Benchmarking the Performance of Machine Translation Evaluation Metrics with Chinese Multiword Expressions

Huacheng Song, Hongzhi Xu
Institute of Corpus Studies and Applications, Shanghai International Studies University
Shanghai China
{hsong, hxu}@shisu.edu.cn

Abstract

To investigate the impact of Multiword Expressions (MWEs) on the fine-grained performance of the state-of-the-art metrics for Machine Translation Evaluation (MTE), we conduct experiments on the WMT22 Metrics Shared Task dataset with a preliminary focus on the Chinese-to-English language pair. We further annotate 28 types of Chinese MWEs on the source texts and then examine the performance of 31 MTE metrics on groups of sentences containing different MWEs. We have 3 interesting findings: 1) Machine Translation (MT) systems tend to perform worse on most Chinese MWE categories, confirming the previous claim that MWEs are a bottleneck of MT; 2) automatic metrics tend to overrate the translation of sentences containing MWEs; 3) most neural-network-based metrics perform better than string-overlap-based metrics. It concludes that both MT systems and MTE metrics still suffer from MWEs, suggesting richer annotation of data to facilitate MWE-aware automatic MTE and MT.

Keywords: machine translation evaluation metric, meta-evaluation, multiword expression, named entity

1. Introduction

Multiword Expressions (MWEs) are commonly recognized as hard nuts noteworthy to many natural language processing tasks, especially Machine Translation (MT) and its evaluation (MTE) (Constant et al., 2017). Despite numerous efforts to enhance MT quality by integrating more attention to MWEs, limited knowledge of how current MTE metrics perform on different MWEs has hindered progress in this field. Additionally, the scarcity of datasets with MWE information is an obstacle to the development of MTE and even MT. Thus, we carry out the fine-grained meta-evaluation of MTE metrics concerned with various types of MWEs. Considering that different types of MWEs exhibit significant heterogeneity in both syntax and semantics (Doren Singh and Bandyopadhyay, 2011), we create a corpus well-annotated with a wide range of categories of Chinese MWEs based on the WMT22 data (Freitag et al., 2022). Our corpus is available online[1].

All data in our experiments are grouped according to MWE categories. The study is hereby conducted at 2 levels: 1) the property level, a coarse-grained grouping of sentences regarding the presence of MWEs in general; and 2) the category level, a grouping of sentences according to the fine-grained categories of MWEs. Our findings reveal that MT systems tend to perform worse on sentences with MWEs as indicated by their lower human evaluation scores and most automatic metrics tend to overrate those translations indicating their insensitivity to poorly translated MWEs and subsequently such MTE metrics perform worse on sentence groups with MWEs. In other words, automatic metrics tend to produce "false positive" scores since they are agnostic of the presence of MWEs which most MT systems are bad at. From another perspective, neural-network-based metrics regardless of whether reference-free or reference-dependent generally outperform string-overlap-based metrics. Further details will be discussed later in the paper. We conclude that our study underscores the importance of addressing MWEs when constructing MT systems and designing MTE metrics by shedding light on the performance and limitations of current models.

The remaining part of the paper proceeds as follows: In Section 2, related work and the motivation of this study are briefly introduced. Section 3 describes the details of our experiments, including methods for data annotation and statistical analysis. In Section 4, we present the experimental results, followed by a detailed analysis in Section 5. Finally, Section 6 concludes the study.

2. Related Work

Due to the multiformity of natural languages and the lack of homogeneity of MWEs, the definition and categorization of MWEs vary in the literature. In broad terms, MWEs are sequences of words that statistically co-occur and function as a single unit across word boundaries (Sag et al., 2002; Calzolari et al., 2002; Carpuat and Diab, 2010). However, this primary description does not fit well with Chinese which lacks word spaces (Wang,

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

LREC-COLING 2024, pages 2204–2216

2020). Another definition commonly adopted by researchers, including us, is provided by Baldwin and Kim (2010) who described MWEs as "lexical items that (a) can be decomposed into multiple lexemes, and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity".

Given the prevalence of MWEs in human languages, accounting for 41% of lexical items in WordNet 1.7 (Fellbaum, 1998), considerable attention has been devoted to leveraging their positive influence on improving MT systems (Deksne et al., 2008; Bouamor et al., 2012; Ghoneim and Diab, 2013; Kordoni and Simova, 2014; Rikters and Bojar, 2017; Zaninello and Birch, 2020). However, the development of MWE-aware MTE metrics for a better diagnosis of MT systems has been a largely under-explored domain and remains an open challenge. Despite BLEU (Papineni et al., 2002) being the predominant choice for assessing translation quality, Constant et al. (2017) argued that it fails to identify specific gaps, regarding e.g. MWEs, between different MT systems.

Some work has focused on examining MT qualities on MWEs by analyzing translation errors within limited types of MWEs and language pairs (Babych and Hartley, 2010; Barreiro et al., 2013; Schottmüller and Nivre, 2014; Esperança-Rodier and Didier, 2016; Esperança-Rodier and Frankowski, 2021). Very few studies probed into the MWE-task-oriented meta-evaluation or design of MTE metrics (Avramidis and Macketanz, 2022; Salehi et al., 2015). Avramidis and Macketanz (2022) found that most MTE metrics perform worse in bidirectional translations between German and English when it comes to some phrasal structures, viz. Named Entities (NEs), terminology, and measuring units. Salehi et al. (2015) carried out a pilot study to integrate the compositionality scores of English noun compounds into a traditional MTE metric and obtained promising results.

Research on Chinese MWEs is notably scarcer in comparison to English (Wang, 2020; Ramisch et al., 2023). Recent contributions to Chinese MWEs include the annotated corpora such as AlphaMWE (Han et al., 2020a), MultiMWE (Han et al., 2020b), and PARSEME 1.2 (Ramisch et al., 2020) and 1.3 (Savary et al., 2023). The HiLMeMe (Han et al., 2020b), a human-in-the-loop MTE metric stressing MWEs, demonstrates the significant potential for enhancing MTE accuracy from the perspective of Chinese MWEs.

In summary, the existing work highlights the insufficient exploration of the impacts of diverse MWEs on mainstream MT systems and MTE metrics along with the growing trend of applying linguistic resources to improve fine-grained MTE (Han, 2022) which appeals efforts on constructing MWE-annotated datasets.

