Multi-Task Learning for Multiple Language Translation

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu and Haifeng Wang
Baidu Inc, Beijing, China
{dongdaxiang, wu_hua, hewei06, yudianhai, wanghaifeng}@baidu.com

Abstract

In this paper, we investigate the problem of learning a machine translation model that can simultaneously translate sentences from one source language to multiple target languages. Our solution is inspired by the recently proposed neural machine translation model which generalizes machine translation as a sequence learning problem. We extend the neural machine translation to a multi-task learning framework which shares source language representation and separates the modeling of different target language translation. Our framework can be applied to situations where either large amounts of parallel data or limited parallel data is available. Experiments show that our multi-task learning model is able to achieve significantly higher translation quality over individually learned model in both situations on the data sets publicly available.

1 Introduction

Translation from one source language to multiple target languages at the same time is a difficult task for humans. A person often needs to be familiar with specific translation rules for different language pairs. Machine translation systems suffer from the same problems too. Under the current classic statistical machine translation framework, it is hard to share information across different phrase tables among different language pairs. Translation quality decreases rapidly when the size of training corpus for some minority language pairs becomes smaller. To conquer the problems described above, we propose a multi-task learning framework based on a sequence learning model to conduct machine translation from one source language to multiple target languages, inspired by the recently proposed neural machine translation(NMT) framework proposed by Bahdanau et al. (2014). Specifically, we extend the recurrent neural network based encoder-decoder framework to a multi-task learning model that shares an encoder across all language pairs and utilize a different decoder for each target language.

The neural machine translation approach has recently achieved promising results in improving translation quality. Different from conventional statistical machine translation approaches, neural machine translation approaches aim at learning a radically end-to-end neural network model to optimize translation performance by generalizing machine translation as a sequence learning problem. Based on the neural translation framework, the lexical sparsity problem and the long-range dependency problem in traditional statistical machine translation can be alleviated through neural networks such as long short-term memory networks which provide great lexical generalization and long-term sequence memorization abilities.

The basic assumption of our proposed framework is that many languages differ lexically but are closely related on the semantic and/or the syntactic levels. We explore such correlation across different target languages and realize it under a multi-task learning framework. We treat a separate translation direction as a sub RNN encode-decoder task in this framework which shares the same encoder (i.e. the same source language representation) across different translation directions, and use a different decoder for each specific target language. In this way, this proposed multi-task learning model can make full use of the source language corpora across different language pairs. Since the encoder part shares the same source language representation

*Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics
and the 7th International Joint Conference on Natural Language Processing*, pages 1723–1732,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

across all the translation tasks, it may learn semantic and structured predictive representations that can not be learned with only a small amount of data. Moreover, during training we jointly model the alignment and the translation process simultaneously for different language pairs under the same framework. For example, when we simultaneously translate from English into Korean and Japanese, we can jointly learn latent similar semantic and structure information across Korea and Japanese because these two languages share some common language structures.

The contribution of this work is three folds. First, we propose a unified machine learning framework to explore the problem of translating one source language into multiple target languages. To the best of our knowledge, this problem has not been studied carefully in the statistical machine translation field before. Second, given large-scale training corpora for different language pairs, we show that our framework can improve translation quality on each target language as compared with the neural translation model trained on a single language pair. Finally, our framework is able to alleviate the data scarcity problem, using language pairs with large-scale parallel training corpora to improve the translation quality of those with few parallel training corpus.

The following sections will be organized as follows: in section 2, related work will be described, and in section 3, we will describe our multi-task learning method. Experiments that demonstrate the effectiveness of our framework will be described in section 4. Lastly, we will conclude our work in section 5.

2 Related Work

Statistical machine translation systems often rely on large-scale parallel and monolingual training corpora to generate translations of high quality. Unfortunately, statistical machine translation system often suffers from data sparsity problem due to the fact that phrase tables are extracted from the limited bilingual corpus. Much work has been done to address the data sparsity problem such as the pivot language approach (Wu and Wang, 2007; Cohn and Lapata, 2007) and deep learning techniques (Devlin et al., 2014; Gao et al., 2014; Sundermeyer et al., 2014; Liu et al., 2014).

