Unsupervised Statistical 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

arXiv:1809.01272v1 [cs.CL] 4 Sep 2018

Abstract

While modern machine translation has relied on large parallel corpora, a recent line of work has managed to train Neural Machine Translation (NMT) systems from monolingual corpora only (Artetxe et al., 2018c; Lample et al., 2018). Despite the potential of this approach for low-resource settings, existing systems are far behind their supervised counterparts, limiting their practical interest. In this paper, we propose an alternative approach based on phrase-based Statistical Machine Translation (SMT) that significantly closes the gap with supervised systems. Our method profits from the modular architecture of SMT: we first induce a phrase table from monolingual corpora through cross-lingual embedding mappings, combine it with an n-gram language model, and fine-tune hyperparameters through an unsupervised MERT variant. In addition, iterative backtranslation improves results further, yielding, for instance, 14.08 and 26.22 BLEU points in WMT 2014 English-German and English-French, respectively, an improvement of more than 7-10 BLEU points over previous unsupervised systems, and closing the gap with supervised SMT (Moses trained on Europarl) down to 2-5 BLEU points. Our implementation is available at https://github.com/artetxem/monoses.

1 Introduction

Neural Machine Translation (NMT) has recently become the dominant paradigm in machine translation (Vaswani et al., 2017). In contrast to more rigid Statistical Machine Translation (SMT) architectures (Koehn et al., 2003), NMT models are trained end-to-end, exploit continuous representations that mitigate the sparsity problem, and overcome the locality problem by making use of unconstrained contexts. Thanks to this additional flexibility, NMT can more effectively exploit large parallel corpora, although SMT is still superior when the training corpus is not big enough (Koehn and Knowles, 2017).

Somewhat paradoxically, while most machine translation research has focused on resource-rich settings where NMT has indeed superseded SMT, a recent line of work has managed to train an NMT system without any supervision, relying on monolingual corpora alone (Artetxe et al., 2018c; Lample et al., 2018). Given the scarcity of parallel corpora for most language pairs, including less-resourced languages but also many combinations of major languages, this research line opens exciting opportunities to bring effective machine translation to many more scenarios. Nevertheless, existing solutions are still far behind their supervised counterparts, greatly limiting their practical usability. For instance, existing unsupervised NMT systems obtain between 15-16 BLEU points in WMT 2014 English-French translation, whereas a state-of-the-art NMT system obtains around 41 (Artetxe et al., 2018c; Lample et al., 2018; Yang et al., 2018).

In this paper, we explore whether the rigid and modular nature of SMT is more suitable for these unsupervised settings, and propose a novel unsupervised SMT system that can be trained on monolingual corpora alone. For that purpose, we present a natural extension of the skip-gram model (Mikolov et al., 2013b) that simultaneously learns word and phrase embeddings, which are then mapped to a cross-lingual space through self-learning (Artetxe et al., 2018b). We use the resulting cross-lingual phrase embeddings to induce a phrase table, and combine it with a language model and a distance-based distortion model to build a standard phrase-based SMT system. The weights of this model are tuned in an unsupervised manner through an iterative Minimum Error Rate Training (MERT) variant, and the entire system

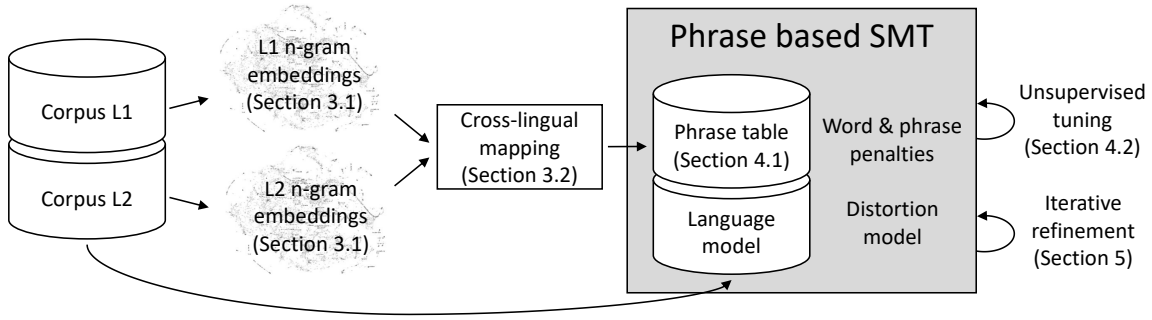

Figure 1: Architecture of our system, with references to sections.

