MMTE: Corpus and Metrics for Evaluating Machine Translation Quality of Metaphorical Language

Shun Wang[1], **Ge Zhang**[3,5], **Han Wu**[4], **Tyler Loakman**[1], **Wenhao Huang**[5], **Chenghua Lin**[1,2]*
[1]Department of Computer Science, The University of Sheffield, UK
[2]Department of Computer Science, The University of Manchester, UK
[3]Department of Computer Science, University of Waterloo, Canada
[4]School of Foreign Studies, UIBE, Beijing, China [5]01.AI, Beijing, China
{swang209, tcloakman1}@sheffield.ac.uk, zhangge@01.ai
wuhan@uibe.edu.cn, chenghua.lin@manchester.ac.uk

Abstract

Machine Translation (MT) has developed rapidly since the release of Large Language Models and current MT evaluation is performed through comparison with reference human translations or by predicting quality scores from human-labeled data. However, these mainstream evaluation methods mainly focus on fluency and factual reliability, whilst paying little attention to figurative quality. In this paper, we investigate the figurative quality of MT and propose a set of human evaluation metrics focused on the translation of figurative language. We additionally present a multilingual parallel metaphor corpus generated by post-editing. Our evaluation protocol is designed to estimate four aspects of MT: Metaphorical Equivalence, Emotion, Authenticity, and Quality. In doing so, we observe that translations of figurative expressions display different traits from literal ones.

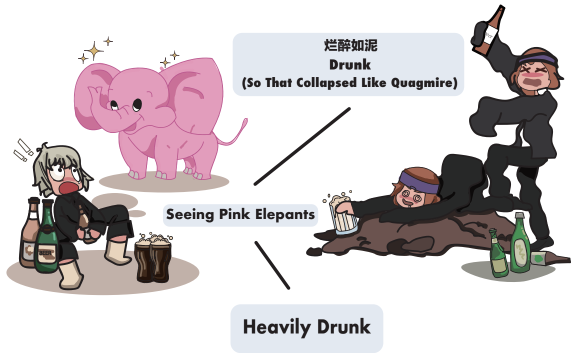


Figure 1: Chinese and English metaphorical expressions of being drunk.

1 Introduction

Metaphorical expressions are widely used in daily life for communication and vivid description, drawing attention from psycholinguistics and computational linguistics due to their key role in the cognitive and communicative functions of language (Wilks, 1978; Lakoff and Johnson, 1980; Lakoff, 1993). Linguistically, a metaphor is defined as a figurative expression that uses one or more words to represent another concept within a given context, rather than taking the literal meaning of the expression (Fass, 1991). For instance, in the sentence "*The scream pierced the night.*", the contextual meaning of *pierced* is to "sound sharply or shrilly", which differs from its literal meaning of "cut or make a way through".[1]

A significant portion (e.g. up to 20%) of our everyday language is delivered in metaphorical terms (Steen, 2010). According to Lakoff and Johnson's study, metaphor is a type of conceptual mapping. The cognitive model, which involves reasoning about one thing in terms of another, has been shown to affect our decision-making and perception (Lakoff and Johnson, 1980; Lakoff, 1993; Boroditsky, 2011). Research also suggests that this concept-to-concept mapping is often language-agnostic, with similar mappings being feasible across different languages (Tsvetkov et al., 2014). For instance, in the aforementioned example, the English word "*pierce*" corresponds to the Chinese word "穿透", which literally means "pass through an object or medium" and is also used as a metaphor to indicate a sudden and sharp sound.

However, direct translations cannot always be found in the target language due to linguistic and cultural differences. To illustrate this issue, we provide an example of metaphorical expressions of being *drunk* in Figure 1, where in Chinese it is common to compare drunkenness to being *collapsed on the ground like quagmire*, whilst in English it is common to compare it to *seeing pink elephants*. These word-sense misalignments caused by dif-

*Corresponding author
[1]http://wordnetweb.princeton.edu/perl/webwn

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 11343–11358
November 12-16, 2024 ©2024 Association for Computational Linguistics

ferent linguistic norms are ubiquitous in practical translation applications.

Metaphor, and especially metaphor *translation*, has received increasing attention in linguistics (Qin and Peng, 2022; Anvarovna, 2022; Li et al., 2023), and its significance has also been highlighted in various NLP tasks similar to translation, such as poem writing (Chakrabarty et al., 2020), story generation (Chakrabarty et al., 2021), and dialogue (Oprea et al., 2022). Unfortunately, the challenge of machine translating metaphorical language remains largely unaddressed due to a scarcity of resources such as parallel data (Mao et al., 2018; Gamonal, 2022a). To remedy this, we propose MMTE - the first systematic study of **M**etaphorical **M**achine **T**ranslation **E**valuation to explore the difficulties inherent in translating metaphorical expressions. Our contributions include:

- ***Corpus***: The first manually annotated multilingual metaphor translation evaluation corpus between English and Chinese/Italian.
- ***Human Evaluation Framework***: The first systematic human evaluation framework for metaphor translation. We also introduce rhetorical **Equivalence** for metaphorical translation Quality Estimation (QE).
- ***Theoretical Linguistic Foundations***: We demonstrate the difficulties of metaphor translation from a multilingual and multi-perspective approach and provide a systematic framework for metaphor translation.

2 Related Work

Metaphor-related Tasks. Metaphors play a crucial role in daily communication and understanding human emotion (Mohammad et al., 2016) and cognition (Tong et al., 2021). Enhancing the quality of our understanding of metaphors has been shown to be crucial for various natural language understanding (NLU) tasks, including natural language inference (NLI) (Stowe et al., 2022), sentiment analysis (Alsiyat and Piao, 2020; Li et al., 2022a), humour explanation (Mittal et al., 2022), and offensive language detection (Tang et al., 2020).

Moreover, adequately conveying metaphor is a staple concern in various natural language generation (NLG) tasks, including poem writing (Liu et al., 2019; Chakrabarty et al., 2020), paraphrasing (Bizzoni and Lappin, 2018; Stowe et al., 2021a), dialogue generation (Zheng et al., 2019; Oprea et al., 2022), and story generation (Chakrabarty et al., 2021). Appropriate use of metaphor has been shown to dramatically improve user satisfaction with such systems (Li et al., 2022b), and exploring the mechanisms behind metaphor generation helps test and verify cognitive theories of how metaphors are created and used (Lederer, 2016; Dankers et al., 2019). Standalone metaphor generation has also been identified as a significant branch of creative NLG in itself (Stowe et al., 2021b; Chakrabarty et al., 2021; Li et al., 2022b; Ge et al., 2023; Shao et al., 2024).

Existing multilingual metaphor research (Mohler et al., 2014; Kozareva, 2015; Gordon et al., 2015) has primarily focussed on multilingual metaphor detection and identification guided by metaphorical mappings (Shutova et al., 2017), the polarity and valence of multilingual metaphors (Kozareva, 2015), and general metaphor frames (Gamonal, 2022b; Aghazadeh et al., 2022). Researchers have also explored metaphor detection and generation mechanisms in languages besides English, including Chinese (Chung et al., 2020; Li et al., 2022b), Malay (Chung, 2005), Arabic (Alsiyat and Piao, 2020), and German (Schneider et al., 2022). Metaphor translation is also a prevalent topic in linguistics (Pranoto, 2021; Qin and Peng, 2022; Anvarovna, 2022), whilst relatively unexplored in the machine translation domain. Whilst some research has explored metaphor's impact on machine translation (Gamonal, 2022a; Li et al., 2024), this has focussed on applying external knowledge bases to enhance cross-lingual metaphor detection. This lack of guidance from professional translators results in translations that lack linguistic nuance.

