Unsupervised Extraction of Partial Translations for Neural Machine Translation

Benjamin Marie **Atsushi Fujita**
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
{bmarie, atsushi.fujita}@nict.go.jp

Abstract

In neural machine translation (NMT), monolingual data are usually exploited through a so-called back-translation: sentences in the target language are translated into the source language to synthesize new parallel data. While this method provides more training data to better model the target language, on the source side, it only exploits translations that the NMT system is already able to generate using a model trained on existing parallel data. In this work, we assume that new translation knowledge can be extracted from monolingual data, without relying at all on existing parallel data. We propose a new algorithm for extracting from monolingual data what we call partial translations: pairs of source and target sentences that contain sequences of tokens that are translations of each other. Our algorithm is fully unsupervised and takes only source and target monolingual data as input. Our empirical evaluation points out that our partial translations can be used in combination with back-translation to further improve NMT models. Furthermore, while partial translations are particularly useful for low-resource language pairs, they can also be successfully exploited in resource-rich scenarios to improve translation quality.

1 Introduction

Neural machine translation (NMT) systems usually require a large quantity of high-quality bilingual parallel data for training. However, for most language pairs, we do not have such resources, or only in very small quantities, mainly because they are costly to produce.

On the other hand, monolingual corpora are readily available in large quantity for many languages. Previous work has proposed various strategies to integrate monolingual data into NMT systems and has confirmed their usefulness to improve NMT systems. The so-called *back-translation* of monolingual data (Sennrich et al., 2016) is undoubtedly the most prevalent one. This approach simply uses a target-to-source MT system to translate monolingual data in the target language into the source language. The produced new synthetic parallel corpus can be used together with the original parallel data to increase the size of the training data, and eventually to improve NMT systems significantly and consistently. However, on the source side, the synthetic data only contain data that can be generated by the back-translation system trained on some existing parallel data.

Previous work has also studied the extraction of translation pairs of source and target sentences from monolingual data in their respective languages. They have been shown to be useful to train better statistical machine translation (SMT) systems, especially in low-resource conditions. Existing methods on sentence pair extraction mainly rely on the availability of comparable corpora as the source of accurate sentence pairs (Abdul Rauf and Schwenk, 2011), or on the robustness of SMT against noise (Goutte et al., 2012) because sentence pairs extracted from unrelated monolingual corpora tend to be noisy (Tillmann and Xu, 2009; Marie and Fujita, 2017). Most of them also require pre-trained accurate translation models, those of SMT systems for instance, that we may not have in low-resource conditions. Moreover, unlike SMT, NMT has been shown to deal very poorly with noisy training data and still largely underperforms SMT for low-resource language pairs (Koehn and Knowles, 2017) for which comparable corpora are usually not available. Even without an accurate translation model, we still have the possibility of extracting sentence pairs from unrelated source and target monolingual data. However, this is very challenging since we have no guarantee that there are sentence pairs actually retrievable from a given

Proceedings of NAACL-HLT 2019, pages 3834–3844
Minneapolis, Minnesota, June 2 - June 7, 2019.

Source sentence	der Mann wurde festgenommen .
Partial translation	a **man was arrested** at the scene **.**
Post-processed	UNKPP **man was arrested** UNKPP UNKPP UNKPP **.**

Figure 1: Example of a source sentence in our German monolingual data, its best partial translation found in our English monolingual data by our algorithm, and its post-processed version. The bold tokens translate tokens of the source sentence.

pair of source and target monolingual corpora.

In this work, we assume (i) that a given pair of monolingual corpora contain sentence pairs that are at least *partial translations*, i.e., pairs of source and target sentences containing phrases (sequences of tokens) that are translations of each other and (ii) that such pairs can help train better NMT systems. On these assumptions, we propose a new algorithm that extracts partial translations from trillions of candidate sentence pairs without any supervision. Relying on an unsupervised phrase table, our algorithm identifies phrases in a source sentence that have likely translations in a target sentence. The extracted partial translations often contain unrelated parts besides aligned phrases. Therefore, we also apply a simple but very effective post-processing to make such noisy sentence pairs exploitable for a target-to-source NMT model, as exemplified in Figure 1.

We report on significant improvements in translation quality for two language pairs and under different experimental conditions when using our extracted partial translations to train NMT systems. While our method is especially designed to provide new training data for low-resource language pairs, we also observed significant improvements over a strong NMT system trained on large quantity of parallel data. Furthermore, we demonstrate the complementarity of our approach with back-translation.

