From Neural Machine Translation to Large Language Models: Analysing Translation Quality of Chinese Idioms

Yafei Zhu[1,2][0000−0002−8283−8725], Daisy Monika Lal[2][0000−0001−6407−6184], Sofiia Denysiuk[2][0009−0009−4476−5922], and Ruslan Mitkov[2][0000−0003−2522−066X]

[1] Shanghai International Studies University, China
0214101651@shisu.edu.cn
[2] Lancaster University, UK
{d.m.lal, s.denysiuk, r.mitkov}@lancaster.ac.uk

Abstract. Idioms present a formidable challenge for machine translation (MT) due to their figurative, culture-specific, and linguistic complexity. In this study, we compiled a corpus of 100 Chinese idioms from the Dictionary of Chinese Idioms and conducted quantitative analyses of nine state-of-the-art MT systems. Recognising the linguistic complexity of idioms, we introduced AIE, a new evaluation metric for translations, derived from its three assessment criteria: Accuracy, Intelligibility, and Elegance. In this framework, we suggest assigning distinct weights to its metrics, supported by empirical evidence. Additionally, we employed automatic metrics ROUGE, BLEU, BLEURT, and METEOR, to assess translation quality. Our analysis revealed that while BLEURT and BLEU exhibited stronger correlations with human scores, the overall correlation remained weak. Furthermore, recognising the significance of automatic evaluation in natural language processing (NLP), we hypothesised that combining existing automatic metrics could yield improved assessment scores compared to individual metrics. To validate this hypothesis, we computed average scores of automatic metrics, which demonstrated a positive correlation with human scores, suggesting a promising alternative. Our findings indicate that GPT4 and GLM4 outperform other state-of-the-art models even in translating less commonly used idioms.

Keywords: Chinese idioms · Machine translation · Large language model.

1 Introduction

Idioms, as a significant component of language, are a class of multiword expressions (MWEs) that can be found abundantly across different domains of natural discourse [20, 21, 26]. Due to semantic opacity and cultural specificity, they have become one of the most important problems in translation [22, 32]. For different languages embedded with similar cultures, literal translation may not pose a huge challenge to the correct understanding. However, it will lead to ambiguity or even utter incomprehension for languages of different families, like Chinese

Proceedings of New Trends in Translation and Technology, pages 247–260
3–6 July, 2024, Varna, Bulgaria.
https://doi.org/10.26615/issn.2815-4711.2024_021

and English. For example, the literal meaning of "Cici Buxiu"(talk incessantly) is "stab without a break". Therefore, while MWEs are crucial, the translation of idioms from Chinese to English could be a significant reference for the quality evaluation or estimation of an MT system [25].

Prior research has predominantly concentrated on strategies to improve idiom translation, including the implementation of substitution methods [26], the development of large-scale dedicated datasets [7], and the introduction of retrieval augmentation and loss weighting techniques [18], among others. While certain studies have introduced novel metrics for automatic evaluation [4, 28], they neglected the importance of human evaluation as the primary metric in MT throughout the Conferences on Machine Translation (WMT) [15]. What is more, due to the rapid development of neural machine translation (NMT) and large language models (LLMs), it is necessary to compare their performance in the task of Chinese-English idiom translation.

In this study, we present a comprehensive analysis of the challenges associated with automatic idiom translation and its evaluation. The research is articulated around several key contributions. Firstly, we introduce a Chinese-English corpus comprising 100 Chinese idioms sourced from the Dictionary of Chinese Idioms. This compilation serves as a valuable resource for studying the translation and interpretation of idiomatic expressions between Chinese and English. Secondly, we compare the quality of six popular NMT systems, three Chinese and three Western, with three LLMs for the task of Chinese-English idiom translation. Thirdly, we propose the AIE metric, which is tailored for assessing the quality of idiom translations from three essential criteria: accuracy, intelligibility, and elegance. We suggest applying different weights to them when calculating the overall translation quality, based on an optimisation model that validates the effectiveness of this weighted approach. Fourthly, recognising the importance of automatic evaluation in natural language processing (NLP), we propose a composite approach to enhance the assessment of translation quality. This approach combines multiple established metrics, including ROUGE, BLEU, BLEURT, and METEOR, to form a more robust evaluation framework. Finally, to confirm the validity of our proposed composite evaluation approach, we conduct a comparative analysis between human judgments and automated evaluation scores. This comparison aims to determine which metric most closely aligns with human perceptions of translation quality.

The rest of the paper is structured as follows. Section 2 surveys related work. Section 3 presents the data and methodology used in this project. Section 4 reports the evaluation results and offers a discussion of the results. Finally, Section 5 lists the conclusions of this study.

