanings and idiomatic meanings, e.g. 戴帽子 *dai mao zi* 'wear a hat' or 'bear a given label'. To correctly translate such MWEs, the MT system needs to be aware of the syntactic relation between the key components and their contexts, as well as understand the idiomatic meaning of such MWEs in order to make a correct judgment.

Ionized words or separable words (IONs) are a phenomenon where a verb is separated into two parts and acts like a verb-object construction. For example, the word 成像 *cheng xiang* (make an image) can be separated as in 成-不了-像 *cheng bu liao xiang* (cannot make an image), while an ION

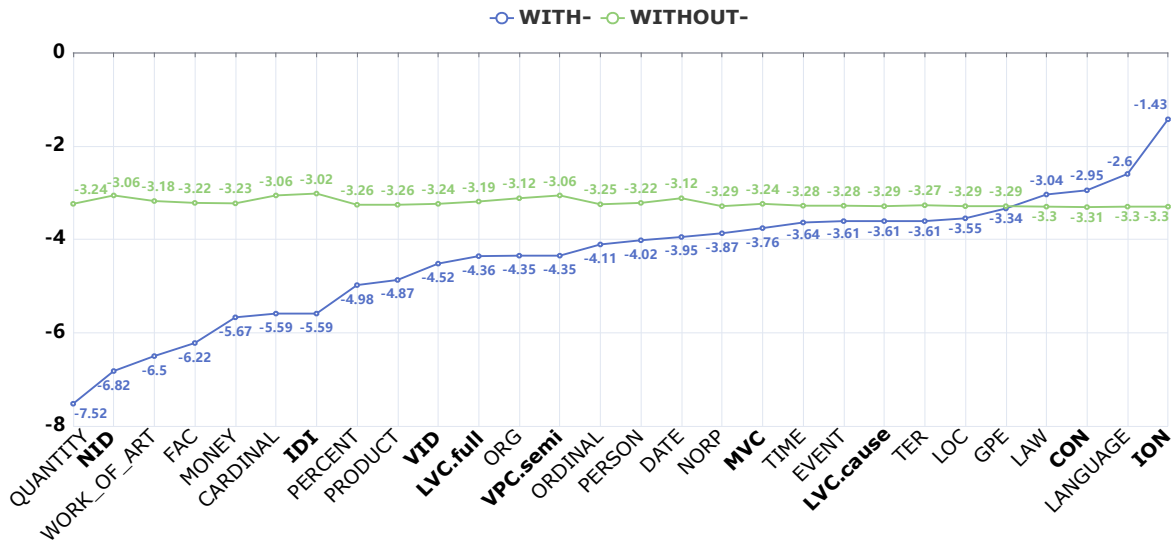

Figure 1: The average human score for each category group of **MWEs** or NEs as shown by blue dots, as well as for their complementary groups as shown by green dots.

can not go through a semantic shift when it is separated. Thus, to correctly translate IONs requires syntactic awareness of the separated components and that the non-separated canonical form is within the vocabulary of the MT system.

Compared to the above categories, the left categories are more fixed including IDI (idioms), most NEs, and others. For example, IDI, which particularly refers to the four-character Chinese idioms, such as 雪中送炭 *xue zhong song tan* 'to send charcoal in snowy weather' (to send help in one's need). Such idioms are always in the same form when used. The vocabularies of these categories are mostly fixed although new items occasionally appear. Whether MT systems can translate them correctly also depends on the vocabulary of the MT systems collected from their training data.

Finally, there is a set of open categories (meaning that they have infinite vocabulary size) including QUANTITY, MONEY, CARDINAL, and so on. Although the forms of such entities are mostly fixed, they require MT systems to grasp the potential numeric generation rules of both source and target languages since the training data is not able to include all possible instances of them. We postulate that flexible MWEs and such open categories of NEs will pose bigger challenges than the other categories. On the other hand, the fixed ones will be a good test of the potential 'vocabulary size' of the MT systems. To test our hypothesis, we focus on two measures: average scores and error rates, in our following experiments.

