24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Learning to Reuse Translations: Guiding Neural Machine Translation with Examples

Qian Cao[1] **Shaohui Kuang**[1] and **Deyi Xiong**[1*]

Abstract. In this paper, we study the problem of enabling neural machine translation (NMT) to reuse previous translations from similar examples in target prediction. Distinguishing reusable translations from noisy segments and learning to reuse them in NMT are non-trivial. To solve these challenges, we propose an Example-Guided NMT (EGNMT) framework with two models: (1) a noise-masked encoder model that masks out noisy words according to word alignments and encodes the noise-masked sentences with an additional example encoder and (2) an auxiliary decoder model that predicts reusable words via an auxiliary decoder sharing parameters with the primary decoder. We define and implement the two models with the state-of-the-art Transformer. Experiments show that the noise-masked encoder model allows NMT to learn useful information from examples with low fuzzy match scores (FMS) while the auxiliary decoder model is good for high-FMS examples. More experiments on Chinese-English, English-German and English-Spanish translation demonstrate that the combination of the two EGNMT models can achieve improvements of up to +9 BLEU points over the baseline system and +7 BLEU points over a two-encoder Transformer.

1 Introduction

Neural machine translation [3, 30, 31, 9] captures the knowledge of the source and target language along with their correspondences as part of the encoder and decoder parameters learned from data. With this embedded and parameterized knowledge, a trained NMT model is able to translate a new source sentence into the target language.

In this paper, we consider a different translation scenario to NMT. In this scenario, in addition to a given source sentence, NMT is also provided with an example translation that contains reusable translation segments for the source sentence. The NMT model can either use the embedded knowledge in parameters or learn from the example translation on the fly to predict target words. This translation scenario is not new to machine translation as it has been studied in example-based machine translation [19] and the combination of statistical machine translation (SMT) with translation memory [16]. However, in the context of NMT, the incorporation of external symbol translations is still an open problem. We therefore propose example-guided NMT (EGNMT) to seamlessly integrate example translations into NMT.

Unlike conventional machine translation formalisms, a trained NMT model is not easy to be quickly adapted to an example translation as the model is less transparent and amenable than SMT models. To address this issue, we use a new encoder (thereafter the example encoder) to encode the example translation in EGNMT, in addition to the primary encoder for the source sentence.

As the example is not identical to the source sentence, only parts of the example translation can be used in the final translation for the source sentence. Hence the challenge is to teach EGNMT to detect and use matched translation fragments while ignoring unmatched noisy parts.

To handle this challenge, we propose two models that guide the decoder to reuse translations from examples. The first model is a noise-masked encoder model (NME). In the example encoder, we pinpoint unmatched noisy fragments in each example translation via word alignments and mask them out with a symbol "$\langle X \rangle$". The noise-masked example translation is then input to the example encoder. This model mimics human translators in paying special attention to reusable parts and ignoring those unrelated parts when an example translation is given.

Different from NME that encodes the noise-masked example translation, in the second model, we directly produce a masked translation from the example translation with an auxiliary decoder (hence the auxiliary decoder model, AD). We compare the reference translation of a source sentence in the training data with its corresponding example translation. The identical parts in the reference translation are retained while other parts are substituted with the symbol "$\langle X \rangle$". The auxiliary decoder is then used to predict the masked reference translation. It is jointly trained with the primary decoder and shares its parameters with the primary decoder. Therefore the primary decoder can learn from the auxiliary decoder to predict reusable words/phrases from the example translation. Notice that the auxiliary decoder is only used during the joint training phase.

In summary, our contributions are threefold.

- We propose an example-guided NMT framework to learn to reuse translations from examples.
- In this framework, we further propose two models: NME that encodes reusable translations in the example encoder and AD that teaches the primary decoder to directly predict reusable translations with the auxiliary decoder via parameter sharing and joint training.
- The proposed EGNMT framework can be used to any encoder-decoder based NMT. In this paper, we define EGNMT over the state-of-the-art NMT architecture Transformer [31] and evaluate EGNMT on Chinese-English, English-German and English-Spanish translation. In our experiments, the best EGNMT model achieves improvements of 4-9 BLEU points over the baseline on the three language pairs. Analyses show that the proposed model can effectively learn from example translations with different similarity scores.

[1] Soochow University, China, email: qcao@stu.suda.edu.cn, shaohuikuang@foxmail.com, dyxiong@suda.edu.cn
* Corresponding Author

2 Related Work

Translation Memory Our work is related to the studies that combine translation memory (TM) with machine translation. Various approaches have been proposed for the combination of TM and SMT. For example, Koehn and Senellart [16] propose to reuse matched segments from TM for SMT. In NMT, Gu et al. [10] propose to encode sentences from TM into vectors, which are then stored as key-value pairs to be explored by NMT. Cao and Xiong [6] regard the incorporation of TM into NMT as a multi-input problem and use a gating mechanism to combine them. Bapna and Firat [4] integrate multiple similar examples into NMT and explore different retrieval strategies. Different from these methods, we propose more fine-grained approaches to dealing with noise in matched translations.