3. Experiments

In this section, we present the details of our experiments, including the data to be used and its annotation, the MTE metrics to be evaluated, and the computational settings.

3.1. Data and Metrics

The original materials for our study are the test sets and the corresponding results officially collected by the WMT22 Metrics Shared Task (Freitag et al., 2022). This dataset comprises 1,875 Chinese sentences (74,616 tokens) from 4 different domains (news, social, e-commerce, and communication) along with references, MT outputs, human scores, and automatic metric scores. We employ the Multidimensional Quality Metrics (MQM) (Lommel et al., 2014) version of expert-based human evaluation scores, which are deemed more reliable than other human-generated scores like Direct Assessment, according to Freitag et al. (2022).

We analyze a total of 31 metrics from the WMT22 Metrics Shared Task, including 9 baselines from SacreBLEU (Post, 2018) and 22 newly-submitted metrics. They are cited in Table 1 with 2 broad criteria: 1) the basis of the model (string-overlap-based versus neural-network-based), and 2) the need for a reference translation (reference-dependent versus reference-free).

Metric	Reference
BLEU	(Papineni et al., 2002)
f101spBLEU	(Goyal et al., 2022)
f200spBLEU	(Team et al., 2022)
chrF	(Popović, 2015)
BERTScore	(Zhang et al., 2019)
BLEURT-20	(Sellam et al., 2020)
COMET-20	(Rei et al., 2020)
COMET-QE*	(Rei et al., 2021)
YiSi-1	(Lo, 2019)
COMET-22 COMETKiwi*	(Rei et al., 2022)
Cross-QE* HWTSC-TLM* HWTSC-Teacher-Sim* KG-BERTScore*	(Liu et al., 2022)
MATESE MATESE-QE*	(Perrella et al., 2022)
MEE	(Mukherjee et al., 2020)
MEE2 MEE4	(Mukherjee and Shrivastava, 2022b)
metricx_xl_DA_2019 metricx_xxl_DA_2019 metricx_xl_MQM_2020 metricx_xxl_MQM_2020	(Freitag et al., 2022)
MS-COMET-22 MS-COMET-QE-22*	(Kocmi et al., 2022)
REUSE*	(Mukherjee and Shrivastava, 2022a)
SEScore	(Xu et al., 2022)
UniTE UniTE-ref UniTE-src*	(Wan et al., 2022)

Table 1: Metrics under comparisons and their categories. Items with underlines (_) are string-overlap-based metrics, otherwise neural-network-based metrics and those with asterisks (*) are reference-free metrics, otherwise reference-dependent metrics. These 2 labels remain consistent in the following content.

3.2. Annotation of Chinese MWEs

This study examines a wide range of MWE types in Chinese. We combine the category system for Chinese verbal MWEs by the PARSEME project (Savary et al., 2023) with the typology of NEs proposed by the OntoNotes project (Weischedel et al., 2012) The final scheme includes 28 categories of Chinese MWEs, encompassing additional phenomena including noun-headed idioms, special lexical structures, multiword terminologies, and separable words (Sag et al., 2002; Pal et al., 2010; Baldwin and Kim, 2010; Constant et al., 2017; Wang, 2020).

Given their large proportions and syntactical idiosyncrasies, NEs merit a separate discussion as a particular subset of MWEs (Jackendoff, 1997; Vincze et al., 2011; Oh, 2022). Therefore, it is important to stress that the term "MWE#" in the subsequent analysis denotes 9 types of non-NE expressions, while the term "NE" covers the other 19 types of expressions as outlined in Appendix A.

We annotate all Chinese MWEs, including MWEs# and NEs, in the WMT22 data through a semi-automatic method. The procedure involves 2 steps: 1) identifying MWEs# in source texts via a pre-made dictionary and recognizing NEs using Stanza (Qi et al., 2020); 2) correcting any errors by native Chinese speakers, including the spans and the categories of all the annotated instances. Ultimately, 4,257 MWEs# and NEs (15,585 tokens) in 1,875 Chinese sentences are labeled, with the proportion and the number of each category displayed in Figure 1 and Appendix A.

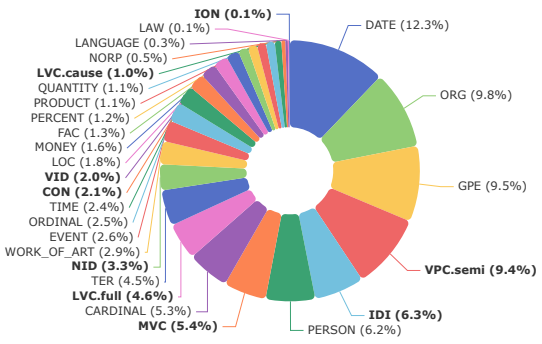

Figure 1: Proportions of **MWEs#** and NEs

To create the MWE# dictionary, we extract MWE# items from 3 sources: 1) *PARSEME 1.3* (zh) (Savary et al., 2023), a multilingual corpus annotated with verbal MWEs; 2) *A Comprehensive Dictionary of Chinese/English Idioms* (Zhang and Zhang, 2014), an authoritative bilingual collection encompassing general three-or-four-character idioms, widely used proverbs, allegorical sayings, and other lexical chunks in Chinese; 3) *ID10M* dataset (Tedeschi et al., 2022), an automatically created dataset annotated with MWEs identified by a Transformer-based model. After manually removing duplicates, errors, and non-Chinese characters, the dictionary contains 39,978 Chinese MWEs#.

3.3. Experimental Settings

The study analyzes all data in 2 dimensions: the property level and the category level. For the property level, source sentences are divided into 12 paired groups based on the presence of MWEs#, NEs, or both. The proportions of all property groups are shown in Figure 2. These results highlight the significance of our study, as the probability for a normal Chinese sentence to contain MWEs (WITH) is accordingly over 72%. For the category level, sentences are further grouped depending on the existence of specific types of MWEs# or NEs.

