An Effective Approach to Unsupervised Machine Translation

Mikel Artetxe, Gorka Labaka, Eneko Agirre
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe, gorka.labaka, e.agirre}@ehu.eus

Abstract

While machine translation has traditionally relied on large amounts of parallel corpora, a recent research line has managed to train both Neural Machine Translation (NMT) and Statistical Machine Translation (SMT) systems using monolingual corpora only. In this paper, we identify and address several deficiencies of existing unsupervised SMT approaches by exploiting subword information, developing a theoretically well founded unsupervised tuning method, and incorporating a joint refinement procedure. Moreover, we use our improved SMT system to initialize a dual NMT model, which is further fine-tuned through on-the-fly back-translation. Together, we obtain large improvements over the previous state-of-the-art in unsupervised machine translation. For instance, we get 22.5 BLEU points in English-to-German WMT 2014, 5.5 points more than the previous best unsupervised system, and 0.5 points more than the (supervised) shared task winner back in 2014.

1 Introduction

The recent advent of neural sequence-to-sequence modeling has resulted in significant progress in the field of machine translation, with large improvements in standard benchmarks (Vaswani et al., 2017; Edunov et al., 2018) and the first solid claims of human parity in certain settings (Hassan et al., 2018). Unfortunately, these systems rely on large amounts of parallel corpora, which are only available for a few combinations of major languages like English, German and French.

Aiming to remove this dependency on parallel data, a recent research line has managed to train unsupervised machine translation systems using monolingual corpora only. The first such systems were based on Neural Machine Translation (NMT), and combined denoising autoencoding and back-translation to train a dual model initialized with cross-lingual embeddings (Artetxe et al., 2018c; Lample et al., 2018a). Nevertheless, these early systems were later superseded by Statistical Machine Translation (SMT) based approaches, which induced an initial phrase-table through cross-lingual embedding mappings, combined it with an n-gram language model, and further improved the system through iterative back-translation (Lample et al., 2018b; Artetxe et al., 2018b).

In this paper, we develop a more principled approach to unsupervised SMT, addressing several deficiencies of previous systems by incorporating subword information, applying a theoretically well founded unsupervised tuning method, and developing a joint refinement procedure. In addition to that, we use our improved SMT approach to initialize an unsupervised NMT system, which is further improved through on-the-fly back-translation.

Our experiments on WMT 2014/2016 French-English and German-English show the effectiveness of our approach, as our proposed system outperforms the previous state-of-the-art in unsupervised machine translation by 5-7 BLEU points in all these datasets and translation directions. Our system also outperforms the supervised WMT 2014 shared task winner in English-to-German, and is around 2 BLEU points behind it in the rest of translation directions, suggesting that unsupervised machine translation can be a usable alternative in practical settings.

The remaining of this paper is organized as follows. Section 2 first discusses the related work in the topic. Section 3 then describes our principled unsupervised SMT method, while Section 4 discusses our hybridization method with NMT. We then present the experiments done and the results obtained in Section 5, and Section 6 concludes the paper.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 194–203
Florence, Italy, July 28 - August 2, 2019.

2 Related work

Early attempts to build machine translation systems with monolingual corpora go back to statistical decipherment (Ravi and Knight, 2011; Dou and Knight, 2012). These methods see the source language as ciphertext produced by a noisy channel model that first generates the original English text and then probabilistically replaces the words in it. The English generative process is modeled using an n-gram language model, and the channel model parameters are estimated using either expectation maximization or Bayesian inference. This basic approach was later improved by incorporating syntactic knowledge (Dou and Knight, 2013) and word embeddings (Dou et al., 2015). Nevertheless, these methods were only shown to work in limited settings, being most often evaluated in word-level translation.

More recently, the task got a renewed interest after the concurrent work of Artetxe et al. (2018c) and Lample et al. (2018a) on unsupervised NMT which, for the first time, obtained promising results in standard machine translation benchmarks using monolingual corpora only. Both methods build upon the recent work on unsupervised cross-lingual embedding mappings, which independently train word embeddings in two languages and learn a linear transformation to map them to a shared space through self-learning (Artetxe et al., 2017, 2018a) or adversarial training (Conneau et al., 2018). The resulting cross-lingual embeddings are used to initialize a shared encoder for both languages, and the entire system is trained using a combination of denoising autoencoding, back-translation and, in the case of Lample et al. (2018a), adversarial training. This method was further improved by Yang et al. (2018), who use two language-specific encoders sharing only a subset of their parameters, and incorporate a local and a global generative adversarial network. Concurrent to our work, Lample and Conneau (2019) report strong results initializing an unsupervised NMT system with a cross-lingual language model.

