Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Réduction des répétitions dans la Traduction Automatique Neuronale

Marko Avila Anabel Rebollo Josep Crego
4 rue du Port aux Vins, F-92 150 Suresnes, France
{mavila,arebollo,jcrego}@chapsvision.com

RÉSUMÉ

Actuellement, de nombreux systèmes TAL utilisent des décodeurs neuronaux pour la génération de textes, qui font preuve d'une capacité impressionnante à générer des textes approchant les niveaux de fluidité humaine. Toutefois, dans le cas des réseaux de traduction automatique, ils sont souvent confrontés à la production de contenu répétitif, également connu sous le nom de diction répétitive ou de répétition de mots, un aspect pour lequel ils n'ont pas été explicitement entraînés. Bien que cela ne soit pas intrinsèquement négatif, cette répétition peut rendre l'écriture monotone ou maladroite si elle n'est pas utilisée intentionnellement pour l'emphase ou des fins stylistiques. La répétition de mots a été traitée par des méthodes post-hoc pendant l'inférence, contraignant le réseau à examiner des hypothèses auxquelles le système avait initialement attribué une plus faible probabilité. Dans cet article, nous implémentons une méthode qui consiste à pénaliser les répétitions lors de l'apprentissage et qui s'inspire des principes du *label smoothing*. Conformément à cette méthode, nous modifions la distribution de la vérité terrain afin d'orienter le modèle de manière à décourager ces répétitions. Les résultats de nos expériences montrent que les méthodes proposées permettent de contrôler le problème de la répétition dans les moteurs neuronaux de traduction automatique sans compromis en termes d'efficacité ou de qualité des traductions.

ABSTRACT

Reducing Repetitions in Neural Machine Translation

Many contemporary NLP systems rely on neural decoders for text generation, which demonstrate an impressive ability to generate text approaching human fluency levels. However, in the case of neural machine translation networks, they often grapple with the production of repetitive content, also known as repetitive diction or word repetition, an aspect they weren't explicitly trained to address. While not inherently negative, this repetition can make writing seem monotonous or awkward if not used intentionally for emphasis or stylistic purposes. Repetitions have been addressed through post-hoc methods during inference, compelling the network to consider hypotheses it initially assigned lower probability. In this paper, we implement a repetition penalty method applied at learning inspired by the principles of label smoothing. In line with label smoothing, we modify the ground-truth distribution to steer the model towards discouraging repetitions. Experiments show the ability of the proposed methods in reducing repetitions within neural machine translation engines, without compromising efficiency or translation quality.

MOTS-CLÉS : Traduction automatique neuronale ; Génération de texte ; Répétitions.

KEYWORDS: Neural machine translation ; Text generation ; Repetitions.

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 198–210.

1 Introduction

This study addresses the issue of word repetition in machine translation, which involves the repeated occurrence of words, phrases, or ideas throughout the translation process. More specifically, we focus on repetitions that typically occur when translating synonyms or semantically equivalent phrases found in the source sentence. These repetitions lead to diminished readability of the text, potentially causing boredom or confusion for the reader and creating the perception of verbose or awkward writing. Consider, for instance, the following French sentence :

— *nous avons **lutté** contre l'infodémie en **combattant** les mythes par des informations fiables.*

and its two corresponding English translations :

— *We have **combated** the infodemia by **combating** myths with reliable information.*

— *We have **fought** the infodemia by **combating** myths with reliable information.*

Since both translations are grammatically correct and semantically equivalent, the first one is clumsy due to word repetition [1], while the second one effectively avoids repetition by suggesting alternative translations *fought* and *combating*, for the French words *lutté* and *combattant*, resulting in a smoother and more preferable translation.

In neural machine translation, repetition often arises when the model faces input sentences containing synonyms, leading these synonymous terms to be translated into identical words. Although lacking numerical support for our observation, the repetition issue becomes more salient when utilizing a model trained across multiple domains, highlighting the dearth of lexical diversity. We attribute this phenomenon to a lack of diversity in the decoder module. Although this type of repetition may occur with low frequency, it is highly concerning as it vividly illustrates the lack of fluency in translations.