2 Related Work

The study of the machine translation of idioms has been a subject of study in the last few decades. One of the main concerns in this field is to improve the translation quality of idioms in many language pairs by implementing various

methods, such as the two-step substitution method used by Salton et al. [26] for English-Brazilian-Portuguese in statistical machine translation (SMT), the creation of a large-scale idiom translation data set by Fadaee et al. [7] for German and English, and the techniques introduced by Liu et al. [18], including upweighting training loss on idiomatic sentences and implementing retrieval-augmented models for translating from French, Finnish, and Japanese to English. To facilitate the translation of Chinese idioms, Bai et al. [2] proposed a high-precision algorithm for extracting translations of a given MWE from parallel corpora based on scores of normalised correlation frequency and demonstrated that the performance of the Moses MT system can be significantly improved for Chinese-English translation. In addition, Qiang et al. [24] proposed CIP (Chinese Idiom Paraphrasing), a novel infill-based approach based on text infilling to rephrase idiomatic sentences to non-idiomatic ones while preserving the original meaning.

Evaluating the performance of machine translation (MT) systems is a significant focus in MT research. Popular global evaluation metrics include BLEU [23], METEOR [3], TER [29], ROUGE [17], and BLEURT [27], among others. Given the unique figurative meanings of idioms, the metrics typically used for non-idiomatic translation often fall short in accurately assessing their translations. Shao et al. [28] introduced a blacklist method to identify the literal translation errors in Chinese idiom translation, which is proved effective but requires manually created word list (the blacklist) including words that should not exist in the translation of the idiom. Inspired by the blacklist method, Baziotis [4] proposed LitTER, a novel metric designed to quantify the literal translation error rate made by a model. This approach automates the creation of the blacklist, enhancing the evaluation process.

Due to drawbacks of the automatic evaluation and the ultimate goal of being valuable to people in natural language generation (NLG), human evaluation is considered the gold standard in a broad range of NLP tasks, such as question answering and machine translation [6]. For example, in recent WMT, the results of human evaluation have been used as the primary metric to correlate with the results of the automatic metrics [5, 8, 14, 15]. For human evaluation, adequacy and fluency, or accuracy and intelligibility [1], have become standard dimensions used widely [6]. While the former measures how much meaning from the source is reproduced, the latter only focuses on the quality of the translated text without taking the source into account. However, those metrics are usually used for the translation of general domains, not fully suitable for special linguistic phenomena like idioms. Due to the absence of human evaluation for idiom translation, we propose the AIE metric as a gold standard to assess the quality of various NMT systems and LLMs in this area. In addition, different global automatic evaluation metrics are also involved for the correlation with the results of human evaluation.

3 Methodology

This section explains the source and structure of the dataset as well as the design of AIE metric for the quality evaluation of idiom translation.

3.1 Chinese-English Idiom Dataset

Our dataset used for evaluation consists of four parts: 100 Chinese idioms, 100 Chinese sentences made by each idiom, their standard human translations (HT), and nine machine translations (MT) for each sentence. What we use to extract Chinese idioms is Zhonghua Chengyu Dacidian (Dictionary of Chinese Idioms), published by Commercial Press International Limited in 2019. As one of the largest Chinese idiom dictionaries, it comprises over 49,000 idioms, with more than 90% being four-character expressions and most of the rest including three or five characters. To preserve the diversity of idioms, our dataset maintains a similar ratio. It is important to highlight that certain Chinese expressions with transparent meaning do not pose a challenge to MT and are therefore excluded from our dataset. The Chinese sentence of each idiom and its corresponding human translation are from the Chinese-English Parallel Corpus of Classic Chinese Literary Works published by Shanghai Foreign Language Education Press[3].

3.2 AIE Metric

There are various ways to run human evaluations. Our AIE metric is designed by referring to several important principles proposed by Khashabi [13]: application-motivated, reproducible, interpretable, scalar, and quantified uncertainty. First of all, the metric is highly applicable for the idiom translation evaluation of any language pairs. Given the two-layer meaning and the fixed form of idiom, the quality of its transformation is assessed from both the function and form, i.e., the meaning and grammar. Accuracy, which assesses the extent to which the original meaning of the idiom is preserved in the translation, and intelligibility, which evaluates how well the figurative meaning of the idiom is comprehended in the target language, are both critical metrics for assessing the semantic quality of translations [1]. Meanwhile, elegance focuses on the grammatical and stylistic propriety of the translation, examining how well the translated text conforms to the grammatical rules and stylistic norms of the target language. Together, these metrics provide a comprehensive framework for evaluating the quality of idiom translations from both semantic and syntactic perspectives.