Example-based MT In the last century, many studies have focused on the impact of examples on translation, or translation by analogy [19, 29]. Wu [36] discuss the relations of statistical, example-based and compositional MT in a three-dimensional model space because of the interplay of them. Our work can be considered as a small step in this space to integrate the example-based translation philosophy with NMT.

Using examples in neural models for other tasks In other areas of natural language processing, many researchers are interested in combining symbolic examples with neural models. Pandey et al. [22] propose a conversational model that learns to utilize similar examples to generate responses. The retrieved examples are used to create exemplar vectors that are used by the decoder to generate responses. Cai et al. [5] also introduce examples into dialogue systems, but they first generate a skeleton based on the retrieved example, and then use the skeleton to serve as an additional knowledge source for response generation. Guu et al. [11] present a new generative language model for sentences that first samples a prototype sentence and then edits it into a new sentence.

External knowledge for NMT Our work is also related to previous works that incorporate external knowledge or information into NMT. Zhou et al. [39] propose to integrate the outputs of SMT to improve the translation quality of NMT while Wang et al. [34] explore SMT recommendations in NMT. Zhang et al. [38] incorporate translation pieces into NMT within beam search. In document translation, many efforts try to encode the global context information by the aid of discourse-level approaches [17, 37, 32]. In addition to these, some studies integrate external dictionaries into NMT [2, 18] or force the NMT decoder to use given words/phrases in target translations [13, 25, 12].

Multi-task learning The way that we use the auxiliary decoder and share parameters is similar to multi-task learning in NMT. Just to name a few, Dong et al. [7] share an encoder among different translation tasks. Weng et al. [35] add a word prediction task in the process of translation. Sachan and Neubig [27] explore the parameter sharing strategies for the task of multilingual machine translation. Wang et al. [33] propose to jointly learn to translate and predict dropped pronouns.

Automatic post-editing Junczys-Dowmunt et al. [14] propose a dual-source Transformer to deal with the automatic post-editing task. In addition to input source sentences, machine translation outputs are also fed to the post-editing model. Pal et al. [20, 21] also propose a multi-encoder Transformer architecture for post-editing. Their model attends to source sentences while translation outputs are encoded. The decoder then takes the output of the multi-encoder into account for post-editing.

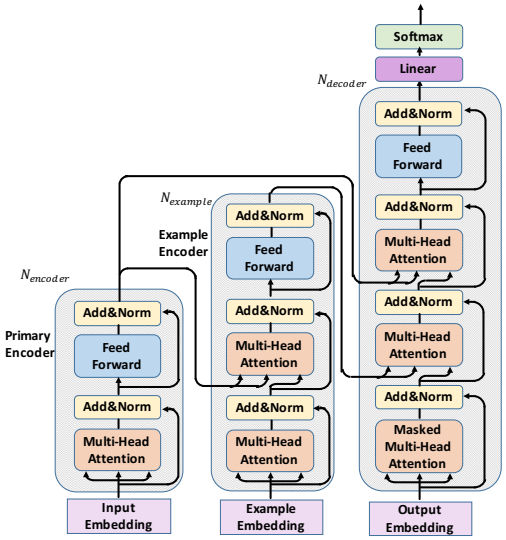

Figure 1. Architecture of the basic model for EGNMT. Positional encodings are omitted to save space.

3 Guiding NMT with Examples

The task here is to translate a source sentence into the target language from the representations of the sentence itself and a matched example translation. In this section, we first introduce how example translations are retrieved and then briefly describe the basic EGNMT model that uses one encoder for source sentences and the other for retrieved example translations. Based on this simple model, we elaborate the proposed two models: the noise-masked encoder model and auxiliary decoder model.

3.1 Example Retrieval

Given a source sentence x to be translated, we find a matched example (x^m, y^m) from an example database $D = \{(x_i, y_i)\}_1^N$ with N source-target pairs. The source part x^m of the matched example has the highest similarity score to x in D. A variety of metrics can be used to estimate this similarity score. In this paper, we first get the top n example translations by off-the-shelf search engine, and then we calculate the cosine similarity between their sentence embeddings and select the highest one as the matched example. Details will be introduced in the experiment section. Later, in order to easy to understand the similarity between the matched example and the source sentence, we also introduce the Fuzzy Match Score [16] as a measurement, which is computed as follows:

$$\mathrm{FMS}(x, x^m) = 1 - \frac{\mathrm{Levenshtein}(x, x^m)}{\max(|x|, |x^m|)} \tag{1}$$

3.2 Basic Model

Figure 1 shows the architecture for the basic model built upon the Transformer. We use two encoders: the primary encoder for encoding the source sentence x and the example encoder for the matched example translation y^m. The primary encoder is constructed following Vaswani et al. [31]:

$$\nu^{\mathrm{src}} = \mathrm{TransformerEncoder}(x) \tag{2}$$

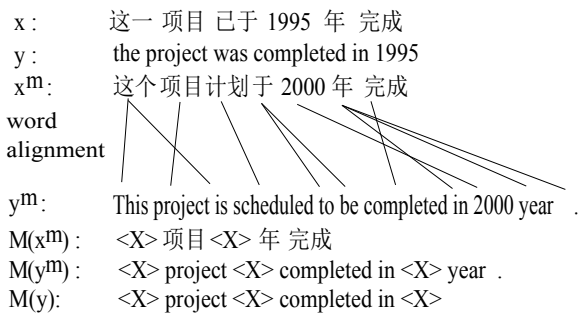

Figure 2. An example demonstrating the masking process.

