Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 46 (2017) 367–373

www.elsevier.com/locate/csl

Introduction to the special issue on deep learning approaches for machine translation☆

Marta R. Costa-jussà*[,a], Alexandre Allauzen[b], Loïc Barrault[c], Kyunghun Cho[d], Holger Schwenk[e]

[a] *TALP Research Center, Universitat Politècnica de Catalunya, Spain*
[b] *LIMSI, CNRS, Univiversité Paris-Sud, Université Paris-Saclay, France*
[c] *LIUM, University of Le Mans, France*
[d] *Courant Institute of Mathematical sciences and Center for Data Science, New York University, France*
[e] *Facebook Artificial Intelligence Research, France*

Received 10 February 2017; Accepted 3 March 2017
Available online 25 May 2017

Abstract

Deep learning is revolutionizing speech and natural language technologies since it is offering an effective way to train systems and obtaining significant improvements. The main advantage of deep learning is that, by developing the right architecture, the system automatically learns features from data without the need of explicitly designing them. This machine learning perspective is conceptually changing how speech and natural language technologies are addressed. In the case of Machine Translation (MT), deep learning was first introduced in standard statistical systems. By now, end-to-end neural MT systems have reached competitive results. This special issue introductory paper addresses how deep learning has been gradually introduced in MT. This introduction covers all topics contained in the papers included in this special issue, which basically are: integration of deep learning in statistical MT; development of the end-to-end neural MT system; and introduction of deep learning in interactive MT and MT evaluation. Finally, this introduction sketches some research directions that MT is taking guided by deep learning.

Keywords: Machine translation; Deep learning,

1. Introduction

Considered as one of the major advance in machine learning, deep learning has been recently applied with success to many areas including Natural Language Processing, Speech Recognition and Image Processing. Deep learning techniques have surprised the entire community, both academy and industry, by its powerful ability to learn complex tasks from data.

Recently introduced to Machine Translation (MT), deep learning was first considered as a new kind of feature, integrated in standard statistical approaches (Koehn et al., 2003). Deep learning has been shown useful in translation

☆ This paper has been recommended for acceptance by Roger K. Moore.

* Corresponding author.
E-mail address: marta.ruiz@upc.edu (M.R. Costa-jussà).

http://dx.doi.org/10.1016/j.csl.2017.03.001
0885-2308/

and language modeling (Schwenk et al., 2007; Le et al., 2012; Vaswani et al., 2013; Devlin et al., 2014a) as well as in reordering (Li et al., 2014), tuning (Shen et al., 2004) and rescoring (Li and Khudanpur, 2009). Additionally, deep learning has been applied to MT evaluation (Gupta et al., 2015) and quality estimation (Kreutzer et al., 2015).

In the last couple of years, a new paradigm proposal has emerged: neural MT (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014). This new paradigm has yielded outstanding results, improving state-of-the art results for several language pairs (Jean et al., 2015; Sennrich et al., 2016; Wu et al., 2016a). This new approach uses an encoder-decoder architecture, along with an attention-based model (Bahdanau et al., 2015) to build an end-to-end neural MT system. This recent line of research opens new research perspectives and sketches new MT challenges, for instance dealing with: large vocabularies (Sennrich et al., 2015a; Costa-jussà and Fonollosa, 2016; Lee et al., 2016); multimodal translation (Elliott et al., 2015); the high computational cost, which implies new issues for large scale optimization Wu et al. (2016a).

This hot topic is raising interest from the scientific community and as a response there have been several related events (e.g. tutorial,[1] winter school.[2]) Moreover, the number of publications on this topic in top conferences (e.g. ACL[3] or EMNLP[4]) has dramatically increased in the last years. The main goal of this pioneer special issue is to gather articles that would give the reader a global vision, insight and understanding of deep learning limits, challenges and impact. This special issue contains high quality submissions on the following topics categories:

- Using deep learning in statistical MT.
- Neural MT.
- Interactive Neural MT.
- MT Evaluation enhanced with deep learning techniques.

