Multi-Task Neural Model for Agglutinative Language Translation

Yirong Pan[1,2,3], **Xiao Li**[1,2,3], **Yating Yang**[1,2,3], and **Rui Dong**[1,2,3]
[1] Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, China
[2] University of Chinese Academy of Sciences, China
[3] Xinjiang Laboratory of Minority Speech and Language Information Processing, China
panyirong15@mails.ucas.ac.cn
{xiaoli, yangyt, dongrui}@ms.xjb.ac.cn

Abstract

Neural machine translation (NMT) has achieved impressive performance recently by using large-scale parallel corpora. However, it struggles in the low-resource and morphologically-rich scenarios of agglutinative language translation task. Inspired by the finding that monolingual data can greatly improve the NMT performance, we propose a multi-task neural model that jointly learns to perform bi-directional translation and agglutinative language stemming. Our approach employs the shared encoder and decoder to train a single model without changing the standard NMT architecture but instead adding a token before each source-side sentence to specify the desired target outputs of the two different tasks. Experimental results on Turkish-English and Uyghur-Chinese show that our proposed approach can significantly improve the translation performance on agglutinative languages by using a small amount of monolingual data.

1 Introduction

Neural machine translation (NMT) has achieved impressive performance on many high-resource machine translation tasks (Bahdanau et al., 2015; Luong et al., 2015a; Vaswani et al., 2017). The standard NMT model uses the encoder to map the source sentence to a continuous representation vector, and then it feeds the resulting vector to the decoder to produce the target sentence.

However, the NMT model still suffers from the low-resource and morphologically-rich scenarios of agglutinative language translation tasks, such as Turkish-English and Uyghur-Chinese. Both Turkish and Uyghur are agglutinative languages with complex morphology. The morpheme structure of the word can be denoted as: ***prefix1*** + … + ***prefixN*** + ***stem*** + ***suffix1*** + … + ***suffixN*** (Ablimit et al., 2010). Since the suffixes have many inflected and morphological variants, the vocabulary size of an agglutinative language is considerable even in small-scale training data. Moreover, many words have different morphemes and meanings in different context, which leads to inaccurate translation results.

Recently, researchers show their great interest in utilizing monolingual data to further improve the NMT model performance (Cheng et al., 2016; Ramachandran et al., 2017; Currey et al., 2017). Sennrich et al. (2016) pair the target-side monolingual data with automatic back-translation as additional training data to train the NMT model. Zhang and Zong (2016) use the source-side monolingual data and employ the multi-task learning framework for translation and source sentence reordering. Domhan and Hieber (2017) modify the decoder to enable multi-task learning for translation and language modeling. However, the above works mainly focus on boosting the translation fluency, and lack the consideration of morphological and linguistic knowledge.

Stemming is a morphological analysis method, which is widely used for information retrieval tasks (Kishida, 2005). By removing the suffixes in the word, stemming allows the variants of the same word to share representations and reduces data sparseness. We consider that stemming can lead to better generalization on agglutinative languages, which helps NMT to capture the in-depth semantic information. Thus we use stemming as an auxiliary task for agglutinative language translation.

In this paper, we investigate a method to exploit the monolingual data of the agglutinative language to enhance the representation ability of the encoder. This is achieved by training a multi-task neural model to jointly perform bi-directional translation and agglutinative language stemming, which utilizes the shared encoder and decoder. We treat stemming as a sequence generation task.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 103–110

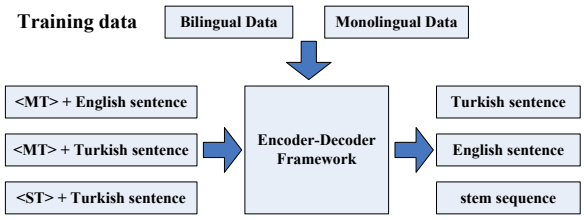

Figure 1: The architecture of the multi-task neural model that jointly learns to perform bi-directional translation between Turkish and English, and stemming for Turkish sentence.

