Investigating techniques for a deeper understanding of Neural Machine Translation (NMT) systems through data filtering and fine-tuning strategies

Lichao Zhu[3], Maria Zimina-Poirot[1]
Maud Bénard[1], Behnoosh Namdar[1], Nicolas Ballier[13], Guillaume Wisniewski[3], Jean-Baptiste Yunès[2]
[1]CLILLAC-ARP, [2]IRIF, [3]LLF
Université Paris Cité, F-75013 Paris, France
lichao.zhu, maria.zimina-poirot, maud.benard, behnoosh.namdarzadeh,
nicolas.ballier, guillaume.wisniewski, jean-baptiste.yunes@u-paris.fr

Abstract

In the context of this biomedical shared task, we have implemented data filters to enhance the selection of relevant training data for fine-tuning from the available training data sources. Specifically, we have employed textometric analysis to detect repetitive segments within the test set, which we have then used for refining the training data used to fine-tune the mBart-50 baseline model. Through this approach, we aim to achieve several objectives: developing a practical fine-tuning strategy for training biomedical in-domain fr<>en models, defining criteria for filtering in-domain training data, and comparing model predictions, fine-tuning data in accordance with the test set to gain a deeper insight into the functioning of Neural Machine Translation (NMT) systems.

1 Introduction

The objective of our contribution to the biomedical shared task this year is to gain deeper insights into the NMT training pipeline, assess the factors influencing performance, and evaluate the robustness of the training system.

Our training strategy was to build tailor-made fine-tuning data with regard to the test data. We calculated repeated segments (Salem, 1986) in the test data and used a selection of them to extract the corresponding data set from the available training data (as outlined in Section 2.1.1). In order to highlight the "reproducibility", we consistently adhered to the same pipeline with minimum settings presented in Section 2.

While our experience encountered various technical obstacles that likely affected our system's performance, these challenges prompted us to prioritize explainability and comparability. This became especially important when the system produced irrelevant results and hallucinations. We elaborated on this in Sections 3 and 4. The remaining sections of the paper are organized as follows: Section 2 outlines our filtering pipeline, Section 3 delves into our results, and Section 4 provides a discussion of these results. Section 5 explores related research and outlines future work.

2 Optimizing Fine-Tuning Data Selection and Pipeline

For this shared task, we exclusively fine-tuned the "mBart-50 large" baseline model (Tang et al., 2020) over 3-epoch, 5-epoch, and 10-epoch training cycles. Our approach prioritized reproducibility and involved a clear distinction between a "horizontal" dimension (facilitating inter-system analysis, especially in comparison with e-translation[1]) and a "vertical" dimension (examining differences in training strategies within the same system). The fine-tuned setting was established with the following parameters set as a minimum: "num_train_epochs", "train_file", "validation_file", "test_file", "per_device_train_batch_size"[2].

2.1 Raw Training Data

As recommended by the WMT biomedical shared task, we employed the corpora listed in Table 1. The entire corpus, intended for fine-tuning, comprises over 90M words in English and more than 100M in French. It is from these corpora that we extracted the filtered aligned sentences in both source and target languages.

2.1.1 Segmental Proximity Analysis for Training Data Filtering

Our training data filtering strategy is rooted in the theoretical principles of segmental proximity analysis. Repeated segments are sequences that are automatically identified as being repeated within

[1] https://commission.europa.eu/resources-partners/etranslation_fr

[2] Because of hardware constraints, we had to significantly reduce the batch dimension to 4

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 275–281
December 6–7, 2023. ©2023 Association for Computational Linguistics

Corpus	Lines	Words
PubMed abstract	13,033	2,429,484 (en) 3,051,103 (fr)
UFAL	2,693,509	89,191,554 (en) 100,024,568 (fr)
Edp	821	92,309 (en) 110,977 (fr)
Khresmoi	1,500	28,454 (en) 33,189 (fr)
Scielo[3]	9,393 9,501	213,684 (en) 262,377 (fr)

Table 1: Raw corpora used

UFAL (en)	30848
UFAL (fr)	16791
Pubmed abstract (fr)	8393
Pubmed abstract (en)	3566
Edp (fr)	54
Edp (en)	10
Khresmoi (fr)	19
Khresmoi (en)	1

Table 2: Matches in raw training corpus

the same text or across different texts.[4]

