Efficient Unsupervised NMT for Related Languages with Cross-Lingual Language Models and Fidelity Objectives

Rami Aly[1], Andrew Caines[2], Paula Buttery[2]
[1] Computer Laboratory, University of Cambridge, U.K.
[2] Computer Laboratory & ALTA Institute, University of Cambridge, U.K.
{rami.aly|andrew.caines|paula.buttery}@cl.cam.ac.uk

Abstract

The most successful approach to Neural Machine Translation (NMT) when only monolingual training data is available, called unsupervised machine translation, is based on back-translation where noisy translations are generated to turn the task into a supervised one. However, back-translation is computationally very expensive and inefficient. This work explores a novel, efficient approach to unsupervised NMT. A transformer, initialized with cross-lingual language model weights, is fine-tuned exclusively on monolingual data of the target language by jointly learning on a paraphrasing and denoising autoencoder objective. Experiments are conducted on WMT datasets for German→English, French→English, and Romanian→English. Results are competitive to strong baseline unsupervised NMT models, especially for closely related source languages (German) compared to more distant ones (Romanian, French), while requiring about a magnitude less training time.

1 Introduction

While traditional end-to-end neural machine translation (NMT) approaches have shown highly promising results when abundant parallel data is available (Barrault et al., 2019), the task remains a considerable challenge when only monolingual training data is available, also called *unsupervised* MT (Artetxe et al., 2018a; Lample et al., 2018a). Unsupervised NMT systems tend to combine back-translation (Sennrich et al., 2016a) with cross-lingual embeddings (Artetxe et al., 2018b; Lample et al., 2018a,b) or, more recently, with weights of a pre-trained cross-lingual language model (XLM) (Conneau and Lample, 2019). Back-translation uses noisy translations, generated by a source-to-target model, as input for a target-to-source model (and vice versa). Although shown to perform well if plenty of monolingual data is available, back-translation is computationally very expensive. It is also highly inefficient as the inference performed to generate the noisy translations is of sequential nature, slowing down the training substantially.

This paper presents a novel unsupervised NMT method that does not require back-translation. Instead, it jointly fine-tunes a transformer (Vaswani et al., 2017), initialized with weights of cross-lingual language models (Conneau and Lample, 2019), on a denoising autoencoder (Vincent et al., 2008; Artetxe et al., 2018b) and paraphrasing objective exclusively on data in the target language. The alignment of the languages in the transformer's encoder means we can learn similar hidden representations for sentences of similar meaning but from different languages. The decoder, fine-tuned to generate a sentence in the target language, can thus generate a translation based on the encoder's representation of a source-language input sentence. Naturally, this method is more suitable for related languages, as the alignment of languages and hidden representations in the cross-lingual encoder is of particular importance to this approach.

Our experiments with the WMT datasets for German→English, French→English, and Romanian→English (Bojar et al., 2016) show that the proposed approach outperforms competitive models – namely Artetxe et al. (2018b), Lample et al. (2018a) and Lample et al. (2018b) – highlighting that the alignment quality achieved by the high-quality cross-lingual language model as a translation signal is superior to aligned embeddings and back-translation. Results for German are substantially higher than for French and Romanian, highlighting that our approach works particularly well for more closely related languages. While achieving competitive results, the proposed approach is substantially more efficient. It converges much quicker while requiring less than 50% time per

Proceedings of the 8th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects, pages 49–59

epoch during fine-tuning which results in around a magnitude less floating point operations for the proposed approach than for an equivalent setup when using back-translation. We further show that the paraphrasing objective improves translation quality considerably compared to using the autoencoder objective in isolation.

2 Method

Given an input sequence X^s in source language s the objective is to generate a sequence Y^t in the target language t, which is semantically equivalent. A model NMT$^{s \rightarrow t}$ models the target function

$$arg\ max_{\mathcal{Y}^t} \prod_{u=1}^{m} p(y_u^t | y_{<u}^t; x_1^s, ..., x_n^s), \quad (1)$$

with $\mathcal{Y}^t$ being the set of all possible sequences in the target language. This paper focuses on the transformer model (Vaswani et al., 2017) to solve Equation 1. The transformer consists of an encoder and a decoder module: both are initialized with weights $W^{s \leftrightarrow t}$ of a cross-lingual language model and a shared subword vocabulary to align languages s and t (§ 2.3).

$$H^s = \text{ENC}_W(X^s) \quad (2)$$

$$\hat{Y}^t = \text{DEC}_W(H^s) \quad (3)$$

The encoder transforms the input into a latent space while the decoder iteratively generates the output sequence $\hat{Y}^t$.

