SPECTRA: Sparse Structured Text Rationalization

Nuno M. Guerreiro, André Martins

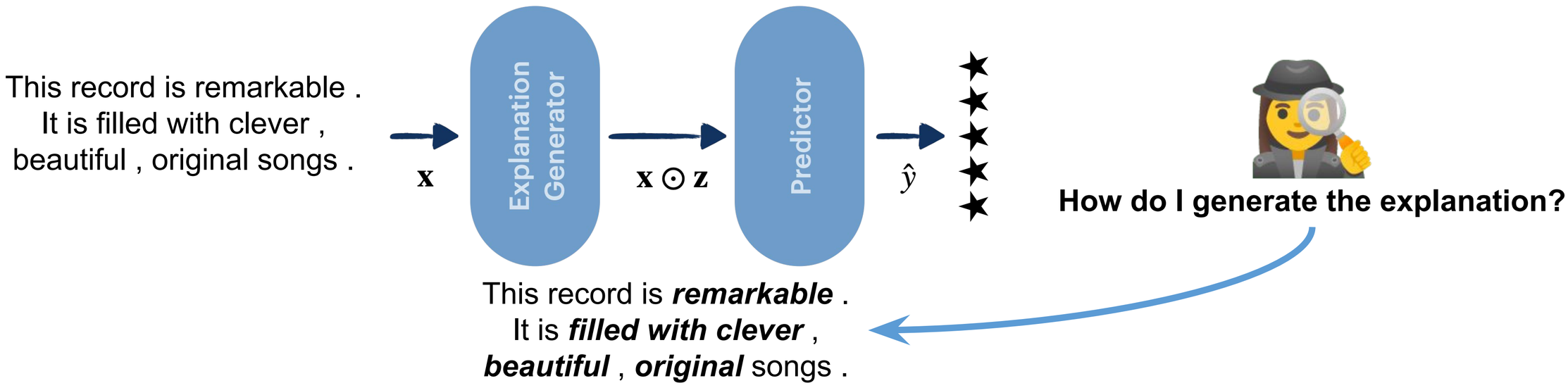
Instituto de Telecomunicações
Unbabel

CMU PORTUGAL SUMMIT 2022

NEW FRONTIERS IN TECH

Background and Motivation

Can we highlight relevant tokens for prediction to explain AI models?



- ***Stochastic*** 🎲 methods exhibit variability on the rationale extraction and are hard to train
- ***Deterministic*** 🎯 methods lack a way to regularize the rationale extraction

Can we propose an easy-to-train deterministic rationalizer with regularized rationale extraction?

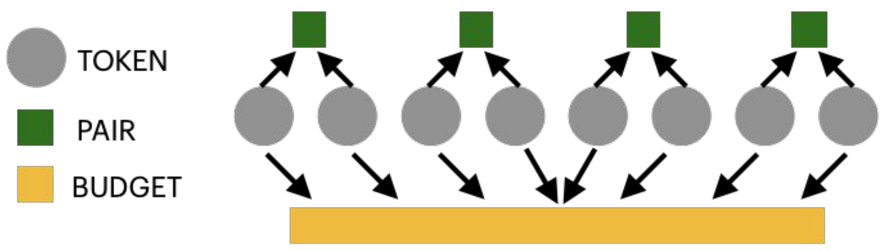
Rationale Extraction as a Structured Prediction Problem

We instantiate **rationales as structures**, whose global scores are given by:

$$\text{score}(\mathbf{z}; \mathbf{s}) = \mathbf{s}^\top \mathbf{z} + \sum_{f \in \mathcal{F}} h_f(\mathbf{z}_f)$$

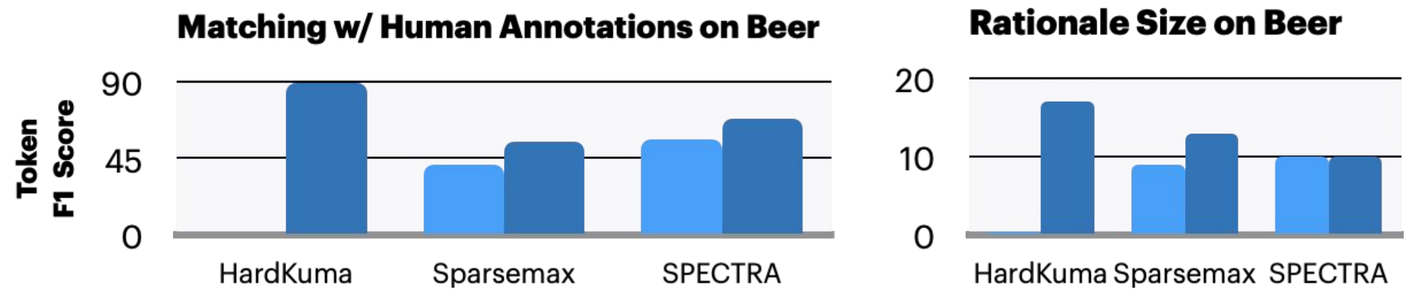
$\mathbf{s} = \text{gen}(\mathbf{x})$ — Generator scores

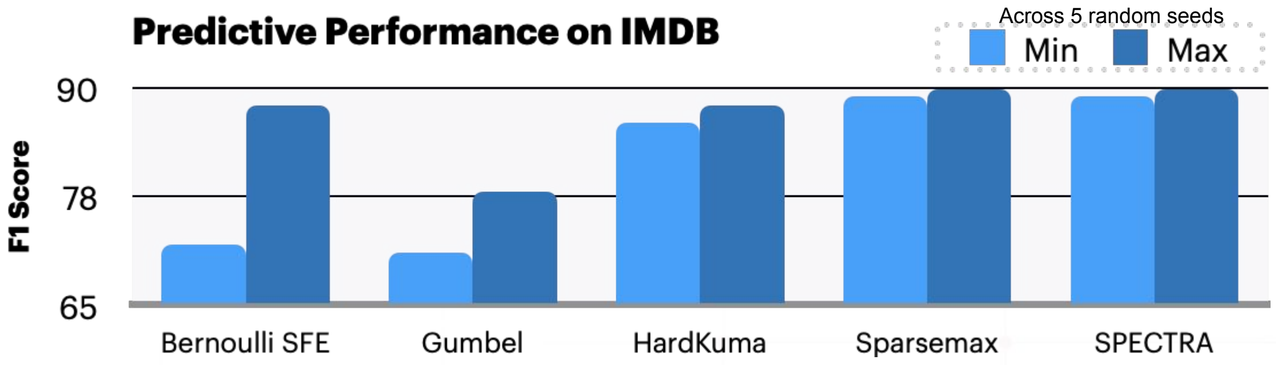
$h_f(\mathbf{z}_f) = \begin{cases} 0 & \text{if } \mathbf{z}_f \in \mathcal{Z}_f \\ -\infty & \text{otherwise} \end{cases}$ — Hard-constraints (e.g: rationale length)



$\mathcal{F} = \{\text{PAIR}(z_i, z_{i+1}; r_{i,i+1}) : 1 \le i < L\} \cup \{\text{BUDGET}(z_1, \ldots, z_L; B)\}$

Results





SPECTRA 🎯 generally outperforms other methods, exhibits less variability and its rationale extraction is regularized.

Acknowledgements

This work was supported by the P2020 programs (MAIA, contract 045909), by the European Research Council (ERC StG DeepSPIN, 758969), and by the Fundação para a Ciência e Tecnologia (contract UIDB/50008/2020).

Carnegie Mellon Portugal

fct Fundação para a Ciência e a Tecnologia

Unbabel

30 YEARS instituto de telecomunicações