is further improved through iterative backtranslation. The architecture of the system is sketched in Figure 1. Our experiments on WMT German-English and French-English datasets show the effectiveness of our proposal, where we obtain improvements above 7-10 BLEU points over previous unsupervised NMT-based approaches, closing the gap with supervised SMT (Moses trained on Europarl) down to 2-5 points.

The remaining of this paper is structured as follows. Section 2 introduces phrase-based SMT. Section 3 presents our unsupervised approach to learn cross-lingual n-gram embeddings, which are the basis of our proposal. Section 4 describes the proposed unsupervised SMT system itself, while Section 5 discusses its iterative refinement through backtranslation. Section 6 describes the experiments run and the results obtained. Section 7 discusses the related work on the topic, and Section 8 concludes the paper.

2 Background: phrase-based SMT

While originally motivated as a noisy channel model (Brown et al., 1990), phrase-based SMT is now formulated as a log-linear combination of several statistical models that score translation candidates (Koehn et al., 2003). The parameters of these scoring functions are estimated independently based on frequency counts, and their weights are then tuned in a separate validation set. At inference time, a decoder tries to find the translation candidate with the highest score according to the resulting combined model. The specific scoring models found in a standard SMT system are as follows:

- **Phrase table**. The phrase table is a collection of source language n-grams and a list of their possible translations in the target language along with different scores for each of them. So as to translate longer sequences, the decoder combines these partial n-gram translations, and ranks the resulting candidates according to their corresponding scores and the rest of scoring functions. In order to build the phrase table, SMT computes word alignments in both directions from a parallel corpus, symmetrizes these alignments using different heuristics (Och and Ney, 2003), extracts the set of consistent phrase pairs, and scores them based on frequency counts. For that purpose, standard SMT uses 4 scores for each phrase table entry: the direct and inverse lexical weightings, which are derived from word level alignments, and the direct and inverse phrase translation probabilities, which are computed at the phrase level.

- **Language model**. The language model assigns a probability to a word sequence in the target language. Traditional SMT uses n-gram language models for that, which use simple frequency counts over a large monolingual corpus with back-off and smoothing.

- **Reordering model**. The reordering model accounts for different word orders across languages, scoring translation candidates according to the position of each translated phrase in the target language. Standard SMT combines two such models: a distance based distortion model that penalizes deviation from a monotonic order, and a lexical reordering model that incorporates phrase orientation frequencies from a parallel corpus.

- **Word and phrase penalties**. The word and phrase penalties assign a fixed score to every generated word and phrase, and are useful to control the length of the output text and the preference for shorter or longer phrases.

Having trained all these different models, a tuning process is applied to optimize their weights in the resulting log-linear model, which typically maximizes some evaluation metric in a separate validation parallel corpus. A common choice is to optimize the BLEU score through Minimum Error Rate Training (MERT) (Och, 2003).

3 Cross-lingual n-gram embeddings

Section 3.1 presents our proposed extension of skip-gram to learn n-gram embeddings, while Section 3.2 describes how we map them to a shared space to obtain cross-lingual n-gram embeddings.

