DeepSPIN: Deep Structured Prediction for Natural Language Processing

André F. T. Martins, Ben Peters, Chrysoula Zerva, Chunchuan Lyu, Gonçalo Correia, Marcos Treviso, Pedro Martins, Tsvetomila Mihaylova

Instituto de Telecomunicações and Unbabel,
Lisbon, Portugal

andre.t.martins@tecnico.ulisboa.pt

Abstract

DeepSPIN 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. We describe in this paper the latest findings from this project.

1 Description

The DeepSPIN project[1] is an ERC Starting Grant (2019–2023) hosted at Instituto de Telecomunicações. Part of the work has been done in collaboration with Unbabel, an SME in the crowdsourcing translation industry. The main goal of DeepSPIN is to bring together deep learning and structured prediction techniques to solve structured problems in NLP. The three main objectives are: developing better decoding strategies; making neural networks more interpretable through the induction of sparse structure; and incorporating of weak supervision to reduce the need for labeled data. We focus here on the applications to MT, including some of the recent results obtained in the project.

Better Decoding Strategies. Our initial work on sparse sequence-to-sequence models (Peters et al., 2019) proposed a new class of decoders (called "entmax decoders", shown in Fig. 1) which operate over a sparse probability distribution over words, which prunes hypotheses automatically. In (Peters and Martins, 2021), we have shown that entmax decoders are better calibrated and less prone to the length bias problem and developed a new label smoothing technique. We also presented entmax sampling for text generation, with improved generation quality (Martins et al., 2020). Another line of work concerns modeling of context in machine translation. We introduced *conditional cross-mutual information* (CXMI), a technique to measure the effective use of contextual information by context-aware systems, and *context-aware word dropout*, which increases its use, leading to improvements (Fernandes et al., 2021). We also compared the models' use of context to that of humans for translating ambiguous words, using the latter as extra supervision (Yin et al., 2021).

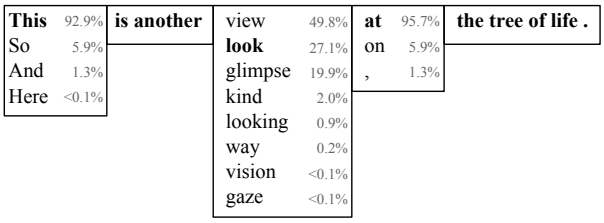

Figure 1: Forced decoding using entmax for the German source sentence "Dies ist ein weiterer Blick auf den Baum des Lebens." Only predictions with nonzero probability are shown at each time step. When consecutive predictions consist of a single word, we combine their borders to showcase *auto-completion* potential.

Sparse Attention and Explainability. A key objective of DeepSPIN is to make neural networks more interpretable to humans. Building upon our work on sparse attention mechanisms (Correia et al., 2019), we presented a framework to predict attention sparsity in transformer architectures, avoiding comparison of queries and keys which

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

MT	DA	COMET	UA-COMET
Она сказала, 'Это не собирается работать.	-0.815	*0.586*	0.149 **[-0.92, 1.22]**
Gloss: "*She said, 'that's not willing to work*"			
Она сказала: «Это не сработает.	0.768	1.047	1.023 [0.673, 1.374]
Gloss: "*She said, «That will not work*"			

Table 1: Example of uncertainty-aware MT evaluation. Shown are two Russian translations of the same English source "*She said, 'That's not going to work.*" with reference "Она сказала: "Не получится." For the first sentence, COMET provides a point estimate (in *red*) that overestimates quality, as compared to a human direct assessment (DA), while our UA-COMET (in **green**) returns a large 95% confidence interval which contains the DA value. For the second sentence UA-COMET is confident and returns a narrow 95% confidence interval. Taken from (Glushkova et al., 2021).

will lead to zero attention probability (Treviso et al., 2022). To model long-term memories, we proposed a new framework based on continuous attention, the ∞-*former* (Martins et al., 2022). We also compared different strategies for explainability of quality estimation scores, which led to an award in the EvalNLP workshop (Treviso et al., 2021).

Transfer Learning. We leveraged large pretrained models to build state-of-the-art models for quality estimation (Zerva et al., 2021) and for machine translation evaluation (Rei et al., 2021). Building upon the recently proposed deep-learned MT evaluation metric COMET (Rei et al., 2020), which tracks human judgements, we presented a new framework for uncertainty-aware MT evaluation (Glushkova et al., 2021), which endows COMET with confidence intervals for segment-level quality assessments (Table 1).

Released Code and Datasets. To promote research reproducibility, the DeepSPIN project has released software code and datasets, including: OpenKiwi,[2] an open-source toolkit for quality estimation (Kepler et al., 2019); the `entmax` package[3] for sparse attention and sparse losses; a dataset with post-editor activity data (Góis and Martins, 2019) and various datasets for quality estimation, used at WMT 2018–2021 shared tasks (Specia et al., 2021).

[2] `http://github.com/Unbabel/OpenKiwi`
[3] `https://github.com/deep-spin/entmax`

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

References

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

Fernandes, Patrick, Kayo Yin, Graham Neubig, and André FT Martins. 2021. Measuring and increasing context usage in context-aware machine translation. In *ACL*.

Glushkova, Taisiya, Chrysoula Zerva, Ricardo Rei, and André FT Martins. 2021. Uncertainty-aware machine translation evaluation. In *Findings of EMNLP*.

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*.

Martins, Pedro Henrique, Zita Marinho, and André FT Martins. 2020. Sparse text generation. In *EMNLP*.

Martins, Pedro Henrique, Zita Marinho, and André FT Martins. 2022. ∞-former: Infinite memory transformer. In *ACL*.

Peters, Ben and André FT Martins. 2021. Smoothing and shrinking the sparse seq2seq search space. In *NAACL*.

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

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Rei, Ricardo, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André FT Martins, and Alon Lavie. 2021. Are references really needed? Unbabel-IST 2021 submission for the metrics shared task. In *WMT*.

Specia, Lucia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André Martins. 2021. Findings of the wmt 2021 shared task on quality estimation. In *WMT*.

Treviso, Marcos, Nuno M Guerreiro, Ricardo Rei, and André FT Martins. 2021. IST-Unbabel 2021 submission for the explainable quality estimation shared task. In *EvalNLP*.

Treviso, Marcos, António Góis, Patrick Fernandes, Erick Fonseca, and André FT Martins. 2022. Predicting attention sparsity in transformers. In *SPNLP Workshop*.

Yin, Kayo, Patrick Fernandes, Danish Pruthi, Aditi Chaudhary, André FT Martins, and Graham Neubig. 2021. Do context-aware translation models pay the right attention? In *ACL*.

Zerva, Chrysoula, Daan van Stigt, Ricardo Rei, Ana C Farinha, Pedro Ramos, José GC de Souza, Taisiya Glushkova, Miguel Vera, Fabio Kepler, and André FT Martins. 2021. Ist-unbabel 2021 submission for the quality estimation shared task. In *WMT*.