Exploiting Linguistic Knowledge for Low-Resource Neural Machine 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
[2] University of Chinese Academy of Sciences
[3] Xinjiang Laboratory of Minority Speech and Language Information Processing
* Correspondence: xiaoli@ms.xjb.ac.cn; yangyt@ms.xjb.ac.cn

Abstract: Exploiting the linguistic knowledge of the source language for neural machine translation (NMT) has recently achieved impressive performance on many large-scale language pairs. However, since the Turkish→English machine translation task is low-resource and the source-side Turkish is morphologically-rich, there are limited resources of bilingual corpora and linguistic information available to further improve the NMT performance. Focusing on the above issues, we propose a multi-source NMT approach that models the word feature in parallel to external linguistic features by using two separate encoders to explicitly incorporate linguistic knowledge into the NMT model. We extend the word embedding layer of the knowledge-based encoder to accommodate for each word's linguistic annotations in the context. Moreover, we share all parameters across encoders to enhance the representation ability of the NMT model on the source language. Experimental results show that our proposed approach achieves substantial improvements of up to 2.4 and 1.1 BLEU scores in Turkish→English and English→Turkish machine translation tasks, respectively, which points to a promising way to utilize the source-side linguistic knowledge for the low-resource NMT.

Keywords: linguistic knowledge; source language; neural machine translation (NMT); low-resource; multi-source NMT

1. Introduction

Neural machine translation (NMT) model is widely used for machine translation tasks [1-5], which directly learns the translation relationship between distinct languages from bilingual corpora. The NMT model employs the encoder to map the source sentence to a continuous representation vector, then it feeds the resulting vector to the decoder to generate the desired target sentence [1]. Recently, by using advanced neural mechanisms, such as GRU [2] and attention [3], the NMT model surpasses the previously dominant statistical machine translation (SMT) model [6], and it achieves the state-of-the-art performance on many natural language processing (NLP) tasks [7-9].

Despite their current success, the existing NMT models reply on large-scale bilingual sentences, and they cannot provide an effective mechanism to utilize the linguistic knowledge of the language. Recently, researchers attempt to explicitly incorporate the source-side linguistic information into the NMT model. Currey and Heafield introduce a multi-source technique that exploits the syntax structure of the source sentence by employing separate encoders, one encoder for the source word sequence and the other encoder for the linearized source parse tree [10].

However, the NMT models still suffer from the low-resource and morphologically-rich scenarios. For Turkish→English machine translation task, the main problem is that the source-side Turkish is a morphologically-rich language with derivational morphology [11]. The vocabulary is very large even in the small-scale training corpus, which leads to many inaccurate translation results [12-13]. Due to the complex morphology, no high-quality syntactic analysis tool is available to provide the linguistic annotations at the sentence-level for Turkish sentences, such as named entity chunking, dependency label, and parsed tree. Thus, it is valuable to fully utilize the linguistic annotations at the word-level to improve the low-resource NMT performance. To address the above issue, Sennrich and Haddow modify the architecture of the encoder to allow for linguistic input features [14]. They generalize the word embedding layer of the encoder to encode each word and its corresponding functional labels. Nevertheless, since their model reduces the word embedding size to adapt to additional linguistic features, it cannot balance the contribution from the word sequence and the linguistic information. Moreover, their model adds the lemma feature as linguistic input while using a fixed-size vocabulary, it suffers from the out-of-vocabulary (OOV) problem [15].

Focusing on the above issues, we propose a multi-source NMT approach for the low-resource NMT to explicitly utilize the source-side linguistic knowledge. The approach models the source word sequence in parallel to linguistic features by using two separate encoders with parameter sharing. The source sequence is encoded by the word-based encoder, the linguistic features is encoded by the knowledge-based encoder. We extend the word embedding layer of the knowledge-based encoder to accommodate for each word's linguistic annotations in the context. Then we combine the hidden states and cell states from each encoder to pass on to the decoder. Experimental results show that our proposed approach achieves substantial improvements in Turkish→English and English→Turkish machine translation tasks. In particular, extensive experiments show that our approach is capable of better utilizing the source-side linguistic knowledge and effective integrating them together.