2 Related Work

Multi-task learning (MTL) aims to improve the generalization performance of a main task by using the other related tasks, which has been successfully applied to various research fields ranging from language (Liu et al., 2015; Luong et al., 2015a), vision (Yim et al., 2015; Misra et al., 2016), and speech (Chen and Mak, 2015; Kim et al., 2016). Many natural language processing (NLP) tasks have been chosen as auxiliary task to deal with the increasingly complex tasks. Luong et al. (2015b) employ a small amount of data of syntactic parsing and image caption for English-German translation. Hashimoto et al. (2017) present a joint MTL model to handle the tasks of part-of-speech (POS) tagging, dependency parsing, semantic relatedness, and textual entailment for English. Kiperwasser and Ballesteros (2018) utilize the POS tagging and dependency parsing for English-German machine translation. To the best of our knowledge, we are the first to incorporate stemming task into MTL framework to further improve the translation performance on agglutinative languages.

Recently, several works have combined the MTL method with sequence-to-sequence NMT model for machine translation tasks. Dong et al. (2015) follow a ***one-to-many*** setting that utilizes a shared encoder for all the source languages with respective attention mechanisms and multiple decoders for the different target languages. Luong et al. (2015b) follow a ***many-to-many*** setting that uses multiple encoders and decoders with two separate unsupervised objective functions. Zoph and Knight (2016) follow a ***many-to-one*** setting that employs multiple encoders for all the source languages and one decoder for the desired target language. Johnson et al. (2017) propose a more simple method in ***one-to-one*** setting, which trains a single NMT model with the shared encoder and decoder in order to enable multilingual translation. The method requires no changes to the standard NMT architecture but instead requires adding a token at the beginning of each source sentence to specify the desired target sentence. Inspired by their work, we employ the standard NMT model with one encoder and one decoder for parameter sharing and model generalization. In addition, we build a joint vocabulary on the concatenation of the source-side and target-side words.

Several works on morphologically-rich NMT have focused on using morphological analysis to pre-process the training data (Luong et al., 2016; Huck et al., 2017; Tawfik et al., 2019). Gulcehre et al. (2015) segment each Turkish sentence into a sequence of morpheme units and remove any non-surface morphemes for Turkish-English translation. Ataman et al. (2017) propose a vocabulary reduction method that considers the morphological properties of the agglutinative language, which is based on the unsupervised morphology learning. This work takes inspiration from our previously proposed segmentation method (Pan et al., 2020) that segments the word into a sequence of sub-word units with morpheme structure, which can effectively reduce language complexity.

3 Multi-Task Neural Model

3.1 Overview

We propose a multi-task neural model for machine translation from and into a low-resource and morphologically-rich agglutinative language. We train the model to jointly learn to perform both the bi-directional translation task and the stemming task on an agglutinative language by using the standard NMT framework. Moreover, we add an artificial token before each source sentence to specify the desired target outputs for different tasks. The architecture of the proposed model is shown in Figure 1. We take the Turkish-English translation task as example. The "**<MT>**" token denotes the bilingual translation task and the "**<ST>**" token denotes the stemming task on Turkish sentence.

3.2 Neural Machine Translation (NMT)

Our proposed multi-task neural model on using the source-side monolingual data for agglutinative language translation task can be applied in any NMT structures with encoder-decoder framework. In this work, we follow the NMT model proposed by Vaswani et al. (2017), which is implemented as Transformer. We will briefly summarize it here.

Task	Data	# Sent	# Src	# Trg
Tr-En	train	355,251	6,356,767	8,021,161
	valid	2,455	37,153	52,125
	test	4,962	69,006	96,291
Uy-Ch	train	333,097	6,026,953	5,748,298
	valid	700	17,821	17,085
	test	1,000	20,580	18,179

Table 1: The statistics of the training, validation, and test datasets on Turkish-English and Uyghur-Chinese machine translation tasks. The "**# Src**" denotes the number of the source tokens, and the "**# Trg**" denotes the numbers of the target tokens.