However, repetitions do not always have a negative impact on readability. Without aiming to be exhaustive : i) repetitions play a role when summarizing information or reinforcing a concept ; ii) common expressions are formed using word repetitions, and altering them to eliminate repetition would alter their intended meaning ; iii) in highly specialized domains, expressions convey precise meanings that disallow being reformulated. The following examples illustrate these observations :

i) *once **closed**, the door stays **closed***
ii) ***over** and **over** ; **to be** or not **to be** ; **step** by **step***
iii) *the congenital **muscular** dystrophy in newborns presenting with **muscular** hypotonia*

As previously introduced, finding suitable alternatives without altering the meaning of a sentence can be a challenging task. In this work, we propose a method applied in training designed to teach the model to discourage certain repetitions, thereby alleviating the need for difficult decisions during inference. Next, we summarize the main contributions of this work :

— We propose a method that discourages repetitions during the training phase by adjusting the ground-truth distribution so as to penalize repetitions more severely.
— We introduce a technique for gathering examples containing both, acceptable repetitions and repetitions that hinder fluency, which are then utilized during the training phase.
— We build a curated test set that includes various types of word repetitions found in machine translations. Evaluation on this test set provides deeper insights into the repetitions issue.

Repetitions can manifest in various forms, including single words, phrases, and larger segments of content. However, this work concentrates on repetitions manifested through the repetition of linguistic

1. Note that repetitions can diminish readability, even when they occur as inflectional variants. Is the case of the repeated words in our example, *combated* and *combating*.

words, which are more commonly observed in machine translations. Note that linguistic words are typically decomposed into multiple tokens as taken into account by neural networks.

2 Related Work

The fluency levels achieved by LLMs are widely acknowledged to be high, primarily owing to the extensive availability of monolingual datasets, which surpasses that of standard neural machine translation (NMT) models trained solely on parallel texts. To the best of our knowledge, no dedicated research has been conducted on addressing the repetition issue tackled in this work within NMT systems. Closely related, (Welleck *et al.*, 2019) describe a method to train neural language models that in addition to maximizing likelihood to model the overall sequence probability distribution, also includes an unlikelihood term in the loss function to correct known biases such as repeated tokens. (Li *et al.*, 2020) use the same approach to control copy effect and repetitions observed in dialogue tasks. (Su *et al.*, 2022) present a contrastive solution to encourage diversity while maintaining coherence in the generated text.

Various studies have addressed diversity in neural MT systems, which is a closely related topic. Sampling predictions from the output distribution can be an effective decoding strategy for back-translation, as described by (Edunov *et al.*, 2018), or sampling from less likely tokens (Holtzman *et al.*, 2020). Results show that such techniques enlarge diversity and richness of the generated translations when compared to data generated by beam or greedy search, but introduce semantic inconsistency in translations. In (Lin *et al.*, 2022) is proposed a multi-candidate optimization framework for augmenting diversity. The authors propose to guide an NMT model to learn more diverse translations from its candidate translations based on reinforcement learning. During training, the model generates multiple candidate translations, of which rewards are quantified according to their diversity and quality.

A different approach attempts to condition the decoding procedure with diverse signals. Typically, (Shu *et al.*, 2019) use syntactic codes to condition the translation process. (Lachaux *et al.*, 2020) replace the syntactic codes with latent domain variables derived from target sentences. Similarly, (Schioppa *et al.*, 2021) use prefix-based control tokens and vector-based interventions for controlling output translations from a NMT system. In the context of paraphrase generation (Vahtola *et al.*, 2023) propose a translation-based guided paraphrase generation model that learns useful features for promoting surface form variation in generated paraphrases.

3 Adjusting the ground-truth distribution

Throughout the training process, at every time-step t, neural machine translation networks generate predictions over the target-side vocabulary based on the input x and previous predictions $y_{<t}$:

$$p_t^i = p(y_t^i | x, y_{<t}),\ i \in [1, ..., V]$$

where V indicates the size of the target vocabulary.

The loss function evaluates the neural network's capacity to model the training data by comparing

t	1	2	3	4	5	6
r	I	like	cookies	and	cookies	.

$\mathcal{M}$

.	0	0	0	0	0	0
I	0	0	0	0	0	0
and	0	0	0	0	0	0
like	0	0	0	0	0	0
cookies	0	0	0	0	1	0

FIGURE 1 – Matrix $\mathcal{M}$ for the ground-truth r ='*I like cookies and cookies.*'. Rows t and r represent respectively the time-step and the corresponding ground-truth token. A reduced model vocabulary (matrix rows) is used to facilitate reading.