2. Multi-Source NMT Model

We follow the NMT architecture proposed by Vaswani *et al.* [5], and we will briefly summarize in this section. The NMT model is implemented as a Transformer with encoder-decoder framework. The encoder maps the source sequence $\boldsymbol{x} = (x_1, \dots, x_m)$ to a representation vector $\boldsymbol{z} = (z_1, \dots, z_m)$. Given z, the decoder produces the target sequence $\boldsymbol{y} = (y_1, \dots, y_n)$ based on all the previously generated symbols. Both the encoder and decoder are composed of a stack of N identical layers. Each layer has two sub-layers of a multi-head self-attention and a fully connected feed-forward network. The decoder additionally inserts a third sub-layer, which performs multi-head attention over the output of the encoder stack.

We denote the encoder in the NMT model as word-based encoder, which is employed to encode the source word sequence. Figure 1 shows our proposed multi-source NMT model. We use two separate encoders to model the word feature and the linguistic features, respectively.

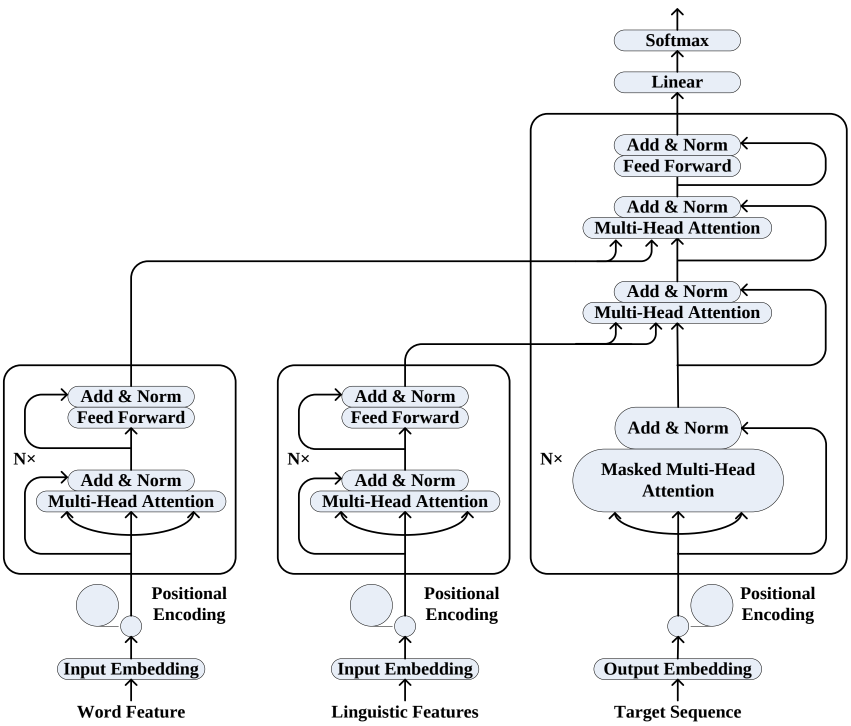

Figure 1. The architecture of our proposed multi-source NMT model.

2.1. Knowledge-based Encoder

Inspired by the NMT model proposed by Sennrich and Haddow [14], we employ additional knowledge-based encoder to incorporate external linguistic information into the NMT model. More specifically, we extend the input embedding layer of the knowledge-based encoder to accommodate for each word's linguistic annotations in the context. Figure 2 shows the input embedding layer.

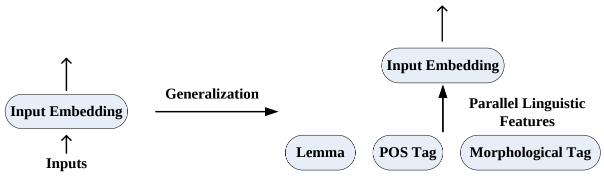

Figure 2. The input embedding layer of the knowledge-based encoder.

Given $|F|$ linguistic annotation sequences $\boldsymbol{k_1} = (k_{11}, \dots, k_{1m})$, $\boldsymbol{k_2} = (k_{21}, \dots, k_{2m})$, …, and $\boldsymbol{k}_{|F|} = (k_{|F|1}, \dots, k_{|F|m})$ corresponding to the source sequence $\boldsymbol{x} = (x_1, \dots, x_m)$, the knowledge-based encoder maps $\boldsymbol{k}_1$, $\boldsymbol{k}_2$, …, $\boldsymbol{k}_{|F|}$ to a continuous representation vector $\boldsymbol{p} = (p_1, \dots, p_m)$, where p_i is computed by:

$$p_i = tanh\left(\boldsymbol{W}\left(\bigcup_{t=1}^{|F|} \boldsymbol{E_t} k_{ti}\right) + \boldsymbol{U} p_{i-1}\right) \tag{1}$$

where $\bigcup$ is the vector concatenation operator, $\boldsymbol{E}_t$ is the feature embedding matrice, both $\boldsymbol{W}$ and $\boldsymbol{U}$ are weight matrices.