2 Extraction of Partial Translations

The whole framework for extracting partial translations is presented in Figure 2. To extract partial translations, we first induce a phrase table that contains phrases in the source language paired with their most probable translations in the target language (Section 2.1). These phrase pairs are collected from the same monolingual data from which we extract partial translations. Given

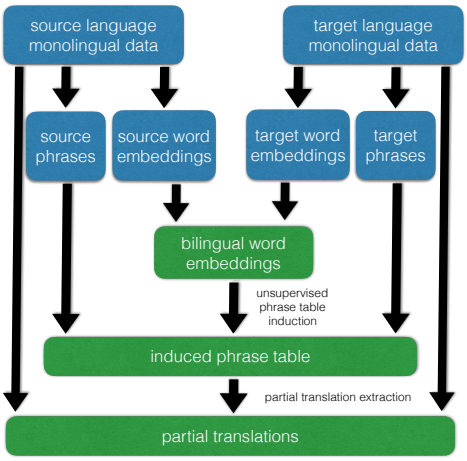

Figure 2: The framework for extracting partial translations from monolingual data.

the induced phrase table, we search for sentence pairs that are the most likely partial translations in the monolingual data (Section 2.2). Finally, the extracted sentence pairs are post-processed (Section 2.3).

2.1 Unsupervised Phrase Table Induction

Recent methods addressed the task of finding word translations from monolingual data without any supervision (Lample et al., 2018a; Artetxe et al., 2018a). On the other hand, Marie and Fujita (2018) presented a method for inducing phrase tables from monolingual data using a weakly-supervised framework. To make our approach useful in as many translation tasks as possible, including very low-resource scenarios, we propose a fully unsupervised version of the method in Marie and Fujita (2018). Using phrases instead of only single tokens promotes the extraction of partial translations containing longer sequences of tokens, rather than those with potentially more but discontinuously translated tokens.

Regarding all n-grams of tokens in the monolingual data as phrases and searching the translations for each phrase can be extremely costly. Therefore, we extract meaningful source and target phrases from their respective monolingual data, following the work by Marie and Fujita (2018).[1] Beside extracting phrases, we also train word embeddings on the same source and target monolingual data, independently. Both source and target embedding spaces are aligned in the same space to make word embeddings bilingual, without any supervision (Artetxe et al., 2018a). Using these

[1] See Section 3.1 of Marie and Fujita (2018) for the details.

bilingual word embeddings, we compute bilingual phrase embeddings for each of the extracted phrases through the element-wise addition of the embeddings of constituent words of the phrases.

Given the source and target phrase sets, we take their Cartesian product to generate all possible pairs of source and target phrases and compute cosine similarity of each pair using their phrase embeddings. Each pair is also associated with a translation probability (Lample et al., 2018b):

$$p(t_j|s_i) = \frac{\exp\left(\beta \cos(emb(t_j), emb(s_i))\right)}{\sum_k \exp\left(\beta \cos(emb(t_k), emb(s_i))\right)}, \quad (1)$$

where t_j is the j-th phrase in the target phrase list and s_i the i-th phrase in the source phrase list, β a parameter to tune the peakiness of the distribution[2] (Smith et al., 2017), $\cos(\cdot, \cdot)$ cosine similarity between two phrase embeddings, and $emb(\cdot)$ a function returning the bilingual embedding of a given phrase. In practice, to keep the set of phrase pairs to score manageable, we first filter the 300k most frequent phrases in each of the source and target phrase sets.[3]

Retrieved phrase pairs may have a very low translation probability, especially when dealing with distant languages and/or noisy monolingual data. Therefore, we keep only the n-best target phrases for each source phrase, according to Eq. (1). While maintaining the coverage by our phrase table of the source monolingual data the same, we ensure that the phrase translations for each source phrase are the most accurate among all the collected target phrases.

2.2 Partial Translation Extraction

NMT architectures expect parallel sentence pairs as training data, even if we have accurate phrase pairs, they cannot be used directly for training.

