Evaluation of Chinese-English Machine Translation of Emotion-Loaded Microblog Texts: A Human Annotated Dataset for the Quality Assessment of Emotion Translation

Shenbin Qian[1], **Constantin Orăsan**[1], **Félix do Carmo**[1],
Qiuliang Li[2], **Diptesh Kanojia**[3]
Centre for Translation Studies, University of Surrey, UK[1]
Xi'an International Studies University, China[2]
Department of Computer Science, University of Surrey, UK[3]
{s.qian, c.orasan, f.docarmo, d.kanojia}@surrey.ac.uk[1,3]
qiuliang0909@gmail.com[2]

Abstract

In this paper, we focus on how current Machine Translation (MT) tools perform on the translation of emotion-loaded texts by evaluating outputs from Google Translate according to a framework proposed in this paper. We propose this evaluation framework based on the Multidimensional Quality Metrics (MQM) and perform a detailed error analysis of the MT outputs. From our analysis, we observe that about 50% of the MT outputs fail to preserve the original emotion. After further analysis of the errors, we find that emotion carrying words and linguistic phenomena such as polysemous words, negation, abbreviation *etc.*, are common causes for these translation errors.

1 Introduction

To express feelings and attitudes is one of language's major functions (Waugh, 1980). In this digital age, people can easily share their emotions or opinions online using social media platforms. This results in the generation of a large amount of emotion-loaded and opinionated texts. It is important to convey the correct emotion or opinion in the text to a large audience from different linguistic or cultural backgrounds for cross-cultural communication. Otherwise, misinformation or even toxic emotions (Frost, 2003) can permeate cross-cultural communication, resulting in harmful implications for the parties involved. Due to the asynchronous nature and sheer quantity of this generated text online, it is impossible for human translators to be present in the loop and perform accurate translations. Hence, machine translation (MT) remains the only viable choice for the task of translating emotion-loaded microblog texts (Carrera et al., 2009).

Social media texts on Sina Weibo[1], the Chinese microblog platform, have their unique characteristics due to certain features of the Chinese language. Since Chinese is a tonal language, there are many characters which share the exact same or very similar pronunciation but with drastically different meanings. Chinese netizens commonly use this language phenomenon to create emotional slang by replacing the original character/word with a homophone character/word to avoid censorship. Similarly, substitution with homographs is another way to create slang, as Chinese is a hieroglyphic language. For example, using "目田", which means "eye field", and substituting them for "自由", meaning "freedom" is an example of homograph substitution (King et al., 2013). We can observe that "目田" looks very similar to "自由", where a few strokes of the two characters are omitted to refer to the lack of freedom. Abbreviation of long expressions or transliteration of Chinese characters is another observed phenomenon in social media texts. Such features in this new online language variant pose severe challenges to the MT of Chinese social media texts, especially the emotion-loaded and opinionated microblogs. These challenges are different from the ones observed in translating tweets with

[1] `https://weibo.com/`

hashtags or non-standard orthography present in the other languages (Saadany et al., 2021b).

There are several studies and datasets which focus on the translation of social media texts, such as TweetMT (San Vicente et al., 2016), the tweet corpus proposed by Mubarak et al. (2020) and the Weibo corpus developed by Ling et al. (2013). However, none of these focus on the translation of emotions. To the best of our knowledge, there is no research which focuses on the Chinese-English machine translation (C-E MT) of emotion-loaded texts. We endeavour to make our contributions to this area as summarised below:

- A quality assessment framework for the machine translation of emotion-loaded texts is proposed for evaluating the MT quality in terms of emotion preservation.
- A detailed error analysis is performed to find out linguistic phenomena that are more likely to cause C-E MT errors in terms of emotions.
- A dataset[2], annotated with translation errors and severity levels, is released to support tasks like error detection and quality estimation of emotion translation.

Section 2 describes the related literature in emotion translation and quality assessment of MT. Our proposed framework for human evaluation of the MT quality of emotion-loaded texts is described in Section 3. In Section 4, we introduce the dataset and methodology for quality assessment. The result of human evaluation and error analysis is presented and analysed in Section 5. Section 6 discusses the conclusion and future plan after summarising the whole paper.

2 Related Work

2.1 Translation of Emotions and Emotion-Loaded Texts

The awareness of emotions in translation has been discussed in the early stages of translation studies when the emotional reaction of the reader was of significance in the translation of the Bible (Lehr, 2020). Nida and Taber (1969) emphasised the importance of transferring emotional elements from source to target and proposed to translate the emotionality of the text with a focus on the final translation product.

