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 bitexts, 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 automatic generation of parallel data by back-translating with a backward model operating in the other direction, and the denoising effect of a language model trained on the target side. 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 27.1 and 23.6 BLEU points without using a single parallel sentence, outperforming the state of the art by more than 11 BLEU points.

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.

A large body of literature has studied the use of monolingual data to boost translation performance when limited supervision is available. This limited supervision is typically provided in the form of a relatively 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), or a large set of parallel sentences but in other related languages (Firat et al., 2016; Johnson et al., 2016; Chen et al., 2017; Zheng et al., 2017), or bilingual dictionaries (Klementiev et al., 2012; Irvine and Callison-Burch, 2014, 2016), or with comparable corpora (Munteanu et al., 2004; Irvine and Callison-Burch, 2013).

Recently, by contrast, two approaches have been proposed that are fully unsupervised (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 ingredients underlying their success. First, they carefully initialize the model 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). In back-translation, the key idea is to maintain two models, one for translating the source into the target and the other to translate the target into the source. The former model generates data to train the latter one and vice versa. The last common prop-

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

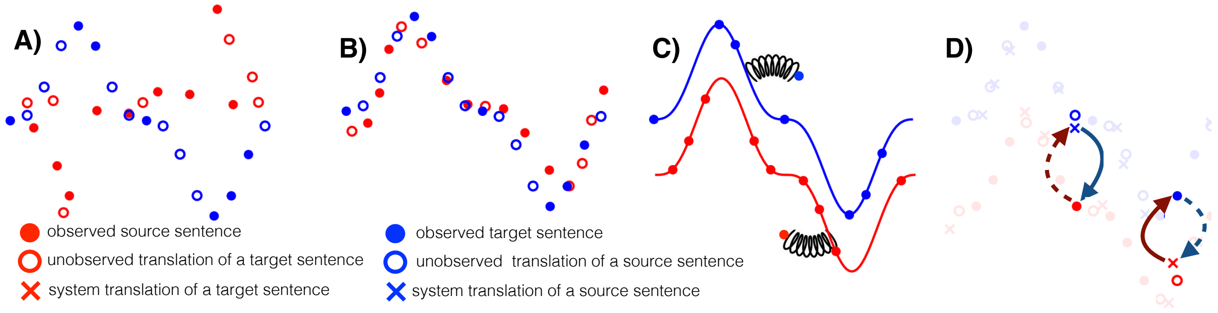
A)
B)
C)
D)
● observed source sentence
○ unobserved translation of a target sentence
× system translation of a target sentence
● observed target sentence
○ unobserved translation of a source sentence
× system translation of a source sentence

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 → 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 → 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 → source model parameters. The same procedure is applied in the opposite direction to train the source → target model.

erty is that these models constrain the latent representations produced by the encoder to be shared across the two languages. Putting these pieces together, the encoder produces similar representations regardless of the input language. The decoder, which is trained both as a language model and as a translator from noisy inputs, learns to produce increasingly better translations in tandem with the backward model (operating from target to source). This iterative process achieves remarkable results in a fully unsupervised setting; for instance, about 15 BLEU points on the WMT'14 English-French benchmark.

In this paper, we propose a 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 are 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. This significantly advances the state of the art in the unsupervised setting.

The rest of this paper is organized as follows. In Section 2 we introduce the key principles underlying our approach to unsupervised machine translation. In Section 3 we introduce NMT and PBSMT models that employ these principles, and evaluate them empirically in Section 4. Finally, we discuss how they relate to other approaches in Section 5.

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 and source sentences. Nevertheless, there has been exciting progress in solving this problem in recent years, as discussed in the related work of Section 5. In this section, we abstract away from the specific assumptions made by this recent work and instead focus on identifying the common principles underlying unsupervised MT.

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

Initialization: Since unsupervised translation is ill-posed, one natural prior we can express over the set of solutions we expect is to initialize the model so that words, short phrases or even subword units (Sennrich et al., 2015b) are aligned. 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 such alignment can be used to perform an initial "word-by-word" translation. And while this may result in a poor translation if languages or corpora are not closely related, it can still preserve 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. They improve the quality of the translation by performing local substitutions and reordering words.