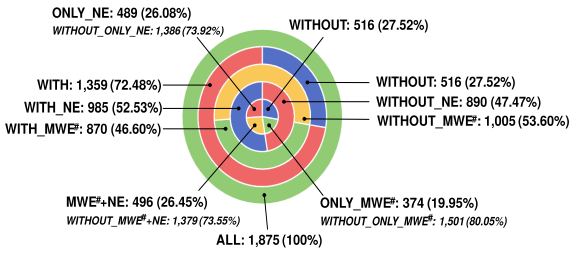

Figure 2: Proportions of different property groups

The present study follows the basic hypothesis of meta-evaluation of MTE metrics which treats human evaluation as the gold standard to assess the correlation between metric scores and human scores (Freitag et al., 2022). Utilizing manual MQM scores in the WMT22 dataset (Freitag et al., 2022), we employ 3 statistical methods to evaluate the performance of selected MTE metrics across various scenarios. Below is a concise overview of these 3 methods.

- **Normalized Average Score**

  Both automatic and human evaluations assigned scores to the English translations of 1,875 Chinese sentences generated by 18 MT systems and a group of references. It results in a total of 35,625 scores from each metric. However, the scoring intervals of these metrics vary significantly, highlighting the need for a normalized average score for each metric.

$$\overline{S_{norm}} = \frac{1}{n_{sys}} \sum_{i=1}^{n_{sys}} \frac{1}{n_{sent}} \sum_{i=1}^{n_{sent}} \frac{s_i - s_{min}}{s_{max} - s_{min}} \quad (1)$$

  In the formula above, "*s*" represents the score given by a specific metric for a sentence, and "*n*" denotes the number of sentences or MT

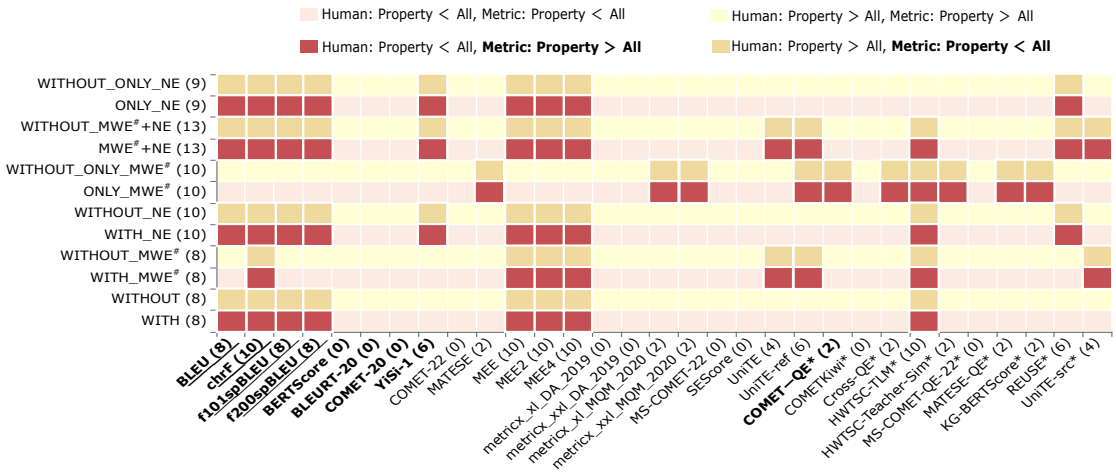

Figure 3: Relationships between the normalized average scores of automatic and human evaluations for each kind of property group. The number in brackets indicates the count of dark-colored squares (red and yellow) in the column or the row. **Dark red** squares represent that the metric **overrates** MT outputs compared to human, while **dark yellow** squares represent that the metric **underrates** MT outputs.

systems. Additionally, "*min*" and "*max*" represent the lowest and highest scores assigned by the metric across all translations. "$\overline{S_{norm}}$" is rightly the normalized average score of the present metric.

- **Kendall Rank Correlation Coefficient**

  The Kendall rank correlation coefficient is a similarity indicator that compares 2 sets of rankings assigned to the same set of objects. To compute Kendall correlations between metric and human scores, we apply MTME[2], a tool recommended by WMT.

- **Paired T-test**

  The paired t-test allows us to determine whether there exists a significant difference between two sets of correlations observed in the distinct property or category groups and the t-statistic value aids in identifying the nature of such difference.

4. Experimental Results

This section presents the quantitative results of our study from the perspectives of property and category level respectively.

4.1. Property Level

4.1.1. Normalized Average Score

Figure 3 presents several interesting points. Here, dark red and light red indicate that the human average score on translations of the corresponding sentence group is lower than that of all translations; dark yellow and light yellow indicate that the human average score on translations of the corresponding sentence group is higher than that of all translations. We can see that most MT systems perform worse on sentence groups with MWEs# and/or NEs from their relatively lower human average scores.

From another perspective, dark-colored squares (red or yellow), representing discrepancies in the judgment of translation quality between metrics and humans, account for 31.2% (116 out of 372) across all property groups. The dark red squares indicate that the metrics give higher evaluation scores on the sentence groups with MWEs# and/or NEs while humans give them lower scores. Conversely, the dark yellow squares show that the metrics give lower evaluation scores on the sentence groups often without MWEs# and/or NEs while humans give higher scores. This suggests that many MTE metrics fail to adequately address translation errors associated with Chinese MWEs.

Regarding the horizontal axis, it is clear that all string-overlap-based metrics (with _), viz. BLEU (8 discrepancies), chrF (10), f101spBLEU (8), and f200spBLEU (8) exhibit low abilities although several neural-network-based metrics, including MEE (10), MEE2 (10), MEE4 (10), and HWTSC-TLM* (10) are also low-performing. On the other hand, out of a total of 27 neural-network-based metrics, 10 (35.7%) metrics achieve the best performance (0) by consistently aligning with human scores across all property groups. Moreover, considering reference-dependent metrics and reference-free metrics (with *), the proportions of discrepancies are 34.1% (86 out of 252) and 25% (30 out of 120), respectively. This implies that reference-free

