Learning From Wrong Predictions in Low-Resource Neural Machine Translation

Jia Cheng Hu, Roberto Cavicchioli, Giulia Berardinelli, Alessandro Capotondi
University of Modena and Reggio Emilia
via G.Campi 213/b, 41125, Modena, Italy
jiacheng.hu@unimore.it, roberto.cavicchioli@unimore.it,
giulia.berardinelli@unimore.it, alessandro.capotondi@unimore.it

Abstract

Resource scarcity in Neural Machine Translation is a challenging problem in both industry applications and in the support of less-spoken languages represented, in the worst case, by endangered and low-resource languages. Many Data Augmentation methods rely on additional linguistic sources and software tools but these are often not available in less favoured language. For this reason, we present USKI (Unaligned Sentences Keytokens pre-training), a pre-training strategy that leverages the relationships and similarities that exist between unaligned sentences. By doing so, we increase the dataset size of endangered and low-resource languages by the square of the initial quantity, matching the typical size of high-resource language datasets such as WMT14 En-Fr. Results showcase the effectiveness of our approach with an increase on average of 0.9 BLEU across the benchmarks using a small fraction of the entire unaligned corpus, suggesting the importance of the research topic and the potential of a currently under-utilized resource and under-explored approach.

Keywords: Low-resource, Translation, Pre-training.

1. Introduction

Over the past years, Deep Learning methods achieved outstanding results in Neural Machine Translation (NMT), however, the performances of these systems are not equally distributed across all languages. Popular languages such as English, Chinese, Russian, and Spanish benefit from the availability of massive amounts of training data in contrast to Endangered Languages (ELs) and Low-Resource Languages (LRLs). To outline the severity of the problem, there are about 7000 currently spoken languages (Moseley, 2010) but more than half of them are estimated to be severely endangered or dead by the year 2100 (Sallabank and Austin, 2011). In addition to that, only 20 are spoken by 50% of the world's population, whereas most of the remaining languages are spoken by less than ten thousand people (Sallabank and Austin, 2011). In these cases, the development of a translation system, or the integration of less-spoken languages into systems based on more popular ones such as English, Chinese, Arabic, and Hindi, can increase the economic opportunities for the speaking minorities as well as provide ways to improve the preservation and revitalization of endangered languages. In addition to that important Natural Language Processing (NLP) tools and projects such as WordNet (Fellbaum, 2010) and CoreNLP (Manning et al., 2014) exist for English but other languages need to be translated first to access them. Finally, in addition to ELs and LRLs, the lack of resources is a common issue in industrial applications, where, for instance, Large Language Models might need to be fine-tuned on small and domain-specific data such as product reviews collected by a young company.

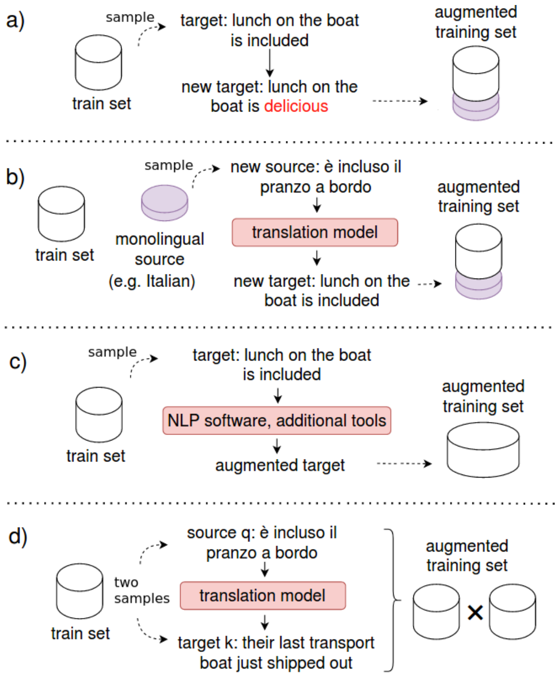
a)
sample
target: lunch on the boat is included
train set
new target: lunch on the boat is delicious
augmented training set
b)
sample
new source: è incluso il pranzo a bordo
train set
monolingual source (e.g. Italian)
translation model
new target: lunch on the boat is included
augmented training set
c)
sample
target: lunch on the boat is included
train set
NLP software, additional tools
augmented target
augmented training set
d)
two samples
source q: è incluso il pranzo a bordo
train set
translation model
target k: their last transport boat just shipped out
augmented training set
×

Figure 1: Overview of most popular approaches compared to our proposal: a) Word replacement. b) Back-translation and Self-Learning; c) Data Augmentation based on NLP tools; d) *Our proposed* approach.

Overall, the lack of data is an important and significant problem in Natural Language Processing

LREC-COLING 2024, pages 10263–10273
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

(NLP) that affects many languages and many possible circumstances. To mitigate the issue works in literature proposed a variety of Data Augmentation techniques. Established works can be categorized into three approaches that are not mutually exclusive. The Figure 1 shows, in brief, these strategies. One approach is based on creating synthetic training data (Fadaee et al., 2017), by replacing a single word in sentences. This method is particularly effective in NMT because models can leverage the translation of semantically improbable sentences as long as the linguistic structure is well preserved. Another stream of work approach consists of generating additional data from monolingual sources using Back-Translation and Self-Learning methods (Bojar and Tamchyna, 2011; Goutte, 2009; Hoang et al., 2018; Li and Specia, 2019). While these methods are proven to be very effective, many of them require additional resources such as monolingual sources, in-domain dictionaries (Peng et al., 2020), alignment models, or NLP software (Duan et al., 2023), which can be a significant limiting factor, especially with endangered and low-resource languages.