The rest of the paper is organized as follows. Section 2 briefly describes the main current alternatives to build a neural MT approach. Section 3 overviews the papers on this special issue ordered by the different categories listed above. Finally, Section 4 discusses the main research perspectives on applying deep learning for MT.

2. Neural MT brief description

Most of the neural MT architectures are based on the so-called *encoder-decoder* approach, where an input sequence in source language is projected into a low-dimensional space from which the output sequence in target language is generated (Sutskever et al., 2014).

Then, many alternatives are possible for designing the encoder. A first approach is to use a simple recurrent neural network (RNN) to encode the input sequence (Cho et al., 2014). However, compressing a sequence into a fixed-size vector appears to be too much reductive to preserve source side information. Then, new systems were developed using bidirectional RNN. Source sequences are encoded into *annotations* by concatenating the two representations obtained with a forward and a backward RNN respectively. In this case, each *annotation* vector contains information from the entire source sequence but focusing on a particular word . An attention mechanism implemented by a feed-forward neural network is then used to attend specific parts of the input and to generate an alignment between input and output sequence. An alternative to the biRNN encoder is the stacked Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) as presented in (Sutskever et al., 2014; Wu et al., 2016b) (Fig. 1).

A major problem with neural MT is dealing with the large softmax normalization at the output which is dependent on the target vocabulary size. Many research works have been done to address this problem, like performing the softmax on a subset of the outputs only (Jean et al., 2014) or using a structured output layer to manage (Le et al., 2011) or self-normalization (Devlin et al., 2014b).

Another possibility is to perform translation at a subword level. This also have the advantage of allowing the generation of out-of-vocabulary words. Character-level machine translation has been presented in several papers

[1] http://naacl.org/naacl-hlt-2015/tutorial-deep-learning.html
[2] http://dl4mt.computing.dcu.ie/
[3] http://acl2016.org/
[4] http://www.emnlp2016.net/

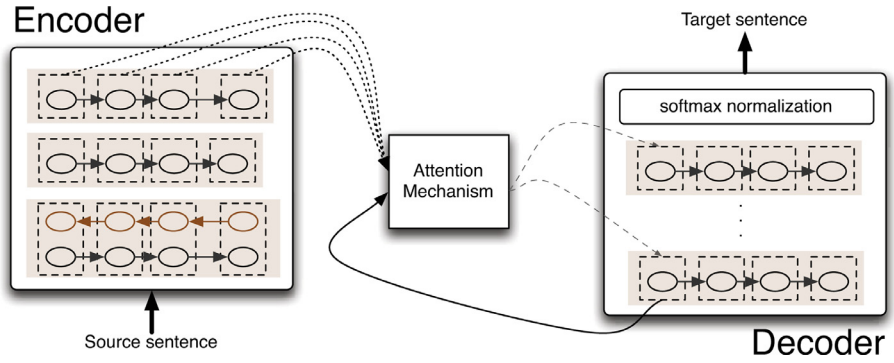

Fig. 1. Standard architecture of actual neural machine translation systems.

(Ling et al., 2015; Costa-Jussà and Fonollosa, 2016; Lee et al., 2016). Byte Pair Encoding (BPE) is a broadly used technique performing very well on many language pairs (Sennrich et al., 2015b).

3. Special issue overview

This section summarizes the papers in this special issue, covering the main idea and contribution of each one. Papers are organized in four categories, which include: using deep learning in statistical MT, neural MT, interactive neural MT and MT evaluation with deep learning techniques (Table 1).

3.1. Using deep learning in statistical machine translation

One of the first approaches to integrate neural networks or deep learning into MT has been through rescoring n-best lists from statistical MT systems (Schwenk et al., 2006; 2007). Given that statistical MT provides state-of-the-art results and deep learning helps in finding the right set of weights for statistical features, the scientific community is still doing research in this direction. As follows, we summarize the main research contributions of the two papers in this special issue that use deep learning to improve statistical MT.