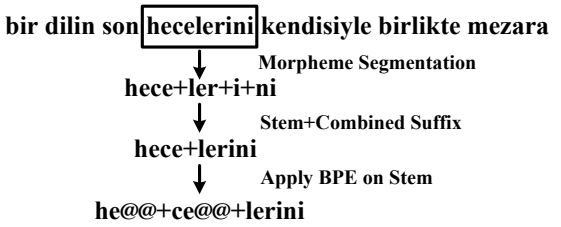

Figure 2: The example of morphological segmentation method for the word in Turkish.

Firstly, the Transformer model maps the source sequence $\boldsymbol{x} = (x_1, \ldots, x_m)$ and the target sentence $\boldsymbol{y} = (y_1, \ldots, y_n)$ into a word embedding matrix, respectively. Secondly, in order to make use of the word order in the sequence, the above word embedding matrices sum with their positional encoding matrices to generate the source-side and target-side positional embedding matrices. The encoder is composed of a stack of *N* identical layers. Each layer has two sub-layers consisting of the multi-head self-attention and the fully connected feed-forward network, which maps the source-side positional embedding matrix into a representation vector.

The decoder is also composed of a stack of *N* identical layers. Each layer has three sub-layers: the multi-head self-attention, the multi-head attention, and the fully connected feed-forward network. The multi-head attention attends to the outputs of the encoder and decoder to generate a context vector. The feed-forward network followed by a linear layer maps the context vector into a vector with the original space dimension. Finally, the *softmax* function is applied on the vector to predict the target word sequence.

4 Experiment

4.1 Dataset

The statistics of the training, validation, and test datasets on Turkish-English and Uyghur-Chinese machine translation tasks are shown in Table 1.

For the Turkish-English machine translation, following (Sennrich et al., 2015a), we use the WIT corpus (Cettolo et al., 2012) and the SETimes corpus (Tyers and Alperen, 2010) as the training dataset, merge the dev2010 and tst2010 as the validation dataset, and use tst2011, tst2012, tst2013, tst2014 from the IWSLT as the test datasets. We also use the talks data from the IWSLT evaluation campaign[1] in 2018 and the news data from News Crawl corpora[2] in 2017 as external monolingual data for the stemming task on Turkish sentences.

For the Uyghur-Chinese machine translation, we use the news data from the China Workshop on Machine Translation in 2017 (CWMT2017) as the training dataset and validation dataset, use the news data from CWMT2015 as the test dataset. Each Uyghur sentence has four Chinese reference sentences. Moreover, we use the news data from the Tianshan website[3] as external monolingual data for the stemming task on Uyghur sentences.

4.2 Data Preprocessing

We normalize and tokenize the experimental data. We utilize the jieba toolkit[4] to segment the Chinese sentences, we utilize the Zemberek toolkit[5] with morphological disambiguation (Sak et al., 2007) and the morphological analysis tool (Tursun et al., 2016) to annotate the morpheme structure of the words in Turkish and Uyghur, respectively.

We use our previously proposed morphological segmentation method (Pan et al., 2020), which segments the word into smaller subword units with morpheme structure. Since Turkish and Uyghur only have a few prefixes, we combine the prefixes with stem into the stem unit. As shown in Figure 2, the morpheme structure of the Turkish word "***hecelerini***" (syllables) is: ***hece*** + ***lerini***. Then the byte pair encoding (BPE) technique (Sennrich et al., 2015b) is applied on the stem unit "***hece***" to segment it into "***he@@***" and "***ce@@***". Thus the Turkish word is segmented into a sequence of sub-word units: ***he@@*** + ***ce@@*** + ***lerini***.