Fine-grained Translation Quality Estimation. Translation Quality Estimation (QE) has received increased attention, yet remains an open challenge due to resource scarcity and difficulties in handling the variation in linguistic forms and cultural norms that is inherent in metaphor (Vamvas and Sennrich, 2022; Lu et al., 2022). Existing translation QE work, including traditional automatic metrics (Martins et al., 2017; Baek et al., 2020), encoder LM-based metrics (Ranasinghe et al., 2020; Zheng et al., 2021), and generative LLM-based metrics (Kocmi and Federmann, 2023; Lu et al., 2023; Zhao et al., 2024), underestimates differences caused by the cultural phenomena that underlie the use of different languages.

Although some exploratory works investigate the influence of cultural norms in the target language (Vela et al., 2014; Eo et al., 2022), cross-lingual patterns (Zhou et al., 2020), and pivot languages (i.e., an intermediary language for translation between many different languages) on machine translation QE (Zou et al., 2022), few works provide specific metrics to analyse how machine translation models perform on maintaining the linguistic phenomena of the target language, including culture-bound figurative description.

Metaphor Quality Estimation. Metaphor Quality Estimation (MQE) mainly adopts the use of human evaluation (Loakman et al., 2023), or directly compares the meanings of tenors (i.e., the subject of a description) and vehicles (i.e., the figurative language used to describe the tenor) (Li et al., 2022b). However, human evaluation of metaphor leads to difficulty in constructing fair contradistinctions between examples (Zayed et al., 2020), and additionally, directly comparing the meanings of tenors and vehicles ignores the reality that metaphors can be generated on the basis of various different aspects of the vehicles and tenors, which may span long stretches of text, leading to the introduction of noise from misjudgments (Stowe et al., 2021b; Wang et al., 2023). Miyazawa and Miyao (2017, 2019) investigate manual metrics for metaphorical expression. The proposed metrics primarily focused on annotators evaluating metaphoricity by assigning one single score, therefore lacking an in-depth exploration of metaphor types and underlying principles. DiStefano et al. (2024) investigates adopting LLMs for metaphor scoring but only for creativity assessment.

In contrast to existing studies, MMTE proposes the first systematic human evaluation framework for performing comprehensive examinations and evaluations of metaphor translation and assessing its complexities and challenges.

3 Metaphorical Translation Quality Annotation Framework

As discussed in §1, metaphorical expressions are not evaluated sufficiently with current MT evaluation metrics. To address this issue, we propose a set of novel MT evaluation metrics based on manual annotation and post-editing. The proposed metrics aim to provide a more accurate and insightful assessment of MT performance in handling metaphors. Our framework allows for the evaluation of MT outputs in terms of their metaphorical expressions, enabling a more comprehensive analysis of their effectiveness in capturing the nuanced meaning conveyed by such expressions, as demonstrated in Figure 2.

3.1 Initial Dataset Translation

Due to the absence of parallel multilingual metaphor datasets, we constructed our own dataset. We employ the **MOH** dataset (Mohammad et al., 2016) as our source, consisting of 315 metaphorical and 332 literal sentences sampled from WordNet (Miller, 1998). In MMTE, **Literal** samples refer to those not containing metaphors.

We utilise four popular MT models to generate translations: the **Google Cloud Translation API**, the **Youdao Cloud Translation API**, the open-source **Helsinki-NLP/opus-mt** model from Hugging Face, and **GPT-4o** to translate English source data into **Chinese** and **Italian**, enabling us to explore and compare the treatment of metaphors in two languages with distinct characteristics. Table 1 presents example metaphors paired with their translations in the two target languages. Additional information regarding preprocessing is presented in Appendix A.1.

3.2 Metaphor Annotation Criteria

Our annotation protocol involves comparing translations with their source sentences. We hire 18 linguistics majors who are native speakers of the target languages to annotate and post-edit 647 English-Chinese (EN-ZH) and English-Italian (EN-IT) translations, with each sample being annotated by 3 individuals. Professional translators cross-checked the results, resolved disagreements in meetings, and recorded final decisions. Additional details are in Appendix A.2. The source instances and their corresponding translations are systematically annotated based on four criteria to evaluate translation quality: Quality, Metaphorical Equivalence, Emotion, and Authenticity. These criteria are outlined and further broken down as follows.

Quality. To estimate the quality of the translation, we adopt criteria inspired by several existing hu-

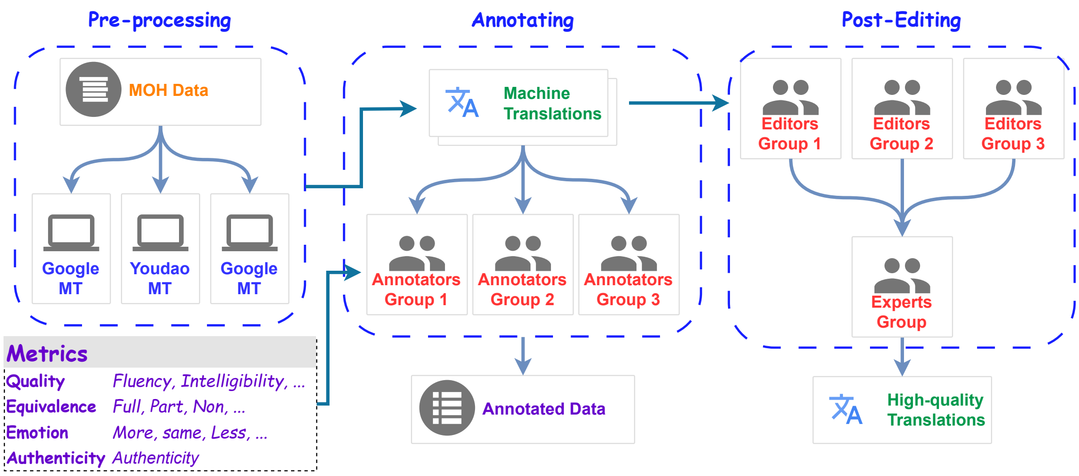


Figure 2: The dataset creation framework. By translating, annotating, and post-editing, we create a cross-lingual metaphor dataset. Specific details of these sub-steps are elaborated in Sections 3.1, 3.2, and 3.3, respectively.

source instances	google-en-zh	youdao-en-zh	opus-mt-en-zh	GPT-4o
The scream **pierced** the night.	尖叫声划破黑夜。	尖叫声划破黑夜。	尖叫声刺穿了夜晚。	尖叫声刺穿了夜晚。
The Senator **steamrollered** the bill to defeat.	参 议 员以压倒性的方式使议案落败。	那位参议员强行使该法案失败。	参议员把法案推倒 了。	参议员将该法案压倒性地击败。
	google-en-it	**youdao-en-it**	**opus-mt-en-it**	**GPT-4o**
The scream **pierced** the night.	L'urlo **squarciò** la notte.	L'urlo **forò** la notte.	L'urlo **ha trafitto** la notte.	L'urlo **ha squarciato** la notte.
The Senator **steamrollered** the bill to defeat.	Il senatore ha **schiacciato** il disegno di legge per sconfiggerlo.	Il senatore ha **buttato via** il disegno di legge per sconfiggerlo.	Il senatore ha **rullato** il conto per sconfiggere.	Il senatore ha **fatto a pezzi** il disegno di legge per sconfiggerlo.

Table 1: Paired samples of source instances and their machine translations from different translation models. Target verbs are in **bold and underlined**.

man assessment methods for MT (Carroll, 1966; Church and Hovy, 1993; White et al., 1994) and consider three primary aspects of quality, including Fluency, Intelligibility, and Fidelity. Detailed definitions are presented in Appendix B.