[2] https://github.com/google-research/mt-metrics-eval

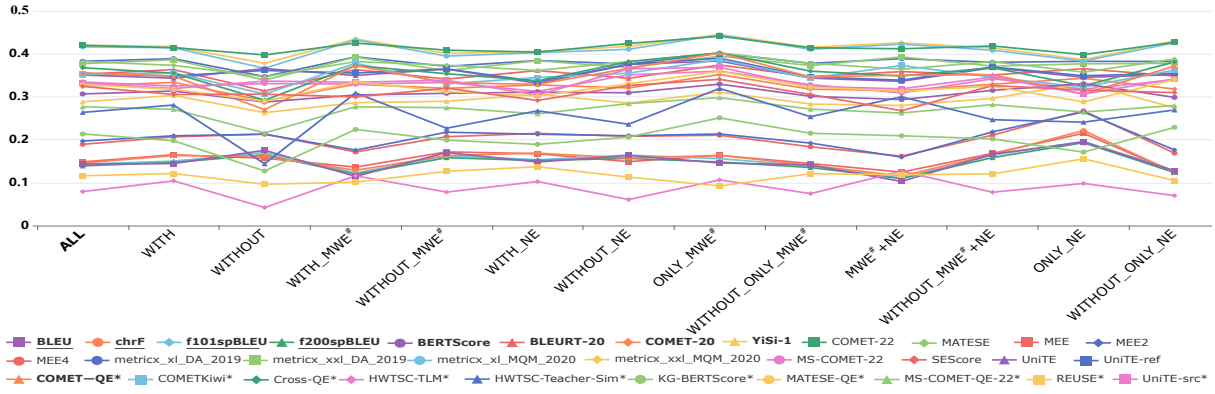

Figure 4: The chart of Kendall correlations between automatic and human evaluations at property level.

metrics perform averagely better than reference-dependent metrics, except for some outliers like HWTSC-TLM* (10) and REUSE* (6).

4.1.2. Kendall Rank Correlation Coefficient

Figure 5 demonstrates that most automatic evaluations have a correlation of around 0.2 to 0.4 with human evaluations in all property groups, indicating ample room for their improvement. It is worth noting that for sentences containing MWEs, except for those with MWE$^{\#}$+NE and WITH_NE, the correlations tend to go higher than those of their complementary property groups. Referring to the number of affected metrics of the MWE$^{\#}$+NE (13) group in Figure 3 which is the highest among all groups, we think that the co-occurrences of MWEs$^{\#}$ and NEs truly cause issues for automatic metrics. However, this finding requires further confirmation in the subsequent discussion.

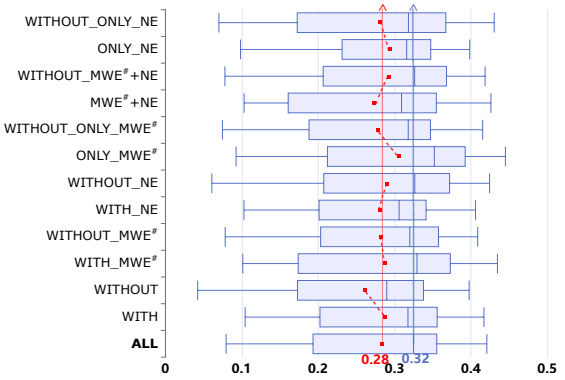

Figure 5: The boxplot of Kendall correlations between automatic and human evaluations at the property level. The blue line and the red dot in the box mark the median and the mean respectively.

Detailed insights into the Kendall correlation between each automatic evaluation and human evaluation are provided by Figure 4. String-overlap-based metrics generally have lower correlations (~0.15) than most neural-network-based metrics (0.3-0.4), with 2 exceptions being underperforming all other metrics (~0.1), i.e. HWTSC-TLM* and REUSE*. COMET-22, metricx_xl_MQM_2020, and metricx_xxl_MQM_2020 consistently rank among the top 3 metrics, with correlations always exceeding 0.4. Overall, most reference-dependent metrics and reference-free metrics have similar correlations which range between 0.2 and 0.3, indicating their comparable performance.

Variable 1	Variable 2	T-statistic	P-value
WITH_MWE$^{\#}$	WITHOUT_MWE$^{\#}$	0.50	6.21E-01
WITH_NE	WITHOUT_NE	-1.86	7.22E-02
ONLY_MWE$^{\#}$	WITHOUT_ONLY_MWE$^{\#}$	10.05	**4.11E-11**
MWE$^{\#}$+NE	WITHOUT_MWE$^{\#}$+NE	-2.77	**9.42E-03**
ONLY_NE	WITHOUT_ONLY_NE	0.94	3.57E-01
WITH	WITHOUT	3.97	**4.20E-04**

Table 2: Results of the paired t-test at the property level. Original data, i.e. Kendall correlations, meet the requirement of normal distribution.

4.1.3. Paired T-test

To validate our previous findings, we conduct paired t-tests based on Kendall correlations. Table 2 presents significant differences ($p<0.05$) in 3 property pairs: 1) ONLY_MWE$^{\#}$ and WITHOUT_ONLY_MWE$^{\#}$ (p=4.11E-11); 2) MWE$^{\#}$+NE and WITHOUT_MWE$^{\#}$+NE (p=9.42E-03); and 3) WITH and WITHOUT (p=4.20E-04). These results underscore the noticeable influence of MWEs on the performance of MTE metrics.

Remarkably, the t-statistic for the pair ONLY_MWE$^{\#}$ and WITHOUT_ONLY_MWE$^{\#}$ (10.05) indicates that the presence of only MWEs$^{\#}$ has an overwhelmingly positive impact on the correlations between automatic and human evaluations. This suggests that the rankings generated by metrics align more closely with those by humans when MWEs are present. Consequently, the correlations for sentences containing MWEs (WITH) are higher (3.97) than those for sentences

