University of Amsterdam

MSc Artificial Intelligence

Master Thesis

Interpretability in sequence tagging models for Named Entity Recognition

by

Sofia Herrero Villarroya

11403861

August 14, 2018

36 ECTS

Feb-Aug 2018

Supervisor UvA:
Dr. Giorgio Patrini

Assessor:
Dr. Zeynep Akata

Supervisor Elsevier:
Deep Kayal

Acknowledgements

I would like to thank my daily supervisor Giorgio Patrini and my supervisor at Elsevier Deep Kayal for their involvement in the project. Throughout the thesis, both of you have always made time to give me valuable advice and feedback. I have learned a lot from both of you and you have been of great support and help for this project.

I would also like to thank my friends and family for their unconditional support and inspiring optimism. Your advice and the time you dedicated to me helped me enormously throughout this thesis.

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.

Contents

Chapter 1

Introduction

The field of Artificial Intelligence has taken enormous steps in the past years and is becoming ubiquitous in the decision-making process of numerous tasks, ranging from applications in our daily lives to more complex tasks. Typical machine learning algorithms that have been used for years could start being replaced by the new deep learning algorithms in common tasks such as recommendation algorithms [1] or malware detection [2]. For more complex tasks, such as image captioning [3] or machine translation [4], deep learning models have shown an impressive performance, making these tasks possible in industry and research. The main reason has been the availability of new resources such as larger data sets, more flexible frameworks and more powerful hardware, which has led to machine learning and deep learning being the crucial component in new applications.

These models in the area of Deep Learning have excelled in a number of complex applications for which they are able to extract hidden representations and patterns and improve prediction accuracy. They are characterized by the complexity of their structure which can include several non-linear layers stacked with hundreds of hidden units. Due to this structure and the size, the decision process that the model follows cannot be fully explained and the models are often used as black boxes. This directly compromises the interpretability of the model which, in turn, impacts its trustworthiness. The decision-making process in humans is backed up by our ability to explain the rationale behind it. We trust a decision more if the decision maker can explain the reasons behind it. For instance, we trust more a doctor's prescription when she or he explains the reasons behind it.

However, the compromise of interpretability for complexity in the new deep learning models (and in the vast majority of the machine learning models) to obtain a higher accuracy is not always affordable. In domains where the consequences of the prediction can be catastrophic, such as the medical or the legal domain, applying models as black boxes is not an option since the model's decision cannot be fully trusted. In a similar way as the human rationale, the line of research in Explainable Artificial Intelligence and interpretable Machine Learning seeks to have transparent models whose behavior can be understandable and explainable and thus, trustworthy, in order for humans to use them. The efforts in this area have been mainly focused on explaining predictions for classification models. Therefore, the aim of this thesis is to broaden the interpretability area by exploring it in models for a different type of task, sequence tagging, instead of classification.

Sequence tagging or sequence labeling is a very common task in applications for Natural Language Processing (NLP) and Information Extraction (IE) such as Part-of-Speech tagging or Named Entity Recognition. The input for sequence tagging models is a sequence of tokens and the output is a sequence of labels, one per token. The task of Named Entity Recognition (NER) consists in finding entities in text data. These entities are proper nouns of places, organizations, people or any other category. It is an important task in NLP and IE since it is usually the first step of a pipeline for more complex tasks such as relation extraction, triplet extraction, assertion classification, summarization and any other application that builds on the entities recognized. To ensure the whole pipeline is explainable the very first step should be explainable. Therefore, this thesis explores, for the first time, the application of explainable AI techniques to sequence tagging models in the context of Named Entity Recognition. An attempt is made to adapt the interpretation methods made for classifiers to this task to determine the viability and the directions for interpretability in sequence tagging models.

1.1 Research questions

The field of Explainable AI has made progress towards the development of interpretation techniques for classifiers. This includes the family of attribution interpretation methods. The interpretation or explanation given by these methods attributes a relevance score to the different features of the input according to how much they impact the prediction. The initial attempts in the field of Explainable AI included model-dependent methods and it has been in the past two years that methods are shifting towards model-agnosticism since this is a desired characteristic to make interpretation possible in any domain. Therefore, for this thesis, we have chosen to use LIME (Linear model-agnostic explanations) since it is one of the few of them that is model-agnostic and is more flexible and general than others in that category.

Named Entity Recognition (NER) involves finding entities which can be composed of several words. Ideally, the interpretation of a prediction would give the user which words the model used to predict the entity, as a group. However, it is not clear whether this would be the most desirable way or how possible it is. With the aim of exploring this aspect, several research questions were formulated and answered in this thesis.