MetaMT, a Meta Learning Method Leveraging
Multiple Domain Data for Low Resource Machine Translation

Rumeng Li,[1*] **Xun Wang,**[1,2,3*] **Hong Yu**[1,2,3,4]

[1]School of Computer Science, University of Massachusetts Amherst, Amherst, MA, United States
[2]Department of Computer Science, University of Massachusetts Lowell, Lowell, MA, United States
[3]Bedford Veterans Affairs Medical Center, Bedford, MA, United States
[4]Department of Medicine, University of Massachusetts Medical School, Worcester, MA, United States
alicerumeng@foxmail.com, wangxun@outlook.com, Hong_Yu@uml.edu

Abstract

Neural machine translation (NMT) models have achieved state-of-the-art translation quality with a large quantity of parallel corpora available. However, their performance suffers significantly when it comes to domain-specific translations, in which training data are usually scarce. In this paper, we present a novel NMT model with a new word embedding transition technique for fast domain adaption. We propose to split parameters in the model into two groups: model parameters and meta parameters. The former are used to model the translation while the latter are used to adjust the representational space to generalize the model to different domains. We mimic the domain adaptation of the machine translation model to low-resource domains using multiple translation tasks on different domains. A new training strategy based on meta-learning is developed along with the proposed model to update the model parameters and meta parameters alternately. Experiments on datasets of different domains showed substantial improvements of NMT performances on a limited amount of data.

Introduction

Neural machine translation (NMT) is a deep learning-based approach for machine translation. A typical NMT model comprises three parts: the encoder, the decoder and the attention model. It relies on large quantities of parallel corpora for effective training of the unprecedented amount of parameters in each of its parts and to avoid overfitting (Bahdanau, Cho, and Bengio 2014; Gehring et al. 2017; Vaswani et al. 2017; Meng et al. 2019). However, finding such data remains challenging for specific domains as the construction of parallel corpus is often too expensive. The scarcity of domain-specific parallel data limits the potentials of NMT models in low-resource scenarios as previous works stated (Zoph et al. 2016; Gu et al. 2018; Koehn and Knowles 2017).

One research direction for resource lean approaches is to leverage data of multiple domains to develop robust translation systems that can be migrated to specific domains easily.

---

*Equal contribution

When migrating NMT models from one domain to another, one of the biggest challenges faced by researchers is the domain divergence. Domain divergence causes difficulties for NMT in at least two aspects. Firstly, different domains tend to have their own distinct set of vocabulary. For example, when referring to the same pathology, a medical corpus would be inclined to use "cardiovascular disease", while "heart disease" would be more common among corpora of general domains. The resulting large vocabulary aggravates the problem of data sparsity. Secondly, even identical words may carry different meanings in the context of their respective domains. For example, "Obama" is widely known as the former president of US, but the same word usually refers to a beautiful seaside city in Japan in a corpus about Japan. As such, embeddings for the same word cannot be shared across domains. This is referred to as the polysemy problem which has been addressed by many works (Yarowsky 1996; Chen, Bian, and Lin 1999; Koehn 2009) and this problem is more severe for domain-specific machine translation.

Existing approaches for domain adaptation NMT generally fall into two categories: data-centric and model-centric (Chu and Wang 2018). Data-centric methods focus on creating more data from either in-domain monolingual corpora, synthetic corpora or parallel corpora (Zhang and Zong 2016; Domhan and Hieber 2017; Chu, Dabre, and Kurohashi 2017). The creation of more in-domain training data could balance the ratio between the in- and out-of-domain data and therefore enable the learnt model to pay more attention to the target domain. The model-centric category focuses on NMT models that are specialized for domain adaptation such as Fine Tuning (Dakwale and Monz 2017) and Instance/Cost Weighting (Wang et al. 2017a; 2017b). Such methods could also leverage out-of-domain parallel corpora or in-domain monolingual corpora. Fine tuning places data of the target domain at the end of the training data stream, forcing the model to pay more attention to the target domain. Instance/Cost Weighting methods force the model to focus on the target domain by explicitly assigning higher weights to data in the target domain (or similar to the data in the target domain) during training. Both instance/cost weighting and fine tuning optimize the model towards a local min-

imum which benefits target-specific performances (Kocmi and Bojar 2017). Previous work shows the above methods yield similar improvements (Wang et al. 2017a).

Inspired by existing work, we propose to manipulate the training data of multiple domains to mimic domain adaptation and train a novel model which addresses the big vocabulary and polysemy problem. Instead of using one large lookup table to store all word representations, the designed model firstly projects all words to a semantic space that is shared by all domains. In this shared semantic space, one word is represented by a selected number of base words. This helps reduce the vocabulary size and also enables words of different domains to have different representations. A transmission layer is used in our model to conduct the mapping of word vectors.