Equivalence. To ascertain how metaphors impact MT, we propose Equivalence to describe how figurative expressions are translated into another language based on two features: 1) *How the meanings of the source and target are conveyed* 2) *Whether or not the translation is still figurative*. By comparing source texts and translations, annotators are asked to determine to what extent the target word is Equivalent in figuration. The annotators label the translation using a set of five distinct tags, encompassing three types of Equivalence and two types of Mistranslation. We elucidate the types of Equivalence in Table 2, based on the following definitions:

- ***Full-Equivalence***: When comparing the source and translation, both the literal meanings and the contextual meanings of the target word are the same.
- ***Part-Equivalence***: When comparing the source and translation, only the contextual meanings of the target word are similar. The literal meaning of the target word between the source and translation is different, but they are both metaphorical.
- ***Non-Equivalence***: When comparing the source and translation, only the contextual meanings of the target word are similar. However, the translation is a non-metaphorical expression, making the literal meaning of the target words between the source and translation different.

We also identify two types of mistranslation:

- ***Misunderstanding***: When the literal meaning of the target word in the source text and translation are similar, but the translation fails to convey the contextual meaning of the target word in the source language.
- ***Error***: When the target word is mistranslated, meaning that not only the contextual meanings are different between the source and translation, but their literal meanings also differ.

If the source instances are non-metaphorical expressions, annotators are instructed to only clas-

Equivalence	Source	Target
Full	The White House **sits** on Pennsylvania Avenue. The ex-slave **tasted** freedom shortly before she died.	白宫坐落在宾夕法尼亚大道上。 l'ex schiava ha **assaporato** la libertà poco prima di morire.
Part	**Wallow** in your success! My personal feelings **color** my judement in this case.	沉浸在你的成功中吧！ i miei sentimenti personali **offuscano** il mio giudizio in questo caso.
Non	This drug will **sharpen** your vision. Fire had **devoured** our home.	这药能改善你的视力。 l'incendio **distrusse** la casa.

Table 2: Instances of various Equivalence types in metaphor translation. *Full* refers to the same literal and contextual meanings; *Part* means similar contextual meanings and different literal meanings while both being metaphorical; and *Non* means similar contextual meanings and different literal meanings with the translation being non-metaphorical.

sify the translations into three categories: ***Literal***, ***Metaphorical***, and ***Error***. The non-metaphorical portion of the data is used for subsequent comparisons with the metaphorical instances.

Emotion. Inspired by Mohammad et al. (2016), we incorporate an analysis of emotion to investigate whether metaphorical expressions in translations convey additional emotional information compared to non-metaphorical expressions. By comparing a source sentence and its translation, the annotators determine to what extent the target word and its translation convey different amounts of emotion. There are four labels to judge emotion: ***Zero***, ***Less***, ***Same***, and ***More***, separately representing that the target word in the source context conveys no emotion, or that the target word in the translation conveys less, the same, or more emotion than the target word in the source sentence.

Authenticity. Authenticity is an extension of existing criteria (Doyon et al., 1999), evaluating: *To what extent the translated metaphor reads like standard, well-edited language, such that the metaphor would be understood by a native speaker of the target language.* The annotators are asked to judge all aforementioned criteria on a 5-point Likert scale (Likert, 1932).

3.3 Post-Editing

Due to the requirement for gold references by automatic evaluation algorithms like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), we introduce a post-editing method to modify the translation results of four MT models to generate a gold standard translation reference, as is common practice (Senez, 1998; Allen, 2003; Somers, 2003). Three groups of annotators, who are native speakers of each target language, are asked to post-edit the translations, resulting in three groups of human-edited translations for both Chinese and Italian. Finally, a panel of expert translators perform final filtering to select the best quality edited translation as the gold reference. Additional details regarding the human annotation process are presented in Appendix B.

We employ several quality control methods to ensure the quality of the dataset obtained through post-editing the machine translations. Annotators compare four different translations, selecting high-quality ones or modifying low-quality ones to provide a reference translation, including translations of both metaphorical and non-metaphorical language. Three separate annotator groups work on each sample. An expert panel of translators then reviewes and refines the selections. Annotators also mark the positions of target words during alignment to avoid issues in word-level processing. The final dataset includes aligned English, Chinese, and Italian translations, with 315 metaphorical and 332 literal instances per language, totalling over 1900 instances.

3.4 Automatic Metrics for Translation Quality

We introduce several automatic metrics to evaluate the quality of translations, which are described below.

BLEU/ROUGE. We use BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) as part of the automatic evaluation metrics, using the selected human-edited translations as our gold standard references.

BERTScore. We use BERTScore (Zhang et al., 2019) as a cross-lingual translation evaluation metric to automatically evaluate translations without a target-language reference, due to BERTScore being shown to be effective in cross-lingual settings (Song et al., 2021).

EN-ZH		Manual Evaluation Metrics					Automatic Evaluation Metrics				
		Fluency	Intelligibility	Fidelity	Authenticity	Overall	BLEU1	BLEU4	Rouge-L	BERTScore	GPT-4o
Google	Metaphorical	4.47	4.31	4.25	4.12	4.34	0.58	0.20	0.62	0.765	4.44
Google	Metaphorical (full)	4.75	4.73	4.72	4.64	4.71	0.52	0.20	0.78	0.766	4.75
Google	Literal	4.53	4.55	4.53	4.49	4.54	0.73	0.38	0.76	0.768	4.67
Opus	Metaphorical	3.87	3.52	3.39	3.22	3.59	0.49	0.10	0.53	0.737	3.56
Opus	Metaphorical (full)	4.40	4.32	4.32	4.25	4.32	0.44	0.13	0.65	0.735	4.14
Opus	Literal	3.93	3.80	3.74	3.75	3.82	0.49	0.14	0.54	0.732	3.77
Youdao	Metaphorical	4.67	4.59	4.53	4.53	4.60	0.64	0.26	0.67	0.759	4.64
Youdao	Metaphorical (full)	4.82	4.81	4.80	4.85	4.82	0.53	0.23	0.82	0.764	4.74
Youdao	Literal	4.66	4.67	4.65	4.62	4.66	0.80	0.57	0.83	0.766	4.74
GPT-4o	Metaphorical	4.05	4.25	4.35	4.05	4.17	0.58	0.26	0.60	0.764	4.69
GPT-4o	Metaphorical (full)	4.59	4.32	4.62	4.59	4.53	0.64	0.30	0.68	0.765	4.87
GPT-4o	Literal	4.54	4.54	4.17	4.22	4.37	0.59	0.29	0.64	0.761	4.90
EN-IT											
Google	Metaphorical	4.57	4.46	4.30	4.32	4.44	0.50	0.22	0.60	0.811	4.51
Google	Metaphorical (full)	4.78	4.77	4.72	4.63	4.73	0.65	0.42	0.74	0.811	4.55
Google	Literal	4.77	4.68	4.58	4.67	4.68	0.68	0.47	0.74	0.807	4.68
Opus	Metaphorical	4.45	4.29	4.14	4.16	4.29	0.48	0.19	0.58	0.808	4.06
Opus	Metaphorical (full)	4.78	4.77	4.74	4.63	4.73	0.64	0.42	0.73	0.809	4.45
Opus	Literal	4.65	4.53	4.45	4.52	4.54	0.65	0.43	0.71	0.803	4.29
Youdao	Metaphorical	4.36	4.16	3.96	4.04	4.16	0.45	0.17	0.54	0.805	3.95
Youdao	Metaphorical (full)	4.73	4.73	4.67	4.56	4.67	0.61	0.38	0.69	0.801	4.34
Youdao	Literal	4.53	4.42	4.29	4.38	4.41	0.58	0.30	0.64	0.799	4.13
GPT-4o	Metaphorical	4.41	4.34	4.14	4.25	4.28	0.52	0.24	0.60	0.812	4.53
GPT-4o	Metaphorical (full)	4.50	4.60	4.64	4.55	4.57	0.59	0.27	0.67	0.810	4.85
GPT-4o	Literal	4.59	4.55	4.50	4.55	4.55	0.55	0.26	0.65	0.811	4.81

Table 3: Metaphorical and literal expression evaluation averages. **Manual Evaluation Metrics** and **GPT** employ a 5-point scale to assess the quality and characteristics of expressions, whilst **Automatic Evaluation Metrics** provide scores ranging 0-1. *Metaphorical (full)* refers to translations annotated as having full-equivalence.