On the problem of how to translate one source language to many target languages within one model, few work has been done in statistical machine translation. A related work in SMT is the pivot language approach for statistical machine translation which uses a commonly used language as a ”bridge” to generate source-target translation for language pair with few training corpus. Pivot based statistical machine translation is crucial in machine translation for resource-poor language pairs, such as Spanish to Chinese. Considering the problem of translating one source language to many target languages, pivot based SMT approaches does work well given a large-scale source language to pivot language bilingual corpus and large-scale pivot language to target languages corpus. However, in reality, language pairs between English and many other target languages may not be large enough, and pivot-based SMT sometimes fails to handle this problem. Our approach handles one to many target language translation in a different way that we directly learn an end to multi-end translation system that does not need a pivot language based on the idea of neural machine translation.

Neural Machine translation is a emerging new field in machine translation, proposed by several work recently (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2014), aiming at end-to-end machine translation without phrase table extraction and language model training. Different from traditional statistical machine translation, neural machine translation encodes a variable-length source sentence with a recurrent neural network into a fixed-length vector representation and decodes it with another recurrent neural network from a fixed-length vector into variable-length target sentence. A typical model is the RNN encoder-decoder approach proposed by Bahdanau et al. (2014), which utilizes a bidirectional recurrent neural network to compress the source sentence information and fits the conditional probability of words in target languages with a recurrent manner. Moreover, soft alignment parameters are considered in this model. As a specific example model in this paper, we adopt a RNN encoder-decoder neural machine translation model for multi-task learning, though all neural network based model can be adapted in our framework.

In the natural language processing field, a notable work related with multi-task learning was proposed by Collobert et al. (2011) which shared common representation for input words and solve different traditional NLP tasks such as part-of-Speech tagging, name entity recognition and semantic role labeling within one framework, where the convolutional neural network model was used. Hatori et al. (2012) proposed to jointly train word segmentation, POS tagging and dependency parsing, which can also be seen as a multi-task learning approach. Similar idea has also been proposed by Li et al. (2014) in Chinese dependency parsing. Most of multi-task learning or joint training frameworks can be summarized as parameter sharing approaches proposed by Ando and Zhang (2005) where they jointly trained models and shared center parameters in NLP tasks. Researchers have also explored similar approaches (Sennrich et al., 2013; Cui et al., 2013) in statistical machine translation which are often refered as domain adaption. Our work explores the possibility of machine translation under the multi-task framework by using the recurrent neural networks. To the best of our knowledge, this is the first trial of end to end machine translation under multi-task learning framework.

3 Multi-task Model for Multiple Language Translation

Our model is a general framework for translating from one source language to many targets. The model we build in this section is a recurrent neural network based encoder-decoder model with multiple target tasks, and each task is a specific translation direction. Different tasks share the same translation encoder across different language pairs. We will describe model details in this section.

3.1 Objective Function

Given a pair of training sentence $\{\mathbf{x}, \mathbf{y}\}$, a standard recurrent neural network based encoder-decoder machine translation model fits a parameterized model to maximize the conditional probability of a target sentence $\mathbf{y}$ given a source sentence $\mathbf{x}$, i.e., $\operatorname{argmax} p(\mathbf{y}|\mathbf{x})$. We extend this into multiple languages setting. In particular, suppose we want to translate from English to many different languages, for instance, French(Fr), Dutch(Nl), Spanish(Es). Parallel training data will be collected before training, i.e. En-Fr, En-Nl, En-Es parallel sentences. Since the English representation of the three language pairs is shared in one encoder, the objective function we optimize is the summation of several conditional probability terms conditioned on representation generated from the same encoder.

$$L(\Theta) = \operatorname*{argmax}_{\Theta}(\sum_{T_p}(\frac{1}{N_p}\sum_{i}^{N_p}\log p(\mathbf{y_i}^{T_p}|\mathbf{x_i}^{T_p};\Theta)) \tag{1}$$

where $\Theta = \{\Theta_{src}, \Theta_{trg_{T_p}}, T_p = 1, 2, \cdots, T_m\}$, Θ_{src} is a collection of parameters for source encoder. And $\Theta_{trg_{T_p}}$ is the parameter set of the T_pth target language. N_p is the size of parallel training corpus of the pth language pair. For different target languages, the target encoder parameters are seperated so we have T_m decoders to optimize. This parameter sharing strategy makes different language pairs maintain the same semantic and structure information of the source language and learn to translate into target languages in different decoders.

