The MAKE-NMTViz System Description for the WMT23 Literary Task

Fabien Lopez[1] and **Gabriela Gonzalez-Saez**[1] and **Damien Hansen**[1 6] and **Mariam Nakhle**[1 5] and **Behnoosh Namdarzadeh**[3] and **Marco Dinarelli**[1] and **Emmanuelle Esperança-Rodier**[1] and **Sui He**[4] and **Sadaf Mohseni**[3] and **Caroline Rossi**[2] and **Didier Schwab**[1] and **Jun Yang**[4] and **Jean-Baptiste Yunès**[3] and **Lichao Zhu**[3] and **Nicolas Ballier**[3]

1 Univ. Grenoble Alpes,
CNRS, Grenoble INP, LIG
38000 Grenoble, France
2 Université Grenoble Alpes
3 Université Paris Cité
4 Swansea University
5 Lingua Custodia, France
6 Université de Liège, CIRTI, 4020 Liège, Belgique
Contact: fabien.lopez@univ-grenoble-alpes.fr

Abstract

This paper describes the MAKE-NMT-Viz's submission to the WMT 2023 Literary task. As a primary submission, we fine-tune the mBART50 model using Train, Valid1, and Test1 as part of the GuoFeng corpus (Wang et al., 2023b).We followed similar training parameters to Lee et al. (2022) when fine-tuning mBART50. For our contrastive1 submission, we used a context-aware NMT system based on the concatenation method (Lupo et al., 2022). The training was performed in two steps: (i) a traditional sentence-level transformer (Vaswani et al., 2017) was trained for 10 epochs using GeneralData, Test2, and Valid2; (ii) second, we fine-tuned such Transformer using document-level data, with 3-sentence concatenation as context, for 4 epochs using Train, Test1, and Valid1 data. We then compared the three translation outputs from an interdisciplinary perspective, investigating some of the effects of sentence- vs. document-based training. Computer scientists, translators and corpus linguists discussed the remaining linguistic issues for this discourse-level literary translation.

1 Introduction

In order to analyse literary translations, we have gathered an interdisciplinary team of translators, linguists and computational scientists. We used this opportunity to explore neural machine translation of literary texts as a test set for test suites and unsolved issues for MMT literary translations, especially for the Chinese-English language pair. While the topic of literary machine translation has gained momentum in the last years, there have still been few attempts to customize systems to literary data, although this idea is also drawing attention (Kenny and Winters, forthcoming). Indeed, research has been carried out on this subject, notably on Catalan (Toral and Way, 2018), but also on Slovenian (Kuzman et al., 2019), German and Russian (Matusov, 2019), and on French (Besacier and Schwartz, 2015), where research suggests that MT systems can be further fine-tuned on specific genres and individual translator styles (Hansen and Esperança-Rodier, 2023).

Of course, these very attempts bring about many issues concerning textual ownership, copyright, translator status and livelihood, possibly lowered quality, cognitive friction, etc. (Taivalkoski-Shilov, 2019). It is therefore important to include these ethical aspects into the research and clarify its objectives: for instance, whether MT should serve as a reading aid (Oliver González, 2017), or as a post-editing tool that may increase the effort needed to translate (Kolb, 2020) and constrain creativity (Guerberof-Arenas and Toral, 2022).

Part research has also focused on evaluating the use of existing tools for literary texts. In the context of Chinese to English, attention has been paid to some of the specific shortcomings of MT systems, such as the translation of adjectival possessive pronouns (Jiang and Yu, 2017), or theme-rheme progressions (Jiang and Niu, 2022). Such limitations can indeed have a drastic impact on readers' acceptance, which Shih (2016) explores in the context of online folktales, confirming that the text's function plays a large role in this respect.

Lastly, Thai et al. (2022) have also pointed the incompatibility of MT metrics, document-level or otherwise, for literary texts, concluding that "hu-

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 287–295
December 6–7, 2023.

man expert evaluation is currently the only way to judge the quality of literary MT".

The rest of the paper is organised as follows: Section 2 details our approaches to the task and the training data of our experiments, Section 3 presents the results and Section 4 discusses them.

2 Data and Tools Used

This section details the toolkits we used and our training data for the three submissions authorised for the task. We first used part of the training data proposed by the organisers (Wang et al., 2023a) to observe the translations from mBART50 from Chinese into English before fine-tuning mBART (primary submission). We then used a fine-tuned context-aware concatenation-based Transformer trained at document level (contrastive1 submission) and a traditional sentence-level Transformer (contrastive2 submission).

