M. Sc. thesis exposé: Interpretable Rationale Extraction within Named Entity Recognition

Richard Herrmann

March 2022

1 Introduction

While neural networks have produced recent breakthroughs in a vast array of Machine Learning tasks, the inner workings of models produced in this way are unfortunately quite opaque which makes it hard for humans to understand the decision making process. As a result, researchers and companies are faced with certain problems. These range from an interest in identifying the shortcomings of a given model (in order to debug and improve it) to compliance with legal requirements such as pushes from EU lawmakers for a 'right to explanation' of machine-made decisions. [1] This includes the identification of discriminatory biases inherent in systems trained on real-world data. [2] These problems have sparked an interest in both creating interpretable systems and making existing systems explainable. Explanations that can serve this purpose are called rationales and the task of acquiring them is called Rationale Extraction. [3]

In the context of Natural Language Processing, a rationale is usually either a list of the k most important tokens wrt their importance in the decision making process or a continuous span of tokens with the length k. The latest research in the field of NLP has already produced a few insights into how NLP models could potentially make explainable decisions. [4] But the focus of this research has so far been mostly on Text Classification tasks whereas other standard NLP tasks such as Named Entity Recognition (NER) have seen little to no publications with regards to interpretability yet.

The goal of this Master's thesis shall be to choose and adapt one of the suggested approaches to create interpretable NER models that can be evaluated in comparison with the performance of their non-interpretable variants and in terms of quality of the resulting explanations. The datasets used will be focussing on the domain of biomedical information extraction.

2 Background

2.1 Definitions of interpretability: Plausibility vs. Faithfulness

Like the search for interpretable models itself, the definition of what interpretability actually means and how the resulting rationales should be evaluated are ongoing research questions.

Early publications have mostly modeled interpretability in terms of plausibility, i.e. on how comprehensible an interpretable rationale visualizing the decision making process is to a human reviewer. A small number of NLP datasets with human reviewers' annotations of plausible rationales have been made available. [4] There is one NER dataset with plausbility annotations called TriggerNER although it does not fall under the biomedical domain. [5]

Lately there have been calls for viewing explainability - additionally or primarily - in the light of faithfulness (while plausibility research is ongoing). Faithfulness concerns the ability of a given rationale to accurately formulate how the decision process was made within the system. [6] It has been suggested that it would be impossible by definition to have human annotations for interpretability as this would require knowledge of the inner workings of an artificial neural network. Therefore all human evaluations would be biased towards plausibility. [6] Indeed, it has been argued that a rationale that is highly plausible but unfaithful could even be seen as the worst case scenario when it appears logical to a human reviewer while failing to explain what actually led to that decision. For example, when a computer assisted decision is applied in a legal or medical setting trust in a decision with an erroneous explanation could have dire consequences. It has been suggested that, in practice, plausibility and faithfulness of a rationale have little to no correlation. [7] While there are approaches that seek to combine both criteria [3], a high faithfulness is chosen as the goal within this thesis, partly because of the unavailability of biomedical NER datasets with gold annotations for plausibility. The aforementioned non-biomedical dataset TriggerNER [5] could be used and its plausibility score evaluated as an optional goal although the explicit focus on this thesis is on faithfulness.

2.2 Rationale Extraction

The task of Rationale Extraction has been attacked from various angles in NLP, most of which concern Transformer-based Language Models (LMs) like BERT and its variants. In a typical pre-processing step the saliency (i.e. word importance scores) of each token is computed on the input sequence with a suitable metric (called saliency metric or attribution algorithm) like TextRank. [8]

In a generalized setting the classification is run over the full input sequence while trying to maximize a declared explainability objective - faithfulness and/or plausibility - of additionally extracted rationales. Post-hoc Rationale Extraction without use of the rationale in training is an option, too. The most important choice that must be made is the selection of the rationale extractor. Generally,

these can be divided into heuristic and learned functions. [3]