3.2 Model Details

Suppose we have several language pairs $(\mathbf{x}^{T_p}, \mathbf{y}^{T_p})$ where T_p denotes the index of the T_pth language pair. For a specific language pair, given a sequence of source sentence input $(x_1^{T_p}, x_2^{T_p}, \cdots, x_n^{T_p})$, the goal is to jointly maximize the conditional probability for each generated target word. The probability of generating the tth target word is estimated as:

$$p(y_t^{T_p}|y_1^{T_p}, \cdots, y_{t-1}^{T_p}, \mathbf{x}^{T_p}) = g(y_{t-1}^{T_p}, s_t^{T_p}, c_t^{T_p}) \tag{2}$$

where the function g is parameterized by a feedforward neural network with a softmax output layer. And g can be viewed as a probability predictor with neural networks. $s_t^{T_p}$ is a recurrent neural network hidden state at time t, which can be estimated as:

$$s_t^{T_p} = f(s_{t-1}^{T_p}, y_{t-1}^{T_p}, c_t^{T_p}) \tag{3}$$

the context vector $c_t^{T_p}$ depends on a sequence of annotations $(h_1, \cdots, h_{L_x})$ to which an encoder maps the input sentence, where L_x is the number of tokens in $\mathbf{x}$. Each annotation h_i is a bidirectional recurrent representation with forward and backward sequence information around the ith word.

$$\mathbf{c_t}^{T_p} = \sum_{j=1}^{L_x} a_{ij}^{T_p}\mathbf{h_j} \tag{4}$$

where the weight $a_{tj}^{T_p}$ is a scalar computed by

$$a_{tj}^{T_p} = \frac{exp(e_{tj}^{T_p})}{\sum_{k=1}^{L_x^{T_p}} exp(e_{tk}^{T_p})} \tag{5}$$

$$e_{tj}^{T_p} = \phi(\mathbf{s_{t-1}}^{T_p}, \mathbf{h_j}) \tag{6}$$

$a_{tj}^{T_p}$ is a normalized score of e_{tj} which is a soft alignment model measuring how well the input context around the jth word and the output word in the tth position match. e_{tj} is modeled through a perceptron-like function:

$$\phi(\mathbf{x}, \mathbf{y}) = \mathbf{v}^T tanh(\mathbf{W}\mathbf{x} + \mathbf{U}\mathbf{y}) \tag{7}$$

To compute $\mathbf{h_j}$, a bidirectional recurrent neural network is used. In the bidirectional recurrent neural network, the representation of a forward sequence and a backward sequence of the input sentence is estimated and concatenated to be a single vector. This concatenated vector can be used to translate multiple languages during the test time.

$$\mathbf{h_j} = [\overrightarrow{\mathbf{h_j}}; \overleftarrow{\mathbf{h_j}}]^T \tag{8}$$

From a probabilistic perspective, our model is able to learn the conditional distribution of several target languages given the same source corpus. Thus, the recurrent encoder-decoders are jointly trained with several conditional probabilities added together. As for the bidirectional recurrent neural network module, we adopt the recently proposed gated recurrent neural network (Cho et al., 2014). The gated recurrent neural network is shown to have promising results in several sequence learning problem such as speech recognition and machine translation where input and output sequences are of variable length. It is also shown that the gated recurrent neural network has the ability to address the gradient vanishing problem compared with the traditional recurrent neural network, and thus the long-range dependency problem in machine translation can be handled well. In our multi-task learning framework, the parameters of the gated recurrent neural network in the encoder are shared, which is formulated as follows.

$$\mathbf{h_t} = (\mathbf{I} - \mathbf{z_t}) \odot \mathbf{h_{t-1}} + \mathbf{z_t} \odot \hat{\mathbf{h_t}} \tag{9}$$

$$\mathbf{z_t} = \sigma(\mathbf{W_z}\mathbf{x_t} + \mathbf{U_z}\mathbf{h_{t-1}}) \tag{10}$$

$$\hat{\mathbf{h_t}} = tanh(\mathbf{W}\mathbf{x_t} + \mathbf{U}(\mathbf{r_t} \odot \mathbf{h_{t-1}})) \tag{11}$$

$$\mathbf{r_t} = \sigma(\mathbf{W_r}\mathbf{x_t} + \mathbf{U_r}\mathbf{h_{t-1}}) \tag{12}$$

Where $\mathbf{I}$ is identity vector and $\odot$ denotes element wise product between vectors. $tanh(x)$ and $\sigma(x)$ are nonlinear transformation functions that can be applied element-wise on vectors. The recurrent computation procedure is illustrated in 1, where x_t denotes one-hot vector for the tth word in a sequence.