4.1 Average Score

Firstly, to investigate whether MT systems will obtain different performances when a certain type of MWEs or NEs is present, we calculate the average score of all translations given by 14 MT systems for all the sentences with a particular type of MWEs or NEs and compare it with that of its complementary group, i.e. sentences without that type of MWEs or NEs. The results are shown in Figure 1. We can see that most categories of MWEs or NEs impose negative effects on MT systems. Specifically, in 24 out of 28 categories, MT systems receive lower average scores than those without them, as indicated by a blue dot below the corresponding green one. Exceptions are the four categories including LAW, CON, LANGUAGE, and ION. However, three of them are very small categories (c.f. Table 1) and the average values thus might not be reliable on them.

On the other hand, the most influential categories are QUANTITY, NID, WORK_OF_ART, FAC, CARDINAL, IDI, PERCENT, PRODUCT, VID, and so on. All of these categories share a common property that they either are open classes or have a very large vocabulary. As we discussed above, MT systems need to manipulate the potential rules of such units in order to translate them correctly. For most of the state-of-the-art MT systems, no such components are equipped to deal with them. The result is also consistent with our analysis that NID and VID, due to their idiosyncrasy nature and flexibility, pose a great challenge

to MT systems. For WORK_OF_ART, IDI, and PRODUCT, although their forms are mostly fixed, their semantics are also non-decompositional and their vocabulary sizes are very large and even infinite, which will cause the OOV issues for the MT systems. Although the most advanced contextual embedding technology can help guess the meaning of the units, it is still a challenge to precisely translate them in most cases.

WITH- / WITHOUT-	*T*-statistic	*P*-value
NID	-13.36	***5.71E-09
ION	**5.61**	***8.52E-05
IDI	-13.58	***4.68E-09
CON	**3.16**	**7.52E-03
VID	-7.58	***4.04E-06
VPC.semi	-11.09	***5.34E-08
LVC.full	-7.13	***7.68E-06
LVC.cause	-1.80	9.59E-02
MVC	-3.71	**2.60E-03
TER	-1.29	2.20E-01
PERSON	-2.41	*3.15E-02
NORP	-2.81	*1.49E-02
FAC	-8.44	***1.24E-06
ORG	-5.74	***6.80E-05
GPE	-0.36	7.25E-01
LOC	-0.97	3.50E-01
PRODUCT	-5.61	***8.43E-05
EVENT	-1.42	1.78E-01
WORK_OF_ART	-7.47	***1.14E-05
LAW	**0.54**	5.97E-01
LANGUAGE	**2.56**	*2.36E-02
DATE	-4.91	***2.86E-04
TIME	-3.37	**5.01E-03
PERCENT	-4.99	***2.49E-04
MONEY	-12.44	***1.35E-08
QUANTITY	-11.68	***2.89E-08
ORDINAL	-5.39	***1.24E-04
CARDINAL	-11.51	***3.44E-08

Table 2: The results of paired t-tests between the average scores of each category group of **MWEs** or NEs, and their correspondingly complementary groups. A significant difference is observed when p is less than 0.05 (*), 0.01 (**) and 0.001 (***).

From another perspective, we want to know whether all MT systems are affected similarly by the presence of different MWEs or NEs. For each of the 14 MT systems, we calculate the average human score of all the sentences it translates in one category group (e.g. WITH-NID) and that of its complementary group (e.g. WITHOUT-NID). Then, we apply paired t-tests to the 14 groups of data to examine if a certain MWE or NE category can significantly affect most MT systems in average scores. As shown in Table 2, most categories of the MWEs and NEs impose significantly negative effects on the MT systems, as indicated by large negative t scores. Again, the four categories that seemingly make positive effects are CON, ION, LAW (non-significant), and LANGUAGE and three of which are very small categories and might be subject to other influential factors associated with them. For CON, upon our detailed examination, most of them behave like fixed lexical items, e.g. 除...外 *chu wai* (except for). Besides, CON has a very limited vocabulary and thus may have been very well learned by MT systems.

The most negatively influential categories identified in this experiment are NID, IDI, VID, VPC.semi, LVC.full, FAC, WORK_OF_ART, MONEY, QUANTITY, and CARDINAL. The results here are consistent with those revealed in Figure 1. Both are consistent with our theoretical analysis above.