Given a high-resource Language dataset and a low-resource one, the size of the first is often in the order of millions, in contrast to the size of the latter, which is typically in the order of thousands. This means that if models can learn from *unaligned sentences*, they can exploit the information from millions of training points in the low-resource as well which can alleviate the data amount gap compared to high-resource datasets. We believe that *unaligned sentences* are currently an under-exploited resource of the Machine Translation field. Motivated by the importance of supporting less favoured languages and the lack of work in that same research direction as our proposed method. In this work, we present USKI, which is short for *Unaligned Sentences Keytokens pre-traIning*, a simple pre-training method that leverages the relationship that exists between unaligned sentences. In particular, we pre-train models to predict matching tokens between unaligned translation sentences. By doing so, results showcase an increase in accuracy over groups of sub-words arbitrarily distributed in sentences, leading to an overall improvement of 0.9 BLEU on average across the resource-lacking setups.

In Section 2, we provide an overview of the related works for the topic. In Section 3, we present our method, the experimental setup is introduced in Section 4, and results are showcased in Section 5. Finally, in Section 6, we draw the conclusions.

2. Related Works

There are many proposals for Data Augmentation in Natural Machine Translation. (Fadaee et al., 2017) proposed to create additional training data by replacing some words in translation sentence pairs without changing the linguistic structure. In SwitchOut (Wang et al., 2018) the authors trained the model with a new optimization formula based on training data in which words in both source and target sentences are replaced with other words sampled from the respective vocabularies. (Nishimura et al., 2018) proposed to fill and replace the sentences in one source with elements from other source languages. (Gao et al., 2019) to combat the potential loss in semantics during word replacing proposed to sample words to be replaced according to a special distribution over the vocabulary that takes into account the similarity of the original word. Another popular method consists of Back-Translation and Self-Learning (Bojar and Tamchyna, 2011; Goutte, 2009; Hoang et al., 2018; Li et al., 2020), which produces a synthetic parallel corpus from additional monolingual datasets for the target and source languages. This method has been proven successful in several NMT systems (Berard et al., 2019; Zheng et al., 2019; Helcl et al., 2019). (Li and Specia, 2019) extends the back-translation by studying the effectiveness of injecting different forms of noise to increase robustness. (Li et al., 2020) refined these methods using sampling strategies that encourage diversity. (Peng et al., 2020) leveraged an in-domain dictionary-based data augmentation to reduce the model performance gap on in-domain and out-of-domain data. (Liu et al., 2021) followed the idea of word replacement but accomplished the goal with an alignment model and a masked language model. (Kondo et al., 2021) proposed a simple but effective data augmentation based on the concatenation of unrelated training sentences. MTL (Sánchez-Cartagena et al., 2021) proves the effectiveness of combining multiple Data Augmentation strategies. (Duan et al., 2023) adopts the distance in a dependency tree to operate word-level editing strategies.

In contrast to most previous works, our proposal does not rely on additional monolingual resources. It does not create different targets using noise or sampling strategies such as word replacement, nor does it require *external* tools or models for alignment or dependency parsing. Our model proposes a novel pre-training stage that leverages the syntactic relationships in the training set between *unaligned sentences*. By doing so, the pre-training step increases the size of the dataset *by the square of its quantity*, and the model can better extract the valuable content that is currently underutilized by the standard training practice and the established Data Augmentation methodologies.

This approach contrasts with the popular methods in the literature of extracting aligned sentences from unaligned corpus (Smith et al., 2010; Azpeitia et al., 2018; Ding et al., 2021; Chen, 1993). The most similar work to ours is (Kondo et al., 2021), where the concatenation practice is used to massively increase the number of training samples. However, while their method focuses on improving the quality of long translations, it leads to mixed results in the case of shorter sequences. Overall, our method is orthogonal to the previous approaches, and it can be combined with existing methods to increase models' performance in resource-lacking conditions.

3. Method

Given a parallel training corpus D_{L_X,L_Y} where L_X is the source language and L_Y denotes the target language. The standard translation training consists of sampling paired sentences $(X, Y) \in D_{L_X,L_Y}$ and requires a model, whose parameters are denoted with θ, to minimize the Cross-Entropy loss function:

$$-\sum_{t}^{T} log(p_\theta(y_t = g_t|y_{1:t-1}, X)) \qquad (1)$$

where $y_1, \ldots, y_T$ are tokens in the predicted sequence, g_t is the ground-truth token for time step t and p_θ denotes the output distribution of the model. In contrast, the main idea of our pre-training method consists of leveraging the information contained in arbitrary source and target sentences $(X_i, Y_j) \in D_{L_X} \times D_{L_Y}$. As a result, sampled pairs are mostly unaligned, in other words, the predictions are incorrect or wrong in a translation setup. However, we believe that involving such data in the training of translation models can be beneficial for two reasons:
- *Full Data Exploitation*. In real applications, the amount of translation data can be scarce for a variety of reasons, such as in the case of endangered languages. Therefore, it is desirable to get as much information content as possible out of the available data. Given all permutations of two independent and identically distributed translation samples $\forall(X_i, Y_i), (X_j, Y_j) \in D_{L_X,L_Y} \times D_{L_X,L_Y}$, it is reasonable to expect the presence of similar linguistic structures or sub-sets of correct alignments between the two pairs for $i \neq j$ as depicted in Figure 2. These structures are only partially exploited or leveraged in an indirect manner when models are trained exclusively on correct pairs.
- *Robustness*. Neural networks suffer from the train-inference discrepancy. In the translation case, the problem is amplified by the fact that erroneous predictions in the early steps increase the likelihood of worse predictions in the subsequent steps. By showing incorrect pairs to the networks, the latter are encouraged to make correct predictions out of noisy and unexpected or incorrect sentence fragments.
While these aspects describe the potential and theoretical goals of leveraging unaligned translation pairs, in Section 3.1 we propose our implementation which represents only one possible solution. We test our method on the popular Transformer architecture (Vaswani et al., 2017)