GPT score. We also employ GPT-4o[2] as an annotator to score the translation results using the same scoring criteria as human annotators.

4 Results

We first conduct a comprehensive comparative analysis of the performance of MT on metaphorical and literal expressions based on both manual and automatic evaluation scores in §4. Specifically, by analysing the distribution of labels from the fine-grained human evaluation protocol, we verify that metaphor translation is more challenging than literal translation in § 4.1. Moreover, we examine the correlations between the suggested fine-grained human evaluation protocol in §4.2, the correlations between Emotional and Metaphorical Expressions in §4.4, and the crucial role of Metaphor Equivalence in metaphor translation Quality Estimation (QE) in §4.5. Additionally, we analyse the translation quality between typologically different languages in §4.6. We also provide a case study indicating that translating between more typologically distant language pairs is harder, by comparing EN-ZH and EN-IT pairs in Appendix C.

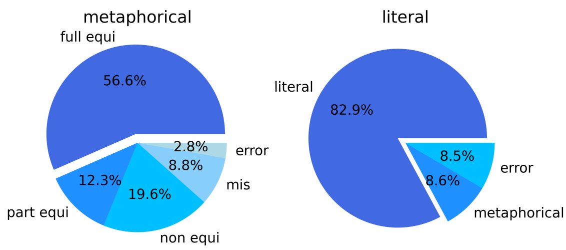


Figure 3: Equivalence distributions of metaphorical and literal expression translations from annotators. *non equi*, *part equi*, and *full equi* refer to non-, part-, and full-equivalence, respectively. *mis* denotes mistranslation.

4.1 Metaphorical vs. Literal Translation

As shown in Figure 3, the analysis of equivalence labels in metaphorical and literal expression translations highlights the varying degrees of equivalence and accuracy in translating metaphorical and literal expressions. Approximately 20% of metaphorical expressions are found to be translated without proper correspondence to the intended metaphorical meaning *(non-equi)*. Furthermore, more than 10% of metaphorical translations exhibit a failure to comprehend the intended metaphor or contain mistakes or inaccuracies (*mis* and *error*). These results emphasise the challenges associated with translating metaphorical expressions.

Table 3 presents scores from manual and au-

[2]gpt-4o-2024-05-13 https://platform.openai.com/

	GPT-3.5[3]	GPT-4o[4]	Gemini Pro[5]
EN-IT full	86.0	86.7	85.7
EN-IT others	92.5	94.0	91.7
EN-ZH full	76.2	76.5	74.4
EN-ZH others	84.1	86.3	84.7

Table 4: Accuracy of LLMs in classifying metaphor equivalence when compared to human annotations. *full* refers to translations annotated as having full equivalence, whilst *others* refers to translations as having non- or part- equivalence.

tomatic evaluations to compare the translation of metaphorical and literal expressions. It can be seen that translating metaphorical expressions poses greater difficulty compared to translating literal expressions. In both EN-ZH and EN-IT translation, the metaphorical expression translations generally obtained lower scores in all Manual Evaluation Metrics and GPT-4o compared to translations of literal expressions from the same MT system. Our automatic evaluation metrics also support this observation, with lower scores from BLEU1, BLEU4, and Rouge-L for metaphorical expression translations, suggesting a reduced level of similarity and alignment with reference translations.

Most importantly, we separately calculate the evaluation scores for full-equivalence translations. The results show that when metaphors are translated faithfully, their scores are significantly higher. This demonstrates that although translating metaphors is a challenging task, achieving the correct form of translation often results in more satisfactory outcomes, therefore highlighting the importance of having comprehensive translation evaluation metrics.

BERTScore struggles to distinguish the performance between metaphorical and literal translations. This limitation may be due to the methods relying on contextual embeddings and cosine similarity struggles to capture the subtle semantic differences inherent in metaphorical language. This highlights the need for specialised evaluation tailored to the complexities of metaphor.

4.2 LLMs Equivalence Assessment

As shown in Table 4, we employ LLMs to annotate the equivalence of metaphor translations and compare the results with the human-annotated reference data. The LLM-based evaluation results demonstrate a high level of consistency with human annotators. Moreover, we task LLMs with providing explanations for their annotations, offering insights into their interpretation of metaphorical content across different languages. For instance, consider the sentence pair: EN: "She swallowed the last words of her speech" and ZH: "她咽下了最后几句话." Here, "咽下" is a translation with full-equivalence. The explanation from GPT-4 is as follows: "*Both in the source sentence and the translation, 'swallowed' and '咽下' are used metaphorically to mean that she did not say the last words of her speech. The literal meanings of 'swallow' and '咽下' are also the same, referring to the action of making food or drink go from your mouth down through your throat and into your stomach.*" Detailed examples of these explanations can be found in Appendix D. This comparison reveals that LLMs can effectively complement human efforts, providing reliable and insightful evaluations that are crucial for high-quality translation assessments at scale.

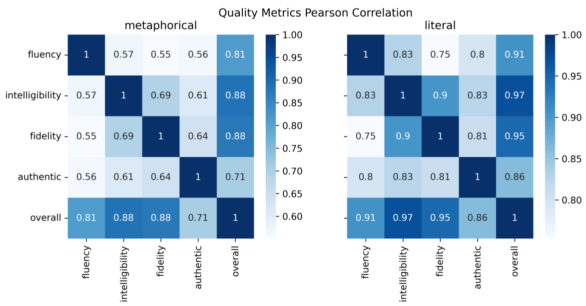


Figure 4: Pearson correlation heatmap of manual evaluation quality.

4.3 Correlation Analysis of Fine-grained Human Evaluation Metrics

Figure 4 shows the fine-grained correlations between human evaluation metrics by calculating pairwise Pearson correlations between the criteria of Fluency, Intelligibility, Fidelity, Authenticity, and Overall Score. Firstly, we observe that the Pearson correlations between each pair are all in the interval between 0.55 and 0.65, indicating a relatively low but positive correlation among them. This observation verifies that Fluency, Intelligibility, Fidelity, and Authenticity represent independent aspects of metaphor translation quality esti-

[3]gpt-3.5-turbo-0125 https://platform.openai.com/
[4]gpt-4o-2024-05-13 https://platform.openai.com/
[5]gemini-1.0-pro-001 https://cloud.google.com/

mation. Secondly, we observe that Fluency, Intelligibility, and Fidelity are all highly positively correlated to the Overall Score, indicating that all three elements of metaphor quality evaluation are paramount in the estimation of overall quality.