Following the initial work on unsupervised NMT, it was argued that the modular architecture of phrase-based SMT was more suitable for this problem, and Lample et al. (2018b) and Artetxe et al. (2018b) adapted the same principles discussed above to train an unsupervised SMT model, obtaining large improvements over the original unsupervised NMT systems. More concretely, both approaches learn cross-lingual n-gram embeddings from monolingual corpora based on the mapping method discussed earlier, and use them to induce an initial phrase-table that is combined with an n-gram language model and a distortion model. This initial system is then refined through iterative back-translation (Sennrich et al., 2016) which, in the case of Artetxe et al. (2018b), is preceded by an unsupervised tuning step. Our work identifies some deficiencies in these previous systems, and proposes a more principled approach to unsupervised SMT that incorporates subword information, uses a theoretically better founded unsupervised tuning method, and applies a joint refinement procedure, outperforming these previous systems by a substantial margin.

Very recently, some authors have tried to combine both SMT and NMT to build hybrid unsupervised machine translation systems. This idea was already explored by Lample et al. (2018b), who aided the training of their unsupervised NMT system by combining standard back-translation with synthetic parallel data generated by unsupervised SMT. Marie and Fujita (2018) go further and use synthetic parallel data from unsupervised SMT to train a conventional NMT system from scratch. The resulting NMT model is then used to augment the synthetic parallel corpus through back-translation, and a new NMT model is trained on top of it from scratch, repeating the process iteratively. Ren et al. (2019) follow a similar approach, but use SMT as posterior regularization at each iteration. As shown later in our experiments, our proposed NMT hybridization obtains substantially larger absolute gains than all these previous approaches, even if our initial SMT system is stronger and thus more challenging to improve upon.

3 Principled unsupervised SMT

Phrase-based SMT is formulated as a log-linear combination of several statistical models: a translation model, a language model, a reordering model and a word/phrase penalty. As such, building an unsupervised SMT system requires learning these different components from monolingual corpora. As it turns out, this is straightforward for most of them: the language model is learned from monolingual corpora by definition; the word and phrase penalties are parameterless; and one

can drop the standard lexical reordering model at a small cost and do with the distortion model alone, which is also parameterless. This way, the main challenge left is learning the translation model, that is, building the phrase-table.

Our proposed method starts by building an initial phrase-table through cross-lingual embedding mappings (Section 3.1). This initial phrase-table is then extended by incorporating subword information, addressing one of the main limitations of previous unsupervised SMT systems (Section 3.2). Having done that, we adjust the weights of the underlying log-linear model through a novel unsupervised tuning procedure (Section 3.3). Finally, we further improve the system by jointly refining two models in opposite directions (Section 3.4).

3.1 Initial phrase-table

So as to build our initial phrase-table, we follow Artetxe et al. (2018b) and learn n-gram embeddings for each language independently, map them to a shared space through self-learning, and use the resulting cross-lingual embeddings to extract and score phrase pairs.

More concretely, we train our n-gram embeddings using *phrase2vec*[1], a simple extension of skip-gram that applies the standard negative sampling loss of Mikolov et al. (2013) to bigram-context and trigram-context pairs in addition to the usual word-context pairs.[2] Having done that, we map the embeddings to a cross-lingual space using VecMap[3] with *identical* initialization (Artetxe et al., 2018a), which builds an initial solution by aligning identical words and iteratively improves it through self-learning. Finally, we extract translation candidates by taking the 100 nearest-neighbors of each source phrase, and score them by applying the softmax function over their cosine similarities:

$$\phi(\bar{f}|\bar{e}) = \frac{\exp\left(\cos(\bar{e}, \bar{f})/\tau\right)}{\sum_{\bar{f}'} \exp\left(\cos(\bar{e}, \bar{f}')/\tau\right)}$$

where the temperature τ is estimated using maximum likelihood estimation over a dictionary induced in the reverse direction. In addition to the phrase translation probabilities in both directions, the forward and reverse lexical weightings are also estimated by aligning each word in the target phrase with the one in the source phrase most likely generating it, and taking the product of their respective translation probabilities. The reader is referred to Artetxe et al. (2018b) for more details.

