The TALP-UPC System for the WMT Similar Language Task: Statistical vs Neural Machine Translation

Magdalena Biesialska **Lluis Guardia** **Marta R. Costa-jussà**
TALP Research Center, Universitat Politècnica de Catalunya, 08034 Barcelona
magdalena.biesialska@upc.edu lluis.guardia@alu-etsetb.upc.edu
marta.ruiz@upc.edu

Abstract

Although the problem of similar language translation has been an area of research interest for many years, yet it is still far from being solved. In this paper, we study the performance of two popular approaches: statistical and neural. We conclude that both methods yield similar results; however, the performance varies depending on the language pair. While the statistical approach outperforms the neural one by a difference of 6 BLEU points for the Spanish-Portuguese language pair, the proposed neural model surpasses the statistical one by a difference of 2 BLEU points for Czech-Polish. In the former case, the language similarity (based on perplexity) is much higher than in the latter case. Additionally, we report negative results for the system combination with back-translation.

Our TALP-UPC system submission won 1st place for Czech→Polish and 2nd place for Spanish→Portuguese in the official evaluation of the 1st WMT Similar Language Translation task.

1 Introduction

Much research work has been done on language translation in the past decades. Given recent success of various machine translation (MT) systems, it is not surprising that some could consider similar language translation an already solved task. However, there are still remaining challenges that need to be addressed, such as limited resources or out-of-domain. Apart from these well-known, standard problems, we have discovered other under-researched phenomena within the task of similar language translation. Specifically, there exist languages from the same linguistic family that have a high degree of difference in alphabets, as it is the case for Czech-Polish, which may pose a challenge for MT systems.

Neural MT has achieved the best results in many tasks, outperforming former statistical MT (SMT) methods (Sennrich et al., 2016a). However, there are tasks where previous statistical MT approaches are still competitive, such as unsupervised machine translation (Artetxe et al., 2018; Lample et al., 2018). Motivated by the close proximity between the languages at hand and limited resources, in this article we aimed to determine whether the neural or the statistical approach is a better one to solve the given problem.

We report our results in the 1st Similar Language Translation WMT task (Barrault et al., 2019). In the official evaluation, our Czech→Polish and Spanish→Portuguese translation systems were ranked 1st and 2nd respectively. The main contributions of our work are the neural and statistical MT systems trained for similar languages, as well as the strategies for adding monolingual corpora in neural MT. Additionally, we report negative results on the system combination by using back-translation and Minimum Bayes Risk (Kumar and Byrne, 2004) techniques.

2 Background

In this section, we provide a brief overview of statistical (phrase-based) and neural-based MT approaches as well as strategies for exploiting monolingual data.

2.1 Phrase-based Approach

Phrase-based (PB) statistical MT (Koehn et al., 2003) translates by concatenating at a phrase level the most probable target given the source text. In this context, a phrase is a sequence of words, regardless if it is a phrase or not from the linguistic point of view. Phrases are extracted from word alignments between both languages in a large parallel corpus, based on the probabilistic study, which identifies each phrase with several features,

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 3: Shared Task Papers (Day 2) pages 187–193
Florence, Italy, August 1-2, 2019.

such as conditional probabilities. The collection of scored phrases constitutes the translation model.

In addition to this model, there are also other models to help achieve a better translation, such as the reordering model, which helps in a better ordering of the phrases; or the language model, trained from a monolingual corpus in the target language helping to obtain a better fluency in the translation. The weights of each of these models are optimized by tuning over a validation set. Based on these optimized combinations, the decoder uses beam search to find the most probable output given an input. Figure 1 shows a diagram of the phrase-based MT approach.

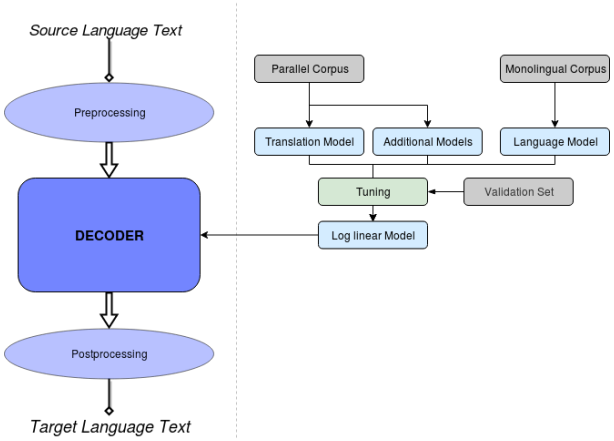

Figure 1: Basic schema of a phrase-based MT system

2.2 Neural Approach

Neural networks (NNs) have been successful in many Natural Language Processing (NLP) tasks in recent years. NMT systems, which use end-to-end NN models to encode a source sequence in one language and decode a target sequence in the second language, early on demonstrated performance on a par with or even outperformed traditional phrase-based SMT systems (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015; Sennrich et al., 2016a; Zhou et al., 2016; Wu et al., 2016).