Many studies focused on the emotional difference or emotion translation between languages, most of which emphasised on the translation of emotion lexica. Russell and Sato (1995) compared 14 emotional words such as 'happy' or 'sad' in English, Chinese and Japanese to observe similarities and differences post-translation. Choi and Han (2008) raised concerns about cross-cultural communication regarding the difficulty of finding the equivalence of some emotional concepts such as *'shimcheong'* (a combination of empathy, sympathy, and compassion) in Korean. Similarly, Hurtado de Mendoza et al. (2010) also raised questions about one-to-one translations of emotion concepts like 'shame' in English and Spanish. For other language pairs like English and Arabic, Kayyal and Russell (2013) did very similar studies and found that only one pair (happiness-farah) passed their equivalence tests, and other lexical pairs differed in terms of culture and language. For English and Indonesian, the emotion 'happy' can be translated into several different words including '*bahagia*', '*senang*', '*suka*', '*lega*', '*kesenangan*', '*gembira ria*', '*riang*', '*ceria*', '*patah hati*', and '*tenteram*' (Suryasa et al., 2019). They are not the same in meaning or style, so translating such words might lead to subtle emotional differences in the target language.

These studies reveal the challenges and importance of translating emotions or emotional lexica in cross-cultural communication. But very few studies focused on machine translation or the quality of machine translation regarding emotion preservation. Mohammad et al. (2016) examined sentiments in social media posts in both Arabic-English and English-Arabic translations, and they found that the change of sentiment was mainly caused by ambiguous words, sarcasm, metaphors, and word-reordering issues. Shalunts et al. (2016) also performed experiments to explore the impact of MT on sentiment analysis in German, Russian and Spanish using general news articles. They surprisingly found that the performance of the sentiment analysis tool on the source and the target was comparable, which

[2] https://github.com/shenbinqian/HADQAET

indicated that the impact of machine translation on sentiment was not obvious. Contrary to their result, Fukuda and Jin (2022) found that sentiment was significantly affected by MT tools. More specifically, positive sentences tended to be more similar in sentiment polarity before and after translation than negative and neutral sentences. Apart from the aforementioned manual or sentiment score-based evaluation of emotion translation, Saadany et al. (2021a) proposed a sentiment-aware measure which can be used to adjust automatic evaluation metrics like BLEU (Papineni et al., 2002) for the evaluation of MT quality of user-generated content.

As can be seen above, most of the work does not focus on proposing a systematic human evaluation framework to assess the MT quality in terms of emotion preservation, especially not for Chinese-English translation. Our work focuses specifically on this particular use case.

2.2 Quality Assessment of Machine Translation

In the MT area, there are several different automatic and human evaluation methods for assessing MT quality. Among those automatic evaluation metrics, BLEU is the most used tool for this purpose. However, BLEU has been criticised for the lack of recall and the "explicit word-matching between translation and references" (Banerjee and Lavie, 2005). Other metrics like ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) were proposed as an alternative to BLEU, but the resultant evaluation has been similar when compared to BLEU in terms of the n-gram matching. More recently, since the rise of BERT-like models (Devlin et al., 2018), metrics like BERTScore (Zhang et al., 2019) have been proposed to calculate the similarity between the candidate/hypothesis and the reference translation to evaluate MT quality.

An alternative way to measure quality is to figure out how much post-editing is needed for the candidate translation to match with the reference translation. Translation Edit Rate (TER), which is defined as "the minimum number of edits needed to change a hypothesis so that it exactly matches one of the references, normalized by the average length of the references" (Snover et al., 2006), is a metric that measures this error based on edit distance.

More recently, Direct Assessment (DA) (Graham et al., 2013) of the translation output, which provides a continuous score within a certain range after the annotator sees a candidate translation and a translation hint, has been used in various ways. It can be used directly to evaluate translation quality as it is obtained from human annotators. It is also used as an input for training quality estimation models in recent Conferences of Machine Translation[3]. Apart from DA, the MQM framework (Lommel et al., 2014) provides a more detailed evaluation methodology. It divides translation errors into six dimensions *i.e.,* accuracy, fluency, design, locale convention style, terminology, and verity. Each dimension consists of several error categories like addition, mistranslation, omission or untranslated under the accuracy dimension, and more fine-grained subcategories (Lommel, 2018). Each error falls into at least one of these categories and contributes to the overall rating of the translation. Error severity could be added as weights to the rating according to the seriousness of these errors. Eventually, an evaluation score can be calculated to measure the overall translation quality using the framework. The practicality, reliability, and validity of this framework (Mariana et al., 2015) have made it the choice of the translation industry and MT evaluation research.

Nevertheless, all the above automatic methods were proposed without taking into account any elements of meaning or emotion, and human evaluation metrics were proposed for the assessment of general MT quality, which might be too generic or over-complicated for specific needs like emotion preservation.