3.2 Adding subword information

An inherent limitation of existing unsupervised SMT systems is that words are taken as atomic units, making it impossible to exploit character-level information. This is reflected in the known difficulty of these models to translate named entities, as it is very challenging to discriminate among related proper nouns based on distributional information alone, yielding to translation errors like *"Sunday Telegraph"* $\rightarrow$ *"The Times of London"* (Artetxe et al., 2018b).

So as to overcome this issue, we propose to incorporate subword information once the initial alignment is done at the word/phrase level. For that purpose, we add two additional weights to the initial phrase-table that are analogous to the lexical weightings, but use a character-level similarity function instead of word translation probabilities:

$$\text{score}(\bar{f}|\bar{e}) = \prod_i \max\left(\epsilon, \max_j \text{sim}(\bar{f}_i, \bar{e}_j)\right)$$

where $\epsilon = 0.3$ guarantees a minimum similarity score, as we want to favor translation candidates that are similar at the character level without excessively penalizing those that are not. In our case, we use a simple similarity function that normalizes the Levenshtein distance $\text{lev}(\cdot)$ (Levenshtein, 1966) by the length of the words $\text{len}(\cdot)$:

$$\text{sim}(f, e) = 1 - \frac{\text{lev}(f, e)}{\max(\text{len}(f), \text{len}(e))}$$

We leave the exploration of more elaborated similarity functions and, in particular, learnable metrics (McCallum et al., 2005), for future work.

3.3 Unsupervised tuning

Having trained the underlying statistical models independently, SMT tuning aims to adjust the weights of their resulting log-linear combination to optimize some evaluation metric like BLEU in a parallel validation corpus, which is typically done through Minimum Error Rate Training or MERT (Och, 2003). Needless to say, this cannot be done in strictly unsupervised settings, but we argue that

[1] https://github.com/artetxem/phrase2vec

[2] So as to keep the model size within a reasonable limit, we restrict the vocabulary to the most frequent 200,000 unigrams, 400,000 bigrams and 400,000 trigrams.

[3] https://github.com/artetxem/vecmap

it would still be desirable to optimize some unsupervised criterion that is expected to correlate well with test performance. Unfortunately, neither of the existing unsupervised SMT systems do so: Artetxe et al. (2018b) use a heuristic that builds two initial models in opposite directions, uses one of them to generates a synthetic parallel corpus through back-translation (Sennrich et al., 2016), and applies MERT to tune the model in the reverse direction, iterating until convergence, whereas Lample et al. (2018b) do not perform any tuning at all. In what follows, we propose a more principled approach to tuning that defines an unsupervised criterion and an optimization procedure that is guaranteed to converge to a local optimum of it.

Inspired by the previous work on CycleGANs (Zhu et al., 2017) and dual learning (He et al., 2016), our method takes two initial models in opposite directions, and defines an **unsupervised optimization objective** that combines a cyclic consistency loss and a language model loss over the two monolingual corpora E and F:

$$L = L_{cycle}(E) + L_{cycle}(F) + L_{lm}(E) + L_{lm}(F)$$

The cyclic consistency loss captures the intuition that the translation of a translation should be close to the original text. So as to quantify this, we take a monolingual corpus in the source language, translate it to the target language and back to the source language, and compute its BLEU score taking the original text as reference:

$$L_{cycle}(E) = 1 - \mathrm{BLEU}(\mathrm{T}_{F\to E}(\mathrm{T}_{E\to F}(E)), E)$$