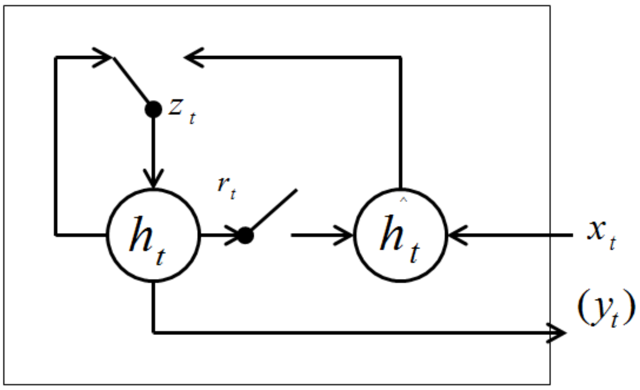

Figure 1: Gated recurrent neural network computation, where r_t is a reset gate responsible for memory unit elimination, and z_t can be viewed as a soft weight between current state information and history information.

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{13}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{14}$$

The overall model is illustrated in Figure 2 where the multi-task learning framework with four target languages is demonstrated. The soft alignment parameters A_i for each encoder-decoder are different and only the bidirectional recurrent neural network representation is shared.

3.3 Optimization

The optimization approach we use is the mini-batch stochastic gradient descent approach (Bottou, 1991). The only difference between our optimization and the commonly used stochastic gradient descent is that we learn several mini-batches within a fixed language pair for several mini-batch iterations and then move onto the next language pair. Our optimization procedure is shown in Figure 3.

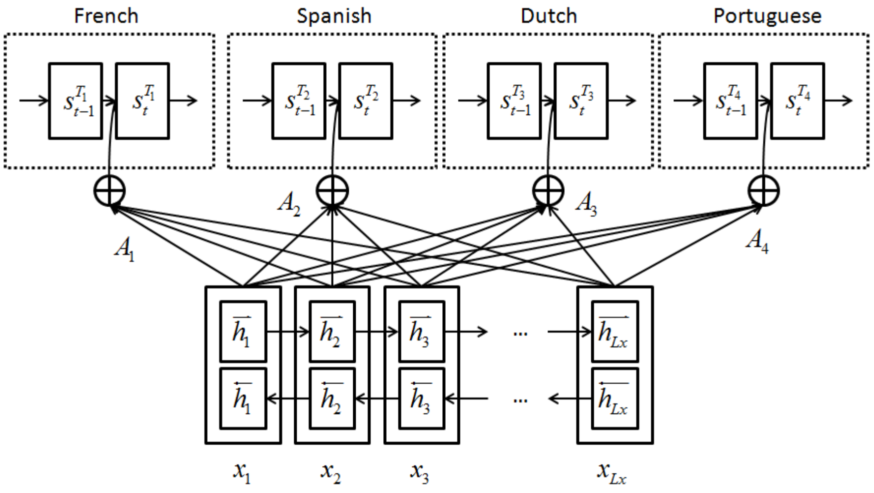
French
Spanish
Dutch
Portuguese
$s_{t-1}^{T_1}$ $s_t^{T_1}$
$s_{t-1}^{T_2}$ $s_t^{T_2}$
$s_{t-1}^{T_3}$ $s_t^{T_3}$
$s_{t-1}^{T_4}$ $s_t^{T_4}$
A_1
A_2
A_3
A_4
$\overrightarrow{h_1}$ $\overrightarrow{h_2}$ $\overrightarrow{h_3}$ ... $\overrightarrow{h_{Lx}}$
$\overleftarrow{h_1}$ $\overleftarrow{h_2}$ $\overleftarrow{h_3}$ ... $\overleftarrow{h_{Lx}}$
x_1 x_2 x_3 x_{Lx}