2.1 Fine-tuning Approach

We propose a fine-tuning approach for this model that solely relies on monolingual data of the target language and the alignment W between s and t. Due to the cross-lingual weights W, the hidden representation of the encoder for both source and target language are aligned and thus expected to be similar[1]:

$$\text{ENC}_W(X^s) \sim \text{ENC}_W(Y^t) \quad (4)$$

This assumption is the essence to our approach and it applies more to closely related source and target languages than to more distant ones. Based on the assumption, our hypothesis is that it is sufficient to train the initialized encoder and decoder on sentence generation tasks in only the target language. More specifically, we explore meaning-preserving training objectives, that focus on monolingual sentence generation objectives so that the meaning of the input sequence is preserved for the generated sequence. We call these *fidelity* objectives. Thus, given a sentence P^t in the target language (specified in § 2.2) with very similar/identical meaning to a sentence Q^t of the same language in the monolingual training data, we optimize the NMT model by calculating the cross-entropy loss of the fidelity task over the shared subword vocabulary:

$$\mathcal{L}_{\text{fid}} = - \sum_{<P^t, Q^t> \in D_{\text{fid}}} log(p(Q^t | P^t)), \quad (5)$$

where D_{fid} is a fidelity training dataset. When confronted with an input sequence X^s in the source language during inference, $ENC_W(X^s)$ generates a hidden representation H^s, which is expected to be similar to the representation H^t for a semantically identical sentence in the target language due to the cross-lingual LM weights W. The similarity between H^s and H^t trains the decoder to generate a meaning-preserving sequence based on the hidden representation of the encoder and enables $\text{DEC}_W(H^s)$ to generate a sentence in the target language similar to $\text{DEC}_W(H^t)$ while preserving the meaning of the source sentence X^s.

2.2 Fidelity Objectives for Fine-tuning

We focus on two learning objectives that are solved by the model for the target language: a *denoising autoencoder* (Artetxe et al., 2018b) and *paraphrase generation* in the target language. The objectives are illustrated in Figure 1 and are learned using Eq. 5.

Denoising Autoencoder We use a straightforward autoencoder objective (Vincent et al., 2008; Artetxe et al., 2018b) to fine-tune the model so that it reconstructs the input Q^t from a noisy version P^t_{denoise} (the noise prevents the model from simply copying the input). We add noise to Q^t by either swapping, omitting or replacing words with a padding token. The number of noise operations on a sentence is a hyperparameter. The denoising autoencoder objective is used in most unsupervised NMT systems in combination with back-translation (Conneau and Lample, 2019), however, in these settings, the autoencoder objective only serves the

[1] This expectation is based on results for zero-shot classification to highlight the sentence similarity across different languages as well as the high cosine-similarity between word translation pairs shown in Conneau and Lample (2019). Languages more similar to the one the model has been trained on have higher sentence similarity and thus achieved higher scores in their experiments.

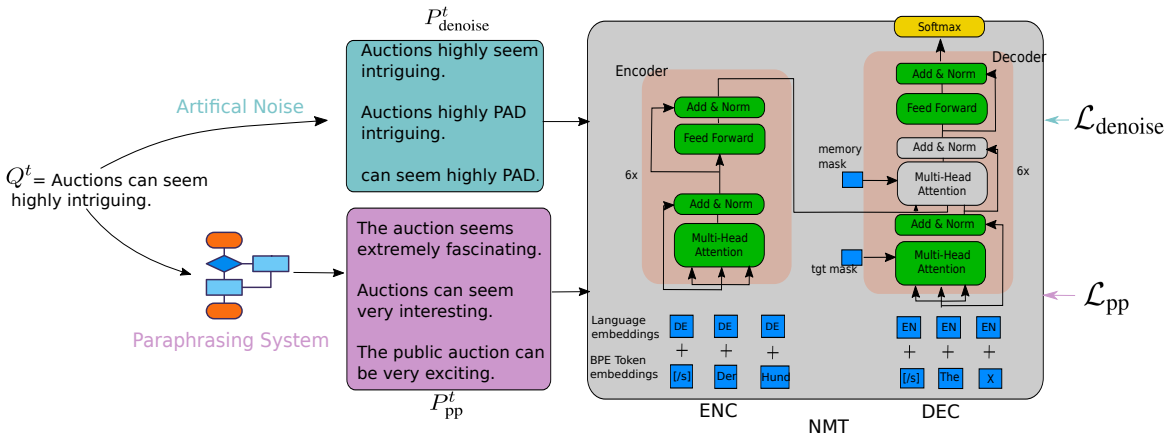

Figure 1: Illustration of the joint fine-tuning approach with denoising autoencoder and paraphrasing objectives. Note that for each sentence multiple paraphrases and noisy inputs are generated for fine-tuning the transformer.

function of making the model familiar with noise in the input that back-translated texts naturally have. Explicitly using alignment in the encoder for unsupervised translation in the way we propose has not yet been explored.