Source sentence simplification for statistical machine translation by *Eva Hasler, Adrià de Gispert, Felix Stahlberg, Aurelien Waite and Bill Byrne.* Long sentences are a major challenge for MT in general. This paper uses text simplification to help hierarchical MT decoding with neural rescoring. Authors combine the full input sentence together with the simplified version of the same sentence. Simplification of the input sentence is done through deletion of most redundant words in the sentence. The corresponding integration is done using a two-step decoding approach to process both inputs. The first step translates the simplified input and produces an n-best list of candidates. The second step uses the n-best list to guide the decoding of the full input sentence.

The main contribution of the work is the procedure of integrating source sentence simplification into the hierarchical MT decoding with neural rescoring. This contribution is interesting for all types of MT and, therefore, further interesting work of this paper includes using source sentence simplification directly in a neural MT system.

Table 1
Summary of papers in this special issue classified by categories.

Category	Papers
Using deep learning in statistical MT	**Source Sentence Simplification for Statistical MT** by *Hasler* et al. **Domain Adaptation Using Joint Neural Network Models** by et al.
Neural MT	**Context-Dependent Word Representation for Neural MT** by *Choi* et al. **Multi-Way, Multilingual Neural MT** by *Firat* et al. **On Integrating a Language Model into Neural MT** by et al.
Interactive neural MT	**Interactive Neural MT** by *Peris* et al.
MT evaluation with deep learning	**MT Evaluation with Neural Networks** by et al.

Domain adaptation using joint neural network models by *Shafiq Joty, Nadir Durrani, Hassan Sajjad and Ahmed Abdelali*. Domain adaptation is still a challenge for MT systems in general, since parallel data can be considered as a scarce resource *wrt* the difference between text types and genres. Neural translation models, such as joint models, have shown an improved adaptation ability, thanks to the continuous representation. This paper investigates different ways to adapt this kind of models. Data selection and mixture modeling is the starting point of this work. The authors then propose a neural model to better estimate model weighting and instance selection in a neural framework. For instance, they introduce a pairwise model that minimizes the cross entropy by regularizing the loss function with respect to an in-domain model. Experimental results on the TED talk (Arabic-to-Engish) task show promising results.

3.2. Neural machine translation

Since the seminal work on neural MT (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014), the encoder-decoder architecture has fastly emerged as an efficient solution, yielding state of the art performance on several translation tasks. Beyond these important results, this kind of architecture renew the perspective of a multilingual approach to MT, but it also has some limitations. For instance, using source context information, together with dealing with highly multilingual frameworks and leveraging the abundant monolingual data remain still difficult challenges.

Context-dependent word representation for neural machine translation by *Heeyoul Choi, Kyunghyun Cho and Yoshua Bengio* deals with two major problems in MT, namely the word sense disambiguation (*i.e.* contextualization), and the symbolization aiming at solving the rare word problem. Contextualization is performed by masking out some dimensions of the target word embedding vectors (feedback) based on the input context, *i.e.* the average of the nonlinearly transformed source word embeddings. Symbolization is performed by introducing position-dependent special tokens to deal with digits, proper nouns and acronyms.

Experiments on the International Evaluation of WMT 2015 (Workshop on Statistical Machine Translation[5]) for two tasks show that the proposed contextualization and symbolic methods impact translation both quantitatively and qualitatively.

Multi-Way, multilingual neural machine translation by *Orhan Firat, Kyunghyun Cho, Baskaran Sankaran, Fatos T. Yarman Vural and Yoshua Bengio* addresses the challenge of efficiently managing highly multilingual environments. The paper presents a multi-way, multilingual neural MT approach with a shared attention mechanism (across language pairs). While keeping multiple encoders and multiple decoders, the main achievement of this paper is that the complexity of adding a language into the system increases the number of parameters only linearly, sharing the advantages of interlingua methods. The approach is tested on 8 language pairs (including linguistically similar and non-similar language pairs, high and low-resource language pairs). The approach improves strong statistical MT system in low-resource language pairs, and it achieves similar performance for other language pairs.