2.2. Combination Method

Both the hidden states and cell states from each encoder are combined to pass on to the decoder. The hidden states of h_1 and h_2 are concatenated. Then, a linear transformation and a tanh nonlinear transformation are applied on the resulting vector:

$$h = tanh\,(\boldsymbol{W_c}[h_1; h_2]) \tag{2}$$

where $\boldsymbol{W}_c$ is the weight matrice. The new cell state is the sum of the cell states of c_1 and c_2:

$$c = c_1 + c_2 \tag{3}$$

3. Linguistic Knowledge for Turkish and English

We utilize three linguistic features for experiments. The first is lemma, which is widely used for information retrieval tasks. Lemmatization can make better generalization by allowing the inflected and morphological variants of the same word to share representations. The second is part-of-speech (POS) tag, which provides the syntactic role for each word in the sentence. It is helpful in extracting information and reducing data ambiguity. The third is morphological tag. Since different word types have distinct sets of morphological features, morphology analysis can reduce data sparseness.

For Turkish→English task, the java toolkit Zemberek[1] with morphological disambiguation [16] is utilized to annotate the Turkish word with lemma, POS tag and morphological feature. Each word's morphological features are concatenated as its morphological tag. For English→Turkish task, the python toolkit NLTK[2] is utilized to annotate the English word with lemma and POS tag.

We use BPE technique [17] to segment the word and lemma in Turkish and English into subword units, and we add "@@" behind each non-final subword. We annotate the lemma annotation sequence by coping the other linguistic annotations' feature value to each lemma's subword units. All the linguistic annotation sequences have the same length. The training examples for Turkish→English and English→Turkish machine translation tasks are shown in Table 1 and Table 2, respectively.

[1] https://github.com/ahmetaa/zemberek-nlp
[2] https://github.com/nltk

Table 1. The training examples for Turkish→English machine translation task.

Turkish Word Sequence	**Ve bunlar sinek@@ kap@@ an an@@ em@@ onlar.**						
Lemma Sequence	ve	bu	sinek@@	kapan	ane@@	mon	.
POS Tag Sequence	Conj	Pronoun	Noun	Noun	Noun	Noun	Punc
Morphological Tag Sequence	<null>	A3pl	<null>	<null>	A3pl	A3pl	<null>

Table 2. The training examples for English→Turkish machine translation task.

English Word Sequence	**And these are fly@@ trap an@@ em@@ ones.**							
Lemma Sequence	and	these	be	fly@@	trap	ane@@	mone	.
POS Tag Sequence	CC	DT	VBP	JJ	JJ	NNS	NNS	.

4. Experiment

4.1. Data Preparation

Following Sennrich *et al.* [18], we merge the WIT corpus [19] that consists of TED talks and the SETimes corpus [20] that consists of news as training corpus, merge dev2010 and tst2010 as validation corpus, and use tst2011, tst2012, tst2013, tst2014 as test corpus. The training and validation corpus statistics in experiments are shown in Table 3.

Table 3. The training and validation corpus statistics in experiments.

Dataset	# Sentences	# Turkish Tokens	# English Tokens
Training	355,251	6,712,018	8,376,414
Validation	2,455	39,272	54,061

4.2. Model Parameter

We implemented our proposed multi-source NMT model by using the OpenNMT-tf[3] toolkit. Both the encoder and decoder have N=6 layers. The number of hidden units is 512. The number of heads for self-attention is 8. Both the source and target word embedding size are 512, and the number of hidden units in feed-forward layers is 1024. The batch size is 48 sentences. The maximum sentence length is 100. The label smoothing is 0.1. The dropout rate in Transformer is 0.1. The length penalty is 0.6, and the clip gradient is 5.0 [21]. The parameters are uniformly initialized in [-0.1, 0.1]. We train the model for 120,000 steps by using the Adam optimizer [22] with the learning rate of 0.0002, and we report the result of averaging the 5 last saved checkpoints (saved every 5,000 steps). Decoding is performed by using the beam search with the beam size of 5.

