Phrase-Based & Neural Unsupervised Machine Translation

Guillaume Lample[†]
Facebook AI Research
Sorbonne Universités
glample@fb.com

Myle Ott
Facebook AI Research
myleott@fb.com

Alexis Conneau
Facebook AI Research
Université Le Mans
aconneau@fb.com

Ludovic Denoyer[†]
Sorbonne Universités
ludovic.denoyer@lip6.fr

Marc'Aurelio Ranzato
Facebook AI Research
ranzato@fb.com

Abstract

Machine translation systems achieve near human-level performance on some languages, yet their effectiveness strongly relies on the availability of large amounts of parallel sentences, which hinders their applicability to the majority of language pairs. This work investigates how to learn to translate when having access to only large monolingual corpora in each language. We propose two model variants, a neural and a phrase-based model. Both versions leverage a careful initialization of the parameters, the denoising effect of language models and automatic generation of parallel data by iterative back-translation. These models are significantly better than methods from the literature, while being simpler and having fewer hyper-parameters. On the widely used WMT'14 English-French and WMT'16 German-English benchmarks, our models respectively obtain 28.1 and 25.2 BLEU points without using a single parallel sentence, outperforming the state of the art by more than 11 BLEU points. On low-resource languages like English-Urdu and English-Romanian, our methods achieve even better results than semi-supervised and supervised approaches leveraging the paucity of available bitexts. Our code for NMT and PBSMT is publicly available.[1]

1 Introduction

Machine Translation (MT) is a flagship of the recent successes and advances in the field of natural language processing. Its practical applications and use as a testbed for sequence transduction algorithms have spurred renewed interest in this topic.

While recent advances have reported near human-level performance on several language pairs using neural approaches (Wu et al., 2016; Hassan et al., 2018), other studies have highlighted several open challenges (Koehn and Knowles, 2017; Isabelle et al., 2017; Sennrich, 2017). A major challenge is the reliance of current learning algorithms on large parallel corpora. Unfortunately, the vast majority of language pairs have very little, if any, parallel data: learning algorithms need to better leverage monolingual data in order to make MT more widely applicable.

While a large body of literature has studied the use of monolingual data to boost translation performance when limited supervision is available, two recent approaches have explored the fully unsupervised setting (Lample et al., 2018; Artetxe et al., 2018), relying only on monolingual corpora in each language, as in the pioneering work by Ravi and Knight (2011). While there are subtle technical differences between these two recent works, we identify several common principles underlying their success.

First, they carefully initialize the MT system with an inferred bilingual dictionary. Second, they leverage strong language models, via training the sequence-to-sequence system (Sutskever et al., 2014; Bahdanau et al., 2015) as a denoising autoencoder (Vincent et al., 2008). Third, they turn the unsupervised problem into a supervised one by automatic generation of sentence pairs via *back-translation* (Sennrich et al., 2015a), i.e., the source-to-target model is applied to source sentences to generate inputs for training the target-to-source model, and vice versa. Finally, they constrain the latent representations produced by the encoder to be shared across the two languages. Empirically, these methods achieve remarkable results considering the fully unsupervised setting; for instance, about 15 BLEU points on the WMT'14 English-French benchmark.

The first contribution of this paper is a model

[†]Sorbonne Universités, UPMC Univ Paris 06, CNRS, UMR 7606, LIP6, F-75005, Paris, France.

[1]https://github.com/facebookresearch/UnsupervisedMT

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 5039–5049
Brussels, Belgium, October 31 - November 4, 2018.