2.2.1 Heuristic Rationale Extractors

Heuristic Rationale Extractors can deploy any attribution algorithm that computes importance scores for each input token. These algorithms are often either gradient-based or perturbation-based. The idea behind perturbation (also called word erasure) is to remove (or mask) tokens from the input sequence and record how the LM output changes in comparison to a classification of the complete, unaltered input. Gradient-based saliency methods compute the importance of a specific input feature based on the first-order derivative with respect to that feature. An advantage of Heuristic Rationale Extractors is the independence from gold rationale supervision. Unfortunately they also have downsides. Usually these heuristics are compute-intensive (especially for more faithful gradient-based methods), perturbation-based methods can become prohibitively expensive for long sequences of thousands of tokens. [3] In comparison with Text Classification this problem should be less pronounced in the context of Named Entity Recognition where mostly single sentences serve as the input sequence as opposed to whole documents that consist of many sentences. Another feature that makes the heuristic approach hard to deploy in production (according to [3]) is that real-world use cases are typically optimized for LM forward passes whereas backward passes require specialized scripted implementations. This might not be a problem for post-hoc Heuristic Rationale Extractors.

2.2.2 Learned Rationale Extractors

On the other side there are Learned Rationale Extractors which can utilize any neural network model that transforms input tokens into importance scores. In practice this means pre-trained Transformer LMs. This method is relatively fast, can share parameters with the model for the actual task and can be deployed into production more easily. [3] Of course this approach has the availability of gold rationales as training data as a hard requirement which unfortunately makes it only suitable for our purposes at the current point if the TriggerNER dataset is used. But since it is not a biomedical dataset, a Heuristic Rationale Extractor is favored.

2.2.3 Special Case: Select-Predict Pipelines

One notable group of interpretability algorithms consists of Select-Predict Pipelines. The idea is to construct inherently faithful rationales by using the rationales themselves as the input for the final classification, thus preventing other information outside of the rationale from impacting the prediction. Saliency scores can optionally be used in training in the shape of a loss function that penalizes a model when the attention distribution deviates from the salience distribution (SaLoss [8]). When making a prediction with the trained model, the saliency score is used to build a binary mask which only retains the top-k most important tokens or a continuous span of length k. A promising method is FRESH [9]

which trains the selector and the predictor separately (not jointly like in other select-predict pipelines). While the faithfulness of all rationales is always perfect in select-predict methods the downside is that the task performance suffers from the input information reduction, often considerably so. [3]

2.3 Evaluation of Faithfulness

For the evaluation of faithfulness, several metrics have been proposed although there is no metric as of now that can be called the agreed-upon standard evaluation metric. DeYoung et al. [4] have proposed the metrics of comprehensiveness and sufficiency which have been picked up comparatively often since their publication. Both measures are based on word erasure to compare (classification) model outputs based on differing input sequences.

2.3.1 Comprehensiveness

Comprehensiveness [4][6][3] measures the saliency of the tokens included in the rationale p by calculating the change in the model's confidence in its prediction when the tokens of p are removed from (or masked in) the input. The change is calculated using cross-entropy loss. The expectation is that a rationale is comprehensive when the model loses confidence in its prediction (i.e. lower class probability) once it can't see the tokens of p anymore. Therefore a high scoring of the value change is targeted.

2.3.2 Sufficiency

Conversely, the complementary measure is Sufficiency [4][6][3] which observes the change in confidence when only the tokens of the rationale p are kept as input for the prediction. For this measure a low scoring is desirable because the idea is that a sufficient explanation of a classification should show little change compared to the class probability of the complete input sequence. To avoid confusion from the fact that a highly sufficient rationale has a low sufficiency score, the reverse difference (1 - suff) can be used so that a higher score is better. [7] Nevertheless, a minimization of the sufficiency scores has its merits as seen in the next subsection.

2.3.3 Comp-Suff-Difference

Both measures can be combined into a single metric which is a simple difference between the comprehensiveness and sufficiency scores, Comp-Suff-Difference (CSD). [3] As a result, a highly faithful rationale is defined as being highly comprehensive and having a low sufficiency score.

2.4 Evaluation of Performance

For the NER task performance (as opposed to the Rationale Extraction task performance) standard evaluation measures such as the macro-averaged F1 score

can be used. This allows for an easy comparison with existing NER models that don't use interpretability as a criterion. Additionally, the speed and data efficiency of the Rationale Extraction task can be measured to compare different RE approaches. These measures are only meaningful in case an ad-hoc Rationale Extractor is implemented as post-hoc extractors don't change the model in question.