We normalize and tokenize the Turkish and English sentences, and we use BPE to segment the word and lemma by learning separate vocabulary with 32K merge operations. Moreover, we report the case-sensitive BLEU [23] score and the ChrF3 [24] score to evaluate the translation performance. The vocabulary size and embedding size of Turkish and English are shown in Table 4. We keep the total embedding size of the linguistic annotations fixed to 512 in experiments.

Table 4. The vocabulary size and embedding size of Turkish and English.

Encoder	Input Feature	Vocabulary Size		Embedding Size	
		Turkish	English	Turkish	English
Word-based Encoder	Word	32,064	31,306	512	512
Knowledge-based Encoder	Lemma	30,637	30,646	352	384
	POS Tag	14	45	64	128
	Morphological Tag	9,176	-	96	-

[3] https://github.com/OpenNMT/OpenNMT-tf

5. Result and Discussion

The experimental results of Turkish→English and English→Turkish machine translation tasks are shown in Table 5 and Table 6, respectively. The NMT model [5] is baseline with only the word feature input. The single-source NMT model proposed by Sennrich and Haddow [14] with additional linguistic feature input is employed for comparison. For Turkish→English machine translation task, we can see from Table 5 that our proposed multi-source NMT model outperforms both the NMT model and the single-source NMT model. It achieves the highest BLEU and ChrF3 scores on all the test datasets, which indicates that our approach is capable of improving machine translation quality. Moreover, it achieves the highest improvements on tst2014 of 2.4 BLEU points and 1.6 ChrF3 points.

For English→Turkish machine translation task, we can see form Table 6 that our proposed model achieves the highest BLEU and ChrF3 scores on tst2012, tst2013, and tst2014. It achieves the highest improvements on tst2012 of 1.1 BLEU points and 1.5 ChrF3 points. As for the test dataset of tst2011, the multi-source NMT model is worse than the NMT model on BLEU score, but it is better on ChrF3. The main reason is that BLEU score is based on the precision of the Turkish words while ChrF3 score is based on both the precision and recall, so the two metrics are disagree. ChrF3 score was found to correlate well with human judgments, especially for the translation results out of English [25]. Therefore, we consider the multi-source NMT model is better than the NMT model on translation performance. Nevertheless, it is still not better than the single-source NMT model. The main reason is that the test data in tst2011 is not suitable for the multi-source NMT model since the linguistic features of English are not accurate or sufficient enough.

Table 5. The experimental results of Turkish→English machine translation task.

Model	**BLEU**				**ChrF3**			
	tst2011	**tst2012**	**tst2013**	**tst2014**	**tst2011**	**tst2012**	**tst2013**	**tst2014**
NMT Model [5]	24.18	25.95	26.60	24.98	47.18	48.61	48.65	48.05
Single-Source NMT Model [14]	24.69	26.65	27.43	25.98	47.99	49.26	49.89	48.87
Multi-Source NMT Model	**25.44**	**26.75**	**28.48**	**27.37***	**48.30**	**49.80**	**50.61**	**49.74***

Table 6. The experimental results of English→Turkish machine translation task.

Model	**BLEU**				**ChrF3**			
	tst2011	**tst2012**	**tst2013**	**tst2014**	**tst2011**	**tst2012**	**tst2013**	**tst2014**
NMT Model [5]	13.82	14.37	13.48	14.73	46.08	46.72	46.15	46.62
Single-Source NMT Model [14]	**13.95**	15.37	14.39	14.90	**46.86**	47.90	47.21	47.50
Multi-Source NMT Model	13.75	**15.48***	**14.44**	**15.36**	46.74	**48.21***	**47.51**	**47.73**

To further evaluate the effectiveness of different linguistic features, we incorporate single feature into our proposed multi-source NMT model. The experimental results of Turkish→English and English→Turkish machine translation tasks are shown in Table 7 and Table 8, respectively.

Table 7. The experimental results of incorporating single feature into the multi-source NMT model in Turkish→English machine translation task.

Encoder Input	**BLEU**				**ChrF3**			
	tst2011	**tst2012**	**tst2013**	**tst2014**	**tst2011**	**tst2012**	**tst2013**	**tst2014**
Word	24.18	25.95	26.60	24.98	47.18	48.61	48.65	48.05
Word + Lemma	**24.60**	**25.75**	**26.69**	25.20	**47.66**	**48.77**	**49.25**	48.33
Word + POS Tag	24.10	25.54	26.20	24.98	47.32	48.66	48.84	47.98
Word + Morph Tag	24.34	25.70	26.39	**25.50**	47.41	48.61	49.08	**48.36**

Table 8. The experimental results of incorporating single feature into the multi-source NMT model in English→Turkish machine translation task.