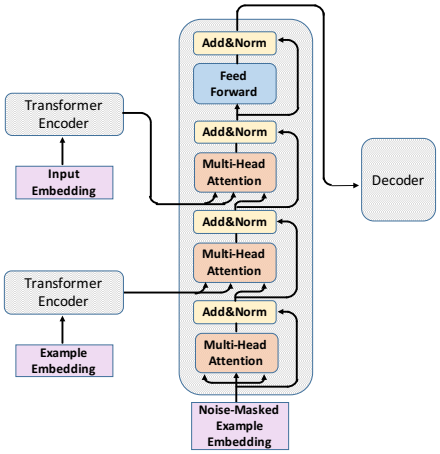

Figure 3. Architecture of the NME model for EGNMT.

The example encoder contains three sub-layers: a multi-head example self-attention layer, a multi-head source-example attention layer and a feed-forward network layer. Each sublayer is followed by a residual connection and layer normalization.

Before we describe these three sublayers in the example encoder, we first define the embedding layer. We denote the matched example translation as $y^m = [y_1^m, ..., y_L^m]$ where L is the length of y^m. The embedding layer is then calculated as:

$$Y^m = [\hat{y}_1^m, ..., \hat{y}_L^m] \tag{3}$$

$$\hat{y}_j^m = \mathrm{Emb}(y_j^m) + \mathrm{PE}(j) \tag{4}$$

where $\mathrm{Emb}(y_j^m)$ is the word embedding of y_j^m and PE is the positional encoding function.

The first sub-layer is a multi-head self-attention layer formulated as:

$$A^m = \mathrm{MultiHead}(Y^m, Y^m, Y^m) \tag{5}$$

The second sub-layer is a multi-head source-example attention which can be formulated as:

$$F^m = \mathrm{MultiHead}(A^m, \nu^{\mathrm{src}}, \nu^{\mathrm{src}}) \tag{6}$$

where ν^{src} is the output of the primary encoder. This sublayer is responsible for the attention between the matched example translation and the source sentence. The third sub-layer is a feed-forward network defined as follows:

$$D^m = \mathrm{FFN}(F^m) \tag{7}$$

Different from the primary encoder with 6 layers, the example encoder has only one single layer. In our preliminary experiments, we find that a deep example encoder is not better than a single-layer shallow encoder. This may be due to the findings of recent studies, suggesting that higher-level representations in the encoder capture semantics while lower-level states model syntax [24, 1, 8]. As the task is to borrow reusable fragments from the example translation, we do not need to fully understand the entire example translation. We conjecture that a full semantic representation of the example translation even disturbs the primary encoder to convey the meaning of the source sentence to the decoder.

In the decoder, different from Vaswani et al. [31], we insert an additional sub-layer between the masked multi-head self-attention and encoder-decoder attention. The additional sublayer is built for the attention of the decoder to the example translation representation:

$$H = \mathrm{Multihead}(\kappa, \nu^{\mathrm{exp}}, \nu^{\mathrm{exp}}) \tag{8}$$

where κ is the output of the masked multi-head self-attention, and ν^{exp} is the output of the example encoder. This sub-layer also contains residual connection and layer normalization.

3.3 Noise-Masked Encoder Model

As the source part of the matched example x^m is not identical to the source sentence x, parts of the example translation cannot be reused in producing the target translation for x. These unmatched parts may act like noisy signals to disturb the translation process of the decoder. In order to prevent these unmatched parts from interrupting the target prediction, we propose a noise-masked encoder to encode the example translation. The idea behind this new encoder is simple. We detect the unmatched parts in the example translation and use a symbol "$\langle X \rangle$" to replace them so as to mask out their effect on translation. The masking process can be defined as a function M, from which we have the noise-masked example translation $M(y^m)$ from y^m.

The masking function can be visualized with an example shown in Figure 2. Comparing the source side x^m of the matched example with the source sentence, we can find repeated source words. Keeping the repeated words and replacing other words with "$\langle X \rangle$", we obtain the masked version $M(x^m)$. Then, we use a pre-trained word alignment model to obtain word alignments between x^m and y^m. We replace words in y^m that are aligned to the masked parts in $M(x^m)$ with "$\langle X \rangle$". In this way, we finally obtain the masked example translation where only reusable parts are retained.