We repeatedly train the model using a (relatively) large dataset of one domain and fine tune it on another domain with a small dataset. We adopt a meta learning strategy which enables fast parameter adaptation on small datasets. Two kinds of parameters are defined in our model as meta parameters and model parameters. The model parameters are used to learn the translation from the source sentences to the target sentences. The meta parameters are used to enhance the generalization ability of the learnt model. At fine tuning, we freeze the model parameters and adjust only the meta parameters. The meta learning strategy (Finn, Abbeel, and Levine 2017) acts as to learn a parameter initialization that can be quickly adopted to new domains.

As there are no language specific features required in the proposed method, it can be applied to any language pairs. For our evaluation we focus on the translation of two mostly widely spoken language pairs as English to Spanish translation, for which we both have a handful of datasets of different domains which can be accessed easily. Experiments show that the proposed method improved results when evaluated using BLEU (Papineni et al. 2002) as compared to existing transfer-learning NMT methods. To further verify the effectiveness of the proposed model, we use a small dataset with only three thousands sentences of electronic health records. Experiments show that the proposed model can produce high quality results for the specific domain when trained on thousands of sentences.

The contribution of this work is two-fold: firstly, a novel domain adaption training strategy based on the meta-learning policy is proposed for neural machine translation. Secondly, a novel word embedding transition technique is proposed to help handle domain divergence.

Background

The Encoder-Decoder Model and NMT

The encoder-decoder model has been used as the backbone for a wide range of NLP generation tasks including machine translation (Bahdanau, Cho, and Bengio 2014; Vaswani et al. 2017), summarization (Rush, Chopra, and Weston 2015; Nallapati et al. 2016) and dialogue generation, (Li et al. 2017; Baheti et al. 2018).

The encoder-decoder model transforms a source sentence $s = (w_1^s, w_2^s, ..., w_m^s)$ into a target sentence $t =$ $(w_1^t, w_2^t, ..., w_n^t)$ with neural networks as follows.

$$P(t|s) = \prod_{i=1}^{n} p(w_i^t|w_{<i}^t, s) \qquad (1)$$

The conditional probability of $P(t|s)$ is parameterized by the encoder-decoder framework. The encoder generates vector representations from a variable-length input sentence, and the decoder outputs a correct translation correspondingly using these vector representations.

Typical structures employed for the encoder-decoder architecture include RNN (Schuster and Paliwal 1997; Mikolov et al. 2010), recursive tree structures (Liu et al. 2014; Li et al. 2015), LSTM (Hochreiter and Schmidhuber 1997) and GRU (Cho et al. 2014) for better handling of long-dependency. A significant characteristic of recent NMT models is the wide use of attention mechanism (Bahdanau, Cho, and Bengio 2014; Li, Monroe, and Jurafsky 2016; Vaswani et al. 2017). Attention mechanism is firstly used in the decoder to enable the decoder to look at the input again and choose the most relevant parts to attend to at each step in translation. Later works like the transformer model (Vaswani et al. 2017) extensively employ the attention mechanism at both the encoder and the decoder sides to capture semantic relations inside sentences. Our proposed model is based on the transformer model, and the detailed description of which can be found in (Vaswani et al. 2017).

Domain adaptation for NMT

Data sparsity has long been a problem for NMT. Domain adaptation methods are employed for NMT when the amount of in-domain parallel corpora is insufficient for training a good NMT system. Conventional ways for domain adaptation are fine-tuning (Luong and Manning 2015; Sennrich, Haddow, and Birch 2016) where models are first trained on a high-resource domain or a mixture of data of different domains to initialize parameters which is further trained on the low-resource domain. In-domain fine-tuning comes with at least two shortcomings: firstly, it depends on the availability of sufficient amounts of in-domain data to avoid over-fitting; secondly, it results in degraded performance for all other domains. Curriculum learning has also been exploited (Zhang et al. 2019) for domain adaptation. As proved in previous work (Bengio et al. 2009), adjusting the order of training data leads to improvements in both the convergence speed and performances. (Wuebker, Simianer, and DeNero 2018) studied the fine tuning process and pointed out that it is possible to do domain adaptation by tuning only a small proportion of the model parameters. This strategy has been adopted by our work by splitting parameters into meta parameters and model parameters. (Zeng et al. 2018) proposed to generate domain-specific and domain-general representations for words. (Vilar 2018) proposed that different neurons play different roles in different domains. It is thus necessary to adjust neurons weights according to data. Instead of manipulating neurons or word representations, we use a neural mapping to consider domain divergence.