As for how to perform the human evaluation based on the metric, we adopt the DA (Direct Assessment)+SQM (Scalar Quality Metrics) used by WMT2022 and WMT2023 [14, 15]. DA facilitates the comparison of a new model to those having been evaluated for studies in the future while SQM stabilises scores across different annotators. Therefore, our evaluation produces an absolute scalar measurement of different models on idiom translation. According to the meaning of each sub-metric, each annotator measures the quality along a five-point Likert scale: very poor, poor, okay, good, and very good (see Table 1).

[3] https://we.sflep.com/cepc/Search/SimpleSearch.aspx

Table 1. Human Assessment Metrics

Metric	Label	Description
Accuracy	1 - inaccurate	fail to preserve either literal or figurative meaning
	2 - poor	manage to preserve part of the literal meaning but not all of it
	3-moderate	preserve the literal meaning well but fail for figurative meaning
	4-good	manage to preserve part of the figurative meaning but not all
	5-high	perfectly preserve the figurative meaning
Intelligibility	1-unintelligible	misleading or hopeless to understand the correct figurative meaning
	2-little	possible to understand a little through guessing but not sure if it is correct
	3-much	possible to understand much through considerable deduction
	4-most	understandable in the context without much effort not native or natural
	5-all	perfectly clear and natural like native text without any effort
Elegance	1-incomprehensible	major mistakes that breaks grammatical rules
	2-disfluent	inconsistent application of grammatical rules with minor mistakes
	3-correct	generally correct but not natural or native in style
	4-good	generally fluent and natural without using a correct idiom
	5-elegant	perfectly elegant by using a correct idiom

4 Results and Discussion

This section outlines the benchmarks used for Chinese idiom translation and presents the results from both human and automatic evaluation. Our analysis utilised nine distinct models, including six NMT systems and three LLMs. The three internationally recognized NMT systems are Google Translator (MT1) [31], DeepL Translator (MT2)[4], and Microsoft Translator (MT3) [12]. For the LLMs, we employed GPT-4 (MT4) [11] by OpenAI, Llama2 (MT5) [30] by Meta, and GLM4 (MT6) [10] by Tsinghua University in China. The three Chinese-developed NMT systems are NiuTrans (MT7)[5], which ranked first in WMT2020 for the Chinese-English language pair [19], VolcTrans (MT8)[6], which held a similar position in WMT2021 [8], and Baidu Translate (MT9) [9], known for its widespread use in China.

[4] https://www.deepl.com/translator
[5] https://niutrans.com/
[6] https://translate.volcengine.com/

4.1 Human Evaluation

We conducted human evaluations using two annotators who assessed 180 translations (20 for each model) and three annotators who evaluated half of them (10 for each model). Additionally, all 900 translations were reviewed by one annotator. All three annotators are native Chinese speakers fluent in English, with professional background in linguistics and translation between Chinese and English. Based on the scores (see Figure 1), we can classify the models into three categories: top, moderate, and low-performing models.

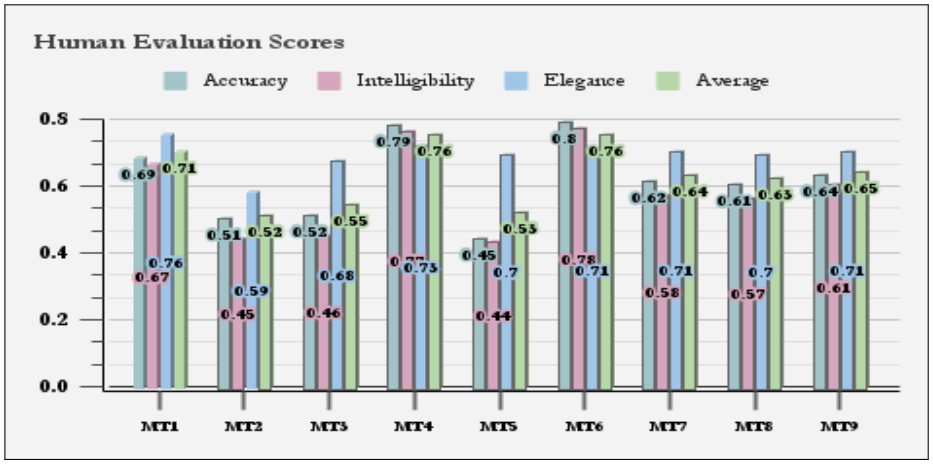

Fig. 1. Human Evaluation Scores for all MT systems.