Paraphrasing In previous work the autoencoder objective has only been shown to be effective in combination with back-translation as a means to make the model familiar with noise in the input (Conneau and Lample, 2019; Artetxe et al., 2018b). This might be attributed to the limitation of the denoising autoencoder that the added noise is artificial and results in ungrammatical sentences. Thus, we additionally explore the task of reconstructing a sentence Q^t from a paraphrased version P^t_{pp} which is complemented by the denoising task. Automatically generated paraphrases are expected to be more grammatical and diverse than the simple rule-based variations used in the autoencoder objective.

2.3 Initialization

To initialize the weights W of the word embeddings, the encoder, and decoder of the transformer model we use the weights of the state-of-the-art pre-trained cross-lingual language model (XLM) of Conneau and Lample (2019). The pre-trained XLM is essentially the encoder part of the transformer model trained on the *masked language modelling* (MLM) objective on a stream of text. Furthermore, XLM uses language embeddings to assist the network in recognizing different languages. Finally, XLM and subsequently the translation model make use of subword tokenized inputs, specifically byte-pair encoding (BPE) (Sennrich et al., 2016b) to reduce the vocabulary size as it is shared by all languages in the model.

3 Evaluation

3.1 Experimental Setup

Data: Our approach uses WMT 2007-8 training data for German→English, as well as French→English (Callison-Burch et al., 2007, 2008), and the training data of WMT 2015 for Romanian→English (Bojar et al., 2015). All languages are Indo-European and therefore related to some extent, but there are differences of relatedness. French and Romanian are Italic whereas German and English are West Germanic: we therefore hold German and English to be the most closely related of our language pairs, with much that is similar in terms of lexicon and morpho-syntax and a recent shared history; followed by French and English (due to extensive lexical borrowings from language contact), and lastly Romanian→English which is the most distantly related language pair.

Note that for all language pairs, our approach converged in the first epoch, with only a few steps difference. Therefore, our approach was implicitly fine-tuned on comparable amount of data for all language pairs. Similar to previous unsupervised MT approaches, we evaluate the models on the WMT 2016 test sets for German→English and Roman→English (Bojar et al., 2016) and use the WMT 2014 test set for French→English (Bojar et al., 2014).

Implementation details: The experiments were run in Python 3.7 and Python 3.5 for the NMT model and paraphrasing system, respectively. The openly accessible repository of the work described

Model	de-en	fr-en	ro-en
(1) Lample et al. (2018a)	13.3	**14.3**	–
(2) Artetxe et al. (2018b)	–	**15.6**	–
(3) Lample et al. (2018b) Transformer	21.0	**24.2**	19.4
(4) Lample et al. (2018b) Transformer + PBSMT	25.1	**27.7**	23.9
(5) Conneau and Lample (2019)	**34.3**	33.3	31.8
(6) Transf. + autoencoder	**20.9**	18.9	18.7
(7) Transf. + autoencoder + paraphrases	**24.2**	22.1	21.2

Table 1: BLEU scores on test sets. (1)–(5) are taken from the respective papers. (6) and (7) refers to the proposed approach. **Bold** numbers indicate the language pair on which the model performs best.

in Conneau and Lample (2019)[2] was used as the basis for the implementation of this paper, and we use their provided pre-trained models [3]. The data is pre-processed into BPE tokens using FastBPE[4] with a vocabulary size of $60,000$. All models were fine-tuned on one GPU (NVIDIA Tesla P100). Since our experiments are conducted in the exact same ecosystem as Conneau and Lample (2019); Lample et al. (2018b), our results are directly comparable to theirs.

We use paraphrases created by the model proposed by Wieting et al. (2017)[5], due to its open-source access. It uses a Seq2Seq architecture to back-translate bilingual sentence pairs. For each sentence in the training data, the most probable three paraphrases are used. It would be preferable to use a fully unsupervised paraphrasing system, e.g. (Roy and Grangier, 2019). Nonetheless, we argue that the employed system does not violate the assumption of an unsupervised NMT system, since the paraphrases are only generated for the target language. Thus, while the target language must be part of at least one bilingual corpus, the source language can be arbitrary (as long as the cross-lingual language model between the source and target language exists).