In this paper we work on the translation of English to Spanish language in different domains. A novel learning pol-

icy based on meta learning is proposed to work with the designed model. Details are explained in the following section.

Meta Learning

Meta learning has been drawing much attention of the NLP research community recently due to its ability in learning to transfer knowledge across tasks and domains (Finn, Abbeel, and Levine 2017; Hochreiter, Younger, and Conwell 2001).

Meta learning, also known as "learn to learn", intends to make machine learning models adaptive to a broader category of tasks/datasets other than the ones they are designed for or trained on. From the perspective of meta learning, training can be regarded as learning a prior over model parameters that is capable for fast adaptation on a new task/dataset.

Current meta learning methods in machine learning refer to a broad category of learning strategies and policies. Works on this topic can be roughly grouped into the following two kinds: 1) Meta learning as a policy, such as transfer learning (Sennrich, Haddow, and Birch 2016; Gu et al. 2018; Finn, Abbeel, and Levine 2017) and learning curriculum(Kocmi and Bojar 2017). 2) Meta learning as a parameter updating algorithm (Hochreiter, Younger, and Conwell 2001; Munkhdalai and Yu 2017; Andrychowicz et al. 2016).
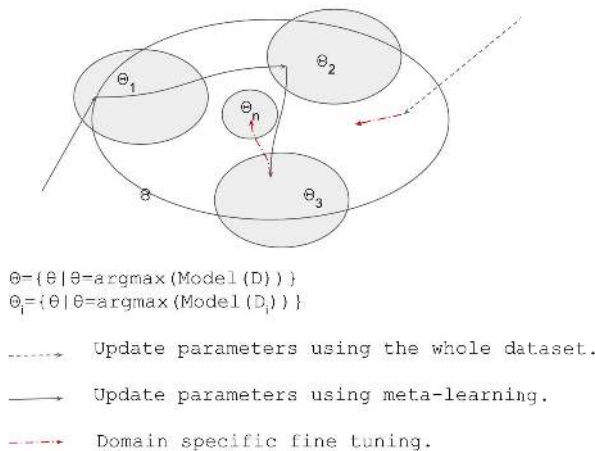

$\Theta = \{\theta \mid \theta = \text{argmax}(\text{Model}(D))\}$
$\Theta_i = \{\theta \mid \theta = \text{argmax}(\text{Model}(D_i))\}$

.... Update parameters using the whole dataset.

⎯⎯⎯ Update parameters using meta-learning.

-- -- -➔ Domain specific fine tuning.

Figure 1: A graphical illustration of the parameter updating procedures in meta learning and fine tuning.

Our proposed method falls into the 1st category and leverages the policies to learn a good machine translation model. We train the proposed model using data of different domains to search for parameters that can be easily adjusted to new domains. This is critical for domain-specific machine translation as the training data usually is limited.

Fig. 1 illustrates how to use meta learning to find a good initial parameter which can be easily adjusted to new domains. Methods like fine tuning firstly use an optimizer like SGD or Adam to learn a parameter θ which minimizes the loss function on a dataset of the general domain. And then starting from θ, the model is further tuned to make the objective loss function reach a local optimum on the targeted

domain and obtain the new parameter θ'. The problem is, as stated above, the dataset of the target domain usually is too small for further tuning. If the starting point θ is far from the "gold" parameter for domain specific NMT, we may end up with a θ' that is not fully optimized for MT on the target domain. Using meta learning policy, we alternately optimize the model on different domains and eventually learn parameters which can be easily adjusted to new domains. Meta learning techniques have been adopted for multilingual machine translation (Gu et al. 2018). This work adopts it for multiple domain machine translation.

Method

The proposed method can be built in different NMT schemes. In this work we adopt the transformer model, a state-of-the-art NMT model for its performance and speed. Figure 2 illustrates the architecture of the proposed model. We first project words into a universal semantic space to reduce domain divergence in text representations. A new learning policy is used to update different parts of the NMT model to search for a parameter which can be easily adjusted to new domains.

Domain-Invariant Word Representation

By training on the wiki crawl dataset with the fastText [1], we obtain the word embeddings of the general domain, E^G, for source words. Note that the E^G here is trained on general domain data and is not optimized for any of the specific domain we are working on nor is E^G optimized for the machine translation task. Starting from E^G, we optimize it towards domain-specific machine translation. We keep the most frequent n words in E^G and use them as the basis in the word embedding space. Here E^G is an $n * d$ matrix. The semantic space defined by E^G is used as the domain-invariant semantic space and all words from different domains will be projected into this space.