Previous state-of-the-art NMT models used predominantly bi-directional recurrent neural networks (RNN) equipped with Long-Short Term Memory (LSTM; Hochreiter and Schmidhuber 1997) units or Gated Recurrent Units (GRU; Cho et al. 2014) both in the encoder and the decoder combined with the attention mechanism (Bahdanau et al., 2015; Luong et al., 2015). There were also approaches, although less common, to leverage convolutional neural networks (CNN) for sequence modeling (Kalchbrenner et al., 2016; Gehring et al., 2017).

In this work, we focus on the most current state-of-the-art NMT architecture, the Transformer (Vaswani et al., 2017), which shows significant performance improvements over traditional sequence-to-sequence models. Interestingly, while the Transformer employs many concepts that were used earlier in encoder-decoder RNN and CNN based models, such as: residual connections (He et al., 2016b), position embeddings (Gehring et al., 2017), attention; the Transformer architecture relies solely on the self-attention mechanism without resorting to either recurrence or convolution.

The variant of the self-attention mechanism implemented by the Transformer, multi-head attention, allows to model dependencies between all tokens in a sequence irrespective of their actual position. More specifically, the representation of a given word is produced by means of computing a weighted average of attention scores of all words in a sentence.

Adding Monolingual Data Although our proposed statistical MT model incorporates monolingual corpora, the supervised neural MT approach is not capable to make use of such data. However, recent studies have reported notable improvements in the translation quality when monolingual corpora were added to the training corpora, either through back-translation (Sennrich et al., 2016b) or copied corpus (Currey et al., 2017). Encouraged by those results and given the similarity of languages at hand, we propose to exploit monolingual data by leveraging back-translation as well as by simply copying target-side monolingual corpus and use it together with the original parallel data.

3 System Combination with Back-translation

In this paper, we propose to combine the results of both phrase-based and NMT systems at the sentence level. However, differently from the previous work of Marie and Fujita (2018), we aimed for a conceptually simple combination strategy.

In principle, for every sentence generated by the two alternative systems we used the BLEU score (Papineni et al., 2002) to select a sentence with the highest translation quality. Each of the translations was back-translated (i.e. translated from the target language to the source language). In-

stead of using only one system to perform back-translation, we used both PB and neural MT systems and weighted them equally. See Figure 2 for a graphical representation of this strategy.

This approach was motivated by the recent success of different uses of back-translation in neural MT studies (Sennrich et al., 2016b; Lample et al., 2018). The final test set was composed of sentences produced by the system that obtained the highest score based on the quality of the combined back-translation.

4 Experimental Framework

In this section we describe the data sets, data preprocessing as well as training and evaluation details for the PB and neural MT systems and the system combination.

4.1 Data and Preprocessing

Both submitted systems are constrained, hence they don't use any additional parallel or monolingual corpora except for the datasets provided by the organizers. For both Czech-Polish and Spanish-Portuguese, we used all available parallel training data, which in the case of Czech-Polish consisted of about 2.2 million sentences and about 4.5 million sentences in the case of Spanish-Portuguese. Also, we used all the target-side monolingual data, which was 1.2 million sentences for Polish and 10.9 million sentences for Portuguese.

Preprocessing Our NMT model was trained on a combination of the original Czech-Polish parallel corpus together with pseudo-parallel corpus obtained from translating Polish monolingual data to Czech with Moses. Additionally, the development corpus was split into two sets: first containing 2k sentences and second containing 1k sentences, where the former was added to the training data and the latter was used for validation purposes.

Our Phrase-Based model was trained on a combination of the original Spanish-Portuguese parallel corpus together with 2k sentences from the dev corpus. Specifically, the development corpus was split into two sets: first containing 2k sentences and second containing 1k sentences, where the former was added to the training data and the latter was used for validation purposes.

Then we followed the standard preprocessing scheme, where training, dev and test data are normalized, tokenized and truecased using *Moses*[1] scripts. Additionally, training data was also cleaned with `clean-corpus-n.perl` script from *Moses*. Finally, to allow open-vocabulary, we learned and applied byte-pair encoding (BPE)[2] for the concatenation of the source and target languages with 16k operations. The postprocessing was done in reverse order and included detruecasing and detokenization.