Hyperparameters: Since for unsupervised NMT the assumption is made that *only* monolingual data is available, selecting the model or hyperparameters on a parallel dataset contradicts this premise. Therefore, the default hyperparameter settings from related work (Conneau and Lample, 2019) and the underlying XLM model are used[6]. The number of training epochs is based on the perplexity scores on the WMT 2013 test sets or, for Romanian, the WMT 2015 development set.

Evaluation metrics While perplexity is used as the metric during training, the performance on the test sets are reported on the commonly used BLEU metric (Papineni et al., 2002), specifically the MOSES evaluation script (Hoang and Koehn, 2008).

4 Results

Table 1 shows the BLEU scores on the test sets for source and target language. The scores for model (1) to (5) are taken from the respective papers with our re-evaluations producing almost identical results when using the respective openly accessible repository. Although the proposed model (7) uses exclusively data of the target language, it performs competitively than the sophisticated approaches in (1), (2), (3), and (4) which all use back-translation (Lample et al., 2018a,b; Artetxe et al., 2018b), especially for German→English. Since (3) uses the transformer architecture as well, the results highlight that the alignment achieved by the high-quality cross-lingual language models of XLM is superior to the gains achieved by the back-translation algorithm. When combining both back-translation and XLM, the results can, however, be further improved substantially, as shown by (5).

While previous attempts to using the denoising objective without back-translation resulted in unusable results (Artetxe et al., 2018b), we observe that our model performs reasonably well already when

[2] https://github.com/facebookresearch/XLM

[3] mlm_ende_1024, mlm_enfr_1024, and mlm_enro_1024

[4] https://github.com/glample/fastBPE

[5] https://github.com/vsuthichai/paraphraser

[6] emb_dim: 1024, #layers 6, #heads 8, dropout: 0.1, attention_dropout: 0.1, tokens_per_batch: 2000, optimizer: adam_inverse_sqrt. Denoising autoencoder parameters: word_shuffle: 3, word_dropout: 0.1, word_blank: 0.1, lr: $7 \cdot 10^{-4}$

Model	(3)	(5)	(7)
time per step (minutes)	89.55	92.24	29.54
total cost (FLOPs)	$1.78 \cdot 10^{18}$	$6.31 \cdot 10^{17}$	$8.07 \cdot 10^{16}$
cost (FLOPs) @ 100K	$5.10 \cdot 10^{17}$	$3.15 \cdot 10^{17}$	$7.47 \cdot 10^{16}$

Table 2: Comparison of average training time between different methods on a single Tesla P100 when fine-tuning for German→English. A step consists of 100K samples. Total cost reports floating-point operations for the entire training process. We use the value 9.5 TFLOP/s for the P100. Costs are shown when trained on the entire training corpus and when exclusively training on 100K sentences. The generation of paraphrases is included in (7) total cost.

training exclusively on this objective (6). Joint modelling of the paraphrasing and denoising objective (7) improves scores of the unsupervised system by about 3 BLEU points over (6).

Furthermore, our approach (7) achieves particularly high results when a source language is related (German) to the training language, compared to a more distant one (French, Romanian). While our model (7) outperforms (3) for German→English by around 3 BLEU points, it scores 2.1 points less for French→English. Moreover, scores between (7) and (3) are only comparable for German, while (3) performs much better for both the less related source languages. This observation is even amplified when the source language is Romanian. Our model appears to be more susceptible to the relatedness of the source language than (5) which uses the same cross-lingual weights, scoring 2.1 points less on French than German, compared to only 1.0 for (5). While our approach solely relies on the alignments based on these weights, the back-translation of (5) adds an important signal especially for more distant languages. This confirms the assumption made for our model: the BLEU score of our model is particularly high for the closely related German→English, compared to the more distant language pairs.