For a word w_k, its representation in domain i is written as $\vec{w_k^i}$. We map $\vec{w_k^i}$ to the space defined by E^G to obtain a new representation, $\vec{w_k}$, for w_k.

$$a_j^i = \vec{w_k^i} * A^i * E^G[j]$$
$$\vec{w_k} = \sum_{j=1}^{n} a_j^i * E^G[j] \tag{2}$$

A^i is a $d * d$ matrix which is to be learned during the training. $\vec{w_k}$ is new vector representation of w_k and will be passed to the encoder. This projection helps reduce the divergence among inputs of different domains. It also makes it possible for identical tokens of different domains to have different representations. The same strategy is also adopted in the decoder.

Encoding-Decoding

The new word representations obtained through the above approach are further passed to the encoder. Though other

---

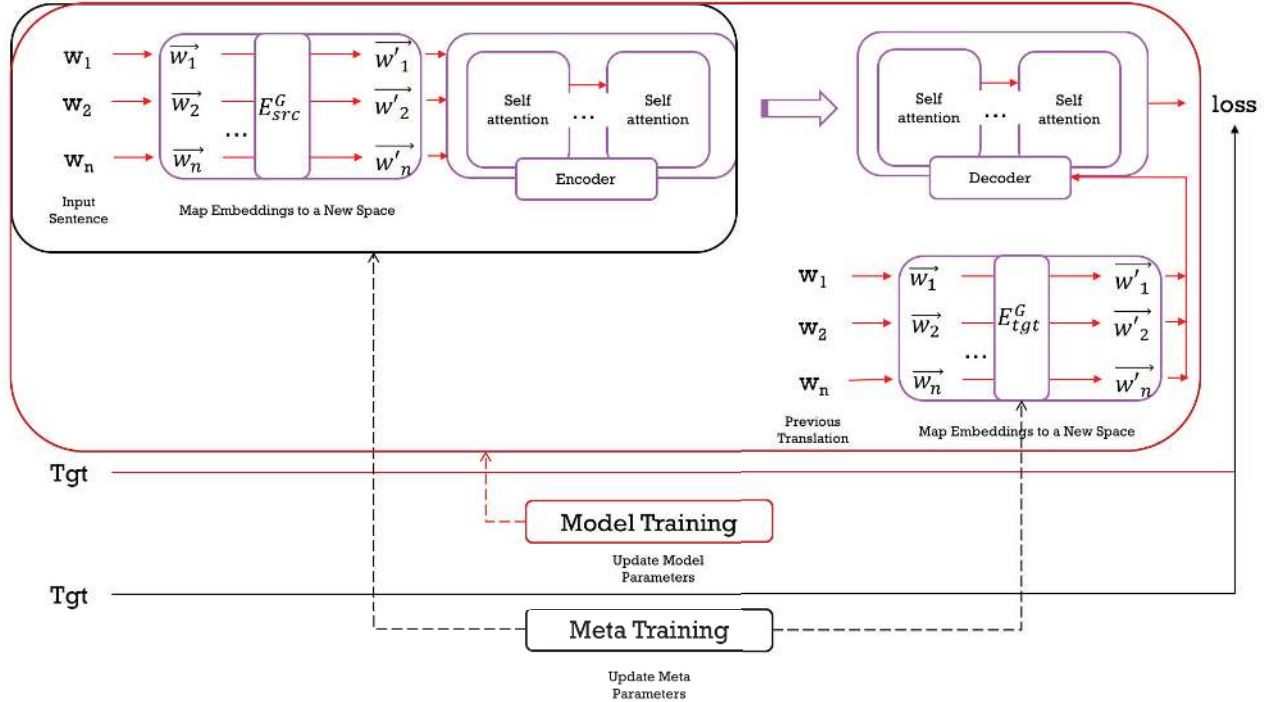[1]https://fasttext.cc/docs/en/unsupervised-tutorial.html

Figure 2: A graphical illustration of the proposed method. In training phase, source and target words are firstly projected to domain-invariant representational spaces and then are encoded/decoded. Parameters in the model are updated alternatively during training.

seq2seq architectures could also be used, here we adopt the Transformer model (Vaswani et al. 2017) with multi-head self-attentions for its effectiveness.

At decoding, we also need to produce the domain-invariant embeddings for the target words. We adopt the same technique which is used for learning source word embeddings. The model first projects target word embeddings into a domain-invariant space in which all words are represented by a selected number of base word embeddings, and then passes the embeddings to the decoder.