The computation of repeated segments (Salem, 1986) is a useful tool in corpus analysis. The computed inventory of repeated segments is of undeniable interest for discourse analysis (Sousa, 2014; Gledhill et al., 2017). This tool facilitates the examination of various discourse phenomena, encompassing the circulation of formulaic expressions, discourse routines, lexico-grammatical patterns, and more. On a cognitive level, the analysis of repeated segments can offer a substantial contribution to the study of knowledge pattern dissemination in specialized discourse. For these reasons, the computation of repeated segments holds substantial value in text profiling, especially in the context of training data selection. Our working hypothesis posits that discerning semantic proximity within a vast dataset can be accomplished through a deliberate selection of repeated segments guided by specific formal criteria, including factors such as frequency and segment length. The method relies on the assumption that related texts share common discourse properties, including phraseology, terminology, and structural patterns. These elements can be effectively "captured" through repeated segments computation and unveiled through segmental proximity analysis (Salem, 1986; Lebart et al., 1997).

The concept of segmental proximity has been thoroughly explored in the work of (Salem, 2006), where the statistical properties of this phenomenon were demonstrated. In this study, Salem (*ibid.*:1) *"considers measures of similarity based on the computation of the frequencies of identical sequences of words among the texts to be compared"*.

In order to identify common sequences within the two sections of a comparable monolingual corpus, it was first necessary to compile a list of segments containing a minimum of four words that were repeated in both parts of the corpus. All segments found exclusively in one part of the corpus were removed from consideration. The sequences that remained were then selected based on their length and their presence in each of the comparable parts of the corpus. By prioritizing relatively longer sequences, we were able to exclude many shorter and more frequent sequences, many of which consisted of common combinations of function words such as "of the" or "by means of."

According to the findings presented in Salem (Salem, 2006), the analysis of lexical distances and proximity indices computed on individual forms (such as the Jaccard index and Chi-square distance) did not reveal any significant affinity between the two sections of the comparable corpus under study. However, when calculations were based on the identification of repeated segments, it became evident that the two parts shared a relatively high number of extended sequences. This methodology enables the study of a range of phenomena related to the circulation of textual units that surpass individual vocabulary items.

Following this line of research, we compiled a systematic inventory of all repeated segments with a length of at least four words, such "acute respiratory syndrome coronavirus", "followed by maintenance therapy", etc. in each test set (English and French), which consisted of texts provided by WMT. We used *iTrameur* (https://itrameur.clillac-arp.univ-paris-diderot.fr) to facilitate this process. We then selected repeated segments with a total frequency of 10 or more. This curated inventory was used for the profiling and filtering of available medical text datasets (training data). We then selected repeated segments having a total frequency

[4] Consortium HN CORpus, Langues et Interactions - Huma-Num: https://corli.huma-num.fr/en/glossaire/repeated-segments/

of 10 or more and used this inventory for profiling and filtering of available medical text sets (training data). This process allowed us to build a dataset that shared common discourse properties and demonstrated semantic similarity with the test set.

2.2 Extraction of Aligned Sentences Containing Filtered Segments

By employing a list of repeated 4-word segments, we implemented a procedure to extract aligned sentences containing these filtered segments (Algorithm 1). To achieve this, we concatenated every delimiter (D) one by one (,;”˜ &|#@=‘-.?!%*$()[]_:+«»§/) of *iTrameur*[5] and every word (g) of each 4-word repeated segment (G) of source language (S, which can be either English or French) to form a regular expression-like pattern GD. We used this pattern to match aligned sentences that contain the segment in both source (S_n, n is the index of matched sentence in raw training corpus of source language) and target (T_n) languages. The extraction result is reported in Table2. The extracted sentences are used to fine-tune the baseline model of mBart50.

```
Data: segment $G$, delimiter $D$ and aligned
      sentences S and T
Result: aligned sentences $S_n$ and $T_n$
        containing $G$
for $g$ in $G$ do
    $GD \leftarrow$ concatenate($g$,$D$)
    if $GD$ in $S_n$ then
        extract $T_n$
    end
end
```

Algorithm 1: Filtering and extraction algorithm

3 Results

In the absence of formal evaluation scores, our approach was largely based on textometric browsing techniques (Zimina, 2005) and qualitative analysis of our submitted translations. These translations were produced using a 3-epoch and 5-epoch training for the en-fr corpus, and a 3-epoch training for the fr-en corpus.