4.2 Parameter Details

Phrase-based For the Phrase-based systems we used Moses (Koehn et al., 2007), which is a statistical machine translation system. In order to build our model, we used generally the default parameters which include: grow-diagonal-final-and word alignment, lexical msd-bidirectional-fe reordering model trained, lexical weights, binarized and compacted phrase table with 4 score components and 4 threads used for conversion, 5-gram, binarized, loading-on-demand language model with Kneser-Ney smoothing and trie data structure without pruning; and MERT (Minimum Error Rate Training) optimisation with 100 n-best list generated and 16 threads.

Neural-based Our neural network model is based on the Transformer architecture (as described in section 2.2) implemented by Facebook in the *fairseq* toolkit[3]. The following hyperparameter configuration was used: 6 attention layers in the encoder and the decoder, with 4 attention heads per layer, embedding dimension of 512, maximum number of tokens per batch set to 4000, Adam optimizer with $\beta_1 = 0.90$, $\beta_2 = 0.98$, varied learning rate with the inverse square root of the step number (warmup steps equal 4000), dropout regularization and label smoothing set to 0.1, weight decay and gradient clipping threshold set to 0.

System Combination The key parameter in the system combination with back-translation, explained in section 3, is the score. Hence, we used the BLEU score (Papineni et al., 2002) at the sentence level, implemented as *sentence-bleu* in *Moses*. Furthermore, we assigned equal weights to both phrase and neural-based translations and back-translations.

[1] https://github.com/moses-smt/mosesdecoder
[2] https://github.com/rsennrich/subword-nmt
[3] https://github.com/pytorch/fairseq

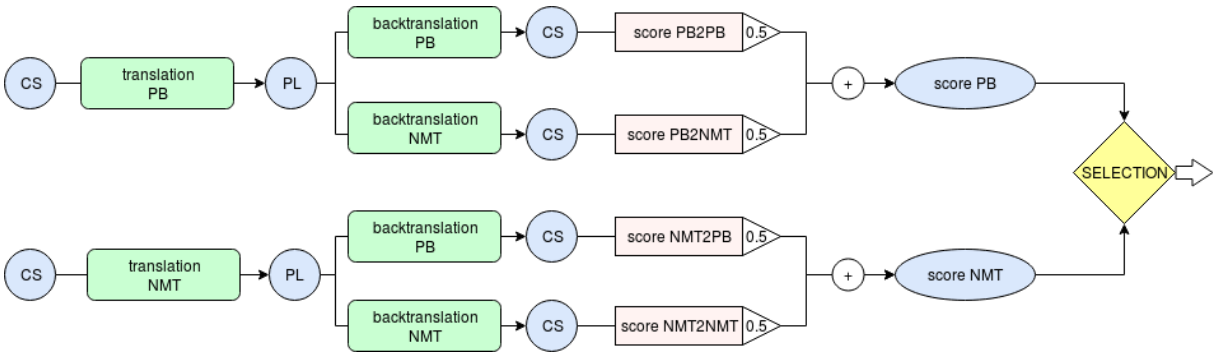

Figure 2: Scheme of the system combination approach

As contrastive approaches for system combination, we used two additional strategies: Minimum Bayes Risk (Kumar and Byrne, 2004) and the length ratio. In the former case, we used the implementation available in *Moses*. In the latter approach, the ratio was computed as the number of words in the translation divided by the number of words in the source input. Sentence translations that gave a length ratio closer to 1 were selected. In the case of ties, we kept the sentence from the system that scored the best according to Table 3.

5 Results

The results provided in Table 1 show BLEU scores for the direct phrase-based and neural-based MT systems. Also, we report on experiments with incorporating monolingual data in two ways: either using a monolingual corpus both on the source and target sides (*monolingual*) or using the back-translation system to produce a translation of a monolingual corpus (*pseudo corpus*). Interestingly, we observe that the *monolingual* approach harms the performance of the system even in the case of similar languages. With regard to the Spanish-Portuguese language pair, due to the large size of the monolingual corpora as well as the time constraint, we were unable to finish training of our NMT model with the pseudo corpus.

Table 1: Phrase-based (PB) and Neural-based (NMT) results.

	CS-PL	**ES-PT**
PB	9.87	64.96
NMT	11.69	58.40
NMT + monolingual	10.91	52.37
NMT + pseudo corpus	12.76	–

As presented in Table 3, our proposed system combinations, employing either MBR or the back-translation approach, did not achieve any significant improvements. The MBR strategy was applied to all systems from Table 1, which means that for the Czech-Polish pair we used 4 systems and for Spanish-Portuguese we used 3 systems. Back-translation results were evaluated with respect to the systems in Table 2 and the system combination with back-translation was created using the best two systems from Table 1.

Table 2: Back-translation system results.

1st system	2nd system	**PL-CS**	**PT-ES**
PB	PB	44.34	84.62
	NMT	24.51	66.15
NMT	PB	32.47	63.37
	NMT	27.31	60.01

Table 3: System Combination results.