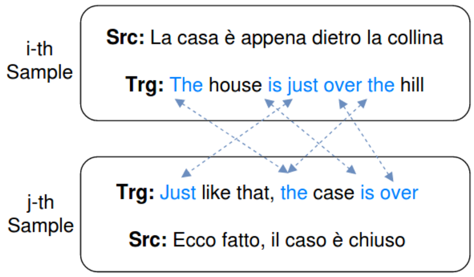
i-th Sample
Src: La casa è appena dietro la collina
Trg: The house is just over the hill
j-th Sample
Trg: Just like that, the case is over
Src: Ecco fatto, il caso è chiuso

Figure 2: Example of similarity between two arbitrary translation samples (from Italian to English).

3.1. USKI

Pre-training is a popular technique leveraged by many Large Language Models, which are trained over a massive collection of linguistic sources and then fine-tuned on translation over smaller datasets (Liu et al., 2020; Raffel et al., 2020; Costa-jussà et al., 2022; Junczys-Dowmunt et al., 2018). Inspired by this approach, in this Section, we present a pre-training method that leverages unaligned pairs of source and target predictions and is applied to the model before the training on the translation task.

Dataset
Given the training corpus D_{L_X,L_Y}, we define D_{L_X} and D_{L_Y} the collections of sentences in the source language L_X and target language L_Y. We then construct the *Augmented Dataset* $\tilde{D} = D_{L_X} \times D_{L_Y}$ made of all possible pairings of source and target language sentences. Since $|\tilde{D}| = |D_{L_X,L_Y}|^2$ training directly on $\tilde{D}$ can be very expensive and time consuming. For this reason, in practice, we will consider a portion of the augmented training set $\tilde{D}$ in the experiments.

Pre-Training Strategy
Given two arbitrary samples in source and target language $X_i \in D_{L_X}$ and $Y_j \in D_{L_Y}$ we define a *Key-Token* as a matching token between two different sentences Y_j and Y_i in the target language. We denote with $\mathbb{K}_{i,j}$ the set of all key-tokens between the target Y_i and Y_j, formally $K_{i,j} = Y_i \cap Y_j$. For instance, assuming the target sentences of two generic unaligned pairs being: Y_i={'a', 'piece', 'of', 'furniture', 'with', 'four', 'legs', 'used', 'for', 'eat-

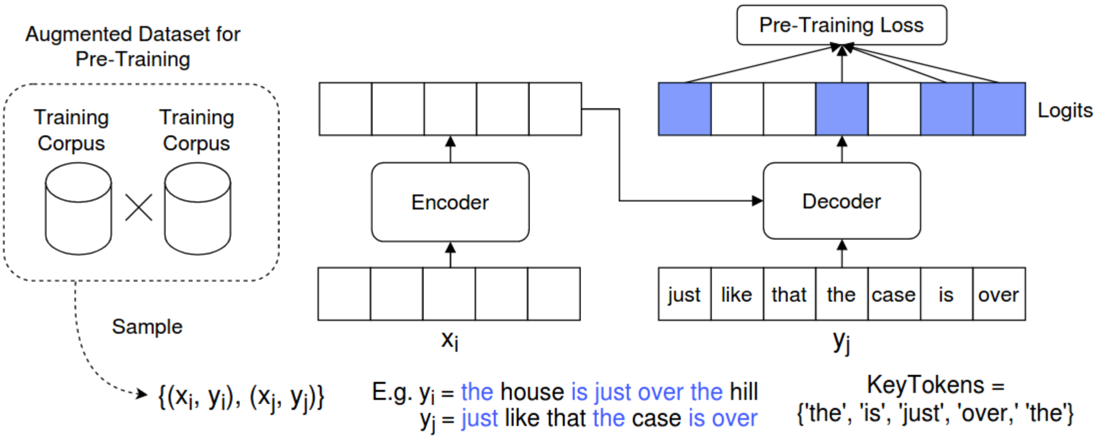
Augmented Dataset for Pre-Training
Training Corpus
Training Corpus
×
Sample
{(xi, yi), (xj, yj)}
Encoder
xi
Decoder
Pre-Training Loss
Logits
just
like
that
the
case
is
over
yj
E.g. yi = the house is just over the hill
yj = just like that the case is over
KeyTokens =
{'the', 'is', 'just', 'over,' 'the'}

Figure 3: Pre-Training operating principle.