3.1 Learning n-gram embeddings

Negative sampling skip-gram takes word-context pairs (w, c), and uses logistic regression to predict whether the pair comes from the true distribution as sampled from the training corpus, or it is one of the k draws from a noise distribution (Mikolov et al., 2013b):

$$\log \sigma\left(w \cdot c\right) + \sum_{i=1}^{k} \mathbb{E}_{c_N \sim P_D}\left[\log \sigma\left(-w \cdot c_N\right)\right]$$

In its basic formulation, both w and c correspond to words that co-occur within a certain window in the training corpus. So as to learn embeddings for non-compositional phrases like *New York Times* or *Toronto Maple Leafs*, Mikolov et al. (2013b) propose to merge them into a single token in a pre-processing step. For that purpose, they use a scoring function based on their co-occurence frequency in the training corpus, with a discounting coefficient δ that penalizes rare words, and iteratively merge those above a threshold:

$$\text{score}(w_i, w_j) = \frac{\text{count}\left(w_i, w_j\right) - \delta}{\text{count}\left(w_i\right) \times \text{count}\left(w_j\right)}$$

However, we also need to learn representations for compositional n-grams in our scenario, as there is not always a 1:1 correspondence for n-grams across languages even for compositional phrases. For instance, the phrase *he will come* would typically be translated as *vendrá* into Spanish, so one would need to represent the entire phrase as a single unit to properly model this relation.

One option would be to merge all n-grams regardless of their score, but this is not straightforward given their overlapping nature, which is further accentuated when considering n-grams of different lengths. While we tried to randomly generate multiple consistent segmentations for each sentence and train the embeddings over the resulting corpus, this worked poorly in our preliminary experiments. We attribute this to the complex interactions arising from the stochastic segmentation (e.g. the co-occurrence distribution changes radically, even for unigrams), severely accentuating the sparsity problem, among other issues.

As an alternative approach, we propose a generalization of skip-gram that learns n-gram embeddings on-the-fly, and has the desirable property of unigram invariance: our proposed model learns the exact same embeddings as the original skip-gram for unigrams, while simultaneously learning additional embeddings for longer n-grams. This way, for each word-context pair (w, c) at distance d within the given window, we update their corresponding embeddings w and c with the usual negative sampling loss. In addition to that, we look at all n-grams p of different lengths that are at the same distance d, and for each pair (p, c), we update the embedding p through negative sampling. In order to enforce unigram invariance, the context c and negative samples c_N, which always correspond to unigrams, are not updated for (p, c). This allows to naturally learn n-gram embeddings according to their co-occurrence patterns as modeled by skip-gram, without introducing subtle interactions that affect its fundamental behavior.

We implemented the above procedure as an extension of *word2vec*, and use it to train monolingual n-gram embeddings with a window size of 5, 300 dimensions, 10 negative samples, 5 iterations and subsampling disabled. 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.

3.2 Cross-lingual mapping

Cross-lingual mapping methods take independently trained word embeddings in two languages, and learn a linear transformation to map them to a shared cross-lingual space (Mikolov et al., 2013a; Artetxe et al., 2018a). Most mapping methods are supervised, and rely on a bilingual dictionary, typically in the range of a few thousand entries, although a recent line of work has managed to achieve comparable results in a fully unsupervised manner based on either self-learning (Artetxe et al., 2017, 2018b) or adversarial training (Zhang et al., 2017a,b; Conneau et al., 2018).

In our case, we use the method of Artetxe et al.

(2018b) to map the n-gram embeddings to a shared cross-lingual space using their open source implementation VecMap[1]. Originally designed for word embeddings, this method builds an initial mapping by connecting the intra-lingual similarity distribution of embeddings in different languages, and iteratively improves this solution through self-learning. The method applies a frequency-based vocabulary cut-off, learning the mapping over the 20,000 most frequent words in each language. We kept this cut-off to learn the mapping over the most frequent 20,000 unigrams, and then apply the resulting mapping to the entire embedding space, including longer n-grams.

4 Unsupervised SMT

As discussed in Section 2, phrase-based SMT follows a modular architecture that combines several scoring functions through a log-linear model. Among the scoring functions found in standard SMT systems, the distortion model and word/phrase penalties are parameterless, while the language model is trained on monolingual corpora, so they can all be directly integrated into our unsupervised system. From the remaining models, typically trained on parallel corpora, we decide to leave the lexical reordering out, as the distortion model already accounts for word reordering. As for the phrase table, we learn cross-lingual n-gram embeddings as discussed in Section 3, and use them to induce and score phrase translation pairs as described next (Section 4.1). Finally, we tune the weights of the resulting log-linear model using an unsupervised procedure based on back-translation (Section 4.2).