5 Efficiency

One major advantage of the proposed method over existing methods is its efficiency in terms of computational time. In Table 2 we report the average time and cost for the German→English experiments. The time to fine-tune on a full epoch is measured in the same environment under identical conditions. For our model, the measured times include the computational cost to generate the paraphrases[7]. We

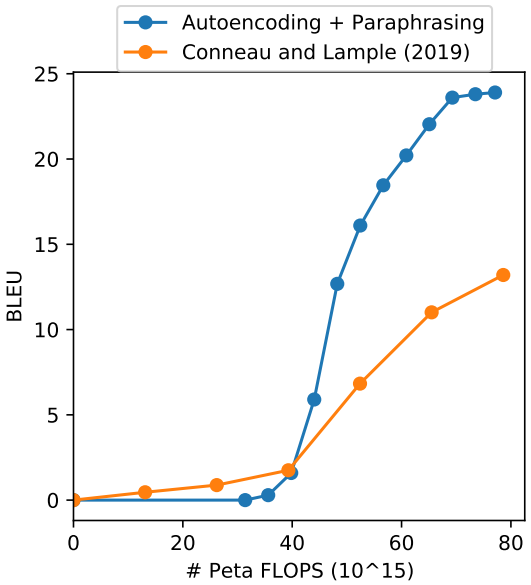

Figure 2: BLEU scores over number of FLOPs. FLOPs include the computation of the paraphrases before our model is trained.

find that by using the paraphrasing objective instead of back-translation the computational time can be reduced by a factor of three. Moreover, the entire training process requires around an order of magnitude less floating point operations due to our approach converging much quicker.

Our approach outperforms Lample et al. (2018b) while being much more efficient, however, one might suggest that the shown efficiency of the proposed model can also be achieved with the better scoring Conneau and Lample (2019) by trading off some of its performance advantages. Therefore, we explored to which extent either less training time or training data to achieve faster convergence improves its efficiency. Regarding training times, our

[7]Training the paraphrasing generator (Wieting et al., 2017) is not included in the costs as the employed paraphrasing model for English is already openly accessible. Moreover, the paraphrasing model uses a (Bi-LSTM) that was trained for 3 epochs on only 24,000 sentence pairs, which is less than 2% of the translation data used for the NMT models.

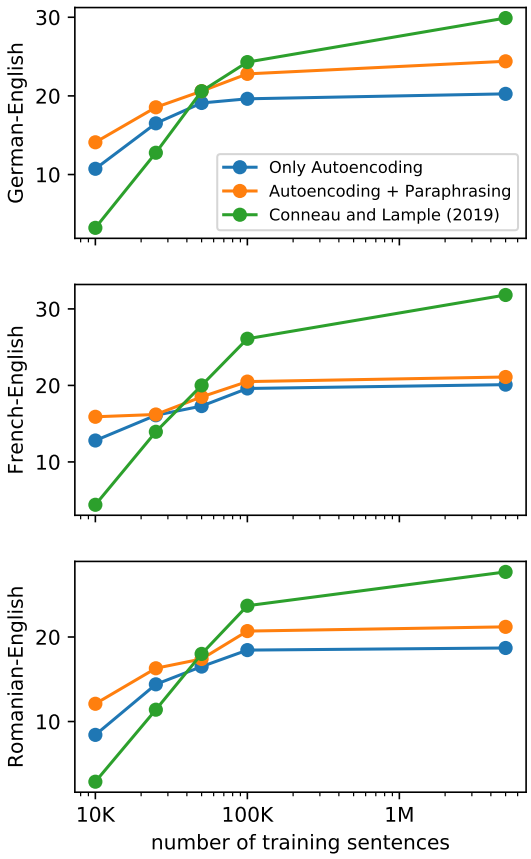

Figure 3: BLEU scores of the proposed model as well as the state-of-the-art approach of Conneau and Lample (2019) for varying number of monolingual training sentences: 10K, 25K, 50K 100K, and 5M.

proposed approach (7) scored consistently higher[8] than the other models until it's convergence, see figure 2. Only after our model converged, the model of Conneau and Lample (2019) surpasses its scores. Thus, stopping the training process earlier would not lead to better efficiency than our model. We then also analyzed the models' scores when modifying the amount of data used for fine-tuning. Results using 10K, 25K, 50K, 100K, and 5M training sentences are shown in Figure 3. As seen, when using little training data, Conneau and Lample (2019) performs much worse than the proposed model since errors and noise from translating into one direction are propagated when translating back. The less monolingual data used, the stronger the effect of this error-propagation issue. The back-translation model of Conneau and Lample (2019) starts to outperform the proposed model starting from 100K training sentences. Yet, even in this setting the proposed model remains much more efficient, as seen in the column *cost (FLOPs) @ 100K* in Table 2. The back-translation model is thus required to train on more data and ultimately longer until convergence and cannot achieve similar efficiency with scores comparable to our approach.

[8]Excluding the initial phase as our model computes the paraphrases first. Since our model catches up at a BLEU of around 1.7, we ignored this special case.