2.1 Primary model: mBART50 fine-tuning

As a primary submission, we used GuoFeng corpus (Wang et al., 2023a) to fine-tune the mBART50 model with Chinese-English data, using the Train set for training, Test1 as test set, and Valid1 as validation set. We followed similar training parameters to (Lee et al., 2022) when fine-tuning mBART50. As Lee et al. (2022), we trained for 3 epochs, using gelu as an activation function, with a learning rate of 0.05, dropout of 0.1 and a batch size of 16 (we parallelised two A100 GPUs with batch size 8 per device). We decoded using a beam search of size 5.

2.2 Contrastive models

We submit two contrastive models, the first is a context-aware model (*contrastive1*) built on the second system, a sentence-level model (*contrastive2*).

For our contrastive1 submission, we used a context-aware NMT system based on the concatenation method (Lupo et al., 2023). The training was performed in two steps: (i) a sentence-level transformer (Vaswani et al., 2017) was trained for 10 epochs[1] using General Data as train set, Test2 as test set, and Valid2 as validation set ; (ii) second, we fine-tuned at document-level using 3-sentence concatenation for 4 epochs[2] using Train as train set, Valid1 as validation set and Test1 as test set. During the fine-tuning, we used ReLU as an activation function, with an inverse square root learning rate decay, dropout of 0.1, and a batch size of 64. We decoded using a beam search of size 4. For our contrastive2, we used the model trained at step (i) (sentence-level). The training parameters were an inverse square root learning rate decay, a dropout of 0.1, and a batch size of 64. We decoded using a beam search of size 4.

[1]We used only 10 epochs because of time constraints

[2]We used only 4 epochs because of time constraints

2.3 Evaluation Metrics

To evaluate our models, we use the BLEU score metric (Papineni et al., 2002) as implemented in the Moses package.

We performed a human annotation of errors in the translation obtained by our primary submission. 109 segments were selected and annotated by three evaluators that are Chinese native speakers. To measure the inter-annotator agreement, we used Fleiss' kappa (Fleiss et al., 1971). The score is calculated to measure the inter-rater reliability of the annotations as the following equation

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where $P_o - P_e$ measures the real concordance of annotations that are not achieved above chance, while $1 - P_e$ measures the achievable concordance of annotations above chance. In our case, we computed errors by type as well as error types by segment (6 types and 109 segments, cf. 4.3.2).

3 Experiments and Results

We provide a human analysis of the primary model by discussing the improvements observed with the mBART fine-tuning with respect to the baseline. Additionally, we report the BLEU scores of our three systems.

3.1 Baseline of primary model: mBART50

During the training phase of the competition, with the standard HuggingFace implementation of mBART50, we observed the following issues when we translated Test1 from Chinese to English, which was part of the data provided for training by the organisers:

- hallucinations
- discrepancy between the Chinese input and the English translations
- tense concord
- co-referentiality issues for pronouns

Most textual discrepancies between the sizes of the sentences in the two languages were fixed by the fine-tuning as well as hallucinations and Chinese characters in the English translations. We nevertheless noticed a certain number of Chinese characters in the mBART50 translations, which decreased after our fine-tuning, and we only found 18 examples for all the 16,742 sentences, mostly for the fantasy genre, when referring to named entities or specific attributes of the universe (*Skills: Blade Technique, Wing Protection*).

3.2 Fine-tuning with Literary Data

In this section, we analyse the outputs qualitatively. This analysis consists of an initial description of the baseline and fine-tuned outputs, followed by a deeper examination of the syntactic and semantic functions of the produced outputs by both models.

