Challenges of Human *vs* Machine Translation of Emotion-Loaded Chinese Microblog Texts

Shenbin Qian s.qian@surrey.ac.uk
Constantin Orăsan c.orasan@surrey.ac.uk
Félix do Carmo f.docarmo@surrey.ac.uk
Centre for Translation Studies, University of Surrey, Guildford, UK

Diptesh Kanojia d.kanojia@surrey.ac.uk
Department of Computer Science, University of Surrey, Guildford, UK

Abstract

This paper attempts to identify challenges professional translators face when translating emotion-loaded texts as well as errors machine translation (MT) makes when translating this content. We invited ten Chinese-English translators to translate thirty posts of a Chinese microblog, and interviewed them about the challenges encountered during translation and the problems they believe MT might have. Further, we analysed more than five-thousand automatic translations of microblog posts to observe problems in MT outputs. We establish that the most challenging problem for human translators is emotion-carrying words, which translators also consider as a problem for MT. Analysis of MT outputs shows that this is also the most common source of MT errors. We also find that what is challenging for MT, such as non-standard writing, is not necessarily an issue for humans. Our work contributes to a better understanding of the challenges for the translation of microblog posts by humans and MT, caused by different forms of expression of emotion.

1 Introduction

User-generated texts (UGT) on social media commonly express sentiments and emotions. The emotion-loaded nature and the non-standard characteristics of UGT make translation difficult. This is true for texts on Sina Weibo[1], the largest Chinese microblog platform, which have their own characteristics due to the unique features of the Chinese language. Since Chinese is a tonal language, there are many characters sharing the same or similar pronunciation, but with quite different meanings. Similar to this feature, homographs that look very similar in writing but with different meanings and pronunciations are also common in Chinese, as there are many logograms and morphograms in Chinese. Netizens use this feature to create emotion-carrying slang by replacing the original character/word with a homophone or homograph character/word to avoid censorship. For example, "草泥马" *caonima*, literally meaning "grass mud

[1] https://weibo.com/

horse", is coined to refer to a swear word as it shares a similar pronunciation. It is used to show a strong angry emotion, but with some humour (Meng, 2011).

These features of emotion-loaded Chinese microblog texts might pose challenges for human and machine translation (MT), which are different from translating tweets such as hashtags or non-standard orthography in other languages or other types of texts (Saadany et al., 2023). In this paper, we endeavour to investigate the challenges of translating emotion-loaded Chinese microblog texts for humans and MT, with a special focus on answers from professional translators to the question: What are the challenges for humans and MT in the translation of emotion-loaded texts? We also evaluated outputs of an MT engine and compared the problems identified in MT outputs with the challenges mentioned by professional translators.

In the rest of this paper, Section 2 reviews related work in translation studies and MT studies on emotion. Section 3 starts with a description of the dataset used in this paper and explains the methodology for finding out the challenges from both professional translators' perspective and the output of MT. Section 4 shows the results of interviews with translators and the analysis of MT outputs. Section 5 concludes the paper by summarising our findings and indicating future research directions.

2 Related Work

2.1 Translation Studies on Emotion

Early studies related to the translation of emotion focused mainly on the translation of emotional lexical items. For example, Russell and Sato (1995) compared 14 emotional words such as 'happy', 'sad', or 'angry' in English, Chinese and Japanese to study if they were similar or equivalent in these languages. Choi and Han (2008) raised concerns about 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) questioned the possibility of one-to-one translation of some emotional concepts like 'shame' in English and Spanish. For other language pairs like English and Arabic, Kayyal and Russell (2013) carried out very similar studies, and they found that only one pair (happiness-farah) of emotional words passed their equivalence tests, whereas others somewhat differed in terms of culture and language.

Different from product-oriented translation studies, which focused on the translation of emotional lexica, process-oriented studies paid closer attention to the influence of emotion on translators, which then further affected their translation result. Rojo López and Ramos Caro (2016) carried out an experiment to measure the impact of emotion on translators' performance. They asked students to translate an emotion-loaded text from English to Spanish and gave positive and negative feedback on their translations to different groups. Then, they asked students to translate another text. They found that positive or negative feedback can elicit different processing styles of the text and may affect translation quality, in aspects like accuracy and creativity. Kimovska and Cvetkoski (2021) replicated their experiment in the English-Macedonian pair, and found that positive feedback has a positive impact on creativity and negative feedback has a negative impact on meaning and style. These studies indicate that emotion is an important but difficult phenomenon to deal with during translation.

2.2 MT Studies on Emotion

Most of the research in Natural Language Processing related to emotion focuses on detection and classification of emotions in texts. However, there are some studies investigating the performance of MT systems in preserving sentiment or emotion. Among these studies, Mohammad et al. (2016) examined sentiments in Arabic-English translation of social media texts and evaluated the difference of sentiments before and after translation through human annotation. They found that the change of sentiment was mainly caused by ambiguous words, sarcasm, metaphors, and word-reordering issues. Shalunts et al. (2016) explored the impact of MT on sentiment in German, Russian and Spanish for general news articles. They found that the performance of the sentiment analysis tool on the source and the target was comparable, which means that MT tools do not impact dramatically on the transfer of sentiments. Contrary to their result, Troiano et al. (2020) found that emotions were at least partially lost after back-translation. Similarly, Fukuda and Jin (2022) found that sentiments were distorted by MT tools, with positive sentences tending to stay the same before and after translation, rather than negative and neutral sentences.

Both translation and MT studies on emotion suggest that emotion is difficult to translate and it could be affected by human or machine translation. However, previous studies rarely cover the challenges of emotion translation from the view of professional translators. To the best of our knowledge, no previous study has compared the challenges of emotion translation between human and machine translation in social media texts. This paper intends to contribute to bridging the gap in this area.

3 Data and Methodology

This section introduces the dataset used in the research, and explains the methodology followed to identify challenges when translating emotion-loaded texts.

3.1 Data Description

In order to identify challenges when translating emotion-loaded texts, it is necessary to have access to a dataset that contains numerous emotion expressions. The dataset used in the *Evaluation of Weibo Emotion Classification Technology on the Ninth China National Conference on Social Media Processing* is a good source for our purposes. It was already annotated with six emotion categories, namely, *anger*, *fear*, *joy*, *sadness*, *surprise* and *neutral* (Guo et al., 2021). The dataset had 34,768 Weibo posts, some classified with neutral emotion. We filtered out posts with neutral emotion and randomly sampled 20% of the remaining (about 5500 entries) as the dataset used in this research. Then, we randomly chose 30 entries (each entry is a Weibo post with about 40 Chinese characters), ensuring that in the end we had six entries for each of the five emotion categories, for use in the next stage, which was a translation and interview task.

3.2 Methodology for Studying Challenges for Human Translators

We interviewed professional translators in order to understand the challenges they face when translating emotion-loaded texts. Ten Chinese-English translators with at least one-year professional translation experience[2] were recruited to translate the same 30 selected entries. They were instructed to pay attention to the emotion in the source, and asked to use *PosEdiOn*[3] (Oliver, 2020) to record their actions and translation time during the process. This enabled us to analyse how different their translations were and whether it was difficult for them to translate emotion-loaded texts.

[2] Nine have more than two-year work experience as a translator; one has half-a-year work experience in translating social media texts and half-a-year translation training experience.

[3] An open-sourced software designed for the convenience of translation and post-editing. It is available at https://github.com/aoliverg/PosEdiOn.

These translations were compared by using two types of scores: an n-gram precision score created by BLEU (Papineni et al., 2002); and a cosine similarity score between embeddings obtained from sentence-BERT (Reimers and Gurevych, 2019).