Learning Policy

When migrated to a new domain, a good NMT model should be quickly adjusted using a limited amount of training data. To achieve this goal, we define two kinds of parameters, one kind being model parameters θ_0 including all the parameters in the model, and the other kind being meta parameters θ_1 which includes parameters in the transmission layers and the encoder. Meta parameters θ_1 are tuned to reduce the domain divergence. The following learning policy is adopted to update the two kinds of parameters.

For the training, we use datasets of several different domains D^* of the same language pair. The data D^i of each domain i is split into three parts, a training set D_{tr}^i, a development set D_{dev}^i and a test set D_{te}^i.

The learning procedure involves several training iterations with each iteration involving two parameter updates. Firstly

we sample a pair of domains $\{i, j\}$. Then we use D_{tr}^i, D_{dev}^i to update the model parameter θ_0 in the model M. This step is referred to as model training. In the next step, using D_{dev}^j we update only the meta parameter θ_1 in M and freeze the remains. As stated above, θ_1 includes parameters in the transmission layers (of the source sides and target sides) and the encoder. As stated, meta parameters are used to fine tune the model for different domains. Intuitively it is sufficient to includes only transmission layers as meta parameters. Here we also include the encoder in the meta parameters θ_1 as experiments and previous work (Li et al. 2016; Gu et al. 2018) prove the usefulness of this strategy. Figure 1 illustrates how parameters are adjusted in this procedure.

After several iterations, we obtain a series of new parameters θ_i'. We collect all the gradients with respect to these parameters and use them to update the model as is done by (Finn, Abbeel, and Levine 2017).

Algorithm 1 describes the training procedure. Note that each training procedure involves two consecutive training processes. First, we conduct the model training using D_{tr}^i, D_{dev}^i as the training and validation datset. Then we conduct the meta training with D_{dev}^j split into two parts and used as the training set (90%) and validation set (10%). Note that although we use part of D_{dev}^j as training data in meta training, it is essentially a redistribution of training data and no extra data has been exposed to M.

Each iteration mimics a domain adaptation of the pro-

Algorithm 1 Learning Policy, Train MetaMT Using Data of Multiple Domains

---

Datasets: $\{D^0, D^1, ..., D^{n-1}\}$
Translation Model: $M(\theta)$
for $D^i, D^j \in Datasets$ **do**
  $\theta_i \longleftarrow \arg\min Loss(D_{tr}^i, D_{dev}^i; M(\theta))$ (Model training, update the translation model.)
  $\theta_i' \longleftarrow \arg\min Loss(D_{dev}^j; M(\theta_i))$ (Meta training, update the transmission layers and the encoder.)
end for
$\theta \longleftarrow \arg\min \sum_{i=0}^{n-1} Loss(D_{tr}^i; M(\theta_i'))$
$\theta \longleftarrow \arg\min Loss(D_{tr}^d, D_{dev}^d; M(\theta))$ (Fine-tune on D_d, the dataset of domain d.)
return $M(\theta)$

---

posed model. The meta parameter learns to handle domain divergence. Using meta learning, we obtain a model $M(\theta)$ which can be easily adjusted to match data of new domains. For NMT on a new domain d, we use θ as the initial parameters for M and use D^d to further train M to get a domain specific model $M(\theta_d)$. Note that when fine-tuning onto the target domain, we update all the parameters θ in the model M.

Experiments & Analysis

We conduct experiments to verify the effectiveness of the proposed model. We follow the same pre- and post- processing procedures for all the experiments unless otherwise stated.

Data

We use 7 En-Es parallel datasets of different domains for the evaluation. These datasets are all public available. For fair comparison, only subsets of these datasets with the same amount of sentence pairs are used to simulate NMT with limited data. However, this does not mean that the training data for our proposed model needs to be strictly balanced. The statistics of the datasets used in this work are shown in Table 1.

The JRCAcquis (Steinberger et al. 2006) is a legal document dataset. Global Voices (Prokopidis, Papavassiliou, and Piperidis 2016) is a collection of blogs on various topics. OpenSub is a dataset of Movie and TV subtitles (Lison and Tiedemann 2016). The Europarll (Koehn 2005) and UN Para (Rafalovitch, Dale, and others 2009) come from EU and UN proceedings. Medline (Liu and Cai 2015) is a dataset constructed from biomedical articles from NIH's MedlinePlus website. EU Bookshops dataset (Skadiņš et al. 2014) is a collection of publications in EU. All datasets are available on OPUS (Tiedemann 2012) [2], the open parallel corpus, except the Medline dataset which is the biomedical domain $ESPAC_{MedlinePlus}$ corpus built in (Liu and Cai 2015).