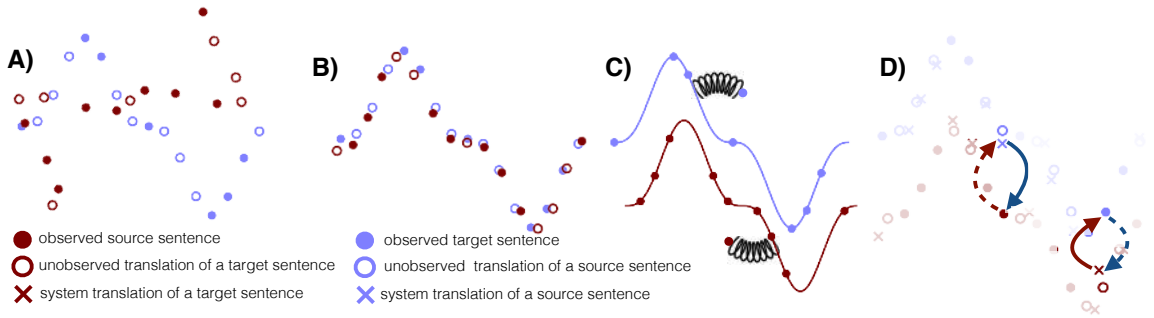

Figure 1: Toy illustration of the three principles of unsupervised MT. **A)** There are two monolingual datasets. Markers correspond to sentences (see legend for details). **B)** First principle: **Initialization**. The two distributions are roughly aligned, e.g. by performing word-by-word translation with an inferred bilingual dictionary. **C)** Second principle: **Language modeling**. A language model is learned independently in each domain to infer the structure in the data (underlying continuous curve); it acts as a data-driven prior to denoise/correct sentences (illustrated by the spring pulling a sentence outside the manifold back in). **D)** Third principle: **Back-translation**. Starting from an observed source sentence (filled red circle) we use the current source $\rightarrow$ target model to translate (dashed arrow), yielding a potentially incorrect translation (blue cross near the empty circle). Starting from this (back) translation, we use the target $\rightarrow$ source model (continuous arrow) to reconstruct the sentence in the original language. The discrepancy between the reconstruction and the initial sentence provides error signal to train the target $\rightarrow$ source model parameters. The same procedure is applied in the opposite direction to train the source $\rightarrow$ target model.

that combines these two previous neural approaches, simplifying the architecture and loss function while still following the above mentioned principles. The resulting model outperforms previous approaches and is both easier to train and tune. Then, we apply the same ideas and methodology to a traditional phrase-based statistical machine translation (PBSMT) system (Koehn et al., 2003). PBSMT models are well-known to outperform neural models when labeled data is scarce because they merely count occurrences, whereas neural models typically fit hundred of millions of parameters to learn distributed representations, which may generalize better when data is abundant but is prone to overfit when data is scarce. Our PBSMT model is simple, easy to interpret, fast to train and often achieves similar or better results than its NMT counterpart. We report gains of up to +10 BLEU points on widely used benchmarks when using our NMT model, and up to +12 points with our PBSMT model. Furthermore, we apply these methods to distant and low-resource languages, like English-Russian, English-Romanian and English-Urdu, and report competitive performance against both semi-supervised and supervised baselines.

2 Principles of Unsupervised MT

Learning to translate with only monolingual data is an ill-posed task, since there are potentially many ways to associate target with source sentences. Nevertheless, there has been exciting progress towards this goal in recent years, as discussed in the related work of Section 5. In this section, we abstract away from the specific assumptions made by each prior work and instead focus on identifying the common principles underlying unsupervised MT.

We claim that unsupervised MT can be accomplished by leveraging the three components illustrated in Figure 1: (i) suitable initialization of the translation models, (ii) language modeling and (iii) iterative back-translation. In the following, we describe each of these components and later discuss how they can be better instantiated in both a neural model and phrase-based model.

Initialization: Given the ill-posed nature of the task, model initialization expresses a natural prior over the space of solutions we expect to reach, jump-starting the process by leveraging approximate translations of words, short phrases or even sub-word units (Sennrich et al., 2015b). For instance, Klementiev et al. (2012) used a provided bilingual dictionary, while Lample et al. (2018) and Artetxe et al. (2018) used dictionaries inferred in an unsupervised way (Conneau et al., 2018; Artetxe et al., 2017). The motivating intuition is that while such initial "word-by-word" translation may be poor if languages or corpora are not closely related, it still preserves some of the original semantics.

Language Modeling: Given large amounts of monolingual data, we can train language models on both source and target languages. These models express a data-driven prior about how sentences should read in each language, and they improve the quality of the translation models by per-

Algorithm 1: Unsupervised MT

Language models: Learn language models P_s and P_t over source and target languages;
Initial translation models: Leveraging P_s and P_t, learn two initial translation models, one in each direction: $P^{(0)}_{s\to t}$ and $P^{(0)}_{t\to s}$;
for k=1 **to** N **do**
  Back-translation: Generate source and target sentences using the current translation models, $P^{(k-1)}_{t\to s}$ and $P^{(k-1)}_{s\to t}$, factoring in language models, P_s and P_t;
  Train new translation models $P^{(k)}_{s\to t}$ and $P^{(k)}_{t\to s}$ using the generated sentences and leveraging P_s and P_t;
end

forming local substitutions and word reorderings.

Iterative Back-translation: The third principle is back-translation (Sennrich et al., 2015a), which is perhaps the most effective way to leverage monolingual data in a semi-supervised setting. Its application in the unsupervised setting is to couple the source-to-target translation system with a backward model translating from the target to source language. The goal of this model is to generate a source sentence for each target sentence in the monolingual corpus. This turns the daunting unsupervised problem into a supervised learning task, albeit with noisy source sentences. As the original model gets better at translating, we use the current model to improve the back-translation model, resulting in a coupled system trained with an iterative algorithm (He et al., 2016).