Therefore, we propose an algorithm for extracting sentence pairs from the monolingual data that matches the best possible combinations of phrase pairs from the induced phrase table. The pseudo-code of this algorithm is presented in Algorithm 1. For each source sentence S (l.2), the algorithm first selects from the phrase table pt all the phrase pairs P_s whose source side appears in S (l.3). It

Algorithm 1: Partial Translation Extraction

Input : pt: a phrase table,
M_s: source monolingual data, and
M_t: target monolingual data
Output : $P_{\text{translations}}$: a set of the m_{top}-best partial translations
Parameters: m: the number of target sentence to retrieve for each source sentence for the PBFD algorithm, and
m_{top}: the size of $P_{\text{translations}}$

1 $P_{\text{translations}} \leftarrow \{\}$;
2 **foreach** *sentence* S *in* M_s **do**
3 $P_s \leftarrow$ `Intersection(`pt, S`)` ;
4 $B_t \leftarrow$ `Bow(`P_s, pt`)` ;
5 $T_{\text{candidates}} \leftarrow \{\}$;
6 **foreach** *sentence* T *in* M_t **do**
7 **if** $\exists w \in T$: $w \in B_t$ **then**
8 $T_{\text{candidates}} \leftarrow T_{\text{candidates}} \cup T$;
9 **end**
10 **end**
11 $T_{\text{m}} \leftarrow$ `Prune(`m, $T_{candidates}$`)` ;
12 $T_{\text{best}} \leftarrow$ `PBFD(`pt, S, T_m`)` ;
13 $P_{\text{translations}} \leftarrow P_{\text{translations}} \cup (T_{\text{best}}, S)$;
14 **end**
15 $P_{\text{translations}} \leftarrow$ `GetBest(`m_{top}, $P_{translations}$`)` ;
16 **return** $P_{\text{translations}}$

then creates a bag of target words B_t via collecting all the words from the target phrases of P_s (l.4). Subsequently, we keep the m-best target sentences T_m (ll.5–11) according to the following score:

$$F_t(S, T) = \frac{2c_s c_t}{c_s + c_t}, \quad (2)$$

where $c_s = k_t/len(S)$, $c_t = k_t/len(T)$, with S and T respectively a source and a target sentence, k_t the number of tokens in T that are covered by B_t, and $len(\cdot)$ a function that returns the number of tokens in a sentence. With the harmonic mean F_t between c_s and c_t, the algorithm searches for target sentences containing as many words translating source tokens as possible while penalizing the retrieval of very long target sentences that may also contain many tokens having no counterparts in the source sentence.

Then, the algorithm re-ranks the target sentences in T_m with the phrase-based forced decoding (PBFD) algorithm (Zhang et al., 2017) (l.12). PBFD searches for the best combination of phrase pairs covering S and each target sentence in T_m, using the phrase translation probability computed by Eq. (1). However, the original PBFD algorithm penalizes sentence pair with words that are not covered by any phrase pair. It tends to favor very short sentences that are potentially less exploitable for NMT systems, or not even be sentences as it may happen when dealing with noisy

[2] We set $\beta = 30$ since it gives consistently good results in our preliminary experiments.

[3] This means that we still have to compute the cosine similarity for 90 billion phrase pairs (300k×300k). This can be done very efficiently on GPU: https://github.com/facebookresearch/faiss.

monolingual data.[4] Therefore, we use a slightly modified version which does not penalize uncovered words on the target side in order to favor the extraction of longer target sentences that may contain more translated tokens. Finally, we retain only $P_{\text{translations}}$: the m_{top} sentence pairs with the highest PBFD scores (l.15).

2.3 Post-Processing of Partial Translations

Since the extracted sentence pairs are only partial translations, incorporating them as they are into the training data for NMT may mislead the training of the model due to their noisiness. Since our PBFD algorithm does not penalize target words not covered by any phrase pair, the target side of our partial translations contains longer sentences than the source side, potentially with many words unaligned with any word in the source sentences. Nonetheless, the back-translation approach has proven that NMT can be trained on noisy data at the source side and fluent sentences at the target side. Following this, we use partial translations only to train target-to-source NMT systems, as for back-translated data. It means that the source and target languages of our extraction algorithm becomes respectively the target and source languages of the NMT system.

We want the NMT system to give as much attention as possible to the phrases of the source sentence that have a translation in the target sentence, while ignoring as much as possible the remaining tokens of the source sentence that are likely to be noise, i.e., not translated by the target sentence. Dropping these unaligned tokens is not an appropriate solution since it would produce unlikely sequences of translated tokens. Then, to make the decoder paying its attention to the translated part in the source sentence, we simply replaced all the tokens not covered by the best combination of phrase pairs found by the PBFD algorithm with a made-up token, `UNKPP`.[5]

We expect this post-processing step to particularly well suit the training of a Transformer NMT model (Vaswani et al., 2017), because it can easily learn to pay no attention to `UNKPP` and more attention to correctly translated tokens thanks to the multi-head attention mechanism. Moreover, the Transformer model does not memorize complete sequences and makes time steps independent, unlike recurrent neural networks (RNN). It uses instead positional encodings and has a better ability in linking important features from the entire sequence (Chen et al., 2018), which may make easier the learning from noisy sequences, such as the ones created by introducing `UNKPP`.