4.2 Error Rate

Here, we focus our discussion primarily on the subtypes of errors, that is, the most fine-grained level. The 'error rate' refers to the percentage of a particular type of error that occurred in the translations of a group of MWE/NE-featured sentences (e.g. WITH-NIDs) out of the total errors. In cases where an error spans over multiple MWEs or NEs, we evenly distribute the count among them. There are three dimensions to consider: the error rate of each severity (*Minor* vs. *Major*), that of each subtype of errors, and that within each category group of MWEs or NEs.

4.2.1 Error Rate of Severity

The overall distribution of translation errors is presented in Figure 2. Among the total 28,745 *Major* and *Minor* errors in all translations, professional translators have identified similar proportions of *Major* and *Minor* errors as 51.85% and 48.15% respectively. We can also see that 28.46% (17.95%+10.51%) of translation errors are caused by MWEs and NEs. Considering that MWEs and NEs only contribute about 20.9% of tokens (15,585 out of 74,616), it should be claimed that they cause a relatively large proportion of errors, indicating that MWEs and NEs are more challenging for MT systems to handle than other parts of sentences. Moreover, both MWEs and NEs result in more *Major* errors compared to *Minor* errors, with *Major* occurring nearly twice as often as *Minor* in translations of MWEs and/or NEs. This could be due to shifts in meaning compared to the literal

meanings of their components and implicit connotations (Sag et al., 2002; Constant et al., 2017; Han et al., 2020a), potentially leading to misunderstandings for MT systems.

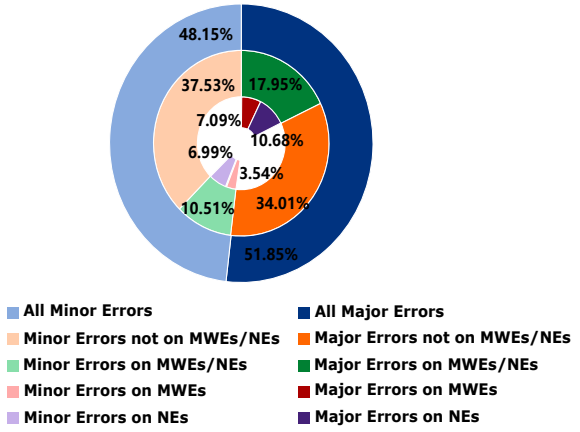

Figure 2: The error rates of different levels of severity. All percentages are computed based on the total number of *Major* and *Minor* errors in all translations, i.e. 28,745.

Figure in Appendix A shows the numbers of different types of errors in detail. It is easy to notice that *Accuracy/Mistranslation* has dominated the *Major* errors, among which NID, IDI, VID, VPC.semi, FAC, ORG, WORK_OF_ART, and TER have contributed the most.

4.2.2 Error Rate of Subtypes

The error rates of different subtypes in both *Major* and *Minor* errors are shown in Figure 3 and Figure 4 respectively. In both *Major* and *Minor* parts, most subtypes of translation errors on NEs show a higher rate than those on MWEs. Although a significant number of errors related to *Accuracy/Mistslation*, *Fluency/Grammar* and *Style/Awkward* appear in the translations of Chinese MWEs and NEs as shown in Appendix A, the interesting fact is that their error rates are not as high as expected. Conversely, the subtypes that take a small amount always account for a large proportion, such as *Locale convention/Date format*, *Locale convention/Address format*, *Locale convention/Currency format*, and *Locale convention/Name format* in both *Major* and *Minor* parts.

The reason for this result is twofold. Firstly, *Accuracy/Mistslation*, *Fluency/Grammar* and *Style/Awkward* represent common errors related to some broader linguistic concepts, like semantics, pragmatics, and syntax, which frequently occur in all parts of translations. Their generality leads to a large number but a relatively small proportion of translation errors on MWEs or NEs. Secondly, in contrast, the *Locale* subtypes denote some special format errors. They exactly match with some properties of MWEs and NEs to a certain degree, like formal rigidity, institutionalization, and non-substitutability (Sag et al., 2002; Baldwin and Kim, 2010) that are not present in other parts of a sentence. This explains why the subtypes of Locale always have high error rates regarding MWEs and NEs. For example, the formal rigidity of NEs requires expressing '张迪鸣', a personal name in Chinese, as 'Zhang, Diming', 'Diming Zhang' or 'ZHANG Diming' to emphasize the last name '张' which appears at the beginning of the Chinese name. At the same time, 'Zhang Diming', the translation from MT systems is marked with the error label *Locale convention/Name format*.