its predictions to a reference target vector $r = [r_1, r_2, ..., r_T]$, where T denotes the sequence length. This loss is utilized to update the network's parameters, aiming to minimize the observed error in the model. The loss at time-step t is usually computed as the cross-entropy between the model predictions $p_t = [p_t^1, ..., p_t^V]$ and the ground-truth distribution $q_t = [q_t^1, ..., q_t^V]$:

$$\mathcal{L}_t = -\sum_{i=1}^{V} q_t^i \, log(p_t^i) \tag{1}$$

Note that the vector q_t is a one-hot encoding representation of r_t, with all entries set to 0 except for the token indicated by r_t, which is set to 1. Addressing the over-fitting risk illustrated by the previous q_t distribution, label smoothing (Szegedy *et al.*, 2015; Müller *et al.*, 2019) (LS) is widely employed to achieve a smoother distribution :

$$q_t^{\epsilon LS} = (1 - \epsilon) q_t + \frac{\epsilon}{V} \tag{2}$$

with ϵ being a commonly small hyper-parameter. [2]

LS can be interpreted as penalizing the probability of the ground-truth class by a factor of $1 - \epsilon$, while evenly distributing the removed probability mass among all classes, ϵ/V. Building upon a strategy akin to label smoothing, we make additional adjustments to the ground-truth distribution and reduce the likelihood of repeated tokens, with the goal of enabling the model to learn to predict repetitions with lower probability. We introduce a matrix, denoted as $\mathcal{M}_{V \times T}$, which indicates whether the ground-truth token r_t is also present in the preceding time-steps. [3] Figure 1 illustrates an example of matrix $\mathcal{M}$ with ground-truth *I like cookies and cookies.* as translation of the French sentence *J'aime les cookies et les biscuits.* with a model vocabulary of 5 tokens (matrix rows). Both French terms *cookies* and *biscuits* are correctly translated into English as *cookies*, yet this choice clearly reduces the fluency and clarity of the translation. As it can be seen, only $\mathcal{M}_{[i=5,t=5]}$ is set to 1 since only r_5 ='*cookies*' occurs in a preceding time-step ($t = 3$).

We consequently update the ground-truth distribution following :

$$q_t^{\epsilon LS \alpha \mathcal{M}} = (1 - \epsilon)(1 - \alpha \mathcal{M}_t) \; q_t + \frac{\epsilon}{V} \tag{3}$$

2. $\epsilon = 0$ yields the initial distribution q_t, whereas $\epsilon = 1$ implies a uniform distribution.
3. Note that repetitions are computed over words while matrix $\mathcal{M}$ refers to tokens $r \in V$ for each time-step $t \in T$.

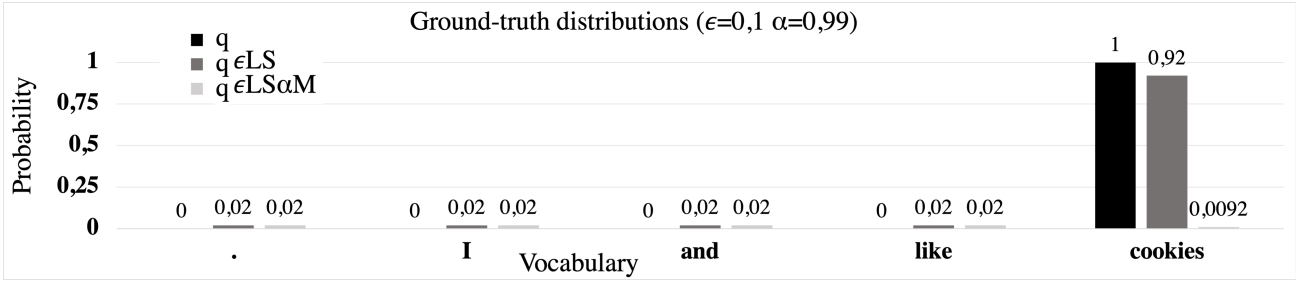

FIGURE 2 – Ground-truth distributions for the 5^{th} time-step of our example : the original one-hot encoding q ; adjusted with label smoothing $q^{\epsilon LS}$; and further adjusted with repetitions $q^{\epsilon L\alpha\mathcal{M}}$.

where α is a hyper-parameter, and $\alpha\mathcal{M}$ is used as a penalty, much like ϵ in the case of LS. Note that only the label smoothing probabilities discounted are distributed among all classes. As a result, time-steps with repeated tokens (such as $t = 5$ in our example) do not constitute proper probability distributions, as their sum does not add to 1. Figure 2 illustrates ground-truth distributions for our example at time-step $t = 5$: the original one-hot encoding q ; the original distribution adjusted using label smoothing $q^{\epsilon LS}$, and further adjusted using repetitions $q^{\epsilon LS\alpha\mathcal{M}}$.[4] A significant challenge with the aforementioned techniques that modify q distribution with repetitions is their limited impact on the training process, primarily caused by the scarcity of repeated tokens in datasets. In the following section, we present alternative approaches to address this challenge.