3 Unsupervised MT systems

Equipped with the three principles detailed in Section 2, we now discuss how to effectively combine them in the context of a NMT model (Section 3.1) and PBSMT model (Section 3.2).

In the reminder of the paper, we denote the space of source and target sentences by $\mathcal{S}$ and $\mathcal{T}$, respectively, and the language models trained on source and target monolingual datasets by P_s and P_t, respectively. Finally, we denote by $P_{s\to t}$ and $P_{t\to s}$ the translation models from source to target and vice versa. An overview of our approach is given in Algorithm 1.

3.1 Unsupervised NMT

We now introduce a new unsupervised NMT method, which is derived from earlier work by Artetxe et al. (2018) and Lample et al. (2018). We first discuss how the previously mentioned three key principles are instantiated in our work, and then introduce an additional property of the system, the sharing of internal representations across languages, which is specific and critical to NMT. From now on, we assume that a NMT model consists of an encoder and a decoder. Section 4 gives the specific details of this architecture.

Initialization: While prior work relied on bilingual dictionaries, here we propose a more effective and simpler approach which is particularly suitable for related languages.[2] First, instead of considering words, we consider byte-pair encodings (BPE) (Sennrich et al., 2015b), which have two major advantages: they reduce the vocabulary size and they eliminate the presence of unknown words in the output translation. Second, instead of learning an explicit mapping between BPEs in the source and target languages, we define BPE tokens by *jointly* processing both monolingual corpora. If languages are related, they will naturally share a good fraction of BPE tokens, which eliminates the need to infer a bilingual dictionary. In practice, we i) join the monolingual corpora, ii) apply BPE tokenization on the resulting corpus, and iii) learn token embeddings (Mikolov et al., 2013) on the same corpus, which are then used to initialize the lookup tables in the encoder and decoder.

Language Modeling: In NMT, language modeling is accomplished via denoising autoencoding, by minimizing:

$$\mathcal{L}^{lm} = \mathbb{E}_{x\sim\mathcal{S}}[-\log P_{s\to s}(x|C(x))] + \mathbb{E}_{y\sim\mathcal{T}}[-\log P_{t\to t}(y|C(y))] \quad (1)$$

where C is a noise model with some words dropped and swapped as in Lample et al. (2018). $P_{s\to s}$ and $P_{t\to t}$ are the composition of encoder and decoder both operating on the source and target sides, respectively.

Back-translation: Let us denote by $u^*(y)$ the sentence in the source language inferred from $y \in \mathcal{T}$ such that $u^*(y) = \arg\max P_{t\to s}(u|y)$. Similarly, let us denote by $v^*(x)$ the sentence in the target language inferred from $x \in \mathcal{S}$ such that $v^*(x) = \arg\max P_{s\to t}(v|x)$. The pairs $(u^*(y), y)$ and $(x, v^*(x)))$ constitute automatically-generated parallel sentences which, following the back-translation principle, can be

[2] For unrelated languages, we need to infer a dictionary to properly initialize the embeddings (Conneau et al., 2018).

used to train the two MT models by minimizing the following loss:

$$\mathcal{L}^{back} = \mathbb{E}_{y\sim\mathcal{T}}[-\log P_{s\to t}(y|u^*(y))] + \mathbb{E}_{x\sim\mathcal{S}}[-\log P_{t\to s}(x|v^*(x))]. \quad (2)$$

Note that when minimizing this objective function we do not back-prop through the reverse model which generated the data, both for the sake of simplicity and because we did not observe improvements when doing so. The objective function minimized at every iteration of stochastic gradient descent, is simply the sum of $\mathcal{L}^{lm}$ in Eq. 1 and $\mathcal{L}^{back}$ in Eq. 2. To prevent the model from cheating by using different subspaces for the language modeling and translation tasks, we add an additional constraint which we discuss next.

Sharing Latent Representations: A shared encoder representation acts like an interlingua, which is translated in the decoder target language regardless of the input source language. This ensures that the benefits of language modeling, implemented via the denoising autoencoder objective, nicely transfer to translation from noisy sources and eventually help the NMT model to translate more fluently. In order to share the encoder representations, we share all encoder parameters (including the embedding matrices since we perform joint tokenization) across the two languages to ensure that the latent representation of the source sentence is robust to the source language. Similarly, we share the decoder parameters across the two languages. While sharing the encoder is critical to get the model to work, sharing the decoder simply induces useful regularization. Unlike prior work (Johnson et al., 2016), the first token of the decoder specifies the language the module is operating with, while the encoder does not have any language identifier.

3.2 Unsupervised PBSMT

In this section, we discuss how to perform unsupervised machine translation using a Phrase-Based Statistical Machine Translation (PBSMT) system (Koehn et al., 2003) as the underlying backbone model. Note that PBSMT models are known to perform well on low-resource language pairs, and are therefore a potentially good alternative to neural models in the unsupervised setting.