This masking method is based on word alignments. In practice, inaccurate word alignments will cause reusable words to be filtered out and noisy words retained. In order to minimize the negative impact of wrong word alignments as much as possible, we employ a standard transformer encoder module to encode the original example translation:

$$\nu^{\mathrm{oriexp}} = \mathrm{TransformerEncoder}(y^m) \tag{9}$$

Hence the differences between the example encoder in the basic model and NME model are twofold: (1) we replace the input y^m with $M(y^m)$; (2) we add a sub-layer between the multi-head self-attention and source-example attention, to attend to the original example translation:

$$K = \mathrm{MultiHead}(\iota, \nu^{\mathrm{oriexp}}, \nu^{\mathrm{oriexp}}) \tag{10}$$

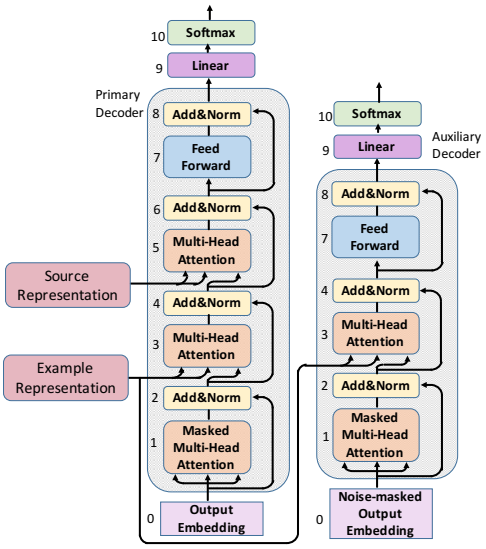

Figure 4. Architecture of the auxiliary decoder model. The modules with the same indicator numbers in the primary and auxiliary decoder share parameters.

where ι is the output of the multi-head self-attention. The architecture can be seen in Figure 3.

3.4 Auxiliary Decoder Model

In order to better leverage useful information in original example translations, we further propose an auxiliary decoder model. In this model, we directly compare the example translation y^m with the target translation y. We can easily detect translation fragments that occur both in the example and real target translation. Similarly, we mask out other words to get a masked version $M(y)$ of the target translation (see the last row in Figure 2).

As the gold target translation y is only available during the training phase, we employ an auxiliary decoder in the new model which is shown in Figure 4. The purpose for the auxiliary decoder is to predict the masked target translation $M(y)$ during the training phase from the example translation y^m and x. It can be formulated as:

$$p(M(y)|x, y^m) = \prod p(M(y)_t|M(y)_{<t}, x, y^m) \quad (11)$$

For this, we need to train an auxiliary NMT system with training instances $\{(x, y^m, M(y))\}$. The primary NMT system is trained with $\{(x, y^m, y)\}$. We jointly train these two systems to minimize a joint loss as follows:

$$L_{\text{joint}} = L_{\text{pri}} + L_{\text{aux}} \quad (12)$$

where L_{pri} is the loss for the primary NMT system while the latter L_{aux} is for the auxiliary NMT system.

During the testing phase, the auxiliary decoder is removed. We therefore share the parameters of the auxiliary decoder with the primary decoder. This is important as it allows the primary decoder to learn from the auxiliary decoder in the training phase to generate reusable parts. The joint training makes the primary decoder pay more attention to the reusable parts in the example translation by adjusting parameters in the attention network between the example encoder and the primary decoder to right directions.

3.5 Assembling NME and AD

The noise-masked encoder model and auxiliary decoder model can be combined together. In this assembling, we not only mask out noise parts in example translations in the encoder but also use the masked example translation to predict the masked target translation in the auxiliary decoder.

4 Experiments

We conducted experiments on Chinese-English, English-German and English-Spanish translation to evaluate the proposed models for EGNMT.

4.1 Experimental Settings

We implemented our example-guided NMT systems based on Tensorflow. We obtained word alignments with the tool fast-align[2]. The maximum length of training sentences is set to 50 for all languages. We applied byte pair encoding [28] with 30k merging operations. We used the stochastic gradient descent algorithm with Adam [15] to train all models. We set the beam size to 4 during decoding. We used two GPUs for training and one for decoding. We used case-insensitive 4-gram BLEU as our evaluation metric [23] and the script "multi-bleu.perl" to compute BLEU scores.

For Chinese-English corpus, we used the United National Parallel Corpus [26] from Cao and Xiong [6], which consists of official records and other parliamentary documents. The numbers of sentences in the training/development/test sets are 1.1M/804/1,614.

We also experimented our methods on English-German and English-Spanish translation. We used the JRC-Acquis corpus[3] following previous works [16, 10, 4]. We randomly selected sentences from the corpus to build the training/development/test sets. The numbers of sentences in the training/development/test sets for English-German are 0.5M/676/1,824 and 0.8M/900/2,795 for English-Spanish. We used the training sets as the example database. We firstly used the Lucene[4] to retrieve top 10 example translations from the example database excluding the sentence itself. Then we obtained the sentence embeddings of these retrieved examples with the fasttext tool[5] and calculated the cosine similarity between the source sentence and each retrieved example. Finally we selected the example with the highest similarity score as the matched example.