The shared attention mechanism is the main contribution of this paper compared to previous existing works on multilingual neural MT. This contribution is specially helpful when the number of language pairs is dramatically increased (e.g. highly multilingual contexts like the European) and/or for low-resource language pairs.

On Integrating a language model into neural machine translation by *Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Yoshua Bengio*. Neural MT training relies on the availability of parallel corpora. For several language pairs, this kind of resources are scarce. While conventional MT systems can leverage the abundant amount of monolingual data by means of target language models, neural MT systems are limited in their ability to benefit from this kind of resources. This paper explores two strategies to integrate recurrent neural language models in neural MT: a shallow fusion simply combine the scores of the neural MT and target language models, and a deep strategy explores the fusion of the hidden states of both models. Experimental results show promising improvements in terms of translation quality for both low and high-resource language pairs to be compared to state of the art MT systems thanks to the correctly exploitation of monolingual resources.

[5] http://www.statmt.org/wmt15/

3.3. Interactive neural machine translation

Despite the promising results achieved in last decades by MT, this technology is still error prone for some domains and applications. Interactive MT is in such cases an efficient solution, defining a collaboration between a human translator and a MT system, especially when the quality of fully-automated systems is insufficient. This approach consists in an iterative prediction–correction process in which, the MT system reacts offering a new translation hypothesis after each user correction. The recent emergence of neural MT has renewed the perspective for interactive MT, hence setting new challenges.

Interactive neural machine translation, by *Alvaro Peris, Miguel Domingo, Francisco Casacuberta* investigates the integration of a neural MT system in an interactive system. The authors propose a new interactive protocol which allows the user an enhanced and more efficient interaction with the system. First, a neural MT system is adapted to fit the prefix-based interactive scenario. In this conventional approach the user corrects the translation hypothesis by reading it from left to right creating a translation prefix that the MT system completes with a new hypothesis. This scenario is then extended by using the peculiarities of neural MT systems: the user can validate word segments and the neural MT system fills the gap by generating a new hypothesis. For these both scenarios, a tailored decoding strategy is proposed. Simulated experiments are carried out on four different translation tasks (user manuals, medical texts and TED talk translations) involving 4 language pairs. The results show a significant reduction of the human effort.

3.4. Machine translation evaluation with deep learning techniques

Progress in the field of MT heavily relies on the evaluation of a proposed translation. However, translation quality and its assessment is still an open question and a scientific challenge which have generated a lot of debates within the scientific community. For MT system development, the goal is to define an automatic metric that can both rank different approaches to measure progress and provide a replicable measure to be optimized. As an illustration, different shared tasks of the WMT evaluation campaigns are organized every year on this topic since 2008, showing its importance.

Machine translation evaluation with neural networks, by *Francisco Guzman, Shafiq Joty, Lluis Marquez and Preslav Nakov.* Given a reference translation, the goal is to select the best translation from a pair of hypotheses. The paper proposes a neural architecture able to represent in distributed vectors the lexical, syntactic and semantic properties from the reference and the two hypotheses.

The experimental setup relies on the WMT metrics shared task and the new flexible model highly correlates with human judgments. Additional contributions include task-oriented tuning of embeddings and sentence-based semantic representations.

4. Research perspectives

Neural MT is a very recent line of work which has already shown great results in many translation tasks. The community, however, lacks of hindsight about how research in the area will evolve in the upcoming year. In comparison, more than ten years were necessary to establish the phrase-based approach as the widespread, robust and intensively tuned solution for MT. Neural MT questions this statement by providing a unified and new framework, which to some extent, renders obsolete the inter-dependant components of statistical MT systems (word alignments, reordering models, phrase extraction). It is worth noticing that we are only at the beginning and that neural MT opens a wide range of research perspectives.