3 Framework for Quality Assessment of Emotion-Loaded Texts

To evaluate the preservation of emotions, we modify the MQM framework (Lommel et al., 2014) for the assessment of MT quality of emotion-loaded microblog texts. Since our focus is on the emotion preservation, we simplify the multidimensional metrics into one dimen-

[3] https://www.statmt.org/

sion, *i.e.,* the accuracy of translating emotions. Our error types follow the accuracy dimension of MQM, *i.e.,* addition, mistranslation, omission, untranslated and source error, but we only consider errors that affect emotion. For instance, an addition error is an error in translation that adds information which does not exist in the source and the addition of this information affects the emotion in the target. Our severity levels are defined based on MQM suggestion: critical, major, and minor, which indicates how severely the source emotion is affected by the error. We define them as follows:

- **a critical error** leads to an absolute change of emotion into a very different or even opposite emotion category;
- **a major error** pertains to a change of emotion into one that is not very different from the original emotion or one that is somewhere between the original emotion category and another different category;
- **a minor error** results in a slight change of emotion with uncertainties about the MT emotion label but certainties about the slight difference between the emotions of the source and the MT text.

Similar to the MQM translation quality score (Lommel et al., 2014), we can also compute evaluation scores regarding emotion preservation by summing up all errors as per their severity level weights. Severity level weights are defined in the MQM framework and for this study, we define them as follows: 10 for critical errors, 5 for major errors and 1 for minor errors. The error rate or evaluation score of emotion translation can now be computed using Equation 1. Examples of error annotation can be seen in the Appendix.

$$\begin{array}{l} Error\ Rate = \dfrac{\sum_{n=1}^{n} Error_n * Weight_s}{Text\ Length} \\ Weight_s:\ weight\ given\ to\ each\ error \\ according\ to\ its\ severity\ level \\ Text\ Length:\ count\ of\ all\ words\ and \\ punctuations\ in\ the\ target\ text \end{array} \qquad (1)$$

4 Data and Methodology

4.1 Data Description

To evaluate the transfer of emotions, we need the source text to be full of emotions. The dataset for the *Evaluation of Weibo Emotion Classification Technology on the Ninth China National Conference on Social Media Processing*[4] (SMP2020-EWECT) is an ideal source for our purposes. It was annotated with six emotion categories, namely, anger, fear, joy, sadness, surprise and neutral, which was provided by the Harbin Institute of Technology and sourced from Sina Weibo (Guo et al., 2021).

Since the dataset is as large as 34,768 entries and it includes Weibo posts with neutral emotions as well, we filter out those posts with neutral emotions and randomly sample 20 percent (about 5500 entries) for machine translation and quality assessment. The distributions of the emotion labels of our sampled dataset and the original SMP2020-EWECT dataset can be seen in Figure 1. We can see that our sampled dataset keeps the original data distribution. We use Google Translate [5] to translate the source text of our sampled dataset and the output is used for quality assessment.

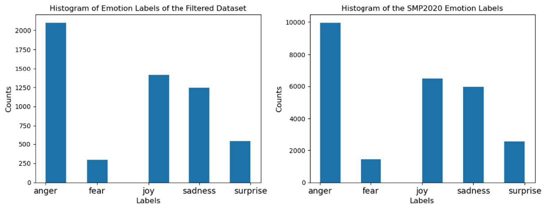

Figure 1: Distributions of Emotion Categories for the Filtered VS Original Dataset

4.2 Methodology

Re-annotation of the emotions in the MT output may prove difficult in some cases due to the fact that some outputs do not make any sense for humans. For example, the MT output "Playing this old game, I just have no friends..." may not make much sense and it is difficult to annotate it with an emotion label. However, a bilingual annotator can easily see that the emotion of the source "玩这个老游戏，我简直是叼到没朋友…" which means "Playing this old game, I'm just too good to have

[4] `https://smp2020ewect.github.io/`

[5] Results from "https://translate.google.co.uk/" on the 30th of May 2022.

rivals", is not present in the target. Therefore, we do not re-annotate the raw MT with emotion labels to check possible loss of emotions. Instead, we assess the quality of MT using the framework in Section 3.

Two annotators with Chinese-English translation qualifications were recruited to annotate error types and severity levels. All translation errors coupled with severity levels that affect the transfer of original emotions were annotated in the MT output. Words or parts of the text in both source and target in relation to the translation errors were highlighted so that they can be used for error analysis. The annotators were given clear and detailed instructions about the decision process behind the annotations. We released the annotation guidelines along with the annotated dataset in our GitHub repository for inspection and reproducibility.