We show in Section 4 that using `UNKPP` tokens instead of dropping uncovered tokens leads to a better model. Nonetheless, a better strategy that we leave for future work could be to apply some forced-decoding, with an SMT system for instance, that translates in the source sentence the parts of the target sentence that are not translated, while preserving the partial translations detected by our algorithm in the source sentence.

3 Experiments

We experimented on three language pairs with different degrees of relatedness between the languages of each pair: English→German (en→de), English→Turkish (en→tr), and Bengali→Malay (bn→ms). While our approach is dedicated to improve translation quality for low-resource language pairs, we included the en→de pair for a detailed analysis on the impact of using partial translations in addition to much more training data. bn→ms is expected to be an extremely difficult translation task, because only small quantity of parallel data are available to start with (Section 3.1) and Bengali and Malay are very distant languages which makes very difficult the training of useful unsupervised bilingual word embeddings (Søgaard et al., 2018), which is a key element for inducing the phrase table.

3.1 Data

To train baseline NMT systems, we used for en→de and en→tr 100k parallel sentences randomly extracted from the parallel data provided for the WMT18 News Translation Task,[6] except the ParaCrawl corpus for en→de. For bn→ms, we used the 18k sentence pairs released by the Asian Language Treebank (ALT) project (Riza et al., 2016).[7] As validation and test data for en→de and en→tr, we used Newstest2016 and Newstest2017, respectively, provided by WMT18. For bn→ms,

[4] This includes equations, rows of a table, titles, etc.

[5] We made this token different from the usual token reserved for unknown word in the vocabulary, since they are of a different nature.

[6] http://www.statmt.org/wmt18/translation-task.html

[7] http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/

we used the official development and test data provided by the ALT project.

We chose this amount of training data for en→de and en→tr, since it suits our need for a low-resource translation task while we can still use the available parallel data for further analysis. Moreover, we needed enough parallel data to train an NMT model that can produce useful back-translated data. In our preliminary experiment, we found out that 100k sentence pairs satisfy the minimum amount to train NMT models for useful back-translation. As such, we did not succeed in training useful models for bn→ms, due to the difficulty of the task, but still decided to report on the results to provide insights and matters for future work.

As for monolingual data, we used the English (239M lines) and German (237M lines) NewsCrawl corpora provided by WMT18, and the NewsCrawl and Common Crawl corpora for Turkish (104M lines). We extracted monolingual data ourselves from the Common Crawl project[8] for Bengali (5.3M lines) and Malay (4.6M lines).

3.2 Methods and Systems

To build an unsupervised phrase table, we first extracted 300k most frequent phrases of up to 6 tokens from the entire monolingual data. We also trained 200-dimensional word embeddings on the same data with `fasttext`.[9] For each source phrase, we retained only the 1-best target phrase in the induced phrase table.

The search for partial translations was performed only for a random sample of 1M lines from the monolingual data for each target language. For each of these lines, we searched for the best partial translation with our algorithm in up to 10M lines randomly extracted from the monolingual data for the source language. Then, to maintain a 1:1 ratio with the parallel data, we retained the 100k best partial translations for en→de and en→tr,[10] and 18k best partial translations for bn→ms.

All our NMT systems, including baselines, were the Transformer model (Vaswani et al., 2017) trained with Marian (Junczys-Dowmunt et al., 2018). Note that we fine-tuned the hyperparameters for training on our validation data for each language pair in order to get best possible baseline systems. We then apply the same hyperparameters in all the experiments for the given language pair. To train systems with partial translations (`partial`), we simply mixed them with the original parallel data during training. We also evaluated the systems using back-translated (`backtr`) and copied[11] (`copy`) (Currey et al., 2017) data, separately mixed with the original parallel data. Note that these data were generated from the same target sentences sampled for extracting partial translations: `partial`, `backtr`, and `copy` had the same target side but different source side.[12] Our NMT systems were evaluated with detokenized BLEU-cased.

