DeepSPIN: Deep Structured Prediction for Natural Language Processing

**André F. T. Martins, Vlad Niculae, Erick Fonseca, Ben Peters,
Gonçalo Correia, Tsvetomila Mihaylova, Marcos Treviso, Pedro Martins**

Instituto de Telecomunicações and Unbabel,
Lisbon, Portugal

andre.t.martins@tecnico.ulisboa.pt

Abstract

DeepSPIN[1] is a research project funded by the European Research Council (ERC) whose goal is to develop new neural structured prediction methods, models, and algorithms for improving the quality, interpretability, and data-efficiency of natural language processing (NLP) systems, with special emphasis on machine translation and quality estimation applications.

1 Description

Neural network models became the standard in NLP applications, with impressive results in machine translation (Bahdanau et al., 2015; Vaswani et al., 2017). New language interfaces (digital assistants, customer service bots) are emerging as the next technologies for seamless, multilingual communication among humans and machines. From a machine learning perspective, many problems in NLP can be characterized as *structured prediction*: they involve predicting structurally rich and interdependent outputs. In spite of this, current neural NLP systems ignore the structural complexity of human language, relying on simplistic and error-prone greedy search procedures. This leads to critical mistakes in MT, such as words being dropped or named entities mistranslated.

The DeepSPIN project attacks these fundamental problems by bringing together deep learning and structured prediction. This is done in three fronts: better generation strategies, beyond left-to-right search; induction of sparse latent structure to make networks more interpretable; and incorporation of weak supervision to reduce the need for labeled data. We focus here on the applications to machine translation, including some results that have already been obtained in the project.

Alternate Generation Strategies. In Peters et al. (2019), we introduced *sparse sequence-to-sequence models*, with encouraging results in MT. These sparse losses endow MT systems with a small set of choices for the next word to be generated. When the model is fully confident, it autocompletes longer phrases (Figure 1). This property is appealing for building interactive MT systems. A current problem with left-to-right decoders is exposure bias (MT systems not exposed to their own predictions at training time). To mitigate this problem, we proposed a new scheduled sampling technique to avoid teacher forcing in transformers Mihaylova and Martins (2019). Future work will look at alternate generation strategies for MT, inspired by preliminary work in sequence tagging (Martins and Kreutzer, 2017).

Sparse Attention and Interpretability. One big goal of the DeepSPIN project is to make neural networks amenable to interpretation by humans. This is particularly useful in a scenario mixing MT and human post-editing. Our recent work on *sparse and structured attention* (Martins and Astudillo, 2016; Niculae et al., 2018) presents a promising avenue for enhancing interpretability (see Figure 1 for sparse word alignments), and we built on this idea in two directions: to reduce repetitions in neural MT by using constrained sparse attention to capture fertility (Malaviya et al., 2018); and using hierarchical sparse attention for document-level MT (Maruf et al., 2019). This idea has also been applied successfully in both RNN

[1]Project website: https://deep-spin.github.io.

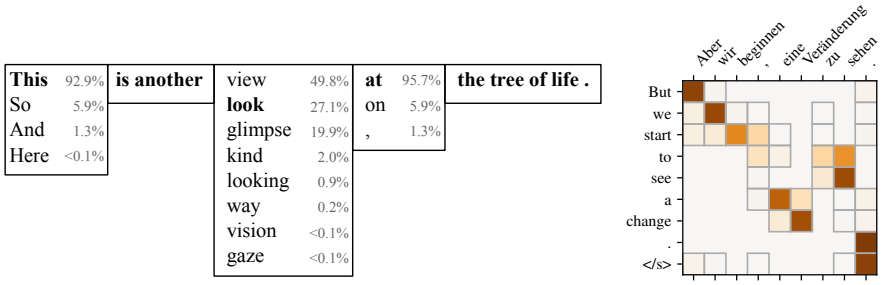

Figure 1: Left: Forced decoding using 1.5-entmax for the German source sentence "Dies ist ein weiterer Blick auf den Baum des Lebens." Only predictions with nonzero probability are shown at each time step. When consecutive predictions consist of a single word, we combine their borders to showcase *auto-completion* potential. The selected gold targets are in boldface. Right: Attention weights produced by a De-En 1.5-entmax model. Nonzero weights are outlined (Peters et al., 2019).

and transformer architectures (Peters et al., 2019; Correia et al., 2019). Future research will push this direction to improve efficiency and compactness of models as well as making MT systems more amenable to interpretation.

Weak Supervision. To avoid the data hunger of neural MT models, DeepSPIN will pursue techniques for weak supervision, such as transfer learning and leveraging of human post-editor activity data. This idea has been explored by Correia and Martins (2019) for automatic post-editing, by fine-tuning a pre-trained BERT model (Devlin et al., 2019). In collaboration with Unbabel, we extracted keystroke information from human post-editors in a crowd-sourced MT platform, and created a new dataset with keystroke sequences (Góis and Martins, 2019). This data[2] has proved extremely informative to understand the behavior of post-editors and predict translation quality.

Released Code and Datasets. To promote research reproducibility, the DeepSPIN project has released software code and datasets that may be useful to other researchers. This includes: OpenKiwi,[3] an open-source toolkit for quality estimation (Kepler et al., 2019); the `entmax` package[4] for sparse attention and losses; a new dataset with post-editor activity data (Góis and Martins, 2019);[2] and a new dataset for document-level quality estimation, used at WMT 2018 and 2019 shared tasks (Fonseca et al., 2019). New datasets are being prepared for shared tasks in WMT 2020.

[2]Available at `https://github.com/Unbabel/translator2vec/releases/download/v1.0/keystrokes_dataset.zip`

[3]`http://github.com/Unbabel/OpenKiwi`

[4]`https://github.com/deep-spin/entmax`

Acknowledgments. This work was supported by ERC StG DeepSPIN 758969 with AM as PI.

References

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Correia, Gonçalo M and André F. T. Martins. 2019. A simple and effective approach to automatic post-editing with transfer learning. In *ACL*.

Correia, Gonçalo, Vlad Niculae, and André F. T. Martins. 2019. Adaptively sparse Transformers. In *EMNLP*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Fonseca, Erick, Lisa Yankovskaya, André FT Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the wmt 2019 shared tasks on quality estimation. In *WMT*.

Góis, António and André FT Martins. 2019. Translator2vec: Understanding and representing human post-editors. In *MT Summit*.

Kepler, Fabio, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. Openkiwi: An open source framework for quality estimation. In *ACL System Demonstrations*.

Malaviya, Chaitanya, Pedro Ferreira, and André FT Martins. 2018. Sparse and constrained attention for neural machine translation. In *ACL*.

Martins, Andre and Ramon Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *ICML*.

Martins, André F. T. and Julia Kreutzer. 2017. Learning what's easy: Fully differentiable neural easy-first taggers. In *EMNLP*.

Maruf, Sameen, André FT Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. In *NAACL*.

Mihaylova, Tsvetomila and André F. T. Martins. 2019. Scheduled sampling for transformers. In *ACL Student Research Workshop*.

Niculae, Vlad, Andre Martins, Mathieu Blondel, and Claire Cardie. 2018. Sparsemap: Differentiable sparse structured inference. In *ICML*.

Peters, Ben, Vlad Niculae, and André F. T. Martins. 2019. Sparse sequence-to-sequence models. In *ACL*.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.