Since the perception of emotion usually varies a lot among people and across time, we randomly sampled 10% (about 550 entries) of the whole dataset for the inter-annotator agreement check and 100 entries for the intra-annotator agreement check to measure how well annotators agree with each other and themselves. The intra-annotator agreement was done by one annotator annotating the same 100 samples twice two months apart.

5 Result of Human Evaluation

This section shows the result of human evaluation on our Weibo dataset based on the framework and methodology proposed in previous sections. We first show the result of inter and intra-annotator agreement and then analyse the evaluation result from two aspects: 1) how many errors there are and how severe these errors are in terms of emotion category and error type; 2) what are the linguistic phenomena that are the likely cause for these errors.

5.1 Result of the Inter and Intra-Annotator Agreement

We use the Cohen Kappa score (Cohen, 1960) to calculate the inter and intra-annotator agreements. Table 1 shows that the Kappa scores for intra-annotator agreement are very high, which means the annotator is consistent with himself/herself during annotation. Inter-annotator agreement is relatively lower, especially for the error severity. So we compared the severity levels of the two annotators and found they are more likely to disagree on whether there is a minor error (or no error). Disagreement on major/critical errors comes the second. This may be partially because different people perceive emotions differently. To further analyse the reasons, we collect some examples which annotators disagree.

	Error Existence	Type	Severity
Inter-AA	0.6689	0.5117	0.3691
Intra-AA	0.8991	0.8990	0.7634

Table 1: Cohens Kappa for Inter and Intra-Annotator Agreement (AA) for Error Existence, type and severity.

One of the main causes is the disagreement on the change of the subject of emotion. For example, the MT output of the source "吓死宝宝了" meaning "Scared me to death", is "Scared the baby to death". One annotator annotates it as a minor error, while the other as a major error. In this example, the subject of emotion should be "me" rather than a third party, "the baby", which might result in the reduction of the strong emotion and the transformation of the emotion from "fear" into somewhere between "fear" and "anger". Annotators are likely to disagree on the severity level of this case.

Emotion conflicts caused by mistranslation is another problem which annotators disagree. For instance, the source emotion of this post "我容易嘛我 黑眼圈, 青春痘, 眉毛, 皱纹 全在这两天爆出来了" is sadness, which means "Life is so hard on me. Dark circles, pimples, eyebrows, wrinkles all had an explosive growth in the past two days", but the MT output "I'm easy. I Dark circles, pimples, eyebrows, wrinkles have all exploded in the past two days" may contain both joy and sadness, two conflicting emotions. This causes the disagreement on the severity level, as one annotator annotates it as a critical error, while the other as a major error.

The complete change of meaning in the target but with the similar emotion as the source is another major cause. For example, the emotion of the MT output "His mother got a leg and caught a cold again, mad at me" might be anger or sadness, which is similar to the emo-

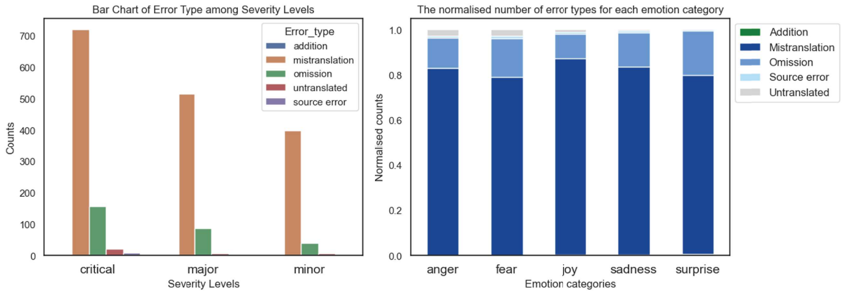

Figure 2: Error Types against Severity Levels and Emotion Categories where the first chart (left) shows the error types among severity levels and the second shows normalised counts for error types among emotion categories.

tion of the source “他娘了个腿的，又感冒了，气死我了”, but the target meaning is completely different from the source “F**k your mother, Cold again! I’m so pissed off”. One annotator annotates it as a critical error, while the other as a major error.

5.2 Error Statistics

After annotating each entry of the dataset, we collect all error entries and display error statistics in the following figures to see 1) how many examples are incorrectly translated; 2) which type of error is most common; 3) which emotion category is less likely to be mistranslated; and 4) which error type is more critical.