6 Model Analysis & Discussion

Alignment and translation quality: To investigate whether the alignment extends from a word-level to a phrase or even sentence level, a qualitative analysis was conducted. Example reference texts and respective translations for German→English are shown in Figure 4. It can be seen that the translation quality is high for shorter sentences but it declines with increasing sentence length. Many words and phrases are translated correctly, which generally leads to preservation of sentence meaning for simple sentences in closely-related languages like English and German. While many simple phrases are grammatical, many longer more complex structures are not.

Furthermore, the model hallucinates content, especially when confronted with numbers and named entities. For example, in sentence 5 the model generates a made-up destination *to northern Croatia* while *New Lloyd* in sentence 6 also never occurs in the source sentence. In the translation for sentence 3, the name is simply omitted and the currency, as well as the amount, is wrong. This observation can be transferred to numbers as well: in sentence 3, the number of 7.5 million was changed to 67.5 million. In an extreme example, the model hallucinates an entire clause for sentence 7. We also observed some artifacts in the output: in sentence 7, *oberflächlich* is translated into *Oberpublic*, instead of *superficial*, merging an German and English term.

Wrong translations frequently contain words that are still closely related to the correct translation. For instance, in sentence 2 the model generates the former prime minister of Israel *Olmert*, instead of *Netanyahu* and translates *meets* instead of *receives*. Or in sentence 5, the model translated into *hurdles* instead of *obstacles*. These very subtle differences are difficult for the model to capture.

Paraphrase quality Although the paraphrasing objective improves the performance of the system substantially, the generated paraphrases are still far from optimal. Figure 5 shows example paraphrases.

X-1: Die Vorbereitung lief gut.
Y: The preparation went well.
$\hat{Y}$**:** The training season went well.

X-2: Obama empfängt Netanyahu.
Y: Obama receives Netanyahu.
$\hat{Y}$**:** Obama meets Olmert.

X-3: Spaniens Nationaltorhüter Iker Casillas hat für 7,5 Millionen Euro seine Hände versichern lassen.
Y: Spain's national goalkeeper Iker Casillas insured his hands for 7.5 million euro.
$\hat{Y}$**:** Spain's King has asked for £ 67.5 million to reinsure his hands.

X-4: Wir wollten gewinnen und wir haben gewonnen, obwohl ich es bedaure, dass wir nicht noch ein oder zwei weitere Tore erzielt haben".
Y: We wanted to win and we did, although I regret that we did not score one or two goals more.
$\hat{Y}$**:** We wanted to win and we have won, although I regret it not yet that we have scored one or two more goals."

X-5: Zagreb kündigt an, die Durchreise über Slowenien nach Norden ohne Hürden zu gewähren.
Y: Zagreb has announced that they will allow passage over Slovenia towards the north without obstacles.
$\hat{Y}$**:** Zagreb has announced it will allow the migration through Slovenia to northern Croatia without hurdles.

X-6: Er könnte sich vorstellen, in Startup-Firmen der Neuen Werf zu investieren und später auch welche zu akquirieren.
Y: He can imagine investing in start-up business in Neuen Werft, and also later acquiring these.
$\hat{Y}$**:** He could also imagine investing in startup-owned companies like the New Lloyd's to invest and later in companies to quiquire.

X-7: Dieser oberflächliche Erfolg wird von dem dominiert, was Psychologen extrinsische Werte nennen; Geld ,Image, sozialer Status, alles auf Kosten intrinsischer Werte - unsere innere Sehnsucht nach persönlicher Entwicklung und Freundschaft, die sich auf unsere geistige Gesundheit viel tiefgehender auswirkt.
Y: This superficial success is dominated by what psychologists term extrinsic values; money, image, social status, all at the expense of intrinsic values - our inner yearning for personal growth and friendship - that more deeply impacts our mental health.
$\hat{Y}$**:** This OberPublic success is dominated by the things that economists call extraterrestrial values to characterize; money, image, social status all on the planet of intrinsatiable values - our inner desire for personal development and friendship, which can be felt on our mental health much deeper than it is in the physical world of human development and friendship, which is a real concern for our human health and friendship, which is in effect a greater word.

Figure 4: Example translations ($\hat{Y}$) by model (7) of sentences (X) from German to English with gold-standard translations (Y).

They show clear noise and are not always grammatical either. The quality of the paraphrases degrades substantially when using more paraphrases per sentence. More sophisticated paraphrasing systems (e.g. Witteveen and Andrews (2019)), might further improve results. We experimented with one, three and five extracted paraphrases per reference sentence. Using more than one paraphrase per sentence boosts performance substantially, highlighting that the paraphrasing objective also serves a data augmentation function. However, there was no noticeable difference between 3 and 5 paraphrases.

