Interpretability in sequence tagging models for Named Entity Recognition
S. Herrero Villarroya

Abstract

The field of Explainable Artificial Intelligence has taken steps towards increasing transparency in the decision-making process of machine learning models for classification tasks. Understanding the reasons behind the predictions of models increases our trust in them and lowers the risks of using them. In an effort to extend this to other tasks apart from classification, this thesis explores the interpretability aspect for sequence tagging models for the task of Named Entity Recognition (NER). This work proposes two approaches for adapting LIME, an interpretation method for classification, to sequence tagging and NER. The first approach is a direct adaptation of LIME to the task, while the second includes adaptations following the idea that entities are conceived as a group of words and we would like one explanation for the whole entity. Given the challenges in the evaluation of the interpretation method, this work proposes an extensive evaluation from different angles. It includes a quantitative analysis using the AOPC metric; a qualitative analysis that studies the explanations at instance and dataset levels as well as the semantic structure of the embeddings and the explanations; and a human evaluation to validate the model's behaviour. The evaluation has discovered patterns and characteristics to take into account when explaining NER models.