Iterative Back-translation: The third component 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 machine 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 can also use the current model to improve the back-translation model, resulting in 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) or 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.

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
    Backtranslation: 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
```

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 important feature of the system, which is specific and critical to NMT.

In general, an NMT model is composed of an encoder and a decoder; the specific details of this architecture is given in Section 4.

Initialization: While prior work relied on bilingual dictionaries, here we propose a more effective and simpler approach which is suitable for related languages.[1] 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, as those we consider in this study, 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 that are used to initialize the lookup tables in the encoder and decoder.

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

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). $P_{s \to s}$ ($P_{t \to t}$) is the composition of the encoder and decoder both operating on the source (target) side.

Back-translation: Let us denote by $u^*(y)$ the sentence in the source language inferred from y 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 such that $v^*(x) = \arg\max P_{s \to t}(v|x)$. The pairs $(u^*(y), y)$ and $(x, v^*(x)))$ can be seen as aligned sentences on which, following the back-translation principle, a new MT model can be learned. Therefore, the back-translation loss is:

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

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 (t) of the learning process, namely gradient step of stochastic gradient descent, is simply the sum of $\mathcal{L}^{back}$ in Eq. 1 and $\mathcal{L}^{lm}$ in Eq. 2. However, this alone would not work very well, because it is too unconstrained. For instance, the decoder operating in the target space has to work well both when fed encoder representations of target sentences as well as encoder representations of source sentences. Unfortunately, the system can cheat and perfectly minimize the denoising and translation loss by *splitting* the latent space in two, and use one subspace for the language modeling task and another subspace for the translation tasks. Clearly, learning to invert the backward model and separately learning a language model are not sufficient to translate well. This leads to an additional constraint required for neural unsupervised machine translation, 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.

Note that the BPE joint tokenization, which removes the need to infer a bilingual dictionary for related languages, and the choice of architecture both differ from prior work (Artetxe et al., 2018; Lample et al., 2018). Moreover, here we share the decoder unlike Artetxe et al. (2018). Compared to Lample et al. (2018), we also do online back translation and lack the adversarial term in the loss, since the architecture and tokenization are sufficient to share the latent representations. Overall, these changes simplify the model and reduce the number of hyper-parameters.

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 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 of a language model.

Given a dataset of bitexts, PBSMT first infers an alignment, and then populates phrase tables. Each entry of a phrase table stores the likelihood that a certain n-gram in the source language is mapped to another n-gram in the target language, an estimation based on normalized counts.

In practice, the actual scoring is a little more involved as other terms are often introduced, such as one to take into account the relative positional misplacement between n-grams, which discourages large phrase re-orderings, one to account for phrase tables in the other direction, etc.

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). These phrases tables are populated with unigrams[2] by setting the scores of the translation of a source word to:

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

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 unigram 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 both generation and training process but in the reverse direction. We repeat these steps as many times as desired, see Algorithm 2.

Algorithm 2: Unsupervised PBSMT