At the same time, the language model loss captures the intuition that machine translation should produce fluent text in the target language. For that purpose, we estimate the per-word entropy in the target language corpus using an n-gram language model, and penalize higher per-word entropies in machine translated text as follows:[4]

$$L_{lm}(E) = \mathrm{LP} \cdot \max(0, \mathrm{H}(F) - \mathrm{H}(\mathrm{T}_{E\to F}(E)))^2$$

where the length penalty $\mathrm{LP} = \mathrm{LP}(E) \cdot \mathrm{LP}(F)$ penalizes excessively long translations:[5]

$$\mathrm{LP}(E) = \max\left(1, \frac{\mathrm{len}(\mathrm{T}_{F\to E}(\mathrm{T}_{E\to F}(E)))}{\mathrm{len}(E)}\right)$$

So as to minimize the combined loss function, we **adapt MERT to jointly optimize** the parameters of the two models. In its basic form, MERT approximates the search space for each source sentence through an n-best list, and performs a form of coordinate descent by computing the optimal value for each parameter through an efficient line search method and greedily taking the step that leads to the largest gain. The process is repeated iteratively until convergence, augmenting the n-best list with the updated parameters at each iteration so as to obtain a better approximation of the full search space. Given that our optimization objective combines two translation systems $\mathrm{T}_{F\to E}(\mathrm{T}_{E\to F}(E))$, this would require generating an n-best list for $\mathrm{T}_{E\to F}(E)$ first and, for each entry on it, generating a new n-best list with $\mathrm{T}_{F\to E}$, yielding a combined n-best list with N^2 entries. So as to make it more efficient, we propose an alternating optimization approach where we fix the parameters of one model and optimize the other with standard MERT. Thanks to this, we do not need to expand the search space of the fixed model, so we can do with an n-best list of N entries alone. Having done that, we fix the parameters of the opposite model and optimize the other, iterating until convergence.

3.4 Joint refinement

Constrained by the lack of parallel corpora, the procedure described so far makes important simplifications that could compromise its potential performance: its phrase-table is somewhat unnatural (e.g. the translation probabilities are estimated from cross-lingual embeddings rather than actual frequency counts) and it lacks a lexical reordering model altogether. So as to overcome this issue, existing unsupervised SMT methods generate a synthetic parallel corpus through back-translation and use it to train a standard SMT system from scratch, iterating until convergence.

[4] We initially tried to directly minimize the entropy of the generated text, but this worked poorly in our preliminary experiments on English-Spanish (note that we used this language pair exclusively for development to be faithful to our unsupervised scenario at test time). More concretely, the behavior of the optimization algorithm was very unstable, as it tended to excessively focus on either the cyclic consistency loss or the language model loss at the cost of the other, and we found it very difficult to find the right balance between the two factors.

[5] Without this penalization, the system tended to produce unnecessary tokens (e.g. quotes) that looked natural in their context, which served to minimize the per-word perplexity of the output. Minimizing the overall perplexity instead of the per-word perplexity did not solve the problem, as the opposite phenomenon arose (i.e. the system tended to produce excessively short translations).

An obvious drawback of this approach is that the back-translated side will contain ungrammatical n-grams and other artifacts that will end up in the induced phrase-table. One could argue that this should be innocuous as long as the ungrammatical n-grams are in the source side, as they should never occur in real text and their corresponding entries in the phrase-table should therefore not be used. However, ungrammatical source phrases do ultimately affect the estimation of the backward translation probabilities, including those of grammatical phrases.[6] We argue that, ultimately, the backward probability estimations can only be meaningful when all source phrases are grammatical (so the probabilities of all plausible translations sum to one) and, similarly, the forward probability estimations can only be meaningful when all target phrases are grammatical.

Following the above observation, we propose an alternative approach that jointly refines both translation directions. More concretely, we use the initial systems to build two synthetic corpora in opposite directions.[7] Having done that, we independently extract phrase pairs from each synthetic corpus, and build a phrase-table by taking their intersection. The forward probabilities are estimated in the parallel corpus with the synthetic source side, while the backward probabilities are estimated in the one with the synthetic target side. This does not only guarantee that the probability estimates are meaningful as discussed previously, but it also discards the ungrammatical phrases altogether, as both the source and the target n-grams must have occurred in the original monolingual texts to be present in the resulting phrase-table. This phrase-table is then combined with a lexical reordering model learned on the synthetic parallel corpus in the reverse direction, and we apply the unsupervised tuning method described in Section 3.3 to adjust the weights of the resulting system. We repeat this process for a total of 3 iterations.[8]

[6]For instance, let's say that the target phrase *"dos gatos"* has been aligned 10 times with *"two cats"* and 90 times with *"two cat"*. While the ungrammatical phrase-table entry *two cat- dos gatos* should never be picked, the backward probability estimation of *two cats - dos gatos* is still affected by it (it would be 0.1 instead of 1.0 in this example).