- **Top Performance:** MT6 or GLM4 (accuracy = 0.80, intelligibility = 0.78, elegance = 0.71), and MT4 or GPT4 (accuracy = 0.79, intelligibility = 0.77, elegance = 0.73) rank highest overall, with relatively high average human evaluation scores for accuracy, intelligibility, and elegance. Both achieve an overall average score of 0.76, notably higher than all other models. MT1 or Google Translator (accuracy = 0.69, intelligibility = 0.67, elegance = 0.76), despite its moderate score for still performs well overall due to its high scores for elegance and ranks third.
- **Moderate Performance:** MT9 or Baidu (accuracy = 0.64, intelligibility = 0.61, elegance = 0.71), MT7 or NiuTrans (accuracy = 0.62, intelligibility = 0.58, elegance = 0.71), and MT8 or Volcano Translator (accuracy = 0.61, intelligibility = 0.57, elegance = 0.70), are moderate performers due to their relatively low score for accuracy and intelligibility. However, these Chinese translation systems demonstrate strength in generating elegant translations, as indicated by their high elegance score.
- **Low Performance:** MT3 or Microsoft Translator (accuracy = 0.52, intelligibility = 0.46, elegance = 0.68), MT5 or Llama2 (accuracy = 0.45, intel-

ligibility = 0.44, elegance = 0.70), and MT2 or DeepL (accuracy = 0.51, intelligibility = 0.45, elegance = 0.59), rank lower due to their low score for accuracy and intelligibility. However, the LLM Llama2 obtains a comparatively high score for elegance suggesting that the model manages to maintain the figurative meaning of idioms and produce translations with a certain level of elegance.

Optimisation Model In our evaluation framework employing the AIE methodology, we advocate for the allocation of different weights to its three metrics to determine the final score. Utilising an optimisation model, we present compelling empirical evidence supporting the necessity for distinct weightings for each metric. To validate this hypothesis, we utilised a statistical tool to identify the optimal weight assignments for each metric within the AIE framework. Our approach unfolds in several steps. Initially, we identified the top-performing models, namely Google Translator, GLM4, and GPT4, through human evaluation. The translations generated by these models were subsequently rated by a single annotator, who provided a scalar score ranging between 0 and 1. Following this, we constructed a dataset to establish the correspondence between AIE scores and the scalar scores provided by the annotator. Through this analysis, we determined the optimal weights for each AIE metric: 0.3832 for accuracy, 0.2950 for intelligibility, and 0.3217 for elegance. These results suggest that accuracy holds the greatest weight in scoring idiom translations, followed by elegance and then intelligibility. However, it's important to acknowledge that, due to the limited size of our dataset, these weights may not be definitive. Nevertheless, our findings underscore the varying importance of each metric and emphasise the necessity of considering weighted scores for idiom translations.

4.2 Krippendorff's Alpha

To assess the reliability and consistency of the human assessment, we employed Krippendorff's Alpha [16] (α_k), a metric that measures the level of agreement or disagreement between the multiple annotators. We performed (α_k) computation at two levels (see Table 2): firstly, between pairs of annotators (α_k^2) who assessed a subset of 180 translations (20 for each MT system), secondly, between three annotators (α_k^3) who assessed a subset of 90 translations (10 for each MT system). The inter-rater agreement was computed for all three assessment metrics: accuracy, intelligibility, and elegance, at both levels. Based on Krippendorff's Alpha α_k^2 scores (see Table 2), it becomes evident that there is a consistent trend of high agreement regarding intelligibility across all systems. Furthermore, for accuracy, MT1, MT2, and MT3 stand out as they exhibit high levels of agreement among annotators. The high agreement on accuracy implies that annotators have a shared perception of the quality of translations produced by these systems, particularly in terms of idiomatic expressions. When analysing the agreement levels among annotators specifically for elegance across different MT systems, we observe a notable pattern. For MT1, MT4, MT6, MT7, and MT8, there emerges a moderate level of agreement among annotators.

Comparing α_k^2 and α_k^3: When we compare both scenarios, in the case of three annotators evaluating 90 translations, there are few instances with a lower Alpha value as compared to two annotators. An increase when moving from two annotators evaluating 180 translations α_k^2 to three annotators evaluating 90 translations α_k^3, is observed. This indicates that the judgments provided by the additional annotator are more consistent or aligned with the judgments of the other annotators, resulting in increased overall agreement. This is because Krippendorff's Alpha takes into account both the number of evaluators and the number of samples being evaluated. With three annotators, there is a better chance of identifying and mitigating individual biases or inconsistencies in the evaluation process, thereby increasing the reliability of the assessments.

Table 2. Krippendorff's Alpha scores for inter-annotator agreement. α_k^2 denotes agreement between two annotators who assessed a subset of 180 translations (20 for each MT). α_k^3 denotes agreement between three annotators who assessed a subset of 90 translations (10 for each MT).