4.2 Chinese-English Results

Table 1 shows the results. In the table, we divide the test set into 9 groups according to the FMS values of matched example translations and show BLEU scores on each group and the entire set. We show the BLEU scores for both the baseline and matched example translations against reference translations for comparison. Additionally, we adapted the gated method proposed by Cao and Xiong [6] to the Transformer and compared with this gated Transformer model. The results of this experiment are also reported in Table 1. From the table, we can observe that

- The basic model obtains an improvement of 2.78 BLEU points over the baseline. This demonstrates the advantage of example-guided NMT: teaching NMT to learn from example translations on

[2] Available at: https://github.com/clab/fast_align
[3] Available at https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis
[4] Available at http://lucene.apache.org/
[5] Available at: https://fasttext.cc/

Table 1. BLEU scores for different models on Chinese-English translation. #S: the number of sentences. T(all data): Transformer, TB: Basic EGNMT model, NME: Noise-Masked Encoder, AD: Auxiliary Decoder, Final: Basic model+Noise-Masked Encoder+Auxiliary Decoder, MET: matched example translations, Gated: method from Cao and Xiong [6]

FMS	#S	T(all data)	TB	TB+NME	TB+AD	Final	MET	Gated
$[0.9, 1.0)$	171	55.71	69.72	66.47	86.58	88.17	**94.23**	79.95
$[0.8, 0.9)$	182	63.09	72.61	69.36	83.49	**86.96**	79.84	81.45
$[0.7, 0.8)$	178	62.56	69.67	67.63	78.55	**79.40**	67.11	74.37
$[0.6, 0.7)$	179	67.76	71.17	71.08	77.04	**77.31**	58.93	71.22
$[0.5, 0.6)$	181	69.21	70.63	70.06	72.75	**73.53**	46.99	69.91
$[0.4, 0.5)$	177	74.28	72.97	74.13	74.10	**74.50**	34.67	73.14
$[0.3, 0.4)$	180	68.39	66.46	**68.58**	66.97	66.85	22.93	65.92
$[0.2, 0.3)$	185	50.57	48.67	**52.58**	49.96	50.97	9.72	49.79
$(0.0, 0.2)$	181	**35.43**	32.05	35.35	33.53	34.66	1.18	32.16
$(0.0, 1.0)$	1,614	60.07	62.85	63.12	68.53	**69.94**	47.32	66.93
#Param	-	92M	102M	102M	102M	102M	-	104M

the fly is better than mixing examples as training data. We also find that the basic model can improve translation quality only when FMS is larger than 0.5, indicating that it suffers from noises in low-FMS example translations.

- The noise-masked encoder model is better than the basic model by 0.27 BLEU points. The model significantly improves translation quality for sentences with low-FMS example translations, which means that masking noise is really helpful. But it also slightly hurts translation quality for high-FMS (e.g., $>$0.5) sentences compared with the basic model. This may be because the noisy parts are much more dominant than the reusable parts in example translations with low FMS, which makes easier to detect and mask out noisy parts via word alignments. However, in high-FMS example translations, many words can be reused with a few unmatched words scattered in them. It is therefore risky to detect and mask out reusable words with inaccurate word alignments. Although we also attend to the original example translation, reusable words that are masked mistakenly may still not be replenished.
- The auxiliary decoder model hugely improves the performance by more than 5.68 BLEU points over the basic model. It significantly improves translation quality for high-FMS sentences by learning to reuse previously translated segments separated by scattered unmatched words. However, in low FMS intervals, its performance is still not satisfactory for that they may not distinguish the unmatched parts accurately.
- Assembling the noise-masked encoder and auxiliary decoder models together, we achieve the best performance, 7.09 BLEU points higher than the basic model and 3.01 BLEU points than the previous gated Transformer model [6]. We can improve translation quality for both high-FMS and low-FMS sentences. This is because, on the one hand, we can mask the noisy information in the example by the NME model, on the other hand, through the AD model, we can learn to let the model use the useful information. The AD model can also guide the NME model in the attendance to the original example.

4.3 Results for English-German and English-Spanish Translation

We further conducted experiments on the English-German and English-Spanish corpus. Results are shown in Table 2 and 3. We have similar findings to those on Chinese-English translation. Our best model achieves improvements of over 4 BLEU points over the basic EGNMT model. The improvements in these two language pairs are

Table 2. BLEU scores of EGNMT on English-German translation.