[7]For efficiency purposes, we restrict the size of each synthetic parallel corpus to 10 million sentence pairs.

[8]For the last iteration, we do not perform any tuning and use default Moses weights instead, which we found to be more robust during development. Note, however, that using unsupervised tuning during the previous steps was still strongly beneficial.

4 NMT hybridization

While the rigid and modular design of SMT provides a very suitable framework for unsupervised machine translation, NMT has shown to be a fairly superior paradigm in supervised settings, outperforming SMT by a large margin in standard benchmarks. As such, the choice of SMT over NMT also imposes a hard ceiling on the potential performance of these approaches, as unsupervised SMT systems inherit the very same limitations of their supervised counterparts (e.g. the locality and sparsity problems). For that reason, we argue that SMT provides a more appropriate architecture to find an initial alignment between the languages, but NMT is ultimately a better architecture to model the translation process.

Following this observation, we propose a hybrid approach that uses unsupervised SMT to warm up a dual NMT model trained through iterative back-translation. More concretely, we first train two SMT systems in opposite directions as described in Section 3, and use them to assist the training of another two NMT systems in opposite directions. These NMT systems are trained following an iterative process where, at each iteration, we alternately update the model in each direction by performing a single pass over a synthetic parallel corpus built through back-translation (Sennrich et al., 2016).[9] In the first iteration, the synthetic parallel corpus is entirely generated by the SMT system in the opposite direction but, as training progresses and the NMT models get better, we progressively switch to a synthetic parallel corpus generated by the reverse NMT model. More concretely, iteration t uses $N_{smt} = N \cdot \max(0, 1 - t/a)$ synthetic parallel sentences from the reverse SMT system, where the parameter a controls the number of transition iterations from SMT to NMT back-translation. The remaining $N - N_{smt}$ sentences are generated by the reverse NMT model. Inspired by Edunov et al. (2018), we use greedy decoding for half of them, which produces more fluent and predictable translations, and random sampling for the other half, which produces more varied translations. In our experiments, we use $N = 1,000,000$ and $a = 30$, and perform a total of 60 such iterations. At test time, we use beam search decoding with an ensemble of all check-

[9]Note that we do not train a new model from scratch each time, but continue training the model from the previous iteration.

		WMT-14				WMT-16	
		fr-en	en-fr	de-en	en-de	de-en	en-de
NMT	Artetxe et al. (2018c)	15.6	15.1	10.2	6.6	-	-
	Lample et al. (2018a)	14.3	15.1	-	-	13.3	9.6
	Yang et al. (2018)	15.6	17.0	-	-	14.6	10.9
	Lample et al. (2018b)	24.2	25.1	-	-	21.0	17.2
SMT	Artetxe et al. (2018b)	25.9	26.2	17.4	14.1	23.1	18.2
	Lample et al. (2018b)	27.2	28.1	-	-	22.9	17.9
	Marie and Fujita (2018)*	-	-	-	-	20.2	15.5
	Proposed system	28.4	30.1	20.1	15.8	25.4	19.7
	*detok. SacreBLEU**	27.9	27.8	19.7	14.7	24.8	19.4
SMT + NMT	Lample et al. (2018b)	27.7	27.6	-	-	25.2	20.2
	Marie and Fujita (2018)*	-	-	-	-	26.7	20.0
	Ren et al. (2019)	28.9	29.5	20.4	17.0	26.3	21.7
	Proposed system	**33.5**	**36.2**	**27.0**	**22.5**	**34.4**	**26.9**
	*detok. SacreBLEU**	33.2	33.6	26.4	21.2	33.8	26.4

Table 1: Results of the proposed method in comparison to previous work (BLEU). Overall best results are in bold, the best ones in each group are underlined.
*Detokenized BLEU equivalent to the official `mteval-v13a.pl` script. The rest use tokenized BLEU with `multi-bleu.perl` (or similar).

points from every 10 iterations.