	$\alpha_\mathbf{k}^\mathbf{2}$			$\alpha_\mathbf{k}^\mathbf{3}$		
	Accuracy	**Intelligibility**	**Elegance**	**Accuracy**	**Intelligibility**	**Elegance**
MT1	**0.815**	**0.895**	0.066	**0.850**	**0.852**	0.402
MT2	0.305	**0.869**	0.101	**0.793**	**0.755**	0.281
MT3	0.446	0.493	-0.029	**0.627**	**0.736**	0.161
MT4	0.292	0.163	0.505	0.561	**0.860**	**0.605**
MT5	0.339	**0.890**	**0.664**	0.575	**0.815**	0.271
MT6	0.020	0.462	0.509	0.447	**0.849**	0.530
MT7	0.267	**0.817**	0.314	**0.666**	**0.906**	0.450
MT8	0.538	**0.628**	0.717	0.598	**0.950**	**0.653**
MT9	0.026	0.352	-0.060	**0.793**	**0.654**	0.128

4.3 Automatic Evaluation

We employed a suite of benchmark automatic evaluation metrics—ROUGE, BLEU, BLEURT, and METEOR—to assess the MT systems, specifically focusing on their ability to accurately translate idioms. We calculated the average scores for 100 idiom translations produced by each system (see Table 3). These metrics gauged the alignment between machine-generated and reference translations. By averaging all scores, we explored whether combining these metrics could effectively assess idiom translation quality in scenarios without human evaluators. This multi-metric approach aims to provide a comprehensive assessment of translation quality, covering fluency, accuracy, grammaticality, and adequacy.

Based on our findings, we categorized the models into three performance tiers: top, moderate, and low.

Table 3. Automatic evaluation scores. The 'Average' denotes the composite score obtained by aggregating the scores for all metrics.

	R1-F1	**R2-F1**	**RL-F1**	**BLEU**	**BLEURT**	**METEOR**	**Average**
MT1	**0.314**	0.105	**0.285**	0.192	-0.475	0.330	**0.1252**
MT2	0.263	0.081	0.236	0.198	-0.557	0.310	**0.0885**
MT3	0.289	0.092	0.260	0.199	-0.577	0.290	**0.0922**
MT4	0.300	0.100	0.268	0.185	-0.444	0.332	**0.1235**
MT5	0.257	0.075	0.234	0.187	-0.613	0.276	**0.0693**
MT6	0.309	0.107	**0.285**	0.181	-0.431	**0.340**	**0.1318**
MT7	0.304	**0.111**	0.277	0.200	-0.505	0.326	**0.1188**
MT8	0.295	0.097	0.266	**0.205**	-0.539	0.307	**0.1052**
MT9	0.302	0.102	0.269	0.195	-0.506	0.326	**0.1313**

- **Top Performance:** MT6 or GLM4 ranks first with the highest average score of 0.1318 followed by MT9 or Baidu (0.1313), MT1 or Google Translator (0.1252), and MT4 or GPT4 (0.1235). These systems achieve moderate-level performance in terms of unigram overlap (as indicated by the R1-F1 score), good fluency (as indicated by the RL-F1 score), a high adequacy and word-order accuracy (as indicated by the METEOR score). However, there are areas for improvement, particularly in capturing bigram overlaps (as indicated by the R2-F1), overall precision and recall (as indicated by the BLEU score), and perceived quality compared to the reference (as indicated by the BLEURT score).
- **Moderate Performance:** MT7 or NiuTrans ranks fifth with an average score of 0.1188 followed by MT8 or Volcano Translator (0.1052). These systems achieve moderate-level performance in terms of unigram overlap (as indicated by the R1-F1 score), good fluency (as indicated by the RL-F1 score), moderate adequacy and word-order accuracy (as indicated by the METEOR score). Moreover, MT8 achieves the highest BLEU score (0.205) while MT7 achieves the second highest BLEU score (0.200). Again, there are notable weaknesses in capturing bigram overlaps, overall precision and recall, and perceived quality (as indicated by the R2-F1 and BLEURT scores).
- **Low Performance:** MT3 or Microsoft Translator, MT2 or DeepL, and MT5 or Llama2, are amongst the low-performers with the lowest average scores (0.0922, 0.0885, and 0.0693, respectively) These models achieve the lowest R2 F1 and BLEURT scores.

4.4 Automatic vs Human Evaluation

To determine which models perform best in terms of idiom translation, we conducted a comparison between human evaluation scores and automatic evaluation results. Our analysis reveals that GPT4 and GLM4 emerge as top performers, as both human evaluators and automatic evaluation techniques consistently classify these models in that category. Conversely, DeepL, Microsoft Translator, and Llama2 are consistently rated as low-performing models by both assessment methods. The superior performance of GPT4 and GLM4 can be further elucidated through examples from our dataset (refer to Figures 2 and 3).

