Data Corruption Impact on Named Entity Recognition for Low Resourced Languages

Anonymous
Anonymous
anonymous

Abstract

1 Introduction

(1 column)

- Data acquisition in low resource languages is not th best
- Therefore we may get errors on the data
- These languages are some times not collected by experienced speakers as experienced speakers will tend bo either verbal and based in rural areas.
- This therefore induces the need to train them which can be cost ineffiecient.
- This induces the need quantify the impact that such errors has on our models
- To perfrom this we choose to design different corruption that reduces the quality of NER datasets
- mentoin NER
- Massiv correlation between the quantity and the quality of the data
- Because we then look at corruption at a token len. Since NER is a token classification, annotators usually have to classify a span of words from a sentence into different catergories
- Therefore a situation where the annotors misclassifies a token is not far from happening.
- Therefore we look at the performance of the model and the certainty of its prediction
- Explain why look at these two metrics
- Then talk about about the contributions of the paper
- Breifly give the results obtained
- Finally, talk about the outline of the paper

2 Background

(1 column)

- Talk about dataset for low resource language

- Methods to construct them (acquisition, annotations, cleaninig)
- Talk about these datasets from a point of view quantity vs quality
- Talk about NER and different methods

3 Approach

(2 columns)

3.1 Data Corruption Strategies

- Having two paragraphs where the first explains quantity based corruption
- Qualty based corruption als explainde (remove sentences)

3.2 Uncertainty Prediction

Describe the metrics used to measure the encertainty of a prediction

- entropy of softmax
- don forget to mention the simplicity vibe

4 Experimental Setup

(1 column)

- We choose NER due to the simplicity of designing corruption strategies. Also to due the availability of datasets
- Indeed, for translation, it is unclear how the quality of data sets can be altered besides altering the target sentence.
- How ever there are virtually infinite ways to build target sentences different from source sentences with varying level of similarity
- However this makes it difficult to design an unbaised metric that quantify the level fo corruption in the data based its semantic property.
- This is why we focus on NER.

4.1 NER Corpora

4.2 Models

5 Results and Discussions

(2 columns)

6 Conclusion

(0.5 column)

7 Ablations

- lowercase factors seems to affects to these models. Would be nice

References