4.4 Correlation Analysis of Emotion and Equivalence Metaphor

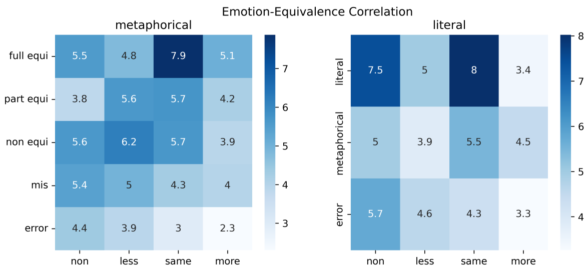


Figure 5: Emotion-Equivalence correlation heatmap based on co-occurrences.

We investigate the correlation between how much emotion the translated version retains and how figurative expressions are translated in Figure 5, which presents a correlation heatmap based on a logarithmic function of the number of co-occurrences of Emotion and Equivalence defined in §3.2. It is noticeable that emotion levels perceived by annotators tend to remain constant if the original metaphorical expression is translated to a fully equivalent version, and the original literal expression is translated to a literal version. This observation indicates that maintaining the figurative status translations is a reasonable strategy for keeping the emotional expression authentic. For example, the metaphorical expressions "*swallow the sentence*" and "咽下这句话" both convey reluctance, whilst the Chinese literal translation "没说这句话" does not. We also observe that non-equivalent translations tend to keep little of the emotion contained in the original metaphorical expressions. In contrast, fully equivalent and part-equivalent translations show weaker, yet similar trends. This finding reveals the difficulty in maintaining emotion through the translation procedure and demonstrates that equivalence and whether the translation is figurative are essential for maintaining levels of emotion.

4.5 Impact of Metaphor Equivalence

Besides maintaining emotional salience, fully equivalent metaphor translations and literal translations of literal expressions demonstrate higher translation quality. This is revealed in Figure 6, which shows that fully equivalent translations of metaphorical expressions outperform others in the dimensions of Fluency, Intelligibility, Fidelity, and Authenticity, whilst literal translations of literal expressions outperform other versions in the four dimensions. We also observe that part-equivalent and non-equivalent translations of metaphors cause more severe translation quality degradation than metaphorical translations of literal expressions. We hypothesise that literal translations of metaphorical expressions between languages spoken by different communities result in unnatural literal statements. which also supports the observation that the *translation of metaphorical expressions is harder than that of literal expressions*.

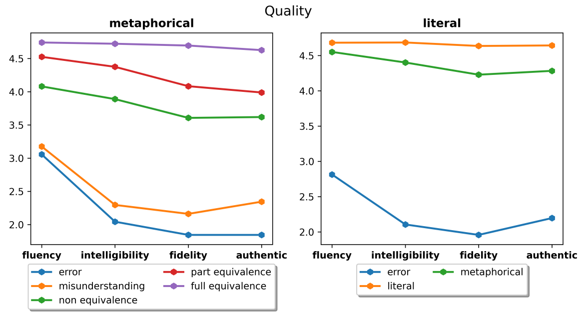


Figure 6: Average quality scores of manual evaluation of metaphorical and literal expression translation.

4.6 Impact of Different Language

Figure 7 presents a comparison of the average evaluation scores of EN-ZH and EN-IT translations across all models. The results show that the average translation quality is lower for Chinese compared to Italian, despite both being translated from English. This can be attributed to several factors. Firstly, Chinese and English belong to different language families and possess distinct linguistic structures, with the grammatical disparities posing challenges for accurate translation. Secondly, cultural differences also play a significant role in translation quality. Translating metaphors accurately requires a deep understanding of cultural nuances and idiomatic expressions between the source and target languages. Failure to grasp these nuances can lead to mistranslation or loss of the intended meaning. Furthermore, the availability and quality of language resources and machine translation models differs for Chinese and Italian.

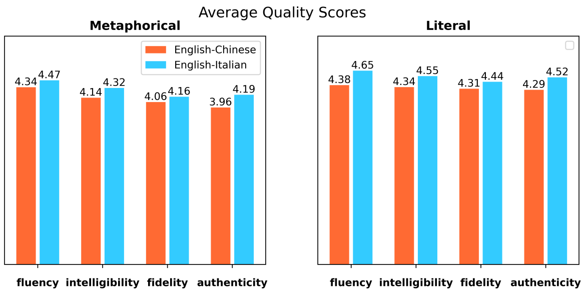


Figure 7: Average quality scores of manual evaluation for EN-ZH and EN-IT.

5 Conclusion

MMTE is the first work to systematically investigate how translations are affected by metaphor in a fine-grained and multi-lingual setting. MMTE also introduces Equivalence as a new dimension of metaphor translation evaluation and verifies its relationship with emotional salience and translation quality. Moreover, we conducted thorough experiments on the proposed evaluation dimensions and verified the increased difficulty of translating metaphorical expressions compared to literal expressions. We further release MMTE, a high-quality metaphor translation corpus which can be adopted for automatic metric design for metaphor translation. Future work intends to combine MMTE with additional well-designed automatic metrics aligning with specific human evaluation dimensions proposed in the paper.

6 Limitations

We summarise several limitations of MMTE which can be explored in future works. Firstly, MMTE only conducts experiments on commercial state-of-the-art translation systems and the most well-known open-source translation packages, rather than models from research works. Secondly, due to resource scarcity for Italian language models and reliable Italian and Chinese metaphor detection models, we only provide thoughts on designing automatic metrics of metaphor translation evaluation based on our corpus, which we will release, rather than presenting plug-and-play automatic metrics. Thirdly, we do not explore language typology in depth in Appendix C as it is an interesting side observation of MMTE. Additionally, it is only our working hypothesis that parallel corpus size is more critical for metaphor translation quality than linguistic typology, rather than a verified conclusion.

7 Ethics Statement

In conducting this research, we adhered to the highest ethical standards to ensure the integrity and responsibility of our work. The data used in our study were sourced from publicly available datasets, and no private or sensitive information was included. All human annotators involved in the study were fully informed about the research objectives and provided their consent prior to participation.

We ensured that the annotations and evaluations were conducted with fairness and respect for linguistic and cultural diversity. Additionally, the use of large language models (LLMs) was guided by ethical considerations, ensuring that the models were applied responsibly and their outputs were critically evaluated.

Our research aims to contribute positively to the field of computational linguistics by improving the quality and reliability of machine translation, particularly in the nuanced area of figurative language. We are committed to transparency, and our methodologies and findings are shared openly for peer review and further research.

Acknowledgments

Tyler Loakman is supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1].

References

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. *arXiv preprint arXiv:2203.14139*.

Jeffrey Allen. 2003. Post-editing. *Benjamins Translation Library*, 35:297–318.

Israa Alsiyat and Scott Piao. 2020. Metaphorical expressions in automatic Arabic sentiment analysis. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4911–4916, Marseille, France. European Language Resources Association.

Fayziyeva Aziza Anvarovna. 2022. Conceptual metaphor universals in english and uzbek. *Open Access Repository*, 8(04):54–57.

Yujin Baek, Zae Myung Kim, Jihyung Moon, Hyunjoong Kim, and Eunjeong Park. 2020. PATQUEST: Papago translation quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 991–998, Online. Association for Computational Linguistics.

Yuri Bizzoni and Shalom Lappin. 2018. Predicting human metaphor paraphrase judgments with deep neural networks. In *Proceedings of the Workshop on Figurative Language Processing*, pages 45–55, New Orleans, Louisiana. Association for Computational Linguistics.

Lera Boroditsky. 2011. How language shapes thought. *Scientific American*, 304(2):62–65.

John B Carroll. 1966. An experiment in evaluating the quality of translations. *Mech. Transl. Comput. Linguistics*, 9(3-4):55–66.