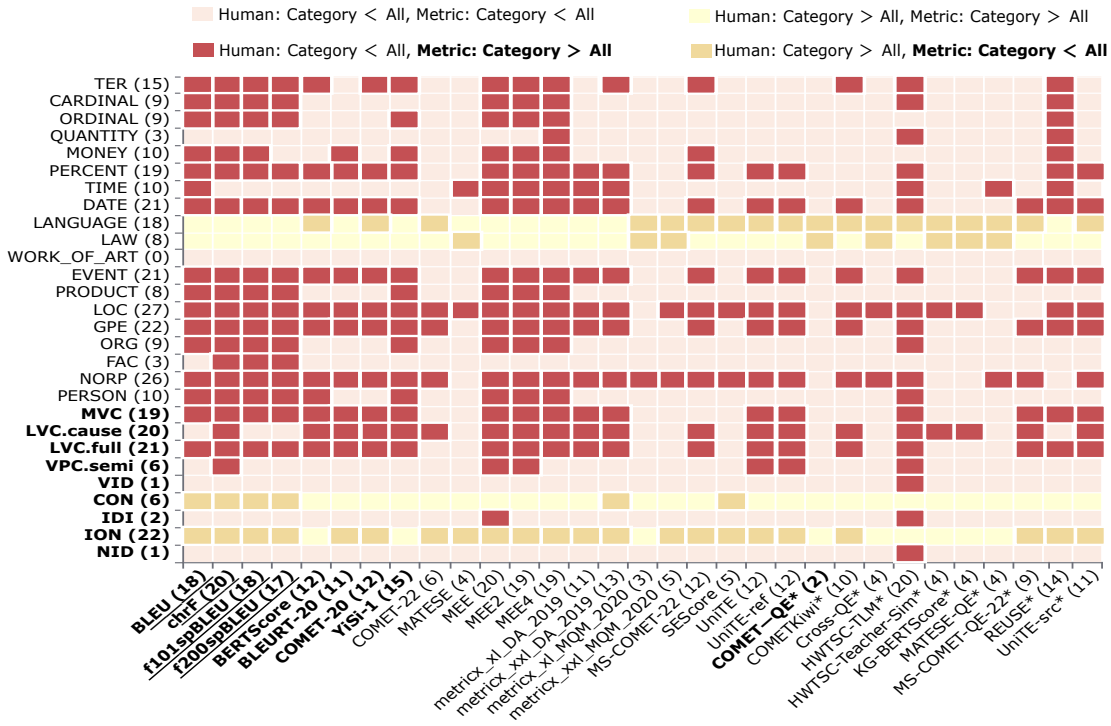

Figure 6: Relationships between the normalized average scores of automatic and human evaluations for each kind of category group. The number in brackets indicates the count of dark-colored squares (red and yellow) in the column or the row. **Dark red** squares represent that the metric **overrates** MT outputs compared to human, while **dark yellow** squares represent that the metric **underrates** MT outputs.

without any annotations (WITHOUT). On the other hand, the co-occurrence of MWEs# and NEs in a sentence (-2.77) dramatically deteriorates the metrics' performance. Overall, it seems that NEs consistently decrease the correlations between automatic and human evaluations, while MWEs# tend to increase them. However, these puzzling observations require profound discussion in conjunction with the results of the paired t-tests in Section 4.2.3.

4.2. Category Level

4.2.1. Normalized Average Score

What stands out in Figure 6 is the significant area in red, indicating that MT systems tend to perform poorly when confronting sentences with various types of MWEs# and NEs. This often leads to low scores assigned by human evaluators. In contrast, a considerable proportion of dark red squares shows a great tendency of automatic metrics to give over-high evaluations on translations with all kinds of MWEs# and NEs, only except for LANGUAGE, LAW, CON, and ION which tend to be underestimated. This contrast claims an interesting observation: MT systems struggle with sentences containing different MWEs, while automatic metrics often overlook the mistakes and assign higher scores to these translations.

Among all categories, 11 types of MWEs (4 MWEs# and 7 NEs) impact over half of automatic metrics, that is, each with a number over 16, such as ION (22), LVC.full (21), LVC.cause (20), MVC (19) and NORP (26). Additionally, it is surprising that WORK_OF_ART (0) triggers no discrepancy and NID (1), IDI (2), and VID (1) have minor impacts on automatic metrics, affecting no more than 3 metrics.

Analyzing the number of MWE categories causing discrepancies for each metric yields similar results to the property-level analysis. Specifically, 4 string-overlap-based metrics are affected by over 16 types of MWEs, showing poor performance. Among the neural-network-based metrics, MEE (20), MEE2 (19), MEE4 (19), and HWTSC-TLM* (20) exhibit particularly low performance. Instead, 10 metrics based on neural networks achieve better performance with discrepancies of less than 7, for instance, COMET-22 (7), MATESE (4), and others. Reference-dependent metrics experience discrepancies in 44.9% (264 out of 588) of cases, while reference-free metrics see discrepancies in 29.3% (82 out of 280) of cases. This suggests that, overall, reference-free metrics tend to provide more accurate evaluations compared to reference-dependent metrics. Interestingly, HWTSC-TLM* (20) still remains an exception in this regard.

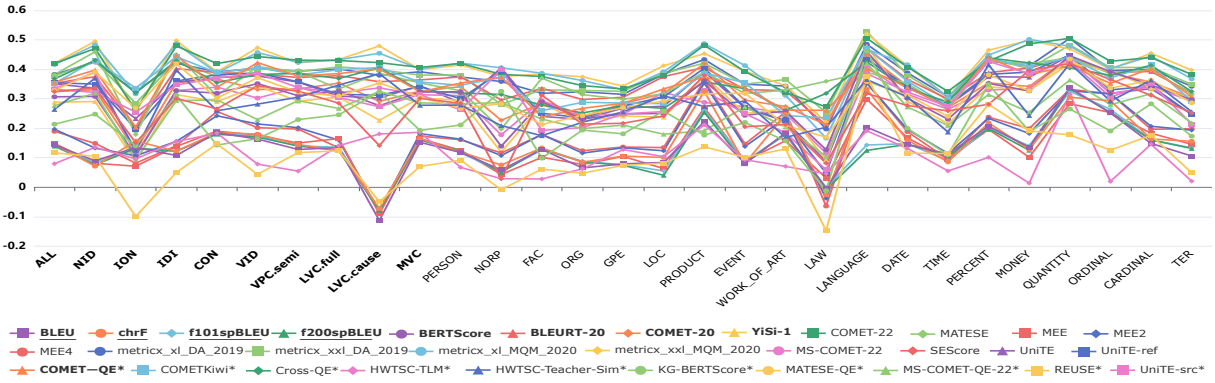

Figure 7: The chart of Kendall correlations between automatic and human evaluations at category level.