For example, Figure 1 shows a parallel section map generated by *iTrameur*, which helps visualize the alignment of parallel sections from two fr-en translations generated by mBart-50 baseline and mBart-50 5-epoch. The map highlights the presence or absence of the token "violence" in both translations. In the contexts displayed below the map, occurrences of repeated segments are underlined. We employ this tool to compare the translation outputs of various systems.

en-fr	Number	Frequency
train set (en)	60	482.144.115
test data (en)	41	98
fr-en	Number	Frequency
train set (fr)	70	176.229.164
test data (fr)	41	109

Table 3: The occurrence of 4-word sequences used for training in the train set corpus and test data (en, fr)

Texts in French	Number	Frequency
mBart50 (3-epoch)	2	9
mBart50 (5-epoch)	2	4
Texts in English	Number	Frequency
mBart50 (3-epoch)	0	0

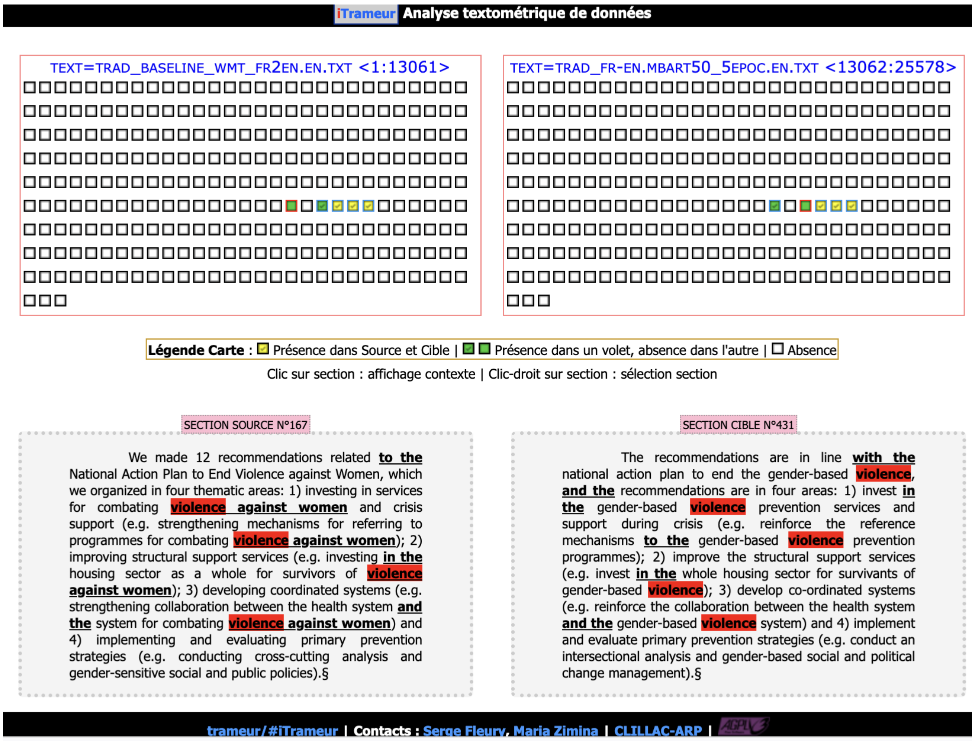
Table 4: The presence of 4-word sequences (en, fr) in translations generated by our systems

In accordance with our training strategy, which is based on segmental proximity analysis (as described in Section 2.1.1), we expected the test data and the test set to have a substantial number of long sequences in common, assuming that many of the long sequences used for training would be shared. To confirm this, we examined the presence of the 136 four-word sequences (repeated segments) used in training within the test data. The results align with our research strategy, as shown in Table 3): 68% (41 sequences) of the English sequences are found in the en-fr test data and 59% (41 sequences) of the French sequences are present in the fr-en test data.

Continuing along this line of investigation, we examined our submitted translations, yet the analysis revealed that our translated texts had very little overlap with the repeated segments employed in training, as demonstrated in Table 4.

In the following paragraphs, we narrow our focus to two sequences taken from the repeated segments used for training. These examples help illustrate the challenges encountered by our systems and highlight the complexities involved in drawing

[5] https://itrameur.clillac-arp.univ-paris-diderot.fr



iTrameur Analyse textométrique de données

TEXT=TRAD_BASELINE_WMT_FR2EN.EN.TXT <1:13061>

TEXT=TRAD_FR-EN.MBART50_5EPOC.EN.TXT <13062:25578>

Légende Carte : 🟨 Présence dans Source et Cible | 🟩 🟩 Présence dans un volet, absence dans l'autre | □ Absence

Clic sur section : affichage contexte | Clic-droit sur section : sélection section

