

# 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 the best
- Therefore we may get errors on the data
- These languages are some times not collected by experienced speakers as experienced speakers will tend to be either verbal and based in rural areas.
- This therefore induces the need to train them which can be cost inefficient.
- This induces the need quantify the impact that such errors has on our models
- To perform this we choose to design different corruption that reduces the quality of NER datasets
- 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.
- However 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 unbiased metric that quantify the level of corruption in the data based its semantic property.
- This is why we focus on NER.
- Because we then look at corruption at a token level. Since NER is a token classification, annotators usually have to classify a span of words from a sentence into different categories
- Therefore a situation where the annotators 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
- Briefly 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, cleaning)
- Talk about these datasets from a point of view
- Talk about NER and different methods

## 3 Approach

(2 columns)

## 4 Experimental Setup

(1 column)

### 4.1 NER Corpora

### 4.2 Models

### 4.3 Evaluation Metric

## 5 Results and Discussions

(2 columns)

## 6 Conclusion

(0.5 column)

## References