5 Experiments and results

In order to make our experiments comparable to previous work, we use the French-English and German-English datasets from the WMT 2014 shared task. More concretely, our training data consists of the concatenation of all News Crawl monolingual corpora from 2007 to 2013, which make a total of 749 million tokens in French, 1,606 millions in German, and 2,109 millions in English, from which we take a random subset of 2,000 sentences for tuning (Section 3.3). Preprocessing is done using standard Moses tools, and involves punctuation normalization, tokenization with aggressive hyphen splitting, and truecasing.

Our SMT implementation is based on Moses[10], and we use the KenLM (Heafield et al., 2013) tool included in it to estimate our 5-gram language model with modified Kneser-Ney smoothing. Our unsupervised tuning implementation is based on Z-MERT (Zaidan, 2009), and we use FastAlign (Dyer et al., 2013) for word alignment within the joint refinement procedure. Finally, we use the big transformer implementation from fairseq[11] for our NMT system, training with a total batch size of 20,000 tokens across 8 GPUs with the exact same hyperparameters as Ott et al. (2018).

We use newstest2014 as our test set for French-English, and both newstest2014 and newstest2016 (from WMT 2016[12]) for German-English. Following common practice, we report tokenized BLEU scores as computed by the `multi-bleu.perl` script included in Moses. In addition to that, we also report detokenized BLEU scores as computed by SacreBLEU[13] (Post, 2018), which is equivalent to the official `mteval-v13a.pl` script.

We next present the results of our proposed system in comparison to previous work in Section 5.1. Section 5.2 then compares the obtained results to those of different supervised systems. Finally, Section 5.3 presents some translation examples from our system.

5.1 Main results

Table 1 reports the results of the proposed system in comparison to previous work. As it can be seen, our full system obtains the best published results in all cases, outperforming the previous state-of-the-art by 5-7 BLEU points in all datasets and translation directions.

A substantial part of this improvement comes from our more principled unsupervised SMT ap-

[10] http://www.statmt.org/moses/
[11] https://github.com/pytorch/fairseq

[12] Note that it is only the test set that is from WMT 2016. All the training data comes from WMT 2014 News Crawl, so it is likely that our results could be further improved by using the more extensive monolingual corpora from WMT 2016.

[13] SacreBLEU signature: BLEU+case.mixed+lang.LANG +numrefs.1+smooth.exp+test.TEST+tok.13a+version.1.2.11, with LANG ∈ {fr-en, en-fr, de-en, en-de} and TEST ∈ {wmt14/full, wmt16}

		WMT-14		WMT-16	
		fr-en	en-fr	de-en	en-de
Lample et al. (2018b)	Initial SMT	27.2	28.1	22.9	17.9
	+ NMT hybrid	27.7 (+0.5)	27.6 (-0.5)	25.2 (+2.3)	20.2 (+2.3)
Marie and Fujita (2018)	Initial SMT	-	-	20.2	15.5
	+ NMT hybrid	-	-	26.7 (+6.5)	20.0 (+4.5)
Proposed system	Initial SMT	28.4	30.1	25.4	19.7
	+ NMT hybrid	**33.5** (+5.1)	**36.2** (+6.1)	**34.4** (+9.0)	**26.9** (+7.2)

Table 2: NMT hybridization results for different unsupervised machine translation systems (BLEU).

		WMT-14			
		fr-en	en-fr	de-en	en-de
Unsupervised	Proposed system	33.5	36.2	27.0	22.5
	*detok. SacreBLEU**	33.2	33.6	26.4	21.2
Supervised	WMT best*	35.0	35.8	29.0	20.6†
	Vaswani et al. (2017)	-	41.0	-	28.4
	Edunov et al. (2018)	-	45.6	-	35.0

Table 3: Results of the proposed method in comparison to different supervised systems (BLEU).
*Detokenized BLEU equivalent to the official `mteval-v13a.pl` script. The rest use tokenized BLEU with `multi-bleu.perl` (or similar).
†Results in the original test set from WMT 2014, which slightly differs from the full test set used in all subsequent work. Our proposed system obtains 22.4 BLEU points (21.1 detokenized) in that same subset.