SECTION SOURCE N°167

We made 12 recommendations related **to the** National Action Plan to End Violence against Women, which we organized in four thematic areas: 1) investing in services for combating **violence against women** and crisis support (e.g. strengthening mechanisms for referring to programmes for combating **violence against women**); 2) improving structural support services (e.g. investing **in the** housing sector as a whole for survivors of **violence against women**); 3) developing coordinated systems (e.g. strengthening collaboration between the health system **and the** system for combating **violence against women**) and 4) implementing and evaluating primary prevention strategies (e.g. conducting cross-cutting analysis and gender-sensitive social and public policies).§

SECTION CIBLE N°431

The recommendations are in line **with the** national action plan to end the gender-based **violence**, **and the** recommendations are in four areas: 1) invest **in the** gender-based **violence** prevention services and support during crisis (e.g. reinforce the reference mechanisms **to the** gender-based **violence** prevention programmes); 2) improve the structural support services (e.g. invest **in the** whole housing sector for survivants of gender-based **violence**); 3) develop co-ordinated systems (e.g. reinforce the collaboration between the health system **and the** gender-based **violence** system) and 4) implement and evaluate primary prevention strategies (e.g. conduct an intersectional analysis and gender-based social and political change management).§

trameur/#iTrameur | **Contacts :** Serge Fleury, Maria Zimina | CLILLAC-ARP |

Figure 1: Parallel section map generated by *iTrameur*.

significant conclusions from a qualitative analysis of our translations.

With 13 occurrences in the fr-en test data, the sequence *"violence envers les femmes"* presents an intriguing case of terminology. A closer examination of the raw training corpus reveals that several other expressions are commonly employed in French to express the same concept, such as *"violence faite aux femmes"* and *"violence contre les femmes"*, among others. In English, we note a reduced degree of variation, primarily using "violence against women" and, to a lesser extent, "violence directed against women", which is not as prevalent as the former. According to the European Institute for Gender Equality (EIGE), an EU agency, the most accurate translation is "violence against women".[6]. However, in the translations generated by the 3-epoch training, this accurate translation is absent, and the proposed translations ("the women's domestic violence" and "gender-based violence") do not correspond to the same concepts. The medical term "blood flow" appears 10 times in the en-fr test data and is part of one of the repeated segments used for training, specifically "blood flow in dogs." It is also a frequent component in complex noun phrases, including "the dermal blood flow," "regional blood flow," or "auricular dermal blood flow" (with 8 occurrences). Both the 3-epoch and 5-epoch systems encounter challenges when translating this repeated segment and the complex noun phrases. Firstly, there's a high degree of terminological variation in the French texts (*"écoulement sanguin"*, *"débit sanguin"*, *"flux sanguin"*, *"circulation sanguine"*), given the limited occurrences of the term in English. Additionally, we observe instances of hallucinations and incomplete outputs in the 3-epoch system. The 5-epoch system, on the other hand, omits one element in the translation of the complex noun phrase, even though the 3-epoch system accurately translated it.

In a specific case, "regional blood flow," the 5-epoch system incorrectly deduced the semantic relationship between the head and a modifier, yielding "*l'écoulement régional du sang*" instead of the correct "*l'écoulement sanguin régional*", which the 3-epoch system produced.

4 Discussion

4.1 Errors and Inconsistencies Arising From Variations in the Train Set

In general, the output exhibits numerous errors and inconsistencies primarily arising from terminological variations within the train set and the inherent

[6] https://eige.europa.eu/publications-resources/thesaurus/terms

heterogeneity of the selected training data. For instance, the terms "gender-based violence" and "violence against women" are both employed in comparable contexts within the train set, as illustrated by segments like "Many women who experience gender-based violence may never seek any formal help..." and "Violence against women is a global phenomenon" (source: PubMed abstracts, train set: en).

4.2 Hallucinations with mBart-50

Analyzing the translations generated by mBart-50 at different epochs (5 and 10) proves to be an interesting area of research. According to Lee et al. (2018), hallucinations occur when the model produces significantly different and inadequate outputs when the source is subjected to specific noise models. Therefore, we can suggest that there may be instances where the model ceases to translate the source text and instead generates an output composed solely of a continuous sequence of tokens from the present invention. This could be seen as an alternative form of hallucination in epoch 10. Moreover, in epoch 10, there are also examples of incomplete translations produced by the system. Based on our analysis, we observed that these errors tend to be resolved during the training process. Consequently, in mBart-50 (5-epoch) fr-en, there are no "X" tokens present, in contrast to the baseline model translation. This result is highlighted by the calculation of generalized co-occurrence networks conducted using *iTrameur*. Figures 2 and 3 depict co-occurrence networks that represent the most characteristic lexical attractions in the fr-en translations produced by two models: mBart-50 baseline and mBart-50 5-epoch. The numbers on the edges represent the strength of lexical attraction: *Specificity Index > 9* (Co-frequency > 1).[7] Co-occurrence networks serve as a monitoring tool for tracking changes across various training stages.