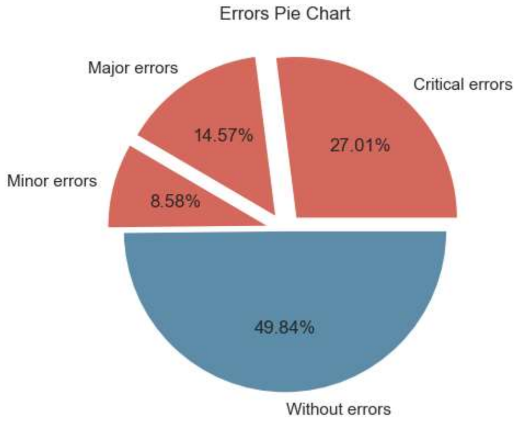

Figure 3: Error Severity in overall MT output

From Figure 3, we know the MT quality of these texts is not acceptable as about 50% of the entries have errors in preserving emotions and 41.58% have major or critical errors.

Among these error severity levels, mistranslation is the most common error type followed by omission according to the left chart in Figure 2. In the right bar chart of Figure 2, we normalise the number of error types of each emotion category against the total number of errors. We can see the pattern is very similar for all emotion categories, which suggests mistranslation is the most common error type and omission comes the second.

In the left bar chart of Figure 4, we normalise the number of errors in each emotion category against the overall number of the dataset. We see that ‘joy’ accounts for the least errors despite it having the second largest number of total entries, which means that those social media texts with the emotion of ‘joy’ are more likely to be translated correctly by Google Translate, compared with other emotion categories. This can be further proved by the right chart of Figure 4, where normalised counts of severity levels are plotted for each emotion category. We can see from critical errors to no error, as the severity level decreases, the number of ‘joy’ increases. This suggests errors in the ‘joy’ category are more likely to be minor. For those entries without errors, ‘joy’ takes the largest percentage among all emotion categories. This result corresponds with the study by Fukuda and Jin (2022), which indicated that positive sentences are less likely to be affected by MT compared with negative and neutral sentences.

In Figure 5, we normalise the number of error severity for each error type against the total number of errors. We can see that for all error types, critical errors take the largest percentage except for addition. In the addition category, minor errors are much more than critical errors, which means addition errors are less likely to have severe impact on emotions. That is maybe because the original emotion would not be changed a lot if we just add some extra words in the target text. For the untranslated

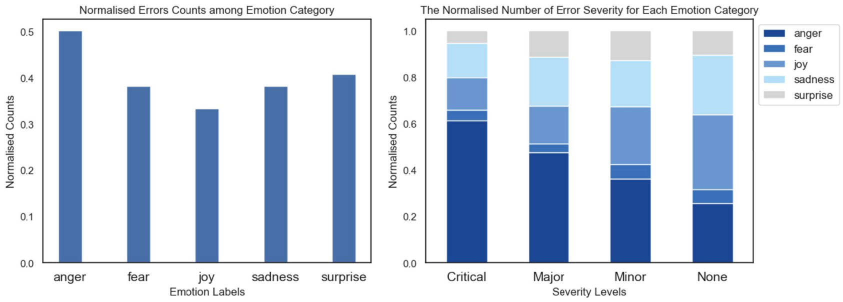

Figure 4: Errors among Emotion Categories where the first chart (left) shows normalised error counts among emotion categories whereas the second chart shows normalised counts of severity levels among emotion categories.

category, critical errors are far more than other types. This suggests that untranslated errors affect the transfer of emotion quite severely.

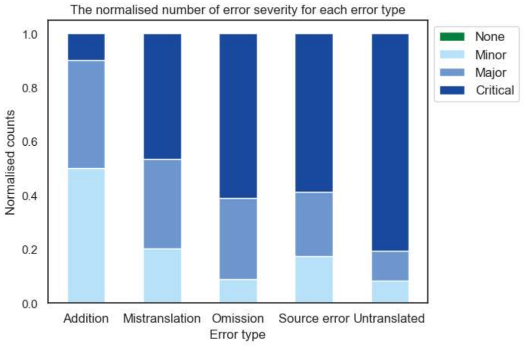

Figure 5: Normalised Error Severity in Error Types

5.3 Analysis of Error Causes

In this section, we investigate linguistic phenomena that are responsible for the translation errors in the MT output based on annotation described in Section 4. We first discuss errors caused by emotion carrying words and then by other linguistic phenomena.

5.3.1 Emotion Carrying Words

To find out the most common cause of these translation errors, we collect all the words and sentences identified during annotation as corresponding to an error and then find out where the error occurs. We count the frequency of these words and sentences, and calculate the percentage of the words in total erroneous entries as shown in Table 2 and Table 3.