Instances of hallucinations were observed in the outputs of our baseline model. The hallucinated elements are present in the source text, so they are not elements which are not present in the source text. According to Lee et al. (2018), hallucinations can be defined as the model producing a vastly different and inadequate output when the source is perturbed under a specific noise model. Thus, we may suggest that there exist other instances where the model ceases translation of the source text and proceeds with generating output punctuated solely by a sequence of continuous commas (,,,,,,,,,), which may represent an alternative manifestation of hallucination. Interestingly, it is noteworthy that the fine-tuned outputs did not exhibit any instances of hallucination. However, it should be mentioned that few Chinese tokens were observed in the fine-tuned outputs. In the Chinese source text, the equivalent of the word "businessmen" is placed at the left periphery of the sentence, having a pragmatic effect that involves topic introduction or re-introduction, based on the context. Both the baseline and fine-tuned models take the left dislocated element to the right periphery of the sentence, thereby inducing an alternation in the sentence's intended meaning. As we observed, the baseline models chunk the sentences and use commas instead of employing coordinations, relative clauses, or more complex structures. In this example, the baseline model produces "Ten minutes later. consciousness is exhausted. scattered" by separating each chunk or even token with a period. In contrast, the fine-tuned model generates "Ten minutes later, his consciousness was exhausted and dissipated.", using coordination to form a united sentence. This represents another instance of the fine-tuned model's proficient manipulation of structures, wherein it employs a relative clause "which" to interconnect the sentences. Ex: "Wang lived in the 413 bedrooms of the West school district, Lins lived in the 413 bedrooms of the East school district." Fine-tuned: 09primary: "Wang Yicheng stayed in 413, which was in the West campus. Lin Sisi stayed in 413, which was in the East campus." Furthermore, the choice of tense seems to be different in the two models: As for the fine-tuned model, a preference for the past tense becomes evident. Conversely, as for the baseline model, an over-use of the present tense is observed in its outputs. We may also add that baseline models tend to favour the indicative mood, which indicates assertion, as seen in an example like "What's wrong with the game?". On the other hand, fine-tuned models have been *trained* to produce sentences in moods that exhibit a reduced level of assertiveness, as evidenced by constructions like "Could there be a problem with the game?".

3.3 BLEU scores

In this section, we report the results of our primary, constrastive1, and contrastive2 in terms of BLEU score computed using Test1 and Test2 datasets at the end of the full training process of each model. The official results of the competition on test3 were not computed as the reference translations were not provided (at the time of writing this article).

Model	Test1	Test2
primary	22.31	–
contrastive1 (document-level)	19.03	17.58
contrastive2 (sentence-level)	22.31	18.22

Table 1: BLEU score for primary, contrastive1 and contrastive2 systems.

Table 1 shows that our primary system achieves the same BLEU score as contrastive2[3], the sentence-level transformer implementation. We notice that the document-level system (contrastive1) is not better than the sentence-level model. This

[3]Primary and contrastive2 scores on Test1 are identical due to coincidence.

might be explained by the few epochs used for training.

4 Discussion

4.1 Lexical Complexity

To appreciate the relative complexity of the terms used in the translations we first qualitatively compared the translations and contrastive2 seemed to be more elaborate, so we tested this impression with more quantitative means. We investigated the vocabulary growth curves of the three translations using the functions available from the languageR package (Baayen and Shafaei-Bajestan, 2019) to find out that the number of different types progress on the same rhythm for the different translations. In this type of representation, the horizontal axis corresponds to the expansion of the translation corpus (number of tokens) and the vertical axis corresponds to the number of types. The first lower series of curves corresponds to the number of hapaxes. As can be seen in Figure 1, the progression is very similar for the different translations we produced, while the mBART fine-tuning translation (primary) seems to be more verbose as the translation contains pore tokens than the two contrastive translations. The difference between our different models is clearly not lexical.

4.2 Challenging Literary Aspects of the Test Set

The first challenge was the size of the testing data, which resorted to different text genres, but was 30 times bigger than other challenge datasets like for the biomedical task in 2021. An additional difficulty was the paucity of metadata for the 14 genres or for chapter attributions (22 announced and 12 found).

4.3 Translation Quality analysis based on Error Annotation

4.3.1 Quality overview

In total, 109 sample segments were randomly selected from the twelve translated texts generated by the fine-tuned mBART50 model. Based on these sample segments, each translated text was assigned an overall grade individually by three annotators on a scale of 1 to 10, with 1-3 denoting "Very Poor", 4-6 denoting "Poor", 7-8 denoting "Moderate", and 9-10 denoting "Good". The annotators are native Chinese speakers with near native level of English competence. They work in the domain of translation training and linguistics with an advanced proficiency of Chinese-English translation. The three grades given by the annotators for each text were then averaged to obtain a relative ranking of each translation. Overall, the twelve translations achieved an average score of 5 out of 10 in general, with a standard deviation of 0.87. Specifically, seven subgenres were identified among the twelve texts, namely: fantasy (4 texts), ancient romance (2 texts), military (1 text), thriller (1 text), modern romance (2 texts), sci-fi (1 text), and online games (1 text). All the sub-genres are typical in contemporary web novels. Notably, there is not a clear cut between different sub-genres and this categorisation is for analytical purposes only. Among the identified subgenres, the ranking from high to low quality is as follows: thriller (6.0 out of 10), fantasy (5.7), online games (5.4), sci-fi (5.0), ancient romance (4.7), modern romance (4.6), and military (3.8). While subgenre types might be a factor in influencing the quality of the translation given their language styles (e.g., the proportion of conversational segments, terminologies, formality, etc.), this line of discussion requires further evidence. Among the sample segments, the quality and language style of individual source text seem to play a more vital role in the overall quality of the translations. Several prominent error types linking to the stylistic features of the texts were identified, as detailed below.