FMS	#S	T(all data)	TB	Final	MET
$[0.9, 1.0)$	199	68.07	77.91	82.26	**83.69**
$[0.8, 0.9)$	210	63.02	70.43	**72.89**	68.33
$[0.7, 0.8)$	205	62.20	66.23	**69.62**	61.43
$[0.6, 0.7)$	203	58.02	59.17	**63.88**	51.89
$[0.5, 0.6)$	207	57.65	**62.44**	62.30	44.55
$[0.4, 0.5)$	183	52.34	51.97	**56.49**	32.83
$[0.3, 0.4)$	205	48.72	45.36	**51.43**	23.47
$[0.2, 0.3)$	206	43.15	40.1	**44.49**	16.47
$(0.0, 0.2)$	206	**37.49**	32.02	37.03	6.35
$(0.0, 1.0)$	1,824	54.19	55.85	**59.25**	36.51

Table 3. BLEU scores of EGNMT on English-Spanish translation.

FMS	#S	T(all data)	TB	Final	MET
$[0.9, 1.0)$	367	66.94	68.38	79.94	**80.43**
$[0.8, 0.9)$	363	68.63	69.57	**73.20**	66.30
$[0.7, 0.8)$	364	68.18	69.38	**71.42**	54.12
$[0.6, 0.7)$	363	68.68	69.45	**70.21**	48.11
$[0.5, 0.6)$	274	62.04	**62.91**	62.79	32.94
$[0.4, 0.5)$	161	58.41	**58.71**	58.02	28.96
$[0.3, 0.4)$	230	58.29	57.06	**61.91**	24.09
$[0.2, 0.3)$	343	53.48	53.98	**54.02**	15.36
$(0.0, 0.2)$	330	49.68	49.80	**50.47**	9.53
$(0.0, 1.0)$	2,795	60.31	60.90	**64.35**	38.72

Table 4. The numbers of matched and unmatched noisy words in example translations. O: original matched example translations. M: noise-masked example translations. n: noisy words. m: matched words.

FMS	O(m)	O(n)	M(m)	M(n)
$[0.0, 0.2)$	148	2,404	99	265
$[0.2, 0.3)$	476	1,739	380	190
$[0.3, 0.4)$	1,007	1,370	893	225
$[0.4, 0.5)$	1,251	1,227	1,146	205
$[0.5, 0.6)$	1,559	885	1,410	228
$[0.6, 0.7)$	2,029	740	1,888	210
$[0.7, 0.8)$	2,154	536	1,987	155
$[0.8, 0.9)$	2,340	352	2,210	116
$[0.9, 1.0)$	2,424	100	2,294	33
$(0.0, 1.0)$	13,388	9,353	12,307	1,627

Table 5. A translation sample from the test set. Reusable parts are highlighted in bold.

source	feizhoudalu de wuzhuangchongtu , genyuan daduo yu pinkun ji qianfada youguan .
reference	**most armed conflicts** on the african continent **are rooted in poverty** and under-development .
x^m	youguan feizhou guojia de wuzhuangchongtu , jiu@@ qi@@ genyuan daduo yu pinkun he qianfada youguan .
y^m	**most armed conflicts** in and among african countries **are rooted in poverty** and lack of development .
Transformer	**most of the armed conflicts** on the continent **are related to poverty** and the less developed countries .
Basic Model	**most armed conflicts** in the african continent **are related to poverty** and lack of development .
Final model	**most armed conflicts** in the african continent **are rooted in poverty** and lack of development .

not as large as those in Chinese-English translation. The reason may be that the retrieved examples are not as similar to German/Spanish translations as those to English translations in the Chinese-English corpus. This can be verified by the BLEU scores of matched example translations in Chinese-English, English-German and English-Spanish corpus, which are 47.32/36.51/38.72 respectively. The more matched example translations are similar to target translations, the higher improvements our model can achieve.

5 Analysis

We look into translations generated by the proposed EGNMT models to analyze how example translations improve translation quality in this section.

5.1 Analysis on the Generation of Reusable Words

We first compared matched example translations against reference translations in the Chinese-English test set at the word level after all stop words are removed. Table 4 shows the number of matched and unmatched noisy words in example translations. The noise-masking procedure can significantly reduce the number of noisy words (9,353 vs. 1,627). 8.1% of matched words in the original example translations are filtered out due to wrong word alignments.

We collected a set of reusable words R that are present in both example and reference translations (all stop words removed). Similarly, we obtained a set of words S that occur in both example and system translations. The words in S can be regarded as words generated by EGNMT models under the (positive or negative) guidance of example translations. The intersection of R and S is the set of words that are correctly reused from example translations by EGNMT models. We computed an F_1 metric for reusable word generation as follows:

$$p = |R \cap S|/|S| \qquad r = |R \cap S|/|R|$$
$$F_1 = 2 * p * r/(p + r) \tag{13}$$

Figure 5 shows the F_1 scores for different EGNMT models. It can be seen that the proposed EGNMT models is capable of enabling the decoder to generate matched words from example translations while filtering noisy words.

The reason that the auxiliary decoder model achieves the lowest F_1 for low-FMS sentences is because the model reuses a lot of noisy words from low-FMS example translations (hence the precision is low). This indicates that low-FMS example translations have a negative impact on the AD model. The NME model is able to achieve a high precision by masking out noisy words but with a low recall for high-FMS examples by incorrectly filtering out reusable words. Combining the strengths of the two models, we can achieve high F_1 scores for both low- and high-FMS examples as shown in Figure 5 (the final model).