Source	Frequency	Human Translation	Word Percentage
尼玛	50	(f**k) your mother	2.19%
居然	42	actually	4.37%
竟然	22	surprisingly	2.96%
特么	20	what's the f**k	1.86%
TMD	14	WTF	0.58%
TM	14	WTF	1.29%
还是	12	still	5.59%
真是醉了	12	rcally spcechlcss	0.45%
日了狗了	10	f**kcd up	0.39%
折腾	10	mess around	0.64%
草泥马	10	f**k your mother	0.71%

Table 2: Most Frequent Words in Erroneous Examples

We can see from "Human Translation"[6] column in Table 2 that almost all the frequent words are emotion carrying words. Some of them, including the most frequent word "尼玛", are emotional slang created by homophone character substitution (Chu and Ruthrof, 2017). Others such as "居然", "竟然" are emotional adverbs used to show strong feelings. Many of these emotion carrying words (top five) take a large percentage among all erroneous entries. For example, "尼玛" appears in 2.19% of the erroneous entries in emotion translation.

Source	Frequency	Human Translation
我也是醉了	12	I'm really speechless
吓死宝宝了	8	scared me to death
我 tm 快炸了	4	I'm f**king exploding
不要不要的	4	super/very
服了自己了	4	disappointed to myself

Table 3: Most Frequent Short Sentences in Erroneous Examples

Table 3 shows the most frequent 5 sentences among those erroneous examples. We can see that these short sentences also contain slang or adverbial clauses that convey strong emotions. From both tables, we observe that emotion carrying words pose a strong challenge to translation.

[6] Human translations here and in the rest of the paper are provided by a professional translator.

5.3.2 Other Linguistic Phenomena

Other linguistic phenomena like polysemous words, abbreviation, negation, subject/object issues, subjunctive mood and punctuation problems *etc.*, also play a role in causing these errors in emotion translation.

5.3.2.1 Polysemous Words

Polysemous words especially those having several different meanings can be easily mistranslated, which might result in the change of the original emotion. In the following example, the character "疼" in the source literally means "hurt", but in the Chinese culture, it can represent an emotion called "heart-aching love" which refers to the love that children get from their doting parents or lovers get from their partners (Sundararajan, 2015). MT clearly mistranslates the source emotion.
Source Text (ST): 介个女人说会**疼**我一辈子
Machine Translation (MT): Tell a woman that she will **hurt** me for the rest of my life
Human Translation (HT): This woman said she will **love** me for the rest of her life.

5.3.2.2 Abbreviation

Internet slang in Chinese can be created by abbreviation, which shortens a longer expression into a word/phrase. In the source of the following example, "活久见" literally meaning "live long see" is an abbreviation of "**活**的时间**久**什么事都可能**见**到", which is often used to imply surprise. Mistranslation of this abbreviation by MT leads to the misunderstanding and change of the source emotion.
ST: **活久见**，我还是比较适合高冷。就一个人喜欢我萌。晚安
MT: **See you for a long time**, I am still more suitable for high cold. The only one who likes me is cute. Good night
HT: **If you live long enough, you can see anything unexpected**. I am more suitable for being cool. Only one person sees me as cute. Good night.

5.3.2.3 Negation

Mistranslation of negation is a known problem for MT affecting both the emotion preservation and the understanding of a text. In the following example, the source character "好" means "very" not the common meaning of "good" and "不" is the negative word, but in the MT result, only "好" is kept as "good" not the correct meaning of "very" and the negation is omitted.
ST: 心情**好不**爽
MT: I'm in a **good** mood
HT: I'm in a **very bad** mood.

5.3.2.4 Subject/Object Issues

Since Chinese is not a subject prominent language (Tan, 1991), omission of subject is a quite common phenomenon in Chinese especially in informal texts. The omission of the subject in the source causes the swap of the subject and object in MT and results in a change of the emotion subject. This further affects the emotion of the MT as it becomes closer to fear rather than anger.
ST: 拉我一下能死吗
MT: Can I die if I pull
HT: Will you die if you pull me up?

5.3.2.5 Subjunctive Mood

Chinese does not have syntactic markers for counterfactual conditionals as the subjunctive mood in English (Feng and Yi, 2006). The source text expresses the wish to run the first place, but machine translation does not render it into the English subjunctive mood, affecting the transfer of the original anger emotion.
ST: 再跑不到第一把在我前面的都删了
MT: I can't run the first one. I deleted the one in front of me.
HT: If I didn't run the first place, I would delete all those who run ahead of me.

5.3.2.6 Punctuation Problems

Nonstandard use of punctuation in Chinese microblogs is another challenge posed to emotion translation. Here, the following source text is separated by exclamation marks, which shows strong emotions. But in the MT output, each separated character is regarded as an independent sentence. Such mistranslations change the original emotion, as the character "好" meaning "very" is translated as "good".
ST: 我！好！饿!!!!!
MT: I! it is good! hungry! ！！！！

HT: I AM SO HUNGRY!!!!!