4 Gathering Examples with Repetitions

As previously depicted, our intention is to instruct the model to minimize certain repetitions while preserving others deemed necessary for an accurate translation. To achieve this, we must compile a relatively large dataset of examples that demonstrate this behavior to the model. We initially focus on repetitions of content words such as *nouns*, *adjectives*, *verbs*, and *adverbs*. Function words, which serve a distinct grammatical role in a sentence, are excluded from this analysis. Current MT networks reliably generate these words based on their understanding of grammatical correctness.

Given that training corpora contain a relatively small number of repetitions, and these are manually curated, often comprising only acceptable repetitions, we opt to focus solely on repetitions found in machine translations. Accordingly, we translate the source sentences (src) of our training examples to generate synthetic translations (hyp). Repetitions that detrimentally impact fluency are only considered if they appear in such translation. Identifying examples containing repetitions following the previous morpho-syntactic patterns is straightforward. However, the challenge lies in discerning which repetitions degrade the fluency of translations and which do not. We follow the next filtering steps to select repetitions degrading fluency :

— We first word align the source (src) and synthetic (hyp) sentences and eliminate repetitions in the synthetic sentences that also align with repeated words in the source sentence. Word alignments are performed by the Giza++ (Och & Ney, 2003) toolkit[5]. This approach is

4. As previously discussed, distribution $q^{\epsilon LS\alpha\mathcal{M}}$ does not form a proper distribution since probabilities do not add to 1 $(0,02 + 0,02 + 0,02 + 0,02 + 0,0092 = 0,0892)$. We leave for future experiments the normalization of the output scores in order to allow for a valid probability distribution.

5. https://github.com/moses-smt/giza-pp.

	Degrading	*Acceptable*
src	The home was modest and frugal	I want to know what you mean
hyp	La maison était modeste et modeste	Je veux savoir ce que tu veux dire
tgt	La maison était modeste et économe	Je veux savoir ce que tu veux dire

TABLE 1 – Two synthetic translations containing repeated words (underlined) and their corresponding source (hyp) and reference translation (tgt) sentences. Shades of grey indicate word alignments between repeated words and their alignments in the src/tgt sentences.

based on the premise that if repeated words are necessary in a human-generated sentence, the corresponding translation may similarly require the repetition. This holds true for repetitions used to reinforce a concept or in highly specialized domains.

— Next, we also align the synthetic (hyp) and target (tgt) sentences word by word and remove repetitions of the synthetic sentences (hyp) aligned to repeated words in the reference translation. The same previous rationale remains consistent, now encompassing examples of common expressions that necessitate word repetitions.

The resulting set of examples with repetitions from src/hyp training pairs will be regarded as instances that the model needs to learn to discourage. Consequently, we utilize them for training after annotating the repeated words in their respective $\mathcal{M}$ matrices. Table 1 illustrates the procedure previously outlined for two synthetic translations (hyp) containing repetitions.

The repeated word on the left-side example is the French adjective *modeste* while the verb *veux* is repeated in the right-side example. Once word alignments are computed between synthetic translations and their respective source and target sentences, the example on the right is not classified as a repetition degrading fluency. This is because both instances of *veux* are aligned with repeated words in the target reference translation, suggesting that the repetition is motivated. Concerning the example on the left side, neither the source nor the reference target sentences contain repeated words, suggesting that the repetition stems from a lack of diversity when translating *modest* and *frugal*, thus hindering fluency.

We employ the left-side example to train the model to identify it as a repetition to be avoided. The corresponding matrix $\mathcal{M}$ marks the second occurrence of the word *modeste* with a 1, thereby incurring in a significant loss if the model predicts it with high probability. The example on the right is used without penalization, thus instructing the model to reproduce the repetition.

It's worth noting that the presented approach does not require any alterations to the network architecture and maintains the same training and inference efficiency.

5 Experimental Framework

We evaluate the proposed methods in an English-to-French translation task. Thus, we utilize English-French parallel corpora freely obtained from the Opus website [6]. We strive for balanced utilization across various domains and ensure the inclusion of clean parallel data whenever possible. Due to the extensive volume of French-English parallel sentences accessible we randomly choose a subset exceeding 7 million examples that we employ as *Training set*.