4.3.2 Error typology

To obtain a more detailed insight into the quality of these translations, the sample segments were annotated based on the error typology introduced by Hansen and Esperança-Rodier (2023). The original typology was further categorized for the Chinese - English language pair and inter-rater validation purposes. Specifically, six level-one error types were identified:

- semantic errors (SEM): errors that directly affect the meaning of the text, involving issues like omission, addition, or wrong translation of content/nuance of content;
- logical, structural and cohesion errors (LSC): errors related to the logical flow and coherence of the text, affecting how different parts relate to each other;
- grammatical errors (GRM): errors related to

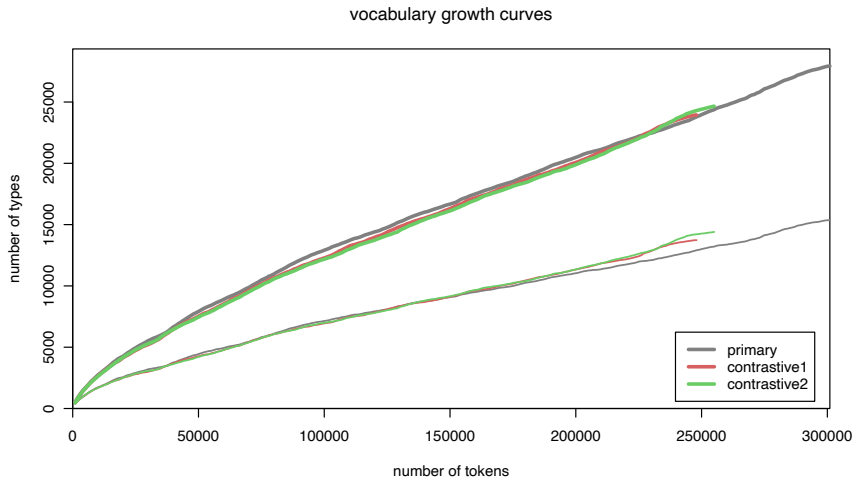

Figure 1: Vocabulary growth curves of our three translations (primary, contrastive1, contrastive2). The lower series of curves corresponds to the hapaxes for primary, ontrastive1 and contrastive2.

the rules of language such as gender, number, tense, and person etc.;

- stylistic errors (STY): errors regarding the style, tone, and appropriateness of the language used;
- stuttering (STU): words repeated for no apparent reason by the MT system;
- non-translation (NTR): source term left untranslated in the target.

Each level one error type contains specific level two and sometimes level three error types. The complete error typology tailored for this task can be found in the appendix.

We use Fleiss' kappa to measure the Level 1 error type inter-rater agreement, and the overall Fleiss' kappa score is 0.288, which can be interpreted as "Moderate agreement" according to Landis and Koch (1977)'s classifications. Fleiss' kappa of Level 1 subgenre annotations is presented in Table 2.

Among all annotated segments, 30.58% segments are considered error-free. 47.71% of them belong to the SEM error type, with the remainder of 11.31% on STY, 4.89% on LSC, and 3.98% on GRM.

4.3.3 Prominent Error Types

Understanding the text in its original language is the basis for literary translation, which requires multi-faceted considerations pinned by context, literary style and cultural nuance. The fine-tuned system attempts to address the greater-than-sentence-level textual features. However, human annotation results have shown that it continues to struggle with contextual analysis, which leads to prominent errors such as non-translation, mistranslation and inconsistent translation or reference of proper nouns and terms, mistranslation of idioms, etc.