proach, which outperforms all previous SMT-based systems by around 2 BLEU points. Nevertheless, it is the NMT hybridization that brings the largest gains, improving the results of this initial SMT systems by 5-9 BLEU points. As shown in Table 2, our absolute gains are considerably larger than those of previous hybridization methods, even if our initial SMT system is substantially better and thus more difficult to improve upon. This way, our initial SMT system is about 4-5 BLEU points above that of Marie and Fujita (2018), yet our absolute gain on top of it is around 2.5 BLEU points higher. When compared to Lample et al. (2018b), we obtain an absolute gain of 5-6 BLEU points in both French-English directions while they do not get any clear improvement, and we obtain an improvement of 7-9 BLEU points in both German-English directions, in contrast with the 2.3 BLEU points they obtain.

More generally, it is interesting that pure SMT systems perform better than pure NMT systems, yet the best results are obtained by initializing an NMT system with an SMT system. This suggests that the rigid and modular architecture of SMT might be more suitable to find an initial alignment between the languages, but the final system should be ultimately based on NMT for optimal results.

5.2 Comparison with supervised systems

So as to put our results into perspective, Table 3 reports the results of different supervised systems in the same WMT 2014 test set. More concretely, we include the best results from the shared task itself, which reflect the state-of-the-art in machine translation back in 2014; those of Vaswani et al. (2017), who introduced the now predominant transformer architecture; and those of Edunov et al. (2018), who apply back-translation at a large scale and, to the best of our knowledge, hold the current best results in the test set.

As it can be seen, our unsupervised system outperforms the WMT 2014 shared task winner in English-to-German, and is around 2 BLEU points behind it in the other translation directions. This shows that unsupervised machine translation is already competitive with the state-of-the-art in supervised machine translation in 2014. While the field of machine translation has undergone great progress in the last 5 years, and the gap between our unsupervised system and the current state-of-the-art in supervised machine translation is still large as reflected by the other results, this suggests that unsupervised machine translation can be a usable alternative in practical settings.

Source	Reference	Artetxe et al. (2018b)	Proposed system
D'autres révélations ont fait état de documents divulgués par Snowden selon lesquels la NSA avait intercepté des données et des communications émanant du téléphone portable de la chancelière allemande Angela Merkel et de ceux de 34 autres chefs d'État.	Other revelations cited documents leaked by Snowden that the NSA monitored German Chancellor Angela Merkel's cellphone and those of up to 34 other world leaders.	Other disclosures have reported documents disclosed by Snowden suggested the NSA had intercepted communications and data from the mobile phone of German Chancellor Angela Merkel and those of 32 other heads of state.	Other revelations have pointed to documents disclosed by Snowden that the NSA had intercepted data and communications emanating from German Chancellor Angela Merkel's mobile phone and those of 34 other heads of state.
La NHTSA n'a pas pu examiner la lettre d'information aux propriétaires en raison de l'arrêt de 16 jours des activités gouvernementales, ce qui a ralenti la croissance des ventes de véhicules en octobre.	NHTSA could not review the owner notification letter due to the 16-day government shutdown, which tempered auto sales growth in October.	The NHTSA could not consider the letter of information to owners because of halting 16-day government activities, which slowed the growth in vehicle sales in October.	NHTSA said it could not examine the letter of information to owners because of the 16-day halt in government operations, which slowed vehicle sales growth in October.
Le M23 est né d'une mutinerie, en avril 2012, d'anciens rebelles, essentiellement tutsi, intégrés dans l'armée en 2009 après un accord de paix.	The M23 was born of an April 2012 mutiny by former rebels, principally Tutsis who were integrated into the army in 2009 following a peace agreement.	M23 began as a mutiny in April 2012, former rebels, mainly Tutsi integrated into the national army in 2009 after a peace deal.	The M23 was born into a mutiny in April 2012, of former rebels, mostly Tutsi, embedded in the army in 2009 after a peace deal.
Tunks a déclaré au Sunday Telegraph de Sydney que toute la famille était «extrêmement préoccupée» du bien-être de sa fille et voulait qu'elle rentre en Australie.	Tunks told Sydney's Sunday Telegraph the whole family was "extremely concerned" about his daughter's welfare and wanted her back in Australia.	Tunks told The Times of London from Sydney that the whole family was "extremely concerned" of the welfare of her daughter and wanted it to go in Australia.	Tunks told the Sunday Telegraph in Sydney that the whole family was "extremely concerned" about her daughter's well-being and wanted her to go into Australia.