When translating from x to y, a PBSMT system scores y according to: $\arg\max_y P(y|x) = \arg\max_y P(x|y)P(y)$, where $P(x|y)$ is derived from so called "phrase tables", and $P(y)$ is the score assigned by a language model.

Given a dataset of bitexts, PBSMT first infers an alignment between source and target phrases. It then populates phrase tables, whose entries store the probability that a certain n-gram in the source/target language is mapped to another n-gram in the target/source language.

In the unsupervised setting, we can easily train a language model on monolingual data, but it is less clear how to populate the phrase tables, which are a necessary component for good translation. Fortunately, similar to the neural case, the principles of Section 2 are effective to solve this problem.

Initialization: We populate the initial phrase tables (from source to target and from target to source) using an inferred bilingual dictionary built from monolingual corpora using the method proposed by Conneau et al. (2018). In the following, we will refer to phrases as single words, but the very same arguments trivially apply to longer n-grams. Phrase tables are populated with the scores of the translation of a source word to:

$$p(t_j|s_i) = \frac{e^{\frac{1}{T}\cos(e(t_j),We(s_i))}}{\sum_k e^{\frac{1}{T}\cos(e(t_k),We(s_i))}}, \quad (3)$$

where t_j is the j-th word in the target vocabulary and s_i is the i-th word in the source vocabulary, T is a hyper-parameter used to tune the peakiness of the distribution[3], W is the rotation matrix mapping the source embeddings into the target embeddings (Conneau et al., 2018), and $e(x)$ is the embedding of x.

Language Modeling: Both in the source and target domains we learn smoothed n-gram language models using KenLM (Heafield, 2011), although neural models could also be considered. These remain fixed throughout training iterations.

Iterative Back-Translation: To jump-start the iterative process, we use the unsupervised phrase tables and the language model on the target side to construct a seed PBSMT. We then use this model to translate the source monolingual corpus into the target language (back-translation step). Once the data has been generated, we train a PBSMT in supervised mode to map the generated data back to the original source sentences. Next, we perform

[3]We set $T = 30$ in all our experiments, following the setting of Smith et al. (2017).

both generation and training process but in the reverse direction. We repeat these steps as many times as desired (see Algorithm 2 in Section A).

Intuitively, many entries in the phrase tables are not correct because the input to the PBSMT at any given point during training is noisy. Despite that, the language model may be able to fix some of these mistakes at generation time. As long as that happens, the translation improves, and with that also the phrase tables at the next round. There will be more entries that correspond to correct phrases, which makes the PBSMT model stronger because it has bigger tables and it enables phrase swaps over longer spans.

4 Experiments

We first describe the datasets and experimental protocol we used. Then, we compare the two proposed unsupervised approaches to earlier attempts, to semi-supervised methods and to the very same models but trained with varying amounts of labeled data. We conclude with an ablation study to understand the relative importance of the three principles introduced in Section 2.

4.1 Datasets and Methodology

We consider five language pairs: English-French, English-German, English-Romanian, English-Russian and English-Urdu. The first two pairs are used to compare to recent work on unsupervised MT (Artetxe et al., 2018; Lample et al., 2018). The last three pairs are instead used to test our PBSMT unsupervised method on truly low-resource pairs (Gu et al., 2018) or unrelated languages that do not even share the same alphabet.

For English, French, German and Russian, we use all available sentences from the WMT monolingual News Crawl datasets from years 2007 through 2017. For Romanian, the News Crawl dataset is only composed of 2.2 million sentences, so we augment it with the monolingual data from WMT'16, resulting in 2.9 million sentences. In Urdu, we use the dataset of Jawaid et al. (2014), composed of about 5.5 million monolingual sentences. We report results on *newstest* 2014 for $en-fr$, and *newstest* 2016 for $en-de$, $en-ro$ and $en-ru$. For Urdu, we use the LDC2010T21 and LDC2010T23 corpora each with about 1800 sentences as validation and test sets, respectively.

We use Moses scripts (Koehn et al., 2007) for tokenization. NMT is trained with 60,000 BPE codes. PBSMT is trained with true-casing, and by removing diacritics from Romanian on the source side to deal with their inconsistent use across the monolingual dataset (Sennrich et al., 2016).

Source	Target	$P(s\|t)$	$P(t\|s)$
heureux	happy	0.931	0.986
	delighted	0.458	0.003
	grateful	0.128	0.003
	thrilled	0.392	0.002
	glad	0.054	0.001
Royaume-Uni	Britain	0.242	0.720
	UK	0.816	0.257
	U.K.	0.697	0.011
	United Kingdom	0.770	0.010
	British	0.000	0.002
Union européenne	European Union	0.869	0.772
	EU	0.335	0.213
	E.U.	0.539	0.006
	member states	0.007	0.006
	27-nation bloc	0.410	0.002

Table 1: **Unsupervised phrase table.** Example of candidate French to English phrase translations, along with their corresponding conditional likelihoods.