Idiom	Examples	HT	Google	Deepl	Microsoft	GPT4	Llama2	GLM4	Niu	Volc	Baidu
如履薄冰	李主任抱住她，不觉有些动情，说道：“我这一生，是如履薄冰，如临深渊的一生。	Director Li was moved. He took her in his arms. "My whole life I feel like I have been walking on thin ice and now I'm at the edge of an abyss."	Director Li hugged her and felt a little emotional. He said that his life was like walking on thin ice or facing an abyss.	Director Li hugged her, do not feel some emotion, said, his life, is like walking on thin ice, such as the life of the abyss.	Director Li hugged her, didn't feel a little emotional, and said, his life is like walking on thin ice, like a life on the edge of the abyss.	Director Li hugged her, feeling moved, and said that his life has been like walking on thin ice, like standing at the edge of a precipice.	Mr. Li embraced her, feeling somewhat emotional, saying that his life was like walking on thin ice and standing at the edge of a deep abyss.	Director Li embraced her and couldn't help but feel moved. He said that his life had been like walking on thin ice, like standing on the edge of an abyss.	Director Li hugged her, feeling a little emotional, and said that his life has been walking on thin ice and facing an abyss.	Director Li hugged her, feeling a little emotional, and said, his life is like walking on thin ice, like a life in the abyss.	Director Li hugged her and couldn't help but feel a little emotional. He said that his life was like walking on thin ice and facing an abyss.
Human	annotator 1		1	1	1	1	1	1	1	1	1
	annotator 2		1	1	1	1	1	1	1	1	1
	annotator 3		1	1	1	1	1	1	1	1	1
	average		1	1	1	1	1	1	1	1	1
Automatic	average		0.25746327	0.152927321	0.21854042	0.235451137	0.286622195	0.325431268	0.242365421	0.188734099	0.265118998

Fig. 2. Example 1: Frequently used Chinese idiom.

Idiom	Examples	HT	Google	Deepl	Microsoft	GPT4	Llama2	GLM4	Niu	Volc	Baidu
刺刺不休	三人絮语闲杂，刺刺不休。	The three women chattered on at great length.	The three of them chattered incessantly and stabbed each other endlessly.	The three of them were talking in a mishmash of words, stabbing and stabbing.	The three of them chattered and stabbed endlessly.	The three of them chattered incessantly.	The three of them chatted endlessly, stabbing each other with their words.	The three of them chatted incessantly, chattering away.	The three of them chatted intermittently and stabbed endlessly.	The three of them were whispering, prickling endlessly.	The three of them chattered intermittently, piercing incessantly.
Human	annotator 1		0.416666667	0.416666667	0.416666667	0.916666667	0.416666667	0.916666667	0.25	0.25	0.25
	annotator 2		0.25	0.25	0.25	0.75	0.25	0.75	0.25	0.25	0.25
	annotator 3		0.25	0.25	0.25	0.916666667	0.25	0.916666667	0.25	0.25	0.25
	average		0.305555556	0.305555556	0.305555556	0.861111111	0.305555556	0.861111111	0.25	0.25	0.25
Automatic	average		0.079349	-0.001959393	0.086362889	0.195370716	0.049253986	0.114159446	0.043857049	0.012420153	0.156651235

Fig. 3. Example 2: Less frequently used Chinese idiom.

In Example 1, which features a frequently used idiom, all models demonstrate an ability to accurately capture its figurative meaning. However, Example 2 presents a less commonly used idiom, where most models struggle to produce fluent and correct translations. Here, the exemplary performance of GPT4 and GLM4 can be distinguished as both consistently deliver accurate and eloquent translations, highlighting their high-quality output even in challenging scenarios. This comprehensive analysis highlights the effectiveness of GPT4 and GLM4 in handling idiomatic expressions, affirming their status as top contenders in the field of machine translation for idioms.

4.5 Pearson Correlation

To gauge how well automatic evaluation metrics correspond with human assessments, Pearson correlation, r, was computed between human and automatic evaluation scores, using average human evaluation scores for accuracy, intelligibility, and elegance.

For most MT systems, a disagreement between human and automatic evaluations is observed, indicating potential limitations or biases in the automatic evaluation metrics. This suggests that the automatic evaluation metrics may not fully capture all aspects of translation quality as perceived by humans, highlighting the need for further refinement and development of evaluation methods. Some correlations (see Table 4) are statistically significant ($p < 0.05$), indicating a meaningful relationship between the MT system's performance and the automatic evaluation metrics. From the correlation coefficients, we can deduce valuable insights regarding the relationship between automatic evaluation metrics and human evaluation scores across all MT systems for idiom translations. Specifically, both BLEU and BLEURT exhibit positive correlations with human evaluation scores, suggesting that these metrics provide reliable indications of translation quality. However, it's crucial to note that several correlations, including R1-F1, R2-F1, RL-F1, and METEOR, do not reach statistical significance ($p > 0.05$). This lack of significance implies that the observed relationships may be attributed to random chance rather than a true association.