Unless otherwise specified, we use Moses[2] with default hyperparameters to implement these different components of our system. We use KenML (Heafield et al., 2013), bundled in Moses by default, to estimate our 5-gram language model with modified Kneser-Ney smoothing, pruning n-grams longer than 3 with a single occurrence.

4.1 Phrase table induction

Given the lack of a parallel corpus from which to **extract phrase translation pairs**, every n-gram in the target language could be taken as a potential translation candidate for each n-gram in the source language. So as to keep the size of the phrase table within a reasonable limit, we train cross-lingual phrase embeddings as described in Section 3, and limit the translation candidates for each source phrase to its 100 nearest neighbors in the target language.

In order to estimate their corresponding **phrase translation probabilities**, we apply the softmax function over the cosine similarities of their respective embeddings. More concretely, given the source language phrase $\bar{e}$ and the translation candidate $\bar{f}$, their direct phrase translation probability is computed as follows[3]:

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

Note that, in the above formula, $\bar{f}'$ iterates across all target language embeddings, and τ is a constant temperature parameter that controls the confidence of the predictions. In order to tune it, we induce a dictionary over the cross-lingual embeddings themselves with nearest neighbor retrieval, and use maximum likelihood estimation over it. However, inducing the dictionary in the same direction as the probability predictions leads to a degenerated solution (softmax approximates the hard maximum underlying nearest neighbor as τ approaches 0), so we induce the dictionary in the opposite direction and apply maximum likelihood estimation over it:

$$\min_{\tau} \sum_{\bar{f}} \log \phi(\bar{f}|\,\mathrm{NN}_{\bar{e}}(\bar{f})) + \sum_{\bar{e}} \log \phi(\bar{e}|\,\mathrm{NN}_{\bar{f}}(\bar{e}))$$

So as to optimize τ, we use Adam with a learning rate of 0.0003 and a batch size of 200, implemented in PyTorch.

In order to compute the **lexical weightings**, we align each word in the target phrase with the one in the source phrase most likely generating it, and take the product of their respective translation probabilities:

$$\mathrm{lex}(\bar{f}|\bar{e}) = \prod_i \max\left(\epsilon, \max_j w(\bar{f}_i|\bar{e}_j)\right)$$

The constant ϵ guarantees that each target language word will get a minimum probability mass, which is useful to model NULL alignments. In our experiments, we set $\epsilon = 0.001$, which we find to

[1] https://github.com/artetxem/vecmap
[2] http://www.statmt.org/moses/

[3] The inverse phrase translation probability $\phi(\bar{e}|\bar{f})$ is defined analogously.

Algorithm 1 Unsupervised tuning

Input: $m_{s \to t}$ (source-to-target models)
Input: $m_{t \to s}$ (target-to-source models)
Input: c_s (source validation corpus)
Input: c_t (target validation corpus)
Output: $w_{s \to t}$ (source-to-target weights)
Output: $w_{t \to s}$ (target-to-source weights)

$w_{t \to s} \leftarrow$ DEFAULT_WEIGHTS
repeat
  $bt_s \leftarrow \text{TRANSLATE}(m_{t \to s}, w_{t \to s}, c_t)$
  $w_{s \to t} \leftarrow \text{MERT}(m_{s \to t}, bt_s, c_t)$
  $bt_t \leftarrow \text{TRANSLATE}(m_{s \to t}, w_{s \to t}, c_s)$
  $w_{t \to s} \leftarrow \text{MERT}(m_{t \to s}, bt_t, c_s)$
until convergence

work well in practice. Finally, the word translation probabilities $w(\bar{f}_i|\bar{e}_j)$ are computed using the same formula defined for phrase translation probabilities (see above), with the difference that the partition function goes over unigrams only.