Subgenre	Score ↑
Modern Military	0.534
Science Fiction	0.344
Ancient Romance	0.321
Fantasy	0.283
Modern Romance	0.283
Thriller	0.152
Online Game	0.143

Table 2: Fleiss' kappa of subgenres. $\kappa = 1$ is perfect concordance, $\kappa = 0$ is no concordance between annotators.

Transliteration is the main way of addressing the character names from Chinese into English (in this case, standard Pinyin is used). Surprisingly, the system failed to maintain consistency of reference to name entities, for example, "宋扶" (song fu) was translated as "Song Fu", "Song Fudge" and "Song Yidao" at places. The character "宋扶" is also mentioned as "宋师弟" or "宋师兄", which were translated literally (see examples in table 3, hand-annotated in bold). Given the nature of fantasy (xianxia) novels, "师兄" (senior brother) or "师弟" (junior brother) is a common way of addressing

王子法一脸惊讶道：“师兄此话怎讲？”	"What do you mean, senior brother?" Prince Charming asked in surprise.
郑金龙笑眯眯道：“师弟，你是在跟我装糊涂吗？宋师弟的死，你们不准备给师门一个交代？”	"Junior brother, are you playing dumb?" Zheng Jin Long said with a smile.“You don’t want to give your sect an account for **Junior brother Song**’s death?"
王子法面容一肃，沉声道：“宋师兄差点坏我蓝玉门好事，宋扶该死！再给我们一次机会，我们还是会这样做！”	"**Senior Brother Song** almost ruined our Lanyu Sect’s business. **Song Fudge** deserves to die! Give us another chance, and we’ll still do this!" the prince said solemnly.

Table 3: Examples for illustration

people under the same sect. Literal translation in this particular context might reduce textual cohesion and such inconsistent reference might confuse target language readers given the numerous consecutive mentions of “brother” in the text. The same issue was observed in the document-level model (contrastive1) result too.

It is difficult for the system to identify a named entity if the name itself or part of the name can be used as a proper noun. For example, “王子法” (wang zi fa) was mistranslated as “Prince Charming”, which was because the system misidentified the first two Chinese characters “王子” (wang zi, literal meaning: prince) as a named entity.

Other inconsistency regarding proper nouns lies in the formality of presentation, i.e., case error, meaning translation going against previous choices regarding the capitalization of series-specific terms. In fantasy novels, sect names and martial arts techniques are prominent terms. However, the capitalization of these terms was not always consistent.

It is challenging for the current system to capture ideas or emotions in culturally specific expressions. For example, the idiom “天下没有不散的宴席” is translated as “there is no such thing as a banquet in the world”. As a literal translation, it omitted the important part of the idiom “不散的” (literal meaning: non-separable / never-ending), which leads to the failure of conveying its figurative meaning “All good things must come to an end”. On the contrary, it did well in translating “哑巴吃黄连” (literal meaning: a mute person eats bitter melons) as “speechless”. The discrepancy between the translation quality of idioms shows that more culture-specific training data is needed to improve the accuracy and idiomaticity of literary machine translation.

4.4 Sentence- vs. Document-based Training Strategies

An important aspect of the competition was the choice to use full chapters with contextualised successive sentences instead of (more) limited contexts usually retained for translation competitions. This resulted in a much bigger dataset than for more standard competitions (in the vicinity of 400 sentences for biomedical tasks). We submitted 2 models based on a similar architecture: *Contrastive1* and *Contrastive2*.

We used as *Contrastive2* a context-agnostic sentence-level transformer model as in Vaswani et al. (2017) trained on 10 epochs.

We used as *Contrastive1* an on-context transformer model with the exact same architecture as *Contrastive2* but that adopts sliding windows of 3 concatenated sentences pre-trained on 10 epochs to the sentence-level and trained on 4 epochs with concatenated sentences.

Concatenation of 3to3 implies that the source sentence is concatenated to the two previous sentences using end-of-sentence tokens between each of them. A *sliding windows* is when sliding-KtoK model encodes the source windows sentences x^i_K using the end to sentence tokens <eos> and a special token <S> used to mark sentence boundaries in the concatenation then decode the translation y^i_K

$$x^i_K = x^{i-K+1}\text{<}S\text{>}x^{i-K+2}\text{<}S\text{>}...\text{<}S\text{>}x^i\text{<}eos\text{>}$$

$$y^i_K = y^{i-K+1}\text{<}S\text{>}y^{i-K+2}\text{<}S\text{>}...\text{<}S\text{>}y^i\text{<}eos\text{>}$$

Another Contrastive model was trained, but unfortunately too late for the submission, based on Lupo et al. (2022) it has the same specificity than *Contrastive1* with a context discount of 0.01. Context-discount means that the loss function is defined as :

$$\mathcal{L}_{CD}(x^j_K, y^j_K) = CD \cdot \mathcal{L}_{context} + \mathcal{L}_{current}$$

After the submission period, we continued training our contrastive systems. After 55 epochs

of sentence-level pre-training and 14 epochs of document-level training, the system achieved a BLEU score of 21.46 on Test1 test set.