R: They have not been charged or formally arrested.
P1: they were not charged or officially arrested.
P2: they didn't have an arrest or official.

R: The Japanese-made tin robots have blocky heads and moveable arms and legs.
P1: the japanese robots have blocky robots, and their arms and feet.
P2: the japanese robots have blocky heads, their hands and feet.

R: The A.P. said it hoped for a resolution so it could return to full coverage of the six-week tournament before the opening match Friday between France and Argentina.
P1: the organisers said they hoped to find a resolution so he could return to full coverage of the six-week tournament before the opening - up friday between france and argentina.
P2: the panasonic said that it hoped for a resolution for an order to return to full coverage of the six-week tournament to keep an eye on friday between france and argentina.

Figure 5: Example references (R) and respective paraphrases (P) of the employed paraphrasing system (Wieting et al., 2017).

7 Literature Review

7.1 Unsupervised Machine Translation

Initial approaches for unsupervised MT focus on NMT systems in combination with back-translation, denoising autoencoding, and cross-lingual embeddings (Lample et al., 2018b; Artetxe et al., 2018b). Artetxe et al. (2018a); Lample et al. (2018b) improve over NMT approaches by focusing on phrase-based Statistical Machine Translation (PBSMT) with phrase tables from cross-lingual embedding mappings and iterative back-translation. Lample et al. (2018b) improves on these attempts by more careful initialization and language models. Lample et al. (2018b) attempt to combine both PBSMT and NMT, by tuning the NMT model on data generated by the PBSMT model and explore additional tweaks, e.g. byte-pair encodings (Sennrich et al., 2016b). Artetxe et al. (2019) also focus on PBSMT for unsupervised

machine translation. They propose a more sophisticated hybridization approach and unsupervised optimization technique for the PBSMT model. Comparable performance to Artetxe et al. (2019) was achieved by Conneau and Lample (2019) by simply initializing the NMT model of Lample et al. (2018b) with the weights of a pre-trained cross-lingual transformer. An analysis on the the practicality of unsupervised machine translation systems by Kim et al. (2020) concludes that linguistic dissimilarity and a domain mismatch between source and target data pose a substantial challenge for current state-of-the-art systems. They attribute these challenges to a lack of sufficient monolingual corpora for these domains, especially if one of the languages is under-resourced. The success of unsupervised methods has led to the first WMT shared subtask on unsupervised MT in 2019 on German-Czech with system submissions being very similar to existing approaches adapted for Czech (Kvapilíková et al., 2019; Liu et al., 2019). Moreover, since 2019 WMT organizes a similar language translation task for Spanish→Portuguese, Czech→Polish, and Hindi→Nepali (Barrault et al., 2019, 2020).

7.2 Cross-lingual Learning with Transformers

The success of fine-tuning pre-trained language models, such as GPT or BERT (Radford et al., 2019; Devlin et al., 2019) has also led to various cross-lingual versions of these pre-trained models, such as M-BERT (Devlin et al., 2019) and XLM (Conneau and Lample, 2019). These cross-lingual transformers learn to align multiple languages in their latent space by concatenating the monolingual training data on a shared subword vocabulary. They have created state-of-the-art results in multiple low-resource languages. Although monolingual models still outperform these cross-lingual ones if enough training data are available (Virtanen et al., 2019), a large-scale version of XLM has been shown to produce comparable results to monolingual LMs for high-resource languages (Conneau et al., 2020).

7.3 Paraphrase Generation

Paraphrase generation is concerned with the creation of phrases/sentences that use different words to express similar information (Bhagat and Hovy, 2013). It is a heavily researched task, with recent approaches focusing on the use of deep learning treating it as a Seq2Seq problem, by either using paraphrase databases (Cao et al., 2017; Li et al., 2018; Witteveen and Andrews, 2019), exploring NMT by pivoting between languages or pairs (Mallinson et al., 2017; Wieting et al., 2017; Federmann et al., 2019), or by treating it as a sentence simplification task (Zhang and Lapata, 2017; Niu et al., 2019). Most paraphrasing systems create a list of k most probable paraphrases. While paraphrase generation focuses on English, paraphrase datasets for other languages exist (Ganitkevitch and Callison-Burch, 2014). Extremely low-resource settings for paraphrasing have also been explored using parallel corpora (Maruyama and Yamamoto, 2019) or in a fully unsupervised setting (Roy and Grangier, 2019).