4.2 Unsupervised tuning

As discussed in Section 2, standard SMT uses MERT over a small parallel corpus to tune the weights of the different scoring functions combined through its log-linear model. Given that we only have access to monolingual corpora in our scenario, we propose to generate a synthetic parallel corpus through backtranslation (Sennrich et al., 2016) and apply MERT tuning over it, iteratively repeating the process in both directions (see Algorithm 1). For that purpose, we reserve a random subset of 10,000 sentences from each monolingual corpora, and run the proposed algorithm over them for 10 iterations, which we find to be enough for convergence.

5 Iterative refinement

The procedure described in Section 4 suffices to train an SMT system from monolingual corpora which, as shown by our experiments in Section 6, already outperforms previous unsupervised systems. Nevertheless, our proposed method still makes important simplifications that could compromise its potential performance: it does not use any lexical reordering model, its phrase table is limited by the underlying embedding vocabulary (e.g. it does not include phrases longer than trigrams, see Section 3.1), and the phrase translation probabilities and lexical weightings are estimated based on cross-lingual embeddings.

Algorithm 2 Iterative refinement

Input: c_s (source language corpus)
Input: c_t (target language corpus)
Input/Output: $m_{t \to s}$ (target-to-source models)
Input/Output: $w_{t \to s}$ (target-to-source weights)
Output: $m_{s \to t}$ (source-to-target models)
Output: $w_{s \to t}$ (source-to-target weights)

$train_s, val_s \leftarrow \text{SPLIT}(c_s)$
$train_t, val_t \leftarrow \text{SPLIT}(c_t)$
repeat
  $btt_s \leftarrow \text{TRANSLATE}(m_{t \to s}, w_{t \to s}, train_t)$
  $btv_s \leftarrow \text{TRANSLATE}(m_{t \to s}, w_{t \to s}, val_t)$
  $m_{s \to t} \leftarrow \text{TRAIN}(btt_s, train_t)$
  $w_{s \to t} \leftarrow \text{MERT}(m_{s \to t}, btv_s, val_t)$
  $btt_t \leftarrow \text{TRANSLATE}(m_{s \to t}, w_{s \to t}, train_s)$
  $btv_t \leftarrow \text{TRANSLATE}(m_{s \to t}, w_{s \to t}, val_s)$
  $m_{t \to s} \leftarrow \text{TRAIN}(btt_t, train_s)$
  $w_{t \to s} \leftarrow \text{MERT}(m_{t \to s}, btv_t, val_s)$
until convergence

In order to overcome these limitations, we propose an iterative refinement procedure based on backtranslation (Sennrich et al., 2016). More concretely, we generate a synthetic parallel corpus by translating the monolingual corpus in one of the languages with the initial system, and train and tune a standard SMT system over it in the opposite direction. Note that this new system does not have any of the initial restrictions: the phrase table is built and scored using standard word alignment with an unconstrained vocabulary, and a lexical reordering model is also learned. Having done that, we use the resulting system to translate the monolingual corpus in the other language, and train another SMT system over it in the other direction. As detailed in Algorithm 2, this process is repeated iteratively until some convergence criterion is met.

While this procedure would be expected to produce a more accurate model at each iteration, it also happens to be very expensive computationally. In order to accelerate our experiments, we use a random subset of 2 million sentences from each monolingual corpus for training[4], in addition to the 10,000 separate sentences that are held out as a validation set for MERT tuning, and perform a fixed number of 3 iterations of the above algorithm. Moreover, we use FastAlign (Dyer et al., 2013) instead of GIZA++ to make word alignment faster. Other than that, training over the synthetic

[4]Note that we reuse the original language model, which is trained in the full corpus.