BLEU scores were calculated using the SacreBLEU method (Post, 2018), a variant of BLEU that uses standard tokenisation for the Conference on Machine Translation[4] to improve reproducibility and comparability. To calculate the BLEU scores, one translation from a translator was selected as a reference, and compared with other translations of the same entry from the other translators one by one. This was done iteratively for each translation and each entry, to get 30 matrices of BLEU scores. As BLEU compares the n-grams in the candidate translation with the reference, it focuses more on form and position of words than on meaning. To assess how varied these translations are in terms of meaning, all translations of these 30 entries were embedded into a "semantic" vector space using Sentence-BERT. Cosine similarity was calculated between embeddings of different translations of the same source entry. Then, 30 matrices of similarity scores were obtained using the same process as above. The similarity scores were analysed to see how human translations varied in terms of the "meaning" captured by the embeddings.

After the translation task, one-to-one interviews were conducted online with the translators via Microsoft Teams to ask their opinions about this task. The interviews were semi-structured, and questions used can be found in Appendix B. During the interviews, translators were mainly asked about the challenges of translating emotion-loaded texts and the usefulness of MT for this type of texts. Each interview lasted approximately 30 minutes and was recorded using the Microsoft Teams cloud recording service. The recordings were transcribed by an automatic speech recognition tool, *Whisper* (Radford et al., 2022), and manually checked by the research team. Transcripts of these recordings were imported into an open-sourced software *QualCoder* (Curtain, 2023) for thematic analysis (Braun and Clarke, 2006). All transcripts were first segmented based on interview questions and then coded with different themes[5]. One translator might mention one theme several times for the same question. Similar themes for the same question were merged into broader ones. These themes were extracted and exported for qualitative analyses to identify challenges humans have when translating emotion-loaded texts and problems they think machines might have when translating the same texts. A screenshot of using *QualCoder* can be seen in Figure A.1 in Appendix A.

3.3 Methodology for Studying Problems in MT Outputs

To see the errors present in MT outputs, we used Google Translate[6] to translate the source text of our dataset and recruited two translators to annotate MT errors in terms of emotion preservation, following the framework proposed by Qian et al. (2023). Errors irrelevant to emotions were discarded, as we only focused on translation of emotion. Words or parts of sentences in the source text that relate to the errors were highlighted to analyse the cause for errors in MT.

In order to see how annotators agree with each other or themselves, we randomly sampled 10% (about 550 entries) of the dataset for the inter-annotator agreement check and 100 entries for the intra-annotator agreement check. More details about the framework, error annotation and agreement checks can be found in Qian et al. (2023).

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

[5] We identified some patterned meaning in the transcripts as initial themes or codes, and then refined them during the process.

[6] Results from "https://translate.google.co.uk/" on 30 May 2022.

4 Results and Discussions

This section analyses the data collected in the experiments described in the previous section. Section 4.1 shows results of the analysis of translations and interviews from translators. Section 4.2 summarises results of the analysis of MT outputs and compares similarities and differences of the challenges for translators and MT.

4.1 Analysis of Challenges for Human Translators

4.1.1 Analysis of Translation Data

Before looking into the interview data, we analysed the variations in human translations, by looking at the BLEU and embedding similarity scores.

BLEU Scores Figure A.2 in Appendix A shows 30 heatmaps (one for each entry) of the BLEU score matrix between the 10 translations (excluding comparing with themselves). The numbers on the x and y axes represent different translations (starting from 0). We might expect that, given the same text and enough time to translate, professional translators might produce translations with similar forms. However, most of these matrices are mixed with very light and dark colours, meaning the BLEU scores vary a lot for the different translations of all 30 Weibo posts. Most of these 30 heatmaps are dominated by dark colours—these are associated with low BLEU scores, hence very low similarity between the different translations of each entry.

The most noticeable example is Entry 25, which shows the largest area of black squares. This indicates that the translations of this particular entry are very different among the 10 translators. The source text "淅淅沥沥，沥沥淅淅。也许长大了，胆子就变得小了，想的也愈发多了。内心时常感到惶恐不安，风吹草动，兵荒马乱。心不够沉静，志不够坚强，繁花似锦，稍纵即逝" contains classic Chinese and idiomatic expressions. The first eight characters, bearing the resemblance of rain drops, have been quite often used in literary work to show subtle sad feelings. Many words and phrases such as "风吹草动，兵荒马乱" (showing insecure feelings) should not be interpreted and translated literally. The translators tried to explain them in their own way, and this potentially causes the variation of their translations, as shown in Table A.1 in Appendix A.

Apart from this particular entry, there are other cases in which one version is very different from the other translations of the same entry. Take Translation 7 of Entry 10, for example. BLEU scores between this translation and other translations are very low. This shows that different emotion-loaded texts might influence translators differently, reflected in different levels of variation in the forms chosen by the translators.

Figure A.3 shows the boxplots of these BLEU scores to see the average value (the orange line) and outliers (white circles). Since the matrix of BLEU scores is almost symmetric[7], only the lower half below the diagonal is kept to avoid repetitive plots. Apart from Entries 5, 10, 15 and 29, all entries have outliers. An outlier means that the BLEU score differs significantly and abnormally from others in statistics. Most of the outliers in these entries have abnormally high BLEU scores. This suggests that high agreement among these translations are rare and abnormal. Most of the BLEU scores in each entry have relatively low score values, but a few high-value outliers raise the average level. Entries 5, 10, 15 and 29 are the exceptions, where the emotion-carrying words are straightforward and easy to translate, unlike cultural-loaded words or slang

[7] BLEU score of Translation 1 (T1) as reference and Translation 2 (T2) as candidate can be slightly different (but very similar) from T2 as reference and T1 as candidate, since the distributions of n-grams in T1 and T2 can differ, and their order and frequencies can affect the calculation of precision.

present in the other entries. Take Entry 5 “这个小狗死了没有? 传说疯狗 10 日内必死可怕这个不可大意没事儿少招猫逗狗！有点恐怖......” for example. The emotion-carrying word “恐怖” literally means “horrible” or “scary”, so the translation is less arguable and varied. This can also be seen in Table A.2[8], which shows translations of the source “有点恐怖......”. This is also true for the other three entries, where the emotion is strong and clear, and emotion-carrying words are easy to translate.

Embedding Similarity Scores Figure A.4, Appendix A contains 30 heatmaps (one for each entry) of the similarity score matrix between the embeddings of the 10 versions of translations (including comparing with themselves on the diagonal). The colours of these matrices are much lighter than those in the BLEU matrices, which indicates less variation and more consistency in this form of analysis of the different translations of the same entries. As translation is more about rendering the meaning rather than the sentence form or structure, higher similarity scores between embeddings are expected.

Matrices with the lowest similarity scores are Entries 8, 11, and 14. A quick analysis of the source text of these entries reveals that they all contain emotion-carrying slang words, for which it is difficult to find equivalent expressions in English. For example, in Entry 14, there are multiple slang expressions such as “小乃心”, “小棉袄”, “小傻瓜”, and “小宝宝”, which all mean “sweetheart” and show similar happy feelings, but in slightly different tones. Some of them are quite unique to the Chinese culture. This might be one of the reasons why similarity scores for the embeddings in this entry are low. Another noticeable point which corresponds to the phenomenon observed in BLEU scores is that similarity scores between Translation 7 and other translations are quite low for Entry 10. This means Translation 7 is dissimilar with others in both word forms or sentence structure and “meaning”. This is also the case for Translation 2 and other translations in Entry 10.

Similar to making the boxplots of BLEU scores, the lower half below the diagonal of the similarity matrix is kept for making the boxplots in Figure A.5. Same as the result from the heatmaps, it is easy to see in Figure A.5 that Entries 8 and 14 have low average scores compared to others. Entries with high average scores are also quite noticeable. For example, Entries 20 and 28 have exceptionally high average scores close to 0.9. An investigation of the two entries suggests that both are relatively long compared to others, which means the translators have more context. Another thing they have in common is that most of the text is more informative than emotional. They do contain emotion somewhere in the text, but they mainly describe the event the blogger saw or experienced. This is probably why these translations are more similar in “meaning”, as expressed by the embeddings.

From Figure A.5, it can be seen that Entry 15 has more outliers than the others. An analysis of this entry indicates that there are two cultural-specific words “暧昧” and “转正” in the source text. The second word is also a polysemous word, literally meaning “becoming an official member of”, which now has been bestowed a new meaning “becoming someone’s partner or wife from a mistress” under the modern Chinese culture. Translators may use different translation strategies to translate them in different ways and this may lead to low similarity scores between the embeddings of these translations.

To verify whether emotion-carrying words are the culprit for these variations, Entry 14 was selected since both its average BLEU score and similarity score are low, and it contains many emotion-carrying words. The emotion-carrying words in all translations

[8] Except Translation 7, which is also dissimilar with others in Entry 10.

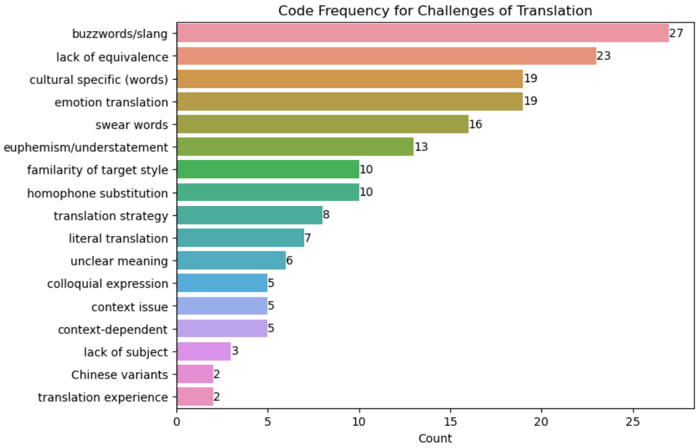


Figure 4.1: Themes Related to What Humans Find Challenging Ordered by Frequency

of Entry 14 were deleted and the BLEU scores[9] between each translation were re-computed to see whether they vary less. Figure A.6 shows the BLEU scores after the deletion of emotion-carrying words in the right boxplot, compared with before deletion in the left. It can be seen that average BLEU scores are improved and there are more high-value outliers, including two extremely high BLEU scores after the deletion of emotion-carrying words. This suggests translators are more likely to agree with each other translating texts without emotion-carrying words.

The analysis above shows that there was ample variation in the translation of Weibo posts by the 10 translators. This suggests that this type of content presents challenges for professional translators, and that emotion-carrying words are one of the sources of such challenges.

4.1.2 Analysis of Interview Data

As described in Section 3.2, interview transcripts were analysed to identify themes mentioned by interviewees in each interview question. The following figures present different themes mentioned in terms of the challenges translators face translating emotion-loaded texts; which aspect in relation to emotion translators find difficult to translate and what problems translators believe MT would have, given emotion-loaded texts.

Figure 4.1 shows the frequency of themes as for the interview question: **Do you think it is difficult to translate emotion-loaded microblog texts? Why?**

The five most frequent themes in the translators' answers are "buzzwords/slang", "lack of equivalence", "cultural specific (words)", "emotion translation" and "swear words". As these Weibo posts are full of buzzwords or slang, and most of them use swear words to show strong emotions, it is clear why these were the top five themes.

"Euphemism/understatement" in the figure is also related to "buzzwords/slang", as Chinese netizens use euphemism or understatement, instead of explicitly using swear words, to make their language more polite and less offensive, but at the same time keeping the same strong emotion. Understatement of emotions on public venues is, as described by interviewees, commonly seen under *"East Asian culture"*, where *"people*

[9] Only BLEU scores were re-computed, not cosine similarity scores, because the deletion of emotion-carrying words made this entry no longer semantically meaningful.

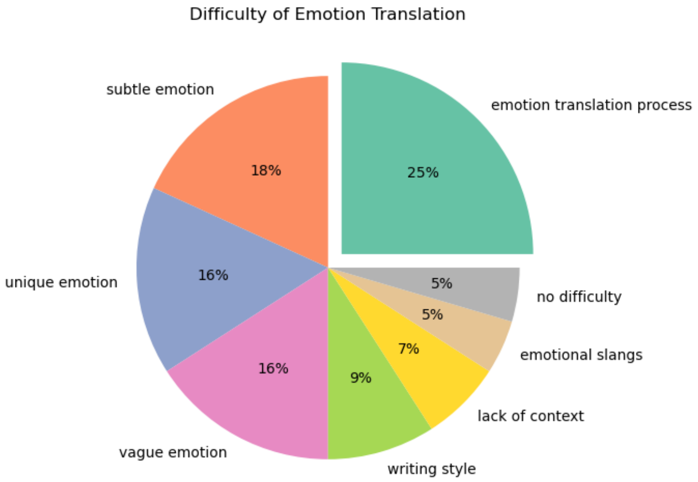


Figure 4.2: Theme Frequency for the Difficulty of Translating Emotions

express feelings in a cute and reserved way". One popular way to create slang expressions or euphemisms in Chinese is "homophone substitution" (King et al., 2013), which is also mentioned as a challenge for translation. Another way to create slang expressions is to use variants of Chinese such as some local dialects to achieve a humorous effect.

However, the use of "buzzwords/slang" may lead to other problems which translators see as challenges as well. "Unclear meaning" and "context issue" are partially caused by the overuse of slang. Some translators indicated that the meaning of some slang expressions are *"context-dependent"*, and that sometimes a blogger uses slang just to show a certain emotion, since he/she *"does not even know what he or she means by saying it"*. This poses some challenges for readers to get the real intention/meaning behind it. Another big challenge for the translation of these texts is "lack of equivalence", because many of these emotion-carrying "buzzwords/slang" are cultural specific. There is no exact equivalent expression in the target context due to cultural differences. Other factors such as "familiarity of the target style", "translation experience" in social media texts, "colloquial expression" and "lack of subject" in the source may also pose challenges for the human translation of emotion-loaded texts.

Two themes worth noticing here are "translation strategy" and "literal translation", which are challenges associated with the choice of the best translation strategy for this type of content. Since cultural specific words are quite common in these texts, translators find it challenging to decide whether to use foreignization, a strategy that is more prone to render the literal meaning, or domestication, which tends to modify the translated text according to the target style and culture. As some of the translators described in the interview, they had to choose *"whether to localise or to keep the features of source"* in the target text.