Encoder Input	BLEU				ChrF3			
	tst2011	tst2012	tst2013	tst2014	tst2011	tst2012	tst2013	tst2014
Word	13.82	14.37	13.48	14.73	46.08	46.72	46.15	46.62
Word + Lemma	13.79	14.59	**13.06**	14.55	45.99	47.11	**46.18**	46.51
Word + POS Tag	**14.06**	**15.01**	12.93	**14.93**	**46.52**	**47.98**	45.82	**46.73**

For Turkish→English machine translation task, we can see from Table 7 that incorporating the lemma feature of Turkish into the multi-source NMT model achieves the highest BLEU and ChrF3 scores on tst2011, tst2012, and tst2013 while incorporating the morphological tag achieves the highest BLEU and ChrF3 scores on tst2014. For English→Turkish machine translation task, we can see from Table 8 that incorporating the lemma feature of English into the multi-source NMT model achieves the highest BLEU and ChrF3 scores on tst2013 while incorporating the POS tag achieves the highest BLEU and ChrF3 scores on tst2011, tst2012, and tst2014. Experimental results show that different linguistic features are appropriate for different translation tasks and test datasets.

In addition, we find that incorporating single linguistic feature into the multi-source NMT model sometimes cannot yield improvements on BLEU or ChrF3 scores. The main reason is that the single feature cannot provide enough linguistic information of the source language for model training, so the multi-source encoder framework is not conducive to improving the NMT performance on source language representation. This fact indicates that our proposed approach is capable of better utilizing the source-side linguistic knowledge for the low-resource NMT and effective integrating linguistic information together at the word-level.

6. Related Work

Machine translation, which aims to perform transition between distinct languages, is a major focus of NLP research [26-27]. Recently, researchers show great interest in utilizing the source-side linguistic information as prior knowledge to improve machine translation quality. In phrase-based SMT model, the factored translation models are employed to incorporate external linguistic knowledge into the translation process [28]. Aqlan *et al.* integrate the linguistic features on top of the word surface form in translation model, and they iteratively train the SMT model to find the most optimized parameters [29]. However, the encoder-decoder based NMT model is more flexible to exploit additional information. The popular method is to modify the architecture of the encoder in the NMT model. Eriguchi *et al.* present a tree-to-sequence NMT model that encodes each phrase in the source parse tree, and they use an attention mechanism that allows the encoder to align both the input words and the input phrases with the output words [30]. Yang *et al.* improve the above work by encoding each node in the source parse tree with the local and global context information, and they introduce a weighted variant of the attention mechanism to adjust the proportion of the conditional information [31]. Li *et al.* combine the source sequence with its linearized syntactic structure into a single sequence to make the NMT model automatically learning useful language information [32].

Multi-source neural model is firstly proposed by Zoph and Knight for multilingual translation [33]. It is a many-to-one setting in the multi-task learning (MTL) method [34]. The model consists of multiple encoders with one encoder per source language. Then it combines the representation vectors from each encoder and feeds the resulting vector to the decoder to generate the desired target language. Following the above model, Li *et al.* build a knowledge gate and an attention gate to control the information from the source sentence and the linguistic features [35]. However, since their model simply concatenates each word's linguistic annotations into one single token, it cannot enrich the word representations in different context to capture semantic correlation. Different from their method, our approach mainly focus on the low-resource NMT, and it provides an effective way to fully exploit the source-side linguistic features at the word-level. Moreover, all the linguistic features are shared in the knowledge-based encoder to make better generalization.

7. Conclusion

In this paper, we propose a multi-source NMT approach for the low-resource NMT to explicitly utilize the source-side linguistic knowledge, which models the word sequence in parallel to the linguistic features by using two separate encoders with parameter sharing. Compared with the single encoder based NMT model, our approach employs additional knowledge-based encoder and extends its input embedding layer to accommodate for each word's linguistic annotations at the word-level, which can effectively enrich the word representations in different context. Experimental results show that our proposed approach is benefit for improving the low-resource NMT performance, and it achieves substantial improvements in Turkish→English and English→Turkish machine translation tasks. In particular, we demonstrate that different linguistic features are appropriate for different translation tasks and test datasets.