[1] https://wit3.fbk.eu/archive/2018-01/additional_TED_xml/
[2] http://data.statmt.org/wmt18/translation-task/
[3] http://uy.ts.cn/
[4] https://github.com/fxsjy/jieba
[5] https://github.com/ahmetaa/zemberek-nlp

Task	Training Sentence Samples
En-Tr Translation	**<MT>** We go through initiation rit@@ es.
	Başla@@ ma ritüel@@ lerini yaş@@ ıyoruz.
Tr-En Translation	**<MT>** Başla@@ ma ritüel@@ lerini yaş@@ ıyoruz.
	We go through initiation rit@@ es.
Turkish Stemming	**<ST>** Başla@@ ma ritüel@@ lerini yaş@@ ıyoruz.
	Başla@@ ritüel@@ yaş@@

Table 2: The training sentence samples for multi-task neural model on Turkish-English machine translation task. We add "**<MT>**" and "**<ST>**" before each source sentence to specify the desired target outputs for different tasks.

Lang	Method	# Merge	Vocab	Avg.Len
Tr	Morph	15K	36,468	28
Tr	BPE	36K	36,040	22
En	BPE	32K	31,306	25
Uy	Morph	10K	38,164	28
Uy	BPE	38K	38,292	21
Ch	BPE	32K	40,835	19

Table 3: The detailed statistics of using different word segmentation methods on Turkish, English, Uyghur, and Chinese.

In this paper, we utilize the above morphological segmentation method for our experiments by applying BPE on the stem units with 15K merge operations for the Turkish words and 10K merge operations for the Uyghur words. The standard NMT model trained on this experimental data is denoted as "**baseline NMT model**". Moreover, we employ BPE to segment the words in English and Chinese by learning separate vocabulary with 32K merge operations. Table 2 shows the training sentence samples for multi-task neural model on Turkish-English machine translation task.

In addition, to certify the effectiveness of the morphological segmentation method, we employ the pure BPE to segment the words in Turkish and Uyghur by learning a separate vocabulary with 36K and 38K merge operations, respectively. The standard NMT model trained on this experimental data is denoted as "**general NMT model**". Table 3 shows the detailed statistics of using different word segmentation methods on Turkish, English, Uyghur, and Chinese. The "**Vocab**" token denotes the vocabulary size after data preprocessing. The "**Avg.Len**" token denotes the average sentence length.

4.3 Training and Evaluation Details

We employ the Transformer model implemented in the *Sockeye* toolkit (Hieber et al., 2017). The number of layer in both the encoder and decoder is set to N=6, the number of head is set to 8, and the number of hidden unit in the feed-forward network is set to 1024. We use an embedding size of both the source and target words of 512 dimension, and use a batch size of 128 sentences. The maximum sentence length is set to 100 tokens with 0.1 label smoothing. We apply layer normalization and add dropout to the embedding and transformer layers with 0.1 probability. Moreover, we use the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 0.0002, and save the checkpoint every 1500 updates.

Model training process stops after 8 checkpoints without improvements on the validation perplexity. Following Niu et al. (2018a), we select the 4 best checkpoint based on the validation perplexity values and combine them in a linear ensemble for decoding. Decoding is performed by using beam search with a beam size of 5. We evaluate the machine translation performance by using the case-sensitive BLEU score (Papineni et al., 2002) with standard tokenization.

4.4 Neural Translation Models

In this paper, we select 4 neural translation models for comparison. More details about the models are shown below:

General NMT Model: The standard NMT model trained on the experimental data segmented by BPE.

Baseline NMT Model: The standard NMT model trained on the experimental data segmented by morphological segmentation. The following models also use this word segmentation method.

Bi-Directional NMT Model: Following Niu et al. (2018b), we train a single NMT model to perform bi-directional machine translation. We concatenate the bilingual parallel sentences in both directions. Since the source and target sentences come from the same language pairs, we share the source and target vocabulary, and tie their word embedding during model training.

Multi-Task Neural Model: We simply use the monolingual data of the agglutinative language from the bilingual parallel sentences. We use a joint vocabulary, tie the word embedding as well as the output layer's weight matrix.