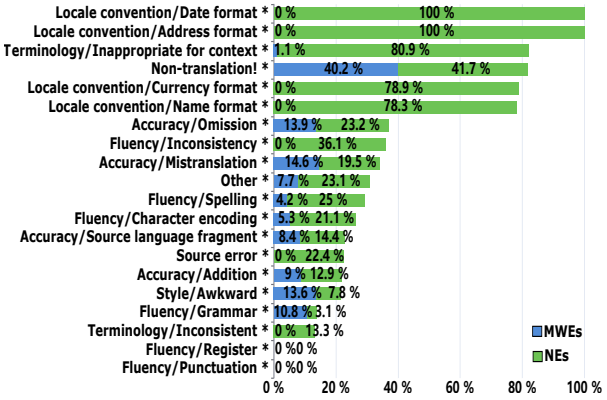

Figure 3: The rate of each subtype in *Major* errors, computed based on the number of the corresponding subtype of errors in the *Major* part.

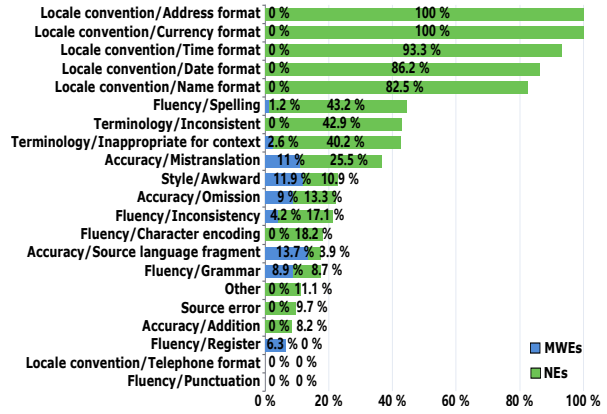

Figure 4: The rate of each subtype in *Minor* errors, computed based on the number of the corresponding subtype of errors in the *Minor* part.