Figure 2: Multi-task learning framework for multiple-target language translation

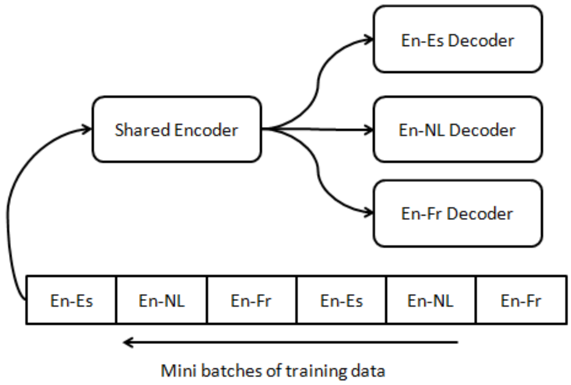

Figure 3: Optimization for end to multi-end model

3.4 Translation with Beam Search

Although parallel corpora are available for the encoder and the decoder modeling in the training phrase, the ground truth is not available during test time. During test time, translation is produced by finding the most likely sequence via beam search.

$$\hat{\mathbf{Y}} = \operatorname*{argmax}_{\mathbf{Y}} p(\mathbf{Y}^{T_p}|\mathbf{S}^{T_p}) \qquad (15)$$

Given the target direction we want to translate to, beam search is performed with the shared encoder and a specific target decoder where search space belongs to the decoder T_p. We adopt beam search algorithm similar as it is used in SMT system (Koehn, 2004) except that we only utilize scores produced by each decoder as features. The size of beam is 10 in our experiments for speedup consideration. Beam search is ended until the end-of-sentence **eos** symbol is generated.

4 Experiments

We conducted two groups of experiments to show the effectiveness of our framework. The goal of the first experiment is to show that multi-task learning helps to improve translation performance given enough training corpora for all language pairs. In the second experiment, we show that for some resource-poor language pairs with a few parallel training data, their translation performance could be improved as well.

4.1 Dataset

The Europarl corpus is a multi-lingual corpus including 21 European languages. Here we only choose four language pairs for our experiments. The source language is English for all language pairs. And the target languages are Spanish (Es), French (Fr), Portuguese (Pt) and Dutch (Nl). To demonstrate the validity of our learning framework, we do some preprocessing on the training set. For the source language, we use 30k of the most frequent words for source language vocabulary which is shared across different language pairs and 30k most frequent words for each target language. Out-of-vocabulary words are denoted as unknown words, and we maintain different unknown word labels for different languages. For test sets, we also restrict all words in the test set to be from our training vocabulary and mark the OOV words as the corresponding labels as in the training data. The size of training corpus in experiment 1 and 2 is listed in Table 1 where

Training Data Information						
Lang	En-Es	En-Fr	En-Nl	En-Pt	En-Nl-sub	En-Pt-sub
Sent size	1,965,734	2,007,723	1,997,775	1,960,407	300,000	300,000
Src tokens	49,158,635	50,263,003	49,533,217	49,283,373	8,362,323	8,260,690
Trg tokens	51,622,215	52,525,000	50,661,711	54,996,139	8,590,245	8,334,454

Table 1: Size of training corpus for different language pairs

En-Nl-sub and En-Pt-sub are sub-sampled data set of the full corpus. The full parallel training corpus is available from the EuroParl corpus, downloaded from EuroParl public websites[1]. We mimic the situation that there are only a small-scale parallel corpus available for some language pairs by randomly sub-sampling the training data. The parallel corpus of English-Portuguese and English-Dutch are sub-sampled to approximately 15% of the full corpus size. We select two data sets as our test data. One is the EuroParl Common test set[2] in European Parliament Corpus, the other is WMT 2013 data set[3]. For WMT 2013, only En-Fr, En-Es are available and we evaluate the translation performance only on these two test sets. Information of test sets is shown in Table 2.

Language pair	En-Es	En-Fr	En-Nl	En-Pt
Common test	1755	1755	1755	1755
WMT2013	3000	3000	-	-

Table 2: Size of test set in EuroParl Common testset and WMT2013

4.2 Training Details

Our model is trained on Graphic Processing Unit K40. Our implementation is based on the open source deep learning package Theano (Bastien et al., 2012) so that we do not need to take care about gradient computations. During training, we randomly shuffle our parallel training corpus for each language pair at each epoch of our learning process. The optimization algorithm and model hyper parameters are listed below.