5 Further Research

5.1 Related Research

This subsection discusses related papers.

For fine-tuning mBART, we replicated the parameters tested by Lee et al. (2022), namely retraining for three epochs. With the same parameter, Namdarzadeh et al. (2023) have fine-tuned Persian→English and Persian→French with a single short story but nevertheless observed dramatic improvement for Persian→French translations in terms of elimination of hallucinations, English words and morpho-syntactic correction. We have not tried other multilingual Large Language Models such as mBERT (Wu and Dredze, 2019) (based on BERT), mT5 (Xue et al., 2020), XLM-R (Conneau et al., 2019) based on RobertA or the more recent (and bigger) Bloom model (Scao et al., 2022).

For concatenation Transformer, we used some parameters tested by Lupo et al. (2022) that translated English→German and English→Russian to observe dramatic improvement on Contrapro set (Müller et al., 2018) and English→Russian set (Voita et al., 2019) although with only a slight improvement in BLEU score.

5.2 Future Research

This first collaboration between several universities and backgrounds has discussed English input and was an opportunity to discuss the findings of the competition on literary data and also our insights into the fine-tuning of mBART50 with literary data. We aim to replicate this analysis on Farsi data, as Farsi is one of the 50 languages of mBART50. As is often the case in competitions, we did not train as much as we expected. For the fine-tuning of mBART, we managed to train for three epochs, which is what we found in previous studies (Lee et al., 2022), but for two other submissions, we were training from scratch and could only manage to train for 10 epochs for constrastive2 (sentence-level) and fine-tune for 4 epochs for contrastive1 (document-level). This impacted our results. Evaluating our BLEU score on Test1, we got 22.31 BLEU score for both primary and contrastive2 meanwhile 19.03 BLEU score for constrative1.

6 Conclusion

This paper presented the MAKE-NMTViz system description for the WMT2023 Literary Shared Task. We participated in the Chinese-to-English task with a model trained at sentence level and at document level. We only used the data provided by the organisers but also analysed the translations produced with mBART50 before our submissions. As we did not receive scores from the organisers of the task, we mostly focused on the qualitative analysis of our translations. We resorted to a typology of translation errors and highlighted prominent error types that remained in our translations.

Limitations

During this translation task, we met one limitation with respect to the document-level translation system. In this case, we did not adapt the system to process in Chinese→English language pair. We employed the same setup described in previous works, where the system was trained for English→Russian, English→German and English→French languages.

Acknowledgements

This paper emanated from research partly supported by the MAKE-NMTVIZ project, funded under the 2022 Grenoble-Swansea Centre for AI Call for Proposals/ GoSCAI - Grenoble-Swansea Joint Centre in Human Centred AI and Data Systems (MIAI@Grenoble Alpes (ANR-19-P3IA-0003)), and by a 2021 research equipment grant from the Scientific Platforms and Equipment Committee (PAPTAN project) under the ANR grant (ANR-18-IDEX-0001, Financement IdEx Université de Paris). Sadaf Mohseni benefitted from a Collège de France /Université Paris Cité PAUSE scholarship and Nicolas Ballier from a CNRS research leave at LLF (Laboratoire de Linguistique Formelle), which are gratefully acknowledged. This work was also supported by the CREMA project (Coreference REsolution into MAchine translation) funded by the French National Research Agency (ANR), contract number ANR-21-CE23-0021-01.

References

R. H. Baayen and Elnaz Shafaei-Bajestan. 2019. *languageR: Analyzing Linguistic Data: A Practical Introduction to Statistics*. R package version 1.5.0.

Laurent Besacier and Lane Schwartz. 2015. Automated Translation of a Literary Work: A Pilot Study. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 114–122. ACL.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Ana Guerberof-Arenas and Antonio Toral. 2022. Creativity in Translation: Machine Translation as a Constraint for Literary Texts. *Translation Spaces*, 11(2):184–212.