---

[2] http://opus.nlpl.eu/

Implementations

Our proposed model is implemented using Pytorch [3], a flexible framework for neural networks. We base our model on the transformer model (Vaswani et al. 2017) and the released Pytorch implementation [4]. Parameters are set as follows: word vector size = 300, hidden size = 512, number of layers=4, number of heads=6, dropout=0.3, batch size=64, and beam size=5. The pre-trained English and Spanish embedding are obtained using fastText (Mikolov et al. 2018) [5] on the Wikipedia datasets of English and Spanish separately. We use the top 10K En/Es words as base words to construct the base semantic spaces. At testing, we use beam search to find the best translated sentences. Decoding ends when every beam gives an $< EOS >$.

Baselines

Various methods have been proposed for neural machine translation. Among them, we compare our methods against strong baselines.

Transformer We use Transformer as a strong baseline as it has achieved promising performances in several datasets. We use a union of all the training data of different domain datasets for training.

Fine Tuning Fine-tuning is a practice used by transfer learning (Zoph et al. 2016). The model is firstly trained using available data and then fine-tuned using the task-specific dataset. As mentioned above, fine-tuning aims to find local minima for the loss function. It is stable and always achieves results comparable to other state-of-the-art systems (Chu, Dabre, and Kurohashi 2017).

MetaMT MetaMT is our proposed method. We also conduct ablation study by removing encoder side embedding projection and decoder side embedding projection, denoted as (-enc-proj,-dec-proj)

For the proposed model and the baselines, we use the same pre/post-processing and parameter settings for all methods in Table 2 and Table4 unless otherwise stated.

We use byte pair encoding (BPE) to reduce the number of unknown words (Sennrich, Haddow, and Birch 2015) for systems mentioned above (num of ops=20K).

Results & Analysis

Our evaluation is done by a single reference, case-insensitive BLEU score using the Moses package (Papineni et al. 2002). Results are reported in Table 2.

As can be seen, the proposed MetaMT model yields gains in BLEU score of about 1.0 to 2.0 points, comparing with baselines except on a few datasets (UN Data and the EU Bookshop data). Both datasets cover a wide range of topics and contain many infrequent words, which may be one of the reasons that the improvement is not significant.

---

[3] https://pytorch.org/
[4] https://github.com/pytorch/fairseq
[5] https://fasttext.cc/

Data	EU Para	UN Data	GlobalVoices	JRCAcquis	EU Bookshops	OpenSub	Medline
Sent. Pairs	520K	520K	520K	520K	520K	520K	143K
En Tokens	13.2M	13.3M	9.81M	10.11M	13.27M	3.34M	1.67M
En Avg. Length	25.41	25.58	18.87	19.45	25.53	6.44	11.65
Es Tokens	13.69M	15.82M	10.58M	11.36M	14.35M	2.96M	1.89M
Es Avg. Length	26.43	30.44	20.35	21.86	27.60	5.71	13.15

Table 1: Statistics of En-Es Datasets Used in this Work

Data	EU Para	UN Data	GlobalVoices	JRCAcquis	EU Bookshops	OpenSub	Medline
Transformer	34.04	49.88	42.28	52.22	20.78	25.54	50.95
+Fine Tune	37.10	54.81	45.24	60.07	22.29	27.31	57.58
MetaMT	39.02	55.13	47.73	61.04	22.74	29.43	59.06
-enc-proj	38.32	55.01	46.95	60.93	22.25	28.97	58.77
-dec-proj	37.34	54.92	45.73	60.52	21.02	28.35	58.22

Table 2: Performance Comparison (BLEU-4)

Experiments on Very Small Dataset

To further evaluate the performance of our proposed model, we also test our model on a very special dataset.

We built an English-Spanish parallel electronic health record (EHR) notes corpus, which comprises 3,020 paired sentences from 57 de-identified EHR discharge summaries, randomly selected from patients with type 2 diabetes in a hospital. Translation was done by a professional medical translator whose first-language is Spanish, who spent over 1000 hours building the corpus, including back translation, a very costly task. Statistics of the EHR corpus are shown in Table 3. Results show that the proposed method, which is fine tuned by thousands of sentences is able to outperform transformer.

Related Work

NMT models require a large amount of parallel data for training their parameters from the very beginning (Bahdanau, Cho, and Bengio 2014). This becomes a severe problem for translation between low-resources languages or low-resource domains. Limited data results in a large quantity of words with low frequency which is hard to represent and translate. A lot of approaches have also been explored to learn the representations and translation of infrequent words. (Domhan and Hieber 2017; van der Wees, Bisazza, and Monz 2017; Sajjad et al. 2017; Koehn and Knowles 2017).