	CS-PL	**ES-PT**
MBR	12.75	62.17
Back-translation	10.71	64.97

In order to analyze the reason behind the weak performance of the system combination with back-translation, we evaluated the correlation between the quality of each translated sentence (generated using PB and NMT systems) and the quality of back-translations (both for PB and NMT systems) on the validation set. For any combination, Czech-Polish or Spanish-Portuguese, correlation varies between 0.2 and 0.4, which explains the poor performance of back-translation as a quality estimation metric.

6 Discussion

Although Czech and Polish belong to the same family of languages (Slavic) and share the same subgroup (Western Slavic), the BLEU score obtained by our winning system is relatively low comparing to other pairs of similar languages (e.g. Spanish and Portuguese). It may seem surprising considering some common characteristics shared

by both languages, such as 7 noun cases, 2 number cases, 3 noun gender cases as well as 3 tenses among others.

Low performance on this task could be explained by the language distance. Considering the metric proposed by Gamallo et al. (2017), which is based on perplexity as a distance measure between languages, the distance between Czech and Polish is 27 while for Spanish-Portuguese is 7. The very same metric used to evaluate the distance of Czech and Polish from other Slavic languages (i.e. Slovak and Russian) shows that Polish is the most distant language within this group (see Table 4). In general, distances between Latin languages are smaller than between Slavic ones.

Table 4: Distances between Slavic and Latin languages. Examples across families.

Slavic		Latin		Mix	
pair	dist.	pair	dist.	pair	dist.
CS-PL	27	ES-PT	7	ES-CS	37
CS-SL	8	ES-FR	15	ES-PL	44
CS-RU	21	ES-RO	20	PT-CS	31
PL-SL	24	PT-FR	15	PT-PL	38
PL-RU	34	PT-RO	22		

While Czech and Polish languages are highly inflected, which poses a challenge, we hypothesize that one of the reasons for the low BLEU score lies also in the difference of the alphabets. Even though both alphabets are based on the Latin script, they include letters with diacritics – *ą, ć, ę, ł, ń, ó, ś, ź, ż* in Polish, and *á, č, d', é, ě, ch, í, ň, ó, ř, š, t', ú, ů, ý, ž* in Czech. The total number of unique letters in Polish is 32, while in the Czech language there are 42 letters. Moreover, some letters are used only in the case of foreign words, such as *q*, *x* (in Czech and Polish), *w* (in Czech), and *v* (in Polish).

7 Future Work

In the future we plan to extend our research in the following directions. First, we would like to explore how removing diacritics on the source-side would impact the performance of our system for the Czech-Polish language pair. Furthermore, we would like to study the performance of our system combination while applying various quality estimation approaches. We would be interested in experimenting with the reward score introduced by He et al. (2016a), which is a linear combination of language model score and the reconstruction probability of the back-translated sentence, as well as with other quality measures implemented in the *OpenKiwi* (Kepler et al., 2019) toolkit[4].

Acknowledgments

The authors want to thank Pablo Gamallo, José Ramom Pichel Campos and Iñaki Alegria for sharing their valuable insights on their language distance studies.

This work is supported in part by the Spanish Ministerio de Economía y Competitividad, the European Regional Development Fund and the Agencia Estatal de Investigación, through the postdoctoral senior grant Ramón y Cajal, the contract TEC2015-69266-P (MINECO/FEDER,EU) and the contract PCIN-2017-079 (AEI/MINECO).

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation, Volume 2: Shared Task Papers*, Florence, Italy. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar*, pages 1724–1734.

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*, pages 148–156, Copenhagen, Denmark. Association for Computational Linguistics.

[4]https://github.com/Unbabel/OpenKiwi

Pablo Gamallo, José Ramom Pichel, and Iñaki Alegria. 2017. From language identification to language distance. *Physica A: Statistical Mechanics and its Applications*, 484:152 – 162.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 1243–1252.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016a. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, pages 820–828.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016b. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709. Association for Computational Linguistics.

Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aäron van den Oord, Alex Graves, and Koray Kavukcuoglu. 2016. Neural machine translation in linear time. *CoRR*, abs/1610.10099.

Fábio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André FT Martins. 2019. Openkiwi: An open source framework for quality estimation. *arXiv preprint arXiv:1902.08646*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *HLT-NAACL 2004: Main Proceedings*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Benjamin Marie and Atsushi Fujita. 2018. A smorgasbord of features to combine phrase-based and neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 111–124, Boston, MA. Association for Machine Translation in the Americas.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

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

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto

Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Jie Zhou, Ying Cao, Xuguang Wang, Peng Li, and Wei Xu. 2016. Deep recurrent models with fast-forward connections for neural machine translation. *TACL*, 4:371–383.