Tuhin Chakrabarty, Smaranda Muresan, and Nanyun Peng. 2020. Generating similes effortlessly like a pro: A style transfer approach for simile generation. *arXiv preprint arXiv:2009.08942*.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021. Mermaid: Metaphor generation with symbolism and discriminative decoding. *arXiv preprint arXiv:2103.06779*.

Siaw-Fong Chung. 2005. Market metaphors: Chinese, english and malay. In *Proceedings of the 19th Pacific Asia conference on language, information and computation*, pages 71–81.

Siaw-Fong Chung, Meng-Hsien Shih, Yu-Hsiang Shen, and Wei-Ting Tseng. 2020. Metaphoricity rating of Chinese KIND metaphor expressions. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 61–69, Hanoi, Vietnam. Association for Computational Linguistics.

Kenneth W Church and Eduard H Hovy. 1993. Good applications for crummy machine translation. *Machine Translation*, 8(4):239–258.

Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. 2019. Modelling the interplay of metaphor and emotion through multitask learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2218–2229, Hong Kong, China. Association for Computational Linguistics.

Paul V DiStefano, John D Patterson, and Roger E Beaty. 2024. Automatic scoring of metaphor creativity with large language models. *Creativity Research Journal*, pages 1–15.

Jennifer Doyon, Kathryn B Taylor, and John S White. 1999. Task-based evaluation for machine translation. In *Proceedings of Machine Translation Summit VII*, pages 574–578.

Sugyeong Eo, Chanjun Park, Hyeonseok Moon, Jaehyung Seo, Gyeongmin Kim, Jungseob Lee, and Heuiseok Lim. 2022. Quak: A synthetic quality estimation dataset for korean-english neural machine translation. *arXiv preprint arXiv:2209.15285*.

Dan Fass. 1991. met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17(1):49–90.

Maucha Gamonal. 2022a. A descriptive study of metaphors and frames in the multilingual shared annotation task. In *Proceedings of the Workshop on Dimensions of Meaning: Distributional and Curated Semantics (DistCurate 2022)*, pages 1–7, Seattle, Washington. Association for Computational Linguistics.

Maucha Gamonal. 2022b. A descriptive study of metaphors and frames in the multilingual shared annotation task. In *Proceedings of the Workshop on Dimensions of Meaning: Distributional and Curated Semantics (DistCurate 2022)*, pages 1–7.

Mengshi Ge, Rui Mao, and Erik Cambria. 2023. A survey on computational metaphor processing techniques: From identification, interpretation, generation to application. *Artificial Intelligence Review*, 56(Suppl 2):1829–1895.

Jonathan Gordon, Jerry Hobbs, Jonathan May, and Fabrizio Morbini. 2015. High-precision abductive mapping of multilingual metaphors. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 50–55, Denver, Colorado. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Zornitsa Kozareva. 2015. Multilingual affect polarity and valence prediction in metaphors. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, page 1, Lisboa, Portugal. Association for Computational Linguistics.

George Lakoff. 1993. The contemporary theory of metaphor.

George Lakoff and Mark Johnson. 1980. Metaphors we live by: University of chicago press. *Chicago, IL.*

Jenny Lederer. 2016. Finding metaphorical triggers through source (not target) domain lexicalization patterns. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 1–9, San Diego, California. Association for Computational Linguistics.

Yucheng Li, Frank Guerin, and Chenghua Lin. 2022a. The secret of metaphor on expressing stronger emotion. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 39–43.

Yucheng Li, Frank Guerin, and Chenghua Lin. 2024. Finding challenging metaphors that confuse pretrained language models. *arXiv preprint arXiv:2401.16012*.

Yucheng Li, Chenghua Lin, and Frank Guerin. 2022b. Cm-gen: A neural framework for chinese metaphor generation with explicit context modelling. In *Proceedings of the 29th international conference on computational linguistics*, pages 6468–6479.

Yucheng Li, Shun Wang, Chenghua Lin, Frank Guerin, and Loic Barrault. 2023. FrameBERT: Conceptual metaphor detection with frame embedding learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1558–1563.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Zhiqiang Liu, Zuohui Fu, Jie Cao, Gerard de Melo, Yik-Cheung Tam, Cheng Niu, and Jie Zhou. 2019. Rhetorically controlled encoder-decoder for modern chinese poetry generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1992–2001.

Tyler Loakman, Aaron Maladry, and Chenghua Lin. 2023. The iron(ic) melting pot: Reviewing human evaluation in humour, irony and sarcasm generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6676–6689.

Qingyu Lu, Baopu Qiu, Liang Ding, Liping Xie, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt.

Yu Lu, Jiali Zeng, Jiajun Zhang, Shuangzhi Wu, and Mu Li. 2022. Learning confidence for transformer-based neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2353–2364, Dublin, Ireland. Association for Computational Linguistics.

Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word embedding and WordNet based metaphor identification and interpretation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1222–1231, Melbourne, Australia. Association for Computational Linguistics.

André F. T. Martins, Marcin Junczys-Dowmunt, Fabio N. Kepler, Ramón Astudillo, Chris Hokamp, and Roman Grundkiewicz. 2017. Pushing the limits of translation quality estimation. *Transactions of the Association for Computational Linguistics*, 5:205–218.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Anirudh Mittal, Diptesh Kanojia, and Pushpak Bhattacharyya. 2022. Survey on computational humour.

Akira Miyazawa and Yusuke Miyao. 2017. Evaluation metrics for automatically generated metaphorical expressions. In *IWCS 2017 — 12th International Conference on Computational Semantics — Short papers*.

Akira Miyazawa and Yusuke Miyao. 2019. Automatically computable metrics to generate metaphorical verb expressions. *Journal of Natural Language Processing*, 26(2):277–300.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33.

Michael Mohler, Bryan Rink, David Bracewell, and Marc Tomlinson. 2014. A novel distributional approach to multilingual conceptual metaphor recognition. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1752–1763, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Arturo Oncevay, Barry Haddow, and Alexandra Birch. 2020. Bridging linguistic typology and multilingual machine translation with multi-view language representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2391–2406, Online. Association for Computational Linguistics.

Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2022. Should a chatbot be sarcastic? understanding user preferences towards sarcasm generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7686–7700.

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.

Gabriella Victorya Pranoto. 2021. *An Analysis of Metaphor Translation in The Mortal Instruments: City of Bones by Cassandra Clare*. Ph.D. thesis.

Xiao-wan Qin and Ke-ming Peng. 2022. The cognitive linguistic view on the english translations of metaphors in "rainy alley". *Journal of Literature and Art Studies*, 12(10):1031–1039.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. Transquest: Translation quality estimation with cross-lingual transformers. *arXiv preprint arXiv:2011.01536*.

Felix Schneider, Sven Sickert, Phillip Brandes, Sophie Marshall, and Joachim Denzler. 2022. Metaphor detection for low resource languages: From zero-shot to few-shot learning in Middle High German. In *Proceedings of the 18th Workshop on Multiword Expressions @LREC2022*, pages 75–80, Marseille, France. European Language Resources Association.

Dorothy Senez. 1998. Post-editing service for machine translation users at the european commission. In *Proceedings of Translating and the Computer 20*.

Yujie Shao, Xinrong Yao, Xingwei Qu, Chenghua Lin, Shi Wang, Wenhao Huang, Ge Zhang, and Jie Fu. 2024. CMDAG: A Chinese metaphor dataset with annotated grounds as CoT for boosting metaphor generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3357–3366.

Ekaterina Shutova, Lin Sun, Elkin Darío Gutiérrez, Patricia Lichtenstein, and Srini Narayanan. 2017. Multilingual metaphor processing: Experiments with semi-supervised and unsupervised learning. *Computational Linguistics*, 43(1):71–123.