Training data	en→de	en→tr	bn→ms
baseline	7.1	9.3	**6.1**
`backtr`	9.1	11.4	5.4
`copy`	9.0	11.2	6.0
`partial`	9.9	10.4	5.5
`backtr+copy`	11.3	11.6	5.5
`backtr+partial`	11.5	11.6	4.5
`backtr+copy+partial`	**11.9**	**12.2**	5.8

Table 1: BLEU scores of NMT systems. All the evaluated systems used the parallel data used to train the baseline system, and other synthetic parallel data generated by back-translation (`backtr`), copy (`copy`), or from extracted partial translation (`partial`).

3.3 Results

The results of our experiments are presented in Table 1. For en→de and en→tr, the baseline systems resulted in a poor translation quality below 10 BLEU points. This highlights how critical it is to get more training data to better train an NMT model for low-resource translation tasks.

Adding 100k synthetic parallel sentences generated by back-translation (`backtr`) improved translation quality by 2.0 and 2.1 BLEU points for en→de and en→tr, respectively. Surprisingly, the simplest `copy` method brought improvements similar to `backtr`. Furthermore, we also observed the complementarity of `backtr` and `copy` (`backtr+copy`), with 4.2 and 2.3 BLEU points of improvements for en→de and en→tr, re-

[8] http://commoncrawl.org/

[9] https://fasttext.cc/

[10] The extraction of 100k partial translations from 10^{13} sentence pairs (1M×10M) required around 26 hours of computation using 100 CPUs.

[11] The `copy` approach simply copies the target sentences to the source side. This method surprisingly offers good results in low-resource conditions, and a good complementarity with back-translation, for languages with some orthographic similarity.

[12] With some source sentences that may be identical.

spectively, over the baseline system. To verify that it is not the consequence of just giving more weight to the target monolingual data that may be in-domain, we also trained `backtr+backtr` but did not observe any improvements over `backtr`.

For en→de, the system using our extracted partial translations (`partial`) outperformed `backtr` and `copy` by 0.8 and 0.9 BLEU points, respectively, and the baseline system by 2.8 BLEU points. For en→tr, `partial` also significantly outperformed the baseline system, by 1.1 BLEU points. However, `backtr` and `copy` brought larger improvements. We can explain the difference between en→de and en→tr by the fact that Turkish is more distant from English than German. It makes unsupervised bilingual word embeddings more difficult to train for en→tr and are consequently significantly less accurate (Søgaard et al., 2018). Extraction of accurate and useful partial translations from monolingual data is a more difficult task for en→tr.

While these three kinds of synthetic parallel data, `backtr`, `copy`, and `partial`, present the same target sentences, we found out that mixing all of them with the original parallel led to the best system (`backtr+copy+partial`). This result shows the complementarity of these three datasets thanks to the diversity of source sides generated by different means. For instance, `partial` provides new original translations that were not generated by back-translation.

As expected, bn→ms is a very difficult task for NMT due to the small size of the training data. We were not able to train an NMT model that can generate useful back-translations. The `copy` method was also unhelpful since Bengali and Malay have different writing systems. Our partial translations did not help either, presumably due to the difficulty in unsupervised learning of bilingual word embeddings. Note also that we used much less monolingual data for this language pair to train the word embeddings. This last result is disappointing, but it confirmed that unsupervised bilingual word embeddings are still far from being useful for truly low-resource and distant language pairs.

4 Analysis

4.1 Impact of the Phrase Table

We induced the phrase table from the 300k most frequent source and target phrases. An obvious and actually simpler alternative is the use of words instead of phrases. Using 300k words instead of 300k phrases, for instance, results in a similar cost for phrase table induction but involves a larger vocabulary introducing a better coverage of the monolingual data. However, involving more, and consequently less frequent, words means that it also introduces words for which the word embedding will be noisier.

Training data	en→de	en→tr
baseline	7.1	9.3
1-best target word	8.6	9.9
1-best target phrase	**9.9**	**10.4**
2-best target phrases	7.6	9.5
5-best target phrases	7.5	9.3

Table 2: BLEU scores of NMT systems based on a phrase table induced from source and target single words, or using the 1-best (as in our default configuration), 2-best, or 5-best target phrase translations for each source phrase.

Another important decision that we made when inducing the phrase table was to take only the 1-best target phrase for each source phrase. Involving noisier translations would have a larger impact on our algorithm by inflating the size of B_t that would contain more and noisier target tokens, resulting in unworthily high F_t scores for target sentences containing these tokens that are nonetheless less likely to be the translations of the corresponding source phrase.