	WMT-14				WMT-16	
	FR-EN	**EN-FR**	**DE-EN**	**EN-DE**	**DE-EN**	**EN-DE**
Artetxe et al. (2018c)	15.56	15.13	10.21	6.55	-	-
Lample et al. (2018)	14.31	15.05	-	-	13.33	9.64
Yang et al. (2018)	15.58	16.97	-	-	14.62	10.86
Proposed system	**25.87**	**26.22**	**17.43**	**14.08**	**23.05**	**18.23**

Table 1: Results of the proposed method in comparison to existing unsupervised NMT systems (BLEU).

parallel corpus is done through standard Moses tools with default settings.

6 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. As discussed throughout the paper, our system is trained on monolingual corpora alone, so we take the concatenation of all News Crawl monolingual corpora from 2007 to 2013 as our training data, which we tokenize and truecase using standard Moses tools. The resulting corpus has 749 million tokens in French, 1,606 million tokens in German, and 2,109 million tokens in English. Following common practice, the systems are evaluated in newstest2014 using tokenized BLEU scores as computed by the `multi-bleu.perl` script included in Moses. In addition to that, we also report results in German-English newstest2016 (from WMT 2016), as this was used by some previous work in unsupervised NMT (Lample et al., 2018; Yang et al., 2018)[5]. So as to be faithful to our target scenario, we did not use any parallel data in these language pairs, not even for development purposes. Instead, we ran all our preliminary experiments on WMT Spanish-English data, where we made all development decisions.

We present the results of our final system in comparison to other previous work in Section 6.1. Section 6.2 then presents an ablation study of our proposed method, where we analyze the contribution of its different components. Section 6.3 compares the obtained results to those of different supervised systems, analyzing the effect of some of the inherent limitations of our method in a standard phrase-based SMT system. Finally, Section 6.4 presents some translation examples from our system.

6.1 Main results

We report the results obtained by our proposed system in Table 1. As it can be seen, our system obtains the best published results by a large margin, surpassing previous unsupervised NMT systems by around 10 BLEU points in French-English (both directions), and more than 7 BLEU points in German-English (both directions and datasets).

This way, while previous progress in the task has been rather incremental (Yang et al., 2018), our work represents an important step towards high-quality unsupervised machine translation, with improvements over 50% in all cases. This suggests that, in contrast to previous NMT-based approaches, phrase-based SMT may provide a more suitable framework for unsupervised machine translation, which is in line with previous results in low-resource settings (Koehn and Knowles, 2017).

6.2 Ablation analysis

We present ablation results of our proposed system in Table 2. The first row corresponds to the initial system with our induced phrase table (Section 4.1) and default weights as used by Moses, whereas the second row uses our unsupervised MERT procedure to tune these weights (Section 4.2). The remaining rows represent different iterations of our refinement procedure (Section 5), which uses backtranslation to iteratively train a standard SMT system from a synthetic parallel corpus.

The results show that the initial system with default weights (first row) is already better than previous unsupervised NMT systems (Table 1) by a substantial margin (2-6 BLEU points). Our unsupervised tuning procedure further improves results, bringing an improvement of over 1 BLEU

[5]Note that we use the same model trained in WMT 2014 for these experiments, so it is likely that our results could be further improved by using the more extensive data from WMT 2016.

	WMT-14				WMT-16	
	FR-EN	**EN-FR**	**DE-EN**	**EN-DE**	**DE-EN**	**EN-DE**
Unsupervised SMT	21.16	20.13	13.86	10.59	18.01	13.22
+ unsupervised tuning	22.17	22.22	14.73	10.64	18.21	13.12
+ iterative refinement (it1)	24.81	26.53	16.01	13.45	20.76	16.94
+ iterative refinement (it2)	**26.13**	**26.57**	17.30	13.95	22.80	18.18
+ iterative refinement (it3)	25.87	26.22	**17.43**	**14.08**	**23.05**	**18.23**

Table 2: Ablation results (BLEU). The last row corresponds to our full system. Refer to the text for more details.