Harold Somers. 2003. Computers and translation. *Computers and Translation*, pages 1–365.

Yurun Song, Junchen Zhao, and Lucia Specia. 2021. Sentsim: Crosslingual semantic evaluation of machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3143–3156.

Gerard Steen. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing.

Kevin Stowe, Nils Beck, and Iryna Gurevych. 2021a. Exploring metaphoric paraphrase generation. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 323–336, Online. Association for Computational Linguistics.

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021b. Metaphor generation with conceptual mappings. *arXiv preprint arXiv:2106.01228*.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. Impli: Investigating nli models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388.

Xiangru Tang, Xianjun Shen, Yujie Wang, and Yujuan Yang. 2020. Categorizing offensive language in social networks: A chinese corpus, systems and an explanation tool. In *China National Conference on Chinese Computational Linguistics*, pages 300–315. Springer.

Xiaoyu Tong, Ekaterina Shutova, and Martha Lewis. 2021. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4673–4686.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258.

Jannis Vamvas and Rico Sennrich. 2022. As little as possible, as much as necessary: Detecting over- and undertranslations with contrastive conditioning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 490–500, Dublin, Ireland. Association for Computational Linguistics.

Mihaela Vela, Anne-Kathrin Schumann, and Andrea Wurm. 2014. Beyond linguistic equivalence. an empirical study of translation evaluation in a translation learner corpus. In *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation*, pages 47–56.

Shun Wang, Yucheng Li, Chenghua Lin, Loic Barrault, and Frank Guerin. 2023. Metaphor detection with effective context denoising. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1404–1409.

John S White, Theresa A O'Connell, and Francis E O'Mara. 1994. The arpa mt evaluation methodologies: evolution, lessons, and future approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*.

Yorick Wilks. 1978. Making preferences more active. *Artificial intelligence*, 11(3):197–223.

Omnia Zayed, John Philip McCrae, and Paul Buitelaar. 2020. Figure me out: A gold standard dataset for metaphor interpretation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5810–5819, 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. *arXiv preprint arXiv:1904.09675*.

Haofei Zhao, Yilun Liu, Shimin Tao, Weibin Meng, Yimeng Chen, Xiang Geng, Chang Su, Min Zhang, and Hao Yang. 2024. From handcrafted features to llms: A brief survey for machine translation quality estimation. *arXiv preprint arXiv:2403.14118*.

Danning Zheng, Ruihua Song, Tianran Hu, Hao Fu, and Jin Zhou. 2019. “love is as complex as math”: Metaphor generation system for social chatbot. In *Workshop on Chinese Lexical Semantics*, pages 337–347. Springer.

Yuanhang Zheng, Zhixing Tan, Meng Zhang, Mieradilijiang Maimaiti, Huanbo Luan, Maosong Sun, Qun Liu, and Yang Liu. 2021. Self-supervised quality estimation for machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3322–3334, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lei Zhou, Liang Ding, and Koichi Takeda. 2020. Zero-shot translation quality estimation with explicit cross-lingual patterns. *arXiv preprint arXiv:2010.04989*.

Longhui Zou, Ali Saeedi, and Michael Carl. 2022. Investigating the impact of different pivot languages on translation quality. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Workshop 1: Empirical Translation Process Research)*, pages 15–28.

A Framework Details

A.1 Translators and Languages

The **Google Cloud Translation API** (Translation V3 API) is a prominent commercial multilingual translation tool employing neural MT (NMT) techniques, known for its wide-ranging capabilities and comprehensive language support.

The **Youdao Cloud Translation API** is a popular commercial multilingual NMT tool within the Chinese community, proficient in handling Chinese language translation tasks.

The **Helsinki-NLP/opus-mt** models are pre-trained on the open parallel corpus (OPUS), a continuously expanding collection of translated texts sourced from the web. These models are widely used by researchers and practitioners due to their effectiveness and versatility.

The **GPT-4o**, developed by OpenAI, is an advanced language model designed to perform a wide range of natural language processing tasks, including serving as a highly capable translator model that can handle multiple languages with high accuracy and fluency.

Chinese, a Sino-Tibetan language, is renowned for its rich idiomatic expressions and extensive use of metaphors. **Italian**, a Romance language descended from Latin and belonging to the Indo-European language family like English, provides a distinct comparison. The distinction between these target languages enables a more accurate assessment of the models' performance in preserving metaphorical meaning.

A.2 Annotation Setup

Our annotation platform is built on a private server using an open-source annotation tool - Doccano (Nakayama et al., 2018). We hired 18 annotators who are native speakers of the target languages, all of whom are linguistics majors with professional working competency in English. All annotation workers are paid based on the median wage of similar tasks on Amazon Turk, which is 10 dollars/hour. Specifically, the annotators are divided into six groups, each with three annotators. The groups are further equally divided between annotating English-Chinese (**EN-ZH**) instances and English-Italian (**EN-IT**) instances. Each group is tasked with labelling target words and post-editing the entire MOH dataset and its translations with all 647 pairs of data, resulting in each paired instance being annotated three times. In the final step, all annotation results are cross-checked by professional translators. A group meeting was held to discuss instances of disagreement, and final decisions were recorded on an online discussion website for future reference.

B Guideline

In each annotation sample, an **English** sentence is be given as source text, followed by four translations in **Chinese**. Please evaluate each translation based on the criteria listed below. You will also be asked to supply your own translation as the gold reference.

B.1 Sentence Quality

Please compare the source sentence and its translation without reference to the correct translation, and evaluate the translation from following aspects:

- ***Fluency***: To what extent the translation is well-formed and grammatical, ensuring that it sounds like it was originally written in the target language.

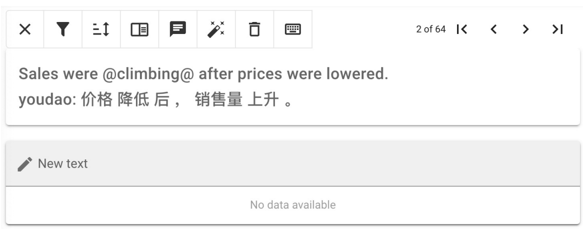
- ***Intelligibility***: To what extent the translation is easily understood and conveys metaphorical meaning sufficiently, such that readers can gain the intended interpretation.
- ***Fidelity***: The extent to which the translation is faithful to the source sentence, such that there is minimal distortion, twisting, or altering of meaning.
- **Overall**: An overall assessment to indicate the quality of the entire sentence seen as a whole.

Please judge these four aspects of quality on a 5-point Likert scale: 5) Very Good; 4) Good; 3) Acceptable; 2) Poor; 1) Very Poor.

B.2 Equivalence

Please compare the source sentence and its translation, and determine to what extent the target word and its translation are Equivalent in figuration. Here are the definitions of the three types of Equivalence and two types of mistakes:

- **Full-Equivalence**: When comparing the source sentence and translation sentence, both the literal meanings and the contextual meanings of the target word are the same.
  EN: He @injected@ new life into the performance.
  ZH: 他给表演注入了新的生命
- **Part-Equivalence**: When comparing the source sentence and translation sentence, only the contextual meanings of the target word are similar. The literal meaning of the target word between the source sentence and translation are different, but they are both metaphorical.
  EN: @Wallow@ in your success!
  ZH: @沉浸@在你的成功中吧！
- **Non-Equivalence**: When comparing the source and translation, only the contextual meanings of the target word are similar. However, the translation is a non-metaphorical expression, making the literal meaning of the target word between the source and translation different.
  EN: Sales were @climbing@ after prices were lowered.
  ZH: 价格下跌后销售额@上升@。
- **Misunderstanding** When the literal meanings are similar between the target word in the source text and the target word in the translation, but the translation conveys no contextual meaning like the target in the source language.
  EN: I @attacked@ the problem as soon as I got out of bed.
  ZH: 我一下床就@攻击@了问题
- **Error**: When the target word is mistranslated, meaning that not only the contextual meanings are different between the source and translation, but their literal meanings also differ.
  EN: @Stamp@ fruit extract the juice.
  ZH: @果果@提取果汁。



(a) unlabeled sample

(b) labeled sample

Figure 8: An example of the annotation process.