3 Goals of this thesis

The goal of this thesis is to evaluate the task of Rationale Extraction as a part of the Named Entity Recognition task. To achieve this, the idea is to implement and evaluate one to three Heuristic Rationale Extractors of the following (non-exhaustive) list of candidate approaches, depending on the required effort: Vanilla Gradient [10], gradient×input [11], SmoothGrad (SG) [12], Integrated Gradients (IG) [13], Layer-wise Relevance Propagation (LRP) [14], LIME [15], TextRank [8] or the perturbation-based methods proposed by Li et al. [16], Kadar et al. [17] or Poerner et al. [18]. The faithfulness evaluation metrics (comprehensiveness and sufficiency) will also have to be implemented as a part of the thesis.

4 Implementation

The implementation will be done within the FLAIR framework [19] which already takes care of many parts of a typical NER workflow such as loading of the datasets in the required format. Thanks to HunFlair [20] a variety of biomedical NER datasets is already available in FLAIR. There is already a wide range of NER implementations available within the framework as well. This hopefully makes it possible to compute extensive results for a selection of datasets and NER models. Special attention will be given to Transformer-based Language models such as BERT as this is what the aforementioned saliency methods have been developed for. If feasible, other models might also be evaluated.

The main implementation task will be to implement the chosen attribution algorithms mentioned the previous section. Post-hoc methods will be favored as they don't require the use of backward passes but optionally it could be made possible to use the resulting saliency information within model training. Furthermore, the metric for faithfulness evaluation, comprehensiveness and sufficiency, need to be computed within the framework.

Experimentation can also be done with both top-k (i.e. a list of the k most salient tokens) and continuous k-span (i.e. a continuous sub-string of the length k within the input) rationales as well as with different values for the parameter k. Additionally, It is necessary to provide a visualization for the resulting rationales which will most likely be a heatmap that highlights the saliency of individual tokens in a given input sequence.

As mentioned above, the focus is on biomedical NER datasets that are already a part of HunFlair.

References

[1] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *CoRR*, abs/1708.08296, 2017.

[2] Felix Friedrich, Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Interactively providing explanations for transformer language models. 2021.

[3] Aaron Chan, Maziar Sanjabi, Lambert Mathias, Liang Tan, Shaoliang Nie, Xiaochang Peng, Xiang Ren, and Hamed Firooz. Unirex: A unified learning framework for language model rationale extraction. *ArXiv*, abs/2112.08802, 2021.

[4] Jay DeYoung, Sarthak Jain, Nazneen Rajani, Eric P. Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. Eraser: A benchmark to evaluate rationalized nlp models. *ArXiv*, abs/1911.03429, 2020.

[5] Bill Yuchen Lin, Dong-Ho Lee, Minghan Shen, Ryan Rene Moreno, Xiao Huang, Prashant Shiralkar, and Xiang Ren. Triggerner: Learning with entity triggers as explanations for named entity recognition. In *ACL*, 2020.

[6] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *ArXiv*, abs/2004.03685, 2020.

[7] George Chrysostomou and Nikolaos Aletras. Flexible instance-specific rationalization of nlp models. 2021.

[8] George Chrysostomou and Nikolaos Aletras. Enjoy the salience: Towards better transformer-based faithful explanations with word salience. *ArXiv*, abs/2108.13759, 2021.

[9] Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. Learning to faithfully rationalize by construction. In *ACL*, 2020.

[10] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2014.

[11] Misha Denil, Alban Demiraj, and Nando de Freitas. Extraction of salient sentences from labelled documents. *ArXiv*, abs/1412.6815, 2014.

[12] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *ArXiv*, abs/1706.03825, 2017.

[13] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *ArXiv*, abs/1703.01365, 2017.

[14] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. *Layer-Wise Relevance Propagation: An Overview*, pages 193–209. Springer International Publishing, Cham, 2019.

[15] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[16] Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure. *ArXiv*, abs/1612.08220, 2016.

[17] Ákos Kádár, Grzegorz Chrupała, and A. Alishahi. Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43:761–780, 2017.

[18] Nina Pörner, Hinrich Schütze, and Benjamin Roth. Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement. In *ACL*, 2018.

[19] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, 2019.

[20] Leon Weber, Mario Sänger, Jannes Münchmeyer, Maryam Habibi, Ulf Leser, and Alan Akbik. HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics*, 37(17):2792–2794, 01 2021.