In the future, we plan to modify the attention mechanism in Transformer to better control the conditional information from different encoders. In addition, we plan to perform machine translation tasks on the other high-resource and morphologically-rich language pairs.

8. Patents

Author Contributions: Investigation, Yirong Pan; Visualization, Yirong Pan; Writing-original draft, Yirong Pan; Writing-review & editing, Xiao Li, Yating Yang and Rui Dong.

Acknowledgments: This work is supported by the High-Level Talents Introduction Project of Xinjiang Uygur Autonomous Region (Y839031201), the National Natural Science Foundation of China (U1703133), the National Natural Science Foundation of Xinjiang under Grant (2019BL-0006), the Open Project of Key Laboratory of Xinjiang Uygur Autonomous Region (2018D04018), and the Youth Innovation Promotion Association of the Chinese Academy of Sciences (2017472).

References

1. Sutskever, Ilya; Vinyals, Oriol; Le, Quoc V. Sequence to sequence learning with neural networks, Proceedings of the 2014 Conference of the Neural Information Processing Systems, Canada, December 13.
2. Cho, Kyunghyun; Van Merriënboer, Bart; Gulcehre, Caglar; Bahdanau, Dzmitry; Bougares, Fethi; Schwenk, Holger; Bengio, Yoshua. Learning phrase representations using RNN encoder-decoder for statistical machine translation, Proceedings of the 2014 Conference of the Empirical Methods on Natural Language Processing, Qatar, October 25-29.
3. Bahdanau, Dzmitry; Cho, Kyunghyun; Bengio, Yoshua. Neural machine translation by jointly learning to align and translate, Proceedings of the 2015 International Conference on Learning Representations, USA, May 7-9.
4. Luong, Minh-Thang; Pham, Hieu; Manning, Christopher D. Effective approaches to attention-based neural machine translation, Proceedings of the 2015 Conference of the Empirical Methods on Natural Language Processing, Portugal, September 17-21.
5. Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N; Kaiser, Łukasz; Polosukhin, Illia. Attention is all you need, Proceedings of the 2017 Conference of the Neural Information Processing Systems, UAS, December 4-9.
6. Koehn, Philipp; Och, Franz Josef; Marcu, Daniel. Statistical phrase-based translation, Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Canada, May 27-June 1.
7. Zhou, Qingyu; Yang, Nan; Wei, Furu; Zhou, Ming. Selective encoding for abstractive sentence summarization, Proceedings of the 2017 Annual Meeting of the Association for Computational Linguistics, Canada, July 30-August 4.
8. Duan, Nan; Tang, Duyu; Chen, Peng; Zhou, Ming. Question generation for question answering, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Denmark, September 9-11.
9. van Noord, Rik; Toral, Antonio; Bos, Johan. Linguistic information in neural semantic parsing with multiple encoders, Proceedings of the 2019 International Conference on Computational Semantics, Netherlands, June 3-5.