ing'} and Y_j={'a', 'dog', 'is', 'eating', 'a', 'piece', 'of', 'chicken'}. KeyTokens $K_{i,j}$ denote a set of tokens contained in both the correct translation Y_i and the arbitrary one Y_j: $K_{i,j}$ = {'a', 'piece', 'eating', 'of'}. Given a training sample $(X_i, Y_j) \subset (X_i, Y_i, X_j Y_j)$, $(X_i, Y_i, X_j Y_j) \in \tilde{D}$ the Pre-Training minimizes the following formula:

$$-\sum_{t}^{T} \mathbb{1}(g_t) \cdot log(p_\theta(y_t = g_t | y_{1:t-1}, X_i)) \cdot w(g_t) \quad (2)$$

where $y_1, \dots, y_T$ are tokens in the prediction, $g_1, \dots g_T$ are the ground-truth tokens of Y_j and $\mathbb{1} : V \rightarrow \mathbb{R}$ and $w : V \rightarrow \mathbb{R}$ are functions that map tokens from the vocabulary V to $\mathbb{R}$. $\mathbb{1}$ is described by the following equations:

$$\begin{cases} 1 & if \;\; g_t \in \mathbb{K}_{i,j} \\ 0 & if \;\; g_t \notin \mathbb{K}_{i,j} \end{cases} \quad (3)$$

wheres w denotes the Inverse-Document-Frequency (IDF) function. To compute the IDF term $w(g), g \in V$ we first define the un-normalized version $\tilde{w}(g)$:

$$\tilde{w}(g) = log(\frac{|D_{L_X, L_Y}| \cdot \alpha}{max(1, \phi(g))}) \quad (4)$$

where α is a corpus size correction coefficient and $\phi : V \rightarrow \mathbb{N}$ is a function that counts the number of occurrences of g in the target training set:

$$f(g) = \sum_{z}^{|D_{L_Y}|} min(1, c_{gz}) \quad (5)$$

where c_{gz} is the number of occurrences of g in the sequence Y_z. The term $w(g)$ is then defined as:

$$w(g) = \frac{\tilde{w}(g)}{\sum_{q \in V} \tilde{w}(q)} + \gamma \quad (6)$$

where γ is an offset weight, and together with α they are configurable parameters.

The IDF term w is introduced to prevent the model from focusing on function words of the language L_Y which are the most likely words to co-occur in two arbitrary sentences. At the same time, since it induces bias over less frequent sub-words it might not always be the optimal choice in some translation tasks and increases the convergence time. Therefore, during the experiments, the IDF term is not always adopted. In these cases, we consider $w(g)$=1 $\forall g \in V$.

The Pre-Training process is depicted in Figure 3.

USKI Benefits

One notable aspect of the pre-training method is that it does not make assumptions about the source and target languages. In particular, the fact that the source and target languages might present very different structures (e.g., Chinese and English) does not affect our proposed approach. Since the KeyTokens are defined by sentences belonging to the same language.

We designed the pre-training method to increase the robustness and accuracy of the model over KeyTokens. In particular, we expect the model to achieve better performances since training on arbitrarily distributed KeyTokens in the decoded output should mitigate the phenomena of error propagation in autoregressive architectures. In detail, when an auto-regressive model is trained to predict the ground-truth token for position q, it is based upon the assumption that all the previous tokens in position $t < q$ are correct. As a result, if the model predicts an incorrect token at certain time j, all the following predictions for $t > j$ will be affected by this mistake. In contrast, during our pre-training strategy, we leverage wrong target sentences and train the model to correctly predict a special set of tokens (the KeyTokens) regardless

Language pair	Type	# training	(# training)^2	# validation	# test	Vocab. size
Sel-Ru	EL-H	7251	52.57 $\cdot 10^6$	200	200	4255
Ev-Ru	EL-H	2136	4.56 $\cdot 10^6$	100	400	3907
Gk-It	EL-H	8136	66.19 $\cdot 10^6$	200	1000	4068
Uz-En	LRL-H	3689	13.60 $\cdot 10^6$	99	199	4356
Wol-It	LRL-H	5916	34.99 $\cdot 10^6$	200	1000	4006

Table 1: Translation datasets statistics. Type refers to EL=Endangered Language, LRL=Low Resource Language, H=High Resource Language.

of the uncorrelated tokens (not KeyTokens) observed in the previous steps. This increases the robustness against unexpected tokens or incorrect predictions in the auto-regressive procedure.

Sel: Täp ponä ča:ǯɨn	⟷	**Ru:** Он на улицу вышел
Ev: Тар нуркōкōсён ирэн	⟷	**Ru:** Этот паренек зашел
Gk: Èbbie i ' stra ' ce nsìgnase na pai .	⟷	**It:** Si mise in strada e s' incamminò .
Uz: Har bir joy uchun har hil neyron bor	⟷	**En:** Different neurons for different locations
Wol: Ku ma gis , gis nga ki ma yónni .	⟷	**Uk:** I хто видить мене , видить Пославшого мене .

Figure 4: Translation examples sampled from the experimental datasets.

4. Experimental setup

4.1. Datasets

To evaluate our method we select several language pairs and dataset sizes to encompass a variety of resource-lacking situations. In addition to more popular languages such as Russian, English, and Ukrainian, we also involve 2 LRLs and 3 ELs. A lack of agreement over the exact definition of LRL was pointed out in (Hämäläinen, 2021). In our work, we will use the term LRL[1] to refer to a language that is not endangered and is involved in a translation dataset made of a few thousand samples, which generally outlines the lack of large, organized and well-known corpora but does not necessarily imply the lack of available data in the "wild". Endangered Languages were identified according to the UNESCO [2].