```
Learn bilingual dictionary using Conneau et al. (2018);
Populate unigram tables using Eq. 3 and learn a language model to build $P^{(0)}_{s\to t}$;
Use $P^{(0)}_{s\to t}$ to translate the source monolingual dataset, yielding $\mathcal{D}^{(0)}_{\mathrm{t}}$;
for i=1 to N do
    Train model $P^{(i)}_{t\to s}$ using $\mathcal{D}^{(i-1)}_{\mathrm{t}}$;
    Use $P^{(i)}_{t\to s}$ to translate the target monolingual dataset, yielding $\mathcal{D}^{(i)}_{\mathrm{s}}$;
    Train model $P^{(i)}_{s\to t}$ using $\mathcal{D}^{(i)}_{\mathrm{s}}$;
    Use $P^{(i)}_{s\to t}$ to translate the source monolingual dataset, yielding $\mathcal{D}^{(i)}_{\mathrm{t}}$;
end
```

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 (Artetxe et al., 2018; Lample et al., 2018), to semi-supervised methods (Gu et al., 2018) and to the very same models but trained with varying amounts of labeled data. We conclude with an ablation study to understand the importance of each component in the system, and some qualitative assessment of the translations.

4.1 Datasets and Methodology

We consider four language pairs: English-French, English-German, English-Romanian and English-Russian. The first two pairs are used to com-

[2] The extension to n-grams is trivial. Experiments with n-grams are reported in Section 4. We could also have considered to work at the level of BPEs as opposed to words (Kunchukuttan and Bhattacharyya, 2016). We leave that to future work.

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

pare to recent work on unsupervised MT (Artetxe et al., 2018; Lample et al., 2018). The last two 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 50 million sentences from the WMT monolingual News Crawl datasets from year 2014 till 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. We report results on *newstest* 2014 for $en - fr$, and *newstest* 2016 for $en - de$, $en - ro$ and $en - ru$.

We use the publicly available implementation of Moses[4] scripts for tokenization. NMT is trained with 60,000 BPE codes. PBSMT is trained with true-casing, and removing diacritics from Romanian on the source side to deal with their inconsistent use across the monolingual dataset (Sennrich et al., 2016).

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)[5] 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 related language pairs like $en - fr$ and $en - de$. We have estimated that more than 95% of the tokens are shared in these two language pairs. Next, we discuss how to initialize the phrase tables in PBSMT.

4.2.1 Phrase Table Initialization

For PBSMT, we generate n-gram embeddings on the source and target corpora independently, and align them using the MUSE library[6] (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.

Source	Target	$P(\text{s}\|\text{t})$	$P(\text{t}\|\text{s})$
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 translations for unigrams and bigrams, along with their corresponding conditional likelihoods $P(\text{s}|\text{t})$ and $P(\text{s}|\text{t})$.

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.

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. For both models, we share all parameters, including the lookup table BPE embeddings. The dimensionality of the embeddings and of the hidden layers is set to 512. For all models, we used 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

For PBSMT, we use Moses with phrase tables initialized as described in Section 4.2.1. The lan-

[4] `http://www.statmt.org/moses/`

[5] `http://fasttext.cc/`

[6] `https://github.com/facebookresearch/MUSE`

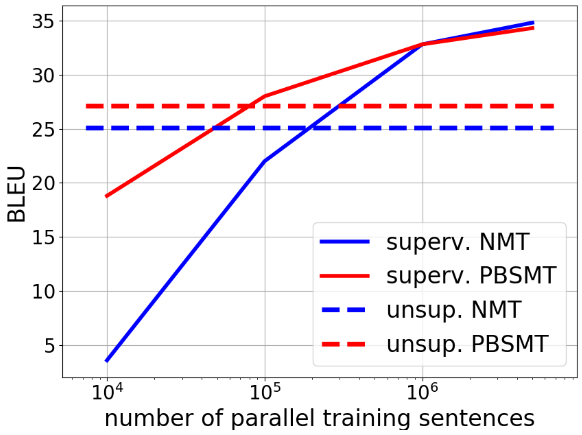

Figure 2: Comparison between supervised and unsupervised approaches on WMT'14 En-Fr, as we vary the number of parallel sentences for the supervised methods.

Model	en-fr	fr-en	de-en	en-de
(Artetxe et al., 2018)	15.1	15.6	-	-
(Lample et al., 2018)	15.0	14.3	13.3	9.6
NMT (LSTM)	24.5	23.7	19.6	14.7
NMT (Transformer)	25.1	24.2	21.0	17.2
PBSMT (Iter. 0)	16.1	15.4	14.5	10.3
PBSMT (Iter. n)	**27.1**	24.7	21.3	16.7
NMT + PBSMT	26.3	25.1	20.2	16.4
PBSMT + NMT	26.7	**27.1**	**23.6**	**19.2**

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.

guage model is a default smoothed n-gram language model and the reordering model is 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 3.

4.4 Model selection

Moses's implementation of PBSMT has 15 hyper-parameters, such as relative weighting of each scoring function, word penalty, etc. In this work, we consider two methods to set these hyper-parameters. 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 hyper-parameter values, and do not use any parallel resource whatsoever.