For testing, we make use of English-French *News-test* (2008 to 2013) datasets made available

6. http://opus.nlpl.eu

	Training set		*Repetition-test*	
Type	*Degrading*	*Acceptable*	*Degrading*	*Acceptable*
Noun	170,169	356,105	25	34
Verb	36,111	41,949	24	33
Adjective	22,834	51,599	26	30
Adverb	4,016	6,097	26	1
Total	233,130	455,750	101	98

TABLE 2 – Number of repetitions that degrade fluency and those which do not, found in both the *Training* and *Repetition-test* corpora. Occurrences are also displayed considering the morpho-syntactic function of repetitions.

through the WMT'2014 translation shared task [7]. In addition, we use a held-out *Repetition-test* composed of reference English and their corresponding French machine translations that feature at least one repeated word on the target (French) side for a more nuanced analysis of repetition. Machine translations were obtained with our baseline NMT model (referred in Appendix B as *baseline*). The *Repetition-test* set primarily serves to assess our models' performance in handling repetition issues, while we employ the *News-test* set to evaluate overall translation accuracy. Further details of the train and test datasets used are given in Appendix A.

We translate the English side of the *Training* and *Repetition-test* sets following the procedure detailed in Section 4 to identify repetitions which hinder fluency (*Degrading*) and those which do not (*Acceptable*). Table 2 displays the frequency of repetitions identified within the French translations. The table presents the occurrence of both types of repetitions, accompanied by an analysis of their frequency concerning the morpho-syntactic function of the repetitions [8]. Sentences are morpho-syntactically analyzed using the spaCy [9] toolkit.

Our NMT model is built using an in-house implementation of the state-of-the-art Transformer architecture (Vaswani *et al.*, 2017). Details of the network and training work are given in Appendix B.

6 Results and Analysis

To evaluate the methods presented in this paper we consider the previous *baseline* model that we update with 15K additional iterations for two different configurations of the ground-truth distribution :

$q^{\epsilon LS}$ follows the same configuration than the *baseline* model with label smoothing set to $\epsilon = 0.1$.

$q^{\epsilon LS\alpha\mathcal{M}}$ further penalizes the ground-truth distribution with repetition penalties as detailed in Section 3 with $\epsilon = 0.1$ and for different values of α.

Note that for both configurations, we use 7.6M reference sentence pairs detailed in Table A (*Training set*) together with the synthetic translations containing repetitions predicted *Degrading* and *Acceptable* of Table 2, summing up to 7.6M + 233K + 455K sentence pairs. It's essential to find a balance between the number of sentences in each training set (reference and synthetic) to uphold overall quality while teaching the model to minimize specific repetitions.

7. https://www.statmt.org/wmt14/translation-task.html
8. Only French adverbs ending with suffix **ment* are considered.
9. https://spacy.io/ with the French fr_core_news_lg model.

Configuration	*Repetition-test* BLEU	COMET	*Degrading*	*Acceptable*	*News-test* BLEU	COMET
$q^{\epsilon LS}$	45.15	37.36	99	95	32.60	26.50
$q^{\epsilon LS\alpha\mathcal{M}}, 1-\alpha = 10^{-2}$	45.54	39.10	81	89	32.45	26.05
$q^{\epsilon LS\alpha\mathcal{M}}, 1-\alpha = 10^{-3}$	45.63	40.31	79	87	32.44	26.18
$q^{\epsilon LS\alpha\mathcal{M}}, 1-\alpha = 10^{-6}$	45.65	40.13	77	86	32.50	26.30
$beam, \beta = 0$	35.03	31.28	0	0	20.66	10.65
$beam, topk = 10$	44.34	37.88	94	85	32.45	25.91
$GPT3.5$	29.70	29.59	25	43	29.98	27.60
$NLLB$	34.13	25.37	51	57	31.98	23.64

TABLE 3 – Translation accuracy results and number of repetitions present in translations performed by models under different configurations. Two different test sets are considered. ϵ is always set to 0.1.

Configuration $beam$ employs the *baseline* model and performs inference following two strategies to reduce repetitions and improve diversity :

β with a penalty applied at each inference time-step t whenever token y_t appears repeated in the hypothesis prefix $y_0, ..., y_{t-1}$. Probability p_t is reduced by factor $0 \leq \beta \leq 1$. Thus, reducing the likelihood of such hypotheses.

$topk$ Sampling predictions from the k most likely tokens of the output distribution. This is an effective decoding strategy typically used for increasing diversity when building back-translation datasets.

We also assess the effectiveness of two large language models (LLM) with translation capabilities to overcome the repetition issue :

$GPT3.5$ consists of the *GPT3.5-turbo* version of the OpenAI LLM. Built upon the Generative Pre-trained Transformer architecture (Radford & Sutskever, 2018) which employs only a transformer decoder. Following an auto-regressive approach, the model ensures that the generated text maintains coherence and relevance to the context provided by the input text. Translations are conducted using the OpenAI API, while emphasizing the importance of minimizing word repetitions through the provided prompt [10].