Figure 4.2 displays the theme frequencies in percentage for the interview question: **Do you think it is difficult to translate the emotion in the source text? Why?**

Excluding the theme related to the impact of emotion in the translation process, the most frequent theme for the difficulty of emotion transferring is "subtle emotion",

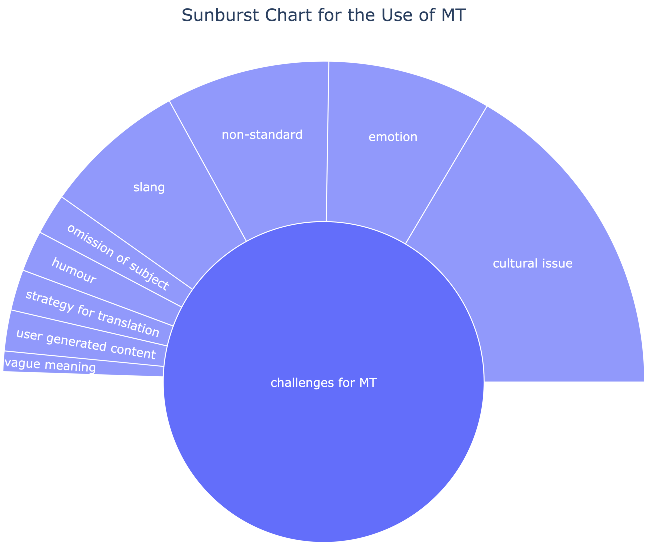


Figure 4.3: Themes Related to Problems for MT in Emotion-Loaded Texts

and then "unique emotion" and "vague emotion". Take Entry 14 for example. The four emotion-carrying words "小乃心", "小棉袄", "小傻瓜" and "小宝宝", meaning "my sweetie", "my warm jacket", "my silly goose" and "my little baby" respectively, are very subtle in expressing emotions. These expressions cause difficulties to translators. Also as discussed in previous literature, some emotions such as "疼", the "heart-aching love" is quite unique to Chinese culture (Sundararajan, 2015). Slang created via homophone substitution is also specific to the Chinese language and culture and very subtle in terms of emotion conveying. These unique and sometimes context dependent emotions raise difficulties for human translation. Likewise, vague or subtle emotions and short context might also pose challenges for understanding and translation, since *"the length of these texts is short"* and some Weibo users accidentally or *"deliberately show their emotion in a vague or subtle way"*. Other frequent themes such as "writing style", "lack of context" and "emotional slang" are also mentioned as difficulties of emotion translation. Only two translators mentioned it is not difficult for them to transfer the emotion as they suggested that they can always find similar expressions of emotion in the target text, without thinking too much about the cultural differences. They mentioned that there are always differences between languages and that translation is a tool for bridging differences, not to eliminate them.

Another frequent theme is "emotion translation process". During the interview, most translators expressed that they were not affected by the strong emotion of the source text, although they felt the same feeling as the source during translation. They remained neutral in the translation process. This does not follow the results of some previous studies, such as Rojo López and Ramos Caro (2016) and Kimovska and Cvetkoski (2021), where emotions affected the translation result.

Figure 4.3 shows themes related to the problems that translators believe MT systems have, when they were asked: **Do you think whether MT tools will be useful for the translation of emotion-loaded social media texts?** The size of segments in the outer layer indicates the frequency of these themes.

The most frequent theme for problems for MT, from the human translators' perspective, are "cultural issue", "emotion" and "non-standard" writing. Translators thought some cultural-specific words are very difficult for MT and end up being translated in a literal way, which severely hinders readers understanding of the emotion of the source. They also stressed that *"machines do not understand human emotions"*, and that unique and subtle emotions in the source are challenging for MT. Different from the challenges for human translators, they think that "non-standard" writing, including the omission of punctuation and subject in some Weibo posts, are difficult for MT, while relatively easier for human translators. Humans can infer the meaning from its context, even if the source does not have punctuation or subject, while machines might not be able to produce good solutions for this. Translators also mentioned that humans might feel relaxed translating the informality as the source is written in a causal way, but machines may have problems in processing this type of content. Another point worth noticing is that one translator mentioned some of the "user generated content" is very *"creative"* in term of the choice of words and writing style. Some of them even use "humour" or other rhetorical devices to convey the meaning or emotion. This also makes the source content challenging for MT.

4.2 Relation between Challenges for Human Translators and Errors in MT Outputs

To study the problems present in MT outputs and test the assumptions of translators regarding the most common sources of MT errors, a task of human evaluation of MT outputs was carried out as described in Section 3.3. This task along with its following results is described in Qian et al. (2023). The results show that besides emotion-carrying words, polysemous words, punctuation, negation, subject/object issues, subjunctive mood and abbreviation are also causes of errors in MT.

In the list of challenges for human translators when translating emotion-loaded texts, and the list of causes of errors in MT outputs of the same type of content, there are two common elements: emotion-carrying words, and subject/object issues.

The differences in these lists of challenges for translators and factors for MT errors can be summarised as follows: 1) MT does not solve correctly issues such as non-standard writing or non-standard use of punctuation; these issues do not pose challenges to human translators, since they can infer meaning from the context; 2) some challenges that humans believe might make themselves and MT struggle, such as the choice of translation strategies, cannot be observed in the MT output; 3) other challenges, such as the translator's familiarity with the target style, are concerns expressed by human translators for themselves, not for the MT. Again, the analysis of MT errors does not give access to such information.

5 Conclusion

Our paper investigates challenges professional translators face when translating emotion-loaded texts, as well as errors MT makes when translating this content. We interviewed 10 professional translators about the challenges they meet when translating this type of texts and the challenges they expect MT to have. We compared results from the interviews with the errors we identified in real MT outputs. We establish that the most challenging problem for both human and MT is emotion-carrying words. We also find that what is more challenging for MT, such as non-standard writing, is not necessarily an issue for human translators. In future work, we plan to explore how these findings can be used to train human translators and improve automatic translation of emotions.

References

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3:77–101.

Choi, S. and Han, G. (2008). SHIMCHEONG PSYCHOLOGY: A CASE OF AN EMOTIONAL STATE FOR CULTURAL PSYCHOLOGY. *International Journal for Dialogical Science Copyright*, 3:205–224.

Curtain, C. (2023). Qualcoder. `https://github.com/ccbogel/QualCoder/releases/tag/3.2`. Last checked on Jun 22, 2023.

Fukuda, K. and Jin, Q. (2022). Analyzing Change on Emotion Scores of Tweets Before and After Machine Translation. In Meiselwitz, G., editor, *Social Computing and Social Media: Design, User Experience and Impact*, volume 13315, pages 294–308. Springer.

Guo, X., Lai, H., Xiang, Y., Yu, Z., and Huang, Y. (2021). Emotion Classification of COVID-19 Chinese Microblogs based on the Emotion Category Description. In *Proceedings of the 20th China National Conference on Computational Linguistics*, pages 916–927. Chinese Information Processing Society of China.

Hurtado de Mendoza, A., Fernández-Dols, J. M., Parrott, W. G., and Carrera, P. (2010). Emotion terms, category structure, and the problem of translation: The case of shame and vergüenza. *Cognition and Emotion*, 24:661–680.

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

Kimovska, S. K. and Cvetkoski, V. (2021). THE EFFECT OF EMOTIONS ON TRANSLATION PERFORMANCE. *Research in Language*, 19:169–186.

King, G., Pan, J., and Roberts, M. E. (2013). How censorship in China allows government criticism but silences collective expression. *American Political Science Review*, 107:326–343.

Meng, B. (2011). From Steamed Bun to Grass Mud Horse: E Gao as alternative political discourse on the Chinese Internet. *Global Media and Communication*, 7:33–51.

Mohammad, S. M., Salameh, M., and Kiritchenko, S. (2016). How Translation Alters Sentiment. *Journal of Artificial Intelligence Research*, 55:95–130.