- Initialization of all parameters are from uniform distribution between -0.01 and 0.01.
- We use stochastic gradient descent with recently proposed learning rate decay strategy Ada-Delta (Zeiler, 2012).
- Mini batch size in our model is set to 50 so that the convergence speed is fast.
- We train 1000 mini batches of data in one language pair before we switch to the next language pair.
- For word representation dimensionality, we use 1000 for both source language and target language.
- The size of hidden layer is set to 1000.

We trained our multi-task model with a multi-GPU implementation due to the limitation of Graphic memory. And each target decoder is trained within one GPU card, and we synchronize our source encoder every 1000 batches among all GPU card. Our model costs about 72 hours on full large parallel corpora training until convergence and about 24 hours on partial parallel corpora training. During decoding, our implementation on GPU costs about 0.5 second per sentence.

4.3 Evaluation

We evaluate the effectiveness of our method with EuroParl Common testset and WMT 2013 dataset. BLEU-4 (Papineni et al., 2002) is used as the evaluation metric. We evaluate BLEU scores on EuroParl Common test set with multi-task NMT models and single NMT models to demonstrate the validity of our multi-task learning framework. On the WMT 2013 data sets, we compare performance of separately trained NMT models, multi-task NMT models and Moses. We use the EuroParl Common test set as a development set in both neural machine translation experiments and Moses experiments. For single NMT models and multi-task NMT models, we select the best model with the highest BLEU score in the EuroParl Common testset and apply it to the WMT 2013 dataset. Note that our experiment settings in NMT is equivalent with Moses, considering the same training corpus, development sets and test sets.

[1] http:www.statmt.orgeuroparl

[2] http://www.statmt.org/wmt14/test.tgz

[3] http://matrix.statmt.org/test_sets

4.4 Experimental Results

We report our results of three experiments to show the validity of our methods. In the first experiment, we train multi-task learning model jointly on all four parallel corpora and compare BLEU scores with models trained separately on each parallel corpora. In the second experiment, we utilize the same training procedures as Experiment 1, except that we mimic the situation where some parallel corpora are resource-poor and maintain only 15% data on two parallel training corpora. In experiment 3, we test our learned model from experiment 1 and experiment 2 on WMT 2013 dataset. Table 3 and 4 show the case-insensitive BLEU scores on the Europarl common test data. Models learned from the multi-task learning framework significantly outperform the models trained separately. Table 4 shows that given only 15% of parallel training corpus of English-Dutch and English-Portuguese, it is possible to improve translation performance on all the target languages as well. This result makes sense because the correlated languages benefit from each other by sharing the same predictive structure, e.g. French, Spanish and Portuguese, all of which are from Latin. We also notice that even though Dutch is from Germanic languages, it is also possible to increase translation performance under our multi-task learning framework which demonstrates the generalization of our model to multiple target languages.

Lang-Pair	En-Es	En-Fr	En-Nl	En-Pt
Single NMT	26.65	21.22	28.75	20.27
Multi Task	28.03	22.47	29.88	20.75
Delta	**+1.38**	**+1.25**	**+1.13**	**+0.48**

Table 3: Multi-task neural translation v.s. single model given large-scale corpus in all language pairs

We tested our selected model on the WMT 2013 dataset. Our results are shown in Table 5 where Multi-Full is the model with Experiment 1 setting and the model of Multi-Partial uses the same setting in Experiment 2. The English-French and English-Spanish translation performances are improved significantly compared with models trained separately on each language pair. Note that this result is not comparable with the result reported in (Bahdanau et al., 2014) as we use much less training corpus. We also compare our trained models with Moses. On the WMT 2013 data set, we utilize parallel corpora for Moses training without any extra resource such as large-scale monolingual corpus. From Table 5, it is shown that neural machine translation models have comparable BLEU scores with Moses. On the WMT 2013 test set, multi-task learning model outperforms both single model and Moses results significantly.

Lang-Pair	En-Es	En-Fr	En-Nl*	En-Pt*
Single NMT	26.65	21.22	26.59	18.26
Multi Task	28.29	21.89	27.85	19.32
Delta	**+1.64**	**+0.67**	**+1.26**	**+1.06**

Table 4: Multi-task neural translation v.s. single model with a small-scale training corpus on some language pairs. * means that the language pair is sub-sampled.

4.5 Model Analysis and Discussion

We try to make empirical analysis through learning curves and qualitative results to explain why multi-task learning framework works well in multiple-target machine translation problem.