4.2 Initialization

Both the NMT and PBSMT approaches require either cross-lingual BPE embeddings (to initialize the shared lookup tables) or n-gram embeddings (to initialize the phrase table). We generate embeddings using fastText (Bojanowski et al., 2017) with an embedding dimension of 512, a context window of size 5 and 10 negative samples. For NMT, fastText is applied on the concatenation of source and target corpora, which results in cross-lingual BPE embeddings.

For PBSMT, we generate n-gram embeddings on the source and target corpora independently, and align them using the MUSE library (Conneau et al., 2018). Since learning unique embeddings of every possible phrase would be intractable, we consider the most frequent 300,000 source phrases, and align each of them to its 200 nearest neighbors in the target space, resulting in a phrase table of 60 million phrase pairs which we score using the formula in Eq. 3.

In practice, we observe a small but significant difference of about 1 BLEU point using a phrase table of bigrams compared to a phrase table of unigrams, but did not observe any improvement using longer phrases. Table 1 shows an extract of a French-English unsupervised phrase table, where we can see that unigrams are correctly aligned to bigrams, and vice versa.

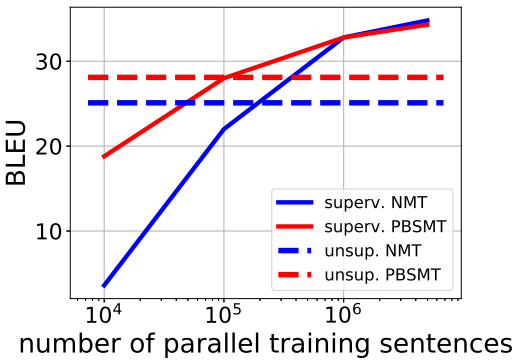

Figure 2: Comparison between supervised and unsupervised approaches on WMT'14 En-Fr, as we vary the number of parallel sentences for the supervised methods.

4.3 Training

The next subsections provide details about the architecture and training procedure of our models.

4.3.1 NMT

In this study, we use NMT models built upon LSTM (Hochreiter and Schmidhuber, 1997) and Transformer (Vaswani et al., 2017) cells. For the LSTM model we use the same architecture as in Lample et al. (2018). For the Transformer, we use 4 layers both in the encoder and in the decoder. Following Press and Wolf (2016), we share all lookup tables between the encoder and the decoder, and between the source and the target languages. The dimensionality of the embeddings and of the hidden layers is set to 512. We used the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 10^{-4}, $\beta_1 = 0.5$, and a batch size of 32. At decoding time, we generate greedily.

4.3.2 PBSMT

The PBSMT uses Moses' default smoothed n-gram language model with phrase reordering disabled during the very first generation. PBSMT is trained in a iterative manner using Algorithm 2. At each iteration, we translate 5 million sentences randomly sampled from the monolingual dataset in the source language. Except for initialization, we use phrase tables with phrases up to length 4.

4.4 Model selection

Moses' implementation of PBSMT has 15 hyperparameters, such as relative weighting of each scoring function, word penalty, etc. In this work, we consider two methods to set these hyperparameters. We either set them to their default values in the toolbox, or we set them using a small validation set of parallel sentences. It turns out that with only 100 labeled sentences in the validation set, PBSMT would overfit to the validation set. For instance, on $en \rightarrow fr$, PBSMT tuned on 100 parallel sentences obtains a BLEU score of 26.42 on *newstest* 2014, compared to 27.09 with default hyper-parameters, and 28.02 when tuned on the 3000 parallel sentences of *newstest* 2013. Therefore, unless otherwise specified, all PBSMT models considered in the paper use default hyperparameter values, and do not use any parallel resource whatsoever.

For the NMT, we also consider two model selection procedures: an *unsupervised criterion* based on the BLEU score of a "round-trip" translation (source $\rightarrow$ target $\rightarrow$ source and target $\rightarrow$ source $\rightarrow$ target) as in Lample et al. (2018), and cross-validation using a small validation set with 100 parallel sentences. In our experiments, we found the unsupervised criterion to be highly correlated with the test metric when using the Transformer model, but not always for the LSTM. Therefore, unless otherwise specified, we select the best LSTM models using a small validation set of 100 parallel sentences, and the best Transformer models with the unsupervised criterion.