Oliver, A. (2020). MTUOC: easy and free integration of NMT systems in professional translation environments. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 467–468, Lisboa, Portugal. European Association for Machine Translation.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Post, M. (2018). A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Qian, S., Orăsan, C., Do Carmo, F., Li, Q., and Kanojia, D. (2023). Evaluation of Chinese-English Machine Translation of Emotion-Loaded Microblog Texts: A Human Annotated Dataset for the Quality Assessment of Emotion Translation. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv preprint*.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rojo López, A. and Ramos Caro, M. (2016). Can emotion stir translation skill? Defining the impact of positive and negative emotions on translation performance. pages 107–130. John Benjamins Publishing Company.

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

Saadany, H., Orasan, C., Do Carmo, F., Zilio, L., and Quintana, R. C. (2023). Analysing Mistranslation of Emotions in Multilingual Tweets by Online MT Tools. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.

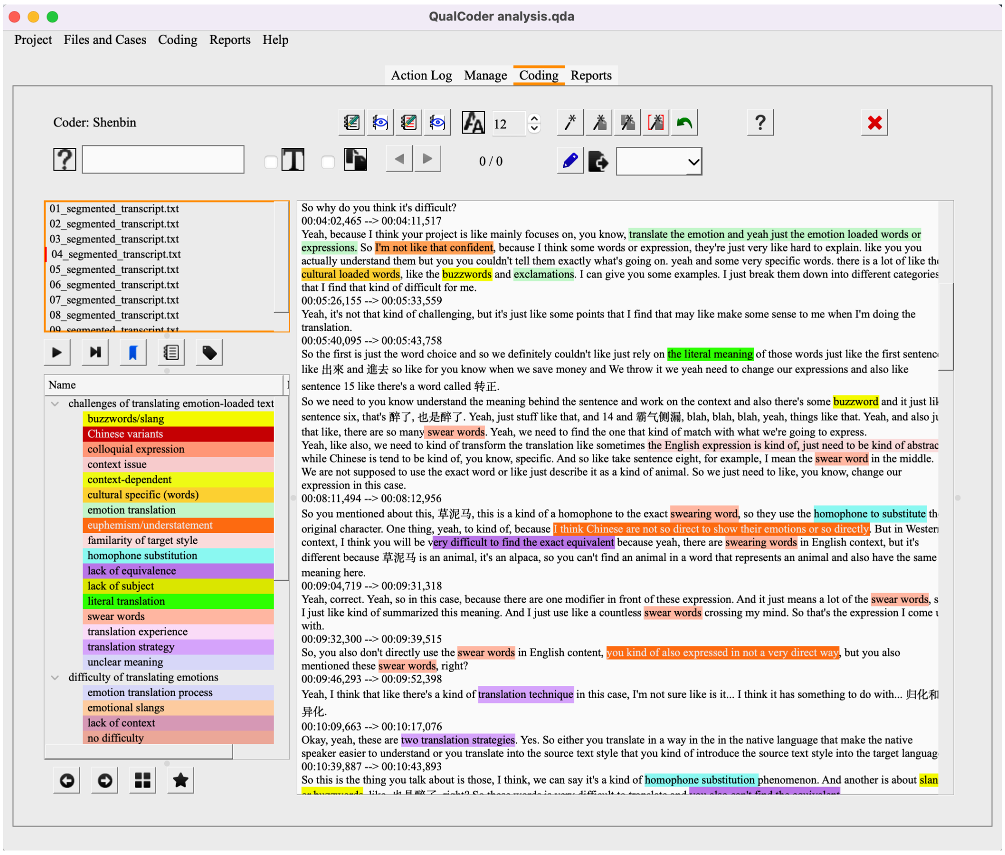
Shalunts, G., Backfried, G., and Commeignes, N. (2016). The Impact of Machine Translation on Sentiment Analysis. In *The Fifth International Conference on Data Analytics*, pages 51–56. IARIA.

Sundararajan, L. (2015). *Understanding Emotion in Chinese Culture: Thinking Through Psychology*. Springer.

Troiano, E., Klinger, R., and Padó, S. (2020). Lost in Back-Translation: Emotion Preservation in Neural Machine Translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4340–4354. International Committee on Computational Linguistics.

Appendices

A Figures and Tables



Coder: Shenbin

01_segmented_transcript.txt
02_segmented_transcript.txt
03_segmented_transcript.txt
04_segmented_transcript.txt
05_segmented_transcript.txt
06_segmented_transcript.txt
07_segmented_transcript.txt
08_segmented_transcript.txt
09_segmented_transcript.txt

- challenges of translating emotion-loaded text
  - buzzwords/slang
  - Chinese variants
  - colloquial expression
  - context issue
  - context-dependent
  - cultural specific (words)
  - emotion translation
  - euphemism/understatement
  - familarity of target style
  - homophone substitution
  - lack of equivalence
  - lack of subject
  - literal translation
  - swear words
  - translation experience
  - translation strategy
  - unclear meaning
- difficulty of translating emotions
  - emotion translation process
  - emotional slangs
  - lack of context
  - no difficulty

So why do you think it's difficult?
00:04:02,465 --> 00:04:11,517
Yeah, because I think your project is like mainly focuses on, you know, translate the emotion and yeah just the emotion loaded words or
expressions. So I'm not like that confident, because I think some words or expression, they're just very like hard to explain. like you you
actually understand them but you you couldn't tell them exactly what's going on. yeah and some very specific words. there is a lot of like th
cultural loaded words, like the buzzwords and exclamations. I can give you some examples. I just break them down into different categories
that I find that kind of difficult for me.
00:05:26,155 --> 00:05:33,559
Yeah, it's not that kind of challenging, but it's just like some points that I find that may like make some sense to me when I'm doing the
translation.
00:05:40,095 --> 00:05:43,758
So the first is just the word choice and so we definitely couldn't like just rely on the literal meaning of those words just like the first sentenc
like 出來 and 進去 so like for you know when we save money and We throw it we yeah need to change our expressions and also like
sentence 15 like there's a word called 转正.
So we need to you know understand the meaning behind the sentence and work on the context and also there's some buzzword and it just lik
sentence six, that's 醉了, 也是醉了. Yeah, just stuff like that, and 14 and 霸气侧漏, blah, blah, blah, yeah, things like that. Yeah, and also ju
that like, there are so many swear words. Yeah, we need to find the one that kind of match with what we're going to express.
Yeah, like also, we need to kind of transform the translation like sometimes the English expression is kind of, just need to be kind of abstrac
while Chinese is tend to be kind of, you know, specific. And so like take sentence eight, for example, I mean the swear word in the middle.
We are not supposed to use the exact word or like just describe it as a kind of animal. So we just need to like, you know, change our
expression in this case.
00:08:11,494 --> 00:08:12,956
So you mentioned about this, 草泥马, this is a kind of a homophone to the exact swearing word, so they use the homophone to substitute th
original character. One thing, yeah, to kind of, because I think Chinese are not so direct to show their emotions or so directly. But in Wester
context, I think you will be very difficult to find the exact equivalent because yeah, there are swearing words in English context, but it's
different because 草泥马 is an animal, it's an alpaca, so you can't find an animal in a word that represents an animal and also have the same
meaning here.
00:09:04,719 --> 00:09:31,318
Yeah, correct. Yeah, so in this case, because there are one modifier in front of these expression. And it just means a lot of the swear words, s
I just like kind of summarized this meaning. And I just use like a countless swear words crossing my mind. So that's the expression I come u
with.
00:09:32,300 --> 00:09:39,515
So, you also don't directly use the swear words in English content, you kind of also expressed in not a very direct way, but you also
mentioned these swear words, right?
00:09:46,293 --> 00:09:52,398
Yeah, I think that like there's a kind of translation technique in this case, I'm not sure like is it... I think it has something to do with... 归化和
异化.
00:10:09,663 --> 00:10:17,076
Okay, yeah, these are two translation strategies. Yes. So either you translate in a way in the in the native language that make the native
speaker easier to understand or you translate into the source text style that you kind of introduce the source text style into the target languag
00:10:39,887 --> 00:10:43,893
So this is the thing you talk about is those, I think, we can say it's a kind of homophone substitution phenomenon. And another is about slan
or buzzwords, like, 也是醉了, right? So these words is very difficult to translate and you also can't find the equivalent