Creating more data helps ease the infrequent word problem. But according to the Zipf's law (Zipf 2013), a few high token-frequency words account for most word occurrences in corpora. We often need a lot more data to increase the token frequencies of several infrequent words.

Besides, as previous work (Koehn and Knowles 2017) showed, sometimes NMT models suffer on corpora with high proportions of frequent words. It seems NMT models favor words with moderate frequencies rather than those with extremely high or low frequencies. Learning from mono-lingual data is another solution (Zhang and Zong 2016; Cheng et al. 2016; Domhan and Hieber 2017). But much noise is introduced when using unsupervised learning.

Meta learning (Vanschoren 2018) explores the ability to learn automatically. It enables the model to learn from a limited number of data which makes it suitable for NMT with limited sources. If the model is facing a new task which overlaps with prior tasks, the model can quickly adjust itself to the new task according to its experiences. In NMT, researchers have tested a broad category of learning policies and obtained promising results (Dakwale and Monz 2017; Munkhdalai and Yu 2017; Gu et al. 2018).

Conclusion

We present a meta learning method for neural machine translation with limited resources. The proposed model uses a collection of base words to represent words from different domains in one semantic space to reduce domain divergence. Parameters in the model are defined into two groups and updated in a new learning policy. With the novel learning policy, data of different domains is used to update different parts of the model. Experiments verify the effectiveness of the proposed method in finding parameters which can be easily adjusted to new domains with only a limited number of training examples.

The proposed training policy is not limited to neural machine translation as it does not rely on any model specific features or strategies. In the future we will further investigate applying the meta training policy to other neural machine learning models.

Acknowledgement

References

Andrychowicz, M.; Denil, M.; Gomez, S.; Hoffman, M. W.; Pfau, D.; Schaul, T.; Shillingford, B.; and De Freitas, N.

	Sent. Pairs	Word Tokens(EN)	Sent. Length (EN)	Word Tokens(ES)	Sent. Length (ES)
Training	2171	34534	15.9	36906	17.0
Development	244	3946	16.2	4191	17.2
Testing	595	10201	17.1	10900	18.3

Table 3: Statistics of EHR dataset.

Transformer	Transformer + Fine Tune	MetaMT
36.38	40.61	42.20

Table 4: Performance Comparison (BLEU-4) on Electronic Health Record Dataset

2016. Learning to learn by gradient descent by gradient descent. In *Advances in neural information processing systems*, 3981–3989.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Baheti, A.; Ritter, A.; Li, J.; and Dolan, B. 2018. Generating more interesting responses in neural conversation models with distributional constraints. *arXiv preprint arXiv:1809.01215*.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48. ACM.

Chen, H.-H.; Bian, G.-W.; and Lin, W.-C. 1999. Resolving translation ambiguity and target polysemy in cross-language information retrieval. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 215–222. Association for Computational Linguistics.

Cheng, Y.; Xu, W.; He, Z.; He, W.; Wu, H.; Sun, M.; and Liu, Y. 2016. Semi-supervised learning for neural machine translation. *arXiv preprint arXiv:1606.04596*.

Cho, K.; Van Merriënboer, B.; Bahdanau, D.; and Bengio, Y. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Chu, C., and Wang, R. 2018. A survey of domain adaptation for neural machine translation. *arXiv preprint arXiv:1806.00258*.

Chu, C.; Dabre, R.; and Kurohashi, S. 2017. An empirical comparison of simple domain adaptation methods for neural machine translation. *arXiv preprint arXiv:1701.03214*.

Dakwale, P., and Monz, C. 2017. Fine-tuning for neural machine translation with limited degradation across in-and out-of-domain data. *Proceedings of the XVI Machine Translation Summit* 117.

Domhan, T., and Hieber, F. 2017. Using target-side monolingual data for neural machine translation through multi-task learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1500–1505.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1126–1135. JMLR. org.

Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; and Dauphin, Y. N. 2017. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*.

Gu, J.; Wang, Y.; Chen, Y.; Cho, K.; and Li, V. O. 2018. Meta-learning for low-resource neural machine translation. *arXiv preprint arXiv:1808.08437*.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Hochreiter, S.; Younger, A. S.; and Conwell, P. R. 2001. Learning to learn using gradient descent. In *International Conference on Artificial Neural Networks*, 87–94. Springer.

Kocmi, T., and Bojar, O. 2017. Curriculum learning and minibatch bucketing in neural machine translation. *CoRR* abs/1707.09533.

Koehn, P., and Knowles, R. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, 79–86. Citeseer.

Koehn, P. 2009. *Statistical machine translation*. Cambridge University Press.