4.2.2. Kendall Rank Correlation Coefficient

The distribution of Kendall correlations in Figure 8 shows how automatic evaluation metrics react to different categories of MWEs. The category labels on the y-axis are arranged according to their mean correlations, and the intersecting point between the red and blue lines in the figure highlights noteworthy findings. They indicate that 15 categories under the intersecting point, including 3 types of MWEs[#] and 12 types of NEs, tend to cause low correlations, particularly in cases of LAW and ION. Some negative correlations can even be observed in those category groups, which are caused by REUSE* and BLEU as shown in Figure 7. By contrast, the other 13 categories, comprising 6 MWEs[#] and 7 NEs above the intersecting point lead to higher correlations, especially PRODUCT, PERCENT, LANGUAGE, and QUANTITY. Moreover, deep into Figure 7, we can observe that the 4 outliers in ORDINAL and LANGUAGE are attributed to HWTSC-TLM*, REUSE*, and f101spBLEU, f200spBLEU respectively.

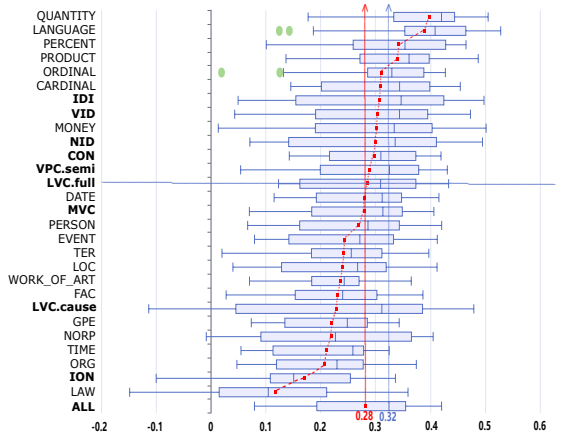

Figure 8: The boxplot of Kendall correlations between automatic and human evaluations at the category level. The blue line and the red dot in the box mark the median and the mean respectively.

Variable 1	Variable 2	T-statistic	P-value	Variable 2	T-statistic	P-value
ALL	**NID**	-1.68	1.03E-01	LOC	5.39	**7.72E-06**
	ION	8.41	**2.19E-09**	PRODUCT	-4.74	**4.81E-05**
	IDI	-3.12	**3.97E-03**	EVENT	6.50	**3.49E-07**
	CON	-1.73	9.34E-02	WORK_OF_ART	3.98	**4.07E-04**
	VID	-3.53	**1.37E-03**	LAW	12.78	**1.14E-13**
	VPC.semi	-3.19	**3.34E-03**	LANGUAGE	-8.36	**2.51E-09**
	LVC.full	-0.40	6.93E-01	DATE	0.50	6.22E-01
	LVC.cause	2.41	**2.25E-02**	TIME	15.20	**1.23E-15**
	MVC	0.51	6.13E-01	PERCENT	-7.26	**4.35E-08**
	PERSON	3.96	**4.27E-04**	MONEY	-2.87	**7.50E-03**
	NORP	4.23	**2.01E-04**	QUANTITY	-12.94	**8.26E-14**
	FAC	7.26	**4.39E-08**	ORDINAL	-3.39	**1.99E-03**
	ORG	14.95	**1.91E-15**	CARDINAL	-6.52	**3.26E-07**
	GPE	11.69	**1.06E-12**	TER	6.44	**4.16E-07**

Table 3: Results of the paired t-test at the category level. Original data, i.e. Kendall correlations, meet the requirement of normal distribution.

The correlation overview depicted in Figure 7 reveals a similar pattern to that in Figure 8. Besides, it is apparent in Figure 7 that all automatic metrics exhibit opposite tendencies of correlations in the categories of LVC.cause and NORP. This divergence indicates the varying abilities of these metrics to detect errors in these 2 kinds of MWEs.

Furthermore, we can find that HWTSC-TLM* performs the worst in VPC.semi, FAC, WORK_OF_ART, TIME, PERCENT, MONEY, ORDINAL, and TER, while REUSE* underperforms in the remaining categories. Conversely, metricx_xl_MQM_2020, metricx_xxl_MQM_2020, and COMET-22 consistently exhibit superior performance across all types of MWEs[#] and NEs. Generally, string-overlap-based metrics have lower correlations (~0.1) compared to neural-network-based metrics, whose correlations tend to range from approximately 0.2 to 0.5. The majorities of reference-dependent metrics and reference-free metrics always show similar correlations, ranging from 0.2 to 0.4.

4.2.3. Paired T-test

The results in Table 3 display significant differences ($p<0.05$) in correlations for 5 of the 9 MWE[#] categories. Sentences containing IDI (-3.12), VID (-3.53), and VPC.semi (-3.19) lead to improved

correlations of evaluations, suggesting that these types of MWEs# can be more reliably evaluated by automatic metrics. However, ION (8.41) and LVC.cause (2.41) pose challenges for automatic metrics to accurately rank translations. These findings are consistent with those in Section 4.2.2.

Regarding NEs, 18 out of 19 types, excluding DATE (p=6.22E-01), have diverse significant impacts. Sentences containing PRODUCT (-4.74), LANGUAGE (-8.36), PERCENT (-7.26), MONEY (-2.87), QUANTITY (-12.94), ORDINAL (-3.39), and CARDINAL (-6.52) have positive impacts on automatic metrics, indicating that the metrics perform well in identifying errors in the translations of these types of NEs. However, sentences containing the remaining 11 types of NEs, such as PERSON (3.96) and NORP (4.23), worsen the performance of automatic metrics.

These fine-grained results provide a clearer picture of the mixed effects from different MWEs# and NEs at the category level. It becomes evident that the findings at the property level are a combination of these effects, which often counteract each other. Therefore, understanding the impact of specific types of MWEs# and NEs is crucial for more accurate MTE.

5. Discussion

The results above provide important insights into the performance of different MTE metrics on evaluation tasks with a variety of Chinese MWEs.

In terms of normalized average scores at both property and category levels, the results show that: 1) MT systems still struggle with most types of MWEs# and NEs, and 2) MTE metrics tend to neglect translation errors in sentences containing MWEs# and NEs, leading to an overestimation of MT performance. The number of discrepancies (represented by dark-colored squares in Figure 3 and Figure 6) reveals the inferior performance of string-overlap-based metrics compared to neural-network-based ones. Meanwhile, reference-dependent metrics exhibit slightly lower performance than reference-free metrics when it comes to Chinese MWEs, hinting at potential bias in evaluations due to reliance on gold reference translations since there can be alternative good translations for the same sentence with different styles, lexical items, and so on.