Task	Model	tst11	tst12	tst13	tst14
Tr-En	general	25.92	26.55	27.34	**26.35**
	baseline	**26.48**	**27.02**	**27.91**	26.33
En-Tr	general	13.73	14.68	13.84	14.65
	baseline	**14.85**	**15.93**	**15.45**	**15.93**

Table 4: The BLEU scores of the general NMT model and baseline NMT model on the machine translation task between Turkish and English.

Task	Model	tst11	tst12	tst13	tst14
Tr-En	baseline	26.48	27.02	27.91	26.33
	bi-directional	26.21	27.17	28.68	26.90
	multi-task	**26.82**	**27.96**	**29.16**	**27.98**
En-Tr	baseline	14.85	15.93	15.45	15.93
	bi-directional	15.08	16.20	16.25	**16.56**
	multi-task	**15.65**	**17.10**	**16.35**	16.41

Table 5: The BLEU scores of the baseline NMT model, bi-directional NMT model, and multi-task neural model on the machine translation task between Turkish and English.

5 Results and Discussion

Table 4 shows the BLEU scores of the general NMT model and baseline NMT model on machine translation task. We can observe that the baseline NMT model is comparable to the general NMT model, and it achieves the highest BLEU scores on almost all the test datasets in both directions, which indicates that the NMT baseline based on our proposed segmentation method is competitive.

5.1 Using Original Monolingual Data

Table 5 shows the BLEU scores of the baseline NMT model, bi-directional NMT model, and multi-task neural model on the machine translation task between Turkish and English. The table shows that the multi-task neural model outperforms both the baseline NMT model and bi-directional NMT model, and it achieves the highest BLEU scores on almost all the test datasets in both directions, which suggests that the multi-task neural model is capable of improving the bi-directional translation quality on agglutinative languages. The main reason is that compared with the bi-directional NMT model, our proposed multi-task neural model additionally employs the stemming task for the agglutinative language, which is effective for the NMT model to learn both the source-side semantic information and the target-side language modeling.

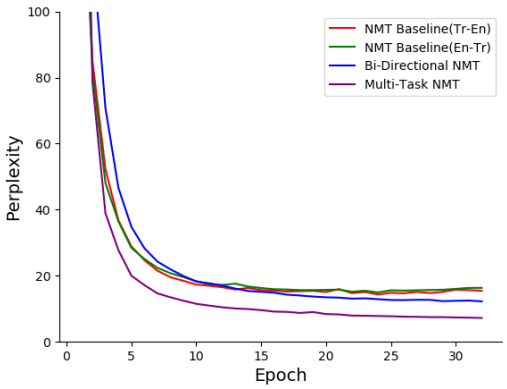

Figure 3: The function of epochs (x-axis) and perplexity (y-axis) values on the validation dataset in different neural translation models for the translation task.

Translation Examples	
source	üniversite hayatı taklit ediyordu.
reference	College was imitating life.
baseline	It was emulating a university life.
bi-directional	The university was emulating its lives.
multi-task	The university was imitating life.

Table 6: A translation example for the different NMT models on Turkish-English.

The function of epochs and perplexity values on the validation dataset in different neural translation models are shown in Figure 3. We can see that the perplexity values are consistently lower on the multi-task neural model, and it converges rapidly.

Table 6 shows a translation example for the different models on Turkish-English. We can see that the translation result of the multi-task neural model is more accurate. The Turkish word "**taklit**" means "imitate" in English, both the baseline NMT and bi-directional NMT translate it into a synonym "emulate". However, they are not able to express the meaning of the sentence correctly. The main reason is that the auxiliary task of stemming forces the proposed model to focus more strongly on the core meaning of each word (or stem), therefore helping the model make the correct lexical choices and capture the in-depth semantic information.