Figure A.1: Screenshot of Using *QualCoder*

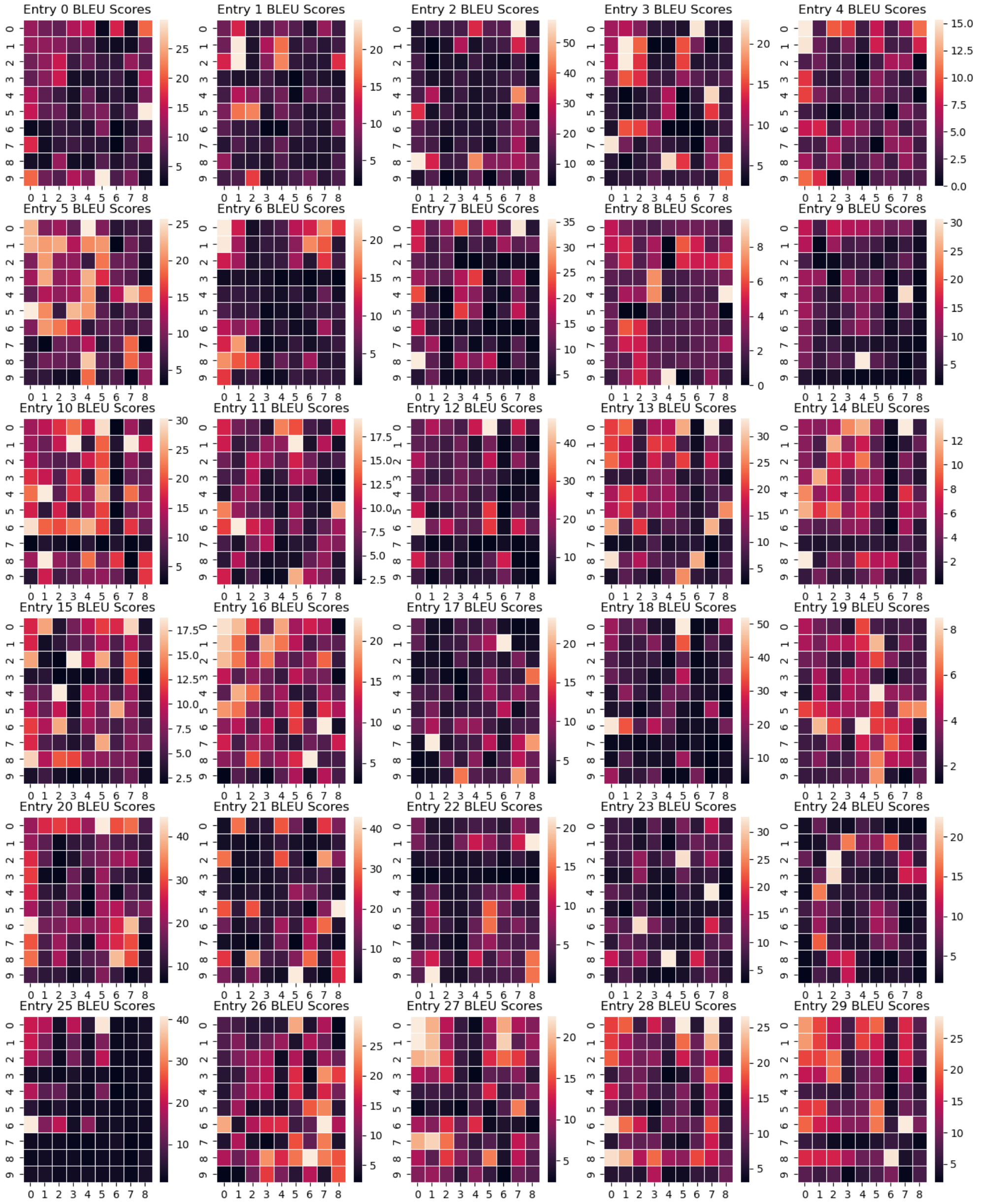


Figure A.2: Heatmaps