From the learning process, we observed that the speed of model convergence under multi-task learning is faster than models trained separately especially when a model is trained for resource-poor language pairs. The detailed learning curves are shown in Figure 4. Here we study the learning curve for resource-poor language pairs, i.e. English-Dutch and En-Portuguese, for which only 15% of the bilingual data is sampled for training. The BLEU scores are evaluated on the Europarl common test set. From Figure 4, it can be seen that in the early stage of training, given the same amount of training data for each language pair, the translation performance of the multi-task learning model is improved more rapidly. And the multi-task models achieve better translation quality than separately trained models within three iterations of training. The reason of faster and better convergence in performance is that the encoder parameters are shared across different language pairs, which can make full use of all the source language training data across the language pairs and improve the source language

	Nmt Baseline	Nmt Multi-Full	Nmt Multi-Partial	Moses
En-Fr	23.89	26.02(**+2.13**)	25.01(**+1.12**)	23.83
En-Es	23.28	25.31(**+2.03**)	25.83(**+2.55**)	23.58

Table 5: Multi-task NMT v.s. single model v.s. moses on the WMT 2013 test set

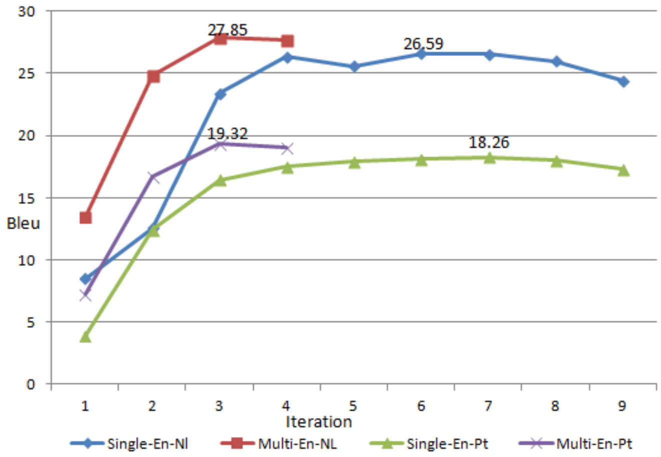

Figure 4: Faster and Better convergence in Multi-task Learning in multiple language translation

representation.

The sharing of encoder parameters is useful especially for the resource-poor language pairs. In the multi-task learning framework, the amount of the source language is not limited by the resource-poor language pairs and we are able to learn better representation for the source language. Thus the representation of the source language learned from the multi-task model is more stable, and can be viewed as a constraint that leverages translation performance of all language pairs. Therefore, the overfitting problem and the data scarcity problem can be alleviated for language pairs with only a few training data. In Table 6, we list the three nearest neighbors of some source words whose similarity is computed by using the cosine score of the embeddings both in the multi-task learning framework (from Experiment two) and in the single model (the resource-poor English-Portuguese model). Although the nearest neighbors of the high-frequent words such as numbers can be learned both in the multi-task model and the single model, the overall quality of the nearest neighbors learned by the resource-poor single model is much poorer compared with the multi-task model.

The multi-task learning framework also generates translations of higher quality. Some examples are shown in Table 7. The examples are from the WMT 2013 test set. The French and Spanish translations generated by the multi-task learning model and the single model are shown in the table.

MultiTask	**Nearest neighbors**
provide	deliver 0.78, providing 0.74, give 0.72
crime	terrorism 0.66, criminal 0.65, homelessness 0.65
regress	condense 0.74, mutate 0.71, evolve 0.70
six	eight 0.98,seven 0.96, 12 0.94
Single-Resource-Poor	**Nearest Neighbors**
provide	though 0.67,extending 0.56, parliamentarians 0.44
crime	care 0.75, remember 0.56, three 0.53
regress	committing 0.33, accuracy 0.30, longed-for 0.28
six	eight 0.87, three 0.69, thirteen 0.65

Table 6: Source language nearest-neighbor comparison between the multi-task model and the single model

5 Conclusion

In this paper, we investigate the problem of how to translate one source language into several different target languages within a unified translation model. Our proposed solution is based on the