Overall, we select Uzbek-English (Uz-En) pairs from the QED (QCRI Educational Domain) Corpus (Tiedemann, 2012), the Wolof-Ukrainian (Wol-Uk) from the Parallel Bible Corpus (Mayer and Cysouw, 2014). Regarding ELs, inspired by the work of (Mossolova and Smaïli, 2022), we choose the Selkup-Russian (Sel-Ru) and the Evenki-Russian (Ev-Ru). The first is represented by the SelkupCorpus (Maria Brykina, Svetlana Orlova and Beáta Wagner-Nagy, 2018)[3], whereas the second dataset is extracted from the "Minority languages of Siberia as our cultural heritage" project (Olga Kazakevich et. al.)[4]. Finally, we consider the Griko-Italian (Gk-It) dataset from (Anastasopoulos et al., 2018). Since we did not find a short identifier for the Griko dialect, with an abuse of notation, we will refer to it with the code "Gk".

Details of each unfiltered dataset can be found in Table 1. Examples of translations are depicted in Figure 4.

4.2. Sentence Processing

Sentences are involved in a simple pre-processing pipeline. First, they are tokenized using Byte-Pair Encoding (BPE) (Sennrich et al., 2016) with 4000 codes and it is applied to both source and target language to create a shared vocabulary. Tokens are lowercase and punctuation is preserved. Sentences whose post-tokenization length is smaller than T_{Min}=2 or higher than T_{Min}=100 are discarded. The length filtering contributes to the removal of a few misaligned sentences and a more computationally friendly dataset.

4.3. Model Configuration

The Transformer architecture (Vaswani et al., 2017) is adopted across all translation tasks and it is configured with a hidden size of 128. The intermediate size is 256 (F=256), one attention head, three encoders and decoders (N=3). The model was designed with limited representation abilities to prevent overfitting on the limited training size. Preliminary experiments showcased that a hidden dimension greater than 384 easily led to overfitting. Whereas a hidden dimension of 256 and increasing the number of layers N did not lead to notable differences.

[1] Endangered languages are also low-resourced, but we will denote them as EL for disambiguation.

[2] https://en.wal.unesco.org

[3] http://hdl.handle.net/11022/0000-0007-CAE5-3

[4] https://siberian-lang.srcc.msu.ru/en/

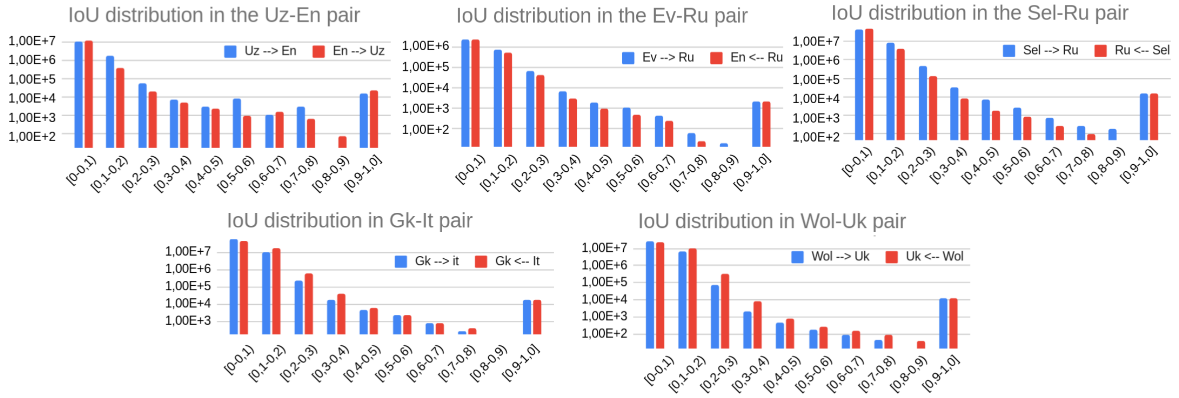

Figure 5: IoU histogram across different translation datasets.

4.4. Training and Evaluation Details

The model is first trained in the pre-training stage with the Adam (Kingma and Ba, 2014) optimizer (β_1=0.99, β_2=0.98), a fixed learning rate of 1e-3 and a sentence batch size of 256. We set α=1.2, γ=0.0001 and for each language we train for a variable number of iterations. In particular, the model is pre-trained up to 5 epochs in the case of Gk-It, Uz-En, Sel-Ru, Wol-Uk and 25 in the case of Ev-Ru. During the training on the translation tasks, the model is trained for 300 epochs with a token batch size of 4096 and the learning rate is set as follows:

$$\gamma(it) = H^{-0.5} \cdot min(it^{-0.5}, it \cdot w^{-1.5}) \qquad (7)$$

where it is the number of iterations, whereas w denotes the warmup and is set to 400. During both pre-training and translation training the Cross-Entropy loss is adopted with a Label Smoothing coefficient of 0.1. During evaluation, predictions are generated with the Beam Search Algorithm using a beam size of 4 and the BPE de-tokenization is applied to the result. In the translation task, the BLEU (Papineni et al., 2002) score is reported as the evaluation metric in particular, we adopt the SacreBLEU (Post, 2018) library[5]. To evaluate each translation, we report the mean and standard deviation of the 10 BLEU scores periodically sampled from the last 100 training epochs.