Table 4: Randomly chosen translation examples from French→English newstest2014 in comparison of those reported by Artetxe et al. (2018b).

5.3 Qualitative results

Table 4 shows some translation examples from our proposed system in comparison to those reported by Artetxe et al. (2018b). We choose the exact same sentences reported by Artetxe et al. (2018b), which were randomly taken from newstest2014, so they should be representative of the general behavior of both systems.

While not perfect, our proposed system produces generally fluent translations that accurately capture the meaning of the original text. Just in line with our quantitative results, this suggests that unsupervised machine translation can be a usable alternative in practical settings.

Compared to Artetxe et al. (2018b), our translations are generally more fluent, which is not surprising given that they are produced by an NMT system rather than an SMT system. In addition to that, the system of Artetxe et al. (2018b) has some adequacy issues when translating named entities and numerals (e.g. *34* → *32*, *Sunday Telegraph* → *The Times of London*), which we do not observe for our proposed system in these examples.

6 Conclusions and future work

In this paper, we identify several deficiencies in previous unsupervised SMT systems, and propose a more principled approach that addresses them by incorporating subword information, using a theoretically well founded unsupervised tuning method, and developing a joint refinement procedure. In addition to that, we use our improved SMT approach to initialize a dual NMT model that is further improved through on-the-fly back-translation. Our experiments show the effectiveness of our approach, as we improve the previous state-of-the-art in unsupervised machine translation by 5-7 BLEU points in French-English and German-English WMT 2014 and 2016. Our code is available as an open source project at `https://github.com/artetxem/monoses`.

In the future, we would like to explore learnable similarity functions like the one proposed by (McCallum et al., 2005) to compute the character-level scores in our initial phrase-table. In addition to that, we would like to incorporate a language modeling loss during NMT training similar to He

et al. (2016). Finally, we would like to adapt our approach to more relaxed scenarios with multiple languages and/or small parallel corpora.

Acknowledgments

This research was partially supported by the Spanish MINECO (UnsupNMT TIN2017-91692-EXP and DOMINO PGC2018-102041-B-I00, co-funded by EU FEDER), the BigKnowledge project (BBVA foundation grant 2018), the UPV/EHU (excellence research group), and the NVIDIA GPU grant program. Mikel Artetxe was supported by a doctoral grant from the Spanish MECD.

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: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018c. Unsupervised neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Qing Dou and Kevin Knight. 2012. Large scale decipherment for out-of-domain machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 266–275, Jeju Island, Korea. Association for Computational Linguistics.

Qing Dou and Kevin Knight. 2013. Dependency-based decipherment for resource-limited machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1668–1676, Seattle, Washington, USA. Association for Computational Linguistics.

Qing Dou, Ashish Vaswani, Kevin Knight, and Chris Dyer. 2015. Unifying bayesian inference and vector space models for improved decipherment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 836–845, Beijing, China. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems 29*, pages 820–828.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified kneser-ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b.

Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Benjamin Marie and Atsushi Fujita. 2018. Unsupervised neural machine translation initialized by unsupervised statistical machine translation. *arXiv preprint arXiv:1810.12703*.

Andrew McCallum, Kedar Bellare, and Fernando Pereira. 2005. A conditional random field for discriminatively-trained finite-state string edit distance. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pages 388–395.

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 26*, pages 3111–3119.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9, Belgium, Brussels. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Sujith Ravi and Kevin Knight. 2011. Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 12–21, Portland, Oregon, USA. Association for Computational Linguistics.

Shuo Ren, Zhirui Zhang, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. Unsupervised neural machine translation with smt as posterior regularization. *arXiv preprint arXiv:1901.04112*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

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 (Volume 1: Long Papers)*, pages 46–55. Association for Computational Linguistics.

Omar Zaidan. 2009. Z-mert: A fully configurable open source tool for minimum error rate training of machine translation systems. *The Prague Bulletin of Mathematical Linguistics*, 91:79–88.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *The IEEE International Conference on Computer Vision (ICCV)*.