The following example shows problems caused by the lack of punctuation. Since there is no space between Chinese characters, it is difficult for MT systems to tokenise the sentence. The lack of punctuation in some entries in the dataset seems to be highly correlated with the quite frequent omission of some emotion loaded parts in the text.
ST: 到底什么时候去考试啊老是忽悠我再拖下去没心情去考试
MT: When are you going to take the test
HT: When are we going to take the exam? Always fooling me. I would be in a bad mood if it postponed again.

5.3.2.7 Hallucination

Hallucination (Lee et al., 2018) is a common problem for neural machine translation, but it is rarely seen in this dataset. We only see the following example of hallucination, which might probably be caused by continuous repetition of some characters since the MT result keeps changing as we edit the repetitive characters. Hallucination is definitely a problem for the preservation of the source emotion.
ST: 次奥真特么是醉了
MT: 200022000
HT: WTF WTF WTF WTF WTF WTF WTF WTF WTF WTF WTF WTF WTF WTF WTF WTF WTF WTF WTF WTF I'm f**king speechless.

6 Conclusion and Future Work

Our work investigates the performance of MT engines on the translation of emotion-loaded texts. We propose a new framework for evaluating MT quality in terms of emotion preservation developed in line with the MQM evaluation framework. We perform a manual evaluation of the MT output and present a detailed error analysis. We observe which type of errors is the most common and which emotion category is more likely to be correctly translated by MT. Our detailed analyses describe which linguistic factors such as emotion carrying words, subject omission and so on, cause these errors in translating microblog texts loaded with emotions. Furthermore, the annotated bilingual dataset can be used for training quality estimators to automatically assess the translation quality while preserving emotions. In future, we aim to extend this dataset with reference translations and use it to train computational models for estimating the translation quality of emotion-loaded texts. We plan to conduct further research and perform more analyses to improve the proposed framework.

References

Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. pages 65–72. Association for Computational Linguistics, 6.

Carrera, Jordi, Olga Beregovaya, and Alex Yanishevsky. 2009. Machine Translation for Cross-Language Social Media.

Choi, Sang-Chin and Gyuseog Han. 2008. SHIMCHEONG PSYCHOLOGY: A CASE OF AN EMOTIONAL STATE FOR CULTURAL PSYCHOLOGY. *International Journal for Dialogical Science Copyright*, 3:205–224.

Chu, Yingchi and Horst Ruthrof. 2017. The social semiotic of homophone phrase substitution in Chinese netizen discourse. *Social Semiotics*, 27:640–655.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint.*

Feng, Gary and Li Yi. 2006. What If Chinese Had Linguistic Markers for Counterfactual Conditionals? Language and Thought Revisited. pages 1281–1286.

Frost, P. 2003. *Toxic emotions at work: How compassionate managers handle pain and conflict.* HBS Press.

Fukuda, Karin and Qun Jin. 2022. Analyzing change on emotion scores of tweets before and after machine translation. volume 13315, pages 294–308. Springer.

Graham, Yvette, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. pages 33–41. Association for Computational Linguistics, 8.

Guo, Xianwei, Hua Lai, Yan Xiang, Zhengtao Yu, and Yuxin Huang. 2021. Emotion Classification of COVID-19 Chinese Microblogs based on the Emotion Category Description. pages 916–927. Chinese Information Processing Society of China, 8.

Hurtado de Mendoza, Alejandra, José Miguel Fernández-Dols, W. Gerrod Parrott, and Pilar Carrera. 2010. Emotion terms, category structure, and the problem of translation: The case of shame and vergüenza. *Cognition and Emotion*, 24:661–680, 6.

Kayyal, Mary H. and James A. Russell. 2013. Language and Emotion: Certain English-Arabic Translations Are Not Equivalent. *Journal of Language and Social Psychology*, 32:261–271, 9.

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2013. How censorship in China allows government criticism but silences collective expression. *American Political Science Review*, 107:326–343.

Lee, Katherine, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in Neural Machine Translation.

Lehr, Caroline. 2020. Translation, emotion and cognition. pages 294–309. Routledge, 5.

Lin, Chin-Yew. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. pages 74–81. Association for Computational Linguistics, 7.

Ling, Wang, Guang Xiang, Chris Dyer, Alan Black, and Isabel Trancoso. 2013. Microblogs as Parallel Corpora. pages 176–186. Association for Computational Linguistics, 8.

Lommel, Arle Richard, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional Quality Metrics: A Flexible System for Assessing Translation Quality. *Tradumàtica: tecnologies de la traducció*, 0:455–463, 12.