Table 1 reports the number of training samples available for the pre-training strategy. However, we note that involving all possible pairings in the augmented dataset can be too expensive. To reduce the training cost and maximize the information carried in each iteration batch, we select only the pairs characterized by an Intersection over Union (IoU) greater than 0.1. Given two samples in source and target language X_i and Y_j we compute the $IoU(X_i, Y_j)$ as the number of unique matching tokens between Y_j and Y_i divided by the sum of unique tokens in Y_j and Y_i. By doing so, we discard a significant portion of the augmented dataset as can be seen from Figure 5. For instance, in the case of Uz→En, we consider only 12.7 % of all possible pairings. However, the discarded portion is made of samples that contribute the least according to the loss defined in our training strategy.

5. Results

5.1. USKI Results

In Table 2, in the second column, we showcase the performances of the baseline architecture. In the third column, we report the scores of the model when our proposed pre-training is applied before the standard translation training. In Table 2, it can be observed that the pre-training method increases the model performances across all languages. The maximum improvement of 1.83 BLEU is observed in the case of Ru←Sel, whereas the lowest one, of 0.37 BLEU, is showcased in the Wol←Uk case. The magnitude of the improvements is variable and does not seem to relate to the data size, the language family, or the similarity between the source and target language family. For instance, the dataset of the Uz-En case is double the amount of Ev-Ru, yet the increase is similar.

In Table 3 it is reported the accuracy computed over the first KeyTokens in the vocabulary, ranked by the IDF-term. In particular, we compute KeyTokens Accuracy in the following way. Let $T = \{(X_i, Y_i)\ \forall i \in \{1, \ldots, |T|\}\}$ be the translation test set. We denote K_n the set of the first n tokens from the vocabulary, sorted according to the weight function of Equation 4 in the definition of KeyTokens prediction pre-training stage (Section 3.1). We define S as the set of translation pairs $(X_i, Y_i) \in T$, such that Y_i is made of at least one

[5] SacreBLEU signature: BLEU+case.mixed+lang.[source-lang]-[target-lang]+numrefs.1+smooth.exp+tok.13a+version.2.0.0

Task	Baseline	w/ USKI	$\delta\uparrow$
Sel→Ru	7.05±0.50	8.19±0.38	1.14
Sel←Ru	4.15±0.45	5.98±0.34	1.83
Ev→Ru	6.58±0.48	7.26±0.39	0.68
Ev←Ru	7.36±0.85	8.65±0.90	1.29
Gk→It	5.91±0.10	6.23±0,.10	0.32
Gk←It	4.43±0.07	5.58±0.17	1.15
Uz→En	18.82±0.74	19.94±0.38	1.12
Uz←En	18.40±0.43	19.07±0.38	0.67
Wol→Uk	4.60±0.13	5.00±0.12	0.40
Wol←Uk	8.25±0.05	8.62±0.14	0.37

Table 2: Model's baseline test set BLEU scores across different language pairs. The source of the arrow denotes the source language and points to the target language of the translation task. Values represent the average of 10 BLEU scores sampled periodically over the last 100 epochs of training ± the standard deviance. $\delta\uparrow$ denotes the difference between the baseline and the model pre-trained with our proposal.

token contained in K_n. The KeyTokens accuracy of a model over the tokens K_n for the dataset T, is given by: $\sum_{i=1}^{|T|} C_i/|S|$ where C_i equals one if the model correctly predicted at least one token that was contained in both K_n and the ground-truth translation Y_i, and zero otherwise. In other words, the KeyTokens accuracy is computed over the n tokens on which USKI focused the most, denoted by K_n. Let A_{250}, A_{500} and A_{1000} denote the KeyTokens Accuracy over the first KeyTokens in the 250, 500 and 1000 rankings respectively (K_{250}, K_{500} and K_{1000}). In Table 3 we observe that while the majority of the language setups showcase an increase of accuracy over the KeyTokens, there are few cases, as depicted in Figure 6, in which the increase is not that significant or even worse, such as the A_{250} in Sel←Ru and A_{1000} in Ev→Ru and Gk→Uk. These instances are important to highlight the difficulty in providing an unbiased interpretation of the accuracy improvements. From one perspective, the increase is expected as a result of the pre-training formulation. On the other hand, the accuracy increase may be simply a consequence of the performance increase, and the case of Ev→Ru and Gk←Uk underlines possibly the phenomena known as "catastrophic forgetting" in Neural Networks. Either way, we conclude that, overall, the pre-training provides a more profitable and robust starting point in the latent space for the translation task, but we hypothesize that, in some instances, the model finds better results in partially forgetting the initial focus over the KeyTokens.

Task	PT	A_{250}	A_{500}	A_{1000}
Sel→Ru	✓	10.00%	15.13%	32.37%
	✗	7.00%	11.51%	31.58%
Sel←Ru	✓	13.45%	10.44%	17.85%
	✗	16.34%	10.14%	16.09%
Ev→Ru	✓	0.7%	1.0%	28.02%
	✗	0.0%	0.5%	30.0%
Ev←Ru	✓	0.71%	1.22%	3.03%
	✗	0.71%	0.61%	2.11%
Gk→It	✓	19.49%	22.80%	23.78%
	✗	18.43%	22.48%	24.28%
Gk←It	✓	4.95%	17.0%	20.67%
	✗	3.96%	14.8%	18.70%
Uz→En	✓	12.63%	18.76%	23.42%
	✗	9.47%	16.62%	22.45%
Uz←En	✓	30.76%	32.03%	29.92%
	✗	29.23%	29.68%	27.96%
Wol→Uk	✓	16.05%	16.45%	28.34%
	✗	14.04%	14.37%	27.39%
Wol←Uk	✓	18.01%	17.35%	32.45%
	✗	17.08%	16.56%	32.29%

Table 3: KeyTokens test set accuracy comparison between a model pre-trained first on the Augmented Dataset and one trained directly on the translation task. A_n denote the accuracy over the first n KeyTokens based on the IDF term. PT indicates whether it is pre-trained with USKI.