Nowadays, most of neural MT systems are based on an auto-encoder architecture which can evolve in many ways by considering for instance different encoders or richer attention-based models to better handle long-range reorderings and syntactic differences between languages. The decoder, or the model generation, is also an important part. The current objective function is based on maximum-likelihood and suffers of several limitations that can be solved within a discriminative framework (Shen et al., 2016) or with a learning-to-rank strategy (Wiseman and Rush, 2016). Neural MT also suffers from the vocabulary limitation issue which is well-known in the field of NLP. The complexity associated to a large output vocabulary hinders the application of such approaches to morphologically rich languages and to non-canonical texts like social media. To circumvent this limitation, several solutions are

under investigation: decoding at the character-level (Ling et al., 2015; Costa-jussà and Fonollosa, 2016; Lee et al., 2016), combining word and character representation (Miyamoto and Cho, 2016), or using subword units (Sennrich et al., 2015a).

Moreover, neural MT systems provide a very promising framework to learn continuous representations for textual data. This creates an important step moving from the word to the sentence level. Along with the introduction of the attention based model, these peculiarities renew how the notion of context can be considered within the translation process. This could allow the model to take into account for instance: a longer context, enabling document or discourse translation; a multi-modal context when translating image captions; or a social anchor to deal with different writing style. In the seminal paper on statistical machine translation (Brown et al., 1990), the authors set out the limit the approach considering that: ”in its most highly developed form, translation involves a careful study of the original text and may even encompass a detailed analysis of the author’s life and circumstances. We, of course, do not hope to reach these pinnacles of the translator’s art”. While this is still valid today, neural MT creates a real opportunity to extend the application field of machine translation in many aspects, beyond ”just” challenging the state-of-the-art performance.

Acknowledgments

The work of the 1st author is supported by the Spanish Ministerio de Economía y Competitividad and European Regional Development Fund, through the postdoctoral senior grant *Ramón y Cajal* and the contract TEC2015-69266-P (MINECO/FEDER, UE). The 4th author thanks the support by Facebook, Google (Google Faculty Award 2016) and NVidia (GPU Center of Excellence 2015−2016).

References

Bahdanau, D., Cho, K., Bengio, Y., 2015. Neural machine translation by jointly learning to align and translate. In: Proceedings of International Conference on Learning Representations.

Brown, P.F., Cocke, J., Pietra, S.D., Pietra, V.J.D., Jelinek, F., Lafferty, J.D., Mercer, R.L., Roossin, P.S., 1990. A statistical approach to machine translation. Comput. Linguist. 16 (2), 79–85.

Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 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. Doha, Qatar, pp. 1724–1734. A meeting of SIGDAT, a Special Interest Group of the ACL.

Costa-jussà, M.R., Fonollosa, J.A.R., 2016. Character-based neural machine translation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Berlin, Germany, pp. 357–361.

Devlin, J., Zbib, R., Huang, Z., Lañar, T., Schwartz, R., Makhoul, J., 2014. Fast and robust neural network joint models for statistical machine translation. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Vol. 1. Association for Computational Linguistics, Baltimore, Maryland, pp. 1370–1380.

Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R., Makhoul, J., 2014. Fast and robust neural network joint models for statistical machine translation. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 1370–1380.

Elliott, D., Frank, S., Hasler, E., 2015. Multi-language image description with neural sequence models. CoRR. abs/1510.04709.

Gupta, R., Orasan, C., van Genabith, J., 2015. Machine translation evaluation using recurrent neural networks. In: Proceedings of the Tenth Workshop on Statistical Machine Translation. Association for Computational Linguistics Lisbon, Portugal, pp. 380–384.

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

Jean, S., Cho, K., Memisevic, R., Bengio, Y., 2014. On using very large target vocabulary for neural machine translation. CoRR. abs/1412.2007.

