Named Entity Recognition: initial analysis

Lodi Dodevska and Viktor Petreski

1 Introduction

Named entity recognition (NER) is one of the first steps in the information extraction process. As the name suggests, the main task is to locate the named entities that are mentioned in the text and classify them into different groups. The entities give us information about the people, locations, organizations, time expressions etc. that are mentioned in the document and they help us to understand the text better. That is exactly why named entity recognition is crucial part in many fields of the natural language processing, such as text summarisation and automatic question answering.

2 Related work

Named entity recognition is a very developed branch of natural language processing. It can be split in two groups: generic (tries to localize names of people, organizations, locations etc.) and domain-specific (mostly used in the pharmaceutical, biological or medical research area, for extracting proteins, genes etc.). In this assignment we will focus on the first category. We checked a few papers in order to familiarize with the existing methodologies. The algorithms that are used can be roughly divided in four groups:

- Rule-based approach: follows a set of predefined rules to extract the named entities from a text. The disadvantage of these methods is that the rules must be created manually, based on the domain we are researching (Kim and Woodland, 2000).

- Supervised learning models: such as Hidden Markov Models, Support Vector Machines (Isozaki and Kazawa, 2002), Decision Trees or Conditional Random Fields. An interesting example of CRF is (Finkel et al., 2005) which explains the baseline for Stanford Named Entity Recognizer. These models are trained on an annotated corpus and they obtain fairly good results when used in the right domain, but their drawback is that they are corpus-dependent and there may not be available corpus for each domain we would like to explore.

- Unsupervised learning models: do not depend on an annotated corpus. They aim to find hidden principles or patterns that represent the data appropriately. An example of an unsupervised model is Google's BERT, which was pretrained on a large text corpus from Wikipedia.

- Deep learning approach: most of the successful state-of-the-art NER models are focused on this area. (Yadav and Bethard, 2019) gives a nice overview of deep neural network architectures and highlights some recent improvements.

3 Dataset

For initial analysis we chose Groningen Meaning Bank (GMB) dataset, which is a freely available annotated corpus of texts mostly used for named entity recognition POS tagging tasks.

The dataset is given in csv format. We used Pandas DataFrame from python for reading and analyzing it. The dataset contains 25 columns, but not all of them are going to be useful for our work in the future. Out of 25 columns, we will probably keep only *sentence_idx*, *pos*, *word*, *lemma* and *tag* for further usage. One row in the dataset corresponds to one word from the text.

First we had to deal with NaN values, or at least we thought. Since the dataset is preprocessed to some extent, there was only one erroneous line

which contained only NaN values, which was a surprise to us. After we removed the line, there were 1050794 rows left. Then we removed the duplicate sentences. After that there are 768956 rows (sentences) left and 30172 total unique words.

Next, we explored the tag column, which contains the appropriate named entity tag for each token. The following types of entities are covered:

1. *geo* = Geographical Entity
2. *org* = Organization
3. *per* = Person
4. *gpe* = Geopolitical Entity
5. *tim* = Time indicator
6. *art* = Artifact
7. *eve* = Event
8. *nat* = Natural Phenomenon

The entities are tagged using IOB format, which means that there are prefixes I and B before a tag. B indicates that the tag is at the beginning of a chunk, I denotes that the tag is inside of a chunk. If the token does not belong to any of the aforementioned entities, its corresponding tag is O. The total values from each type of entity are shown in Table 1.

Entity	Total
geo	32969
org	27136
per	24991
tim	19517
gpe	11989
art	478
eve	433
nat	205

Table 1: Number of appearances of each entity type

We can see that the tags are not uniformly distributed. We will have to take that into account when we will analyze the performance of the NER model that we will use in the future and identify the number of false positives carefully.

We are also trying to obtain the CoNLL-2003 dataset which is free, but we are missing the Reuters Corpus. As soon as we get it we will try to incorporate this dataset too. Also, we would like to try and work with dataset in our native language - Macedonian, but so far we did not have any luck in finding such a dataset.

4 Methods

We have researched the latest approaches in named entity recognition area. These state-of-the-art models are based on deep learning approaches. Our idea for this project is to implement some of these models, try to tune their hyperparameters and then compare their performances and analyze their strengths and weaknesses. One idea that can be implemented is the Bidirectional LSTM + CRF published in (Huang et al., 2015). Even though LSTM neural networks have outperformed the Conditional Random Fields in the NER field, CRF are still used in combination with LSTM in order to allow the model to consider the context of the token and the tag of the previous word to predict the current word. Another approach would be to try and use BERT and/or Flair as one of the most prominent NLP libraries and finetune them based on our dataset.

5 Conclusion

This is just a brief summary of our initial analysis in the area of Named Entity Recognition. We are aware that there is a lot of work to do, but we are willing to explore furthermore as we are developing our project.

References

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, page 363–370, USA. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Hideki Isozaki and Hideto Kazawa. 2002. Efficient support vector classifiers for named entity recognition. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, page 1–7, USA. Association for Computational Linguistics.

Ji-Hwan Kim and Philip C. Woodland. 2000. A rule-based named entity recognition system for speech input. In *INTERSPEECH*.

Vikas Yadav and Steven Bethard. 2019. A survey on recent advances in named entity recognition from deep learning models.