Table 2 presents results obtained when using single words instead of phrases to induce the phrase table, and when used the 1-, 2-, or 5-best target phrases for each source phrase. For en→de, using words instead of phrases still led to useful partial translations, with an improvement of 1.5 BLEU points. However, it was 1.3 BLEU points lower than when using phrases. The use of more than one target phrase for each source phrase in the induced phrase table resulted in the retrieval of noisier partial translations that are much less useful to train better NMT models. We observed the similar tendencies for en→tr.

One of the strongest assumption of this work is that partial translations bring translation knowledge complementary to the manually created parallel data used to the train the NMT system. This new knowledge is unbiased by the existing parallel data since we induce the phrase table without using any given parallel data. On the other hand, we can train a phrase table on the parallel data used

Training data	en→de	bn→ms
baseline	7.1	6.1
induced phrase table	**9.9**	5.5
standard phrase table	9.4	**6.3**

Table 3: BLEU scores of NMT systems based on a phrase table induced from monolingual data or based on a standard phrase table trained on the same parallel data also used to train the baseline system.

Training data	en→de	en→tr
baseline	7.1	9.3
`original`	6.2	7.7
`dropped`	8.8	10.0
`partial`	**9.9**	**10.4**

Table 4: BLEU scores of NMT systems trained on partial translations with (`dropped`, `partial`) and without (`original`) post-processing.

to train the NMT system and use it to extract partial translations. Owing to the supervision, we can expect such a phrase table to be much more accurate than the induced phrase table. However, it would introduce a strong bias that encourages the retrieval of partial translations similar to the parallel data that are already used to train the system. Consequently, the extracted partial translations may be less useful than those extracted with an induced phrase table.

To test the above assumption, we trained a standard phrase table on the given parallel data, extracted partial translations using it, and evaluated their impact on the translation quality. Table 3 shows the results for en→de and bn→ms. The results for en→de supports our hypothesis: using a standard phrase table for extracting partial translations has led to a drop of 0.5 BLEU points compared to the use of an induced phrase table. In contrast, for bn→ms, standard phrase table achieved significantly better results than using the induced phrase table and brought a slight improvement over the baseline system. We speculate that the standard phrase table trained only on 18k sentence pairs is not strong enough to bias the extraction of partial translations.

4.2 Impact of Post-Processing

By replacing unaligned tokens, identified by the PBFD algorithm, with a made-up token `UNKPP`, we aimed to guide the decoder to ignore them during training. This section explores the impact of this post-processing, through comparing the translation quality of NMT systems trained on partial translations without any post-processing (`original`), post-processed by removing unaligned source tokens (`dropped`), and our proposed method that replaces unaligned tokens with `UNKPP` (`partial`).

Table 4 presents our results. Without any post-processing, using partial translations brought a significant drop of translation quality of 0.9 BLEU points for en→de, and 1.6 BLEU points for en→tr, from the baseline system trained only on parallel data. This is expected since the partial translations can be very noisy with many unaligned tokens for which the Transformer model will still try to learn a translation in the target sentence. Removing them helps significantly with, for instance for en→de, a 1.7 BLEU points of improvements over the baseline system. However, their removal hides the existence of unaligned tokens and produces sequences of tokens that are unlikely in the source language, presumably misleading the training of the NMT model. Indeed, replacing them with a made-up token further improved the translation quality by 1.1 BLEU points.

4.3 Impact of Noisier Partial Translations

The baseline systems evaluated in Section 3 were trained on 100k parallel data and augmented with the same amount of back-translated sentences and partial translations. Our algorithm retrieved the best partial translation for each one of the 1M source sentences, and then ranked them according to the PBFD score in order to select the most accurate ones. In fact, partial translations at lower rank contained more unaligned tokens, as shown in Figure 3, and also incorrectly aligned tokens. Using these noisier partial translations may disturb the training of the NMT system. In contrast, we can easily increase the quantity of the back-translated data of a similar quality.

To verify this assumption, we evaluated NMT models for en→de and en→tr trained on different quantities of original parallel data, back-translated data, and partial translations.[13] The results are presented in Table 5. As expected, using more back-translated data was much more helpful than using more partial translations. For en→de, in combination with 100k parallel sentences, using

[13] We did not oversample the original parallel data to match the size of original and synthetic parallel data, as commonly performed in multilingual NMT (Johnson et al., 2017).