For the NMT, we also consider two model selection procedures: one based on the BLEU score of a "round-trip" translation (source → target → source and target → source → target) as in Lample et al. (2018), and one based on a small validation set of 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 defined in Lample et al. (2018).

4.5 Results

The results reported in Table 2 show that both our unsupervised NMT and PBSMT largely outperform previous unsupervised baselines. For instance, on the $en \rightarrow fr$ task, our unsupervised PBSMT obtains a BLEU score of 27.09, while Artetxe et al. (2018) only obtained 15.13 and Lample et al. (2018) 15.04. We report large gains on all languages pairs and in both directions. Even on a more complex task like $en \rightarrow de$, both PBSMT and NMT surpass the baseline score by more than 10 BLEU points. Note that the PBSMT model with the unsupervised phrase table alone (i.e. before starting back-translation), already significantly outperforms previous approaches, and can be generated in a few minutes using MUSE once the embeddings are learned with fastText.

The last rows of Table 2 also show that we can get additional gains by further tuning the NMT model on the data generated by PBSMT (PBSMT + NMT). Here, 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 a BLEU score of 19.16 on the challenging $en \rightarrow de$ translation task, and boost performance on the $de \rightarrow en$ task up to 23.62 points. We also tried to boostrap the PBSMT model with back-translated data gen-

	en → fr	fr→ en	en→ de	de→ en	en→ ro	ro→ en	en→ ru	ru→ en
Unsupervised phrase table	-	15.42	-	14.50	12.99	-	-	7.68
Back-translation - Iter. 1	24.09	23.65	15.06	20.86	20.17	17.12	10.58	14.35
Back-translation - Iter. 2	26.05	24.44	16.66	21.28	20.80	19.25	12.22	15.20
Back-translation - Iter. 3	26.52	24.48	**16.74**	**21.30**	**21.04**	19.84	**12.46**	**15.35**
Back-translation - Iter. 4	26.85	**24.67**	16.71	-	-	**20.07**	12.40	-
Back-translation - Iter. 5	**27.09**	-	-	-	-	-	-	-

Table 3: **Fully unsupervised PBSMT.** We report the BLEU score for PBSMT on 8 directed language pairs. Results are obtained on *newstest* 2014 for $en - fr$ and *newstest* 2016 for every other pair. Models created with the unsupervised phrase table obtain a relatively low performance, but can be used to generate back-translated data and train new models in a supervised way. After one iteration we observe up to 8 BLEU points improvement on $fr \rightarrow en$. The models converge after few iterations of back-translation.

erated by a NMT model (NMT + PBSMT), but this did not improve over the 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 datasets of varying number of training examples. The unsupervised PBSMT model is able to achieve the same performance than its supervised counterpart trained on almost 100,000 sentences. These unsupervised methods produce reasonable translation models at no labeling price, becoming a viable alternative for translating low-resource languages.

This is confirmed on a low-resource language like Romanian (Ro). In particular, on the $ro \rightarrow en$ language pair, our PBSMT model obtains a BLEU score of 21.0 without using a single parallel sentence, and 22.2 when using a small validation set to tune the weights of the model. As a comparison, Gu et al. (2018) obtain 22.9 BLEU by leveraging 6,000 parallel sentences, a seed dictionary, and a multi-NMT system combining parallel resources from 5 different languages.

Finally, we tested our unsupervised PBSMT on a very different language like Russian, and obtained a respectable BLEU score of 15.4 on $ru \rightarrow en$, showing that this approach works reasonably well also on distant languages.

Iterative back-translation: Table 3 illustrates the quality of the PBSMT model during the iterative training process, i.e. after each back-translation step. This highlights the importance of making multiple back-translation iterations.

For instance, in the $en - fr$ task, the $fr \rightarrow en$ model obtains a BLEU score of 15.42 at iteration 0 – i.e after the unsupervised phrase table construction – while it achieves a score of 24.67 at iteration 4. The same improvement can be observed in the different language pairs we have tested. As we iterate, the BLEU score steadily increases until saturation, showing the importance of iterating. Note that, for the $en \rightarrow de$ task, the increase is less pronounced - going from 15.06 at iteration 1 to 16.71 at iteration 4 - but still significant.

4.6 Ablation Study