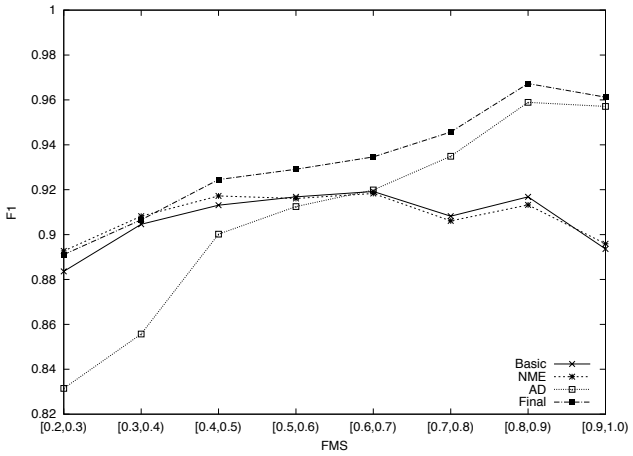

Figure 5. Reusable word generation F1 scores of EGNMT models.

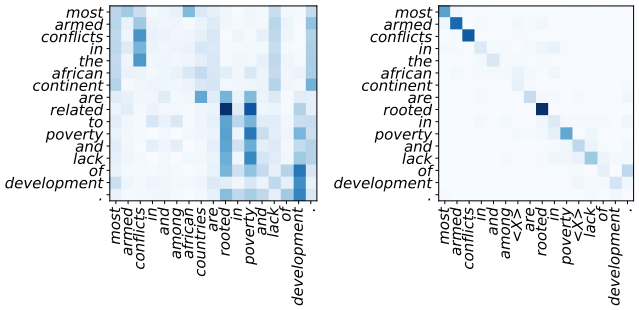

Figure 6. Visualization of attention weights between the example translation (X-axis) and the system translations (Y-axis) generated by the basic model (left) and final model (right).

5.2 Attention Visualization and Analysis

Table 5 provides a sample from the Chinese-English test set. We can see that the example translation provides two fragments that are better than the target translation generated by the baseline model. The fragment "most armed conflicts" is successfully reused by the basic model, but the fragment "are rooted in poverty" does not appear in the target translation generated by the basic model. In contrast to the two models, our final model successfully reuses the two fragments.

We further visualize and analyze attention weights between the example translation and system translation (the example encoder vs.

the primary decoder). The visualization of attention weights for this sample is shown in Figure 6. Obviously, the basic EGNMT model can use only a few reusable words as the attention weights scatter over the entire example translation rather than reusable words. The final EGNMT system that uses both the noise-masked encoder and auxiliary decoder model, by contrast, correctly detects all reusable words and enables the decoder to pay more attention to these reusable words than other words.

6 Conclusions

In this paper, we have presented EGNMT, a general and effective framework that enables the decoder to detect and take reusable translation fragments in generated target translations from the matched example translations. The noise-masking technique is introduced to filter out noisy words in example translations. The noise-masking encoder and auxiliary decoder model are proposed to learn reusable translations from low- and high-FMS example translations. Both experiments and analyses demonstrate the effectiveness of EGNMT and its advantage over mixing example translations with training data.

Acknowledgments

The present research was supported by the National Key Research and Development Project (Grant No. 2019QY1802). We would like to thank the anonymous reviewers for their insightful comments and Pei Zhang for discussion.

REFERENCES

[1] Antonios Anastasopoulos and David Chiang, 'Tied multitask learning for neural speech translation', in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pp. 82–91, (2018).

[2] Philip Arthur, Graham Neubig, and Satoshi Nakamura, 'Incorporating discrete translation lexicons into neural machine translation', in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1557–1567, (2016).

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, 'Neural machine translation by jointly learning to align and translate', in *Proceedings of the International Conference on Learning Representations (ICLR 2015)*, (2015).

[4] Ankur Bapna and Orhan Firat, 'Non-parametric adaptation for neural machine translation', in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1921–1931, (2019).

[5] Deng Cai, Yan Wang, Victoria Bi, Zhaopeng Tu, Xiaojiang Liu, Wai Lam, and Shuming Shi, 'Skeleton-to-response: Dialogue generation guided by retrieval memory', *arXiv preprint arXiv:1809.05296*, (2018).

[6] Qian Cao and Deyi Xiong, 'Encoding gated translation memory into neural machine translation', in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3042–3047, (2018).

[7] Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang, 'Multi-task learning for multiple language translation', 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)*, volume 1, pp. 1723–1732, (2015).

[8] Zi-Yi Dou, Zhaopeng Tu, Xing Wang, Shuming Shi, and Tong Zhang, 'Exploiting deep representations for neural machine translation', in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4253–4262, (2018).

[9] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin, 'Convolutional sequence to sequence learning', in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1243–1252. JMLR. org, (2017).