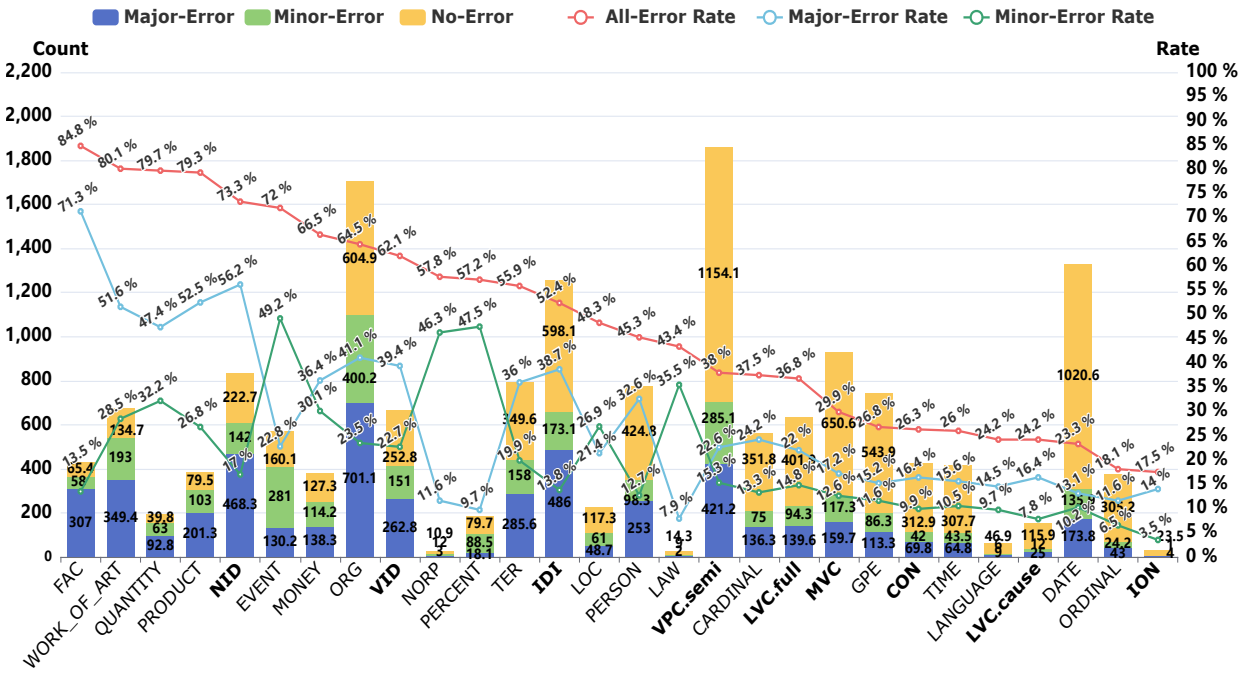

Figure 5: The error rates in different categories of **MWEs** and NEs. The '*Major-Error Rate*' and the '*Minor-Error Rate*' are the proportions of the English instances of a certain type of MWEs or NEs, with *Major* errors and *Minor* errors respectively, and the '*All-Error Rate*' indicates the total proportions of mistranslations for that type of MWEs or NEs, that is, the sum of '*Major-Error Rate*' and the '*Minor-Error Rate*' within the category.

4.2.3 Error Rate in Category

From the perspective of different categories of MWEs and NEs, we compute how many of them are mistranslated, i.e. the error rate per category. Figure 5 shows the results. In addition, the figure also includes the counts of errors that occurred in each category. The categories are sorted with the error rates in descending order. It is striking that for 13 categories of MWEs and NEs, e.g. FAC, WORK_OF_ART, etc., the error rates exceed 50%, with those of FAC and WORK_OF_ART even exceeding 80%. In other words, more than half of their instances are translated with errors. Furthermore, WORK_OF_ART, NID, ORG, VID, TER, IDI, and VPC.semi also cause the most errors when it comes to the counts. Additionally, it is noteworthy that for translations of EVENT, NORP, PERCENT, LOC, and LAW, the rates of *Minor* errors surpass those of *Major* errors, highlighting the inadequacies still substantially existing in their translations that deserve some subtle adjustments. The translations of the other 23 categories of Chinese MWEs and NEs, instead, require extra attention to their *Major* errors that are frequently made by most MT systems.

5 Discussion

With the analysis presented in this study, it is confirmed that current MT systems still face challenges from most kinds of Chinese MWEs and NEs. Although MWEs and NEs only take up one-fifth of the tokens in source texts, nearly one-third of all translation errors are caused by them, and among all the errors, two-thirds are *Major* errors, seriously affecting the accuracy and fluency of translations.

Secondly, according to a detailed error analysis, there are some interesting differences between the features of errors on MWEs and NEs. The errors on MWEs tend to be concentrated on *Major* errors, particularly *Accuracy/Mistranslation*, while those on NEs usually show a larger amount, a higher error rate, and greater diversity. This pattern is consistent with the features of MWEs and NEs themselves in sentences, where NEs tend to have a larger quantity, higher proportion, and richer categories.

Thirdly, we have hypothesized that open NE categories such as QUANTITY, MONEY, CARDINAL, and MWEs categories with high flexibility and idiosyncrasy or semantic non-transparency such as NID, VID, and IDI will cause most problems to MT systems. The experimental results are consistent with our theoretical analyses. The fol-

lowing shows some typical examples of MWEs and NEs that have been mistranslated by MT systems [2].

- **NID (*Major-Accuracy/Mistranslation*)**
  Source: 团队中的“老油条”不知如何应对
  Pinyin: tuan dui zhong de “lao you tiao” bu zhi ru he ying dui
  Lit. of the MWEs: old oil-fried stick
  Reference: there's no way to handle the "sophisticated ones" in the team
  MT: the "old <v>fritters</v>" in the team do not know how to deal with

- **FAC (*Major-Accuracy/Mistranslation*)**
  Source: 居然之家一共11层
  Pinyin: ru ran zhi jia yi gong 11 ceng
  Lit. of the MWEs: Juran Home
  Reference: there are 11 floors in Easyhome
  MT: <v>the Home of the residence</v> a total of 11 floors

- **VID (*Major-Accuracy/Mistranslation*)**
  Source: 研究成果让人脑洞大开
  Pinyin: yan jiu cheng guo rang ren nao dong da kai
  Lit. of the MWEs: open brain holes
  Reference: The research results make people's minds open
  MT: The research <v>generates brain holes</v>