Model	en-fr	fr-en	en-de	de-en
(Artetxe et al., 2018)	15.1	15.6	-	-
(Lample et al., 2018)	15.0	14.3	9.6	13.3
(Yang et al., 2018)	17.0	15.6	10.9	14.6
NMT (LSTM)	24.5	23.7	14.7	19.6
NMT (Transformer)	25.1	24.2	17.2	21.0
PBSMT (Iter. 0)	16.2	17.5	11.0	15.6
PBSMT (Iter. n)	**28.1**	27.2	17.9	22.9
NMT + PBSMT	27.1	26.3	17.5	22.1
PBSMT + NMT	27.6	**27.7**	**20.2**	**25.2**

Table 2: **Comparison with previous approaches.** BLEU score for different models on the $en - fr$ and $en - de$ language pairs. Just using the unsupervised phrase table, and without back-translation (PBSMT (Iter. 0)), the PBSMT outperforms previous approaches. Combining PBSMT with NMT gives the best results.

4.5 Results

The results reported in Table 2 show that our unsupervised NMT and PBSMT systems largely outperform previous unsupervised baselines. We report large gains on all language pairs and directions. For instance, on the $en \rightarrow fr$ task, our unsupervised PBSMT obtains a BLEU score of 28.1, outperforming the previous best result by more than 11 BLEU points. Even on a more complex task like $en \rightarrow de$, both PBSMT and NMT surpass the baseline score by more than 10 BLEU

	en → fr	fr→ en	en→ de	de→ en	en→ ro	ro→ en	en→ ru	ru→ en
Unsupervised PBSMT								
Unsupervised phrase table	-	17.50	-	15.63	-	14.10	-	8.08
Back-translation - Iter. 1	24.79	26.16	15.92	22.43	18.21	21.49	11.04	15.16
Back-translation - Iter. 2	27.32	26.80	17.65	22.85	20.61	22.52	12.87	16.42
Back-translation - Iter. 3	27.77	26.93	17.94	22.87	21.18	22.99	13.13	16.52
Back-translation - Iter. 4	27.84	27.20	17.77	22.68	21.33	23.01	13.37	**16.62**
Back-translation - Iter. 5	**28.11**	27.16	-	-	-	-	-	-
Unsupervised NMT								
LSTM	24.48	23.74	14.71	19.60	-	-	-	-
Transformer	25.14	24.18	17.16	21.00	21.18	19.44	7.98	9.09
Phrase-based + Neural network								
NMT + PBSMT	27.12	26.29	17.52	22.06	21.95	23.73	10.14	12.62
PBSMT + NMT	27.60	**27.68**	**20.23**	**25.19**	**25.13**	**23.90**	**13.76**	**16.62**

Table 3: **Fully unsupervised results.** We report the BLEU score for PBSMT, NMT, and their combinations on 8 directed language pairs. Results are obtained on *newstest* 2014 for $en - fr$ and *newstest* 2016 for every other pair.

points. Even before iterative back-translation, the PBSMT model significantly outperforms previous approaches, and can be trained in a few minutes.

Table 3 illustrates the quality of the PBSMT model during the iterative training process. For instance, the $fr \rightarrow en$ model obtains a BLEU score of 17.5 at iteration 0 – i.e. after the unsupervised phrase table construction – while it achieves a score of 27.2 at iteration 4. This highlights the importance of multiple back-translation iterations. The last rows of Table 3 also show that we get additional gains by further tuning the NMT model on the data generated by PBSMT (PBSMT + NMT). We simply add the data generated by the unsupervised PBSMT system to the back-translated data produced by the NMT model. By combining PBSMT and NMT, we achieve BLEU scores of 20.2 and 25.2 on the challenging $en \rightarrow de$ and $de \rightarrow en$ translation tasks. While we also tried bootstraping the PBSMT model with back-translated data generated by a NMT model (NMT + PBSMT), this did not improve over PBSMT alone.

Next, we compare to fully supervised models. Figure 2 shows the performance of the same architectures trained in a fully supervised way using parallel training sets of varying size. The unsupervised PBSMT model achieves the same performance as its supervised counterpart trained on more than 100,000 parallel sentences.

This is confirmed on low-resource languages. In particular, on $ro \rightarrow en$, our unsupervised PBSMT model obtains a BLEU score of 23.9, outperforming Gu et al. (2018)'s method by 1 point, despite its use of 6,000 parallel sentences, a seed dictionary, and a multi-NMT system combining parallel resources from 5 different languages.

On Russian, our unsupervised PBSMT model obtains a BLEU score of 16.6 on $ru \rightarrow en$, showing that this approach works reasonably well on distant languages. Finally we train on $ur \rightarrow en$, which is both low resource and distant. In a supervised mode, PBSMT using the noisy and out-of-domain 800,000 parallel sentences from Tiedemann (2012) achieves a BLEU score of 9.8. Instead, our unsupervised PBSMT system achieves 12.3 BLEU using only a validation set of 1800 sentences to tune Moses hyper-parameters.