Jean, S., Firat, O., Cho, K., Memisevic, R., Bengio, Y., 2015. Montreal neural machine translation systems for wmt15. In: Proceedings of the Tenth Workshop on Statistical Machine Translation. Association for Computational Linguistics, Lisbon, Portugal, pp. 134–140.

Kalchbrenner, N., Blunsom, P., 2013. Recurrent continuous translation models. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Seattle, Washington, USA, pp. 1700–1709.

Koehn, P., Och, F.J., Marcu, D., 2003. Statistical phrase-based translation. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, pp. 48–54.

Kreutzer, J., Schamoni, S., Riezler, S., 2015. Quality estimation from scratch (quetch): Deep learning for word-level translation quality estimation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation. Association for Computational Linguistics, Lisbon, Portugal, pp. 316–322.

Le, H.-S., Allauzen, A., Yvon, F., 2012. Continuous space translation models with neural networks. In: Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Association for Computational Linguistics, Montréal, Canada, pp. 39–48.

Le, H.-S., Oparin, I., Messaoudi, A., Allauzen, A., Gauvain, J.-L., Yvon, F., 2011. Large vocabulary SOUL neural network language models. In: Proceedings of INTERSPEECH.

Lee, J., Cho, K., Hofmann, T., 2016. Fully character-level neural machine translation without explicit segmentation. CoRR. abs/1610.03017.
Li, P., Liu, Y., Sun, M., Izuha, T., Zhang, D., 2014. A neural reordering model for phrase-based translation. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, pp. 1897–1907.
Li, Z., Khudanpur, S., 2009. MT from text. Forestreranking for machine translation with the perceptron algor ithm. GALE.
Ling, W., Trancoso, I., Dyer, C., Black, A.W., 2015. Character-based neural machine translation. CoRR. abs/1511.04586.
Miyamoto, Y., Cho, K., 2016. Gated word-character recurrent language model. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas, pp. 1992–1997.
Schwenk, H., Costa-Jussà, M.R., Fonollosa, J.A.R., 2006. Continuous space language models for the IWSLT 2006 task. In: Proceedings of 2006 International Workshop on Spoken Language Translation, IWSLT Keihanna Science City. Kyoto, Japan, pp. 166–173.
Schwenk, H., Costa-Jussà, M.R., Fonollosa, J.A.R., 2007. Smooth bilingual n-gram translation. EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Em pirical Methods in Natural Language Processing and Computational Natural Langua ge Learning, June 28−30, 2007. Prague, Czech Republic, pp. 430–438.
Sennrich, R., Haddow, B., Birch, A., 2015. Neural machine translation of rare words with subword units. CoRR. abs/1508.07909.
Sennrich, R., Haddow, B., Birch, A., 2016. Edinburgh neural machine translation systems for wmt 16. In: Proceedings of the First Conference on Machine Translation. Association for Computational Linguistics, Berlin, Germany, pp. 371–376.
Shen, L., Sarkar, A., Och, F.J., 2004. Discriminative reranking for machine translation. In: Susan Dumais, D.M., Roukos, S. (Eds.), Proceedings of HLT-NAACL 2004: Main Proceedings. Association for Computational Linguistics, Boston, Massachusetts, USA, pp. 177–184.
Shen, S., Cheng, Y., He, Z., He, W., Wu, H., Sun, M., Liu, Y., 2016. Minimum risk training for neural machine translation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, pp. 1683–1692.
Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. In: Proceedings of Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014. Montreal, Quebec, Canada, pp. 3104–3112.
Vaswani, A., Zhao, Y., Fossum, V., Chiang, D., 2013. Decoding with large-scale neural language models improves translation. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18−21 October 2013. Grand Hyatt Seattle, Seattle, Washington, USA,, pp. 1387–1392. A meeting of SIGDAT, a Special Interest Group of the ACL.
Wiseman, S., Rush, A.M., 2016. Sequence-to-sequence learning as beam-search optimization. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas, pp. 1296–1306.
Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J., 2016. Google's neural machine translation system: bridging the gap between human and machine translation. CoRR. abs/1609.08144.