Li, J.; Luong, M.-T.; Jurafsky, D.; and Hovy, E. 2015. When are tree structures necessary for deep learning of representations? *arXiv preprint arXiv:1503.00185*.

Li, J.; Monroe, W.; Ritter, A.; Galley, M.; Gao, J.; and Jurafsky, D. 2016. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.

Li, J.; Monroe, W.; Shi, T.; Jean, S.; Ritter, A.; and Jurafsky, D. 2017. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*.

Li, J.; Monroe, W.; and Jurafsky, D. 2016. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv:1611.08562*.

Lison, P., and Tiedemann, J. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Liu, W., and Cai, S. 2015. Translating electronic health record notes from english to spanish: A preliminary study. *Proceedings of BioNLP 15* 134–140.

Liu, S.; Yang, N.; Li, M.; and Zhou, M. 2014. A recursive recurrent neural network for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1491–1500.

Luong, M.-T., and Manning, C. D. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*, 76–79.

Meng, Y.; Ren, X.; Sun, Z.; Li, X.; Yuan, A.; Wu, F.; and Li, J. 2019. Large-scale pretraining for neural machine translation with tens of billions of sentence pairs. *arXiv preprint arXiv:1909.11861*.

Mikolov, T.; Karafiát, M.; Burget, L.; Černockỳ, J.; and Khudanpur, S. 2010. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*.

Mikolov, T.; Grave, E.; Bojanowski, P.; Puhrsch, C.; and Joulin, A. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Munkhdalai, T., and Yu, H. 2017. Meta networks. *CoRR* abs/1703.00837.

Nallapati, R.; Zhou, B.; Gulcehre, C.; Xiang, B.; et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics.

Prokopidis, P.; Papavassiliou, V.; and Piperidis, S. 2016. Parallel global voices: a collection of multilingual corpora with citizen media stories. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 900–905.

Rafalovitch, A.; Dale, R.; et al. 2009. United nations general assembly resolutions: A six-language parallel corpus. In *Proceedings of the MT Summit*, volume 12, 292–299.

Rush, A. M.; Chopra, S.; and Weston, J. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Sajjad, H.; Durrani, N.; Dalvi, F.; Belinkov, Y.; and Vogel, S. 2017. Neural machine translation training in a multi-domain scenario. *arXiv preprint arXiv:1708.08712*.

Schuster, M., and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45(11):2673–2681.

Sennrich, R.; Haddow, B.; and Birch, A. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Edinburgh neural machine translation systems for wmt 16. *arXiv preprint arXiv:1606.02891*.

Skadiņš, R.; Tiedemann, J.; Rozis, R.; and Deksne, D. 2014. Billions of parallel words for free: Building and using the eu bookshop corpus. In *Proceedings of LREC*.

Steinberger, R.; Pouliquen, B.; Widiger, A.; Ignat, C.; Erjavec, T.; Tufis, D.; and Varga, D. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv preprint cs/0609058*.

Tiedemann, J. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, 2214–2218.

van der Wees, M.; Bisazza, A.; and Monz, C. 2017. Dynamic data selection for neural machine translation. *CoRR* abs/1708.00712.

Vanschoren, J. 2018. Meta-learning: A survey. *CoRR* abs/1810.03548.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Vilar, D. 2018. Learning hidden unit contribution for adapting neural machine translation models. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 500–505.

Wang, R.; Utiyama, M.; Liu, L.; Chen, K.; and Sumita, E. 2017a. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1482–1488.

Wang, R.; Utiyama, M.; Liu, L.; Chen, K.; and Sumita, E. 2017b. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1482–1488. Association for Computational Linguistics.

Wuebker, J.; Simianer, P.; and DeNero, J. 2018. Compact personalized models for neural machine translation. *arXiv preprint arXiv:1811.01990*.

Yarowsky, D. E. 1996. Three machine learning algorithms for lexical ambiguity resolution.

Zeng, J.; Su, J.; Wen, H.; Liu, Y.; Xie, J.; Yin, Y.; and Zhao, J. 2018. Multi-domain neural machine translation with word-level domain context discrimination. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 447–457.

Zhang, J., and Zong, C. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1535–1545.

Zhang, X.; Shapiro, P.; Kumar, G.; McNamee, P.; Carpuat, M.; and Duh, K. 2019. Curriculum learning for domain adaptation in neural machine translation. *arXiv preprint arXiv:1905.05816*.

Zipf, G. K. 2013. *The psycho-biology of language: An introduction to dynamic philology*. Routledge.

Zoph, B.; Yuret, D.; May, J.; and Knight, K. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.