To better understand the importance of each component of our model, we performed an ablation study of the NMT-Transformer model on the $fr \rightarrow en$ data. First, if we remove the denoising autoencoder term in the objective function, see Eq. 1, the model does not learn to translate at all. Similar catastrophic failure is observed if we remove the back-translation objective of Eq. 2. If we do not share the decoder, the performance on the validation set increases by half a BLEU point but it decreases by the same amount on the test set. Finally, if we do not initialize the model with pre-trained embeddings, the model does learn, but much slower and to a much lower accuracy, reaching a mere BLEU score of 10.5 as opposed to 25.1 of the model initialized according to Section 4.2.

4.7 Qualitative study

Table 4 shows examples of translations of French sentences from the French-English *newstest* 2014 dataset at different iterations of the learning algorithm for both the NMT and PBSMT models. Before the first iteration of back-translation, using only the unsupervised phrase table, the PBSMT translations are not far from word-by-word translations that do not respect the syntax of the target language, but still contain most of the semantic of the original sentences. As we increase the number of epochs in NMT and as we iterate for PBSMT, we observe a continuous improvement in the quality of the unsupervised translations. In-

Source	**Je rêve constamment d'eux, peut-être pas toutes les nuits mais plusieurs fois par semaine c'est certain.**
NMT Epoch 1	I constantly dream, but not all nights but by several times it is certain.
NMT Epoch 3	I continually dream them, perhaps not all but several times per week is certain.
NMT Epoch 45	I constantly dream of them, perhaps not all nights but several times a week it 's certain.
PBSMT Iter. 0	I dream of, but they constantly have all those nights but several times a week is too much. "
PBSMT Iter. 2	I had dreams constantly of them, probably not all nights but several times a week it is large.
PBSMT Iter. 8	I dream constantly of them, probably not all nights but several times a week it is certain.
Reference	**I constantly dream of them, perhaps not every night, but several times a week for sure.**
Source	**La protéine que nous utilisons dans la glace réagit avec la langue à pH neutre.**
NMT Epoch 1	The protein that we use in the ice with the language to pH.
NMT Epoch 8	The protein we use into the ice responds with language to pH neutral.
NMT Epoch 45	The protein we use in ice responds with the language from pH to neutral.
PBSMT Iter. 0	The protein that used in the ice responds with the language and pH neutral.
PBSMT Iter. 2	The protein that we use in the ice responds with the language to pH neutral.
PBSMT Iter. 8	The protein that we use in the ice reacts with the language to a neutral pH.
Reference	**The protein we are using in the ice cream reacts with your tongue at neutral pH.**
Source	**Selon Google, les déguisements les plus recherchés sont les zombies, Batman, les pirates et les sorcières.**
NMT Epoch 1	According to Google, there are more than zombies, Batman, and the pirates.
NMT Epoch 8	Google's most wanted outfits are the zombies, Batman, the pirates and the evil.
NMT Epoch 45	Google said the most wanted outfits are the zombies, Batman, the pirates and the witch.
PBSMT Iter. 0	According to Google, fancy dress and most wanted fugitives are the bad guys, Wolverine, the pirates and their minions.
PBSMT Iter. 2	According to Google, the outfits are the most wanted fugitives are zombies, Batman, pirates and witches.
PBSMT Iter. 8	According to Google, the outfits, the most wanted list are zombies, Batman, pirates and witches.
Reference	**According to Google, the highest searched costumes are zombies, Batman, pirates and witches.**

Table 4: **Unsupervised translations.** Examples of translations on the French-English pair of *newstest* 2014 at different iterations of training. For PBSMT, we show translations at iterations 0, 1 and 4, where the model obtains BLEU scores of 15.4, 23.7 and 24.7 respectively. For NMT, we show examples of translations after epochs 1, 8 and 42, where the model obtains BLEU scores of 12.3, 17.5 and 24.2 respectively. Iteration 0 refers to the PBSMT model obtained using the unsupervised phrase table, and an epoch corresponds to training the NMT model on 500k monolingual sentences. At the end of training, both models generate very good translations.

terestingly, in the second example, both the PBSMT and NMT models fail to adapt to the polysemy of the French word "langue", which can be translated as "tongue" or "language" in English. These translations were both present in the unsupervised phrase table, but the conditional probability of "language" to be the correct translation of "langue" was very high compared to the one of "tongue": $P(\text{language}|\text{langue}) = 0.92$, while $P(\text{tongue}|\text{langue}) = 0.0005$. As a comparison, the phrase table of a Moses model trained in a supervised way contains $P(\text{language}|\text{langue}) = 0.633, P(\text{tongue}|\text{langue}) = 0.0076$, giving a higher probability for "langue" to be properly translated. This underlines the importance of the initial unsupervised phrase alignment procedure.