5.2 Using External Monolingual Data

Moreover, we evaluate the multi-task neural model on using external monolingual data for Turkish stemming task. We employ the parallel sentences and the monolingual data in a 1-1 ratio, and shuffle them randomly before each training epoch. More details about the data are shown below:

Task	Data	tst11	tst12	tst13	tst14
Tr-En	original	**26.82**	27.96	29.16	27.98
	talks	26.55	27.94	29.13	**28.02**
	news	26.47	**28.18**	28.89	27.40
	mixed	26.60	27.93	**29.58**	27.32
En-Tr	original	15.65	17.10	16.35	16.41
	talks	15.57	16.97	16.22	**16.91**
	news	15.67	17.19	16.26	16.69
	mixed	**15.96**	**17.35**	**16.55**	16.89

Table 7: The BLEU scores of the multi-task neural model on using external monolingual data of talks data, news data, and mixed data.

Task	Model	BLEU
Uy-Ch	general NMT model	35.12
	baseline NMT model	35.46
	multi-task neural model with external monolingual data	**36.47**
Ch-Uy	general NMT model	21.00
	baseline NMT model	21.57
	multi-task neural model with external monolingual data	**23.02**

Table 8: The BLEU scores of the general NMT model, baseline NMT model, and the multi-task neural model with external monolingual data on Uyghur-Chinese and Chinese-Uyghur machine translation tasks.

Original Data: The monolingual data comes from the original bilingual parallel sentences.
Talks Data: The monolingual data contains talks.
News Data: The monolingual data contains news.
Talks and News Mixed Data: The monolingual data contains talks and news in a 3:4 ratio as the same with the original bilingual parallel sentences.

Table 7 shows the BLEU scores of the proposed multi-task neural model on using different external monolingual data. We can see that there is no obvious difference on Turkish-English translation performance by using different monolingual data, whether the data is in-domain or out-of-domain to the test dataset. However, for the English-Turkish machine translation task, which can be seen as agglutinative language generation task, using the mixed data of talks and news achieves further improvements of the BLEU scores on almost all the test datasets. The main reason is that the proposed multi-task neural model incorporates many morphological and linguistic information of Turkish rather than that of English, which mainly pays attention to the source-side representation ability on agglutinative languages rather than the target-side language modeling.

We also evaluate the translation performance of the general NMT model, baseline NMT model, and multi-task neural model with external news data on the machine translation task between Uyghur and Chinese. The experimental results are shown in Table 8. The results indicate that the multi-task neural model achieves the highest BLEU scores on the test dataset by utilizing external monolingual data for the stemming task on Uyghur sentences.

6 Conclusions

In this paper, we propose a multi-task neural model for translation task from and into a low-resource and morphologically-rich agglutinative language. The model jointly learns to perform bi-directional translation and agglutinative language stemming by utilizing the shared encoder and decoder under standard NMT framework. Extensive experimental results show that the proposed model is beneficial for the agglutinative language machine translation, and only a small amount of the agglutinative data can improve the translation performance in both directions. Moreover, the proposed approach with external monolingual data is more useful for translating into the agglutinative language, which achieves an improvement of *+1.42* BLEU points for translation from English into Turkish and *+1.45* BLEU points from Chinese into Uyghur.

In future work, we plan to utilize other word segmentation methods for model training. We also plan to combine the proposed multi-task neural model with back-translation method to enhance the ability of the NMT model on target-side language modeling.

Acknowledgements

We are very grateful to the mentor of this paper for her meaningful feedback. Thanks three anonymous reviewers for their insightful comments and practical suggestions. This work is supported by the High-Level Talents Introduction Project of Xinjiang under Grant No.Y839031201, the National Natural Science Foundation of China under Grant No.U1703133, the National Natural Science Foundation of Xinjiang under Grant No.2019BL-0006, the Open Project of Xinjiang Key Laboratory under Grant No.2018D04018, and the Youth Innovation Promotion Association of the Chinese Academy of Sciences under Grant No.2017472.

References

Mijit Ablimit, Graham Neubig, Masato Mimura, Shinsuke Mori, Tatsuya Kawahara, and Askar Hamdulla. 2010. Uyghur morpheme-based language models and ASR. In *IEEE International Conference on Signal Processing*.