Based on the analysis associated with the given examples, some hints of possibly improving the current MT systems can be derived. Firstly, most MT systems have learned quite a bit of syntax. For example, even though the systems cannot translate some NIDs and VIDs correctly due to the flexibility in their syntactic structures, they still try their best to understand the meaning only through the semantic composition of the components. There are two possible ways of solving the problem. The first is to feed more training data containing examples of such MWEs or NEs, so the MT systems can learn them properly. However, due to the large vocabulary, it is potentially impossible to include all of them in the training data. The second possible solution is to allow the MT models to look up dictionaries in real-time and incorporate real-time information when generating the translations. The in-context learning scheme of large language models (LLMs) might also be possibly integrated into the MT task-specific models as well.

Secondly, for the open categories that show strict inner structures such as QUANTITY, MONEY, DATE, etc., a specific component for parsing such structures and generating some proper intermediate representations might still be necessary for modern deep learning architectures. Correspondingly, linguistic resources that support solving the problems associated with MWEs and NEs as we exposed in this study should be constructed and leveraged.

6 Conclusion

This study completes a deep dive into the performance of state-of-the-art MT systems on the task of translating Chinese MWEs into English. By comparing average scores and the three types of error rates in the automatic translations of 9 types of typical Chinese MWEs and 19 types of Chinese multiword NEs, we confirm that most Chinese MWEs and NEs are still a bottleneck problem for MT systems, causing one-third of translation errors in general. Categories such as NID, VID, FAC, QUANTITY, WORK_OF_ART, PRODUCT, etc. significantly degrade the performance of MT systems due to their high error rates (above 60%). The study thus provides invaluable insights for facilitating further improvement of MT systems from the perspective of integrating and intensifying knowledge of different MWEs.

Limitations

Our analysis is based on the existing WMT22 data and its MWE-annotated version by Song and Xu (2024), which does not contain translations by LLMs. We considered extending the data by including translations from LLMs. However, the obtained translations will need further annotations including human evaluation scores and translation errors. Since the annotators may have different criteria, the direct comparison between the two separate annotations could be misleading. Although LLMs have shown promising results in performing multiple tasks, there is no strong evidence that they outperform any other task-specific models. Additionally, given that the state-of-the-art MT systems and LLMs are all built upon similar architecture, i.e. transformers, our findings may also apply to LLMs in theory, and in our future work, we will address this issue by adding data from WMT23 and

[2]The errors in MT outputs are labeled with '<v>' and '</v>' by human experts.

WMT24 dataset which includes LLM translations.

Acknowledgments

This work is supported by the Supervisor Academic Guidance Program of Shanghai International Studies University (Grant No. 2022113024) and the Linguistics Frontier Research Funding of Shanghai International Studies University (Grant No. 41004525/001).

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, chapter 12, pages 267–292. Chapman and Hall/CRC, New York.

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2012. Identifying bilingual multi-word expressions for statistical machine translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 674–679, Istanbul, Turkey. European Language Resources Association (ELRA).

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Sara Ebrahim, Doaa Hegazy, Mostafa Gadal-Haqq M Mostafa, and Samhaa R El-Beltagy. 2017. Detecting and integrating multiword expression into english-arabic statistical machine translation. *Procedia Computer Science*, 117:111–118.

Emmanuelle Esperança-Rodier and Johan Didier. 2016. Translation quality evaluation of mwes from french into english using an smt system. In *Translating and the Computer 38*.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press, Massachusetts.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kamal Deep Garg, Shashi Shekhar, Ajit Kumar, Vishal Goyal, Bhisham Sharma, Rajeswari Chengoden, and Gautam Srivastava. 2022. Framework for handling rare word problems in neural machine translation system using multi-word expressions. *Applied Sciences*, 12(21).

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020a. AlphaMWE: Construction of multilingual parallel corpora with MWE annotations. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 44–57, online. Association for Computational Linguistics.

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020b. MultiMWE: Building a multi-lingual multi-word expression (MWE) parallel corpora. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2970–2979, Marseille, France. European Language Resources Association.