$NLLB$ is a family of machine translation models based on the Transformer encoder-decoder architecture, enabling translation between any of the 202 language varieties (NLLB Team *et al.*, 2022). We use the *nllb-200-distilled-600M* [11] version and perform translations with the efficient CTranslate2 [12] inference toolkit.

To evaluate the presented methods, we report BLEU and COMET results computed by `sacrebleu` [13] (Post, 2018) and `comet-score` [14] (Rei *et al.*, 2020) respectively over both test sets. Concerning *Repetition-test*, we also report the number of word repetitions that hinder fluency, *Degrading*, and those deemed acceptable, *Acceptable*, measured in translation hypotheses.

10. Prompt = *Translate the following text from English to French, ensuring that the translated output maintains coherence and fluency while minimizing the repetition of words or phrases. Pay attention to using synonyms, varied sentence structures, and appropriate linguistic devices to enhance the overall quality of the translation. Feel free to creatively adapt the language to achieve a natural and engaging tone in the target language.I want you to only reply the traduction, do not write explanations*
11. https://huggingface.co/facebook/nllb-200-distilled-600M
12. https://github.com/OpenNMT/CTranslate2
13. https://github.com/mjpost/sacrebleu
14. https://github.com/Unbabel/COMET

Models fine-tuned from the *baseline* network exhibit nearly identical quality scores across the *News-test* sets. This suggests that training with the method presented to adjust the ground-truth distribution does not compromise translation quality. On the contrary, unlike Configuration $q^{\epsilon LS}$, Configurations $q^{\epsilon LS\alpha\mathcal{M}}$ demonstrate a significant decrease in the number of repetitions that degrade fluency over the *Repetition-test*, while retaining most of the acceptable repetitions in the translated output. Note also the increase in quality over the *Repetition-test* set as measured by COMET score ($\sim 40 > 37.62$). Adjusting α does not seem to have a significant impact on reducing repetitions that degrade fluency. Decreasing its value gradually (lightly) reduces the occurrence of such repetitions. As expected, the number of acceptable repetitions remains unchanged since the training input signal with acceptable repetitions remains constant across all α values.

Regarding inference-based configurations, *beam* with $\beta = 0$ effectively eliminates all repetitions but at the expense of a notable decrease in translation quality, in the case of $topk = 10$ the number of repetitions is lightly reduced as well as global accuracy.

Results from both LLMs demonstrate a reduced number of repetitions, suggesting an elevated level of diversity and fluency of such models. However, the translation quality scores of LLMs do not align with those achieved by the models presented in this study in either of the test sets, especially translations obtained by GPT-3.5. These findings are consistent with those presented by (Bawden & Yvon, 2023) where the authors note the challenge of controlling translations performed by BLOOM [15], a multilingual LLM.

Table 4 illustrates reference translations (src and tgt) together with translations by models $q^{\epsilon LS}$ and $q^{\epsilon LS\alpha\mathcal{M}}$. The first (top) examples exhibit the ability of model $q^{\epsilon LS\alpha\mathcal{M}}$ to avoid *degrading* repetitions. The last examples contain *acceptable* repetitions hypothesized by both models.

7 Conclusions and Further Work

We have introduced a method to reduce the occurrence of repetitions in translation hypotheses, which significantly affects the readability of the generated texts. Additionally, we have proposed a straightforward approach to identify repetitions in machine translations that detract from fluency. The method is solely implemented during fine-tuning at the conclusion of the training phase, without any modifications to the inference process. Experiments indicate the ability of our proposed methods in reducing the repetition problem. Additional experiments are necessary to confirm the applicability of the proposed methods across various language pairs and dataset conditions. We aim to further study the impact of the ratio between the number of reference sentences and synthetic translations that include repetitions during the training process. Additionally, we plan to analyze the influence of the distance (measured in number of words) between repetitions and explore the possibility of replacing the binary penalty in matrix $\mathcal{M}$ with a softer approach.

Remerciements

The work presented in this paper was supported by the EU Horizon 2020 Programme for TRACE project (Grant Agreement No. 101022004).