4.6 Ablation Study

In Figure 3 we report results from an ablation study, to better understand the importance of the three principles when training PBSMT. This study shows that more iterations only partially compensate for lower quality phrase table initialization (Left), language models trained over less data (Middle) or less monolingual data (Right). Moreover, the influence of the quality of the language model becomes more prominent as we iterate. These findings suggests that better initialization methods and more powerful language models may further improve our results.

We perform a similar ablation study for the NMT system (see Appendix). We find that back-translation and auto-encoding are critical components, without which the system fails to learn. We also find that initialization of embeddings is very important, and we gain 7 BLEU points compared to prior work (Artetxe et al., 2018; Lample et al., 2018) by learning BPE embeddings over the concatenated monolingual corpora.

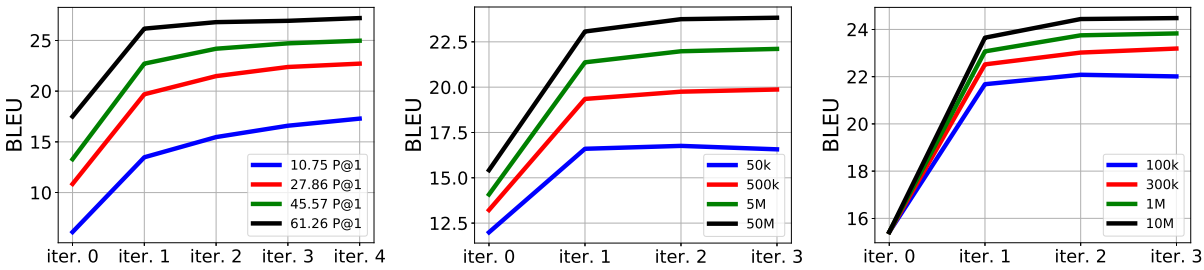

Figure 3: Results with PBSMT on the $fr \rightarrow en$ pair at different iterations. We vary: Left) the quality of the initial alignment between the source and target embeddings (measured in P@1 on the word translation task), Middle) the number of sentences used to train the language models, Right) the number of sentences used for back-translation.

5 Related Work

A large body of literature has studied using monolingual data to boost translation performance when limited supervision is available. This limited supervision is typically provided as a small set of parallel sentences (Sennrich et al., 2015a; Gulcehre et al., 2015; He et al., 2016; Gu et al., 2018; Wang et al., 2018); large sets of parallel sentences in related languages (Firat et al., 2016; Johnson et al., 2016; Chen et al., 2017; Zheng et al., 2017); cross-lingual dictionaries (Klementiev et al., 2012; Irvine and Callison-Burch, 2014, 2016); or comparable corpora (Munteanu et al., 2004; Irvine and Callison-Burch, 2013).

Learning to translate *without* any form of supervision has also attracted interest, but is challenging. In their seminal work, Ravi and Knight (2011) leverage linguistic prior knowledge to reframe the unsupervised MT task as deciphering and demonstrate the feasibility on short sentences with limited vocabulary. Earlier work by Carbonell et al. (2006) also aimed at unsupervised MT, but leveraged a bilingual dictionary to seed the translation. Both works rely on a language model on the target side to correct for translation fluency.

Subsequent work (Klementiev et al., 2012; Irvine and Callison-Burch, 2014, 2016) relied on bilingual dictionaries, small parallel corpora of several thousand sentences, and linguistically motivated features to prune the search space. Irvine and Callison-Burch (2014) use monolingual data to expand phrase tables learned in a supervised setting. In our work we also expand phrase tables, but we initialize them with an inferred bilingual n-gram dictionary, following work from the connectionist community aimed at improving PBSMT with neural models (Schwenk, 2012; Kalchbrenner and Blunsom, 2013; Cho et al., 2014).

In recent years back-translation has become a popular method of augmenting training sets with monolingual data on the target side (Sennrich et al., 2015a), and has been integrated in the "dual learning" framework of He et al. (2016) and subsequent extensions (Wang et al., 2018). Our approach is similar to the dual learning framework, except that in their model gradients are backpropagated through the reverse model and they pretrain using a relatively large amount of labeled data, whereas our approach is fully unsupervised.

Finally, our work can be seen as an extension of recent studies (Lample et al., 2018; Artetxe et al., 2018; Yang et al., 2018) on *fully unsupervised* MT with two major contributions. First, we propose a much simpler and more effective initialization method for related languages. Second, we abstract away three principles of unsupervised MT and apply them to a PBSMT, which even outperforms the original NMT. Moreover, our results show that the combination of PBSMT and NMT achieves even better performance.