Carlos Manuel Hidalgo-Ternero and Xiaoqing Zhou-Lian. 2022. Reassessing gapp: Does mwe discontinuity always pose a challenge to neural machine translation? In *International Conference on Computational and Corpus-Based Phraseology*, pages 116–132. Springer.

Ray Jackendoff. 1995. The boundaries of the lexicon. In André Schenk Martin Everaert, Elisabeth van der Linden and Robert Schreuder, editors, *Idioms, Structural and Psychological Perspectives*, chapter 7, pages 133–166. Psychology Press, New York.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Patrick Lambert and Rafael Banchs. 2005. Data inferred multi-word expressions for statistical machine translation. In *Proceedings of Machine Translation Summit X: Posters*, pages 396–403, Phuket, Thailand.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, and Hans Uszkoreit. 2018. Fine-grained evaluation of German-English machine translation based on a test suite. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 578–587, Belgium, Brussels. Association for Computational Linguistics.

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, He Wang, Renlong Ai, Shushen Manakhimova, Ursula Strohriegel, Sebastian Möller, and Hans Uszkoreit. 2022. A linguistically motivated test suite to semi-automatically evaluate German–English machine translation output. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 936–947, Marseille, France. European Language Resources Association.

Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2023. Linguistically motivated evaluation of the 2023 state-of-the-art machine translation: Can ChatGPT outperform NMT? In *Proceedings of the Eighth Conference on Machine Translation*, pages 224–245, Singapore. Association for Computational Linguistics.

Tsuyoshi Okita and Andy Way. 2011. Given bilingual terminology in statistical machine translation: Mwe-sensitve word alignment and hierarchical pitman-yor process-based translation model smoothing. In *FLAIRS Conference*, pages 269–274.

Paul Rayson, Scott Piao, Serge Sharoff, Stefan Evert, and Begona Villada Moirón. 2010. Multiword expressions: hard going or plain sailing? *Language Resources and Evaluation*, 44:1–5.

Zhixiang Ren, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (MWE 2009)*, pages 47–54, Singapore. Association for Computational Linguistics.

Matīss Rikters and Ondřej Bojar. 2017. Paying attention to multi-word expressions in neural machine translation. In *Proceedings of Machine Translation Summit XVI: Research Track*, pages 86–95, Nagoya Japan.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing 2002 Mexico City, Mexico, February 17–23, 2002 Proceedings 3*, pages 1–15. Springer, Berlin Heidelberg.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, Jolanta Kovalevskaite, Timm Lichte, Nikola Ljubešić, Johanna Monti, Carla Parra Escartín, Mehrnoush Shamsfard, Ivelina Stoyanova, Veronika Vincze, and Abigail Walsh. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Huacheng Song, Yi Li, Yiwen Wu, Yu Liu, Jingxia Lin, and Hongzhi Xu. 2024. How grammatical features impact machine translation: A new test suite for Chinese-English mt evaluation. In *Proceedings of the 2024 Internaltional Conference on Machine Translation (WMT 2024)*.

Huacheng Song and Hongzhi Xu. 2024. Benchmarking the performance of machine translation evaluation metrics with Chinese multiword expressions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2204–2216, Torino, Italia. ELRA and ICCL.

Liling Tan and Santanu Pal. 2014. Manawi: Using multi-word expressions and named entities to improve machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 201–206, Baltimore, Maryland, USA. Association for Computational Linguistics.

Veronika Vincze, István Nagy T., and Gábor Berend. 2011. Multiword expressions and named entities in the wiki50 corpus. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 289–295, Hissar, Bulgaria. Association for Computational Linguistics.

Shan Wang. 2020. *Chinese Multiword Expressions*. Springer Singapore, Singapore.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2012. Ontonotes release 5.0 with ontonotes db tool v0. 999 beta. *Linguistic Data Consortium*, pages 1–53.

Andrea Zaninello and Alexandra Birch. 2020. Multiword expression aware neural machine translation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3816–3825, Marseille, France. European Language Resources Association.

A Translation Error Distribution

Figure 6: The distribution of different translation errors. On the y-axis, items with asterisks (*) are types/subtypes of *Major* errors, while those without any marks are types/subtypes of *Minor* errors. Items with caret (^) in the whole plot represent the sum of errors within the current severity or subtype.