[10] Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li, 'Search engine guided neural machine translation', in *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pp. 5133–5140. AAAI press, (2018).

[11] Kelvin Guu, Tatsunori B Hashimoto, Yonatan Oren, and Percy Liang, 'Generating sentences by editing prototypes', *Transactions of the Association of Computational Linguistics*, **6**, 437–450, (2018).

[12] Eva Hasler, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne, 'Neural machine translation decoding with terminology constraints', in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pp. 506–512, (2018).

[13] Chris Hokamp and Qun Liu, 'Lexically constrained decoding for sequence generation using grid beam search', in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pp. 1535–1546, (2017).

[14] Marcin Junczys-Dowmunt and Roman Grundkiewicz, 'Ms-uedin submission to the wmt2018 ape shared task: Dual-source transformer for automatic post-editing', *arXiv preprint arXiv:1809.00188*, (2018).

[15] Diederik P Kingma and Jimmy Lei Ba, 'Adam: A method for stochastic optimization', in *Proceedings of the International Conference on Learning Representations (ICLR 2015)*, (2015).

[16] Philipp Koehn and Jean Senellart, 'Convergence of translation memory and statistical machine translation', in *Proceedings of AMTA Workshop on MT Research and the Translation Industry*, pp. 21–31, (2010).

[17] Shaohui Kuang and Deyi Xiong, 'Fusing recency into neural machine translation with an inter-sentence gate model', in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 607–617, (2018).

[18] Xiaoqing Li, Jiajun Zhang, and Chengqing Zong, 'Towards zero unknown word in neural machine translation', in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 2852–2858. AAAI Press, (2016).

[19] Makoto Nagao, 'A framework of a mechanical translation between japanese and english by analogy principle', *Artificial and Human Intelligence*, 351–354, (1984).

[20] Santanu Pal, Hongfei Xu, Nico Herbig, Antonio Krueger, and Josef van Genabith, 'The transference architecture for automatic post-editing', *arXiv preprint arXiv:1908.06151*, (2019).

[21] Santanu Pal, Hongfei Xu, Nico Herbig, Antonio Krüger, and Josef van Genabith, 'Usaar-dfki–the transference architecture for english–german automatic post-editing', in *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pp. 124–131, (2019).

[22] Gaurav Pandey, Danish Contractor, Vineet Kumar, and Sachindra Joshi, 'Exemplar encoder-decoder for neural conversation generation', in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pp. 1329–1338, (2018).

[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, 'Bleu: a method for automatic evaluation of machine translation', in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 311–318. Association for Computational Linguistics, (2002).

[24] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, 'Deep contextualized word representations', in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pp. 2227–2237, (2018).

[25] Matt Post and David Vilar, 'Fast lexically constrained decoding with dynamic beam allocation for neural machine translation', in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pp. 1314–1324, (2018).

[26] Alexandre Rafalovitch, Robert Dale, et al., 'United nations general assembly resolutions: A six-language parallel corpus', in *Proceedings of the MT Summit*, volume 12, pp. 292–299, (2009).

[27] Devendra Sachan and Graham Neubig, 'Parameter sharing methods for multilingual self-attentional translation models', in *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 261–

271, (2018).

[28] Rico Sennrich, Barry Haddow, and Alexandra Birch, 'Neural machine translation of rare words with subword units', in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pp. 1715–1725, (2016).

[29] Harold Somers, 'Example-based machine translation', *Machine Translation*, **14**(2), 113–157, (1999).

[30] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, 'Sequence to sequence learning with neural networks', in *Advances in Neural Information Processing Systems*, pp. 3104–3112, (2014).

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, 'Attention is all you need', in *Advances in Neural Information Processing Systems*, pp. 5998–6008, (2017).

[32] Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov, 'Context-aware neural machine translation learns anaphora resolution', in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pp. 1264–1274, (2018).

[33] Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu, 'Learning to jointly translate and predict dropped pronouns with a shared reconstruction mechanism', in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2997–3002, (2018).

[34] Xing Wang, Zhengdong Lu, Zhaopeng Tu, Hang Li, Deyi Xiong, and Min Zhang, 'Neural machine translation advised by statistical machine translation', in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 3330–3336. AAAI Press, (2017).

[35] Rongxiang Weng, Shujian Huang, Zaixiang Zheng, XIN-YU DAI, and CHEN Jiajun, 'Neural machine translation with word predictions', in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 136–145, (2017).

[36] Dekai Wu, 'MT model space: statistical versus compositional versus example-based machine translation', *Machine Translation*, **19**(3-4), 213–227, (2005).

[37] Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu, 'Improving the transformer translation model with document-level context', in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 533–542, (2018).

[38] Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura, 'Guiding neural machine translation with retrieved translation pieces', in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1325–1335, (2018).

[39] Long Zhou, Wenpeng Hu, Jiajun Zhang, and Chengqing Zong, 'Neural system combination for machine translation', in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pp. 378–384, (2017).