		WMT-14				WMT-16	
		FR-EN	**EN-FR**	**DE-EN**	**EN-DE**	**DE-EN**	**EN-DE**
Supervised	NMT (transformer)	-	41.8	-	28.4	-	-
	WMT best	35.0	35.8	29.0	20.6	40.2	34.2
	SMT (europarl)	30.61	30.82	20.83	16.60	26.38	22.12
	+ w/o lexical reord.	30.54	30.33	20.37	16.34	25.99	22.20
	+ constrained vocab.	30.04	30.10	19.91	16.32	25.66	21.53
	+ unsup. tuning	29.32	29.46	17.75	15.45	23.35	19.86
Unsup.	Proposed system	25.87	26.22	17.43	14.08	23.05	18.23

Table 3: Results of the proposed method in comparison to supervised systems (BLEU). Transformer results reported by Vaswani et al. (2017). SMT variants are incremental (e.g. 2nd includes 1st). Refer to the text for more details.

point in both French-English directions, although its contribution is somewhat weaker for German-to-English (almost 1 BLEU point in WMT 2014 but only 0.2 in WMT 2016), and does not make any difference for English-to-German.

The proposed iterative refinement method has a much stronger positive effect, with improvements over 2.5 BLEU points in all cases, and up to 5 BLEU points in some. Most gains come in the first iteration, while the second iteration brings weaker improvements and the algorithm seems to converge in the third iteration, with marginal improvements for German-English and a small drop in performance for French-English.

6.3 Comparison with supervised systems

So as to put our results into perspective, Table 3 comprises the results of different supervised methods in the same test sets. More concretely, we report the results of the Transformer (Vaswani et al., 2017), an NMT system based on self-attention that is the current state-of-the-art in machine translation, along with the scores obtained by the best performing system in each WMT shared task at the time, and those of a standard phrase-based SMT system trained on Europarl and tuned on newstest2013 using Moses. We also report the effect of removing lexical reordering from the latter as we do in our initial system (Section 4), restricting the vocabulary to the most frequent unigram, bigram and trigrams as we do when training our embeddings (Section 3), and using our unsupervised tuning procedure over a subset of the monolingual corpus (Section 4.2) instead of using standard MERT tuning over newstest2013.

Quite surprisingly, our proposed system, trained exclusively on monolingual corpora, is relatively close to a comparable phrase-based SMT system trained on Europarl, with differences below 5 BLEU points in all cases and as little as 2.5 in some. Note that both systems use the exact same language model trained on News Crawl, making them fully comparable in terms of the monolingual corpora they have access to. While more of a baseline than the state-of-the-art, note that Moses+Europarl is widely used as a reference system in machine translation. As such, we think that our results are very encouraging, as they show

Source	Reference	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.
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.
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.
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.

Table 4: Randomly chosen translation examples from French→English newstest2014.

that our fully unsupervised system is already quite close to this competitive baseline.

In addition to that, the results for the constrained variants of this SMT system justify some of the simplifications required by our approach. In particular, removing lexical reordering and constraining the phrase table to the most frequent n-grams, as we do for our initial system, has a relatively small effect, with a drop of less than 1 BLEU point in all cases, and as little as 0.28 in some. Replacing standard MERT tuning with our unsupervised variant does cause a considerable drop in performance, although it is below 2.5 BLEU points even in the worst case, and our unsupervised tuning method is still better than using default weights as reported in Table 2. This shows the importance of tuning in SMT, suggesting that these results could be further improved if one had access to a small parallel corpus for tuning.

6.4 Qualitative results

Table 4 shows some of the translations produced by the proposed system for French→English. Note that these examples where randomly taken from the test set, so they should be representative of the general behavior of our approach.

While the examples reveal certain adequacy issues (e.g. *The Times of London from Sidney* instead of *Sydney's Sunday Telegraph*), and the produced output is not perfectly grammatical (e.g. *go in Australia*), our translations are overall quite accurate and fluent, and one could get a reasonable understanding of the original text from them. This suggests that unsupervised machine translation can indeed be a usable alternative in low resource settings.