5 Related Research and Future Work

Corpus filtering is discussed in many previous and recent works for training data preparation with different approaches : perplexity threshold of text segments(Moore and Lewis, 2010), metric evaluations of raw NMT models' outputs (Duh et al., 2013), acceptability of filtering evaluated by mulitlingual BERT classifier(Zhang et al., 2020), etc. Our approach aims at data relevance between a given test set and in-domain training data.

5.1 Robustness of NMT Models

An essential aspect to consider is that the system generates tokens regardless of whether it possesses the relevant information for translation. However, this raises the need for potential trigger warnings in situations where the system lacks adequate data for accurate translation. This suggests an avenue for developing a confidence index that reflects the system's efforts when generating output. We consider to explore various parameters based on sentence level and on a token level to build such a confidence index, e.g. the scores used by certain large language models to assess the confidence of each token's projection or beam search and the number of competitors for each token to gauge the complexity of a text for translation.

Another aspect to consider is the model's over/underfitting. By plotting our training and test data features in Figure 4, we find out that the model is indeed overfitted from depth 4. That explains partially why the model under-performed and helps us to choose a better data fitting strategy in the future.

5.2 Fine-tuning of mBart-50 and Other Multilingual Systems

For the moment, we only tried to fine-tune mBart-50 as a multilingual large language model, whereas other systems have been developed since, some of them with many more parameters. We may try to replicate or fine-tune experiments with more classical systems such as SYSTRAN Model Studio Advanced (https://u-paris.fr/plateforme-paptan), but also other different multi-lingual large language models such as mT5 (Xue et al., 2020) or Bloom (Scao et al., 2022), a 176-billion parameter language model, in spite of its carbon footprint (at least 24.7 tons of carbon just for the dynamic power consumption) (Luccioni et al., 2022).

Finally, it is worth noting that our fine-tuning efforts were primarily centered around mBart-50, a multilingual large language model. However, since our experiments, various other systems have emerged, some with significantly more parameters. It might be advantageous for us to replicate or fine-tune experiments with more conventional systems based on translation models, such as the SYSTRAN Model Studio Advanced (available at Université Paris Cité: https://u-paris.fr/plateforme-paptan), and explore different multilingual large

[7] For specific details regarding the computation of *Specificity Index*, refer to (Lebart et al., 1997).

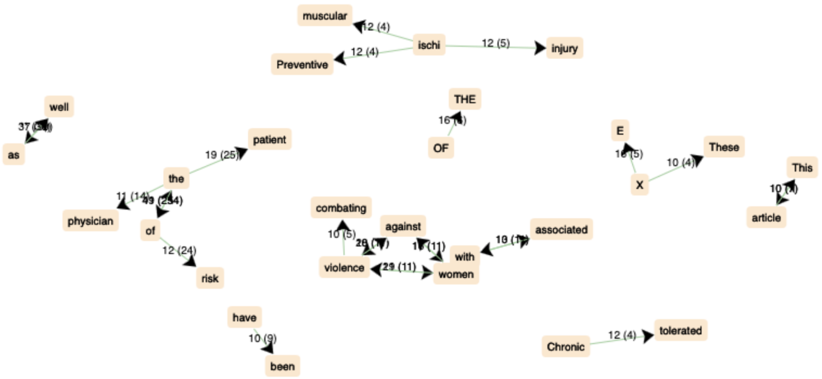


Figure 2: Co-occurrence networks for the baseline.

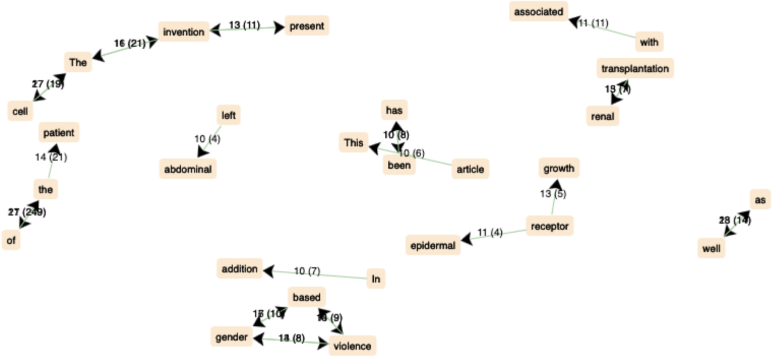


Figure 3: Co-occurrence networks for the 5-epoch system.