15. https://huggingface.co/bigscience/bloom

src	(h) liabilities, including unliquidated obligations ;
tgt	h) les dettes, y compris les engagements non réglés ;
$q^{\epsilon LS}$	(h) les engagements, y compris les engagements non réglés ;
$q^{\epsilon LS\alpha\mathcal{M}}$	(h) les **passifs**, y compris les **engagements** non réglés ;
src	We talked about the tourism, hospitality and hotel sectors.
tgt	On a parlé des secteurs du tourisme, de l'hébergement et de l'hôtellerie.
$q^{\epsilon LS}$	Nous avons parlé des secteurs du tourisme, de l'hôtellerie et de l'hôtellerie.
$q^{\epsilon LS\alpha\mathcal{M}}$	Nous avons parlé des secteurs du tourisme, de l'**accueil** et de l'**hôtellerie**.
src	The home was modest and frugal.
tgt	C'était une maison modeste.
$q^{\epsilon LS}$	La maison était modeste et modeste.
$q^{\epsilon LS\alpha\mathcal{M}}$	La maison était **modeste** et **économe**.
src	Courts will tackle the question anyway, often obliquely or indirectly.
tgt	Les tribunaux vont se poser cette question de toute façon, souvent à mots couverts ou indirectement.
$q^{\epsilon LS}$	Les tribunaux aborderont la question de toute façon, souvent indirectement ou indirectement.
$q^{\epsilon LS\alpha\mathcal{M}}$	Les tribunaux aborderont la question de toute façon, souvent de façon **oblique** ou **indirecte**.
src	Technology travels fast and is swiftly adopted.
tgt	Les technologies se distribuent rapidement et sont rapidement adoptées.
$q^{\epsilon LS}$	La technologie voyage rapidement et est rapidement adoptée.
$q^{\epsilon LS\alpha\mathcal{M}}$	La technologie voyage **vite** et est **rapidement** adoptée.
src	Parliament must be able to exercise its power of scrutiny.
tgt	Le Parlement européen doit exercer son pouvoir de contrôle.
$q^{\epsilon LS}$	Le Parlement doit pouvoir exercer son pouvoir de contrôle.
$q^{\epsilon LS\alpha\mathcal{M}}$	Le Parlement doit **être en mesure d'**exercer son **pouvoir** de contrôle.
src	Interferometer apparatus and interferometric method
tgt	Appareil interféromètre et procédé interférométrique
$q^{\epsilon LS}$	Appareil interférométrique et procédé interférométrique
$q^{\epsilon LS\alpha\mathcal{M}}$	Appareil interférométrique et procédé interférométrique
src	Cardiac murmur, heart rate increased
tgt	Souffle cardiaque, augmentation de la fréquence cardiaque
$q^{\epsilon LS}$	Souffle cardiaque, fréquence cardiaque augmentée
$q^{\epsilon LS\alpha\mathcal{M}}$	Souffle cardiaque, fréquence cardiaque augmentée

TABLE 4 – Models configured with ($q^{\epsilon LS\alpha\mathcal{M}}$) and without ($q^{\epsilon LS}$) penalization exhibit varying performance when encountering repetitions (outlined using blue). Reference source (src) and target (tgt) translations are also indicated. The initial (top) examples include repetitions that *degrade* fluency, whereas the final (bottom) examples feature *acceptable* repetitions.

Références

BAWDEN R. & YVON F. (2023). Investigating the translation performance of a large multilingual language model : the case of BLOOM. In M. NURMINEN, J. BRENNER, M. KOPONEN, S. LATOMAA, M. MIKHAILOV, F. SCHIERL, T. RANASINGHE, E. VANMASSENHOVE, S. A. VIDAL, N. ARANBERRI, M. NUNZIATINI, C. P. ESCARTÍN, M. FORCADA, M. POPOVIC, C. SCARTON & H. MONIZ, Éds., *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, p. 157–170, Tampere, Finland : European Association for Machine Translation.

EDUNOV S., OTT M., AULI M. & GRANGIER D. (2018). Understanding back-translation at scale. In E. RILOFF, D. CHIANG, J. HOCKENMAIER & J. TSUJII, Éds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 489–500, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1045.

HOLTZMAN A., BUYS J., DU L., FORBES M. & CHOI Y. (2020). The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020* : OpenReview.net.

KINGMA D. P. & BA J. (2014). Adam : A method for stochastic optimization. *arXiv preprint arXiv :1412.6980*.

LACHAUX M.-A., JOULIN A. & LAMPLE G. (2020). Target conditioning for one-to-many generation. In T. COHN, Y. HE & Y. LIU, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 2853–2862, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.256.

LI M., ROLLER S., KULIKOV I., WELLECK S., BOUREAU Y.-L., CHO K. & WESTON J. (2020). Don't say that ! making inconsistent dialogue unlikely with unlikelihood training. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 4715–4728, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.428.

LIN H., YANG B., YAO L., LIU D., ZHANG H., XIE J., ZHANG M. & SU J. (2022). Bridging the gap between training and inference : Multi-candidate optimization for diverse neural machine translation. In M. CARPUAT, M.-C. DE MARNEFFE & I. V. MEZA RUIZ, Éds., *Findings of the Association for Computational Linguistics : NAACL 2022*, p. 2622–2632, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-naacl.200.