Since we cannot rely solely on any single automatic metric system, we sought to explore the potential of combining these metrics, considering their varied strengths. Upon computing the correlation with the average scores of all metrics, we observed a moderately positive correlation. This finding suggests that an aggregate of these metrics could offer a more comprehensive assessment of translation quality. Hence, in scenarios where human evaluators are unavailable and there is a need for automatic evaluation of a translation system involving idioms, utilising the average of these metrics could serve as a viable alternative for gauging the efficacy of the system. This approach leverages the diverse perspectives provided by multiple metrics, contributing to a more holistic evaluation of idiom translation quality.

5 Conclusion

Our study conducted a quantitative analysis of nine state-of-the-art MT systems and LLMs. The aim was to evaluate both the performance of MT systems and the ability of automatic metrics to gauge the idiom translation. From our experiments and results, several key conclusions can be drawn. GLM4 and GPT4 emerged as top performers in capturing the figurative meaning of Chinese idioms: these models achieved the highest scores in both human and automatic assessments. On the other hand, DeepL, Microsoft, and Llama2 ranked the lowest and exhibited a need for significant improvement in capturing the implied meaning. Our AIE evaluation metric suggests assigning distinct weights to its metrics, supported by empirical evidence. Among the automatic evaluation metrics, BLEU

Table 4. Pearson correlation (r) with Average of Human Evaluation Scores

		R1-F1	R2-F1	RL-F1	BLEU	BLEURT	METEOR	AVERAGE
MT1	**r**	0.025	0.061	0.049	**0.229**	**0.261**	0.182	**0.197**
	p-value	0.8	0.5	0.6	**0.0**	**0.0**	0.1	**0.0**
MT2	**r**	0.038	-0.005	0.053	**0.266**	**0.320**	0.063	**0.222**
	p-value	0.7	0.9	0.6	**0.0**	**0.0**	0.5	**0.0**
MT3	**r**	0.149	0.035	0.128	0.155	**0.287**	0.116	**0.230**
	p-value	0.1	0.7	0.2	0.1	**0.0**	0.3	**0.0**
MT4	**r**	0.141	0.114	0.181	**0.281**	0.174	0.195	**0.219**
	p-value	0.2	0.3	0.1	**0.0**	0.1	0.1	**0.0**
MT5	**r**	0.176	0.198	0.194	0.042	**0.316**	0.176	**0.262**
	p-value	0.1	0.1	0.1	0.7	**0.0**	0.1	**0.0**
MT6	**r**	0.124	0.099	0.109	**0.233**	**0.225**	**0.209**	**0.224**
	p-value	0.2	0.3	0.3	**0.0**	**0.0**	**0.0**	**0.0**
MT7	**r**	0.040	0.027	-0.017	**0.205**	**0.309**	0.154	**0.202**
	p-value	0.7	0.8	0.9	**0.0**	**0.0**	0.1	**0.0**
MT8	**r**	**0.231**	0.120	0.198	0.173	**0.415**	**0.240**	**0.349**
	p-value	**0.0**	0.2	0.1	0.1	**0.0**	**0.0**	**0.0**
MT9	**r**	**0.299**	**0.291**	**0.293**	0.099	**0.368**	**0.351**	**0.395**
	p-value	**0.0**	**0.0**	**0.0**	0.3	**0.0**	**0.0**	**0.0**

and BLEURT demonstrated moderate correlation with human scores, suggesting their utility in assessing the quality of MT systems. However, it is important to note that no single automatic evaluation metric provided a comprehensive assessment of the translations. Therefore, relying on a combination of all scores could offer a more holistic evaluation of MT systems.

Acknowledgments. The authors would like to thank Damith Dola Mullage, PhD researcher at Lancaster University, for his invaluable assistance with the experiments conducted for this study. Additionally, this work was supported by the China Scholarship Council [grant number 100829].

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arnold, D., Balkan, L., Humphreys, R.L., Meijer, S., Sadler, L.: Machine translation: An introductory guide