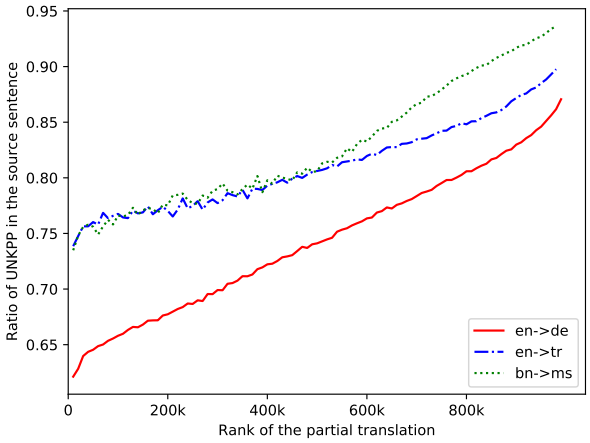

Figure 3: Ratio of UNKPP in the source sentence according to the rank of the partial translation. To smooth the curves, we drew the averaged ratio for every batch of 10k consecutively ranked partial translations.

para.	backtr	partial	en→de	en→tr
100k			7.1	9.3
100k	300k		13.2	13.3
100k		300k	10.8	10.7
100k	300k	300k	**13.6**	**13.7**
100k	1M		16.9	17.9
100k		1M	12.0	11.8
100k	1M	1M	**17.5**	**18.3**
all			26.2	13.6
all	1M		27.7	18.6
all		1M	26.4	14.7
all	1M	1M	**28.2**	**19.0**

Table 5: BLEU scores of NMT systems trained on different combinations of original parallel data (para.), back-translated data (backtr), and partial translations (partial). "all" denotes the use of all the parallel data provided by the WMT18 News Translation Task: around 5.6M and 207k sentence pairs for en→de and en→tr, respectively.

300k back-translated data achieved better results than using 300k partial translations, with an advantage of 2.4 BLEU points, in contrast to our observation with 100k additional data (Table 1). The gap was even more significant when using 1M additional data: back-translated data achieved 4.9 BLEU points higher than using partial translations. Nonetheless, for en→de, over the configuration using 100k partial translations (9.9 absolute BLEU points), using 300k and 1M partial translations improved by 0.9 and 2.1 BLEU points, respectively. We observed a similar tendency for en→tr. Searching for partial translations in more monolingual data would result in the extraction of a larger number of more accurate partial translations, and presumably help obtain even better NMT models.

Mixing partial translations, back-translated data, and parallel data remains our best configuration. We consistently obtained better results than when using only either partial translations or back-translated data as training data. Even when using the full parallel data of 5.6M sentence pairs and 1M back-translated data[14] for en→de, partial translations still brought an improvement of 0.5 BLEU points. Our best system reached 28.2 BLEU points.[15] This last result confirms that using partial translations as additional training data has also the potential to improve a state-of-the-art NMT system, while it is much more effective in low-resource scenarios.

5 Related Work

There are various methods for extracting sentence pairs from monolingual corpora. However, most of them rely on the availability of document-level information, in comparable corpora for instance, and usually for one specific domain, to efficiently extract accurate sentence pairs (Abdul Rauf and Schwenk, 2011). Other methods extract sentence pairs from completely unrelated monolingual corpora (Tillmann and Xu, 2009; Marie and Fujita, 2017). However, they still rely on an existing accurate translation model trained on large parallel data, introducing a strong bias in the retrieval of sentence pairs. Unlike existing methods, our algorithm for retrieving partial translations is efficient enough to work on large unrelated monolingual data without relying on any document-level information, and also fully unsupervised. Without any bias toward some existing parallel data, it is very suitable for low-resource scenarios.

Previous work has also exploited monolingual data in the target language for improving NMT systems (Sennrich et al., 2016; Currey et al., 2017; Hoang et al., 2018). As demonstrated in this paper, our approach is complementary to previous work, since partial translations can introduce novel information into training. To the best of our knowledge, this is the first work to propose a method for extracting sentence pairs from source and target

[14]For this experiment, the NMT system for back-translation was also trained on the full parallel data.

[15]Only 0.1 BLEU points lower than the best reported result at WMT17 for this task. The winning systems used much more back-translated data and an ensemble of several models for decoding.

unrelated monolingual corpora that can be used to train better NMT systems, without requiring any modification of current NMT model architecture.