Lommel, Arle. 2018. Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies. volume 1, pages 109–127. Springer.

Mariana, Valerie, Troy Cox, and Alan Melby. 2015. The Multidimensional Quality Metrics (MQM) Framework: a new framework for translation quality assessment. *The Journal of Specialised Translation Issue*, 23.

Mohammad, Saif M, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55:95–130.

Mubarak, Hamdy, Sabit Hassan, and Ahmed Abdelali. 2020. Constructing a Bilingual Corpus of Parallel Tweets. pages 14–21. European Language Resources Association, 5.

Nida, E. A. and C. R. Taber. 1969. *The Theory and Practice of Translation*. Brill.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. pages 311–318. Association for Computational Linguistics.

Russell, James A. and Kaori Sato. 1995. Comparing Emotion Words between Languages. *Journal of Cross-Cultural Psychology*, 26:384–391.

Saadany, Hadeel, Constantin Orăsan, Emad Mohamed, and Ashraf Tantawy. 2021a. Sentiment-Aware Measure (SAM) for Evaluating Sentiment Transfer by Machine Translation Systems. pages 1217–1226. INCOMA Ltd., 9.

Saadany, Hadeel, Constantin Orăsan, Rocío Caro Quintana, Félix Do Carmo, and Leonardo Zilio. 2021b. Challenges in Translation of Emotions in Multilingual User-Generated Content: Twitter as a Case Study. *arXiv preprint*.

San Vicente, Iñaki, Iñaki Alegría, Cristina España-Bonet, Pablo Gamallo, Hugo Gonçalo Oliveira, Eva Martínez Garcia, Antonio Toral, Arkaitz Zubiaga, and Nora Aranberri. 2016. TweetMT: A parallel microblog corpus. pages 2936–2941. European Language Resources Association (ELRA), 5.

Shalunts, Gayane, Gerhard Backfried, and Nicolas Commeignes. 2016. The impact of machine translation on sentiment analysis. pages 51–56. IARIA.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. pages 223–231. Association for Machine Translation in the Americas, 8.

Sundararajan, Louise. 2015. *Understanding Emotion in Chinese Culture: Thinking Through Psychology*. Springer.

Suryasa, I. Wayan, I. Nengah Sudipa, Ida Ayu Made Puspani, and I. Made Netra. 2019. Translation Procedure of Happy Emotion of English into Indonesian in Kṛṣṇa Text. *Journal of Language Teaching and Research*, 10:738–746.

Tan, Fu. 1991. *Notion of subject in Chinese*. Ph.D. thesis, Stanford University.

Waugh, L.R. 1980. The Poetic Function in the Theory of Roman Jakobson. *Poetics Today*, 2:57–82.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. *arXiv preprint*, 4.

Appendix – Example of Data Annotation

No.	Source	MT output	Human Translation	Original emotion label	Error type	Error severity
1	管理学真是水的一比，努力的想听，依然坚持不过一分钟……考研怎么办呀	Management is really a comparison of water. I want to listen hard, but I still can’t hold on for a minute...What about the postgraduate entrance examination?	Management is really a bunch of fiddle-faddle. I try hard to listen, but still can’t hold on for a minute...What about the postgraduate entrance examination?	anger	mistranslation	critical
2	【山羊也爱玩水，冲浪不输人！！】OMG，简直了～～逆天了～～～	[Goats also love to play in the water, and surfing is not lost! ! 】 OMG, it’s just	【Goats also love to play in the water, and they surf as well as humans! 】 OMG, it’s so crazy, unbelievable!	surprise	mistranslation	critical
3	丢死人了第一次在舞台上唱歌虽然不大但自己也没有准备好开始就忘词	It’s the first time to sing on stage, although it’s not big, but I’m not ready to start and forget the lyrics	So embarrassing! It’s the first time to sing on stage. Although it’s not big, I’m not ready to start and I forget the lyrics at the beginning.	sadness	omission	major
4	尼玛太吓人了今天在家乐福买的肉煮出来汤是黄色的！咱把萝卜煮了萝卜都黄了尼玛要日天啊。。。。。。	Nima is so scary. The meat I bought at Carrefour today turned out to be yellow in soup! We boiled the radish and the radish turned yellow. Nima is going to be every day.	So fucking scary! The meat I bought at Carrefour today turned out to be yellow in soup! We boiled the radish and it turned yellow as well. That’s fucking crazy!	fear	mistranslation; mistranslation	major; critical
5	玩这个老游戏，我简直是叼到没朋友…	Playing this old game, I just have no friends...	Playing this old game, I’m just too good to have rivals!	joy	mistranslation	critical