2. Bai, M.H., You, J.M., Chen, K.J., Chang, J.S.: Acquiring translation equivalences of multiword expressions by normalized correlation frequencies. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. pp. 478–486 (2009)
3. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
4. Baziotis, C., Mathur, P., Hasler, E.: Automatic evaluation and analysis of idioms in neural machine translation. arXiv preprint arXiv:2210.04545 (2022)
5. Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A.J., Koehn, P., Logacheva, V., Monz, C., et al.: Findings of the 2016 conference on machine translation (wmt16). In: First conference on machine translation. pp. 131–198. Association for Computational Linguistics (2016)
6. Celikyilmaz, A., Clark, E., Gao, J.: Evaluation of text generation: A survey. arXiv preprint arXiv:2006.14799 (2020)
7. Fadaee, M., Bisazza, A., Monz, C.: Examining the tip of the iceberg: A data set for idiom translation. arXiv preprint arXiv:1802.04681 (2018)
8. Farhad, A., Arkady, A., Magdalena, B., Ondřej, B., Rajen, C., Vishrav, C., Costa-jussa, M.R., Cristina, E.B., Angela, F., Christian, F., et al.: Findings of the 2021 conference on machine translation (wmt21). In: Proceedings of the Sixth Conference on Machine Translation. pp. 1–88. Association for Computational Linguistics (2021)
9. He, Z.: Baidu translate: Research and products. In: Proceedings of the Fourth Workshop on Hybrid Approaches to Translation (HyTra). pp. 61–62 (2015)
10. Imankulova, A., Kaneko, M., Hirasawa, T., Komachi, M.: Towards multimodal simultaneous neural machine translation (2020)
11. Jiao, W., Wang, W., tse Huang, J., Wang, X., Shi, S., Tu, Z.: Is chatgpt a good translator? yes with gpt-4 as the engine (2023)
12. Junczys-Dowmunt, M.: Microsoft translator at wmt 2019: Towards large-scale document-level neural machine translation. arXiv preprint arXiv:1907.06170 (2019)
13. Khashabi, D., Stanovsky, G., Bragg, J., Lourie, N., Kasai, J., Choi, Y., Smith, N.A., Weld, D.S.: Genie: Toward reproducible and standardized human evaluation for text generation. arXiv preprint arXiv:2101.06561 (2021)
14. Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., et al.: Findings of the 2023 conference on machine translation (wmt23): Llms are here but not quite there yet. In: Proceedings of the Eighth Conference on Machine Translation. pp. 1–42 (2023)
15. Kocmi, T., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Gowda, T., Graham, Y., Grundkiewicz, R., Haddow, B., et al.: Findings of the 2022 conference on machine translation (wmt22). In: Proceedings of the Seventh Conference on Machine Translation (WMT). pp. 1–45 (2022)
16. Krippendorff, K.: Computing krippendorff's alpha-reliability (2011)
17. Lin, C.Y., Cao, G., Gao, J., Nie, J.Y.: An information-theoretic approach to automatic evaluation of summaries. In: Proceedings of the Human Language Technology Conference of the NAACL, Main Conference. pp. 463–470 (2006)
18. Liu, E., Chaudhary, A., Neubig, G.: Crossing the threshold: Idiomatic machine translation through retrieval augmentation and loss weighting. arXiv preprint arXiv:2310.07081 (2023)

19. Loïc, B., Magdalena, B., Ondřej, B., Christian, F., Yvette, G., Roman, G., Barry, H., Matthias, H., Eric, J., Tom, K., et al.: Findings of the 2020 conference on machine translation (wmt20). In: Proceedings of the Fifth Conference on Machine Translation. pp. 1–55. Association for Computational Linguistics, (2020)
20. Mitkov, R., Seretan, V., Corpas Pastor, G., Monti, J.: Multiword units in machine translation and translation technology. Multiword units in machine translation and translation technology pp. 1–269 (2018)
21. Monti, J., Corpas Pastor, G., Mitkov, R., Hidalgo-Ternero, C.M. (eds.): Recent Advances in Multiword Units in Machine Translation and Translation Technology. John Benjamins Publishing Company (2024)
22. Newmark, P.: A textbook of translation, vol. 66. Prentice hall New York (1988)
23. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
24. Qiang, J., Li, Y., Zhang, C., Li, Y., Zhu, Y., Yuan, Y., Wu, X.: Chinese idiom paraphrasing. Transactions of the Association for Computational Linguistics **11**, 740–754 (2023)
25. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword expressions: A pain in the neck for nlp. In: Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing 2002 Mexico City, Mexico, February 17–23, 2002 Proceedings 3. pp. 1–15. Springer (2002)
26. Salton, G., Ross, R., Kelleher, J.D.: Evaluation of a substitution method for idiom transformation in statistical machine translation. 10th Workshop on Multiword Expres- sions (MWE 2014) at 14th Conference of the Euro- pean Chapter of the Association for Computational Linguistics (2014)
27. Sellam, T., Das, D., Parikh, A.P.: Bleurt: Learning robust metrics for text generation. arXiv preprint arXiv:2004.04696 (2020)
28. Shao, Y., Sennrich, R., Webber, B., Fancellu, F.: Evaluating machine translation performance on chinese idioms with a blacklist method. arXiv preprint arXiv:1711.07646 (2017)
29. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers. pp. 223–231 (2006)
30. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models (2023)
31. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.: Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
32. Zitawi, J.: English-arabic dubbed children's cartoons: Strategies of translating idioms. Across languages and cultures **4**(2), 237–251 (2003)