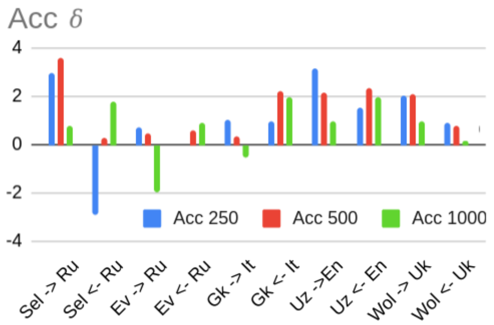

Figure 6: Accuracy differences between the baseline models and pre-trained ones on the test set KeyTokens. A_n denote the accuracy over the first n KeyTokens ranked according to the IDF term.

5.2. KeyTokens Granularity

The number of merge operations in the sub-word tokenization method (in this case the BPE) is proportional to the vocabulary size and affects the granularity of sentences. This directly impacts the size of the pre-training set because it affects the number of aligned sentences with an IoU higher than 0.10. In Table 4 we found out that the pre-training set is inversely proportional to the vocabulary size. However, the increase in the pre-training dataset size (denoted by γ) does not lead to an increase in improvements (δ). This suggests that focusing on KeyTokens made of more complete sub-words or entire words is less effective compared to the case of KeyTokens made of shorter sub-words. This factor might outweigh the increase in the pre-training size. This result provides an explanation of the significant improvement difference, in Table 2, between the translation from the first language into the second and vice versa within the same translation pair, since sub-words are not distributed equally between two languages.

$\lvert V \rvert$	γ	Baseline	w/ USKI	$\delta \uparrow$
4356	1.7M	19.07	19.64	0.57
7983	2.2M	19.68	20.08	0.40
11385	2.5M	19.49	19.95	0.46
13901	2.6M	20.06	20.29	0.23

Table 4: BLEU score on Uz→En task for different vocabulary sizes. $\lvert V \rvert$ denotes the vocabulary size. γ denotes the pre-training dataset size. $\delta \uparrow$ denotes the difference between the fourth and third columns. Results are obtained by averaging the BLEU scores across 5 seeds.

5.3. Comparison With mBART

In this Section, we compare our results with a popular large pre-trained language model called mBART (`mbart-large-cc25`) (Liu et al., 2020), which is a pre-trained model over a massive amount of multi-lingual data. Additionally, we provide an example of how to integrate our method into these large models. In our experiments, the model is fine-tuned on each translation task for 50 epochs with a learning size of 0.0001, weight decay of 0.01, and batch size of 32.

In Table 5 we report a comparison between our best results and mBART. As expected, our proposed architectures perform significantly worse across almost all translations. This highlights the fact that our models were designed for experimental development instead of competitiveness as they are not trained on much less training data and are 440× smaller. Despite worse results, USKI can be integrated and applied to mBART. In particular, to prevent catastrophic forgetting, our method can be applied on a set of properly designed additional layers of learnable parameters (Hu et al., 2021) placed on top of each layer in the language model, so that it can benefit from both the massive multi-lingual knowledge and the KeyTokens pre-training during the fine-tuning task.

Task	Baseline	w/ USKI	mBART
Sel→Ru	7.05	8.19	18.50
Ev→Ru	6.58	7.26	26.03
Gk→It	5.91	6.23	9.12
Uz→En	18.82	19.94	20.88
Wol→Uk	4.69	5.00	6.24

Table 5: BLEU score comparison between the baseline, our best pre-training results, and mBART.

6. Conclusion and Future Works

In this work, we tackled the problem of data scarcity in NMT, with a particular focus on endangered and low-resource languages. In particular, in USKI, we first proposed to construct an augmented version of the initial dataset that exhibits the property of being made of square the number of training elements. We then proposed a pre-training method, based on matching tokens, called KeyTokens, to leverage unaligned sentences in the first dataset. Results showcased that our method led to an average improvement of 0.9 BLEU across all the selected translation tasks despite using about one-tenth of the entire unaligned corpus.

The main limitations of our proposed method are twofold. First, although the pre-training can be made of millions or hundreds of thousands of samples, they are trivially not as effective as the aligned sentences for the translation task. On top of that, most tokens in each sentence are still underutilized in the KeyTokens pre-training formulation. Future works will focus on developing better methods to increase the value of each unaligned pair. Second, the pre-training step can be time-consuming if not carefully addressed, for this reason, we trained only over a smaller portion of the entire dataset. In future works, existing strategies such as Bucketing can be adapted to increase efficiency.

In conclusion, while the performance improvement generated by our solution alone is not impressive, especially when compared with other standard approaches that involve Large Language Models (Liu et al., 2020; Raffel et al., 2020; Costa-jussà et al., 2022; Junczys-Dowmunt et al., 2018), it showcases a small example of the capabilities and potential of leveraging unaligned sentences. Overall, we believe this work makes a step towards the development of important but currently under-utilized resources.