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically motivated vocabulary reduction for neural machine translation from Turkish to English. *Journal of the Prague Bulletin of Mathematical Linguistics*, 108:331-342.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473.*

Mauro Cettolo, Christian Girardi and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation.*

Dongpeng Chen and Brian Kan-Wing Mak. 2015. Multitask learning of deep neural networks for low-resource speech recognition. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing.*

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-Supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of ACL*.

Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078.*

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*.

Tobias Domhan and Felix Hieber. 2017. Using target-side monolingual data for neural machine translation through multi-task learning. In *Proceedings of EMNLP.*

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of ACL.*

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, et al. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535v2*.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2017. A joint many-task model: Growing a neural network for multiple NLP tasks. In *Proceedings of EMNLP.*

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *arXiv preprint arXiv:1712.05690.*

Matthias Huck, Simon Riess, and Alexander Fraser. 2017. Target-side word segmentation strategies for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In *Proceedings of ACL.*

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. In *Proceedings of ACL.*

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2016. Joint CTC-Attention based end-to-end speech recognition using multi-task learning. *arXiv preprint arXiv:1609.06773.*

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR.*

Eliyahu Kiperwasser, Miguel Ballesteros. 2018. Scheduled multi-task learning: From syntax to translation. In *Transactions of the Association for Computational Linguistics.*

Kazuaki Kishida. 2005. Technical issues of cross-language information retrieval: A review. *Journal of the Information Processing and Management*, 41(3):433-455. https://doi.org/10.1016/j.ipm.2004.06.007.

Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *Proceedings of NAACL.*

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015a. Effective approaches to Attention-based neural machine translation. In *Proceedings of EMNLP.*

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015b. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114.*

Minh-Thang Luong, Christopher D. Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of ACL.*

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. 2016. Cross-Stitch networks for multi-task learning. In *Proceedings of CVPR.*

Xing Niu, Sudha Rao, and Marine Carpuat. 2018a. Multi-task neural models for translating between styles within and across languages. In *Proceedings of COLING*.

Xing Niu, Michael Denkowski, and Marine Carpuat. 2018b. Bi-Directional neural machine translation with synthetic parallel data. *arXiv preprint arXiv: 1805.11213.*

Prajit Ramachandran, Peter Liu, and Quoc Le. 2017. Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on EMNLP*.

Yirong Pan, Xiao Li, Yating Yang, and Rui Dong. 2020. Morphological word segmentation on agglutinative languages for neural machine translation. *arXiv preprint arXiv: 2001.01589.*

Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL.*

Ha¸Sim Sak, Tunga Güngör, and Murat Saraçlar. 2007. Morphological disambiguation of Turkish text with perceptron algorithm. In *International Conference on Intelligent Text Processing and Computational Linguistics.*

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015a. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709.*

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. 2015b. *arXiv preprint arXiv: 1508.07909.*

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of NAACL.*

Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of NIPS.*

Ahmed Tawfik, Mahitab Emam, Khaled Essam, Robert Nabil, and Hany Hassan. 2019. Morphology-aware word-segmentation in dialectal Arabic adaptation of neural machine translation. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*.

Eziz Tursun, Debasis Ganguly, Turghun Osman, Yating Yang, Ghalip Abdukerim, Junlin Zhou, and Qun Liu. 2016. A semi-supervised tag-transition-based Markovian model for Uyghur morphology analysis. In *ACM Transactions on Asian and Low-Resource Language Information Processing.*

Francis M. Tyers and Murat Alperen. 2010. South-East European Times: A parallel corpus of the Balkan languages, In *Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages.*

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems.*

Junho Yim, Heechul Jung, ByungIn Yoo, Changkyu Choi, Dusik Park, and Junmo Kim. 2015. Rotating your face using multi-task deep neural network. In *Proceedings of CVPR.*

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of EMNLP.*

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of NAACL.*