5 Related Work

Learning to translate without supervision has been a long standing research problem for the MT community. The first known attempt at fully unsupervised machine translation is the work by Ravi and Knight (2011), who leverage linguistic prior knowledge to reframe the task as an instance of deciphering and demonstrate the feasibility on short sentences with limited vocabulary. Even earlier work by Carbonell et al. (2006) also aimed at unsupervised machine translation, but leveraged a bilingual dictionary to seed the translation. Both works rely on a language model on the target side to correct for fluency of the translation.

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

Using monolingual data on the target side for data augmentation purposes has been a major advance in recent years through a method called back-translation (Sennrich et al., 2015a). In back-translation, a model trained from the target to the source generates translations that are added to the regular training set in order to regularize the model. This method is perhaps the most effective way to leverage monolingual data in the semisupervised setting, and it has been integrated in the "dual learning" framework of He et al. (2016) with later extensions (Wang et al., 2018). Our approach is similar to the dual learning framework, except that their model is pretrained using a relatively large amount of labeled data and gradients are backpropagated through the reverse model, whereas our approach is fully unsupervised.

Finally, recent work by Lample et al. (2018) and Artetxe et al. (2018) have achieved *fully unsupervised* machine translation on large-scale benchmark datasets by leveraging bilingual dictionaries that are also learned without supervision (Conneau et al., 2018; Artetxe et al., 2017). Additionally, these works depend on back-translation, strong language models (implemented via denoising autoencoders), clever initialization of lookup tables, weight sharing between encoders, and an adversarial training loss to align the latent representations (encoder output) between language pairs. Our approach is different in several ways. First, we perform online backtranslation in both directions, similarly to Artetxe et al. (2018) and the dual learning framework. Second, we adopt a simpler scheme for aligning latent representations between language pairs, by recognizing that a significant fraction of words and sub-word (BPE) tokens (Sennrich et al., 2015b; Press and Wolf, 2016) are usually shared between related languages. In particular, we find that this overlap is often sufficient to align the latent representations without requiring complex initialization schemes or any adversarial loss terms.

6 Conclusions and Future Work

In this work, we synthesize three principles underlying recent successes in fully unsupervised machine translation: (1) proper initialization (e.g., by inferring a bilingual dictionary); (2) leveraging a strong language model; and (3) iterative training with artificially-generated parallel data through back-translation. Using these principles, we propose both a simplified neural model and a novel phrase-based model for unsupervised MT. These models achieve state of the art translation performance across multiple benchmark datasets and language pairs, in some cases improving upon the previous best models by +12 BLEU points. We then show that by combining the neural and phrase-based models we can improve performance even further.

In the future, we plan to further investigate the initialization of phrase tables with n-grams aligned in an unsupervised way. Finally, we plan to study how these methods can be extended to the semi-supervised setting and to settings where we may have significant quantities of labeled data in other language pairs.

Acknowledgments

The authors would like to thank Hervé Jegou for his guidance and suggestions, and also Arthur Szlam for helpful discussions.

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.

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 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.

Anoop Kunchukuttan and Pushpak Bhattacharyya. 2016. Orthographic syllable as basic unit for smt between related languages. In Conference on Empirical Methods in Natural Language Processing.

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

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.

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.

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.