Analysis of Kendall rank correlation coefficients reveals that the accuracy of current MTE metrics is far from perfect. In detail, string-overlap-based metrics are worse than neural-network-based ones, while reference-dependent metrics demonstrate similar performance to reference-free metrics. The conflict seems to arise when looking at some specific categories that have positive impacts on correlations but are inclined to receive over-high average scores, like PERCENT and DATE. However, this can be explained by the fact that Kendall correlations are computed based on rankings, where discrepancies in normalized average scores may not affect the order of sentence scores, resulting in higher correlations with human-made rankings.

Paired t-tests further elucidate our findings on the impacts of different Chinese MWEs on MTE metrics. It is important to note that the fine-grained meta-evaluations are necessary because the impacts of different MWEs# and NEs differ significantly and can be easily overlooked when they are mixed at the property level. Generally, most NEs exert a greater influence on MTE metrics than MWEs#. Interestingly, what is out of our intuitions is that NIDs have minimal impacts on automatic metrics in terms of both average scores and correlations. This could be attributed to the high semantic complexity of NIDs (Constant et al., 2017), leading to lower translation quality. Consequently, automatic metrics, which are more sensitive to severe errors (Ma et al., 2019), may find it easier to accurately capture errors in NID translations.

In summary, reference-dependent metrics based on neural networks demonstrate the best performance in the current task, tightly followed by reference-free metrics. On the other hand, string-overlap-based metrics consistently perform the worst among nearly all metrics.

6. Conclusion

In this study, we conducted a comprehensive meta-evaluation on the state-of-the-art MTE metrics in scoring English translations of Chinese sentences, with a special concern on a variety of Chinese MWEs. Our investigation probed into the potential impacts of different types of Chinese MWEs on MTE metrics across 2 levels of analysis. The results reveal a general pattern of poor performance of MT systems on most Chinese MWEs. Similarly, MTE metrics show deficiencies in evaluating translations containing MWEs by consistently overrating them. From another perspective, neural-network-based metrics always have better performance than string-overlap-based metrics and reference-free metrics have comparable effectiveness as reference-dependent metrics in general.

While our study reached a conclusion consistent with that of Freitag et al. (2022), we can offer more detailed insights for the future improvement of MT systems and MTE metrics, particularly regarding different categories of MWEs. Additionally, we extended the annotation of the WMT22 dataset which is publicly available for future studies.

7. Limitations

During the study, we recognized several avenues for improvement in continuous studies. Firstly, the size of basic data can be enlarged to cover more domains, languages, and MWE categories. Secondly, we would like to move beyond the statistics by delving into the specific MQM-based errors in MWE translations. Moreover, the calculation can be more convincing by balancing the proportions of different MWE items.

8. Bibliographical References

Eleftherios Avramidis and Vivien Macketanz. 2022. Linguistically motivated evaluation of machine translation metrics based on a challenge set. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 514–529, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Bogdan Babych and Anthony Hartley. 2010. Automated error analysis for multiword expressions: Using BLEU-type scores for automatic discovery of potential translation errors. *Evaluation of Translation Technology*, 8:81–104.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. *Handbook of Natural Language Processing*, 2:267–292.

Anabela Barreiro, Johanna Monti, Brigitte Orliac, and Fernando Batista. 2013. When multiwords go bad in machine translation. In *Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technologies*, pages 26–33, Nice, France.

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2012. Automatic construction of a MultiWord expressions bilingual lexicon: A statistical machine translation evaluation perspective. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon*, pages 95–108, Mumbai, India. The COLING 2012 Organizing Committee.

Nicoletta Calzolari, Charles J Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, pages 1934–1940, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Marine Carpuat and Mona Diab. 2010. Task-based evaluation of multiword expressions: A pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 242–245, Los Angeles, California. Association for Computational Linguistics.

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.

Daiga Deksne, Raivis Skadiņš, and Inguna Skadiņa. 2008. Dictionary of multiword expressions for translation into highly inflected languages. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 1401–1405, Marrakech, Morocco. European Language Resources Association (ELRA).

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2011. Integration of reduplicated multiword expressions and named entities in a phrase based statistical machine translation system. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1304–1312, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Emmanuelle Esperança-Rodier and Johan Didier. 2016. Translation quality evaluation of MWE from French into English using an SMT system. In *Proceedings of Translating and the Computer 38*, pages 33–41, London, UK. AsLing.

Emmanuelle Esperança-Rodier and Damian Frankowski. 2021. DeepL vs Google translate: who's the best at translating MWEs from French into Polish? a multidisciplinary approach to corpora creation and quality translation of MWEs. In *Translating and the Computer 43*, pages 110–127.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chikiu 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.

Mahmoud Ghoneim and Mona Diab. 2013. Multiword expressions in the context of statistical machine translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1181–1187, Nagoya, Japan. Asian Federation of Natural Language Processing.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Lifeng Han. 2022. *An Investigation into Multi-word Expressions in Machine Translation*. Ph.D. thesis, Dublin City University.

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.

Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. 28. MIT Press.

Tom Kocmi, Hitokazu Matsushita, and Christian Federmann. 2022. MS-COMET: More and better human judgements improve metric performance. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 541–548, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Valia Kordoni and Iliana Simova. 2014. Multiword expressions in machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1208–1211, Reykjavik, Iceland. European Language Resources Association (ELRA).

Yilun Liu, Xiaosong Qiao, Zhanglin Wu, Su Chang, Min Zhang, Yanqing Zhao, Song Peng, Shimin Tao, Hao Yang, Ying Qin, Jiaxin Guo, Minghan Wang, Yinglu Li, Peng Li, and Xiaofeng Zhao. 2022. Partial could be better than whole. HW-TSC 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 549–557, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Chi-kiu Lo. 2019. YiSi-a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

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.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Ananya Mukherjee, Hema Ala, Manish Shrivastava, and Dipti Misra Sharma. 2020. MEE: An automatic metric for evaluation using embeddings for machine translation. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 292–299. IEEE.