10. Currey, Anna; Heafield, Kenneth. Multi-source syntactic neural machine translation, Proceedings of the 2018 Conference of the Empirical Methods in Natural Language Processing, Belgium, October 31-November 4.
11. Bisazza, Arianna; Federico, Marcello. Morphological pre-processing for Turkish to English statistical machine translation, Proceedings of the 2009 Conference of the International Workshop on Spoken Language Translation, Japan, December 1-2.
12. Ataman, Duygu; Negri, Matteo; Turchi, Marco; Federico, Marcello. Linguistically motivated vocabulary reduction for neural machine translation from Turkish to English. *Prague Bulletin of Mathematical Linguistics* **2017**, *108*, 331-342.
13. Pan Yirong; Li Xiao; Yang Yating; Dong Rui. Morphological Word Segmentation on Agglutinative Languages for Neural Machine Translation. *arXiv preprint arXiv:2001.01589, 2020*.
14. Sennrich, Rico; Haddow, Barry. Linguistic input features improve neural machine translation, Proceedings of the 2016 Conference of the Empirical Methods in Natural Language Processing, Germany, August 11-12.
15. Jean, Sébastien; Cho, Kyunghyun; Memisevic, Roland; Bengio, Yoshua. On using very large target vocabulary for neural machine translation, Proceedings of the 2015 Conference of the Association for Computational Linguistics, China, July 26-31.
16. Sak, Haşim; Tunga, Güngör; Murat, Saraçlar. Morphological disambiguation of Turkish text with perceptron algorithm, Proceedings of the 2007 International Conference on Intelligent Text Processing and Computational Linguistics, Berlin.
17. Sennrich, Rico; Haddow, Barry; Birch, Alexandra. Neural machine translation of rare words with subword units, Proceedings of the 2016 Conference of the Association for Computational Linguistics, Germany, August 7-12.
18. Sennrich, Rico; Haddow, Barry; Birch, Alexandra. Improving neural machine translation models with monolingual data, Proceedings of the 2016 Conference of the Association for Computational Linguistics, Germany, August 7-12.
19. Cettolo, Mauro; Girardi, Christian; Federico, Marcello. Wit3: Web inventory of transcribed and translated talks, Proceedings of the 2012 Conference of the European Association for Machine Translation.
20. Tyers, Francis M; Alperen, Murat Serdar. South-east european times: A parallel corpus of balkan languages, Proceedings of the 2010 Conference of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and South-Eastern European Languages.
21. Pascanu, Razvan; Mikolov, Tomas; Bengio, Yoshua. On the difficulty of training recurrent neural networks, Proceedings of the 2013 International Conference on Machine Learning, USA, June 16-21.
22. Kingma, Diederik P; Ba, Jimmy. Adam: A method for stochastic optimization, Proceedings of the 2015 International Conference on Learning Representations, USA, May 7-9.
23. Papineni, Kishore; Roukos, Salim; Ward, Todd; Zhu, Wei-Jing. BLEU: A method for automatic evaluation of machine translation, Proceedings of the 2002 Conference of the Association for Computational Linguistics, USA, July 6-12.
24. Popović, Maja. ChrF: Character n-gram F-score for automatic mt evaluation. Proceedings of the 2015 Workshop on Statistical Machine Translation, Portugal, September 17-18.
25. Stanojević, Miloš; Kamran, Amir; Koehn, Philipp; Bojar Ondřej. Results of the wmt15 metrics shared task. Proceedings of the 2015 Workshop on Statistical Machine Translation, Portugal, September 17-18.
26. Gaspari, Federico; Almaghout, Hala; Doherty, Stephen. A survey of machine translation competences: Insights for translation technology educators and practitioners. *Perspectives* **2015**, *23*, 333-358.
27. Neubig, Graham; Watanabe, Taro. Optimization for statistical machine translation: A survey. *Computational Linguistics* **2016**, *42*, 1-54.
28. Koehn, Philipp; Hoang, Hieu. Factored translation models, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, June 28-30.
29. Aqlan, Fares; Fan, Xiaoping; Alqwbani, Abdullah; Al-Mansoub, Akram. Improved Arabic-Chinese machine translation with linguistic input features. *Future Internet* **2019**, *11*, 22.
30. Eriguchi, Akiko; Hashimoto, Kazuma; Tsuruoka, Yoshimasa. Tree-to-sequence attentional neural machine translation, Proceedings of the 2016 Conference of the Association for Computational Linguistics, Germany, August 7-12.

31. Yang, Baosong; Wong, Derek F; Xiao, Tong; Chao, Lidia S; Zhu, Jingbo. Towards bidirectional hierarchical representations for attention-based neural machine translation, Proceedings of the 2017 Conference of the Empirical Methods in Natural Language Processing, Denmark, September 9-11.
32. Li, Junhui; Xiong, Deyi; Tu, Zhaopeng; Zhu, Muhua; Zhang, Min; Zhou, Guodong. Modeling source syntax for neural machine translation, Proceedings of the 2017 Conference of the Association for Computational Linguistics, Canada, July 30-August 4.
33. Zoph, Barret; Knight, Kevin. Multi-source neural translation, Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics, USA, June 12-17.
34. Luong, Minh-Thang; Le, Quoc V; Sutskever, Ilya; Vinyals, Oriol; Kaiser, Lukasz. Multi-task sequence to sequence learning, Proceedings of the 2016 International Conference on Learning Representations, Puerto Rico, May 2-4.
35. Li, Qiang; Wong, Derek F; Chao, Lidia S; Zhu, Muhua; Xiao, Tong; Zhu, Jingbo; Zhang, Min. Linguistic knowledge-aware neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2018**, *26*, 2341-2354.