6 Conclusions and Future Work

In this work, we identify three principles underlying recent successes in fully unsupervised MT and show how to apply these principles to PBSMT and NMT systems. We find that PBSMT systems often outperform NMT systems in the fully unsupervised setting, and that by combining these systems we can greatly outperform previous approaches from the literature. We apply our approach to several popular benchmark language pairs, obtaining state of the art results, and to several low-resource and under-explored language pairs.

It's an open question whether there are more effective instantiations of these principles or other principles altogether, and under what conditions our iterative process is guaranteed to converge. Future work may also extend to the semi-supervised setting.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, volume 1, pages 451–462.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In International Conference on Learning Representations (ICLR).

D. Bahdanau, K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. In International Conference on Learning Representations (ICLR).

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146.

Jaime Carbonell, Steve Klein, David Miller, Michael Steinbaum, Tomer Grassiany, and Jochen Frey. 2006. Context-based machine translation. In The Association for Machine Translation in the Americas.

Y. Chen, Y. Liu, Y. Cheng, and V.O.K. Li. 2017. A teacher-student framework for zero-resource neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder—decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1724——1734.

A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou. 2018. Word translation without parallel data. In International Conference on Learning Representations (ICLR).

O. Firat, B. Sankaran, Y. Al-Onaizan, F.T.Y. Vural, and K. Cho. 2016. Zero-resource translation with multilingual neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. arXiv preprint arXiv:1503.03535.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. In arXiv:1803.05567.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In Advances in Neural Information Processing Systems, pages 820–828.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In Proceedings of the Sixth Workshop on Statistical Machine Translation, pages 187–197. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural computation, 9(8):1735–1780.

Ann Irvine and Chris Callison-Burch. 2013. Combining bilingual and comparable corpora for low resource machine translation. In Proceedings of the eighth workshop on statistical machine translation, pages 262–270.

Ann Irvine and Chris Callison-Burch. 2014. Hallucinating phrase translations for low resource mt. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning, pages 160–170.

Ann Irvine and Chris Callison-Burch. 2016. End-to-end statistical machine translation with zero or small parallel texts. In Journal of Natural Language Engineering, volume 22, pages 517–548.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2486–2496.

Bushra Jawaid, Amir Kamran, and Ondrej Bojar. 2014. A tagged corpus and a tagger for urdu. In LREC.

M. Johnson, M. Schuster, Q.V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. In Transactions of the Association for Computational Linguistics.

Nal Kalchbrenner and Phil Blunsom. 2013. Two recurrent continuous translation models. In Conference on Empirical Methods in Natural Language Processing.

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

Alex Klementiev, Ann Irvine, Chris Callison-Burch, and David Yarowsky. 2012. Toward statistical machine translation without parallel corpora. In Proceedings of the 13th Conference of the European Chapter of the Association for computational Linguistics, Avignon, France. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Ondrej Bojar Chris Dyer, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In Annual Meeting of the Association for Computational Linguistics (ACL), demo session.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In Proceedings of the First Workshop on Neural Machine Translation, pages 28–39.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL), volume 1, pages 48—54.

G. Lample, A. Conneau, L. Denoyer, and M. Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In International Conference on Learning Representations (ICLR).

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119.

D.S. Munteanu, A. Fraser, and D. Marcu. 2004. Improved machine translation performance via parallel sentence extraction from comparable corpora. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).

Ofir Press and Lior Wolf. 2016. Using the output embedding to improve language models. arXiv preprint arXiv:1608.05859.

S. Ravi and K. Knight. 2011. Deciphering foreign language. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 12–21.

Holger Schwenk. 2012. Continuous space translation models for phrase-based statistical machine translation. In International Conference on Computational Linguistics, pages 1071—1080.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? assessing mt quality with contrastive translation pairs. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, pages 376–382.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015a. Improving neural machine translation models with monolingual data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 86–96.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015b. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1715–1725.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for wmt 16. In Proceedings of the First Conference on Machine Translation, pages 371–376.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In Internaltional Conference on Learning Representations.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems, pages 3104–3112.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In Proceedings of the 8th International Conference on Language Resources and Evaluation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. arXiv preprint arXiv:1706.03762.

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, pages 1096–1103.

Yijun Wang, Yingce Xia, Li Zhao, Jiang Bian, Tao Qin, Guiquan Liu, and Tie-Yan Liu. 2018. Dual transfer learning for neural machine translation with marginal distribution regularization. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. Transactions of the Association for Computational Linguistics, pages 339–351.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018. Unsupervised neural machine translation with weight sharing. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics.

H. Zheng, Y. Cheng, and Y. Liu. 2017. Maximum expected likelihood estimation for zero-resource neural machine translation. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI), pages 4251–4257.