Wang et al. (2017) proposed a method to train an RNN-based NMT system on partially aligned translations only. However, this method cannot straightforwardly be applied to the state-of-the-art Transformer architecture. In contrast, our proposed method does not assume a particular architecture of NMT, nor requires any modifications of the NMT implementation. In addition, they assume not only that a phrase table is given for their low-resource language pairs, but also that the phrase pairs in the given phrase table are very accurate. We rather focus on augmenting the training data, without assuming phrase pairs of high accuracy. Training NMT only on our extracted partial translations could also be worth investigating.

As confirmed in Section 4.1, the quality of induced phrase table nevertheless affects the usefulness of resulting partial translations. Recent advances in unsupervised MT (Artetxe et al., 2018b; Lample et al., 2018b) have shown that we can obtain phrase tables of better quality through iterating generation of synthetic parallel data and (pseudo-)training of a phrase table on such data. We plan to evaluate whether better phrase tables result in more useful partial translations.

6 Conclusion

We presented a new algorithm for extracting partial translations from unrelated monolingual corpora. Our algorithm is fully unsupervised, i.e., it does not rely on any existing human-made bilingual data, making itself suitable for low-resource language pairs. We demonstrated that very noisy partial translations can be transformed into useful training data for NMT systems with a simple post-processing. While we designed our method specifically for low-resource scenarios, we also showed that partial translations are useful for further improving a state-of-the-art NMT system trained on large parallel data and back-translated synthetic parallel data.

In our future work, we will study the impact of using more partial translations of better quality to train NMT systems. We assume that we can collect better partial translations by searching in more monolingual data. Moreover, we also observed that the top-ranked sentence pairs extracted by our algorithm may be translations of very good quality. We will study the possibility of using such sentence pairs as development data to enable the tuning of unsupervised SMT and NMT systems (Lample et al., 2018b). We will also analyze whether our partial translations are useful because of their noisy nature, since noisy synthetic data have recently been proven useful in some specific configurations (Edunov et al., 2018).

Acknowledgments

We would like to thank the reviewers for their useful comments and suggestions, and Jingyi Zhang for providing the implementation of the phrase-based forced decoding for our experiments. A part of this work was conducted under the program "Promotion of Global Communications Plan: Research, Development, and Social Demonstration of Multilingual Speech Translation Technology" of the Ministry of Internal Affairs and Communications (MIC), Japan.

References

Sadaf Abdul Rauf and Holger Schwenk. 2011. Parallel sentence generation from comparable corpora for improved SMT. *Machine Translation*, 25(4):341–375.

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, Melbourne, Australia. 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.

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2018. The best of both worlds: Combining recent advances in neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–86, Melbourne, Australia. Association for Computational Linguistics.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In

Proceedings of the Second Conference on Machine Translation, pages 148–156, Copenhagen, Denmark. 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. Association for Computational Linguistics.

Cyril Goutte, Marine Carpuat, and George Foster. 2012. The impact of sentence alignment errors on phrase-based machine translation performance. In *Proceedings of the 10th Biennial Conference of the Association for Machine Translation in the Americas*, San Diego, USA.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. 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, Canada. Association for Computational Linguistics.

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

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. Association for Computational Linguistics.

Benjamin Marie and Atsushi Fujita. 2017. Efficient extraction of pseudo-parallel sentences from raw monolingual data using word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 392–398, Vancouver, Canada. Association for Computational Linguistics.

Benjamin Marie and Atsushi Fujita. 2018. Phrase table induction using monolingual data for low-resource statistical machine translation. *ACM Transaction on Asian and Low-Resource Language Information Processing*, 17(3):16:1–16:25.

Hammam Riza, Michael Purwoadi, Gunarso, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Rapid Sun, Sethserey Sam, Sophe ap Seng, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama, and Chenchen Ding. 2016. Introduction of the Asian Language Treebank. In *Proceedings of the 2016 Conference of the Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Technique (O-COCOSDA)*, pages 1–6, Bali, Indonesia.

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.

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 *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Christoph Tillmann and Jian-ming Xu. 2009. A simple sentence-level extraction algorithm for comparable data. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 93–96, Boulder, USA. 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 *Proceedings of Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Yining Wang, Yang Zhao, Jiajun Zhang, Chengqing Zong, and Zhengshan Xue. 2017. Towards neural machine translation with partially aligned corpora. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 384–393, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura. 2017. Improving neural machine translation through phrase-based forced decoding. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 152–162, Taipei, Taiwan. Asian Federation of Natural Language Processing.