B.3 Emotion

Please compare the source sentence and its translation, and determine to what extent the target word and its translation convey equal amounts of emotion. There are four labels to judge emotion:

- **Zero**: If the target word in source context conveys no emotion, please fill **Zero**.
  EN: I can not @digest@ milk products.
  ZH: 我不能消化牛奶产品。

- **More**: The target word in the translation conveys ***more*** emotion than the target word in the source sentence.
  EN: The seamstress @ruffled@ the curtain fabric.
  ZH: 裁缝女把窗帘布弄得一团糟.

- **Same**: The target words in the two sentences convey a ***similar*** degree of emotion.
  EN: I @salute@ your courage!
  ZH: 我向你的勇气致敬!

- **Less**: The target word in the translation conveys ***less*** emotion than the target word in the source sentence.
  EN: The spaceship blazed out into space.
  ZH: 太空船飞向太空

B.4 Authenticity Target

Please compare the target in the source sentence and its translation, and evaluate whether the target translation is authentic. In other words, to what extent is the translation idiomatic (i.e. is expressed in a way that a native speaker would express it)? Please judge the target on a 5-point scale: 5) Very Good; 4) Good; 3) Acceptable; 2) Poor; 1) Very Poor.

B.5 Post-Editing

By referring to the source sentence and its translations, in addition to the above Equivalence scale, please give two fluent and high-quality translations: 1) using figurative language (full-equivalence, part-equivalence) and 2) without using figurative language (non-equivalence). You should focus on the given target word, and make sure it is translated into an appropriate expression.

C Influence of Linguistic Typology on Translation Difficulty

Linguistic typological features are known to be able to assist translation and rank candidates for multilingual transfer (Oncevay et al., 2020). The experimental results of Opus in Table 3 support a similar conclusion, that translation between a language pair with a closer typological relationship (EN-IT) is easier than a more distant pair (EN-ZH). However, this conclusion does not hold for the experimental results of Google and Youdao in Table 3. Youdao, a popular commercial multilingual translation tool in the Chinese community, achieves better translation performance in the EN-ZH direction than EN-IT. We hypothesize that the size of the corpus is much more important for translation quality compared to linguistic typology. Due to the above observations, the potentially larger EN-ZH parallel corpus that Youdao and Google have compared to EN-IT, and the relatively balanced sizes of EN-ZH and EN-IT parallel corpora that Opus holds, may aid in explaining the observed difference.

D Metaphor Explanation with LLMs

We used LLMs to annotate metaphor equivalence and attempted to guide the models to provide explanations for their evaluations. The specific formats of the prompts and queries are shown in the Tab. 5. By providing specific examples and explanations for each type of equivalence and including them in the prompts, we aimed to give LLMs references for comparison.

Given the powerful capabilities of large language models (LLMs), we employed LLMs to annotate and explain different metaphor translations. LLMs, particularly GPT-4, demonstrate an understanding that approaches human annotators in terms of both semantic and rhetorical comprehension.

As shown in Tab. 6, LLMs showcased a robust ability to understand and interpret metaphors, providing comprehensive explanations that covered both semantic nuances and rhetorical aspects. This performance indicated a high level of competency in handling cross-linguistic tasks.

By analyzing the explanations provided by LLMs, we were able to validate their effectiveness in metaphor translation tasks. This analysis demonstrated that LLMs could not only understand and interpret metaphors accurately but also articulate the reasoning behind their evaluations. This capability is crucial for ensuring that the subtleties of metaphorical language are preserved in translation. LLMs offered reliable and insightful evaluations that are essential for high-quality translation assessment. Their ability to generate detailed and contextually accurate explanations for their decisions highlights their potential as a robust tool in the translation process.

Prompts	Please compare the source sentence and its translation, and determine to what extent the @target word@ and its translation are Equivalent in figuration. Here are the definitions of the three types of Equivalence and two types of mistake: Full-Equivalence: When comparing the source sentence and translation sentence, both the literal meanings and the contextual meanings of the target word are the same. Example: source: He @injected@ new life into the performance. translation: 他给表演@注入@了新的生命Part-Equivalence: When comparing between the source sentence and translation sentence, only the contextual meanings of the target word are similar. The literal meaning of the target word between the source sentence and translation are different, but they are both metaphorical. Example: source: @Wallow@ in your success! translation: @沉浸@在你的成功中吧！Non-Equivalence: When comparing the source and translation, only the contextual meanings of the target word are similar. However, the translation is a non-metaphorical expression, making the literal meaning of the target word between the source and translation different. Example: source: Sales were @climbing@ after prices were lowered. translation: 价格下跌后销售额@上升@。Misunderstanding: When the literal meanings are similar between the target word in the source text and the target word in the translation, but the translation conveys no contextual meaning like the target in the source language. Example: source: I @attacked@ the problem as soon as I got out of bed. translation: 我一下床就@攻击@了问题Error: When the target word is mistranslated, meaning not only that the contextual meanings are different between the source and translation, but their literal meanings also differ. Example: source: The @stamp@ fruit extract the juice. translation: 果果提取果汁。
Query	Source: **English source sentence** Translation: **Target language translation** Equivalence and Explanation:

Table 5: Prompt and Query Format for LLMs in Evaluating Metaphor Translation Equivalence and Generating Explanations.

	EN: The news were leaked to the paper. **ZH**: 消息泄露 给了报纸。	**EN**: The colors don't harmonize. **IT**: I colori non si armonizzano.
GPT-3.5	Both the source sentence and the translation use a metaphorical expression to convey the same meaning. The target words "leaked" and "泄露" both literally and contextually mean that information was disclosed or released unintentionally or secretly.	The source sentence and the translation both use the word "harmonize" (and its Italian equivalent "armonizzano") in the same context, with both the literal and contextual meanings being the same.
GPT-4	The literal and contextual meanings of "leaked" in the source sentence and "泄露" in the translation are the same. Both terms refer to the unauthorized release or disclosure of information.	The source sentence and the translation have both the same literal and contextual meanings of the target word "harmonize". In both languages, it refers to the act of being in harmony or agreement.
Gemini Pro	**Literal Meaning:** Both "leaked" and "泄露" share the literal meaning of confidential information being unintentionally or secretly revealed. **Contextual Meaning:** Both words fit the context of the sentence, implying an act of disclosing information without authorization. The information flow (news to the paper) is consistent. Therefore, both the literal and contextual meanings are preserved in the translation, making it a case of Full-Equivalence.	The word "harmonize" and its Italian translation "armonizzano" share both literal and contextual meanings. * **Literal Meaning:** Both words refer to the concept of elements coming together in a pleasing and balanced way.* **Contextual Meaning:** In the context of colours, both words imply that the colors used together do not create a visually appealing or cohesive effect. Therefore, the translation accurately conveys the meaning of the source word, making it a case of full equivalence.

Table 6: Examples of explanations given by the LLMs for metaphor translations, selecting one full-equivalence translation each in Chinese and Italian.