English	Students, meanwhile, say the course is one of the most interesting around.
Reference-Fr	Les étudiants, pour leur part, assurent que le cours est l' un des plus intéressants.
Single-Fr	Les étudiants, entre-temps, disent entendu l' une des plus intéressantes.
Multi-Fr	Les étudiants, en attendant, disent qu' il est l' un des sujets les plus intéressants.
English	In addition, they limited the right of individuals and groups to provide assistance to voters wishing to register.
Reference-Fr	De plus, ils ont limité le droit de personnes et de groupes de fournir une assistance aux électeurs désirant s' inscrire.
Single-Fr	En outre, ils limitent le droit des particuliers et des groupes pour fournir l' assistance aux électeurs.
Multi-Fr	De plus, ils restreignent le droit des individus et des groupes à fournir une assistance aux électeurs qui souhaitent enregistrer.

Table 7: Translation of different target languages given the same input in our multi-task model.

recently proposed recurrent neural network based encoder-decoder framework. We train a unified neural machine translation model under the multi-task learning framework where the encoder is shared across different language pairs and each target language has a separate decoder. To the best of our knowledge, the problem of learning to translate from one source to multiple targets has seldom been studied. Experiments show that given large-scale parallel training data, the multi-task neural machine translation model is able to learn good predictive structures in translating multiple targets. Significant improvement can be observed from our experiments on the data sets publicly available. Moreover, our framework is able to address the data scarcity problem of some resource-poor language pairs by utilizing large-scale parallel training corpora of other language pairs to improve the translation quality. Our model is efficient and gets faster and better convergence for both resource-rich and resource-poor language pair under the multi-task learning.

In the future, we would like to extend our learning framework to more practical setting. For example, train a multi-task learning model with the same target language from different domains to improve multiple domain translation within one model. The correlation of different target languages will also be considered in the future work.

Acknowledgement

This paper is supported by the 973 program No. 2014CB340505. We would like to thank anonymous reviewers for their insightful comments.

References

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfellow, Arnaud Bergeron, Nicolas Bouchard, David Warde-Farley, and Yoshua Bengio. 2012. Theano: new features and speed improvements. *CoRR*, abs/1211.5590.

Léon Bottou. 1991. Stochastic gradient learning in neural networks. In *Proceedings of Neuro-Nîmes 91*, Nimes, France. EC2.

KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259.

Trevor Cohn and Mirella Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proc. ACL*, pages 728–735.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Lei Cui, Xilun Chen, Dongdong Zhang, Shujie Liu, Mu Li, and Ming Zhou. 2013. Multi-domain adaptation for SMT using multi-task learning. In *Proc. EMNLP*, pages 1055–1065.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard M. Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *Proc. ACL*, pages 1370–1380.

Jianfeng Gao, Xiaodong He, Wen-tau Yih, and Li Deng. 2014. Learning continuous phrase representations for translation modeling. In *Proc. ACL*, pages 699–709.

Jun Hatori, Takuya Matsuzaki, Yusuke Miyao, and Jun'ichi Tsujii. 2012. Incremental joint approach to word segmentation, POS tagging, and dependency parsing in chinese. In *Proc. ACL*, pages 1045–1053.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proc. EMNLP*, pages 1700–1709.

Philipp Koehn. 2004. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Machine Translation: From Real Users to Research, 6th Conference of the Association for Machine Translation in the Americas, AMTA 2004, Washington, DC, USA, September 28-October 2, 2004, Proceedings*, pages 115–124.

Zhenghua Li, Min Zhang, Wanxiang Che, Ting Liu, and Wenliang Chen. 2014. Joint optimization for chinese POS tagging and dependency parsing. *IEEE/ACM Transactions on Audio, Speech & Language Processing*, 22(1):274–286.

Shujie Liu, Nan Yang, Mu Li, and Ming Zhou. 2014. A recursive recurrent neural network for statistical machine translation. In *Proc. ACL*, pages 1491–1500.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proc. ACL*, ACL 2002, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rico Sennrich, Holger Schwenk, and Walid Aransa. 2013. A multi-domain translation model framework for statistical machine translation. In *Proc. ACL*, pages 832–840.

Martin Sundermeyer, Tamer Alkhouli, Joern Wuebker, and Hermann Ney. 2014. Translation modeling with bidirectional recurrent neural networks. In *Proc. EMNLP*, pages 14–25.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. In *Proc. ACL*, pages 165–181.

Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701.