8 Conclusions & Future work

This work presented a simple fine-tuning method for unsupervised NMT that solely relies on the underlying alignment of a cross-lingual language model and monolingual data in the target language. Joint learning on a denoising autoencoder and paraphrasing objective creates a competitive system to strong baselines, especially for related language pairs, while requiring much shorter training times.

While this work has explored the proposed approach on commonly used language pairs for benchmarking unsupervised MT, future work includes testing the proposed method on other language pairs. This includes i) even more closely related languages (e.g. German→Dutch or Spanish→Portuguese), ii) language pairs without any parallel data, iii) translations between dialects. Moreover, we aim to further explore to which extent the proposed method can be combined with back-translation, especially in the context of distant languages, e.g. by adding back-translation iterations on top of the proposed approach, similar to PBSMT in (Lample et al., 2018b). Other training objectives should be explored, such as sentence simplification, translating between dialects, or even abstractive summarization when scaling this approach to document-level translation. Besides higher efficiency, this approach appears promising when training data for a language pair origins from different domains (e.g. Wikipedia versus News). Since our approach only requires data in the target language, domain mismatches in the training data for the language pair do not affect the proposed method. We are further aiming for human evaluation of translation quality.

Acknowledgements

We wish to thank Dr Andreas Vlachos for his support. We also thank the anonymous reviewers for their time and effort giving us feedback on our paper. This work was supported by the Engineering and Physical Sciences Research Council Doctoral Training Partnership. The second and third authors are supported by Research England via the University of Cambridge Global Challenges Research Fund.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Unsupervised Statistical Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018*, Vancouver, Canada.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, Florence, Italy.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (wmt20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online.

Rahul Bhagat and Eduard Hovy. 2013. Squibs: What Is a Paraphrase? *Computational Linguistics*, 39(3):463–472.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, and Lucia Specia, editors. 2014. *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal.

Chris Callison-Burch, Philipp Koehn, Cameron Shaw Fordyce, and Christof Monz, editors. 2007. *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Prague, Czech Republic.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Josh Schroeder, and Cameron Shaw Fordyce, editors. 2008. *Proceedings of the Third Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Columbus, Ohio.

Ziqiang Cao, Chuwei Luo, Wenjie Li, and Sujian Li. 2017. Joint copying and restricted generation for paraphrase. In *Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, CA, USA.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems 32*, Vancouver, Canada.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, Minneapolis, Minnesota.

Christian Federmann, Oussama Elachqar, and Chris Quirk. 2019. Multilingual whispers: Generating

paraphrases with translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, Hong Kong.

Juri Ganitkevitch and Chris Callison-Burch. 2014. The Multilingual Paraphrase Database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Hieu Hoang and Philipp Koehn. 2008. Design of the Moses Decoder for Statistical Machine Translation. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 58–65, Columbus, Ohio.

Yunsu Kim, Miguel Graça, and Hermann Ney. 2020. When and why is unsupervised neural machine translation useless? In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, Lisboa, Portugal.

Ivana Kvapilíková, Dominik Macháček, and Ondřej Bojar. 2019. CUNI systems for the unsupervised news translation task in WMT 2019. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, Florence, Italy.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*, Vancouver, Canada.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-Based & Neural Unsupervised Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium.

Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. 2018. Paraphrase generation with deep reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium.

Zihan Liu, Yan Xu, Genta Indra Winata, and Pascale Fung. 2019. Incorporating word and subword units in unsupervised machine translation using language model rescoring. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, Florence, Italy.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Valencia, Spain.

T. Maruyama and K. Yamamoto. 2019. Extremely low resource text simplification with pre-trained transformer language model. In *2019 International Conference on Asian Language Processing (IALP)*, Shanghai, China.

Tong Niu, Caiming Xiong, and Richard Socher. 2019. Deleter: Leveraging BERT to perform unsupervised successive text compression. *arXiv*, 1909.03223.

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*, Philadelphia, Pennsylvania, USA.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8).

Aurko Roy and David Grangier. 2019. Unsupervised paraphrasing without translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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)*, Berlin, Germany.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, Vancouver, Canada.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, Helsinki, Finland.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. *arXiv:1912.07076 [cs]*.

John Wieting, Jonathan Mallinson, and Kevin Gimpel. 2017. Learning Paraphrastic Sentence Embeddings from Back-Translated Bitext. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark.

Sam Witteveen and Martin Andrews. 2019. Paraphrasing with Large Language Models. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, Hong Kong.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark.