From Feature to Paradigm: Deep Learning in Machine Translation (Extended Abstract)

Marta R. Costa-jussà
TALP Research Center - Universitat Politècnica de Catalunya, Barcelona
marta.ruiz@upc.edu

Abstract

In the last years, deep learning algorithms have highly revolutionized several areas including speech, image and natural language processing. The specific field of Machine Translation (MT) has not remained invariant. Integration of deep learning in MT varies from re-modeling existing features into standard statistical systems to the development of a new architecture. Among the different neural networks, research works use feed-forward neural networks, recurrent neural networks and the encoder-decoder schema. These architectures are able to tackle challenges as having low-resources or morphology variations.

This extended abstract focuses on describing the foundational works on the neural MT approach; mentioning its strengths and weaknesses; and including an analysis of the corresponding challenges and future work. The full manuscript [Costa-jussà, 2018] describes, in addition, how these neural networks have been integrated to enhance different aspects and models from statistical MT, including language modeling, word alignment, translation, reordering, and rescoring; and on describing the new neural MT approach together with recent approaches on using subword, characters and training with multilingual languages, among others.

1 Introduction[1]

The information society is continuously evolving towards multilinguality: e.g. different languages other than English are gainining more and more importance in the web; and strong societies, like the European, are and will continue to be multilingual. Different languages, domains, and language styles are combined as potential sources of information. In such a context, Machine Translation (MT), which is the task of automatically translating a text from a source language into a target language, is gaining more and more relevance. Both industry and academy are strongly investigating in the field which is progressing at an incredible speed. This progress may be directly attached to the introduction of deep learning. Basically, deep learning is the evolution of neural networks composed by multiple-layered models, and neural networks are machine learning systems capable of learning a task by training from examples and without requiring being explicitly programmed for that task. MT is just one of the applications where deep learning has succeeded recently. Although neural networks were proposed for MT in late nineties [Forcada and Ñeco, 1997; Castaño and Casacuberta, 1997], and have been integrated in different parts of statistical MT since 2006, it was not until 2013 and 2014 that first competitive neural MT systems were proposed [Kalchbrenner and Blunsom, 2013; Sutskever *et al.*, 2014; Cho *et al.*, 2014b], and in 2015, that neural MT reached the state-of-the-art [Bahdanau *et al.*, 2015].

1.1 MT Approaches Before Deep Learning

MT has been approached mainly following a rule-based or corpus-based strategy. Rule-based MT systems date back early 70s with the initiatives of Systran [Philipson, 2014 accessed September 2017] or EUROTRA [Maegaard, 1989]. The idea behind rule-based approaches is that transformation from source to target is done by means of performing an analysis of the source text, transfering (with hand-crafted rules) this new source representation to a target representation and generating the final target text.

Corpus-based approaches learn from large amounts of text. One popular and successful approach is the statistical one and, in particular, the phrase-based MT system [Koehn *et al.*, 2003]. This statistical approach benefits from being trained on large datasets. Normally, statistical MT uses parallel texts at the level of sentences, it uses co-occurrences to extract a bilingual dictionary, and finally, it uses monolingual text to compute a language model which estimates the most fluent translation text in the target language.

The main limitations of statistical MT are that it relies on parallel corpora. In rule-based MT, limitations are that it requires many linguistic resources, and a lot of human expert time. There is a considerable amount of research trying to hybridize these two approaches [Costa-jussà, 2015].

Another type of MT approaches, popular in the decade of the 80s, were interlingua-based, which focus on finding a universal representation of all languages. However, these approaches have fallen into disuse because it is very challenging

[1] This paper is an extended abstract of the JAIR publication [Costa-jussà, 2018]

and expensive to manually find a universal representation for all languages.

1.2 MT and Deep Learning

Recent appearence of new training and optimization algorithms for neural networks, i.e. deep learning techniques [Hinton *et al.*, 2006; Bengio, 2009; Goodfellow *et al.*, 2016], the availability of large quantities of data and the increase of computational power capacity have benefited the introduction of deep learning in MT.

Deep learning is about learning representations with multiple levels of abstraction and complexity [Bengio, 2009]. There has been a lot of excitement around deep learning because of the achieved breakthroughs, e.g. the automatic extraction of composition of images from lines to faces [Lee *et al.*, 2009], the ImageNet classification [Krizhevsky *et al.*, 2012] or reducing the error rate in speech recognition by around 10% [Graves *et al.*, 2013]. There has been a lot of recent activity from the scientific community in using deep learning in MT refelected in, for example, an explosion in the number of works in relevant conferences from 2014 up to date.

Deep learning has started as a feature function in statistical MT [Schwenk *et al.*, 2006] to become an entire new paradigm, which has achieved state-of-the-art results [Jean *et al.*, 2015] within one-year of development.

2 Foundational Works

Early research on this neural MT can be found in works like [Forcada and Ñeco, 1997; Castaño and Casacuberta, 1997], which were mainly limited by the computational power and short data. The former builds a state-space representation of each input string and unfolds it to obtain the corresponding output string. The latter uses an Elman simple RNN [Elman, 1990] to go from source to target.

First proposed neural MT models mainly use the previous encoder-decoder architecture [Sutskever *et al.*, 2014; Cho *et al.*, 2014b]. As explained in previous section **??**, this architecture allows for encoding the source text into a fixed-length vector and decoding this fixed-length vector into the target text. Both encoding and decoding are trained as a single architecture on a parallel corpus. The main problem with this type of architecture is to compress the source sentence into a fixed-length vector. [Cho *et al.*, 2014a] analyse this new approach and show that neural MT performs relatively well on short sentences without unknown words, but its performance degrades rapidly with the increment of sentence length and number of unknown words.

To address the long sentence issues, i.e. mainly caused by encoding the input sentence into a single fixed-lentgh vector, [Bahdanau *et al.*, 2015] propose a new mechanism where the decoder decides which parts of the source sentence to pay attention to. This attention mechanism relieves the encoder from having to compress all the source sentence into a fixed-length vector, allowing the neural translation model to deal better with long sentences. See schematic representation of the encoder-decoder with attention in Figure 1.

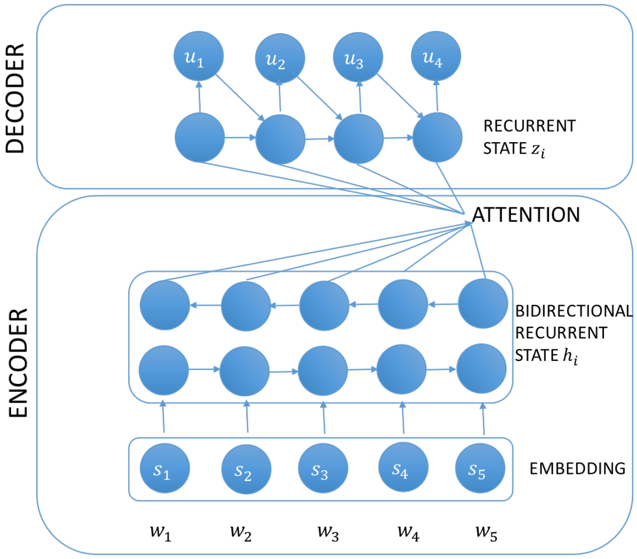

Figure 1: Neural MT architecture.

In the case of [Pouget-Abadie *et al.*, 2014], authors propose a way to address the challenge of long sentences by automatically segmenting an input sentence into phrases that can be easily translated by the neural network translation model.

3 Neural MT Analysis: Strengths and Weaknesses

Deep learning has been introduced in standard statistical MT systems [Costa-jussà, 2018] and as a new MT approach (see previous section 2). This section makes an analysis of the main strengths and weaknesses of the neural MT approach (see a summary in Figure 2). This analysis helps towards planning the future directions of neural MT.

Strengths The main inherent strength of neural MT is that all the model components are jointly trained allowing for an end-to-end optimization.

Another relevant strength is that, given its architecture based on creating an intermediate representation, the neural model could eventually evolve towards a machine-learnt interlingua approach [Johnson *et al.*, 2016]. This interlingua representation would be key to outperform MT on low-resourced language pairs as well as to efficiently deal with MT in highly multilingual environments.

In addition, neural MT has shown to be able to learn from different basic unit granularities. Subword-based representations, e.g. [Sennrich *et al.*, 2016; Costa-jussà and Fonollosa, 2016; Lee *et al.*, 2016], allow neural MT models with open-vocabulary by translating segmented words. Among the different alternatives to build subword units, the byte pair encoding, which is a data compresion technique, has shown to perform efficiently [Sennrich *et al.*, 2016]. Characters allows

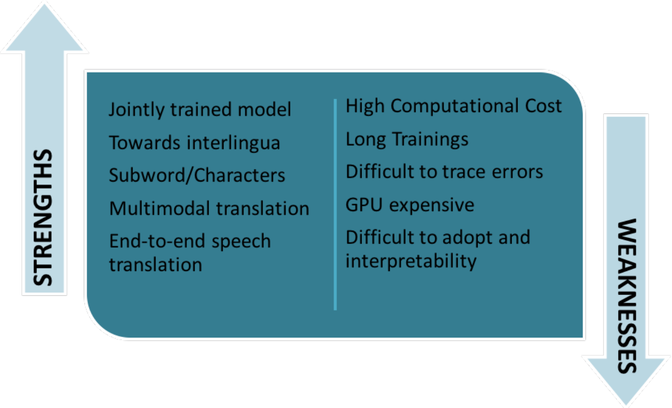

Figure 2: Strengths and Weaknesses analysis for Neural MT.

to take advantage of intra-word information and they have been implemented only in the source side [Costa-jussà and Fonollosa, 2016] and both in the source and target sides [Lee *et al.*, 2016].

Finally, the new paradigm allows for multimodal machine translation [Elliott *et al.*, 2015], allowing to take advantage of image information while translating and end-to-end speech translation architectures [Weiss *et al.*, 2017], which reduces concatenating errors.

Weaknesses The main inherent weaknesses of neural MT are the difficulty to trace errors, long training time and high computational cost. Other weakness is the high computational cost of training and testing the models. Training can only be faced with GPUs (Graphical Processing Units) which are expensive.

Finally, an added weakness is related to interpretability of the model and the fact that the model works with vectors, matrices and tensors instead of words or phrases. Therefore, the ability to train these neural models from scratch requires background in machine learning and computer science and it is not easy that users/companies are able to comprehend/interpret it. It is difficult to adopt the paradigm. Small companies may prefer other more consolidated paradigms like the phrase and rule-based.

4 Summary, Conclusions and Future Work

Deep learning has been integrated in standard statistical MT systems at different levels (i.e. into the language model, word alignment, translation, reordering and rescoring) from different perspectives and achieving significant improvements in all cases [Costa-jussà, 2018]. The field of deep learning is advancing so quickly that it is worth noticing that neural-based techniques that work today may be replaced by new ones in the near future.

In addition, an entire new paradigm has been proposed: neural MT. Curiously, this approach has been proposed almost simultaneaously as the popular phrase-based system in [Forcada and Ñeco, 1997; Castaño and Casacuberta, 1997]. The proposal was named differently *connectionist MT*, and given that the computational power required was prohibitive at that time and data available was not enough to train such complex systems, the idea was abandoned. Nowadays, thanks to GPUs, the computational power is not such a limitation and the information society is providing large quantities of data which allow to train the large number of parameters that these models have.

It is difficult to quantify how much does MT improve with the neural approach. It varies from language pair and task. For example, results on the WMT 2016 evaluation [Bojar *et al.*, 2016] show that neural MT achieved best results (in terms of human evaluation) in some language directions such as German-English, English Romanian, English-German, Czech-English, English-Czech; but not in others like Romanian-English, Russian-English, English-Russian, English-Finnish. Neural MT may be more affected by large language differences, low resources and variations in training versus test domain [Aldón, 2016; Costa-jussà *et al.*, 2017; Costa-jussá, 2017; Koehn and Knowles, 2017]. Interpreting MT systems has never before been more difficult. In the evolution of MT, we have first lost rules (in the transition from the rule to the statistical-based approach) and recently, we have lost translation units (in the transition from the statistical to the neural-based approach). Nowadays, the new neural-based approaches to MT are opening new questions, e.g. is it a machine-learnt interlingua something attainable? which are the minimal units to be translated?

This manuscript recompiles and systematizes the foundational works in using deep learning in MT which is progressing incredibly fast. Deep learning is influencing many areas in natural language processing and the expectations on the use of these techniques are controversial.

It is adventurous to envisage how neural algorithms are going to impact MT in the future but it seems that they are here to stay as proven by recent news on big companies adopting the neural MT approach e.g. Google[2] and Systran[Crego *et al.*, 2016].Furthermore, deep learning is already taking the field dramatically further as shown by the appearence of first end-to-end speech-to-text translation [Weiss *et al.*, 2017] and multimodal MT [Elliott *et al.*, 2015],interlingua-based representations [Firat *et al.*, 2017] and unsupervised MT [Artetxe *et al.*, 2017; Lample *et al.*, 2017].

[2] http://www.nature.com/news/deep-learning-boosts-google-translate-tool-1.20696

Acknowledgements

The authors would like to specially thank Prof. Kobi Gal for his invitation to participate in the IJCAI journal track.

This work is supported 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

[Aldón, 2016] David Aldón. *Sistema de Traducción Neuronal Usando Bitmaps*. B.s. thesis, Universitat Politècnica de Catalunya, 2016.

[Artetxe *et al.*, 2017] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. *CoRR*, abs/1710.11041, 2017.

[Bahdanau *et al.*, 2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2015.

[Bengio, 2009] Yoshua Bengio. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2(1):1–127, January 2009.

[Bojar *et al.*, 2016] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany, August 2016. Association for Computational Linguistics.

[Castaño and Casacuberta, 1997] M. A. Castaño and F. Casacuberta. A connectionist approach to mt. In *Proc. of the EUROSPEECH Conference*, 1997.

[Cho *et al.*, 2014a] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, October 2014. Association for Computational Linguistics.

[Cho *et al.*, 2014b] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 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, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734, 2014.

[Costa-jussà and Fonollosa, 2016] Marta R. Costa-jussà and José A. R. Fonollosa. Character-based neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 357–361, Berlin, Germany, August 2016. Association for Computational Linguistics.

[Costa-jussà *et al.*, 2017] Marta R. Costa-jussà, David Aldón, and José A. R. Fonollosa. Chinese-spanish neural machine translation enhanced with character andword bitmap fonts. *Machine Translation*, page Accepted for publication, 2017.

[Costa-jussà, 2015] Marta R. Costa-jussà. How much hybridization does machine translation need? *Journal of the Association for Information Science and Technology*, 66(10):2160–2165, 2015.

[Costa-jussá, 2017] Marta R. Costa-jussá. Why catalan-spanish neural machine translation? analysis, comparison and combination with standard rule and phrase-based technologies. In *Proceedings of the EACL Workshop: Vardial*, Valencia, April 2017.

[Costa-jussà, 2018] M.R. Costa-jussà. From feature to paradigm: Deep learning in machine translation. *Journal of Artificial Intelligence Research (JAIR)*, 61, 2018.

[Crego *et al.*, 2016] Josep Maria Crego, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, and Patrice Brunelle et al. Systran's pure neural machine translation systems. *CoRR*, abs/1610.05540, 2016.

[Elliott *et al.*, 2015] Desmond Elliott, Stella Frank, and Eva Hasler. Multi-language image description with neural sequence models. *CoRR*, abs/1510.04709, 2015.

[Elman, 1990] Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179 – 211, 1990.

[Firat *et al.*, 2017] Orhan Firat, Kyunghyun Cho, Baskaran Sankaran, Fatos T. Yarman Vural, and Yoshua Bengio. Multi-Way, Multilingual Neural Machine Translation. *Accepted for publication in Computer Speech and Language, Special Issue in Deep learning for Machine Translation*, 2017.

[Forcada and Ñeco, 1997] Mikel L. Forcada and Ramón P. Ñeco. Recursive hetero-associative memories for translation. In *Proceedings of the International Work-Conference on Artificial and Natural Neural Networks: Biological and Artificial Computation: From Neuroscience to Technology*, IWANN '97, pages 453–462, London, UK, UK, 1997. Springer-Verlag.

[Goodfellow *et al.*, 2016] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[Graves *et al.*, 2013] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649. IEEE, May 2013.

[Hinton *et al.*, 2006] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, July 2006.

[Jean *et al.*, 2015] Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Beijing, China, July 2015. Association for Computational Linguistics.

[Johnson *et al.*, 2016] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558, 2016.

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

[Koehn and Knowles, 2017] Philipp Koehn and REbecca Knowles. Six challenges for neural machine translation. In *ACL Workshop on Neural Machine translation*, 2017.

[Koehn *et al.*, 2003] Philipp Koehn, Franz Josef Och, and Daniel Marcu. 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*, NAACL '03, pages 48–54, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc.

[Lample *et al.*, 2017] Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *CoRR*, abs/1711.00043, 2017.

[Lee *et al.*, 2009] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 609–616, New York, NY, USA, 2009. ACM.

[Lee *et al.*, 2016] Jason Lee, Kyunghyun Cho, and Thomas Hofmann. Fully character-level neural machine translation without explicit segmentation. *CoRR*, abs/1610.03017, 2016.

[Maegaard, 1989] B. Maegaard. Eurotra: the machine translation project of the european communities. *Perspectives in artificial intelligence*, II, 1989.

[Philipson, 2014 accessed September 2017] J. Philipson. Systran: A brief history of machine translation, 2014; accessed September 2017.

[Pouget-Abadie *et al.*, 2014] Jean Pouget-Abadie, Dzmitry Bahdanau, Bart van Merrienboer, Kyunghyun Cho, and Yoshua Bengio. Overcoming the curse of sentence length for neural machine translation using automatic segmentation. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 78–85, Doha, Qatar, October 2014. Association for Computational Linguistics.

[Schwenk *et al.*, 2006] Holger Schwenk, Marta R. Costa-Jussà, and José A. R. Fonollosa. Continuous space language models for the IWSLT 2006 task. In *2006 International Workshop on Spoken Language Translation, IWSLT 2006, Keihanna Science City, Kyoto, Japan, November 27-28, 2006*, pages 166–173, 2006.

[Sennrich *et al.*, 2016] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

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

[Weiss *et al.*, 2017] Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. Sequence-to-sequence models can directly translate foreign speech. *CoRR*, abs/1703.08581, 2017.