MÜLLER R., KORNBLITH S. & HINTON G. (2019). *When Does Label Smoothing Help ?*, In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc. : Red Hook, NY, USA.

NLLB TEAM, COSTA-JUSSÀ M. R., CROSS J., ÇELEBI O., ELBAYAD M., HEAFIELD K., HEFFERNAN K., KALBASSI E., LAM J., LICHT D., MAILLARD J., SUN A., WANG S., WENZEK G., YOUNGBLOOD A., AKULA B., BARRAULT L., MEJIA-GONZALEZ G., HANSANTI P., HOFFMAN J., JARRETT S., SADAGOPAN K. R., ROWE D., SPRUIT S., TRAN C., ANDREWS P., AYAN N. F., BHOSALE S., EDUNOV S., FAN A., GAO C., GOSWAMI V., GUZMÁN F., KOEHN P., MOURACHKO A., ROPERS C., SALEEM S., SCHWENK H. & WANG J. (2022). No language left behind : Scaling human-centered machine translation.

OCH F. J. & NEY H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, **29**(1), 19–51.

POST M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation : Research Papers*, p. 186–191, Belgium, Brussels : Association for Computational Linguistics.

RADFORD, ALEC N. K. S. T. & SUTSKEVER I. (2018). Improving language understanding with unsupervised learning. *Technical Report*.

REI R., STEWART C., FARINHA A. C. & LAVIE A. (2020). COMET : A neural framework for MT evaluation. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 2685–2702, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.213.

SCHIOPPA A., VILAR D., SOKOLOV A. & FILIPPOVA K. (2021). Controlling machine translation for multiple attributes with additive interventions. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 6676–6696, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.535.

SENNRICH R., HADDOW B. & BIRCH A. (2016). Neural machine translation of rare words with subword units. In K. ERK & N. A. SMITH, Éds., *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1715–1725, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/P16-1162.

SHU R., NAKAYAMA H. & CHO K. (2019). Generating diverse translations with sentence codes. In A. KORHONEN, D. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 1823–1827, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1177.

SU Y., LAN T., WANG Y., YOGATAMA D., KONG L. & COLLIER N. (2022). A contrastive framework for neural text generation. In A. H. OH, A. AGARWAL, D. BELGRAVE & K. CHO, Éds., *Advances in Neural Information Processing Systems*.

SZEGEDY C., VANHOUCKE V., IOFFE S., SHLENS J. & WOJNA Z. (2015). Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 2818–2826.

VAHTOLA T., CREUTZ M. & TIEDEMANN J. (2023). Guiding zero-shot paraphrase generation with fine-grained control tokens. In A. PALMER & J. CAMACHO-COLLADOS, Éds., *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, p. 323–337, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.starsem-1.29.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems*, volume 30 : Curran Associates, Inc.

WELLECK S., KULIKOV I., ROLLER S., DINAN E., CHO K. & WESTON J. (2019). Neural text generation with unlikelihood training. *ArXiv*, **abs/1908.04319**.

A Corpora Statistics

Table 5 presents various statistics of the corpora used in this work, including the total number of sentences, vocabularies, words, and average sentence length. Statistics are computed after performing a light tokenization aiming to split-off punctuation.

Side	Sentences	Vocabulary	Words	Length
Training set				
English	7.6M	755K	174M	22.9
French		839K	208M	27.3
News-test				
English	16,071	27K	401K	24.1
French		32K	468K	29.1
Repetition-test				
English	199	1,323	2,521	12.6
French		1,352	3,215	16.1

TABLE 5 – Corpora statistics. M and K stand for millions and thousands respectively.

size of word embedding	512
size of hidden layers	512
size of inner feed forward layer	2,048
number of heads	8
number of layers	6
batch size	4,000 (tokens)
batch accumulation	25 (batches)

TABLE 6 – Network hyperparameters.

B NMT Network

Table 6 indicate the hyper-parameters employed to build our translation network.

For optimization work we use the lazy Adam algorithm (Kingma & Ba, 2014). We set warmup steps to 4,000 and update learning rate for every 8 iterations. All models are trained using a single NVIDIA V100 GPU.

We limit the source and target sentence lengths to 150 tokens based on BPE (Sennrich *et al.*, 2016) preprocessing in both source and target sides. We use a joint vocabulary of $32K$ tokens for both source and target sides. In inference we use a beam size of 5.

Our *baseline* English-to-French model is trained during more than 3 million iterations using all the parallel data available in the Opus website (see Appendix A).