for BLEU Score Matrices

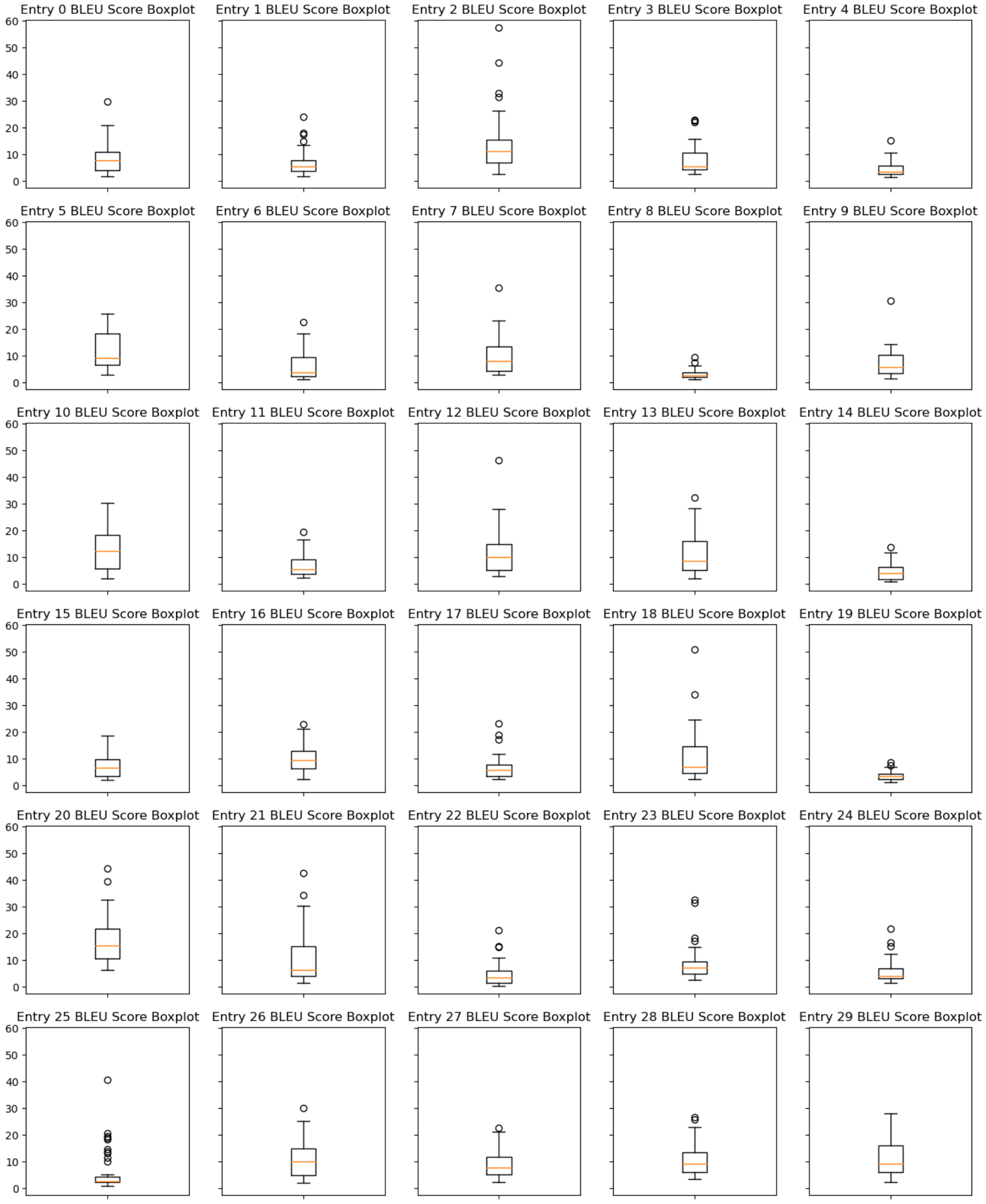
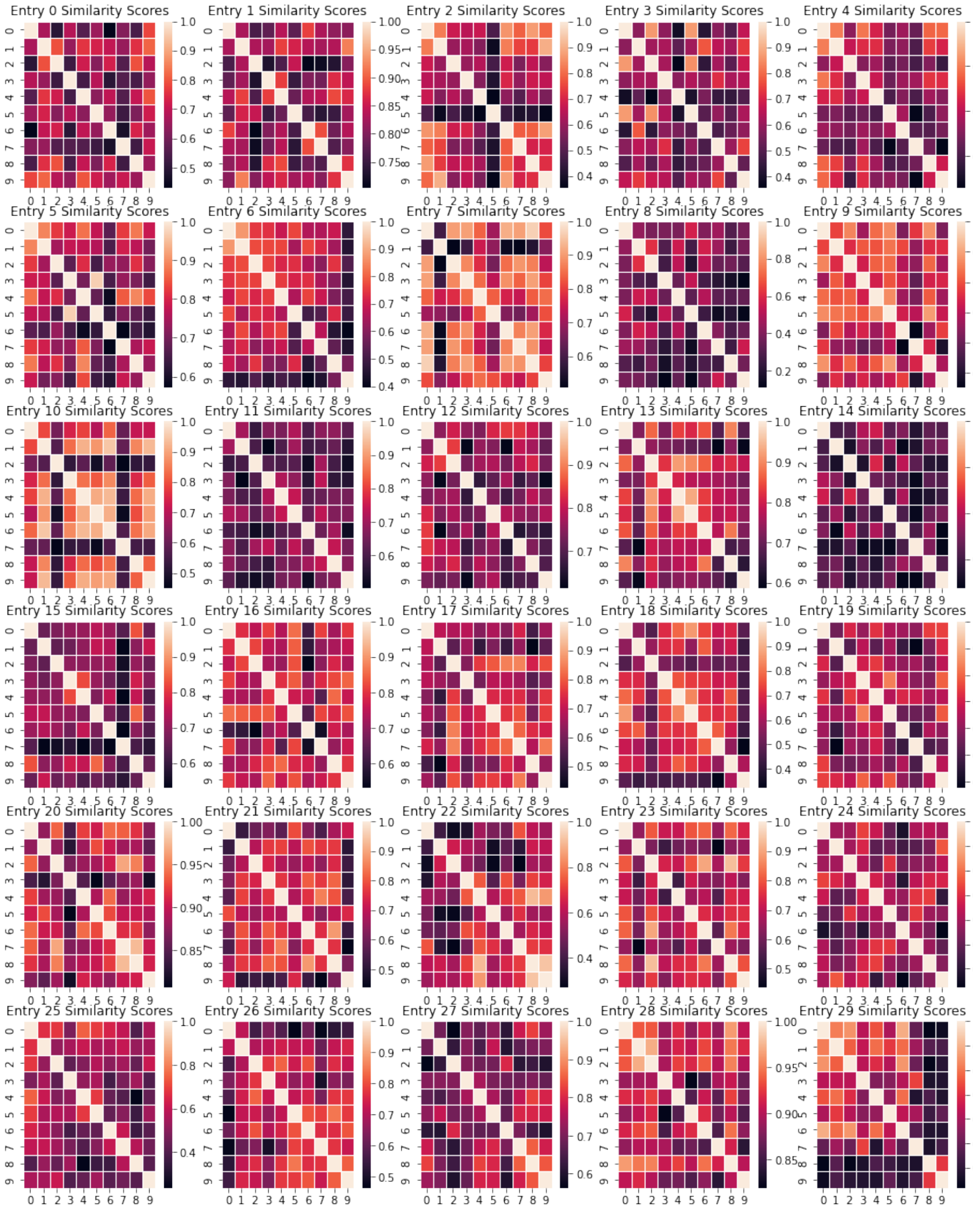