7. Bibliographical References

Antonios Anastasopoulos, Marika Lekakou, Josep Quer, Eleni Zimianiti, Justin DeBenedetto, and David Chiang. 2018. Part-of-speech tagging on an endangered language: a parallel Griko-Italian resource. In *Proc. COLING*.

Andoni Azpeitia, Thierry Etchegoyhen, and Eva Martınez Garcia. 2018. Extracting parallel sentences from comparable corpora with stacc variants. In *Proceedings of the 11th workshop on building and using comparable corpora*, pages 48–52.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Alexandre Berard, Ioan Calapodescu, and Claude Roux. 2019. Naver labs europe's systems for the wmt19 machine translation robustness task. *arXiv preprint arXiv:1907.06488*.

Ondřej Bojar and Aleš Tamchyna. 2011. Improving translation model by monolingual data. In *Proceedings of the sixth workshop on statistical machine translation*, pages 330–336.

Stanley F Chen. 1993. Aligning sentences in bilingual corpora using lexical information. In *31st Annual Meeting of the Association for Computational Linguistics*, pages 9–16.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Ying Ding, Junhui Li, Zhengxian Gong, and Guodong Zhou. 2021. Improving neural sentence alignment with word translation. *Frontiers of Computer Science*, 15:1–10.

Sufeng Duan, Hai Zhao, and Dongdong Zhang. 2023. Syntax-aware data augmentation for neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv:1705.00440*.

Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Fei Gao, Jinhua Zhu, Lijun Wu, Yingce Xia, Tao Qin, Xueqi Cheng, Wengang Zhou, and Tie-Yan Liu. 2019. Soft contextual data augmentation for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5539–5544.

Cyril Goutte. 2009. *Learning machine translation*. MIT Press.

Mika Hämäläinen. 2021. Endangered languages are not low-resourced! *arXiv preprint arXiv:2103.09567*.

Jindřich Helcl, Jindřich Libovickỳ, and Martin Popel. 2019. Cuni system for the wmt19 robustness task. *arXiv preprint arXiv:1906.09246*.

Cong Duy Vu Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *2nd Workshop on Neural Machine Translation and Generation*, pages 18–24. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jia Cheng Hu, Roberto Cavicchioli, and Alessandro Capotondi. 2023. A request for clarity over the end of sequence token in the self-critical sequence training. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 14233 LNCS:39 – 50.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, et al. 2018. Marian: Fast neural machine translation in c++. *arXiv preprint arXiv:1804.00344*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Seiichiro Kondo, Kengo Hotate, Masahiro Kaneko, and Mamoru Komachi. 2021. Sentence concatenation approach to data augmentation for neural machine translation. *arXiv preprint arXiv:2104.08478*.

Yu Li, Xiao Li, Yating Yang, and Rui Dong. 2020. A diverse data augmentation strategy for low-resource neural machine translation. *Information*, 11(5):255.

Zhenhao Li and Lucia Specia. 2019. Improving neural machine translation robustness via data augmentation: Beyond back translation. *arXiv preprint arXiv:1910.03009*.

Qi Liu, Matt Kusner, and Phil Blunsom. 2021. Counterfactual data augmentation for neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 187–197.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Maria Brykina, Svetlana Orlova and Beáta Wagner-Nagy. 2018. *INEL Selkup Corpus. Version 0.1 Publication date 2018-12-31. Archived in Hamburger Zentrum für Sprachkorpora.* Beáta Wagner-Nagy, Alexandre Arkhipov, Anne Ferger and Daniel Jettka. Timm Lehmberg (eds.). 2018.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel bible corpus. *Oceania*, 135(273):40.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. Unesco.

Anna Mossolova and Kamel Smaïli. 2022. The only chance to understand: machine translation of the severely endangered low-resource languages of eurasia. In *The Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT), COLING 2022*.

Yuta Nishimura, Katsuhito Sudoh, Graham Neubig, and Satoshi Nakamura. 2018. Multi-source neural machine translation with data augmentation. *arXiv preprint arXiv:1810.06826*.

Olga Kazakevich et. al. . *Development of the web-site 'Minority languages of Siberia as our cultural heritage'*. Laboratory for Computational Lexicography, Research Computing Centre, Lomonosov Moscow State University.

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.

Wei Peng, Chongxuan Huang, Tianhao Li, Yun Chen, and Qun Liu. 2020. Dictionary-based data augmentation for cross-domain neural machine translation. *arXiv preprint arXiv:2004.02577*.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Julia Sallabank and Peter K Austin. 2011. Endangered languages. *The SAGE Handbook of Sociolinguistics. London: Sage Publications Ltd*, pages 496–513.

Víctor M Sánchez-Cartagena, Miquel Esplà-Gomis, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. 2021. Rethinking data augmentation for low-resource neural machine translation: A multi-task learning approach. *arXiv preprint arXiv:2109.03645*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Jason Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics*, pages 403–411.

Amane Sugiyama and Naoki Yoshinaga. 2019. Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the fourth workshop on discourse in machine translation (DiscoMT 2019)*, pages 35–44.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018. Switchout: an efficient data augmentation algorithm for neural machine translation. *arXiv preprint arXiv:1808.07512*.

Renjie Zheng, Hairong Liu, Mingbo Ma, Baigong Zheng, and Liang Huang. 2019. Robust machine translation with domain sensitive pseudo-sources: Baidu-osu wmt19 mt robustness shared task system report. *arXiv preprint arXiv:1906.08393*.