Ananya Mukherjee and Manish Shrivastava. 2022a. REUSE: REference-free UnSupervised quality Estimation metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 564–568, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ananya Mukherjee and Manish Shrivastava. 2022b. Unsupervised embedding-based metric for MT evaluation with improved human correlation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 558–563, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Min Sik Oh. 2022. kpfriends at SemEval-2022 task 2: NEAMER - named entity augmented multi-word expression recognizer. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 178–185, Seattle, United States. Association for Computational Linguistics.

Santanu Pal, Sudip Kumar Naskar, Pavel Pecina, Sivaji Bandyopadhyay, and Andy Way. 2010.

Handling named entities and compound verbs in phrase-based statistical machine translation. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pages 46–54, Beijing, China. Coling 2010 Organizing Committee.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. 2022. MaTESe: Machine translation evaluation as a sequence tagging problem. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 569–577, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

Carlos Ramisch, Abigail Walsh, Thomas Blanchard, and Shiva Taslimipoor. 2023. A survey of MWE identification experiments: The devil is in the details. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 106–120, Dubrovnik, Croatia. Association for Computational Linguistics.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André FT Martins, and Alon Lavie. 2021. Are references really needed? Unbabel-IST 2021 submission for the metrics thared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. 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 *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15, Mexico City, Mexico. Springer.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983, Denver, Colorado. Association for Computational Linguistics.

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.

Nina Schottmüller and Joakim Nivre. 2014. Issues in translating verb-particle constructions from German to English. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 124–131, Gothenburg, Sweden. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

NLLB Team, Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. ID10M: Idiom identification in 10 languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726, Seattle, United States. Association for Computational Linguistics.

Veronika Vincze, István Nagy, 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.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek F Wong, and Lidia S Chao. 2022. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. 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.

Wenda Xu, Yi-Lin Tuan, Yujie Lu, Michael Saxon, Lei Li, and William Yang Wang. 2022. Not all errors are equal: Learning text generation metrics using stratified error synthesis. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6559–6574, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Xueying Zhang and Hui Zhang. 2014. *A Comprehensive Dictionary of Chinese/English Idioms with English/Chinese Translations*. Tsinghua University Press, Beijing.

Category	Sum1	Sum2	Number	Deduplicated Number	Sentence Number	Explanation	Example
NID			137	81	117	Noun-headed idioms or multiword units function as nouns	巨无霸 (jù wú bà) (lit. too huge to be bullied) the giant
ION			5	4	5	Separable words	成不了像 (chéng bù liǎo xiàng) (lit. to make no image) fail to make an image
IDI		**Number: 1,430 Deduplicated Number: 950**	265	239	200	Fixed Chinese idioms consist of four or more characters	雪中送炭 (xuě zhōng sòng tàn) (lit. to send charcoal in snowy weather) to send help in one's need
CON			87	34	82	Syntactically special constructions	除 _ 外 (chú_ wài) (lit. apart from) except for
VID			83	65	78	Verb-headed idioms or multiword units function as verbs	下决心 (xià jué xīn) (lit. to set down the determination) to make up one's mind
VPC.semi			397	212	343	Semi non-compositional verb-particle constructions	意识到 (yì shí dào) (lit. to be aware of)
LVC.full			195	154	163	Light verb constructions with bleached verbs	发表演讲 (fā biǎo yǎn jiǎng) (lit. to give a speech)
LVC.cause			35	30	34	Light verb constructions with causative verbs	引发破坏 (yǐn fā pò huài) (lit. to lead to damage)
MVC			226	131	198	Multi-verb constructions	试试看 (shì shì kàn) (lit. to try and see) to have a try
PERSON	**Number: 4,257 Deduplicated Number: 2,537**	**Number: 2,827 Deduplicated Number: 1,587**	262	166	176	Specific appellations or names of people	李效良 (lǐ xiào liáng) (lit. Li, Xiaoliang)
NORP			21	7	16	Nationalities, religious, political or ethnic groups	中华民族 (zhōng huá mín zú) (lit. Chinese nation)
FAC			53	29	46	Buildings or other concrete facilities in society	天安门广场 (tiān ān mén guǎng chǎng) (lit. Tian Anmen Square)
ORG			414	223	267	Companies, agencies, or institutions	湖南日报 (hú nán rì bào) (lit. Hunan Daily)
GPE			400	125	225	Countries, cities, or states	青海省 (qīng hǎi shěng) (lit. Qinghai Province)
LOC			76	33	56	Non-GPE locations, like mountain ranges and bodies of water	青藏高原 (qīng zàng gāo yuán) (lit. Qinghai-Tibet Plateau)
PRODUCT			48	20	38	Man-made staff, like vehicles, weapons, or foods	比特币 (bǐ tè bì) (lit. Bitcoin)
EVENT			111	79	83	Specific activities, like named hurricanes, battles, or sports events	民主峰会 (mín zhǔ fēng huì) (lit. Summit of Democracy)
WORK_OF_ART			122	98	67	Titles of books, songs, or paintings	《查理和巧克力工厂》(chá lǐ hé qiǎo kè lì gōng chǎng) (lit. Charlie and the Chocolate Factory)
LAW			6	6	6	Named documents made into laws	《刑法》(xíng fǎ) (lit. Criminal Law)
LANGUAGE			13	4	10	Named languages	日文 (rì wén) (lit. Japanese)
DATE			519	285	387	Absolute or relative dates or periods	5 月 20 日 (5 yuè 20 rì) (lit. on May 20th)
TIME			101	75	83	Times shorter than a day	四分钟 (sì fēn zhōng) (lit. four minutes)
PERCENT			50	37	41	Proportions or percentages	四成以上 (sì chéng yǐ shàng) (lit. more than 40%)
MONEY			68	55	49	Monetary values, including units	十元 (shí yuán) (lit. ten yuan)
QUANTITY			47	42	25	Measurements, as of weight or distance	几千公里 (jǐ qiān gōng lǐ) (lit. thousands of kilometers)
ORDINAL			104	66	96	Numbers used to order	第一次 (dì yī cì) (lit. at the first time)
CARDINAL			223	161	171	Numbers and measure words used to count	两匹 (liǎng pǐ) (lit. two 'horses')
TER			189	76	127	Domain-specific terminology	石英 (shí yīng) (lit. Quartz)

Table 4: Categorization Framework and Annotation Results of 28 types of Chinese MWEs (including 9 types of **MWEs#** and 19 types of NEs).