Figure A.3: Boxplots for BLEU Scores



Entry 0 Similarity Scores

Entry 1 Similarity Scores

Entry 2 Similarity Scores

Entry 3 Similarity Scores

Entry 4 Similarity Scores

Entry 5 Similarity Scores

Entry 6 Similarity Scores

Entry 7 Similarity Scores

Entry 8 Similarity Scores

Entry 9 Similarity Scores

Entry 10 Similarity Scores

Entry 11 Similarity Scores

Entry 12 Similarity Scores

Entry 13 Similarity Scores

Entry 14 Similarity Scores

Entry 15 Similarity Scores

Entry 16 Similarity Scores

Entry 17 Similarity Scores

Entry 18 Similarity Scores

Entry 19 Similarity Scores

Entry 20 Similarity Scores

Entry 21 Similarity Scores

Entry 22 Similarity Scores

Entry 23 Similarity Scores

Entry 24 Similarity Scores

Entry 25 Similarity Scores

Entry 26 Similarity Scores

Entry 27 Similarity Scores

Entry 28 Similarity Scores

Entry 29 Similarity Scores

Figure A.4: Heatmaps for Embedding Similarity Score Matrices

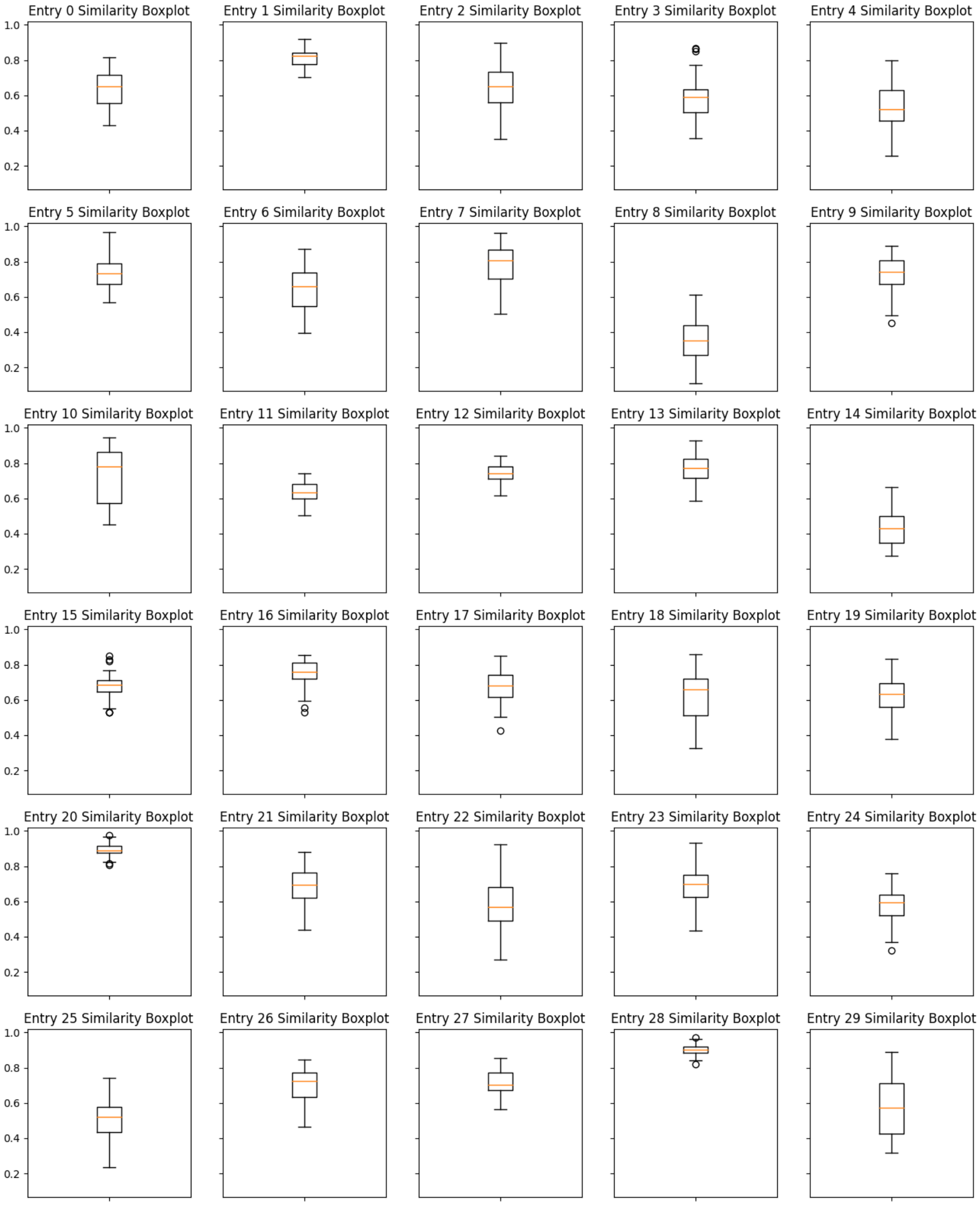


Figure A.5: Boxplots for Embedding Similarity Scores

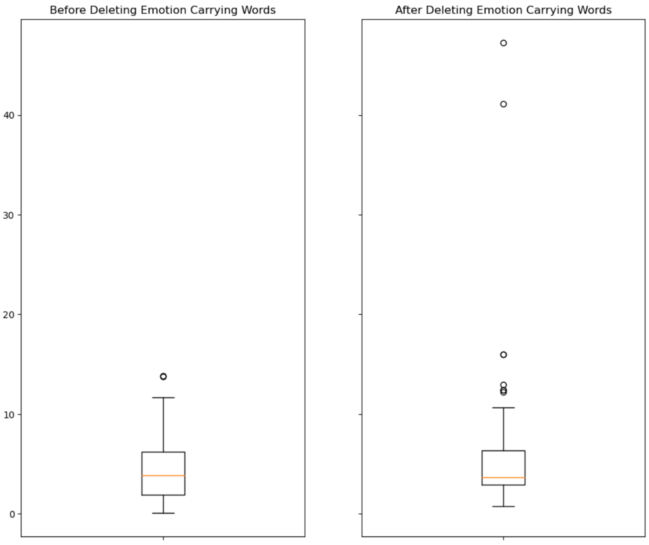


Figure A.6: Entry 14 Boxplots of BLEU Scores **Before (Left)** and **After (Right)** Deletion of Emotion Carrying Words

No.	**Translations**
0	Drizzling, drizzling, drizzling. Maybe when I grow up, I become less courageous and think more. I often feel panic and uneasiness in my heart, turmoil and chaos. My heart is not calm enough, my ambition is not strong enough. Flowers are in blossom now but will die out every soon.
1	It's drizzling. Maybe people will actually be less brave and have more thoughts when they are getting older. In my mind, I am always worried and being sensitive, just like a nervous wreck. My heart is not calm enough and my spirit is not strong enough. There is no much time.
2	Pattering. Pattering. Maybe I became less bold as I grow older and think more often. I often feel unrest, easy picked and panic. My heart is not calm, my will is not strong. Flourishing flowers dies in a blink of an eye.
3	There is breeze and drizzle, the leaves falling. Maybe when people grow older, they tend to be more intimidated and there is more on their mind. They are easily got fidgety and frightened out of anything unexpected. We may likely to lose those wonderful things we are having in the peaceful state instantly if our mind is in chaos and easily destructive.
4	It's been drizzling for a while. I turn to be more cowardly and think more after I grow up. I often feel panic and anxiety. A little breeze could blow the grass in my heart hard. My heart is in chaos of wars. My heart is not calm enough. My life goal is not clear enough. Though my inside is like a picture of thousands of flowers, it vanishes in a second.
5	It's raining outside. Maybe I lose my guts growing up and I have too many thoughts, and often feel insecure. A small thing in the outside world has an impact in my innner world. I don't have inner peace or strong will. Flowers are blooming, but soon they will be gone.
6	Gradually, time gose by. Maybe when I grow up, I become less daring and think more. I often feel panic and very sensitive. The heart is not calm enough, and the will is not strong enough, such as blooming flowers, transiently.
7	As I get older, I find myself thinking more and accomplishing less. Changes around me frequently astounded me. I wasn't calm or strong enough to deal with these sudden changes.
8	Pitter patter, pitter patter. Maybe being a grown-up makes us more timid. We have loads of ideas, making us afraid and confused. With just a sign of disturbance and trouble, we find it hard to calm down and concentrate. We live like a flash in the pan.
9	Patter, patter. Maybe it's because I grew older, I became not that bold and am always careful with lots of thoughts in my mind. Always feel worried. Any sign could mess my mind. Not calm enough. Neither determined. Shiny as golden hours are, they don't last and very soon pass.

Table A.1: Translations of Entry 25

No.	**Translations**
0	It’s a little scary...
1	It’s scary......
2	Bit scary......
3	It’s a bit intimidating...
4	This is a bit spooky.
5	Horrible...
6	It’s scary.....
7	I am not joking.
8	It’ s a bit spooky...
9	This is kind of scaring...

Table A.2: Translations of the Emotion-Loaded Part of Entry 5

B Interview Questions[10]

1. Could you tell me about your experience in translation, specifically in translating this type of Weibo posts?
2. Do you think it is difficult to translate this type of emotion-loaded Weibo posts? Why?
3. Are there elements in the Weibo posts that are more difficult to translate? Why?
4. Look at the example in the chat, do you think it is tricky to translate? Why? 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.
5. How long did it take to translate all the texts?
6. Do you think it is difficult to translate the emotion in the source text? Why?
7. Do you think the strong emotions or those emotional words in the source text makes you concerned more about the overall translation quality? If yes, how might this affect your quality?
8. Do you think whether MT tools will be useful for the translation of emotion-loaded social media texts?
9. Have you tried, as a user, not a translator, the Translate option on social media applications such as Twitter or WeChat?
10. Do you think whether it will be difficult for MT to translate the following sentence in the chat? If yes, which part? 嘤嘤嘤翻牌了开森受宠若惊晚安么么哒.
11. Can you guess the original emotion of the source by looking at the following MT result in the chat? *Tell a woman that she will hurt me for the rest of my life.*

[10] All questions in the interview are listed here, but due to the length of this paper, only questions most relevant to the theme are included in Section 4.1.