7 Related work

Similar to our approach, statistical decipherment also attempts to build machine translation systems from monolingual corpora. For that purpose, existing methods treat the source language as ciphertext, and model its generation through a noisy channel model involving two steps: the generation of the original English sentence and the probabilistic replacement of the words in it (Ravi and Knight, 2011; Dou and Knight, 2012). 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. Subsequent work has attempted to enrich these models with additional information like syntactic knowledge (Dou and Knight, 2013) and word embeddings (Dou et al., 2015). Nevertheless, these systems work in a word-by-word basis and have

only been shown to work in limited settings, being often evaluated in word-level translation. In contrast, our method builds a fully featured phrase-based SMT system, and achieves competitive performance in a standard machine translation task.

More recently, Artetxe et al. (2018c) and Lample et al. (2018) have managed to train a standard attentional encoder-decoder NMT system from monolingual corpora alone. For that purpose, they use a shared encoder for both languages with pre-trained cross-lingual embeddings, and train the entire system using a combination of denoising, backtranslation and, in the case of Lample et al. (2018), adversarial training. This method was further improved by Yang et al. (2018), who use a separate encoder for each language, sharing only a subset of their parameters, and incorporate two generative adversarial networks. However, our results in Section 6.1 show that our SMT-based approach obtains substantially better results.

Our method is also connected to some previous approaches to improve machine translation using monolingual corpora. In particular, the generation of a synthetic parallel corpus through backtranslation (Sennrich et al., 2016), which is a key component of our unsupervised tuning and iterative refinement procedures, has been previously used to improve NMT. In addition, there have been several proposals to extend the phrase table of SMT systems by inducing translation candidates and/or scores from monolingual corpora, using either statistical decipherment methods (Dou and Knight, 2012, 2013) or cross-lingual embeddings (Zhao et al., 2015; Wang et al., 2016). While all these methods exploit monolingual corpora to enhance an existing machine translation system previously trained on parallel corpora, our approach learns a fully featured phrase-based SMT system from monolingual corpora alone.

8 Conclusions and future work

In this paper, we propose a novel unsupervised SMT system that can be trained on monolingual corpora alone. For that purpose, we extend the skip-gram model (Mikolov et al., 2013b) to simultaneously learn word and phrase embeddings, and map them to a cross-lingual space adapting previous unsupervised techniques (Artetxe et al., 2018b). The resulting cross-lingual phrase embeddings are used to induce a phrase table, which coupled with an n-gram language model and distance-based distortion yields an unsupervised phrase-based SMT system. We further improve results tuning the weights with our unsupervised MERT variant, and obtain additional improvements retraining the entire system through iterative backtranslation. Our implementation is available as an open source project at `https://github.com/artetxem/monoses`.

Our experiments on standard WMT French-English and German-English datasets confirm the effectiveness of our proposal, where we obtain improvements above 10 and 7 BLEU points over previous NMT-based approaches, respectively, closing the gap with supervised SMT (Moses trained on Europarl) down to 2-5 points.

In the future, we would like to extend our approach to semi-supervised scenarios with small parallel corpora, which we expect to be particularly helpful for tuning purposes. Moreover, we would like to try a hybrid approach with NMT, using our unsupervised SMT system to generate a synthetic parallel corpus and training an NMT system over it through iterative backtranslation.

Acknowledgments

This research was partially supported by the Spanish MINECO (TUNER TIN2015-65308-C5-1-R, MUSTER PCIN-2015-226 and TADEEP TIN2015-70214-P, cofunded by EU FEDER), the UPV/EHU (excellence research group), and the NVIDIA GPU grant program. Mikel Artetxe enjoys 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. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 5012–5019.

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

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

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.

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.

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, Vancouver. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

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

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. 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.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

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.

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.

Rui Wang, Hai Zhao, Sabine Ploux, Bao-Liang Lu, and Masao Utiyama. 2016. A bilingual graph-based semantic model for statistical machine translation. In *IJCAI*, pages 2950–2956.

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.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017a. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017b. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark. Association for Computational Linguistics.

Kai Zhao, Hany Hassan, and Michael Auli. 2015. Learning translation models from monolingual continuous representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1527–1536, Denver, Colorado. Association for Computational Linguistics.