Damien Hansen and Emmanuelle Esperança-Rodier. 2023. Human-Adapted MT for Literary Texts: Reality or Fantasy? In *Proceedings of the New Trends in Translation and Technology Conference – NeTTT 2022*, pages 178–190. Incoma Ltd.

Yue Jiang and Jiang Niu. 2022. How are neural machine-translated chinese-to-english short stories constructed and cohered? an exploratory study based on theme-rheme structure. *Lingua*, 273:103318.

Yue Jiang and Biyan Yu. 2017. A Contrastive Study on the Rendition of Adjectival Possessive Pronouns in Pride and Prejudice by Human Translation and Online Machine Translation. *Journal of Xidian University*, 2:147–155.

Dorothy Kenny and Marion Winters. forthcoming. Customization, Personalization and Style in Literary Machine Translation. In Marion Winters, Sharon Deane-Cox, and Ursula Böser, editors, *Translation, Interpreting and Technological Changes: Innovations in Research, Practice and Training*. Bloomsburry.

Waltraud Kolb. 2020. Less room for engagement. *Counterpoint*, 4: 26–27.

Taja Kuzman, Špela Vintar, and Mihael Arčan. 2019. Neural Machine Translation of Literary Texts from English to Slovene. In *Proceedings of the Qualities of Literary Machine Translation*, pages 1–9. EAMT.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

En-Shiun Annie Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Ifeoluwa Adelani, Ruisi Su, and Arya D McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'22)*, pages 58–67.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation.

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2022. Focused concatenation for context-aware neural machine translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 830–842, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2023. Encoding sentence position in context-aware neural machine translation with concatenation. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 33–44.

Evgeny Matusov. 2019. The Challenges of Using Neural Machine Translation for Literature. In *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19. EAMT.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Behnoosh Namdarzadeh, Sadaf Mohseni, Lichao Zhu, Guillaume Wisniewski, and Nicolas Ballier. 2023. Fine-tuning mbart-50 with french and farsi data to improve the translation of farsi dislocations into english and french. In *Proceedings of Machine Translation Summit XIX: Users Track*, pages 152–162, Virtual. Association for Machine Translation in the Americas.

Antoni Oliver González. 2017. InLéctor: Automatic Creation of Bilingual E-Books. *Tradumàtica*, 15:21–47.

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

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Chung-ling Shih. 2016. Can Machine Translation Declare a New Realm of Service? Online Folktales as a Case Study. *Theory and Practice in Language Studies*, 6(2):252–259.

Kristiina Taivalkoski-Shilov. 2019. Ethical Issues Regarding Machine(-Assisted) Translation of Literary Texts. *Perspectives*, 27(5):689–703.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring Document-Level Literary Machine Translation with Parallel Paragraphs from World Literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi. ACL.

Antonio Toral and Andy Way. 2018. What level of quality can neural machine translation attain on literary text? In Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, editors, *Translation Quality Assessment: From Principles to Practice*, pages 263–287. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Longyue Wang, Zefeng Du, DongHuai Liu, Deng Cai, Dian Yu, Haiyun Jiang, Yan Wang, Shuming Shi, and Zhaopeng Tu. 2023a. Guofeng: A discourse-aware evaluation benchmark for language understanding, translation and generation.

Longyue Wang, Zefeng Du, Dian Yu, Liting Zhou, Siyou Liu, Yan Gu, Yufeng Ma, Bonnie Webber, Philipp Koehn, Yvette Graham, Andy Wray, Shuming Shi, and Zhaopeng Tu. 2023b. Findings of the wmt 2023 shared task on discourse-level literary translation. proceedings of the eighth conference on machine translation (wmt).

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

A Error Typology

- Semantic Errors: Addition (including Overtranslation); Undertranslation (including Omission); Mistranslation (including Opposite Meaning, Nonsense, and Shift in Meaning); Hallucination; Literal Translation.
- Logical, Structural and Cohesion Errors: Referential Cohesion; Relational Cohesion; Function Words; Logic; Coherence with Previous Volumes; Loss.
- Grammatical Errors: Gender; Number; Tense; Person.
- Stylistic Errors: Language Style; Register; Unfitting Paraphrase; Case; Punctuation; Adaptation; Dialogues.
- Stuttering.
- Non-translation.