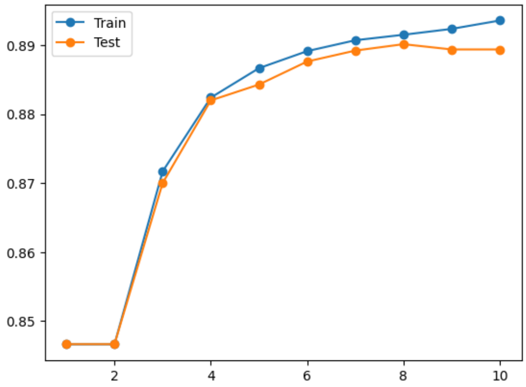


Figure 4: Model's accuracy prediction by training and test data features classification

language models like mT5 (Xue et al., 2020) or Bloom (Scao et al., 2022). Notably, Bloom is a huge 176-billion-parameter language model of major interest, despite its substantial carbon footprint, which amounts to at least 24.7 tons of carbon emissions solely for dynamic power consumption (Luccioni et al., 2022).

6 Conclusion

In this paper, we have described the translation systems used for the submissions in the WMT23 biomedical task6 (our data are available at: https://github.com/lichaozhu/WMT23). Nevertheless, due to certain hardware constraints, we were unable to pinpoint the exact reasons for the model's underperformance.

We also considered our previous participation in the biomedical task. Since 2021, we have recognized that having scores provided in advance and reference texts used for score computation can significantly facilitate our work. These resources enable a more critical evaluation of the translations we generate.

To address the absence of reference translations and evaluation results, translations can undergo spot checks. In our work, these checks involved the use of qualitative examples to assess the model's successes and failures. Additionally, textometric browsing helped to unveil distinctive features within multiple machine translation outputs.

Acknowledgements

This publication has emanated from research supported in part by a 2021 research equipment grant (PAPTAN project)[8] from the Scientific Platforms and Equipment Committee, under the ANR grant (ANR-18-IDEX-0001, Financement IdEx Université de Paris).

References

Kevin Duh, Graham Neubig, Katsuhito Sudoh, and Hajime Tsukada. 2013. Adaptation data selection using neural language models: Experiments in machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 678–683, Sofia, Bulgaria. Association for Computational Linguistics.

Christopher Gledhill, Maria Zimina-Poirot, and Stéphane Patin. 2017. Lexico-grammaire et textométrie : identification et visualisation de schémas lexico-grammaticaux caractéristiques dans deux corpus juridiques comparables en français. *Corpus*.

Ludovic Lebart, André Salem, and Lisette Berry. 1997. *Exploring Textual Data*, volume 4. Springer Science & Business Media.

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

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2022. Estimating the carbon footprint of bloom, a 176b parameter language model. *arXiv preprint arXiv:2211.02001*.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

André Salem. 2006. Proximités segmentales. In *Actes des 8e Journées internationales d'Analyse statistique des Données Textuelles (JADT 2006)*, pages 839–849, Besançon, France. Université de Franche-Comté.

André Salem. 1986. Segments répétés et analyse statistique des données textuelles. *Histoire & Mesure*, 1(2):5–28.

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*.

Serge de Sousa. 2014. De la statistique textuelle à l'analyse des idéologies: l'exemple du discours révolutionnaire en amérique latine (1810-2010). *Corela. Cognition, représentation, langage*, HS(15).

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

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*.

Boliang Zhang, Ajay Nagesh, and Kevin Knight. 2020. Parallel corpus filtering via pre-trained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8545–8554, Online. Association for Computational Linguistics.

Maria Zimina. 2005. Bi-text Topography and Quantitative Approaches of Parallel Text Processing. In *The Corpus Linguistics 2005 conference*, volume 1 Issue 1, Birmingham, United Kingdom. Centre for Corpus Research, Birmingham University.

[8] Plateforme pour l'apprentissage profond pour la traduction automatique neuronale, in English: Deep Learning for Machine Translation at Universite Paris Cité. See the description of the platform on the project website: `https://u-paris.fr/plateforme-paptan`