Distant Learning on Wikipedia Corpus for NER Annotation

Submitted as part of the requirements for

CE-807-Text Analytics

Task: 1

Name: Siddharth Shyamsunder

Lecturer: Prof. Ansgar Scherp

Submission Date: 25thth February 2019

Contents

[INTRODUCTION 3](#_Toc1978153)

[BASIC TERMS: 3](#_Toc1978154)

[BACKGROUND: 4](#_Toc1978155)

[CRITICAL DISCUSSIONS 4](#_Toc1978156)

[CONCLUSION: 5](#_Toc1978157)

[REFERENCES 5](#_Toc1978158)

# INTRODUCTION

Named Entity Recognizer for the English language makes use of different Gold Standards, the Message Understanding Conferences(MUC), the Conference on Computational Natural Language Learning(CoNLL) and most recently the Wikiner or the wikipedia corpuses. The former 2 standards were manually annotated, however this led to complexity in the generation of the dataset, as manually annotating each information was done by a huge term of experts which makes it very time expensive and labor costly. The plan that was incorporated to remove these additional cumbersome costs was to have a group of experts manually annotate or label some part of the data, and what is done next is to have this data be a training model to predict the value of the unseen data. Doing this is called Distant supervised Learning. Although this system seems analogous to that of a traditional supervised Learning, the subtle difference is the accuracy of validating the predicted data on the manually annotated data.

What the main aim to accomplish in this task is to decide which distant learning technique would correctly classify from a previously manually annotated data into an unseen data. When considering manually annotated data, lots of additional features from the annotated data like the Parts of Speech Tag, the Named Entity Recognition factor plays an important role, and so it is of utmost importance to evaluate a predicted model from a previously annotated model to identify the accuracy of the distant Learning model.

# BASIC TERMS:

Named Entity Recognition:[4] The Named Entity Recognizer basically is used to classify text as per their specific entity which could be a Person, or an organization or a Location. When we consider NER for the English Language, we typically make use of 3 models the MUC, CoNLL or BBN and mot recently Wikipedia.

Distant Supervised Learning:[3] When we generally consider a corpus for a language, we manually annotate the salient features of each word so as to perfect our model to any application. However this is a cumbersome task due to the influx of information that keeps coming into any language. Distant Supervision makes use of an existing labeled data to make a prediction of unseen label data, even though the corpus as such has many unique cases. Although Distant Learning can definitely be an approach to handling information, its known to be prone to noisy data.

# BACKGROUND:

When it comes to Named Entity Recognition, Statistical methods have proven to be highly successful methods making use ff the more refined parts of language like as the orthographic, linguistic, contextual and external features. But however successful previous experiments using statistical methods were, the time taken for manually annotating corpuses was enormous and labor costly, so experiments were carried out to have a more adaptable model called Distant Supervision Learning, wherein researchers like Nothman[1] from the University of Sydney, would make use of a previously annotated model to test the adaptability measure of predicting new information. Testing of one model on the others resulted in poor interrelation between models, so further research in the area resulted in the fact that interdependencies in the word structure such as tokenization, POS tag alignment played a major role in the performance of an NER system. So they tested the Curran and Clark model on the gold standard and an F Measure increase of upto 11% was observed.

Another performance standard that was considered by Mikheev[2] in his paper “Named entity recognition without gazetteers”, was that the size of the corpus did not necessarily mean an increase in the performance of the prediction of the NER System and that the Distant Learning measure focused more on the interrelation between patterns found in the manually annotated set. Another measure that was considered was the performance measure, of the model used to do the training, because, every model had its own ways of handling heading data and other entity labels and formats of data, like the CoNLL holds all caps in headlines, and the MUC, could only classify the text as person, organization and location. After looking further into the models used for evaluating, the information generated for each model was as such.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Corpus | # tags | Number of tokens | | |
|  |  | TRAIN | DEV | TEST |
| MUC-7 | 3 | 83601 | 18655 | 60436 |
| CoNLL-03 | 4 | 203621 | 51362 | 46435 |
| BBN | 54 | 901894 | 142218 | 129654 |

Also based on the model another parameter that was to be taken into consideration was the type of relationship agreed by the sentences in the corpora, in other words in the case of CoNLL model, there was no direct interrelation with the words so reducing the number of grams would result in a higher accuracy, whereas the MUC and the BBN Model would be a boundary specific model. In such way many features contributed to the evaluation standard of the NER Method.

# CRITICAL DISCUSSIONS

Further analysis on the 3 models tells us, how our aim to provide an efficient method for making use of Machine Learning to automatically identify patterns in a few annotated labeled datasets can bring about a much higher accuracy through many factors. First and foremost, we agree with the fact that the alignments of the words play an important role in the determining of the output. In other words, having considering a unigram relationship rather than a bigram relation ship can cause a drastic change in the evaluation result of our system.

As for what we have considered in terms of Mikheev’s statement that the size of the corpus does not depend upon the performance measure of the evaluation, we can agree upon this with a set of limitations. As we train the corpus, to handle huge number of data, initially there might be a fluctuation on the accuracy, but bringing together a new entity will cause a sudden improvement in the accuracy of the system. Furthermore, we believe using a measure of uniqueness and pattern recognition in obtaining the predicted labels, we can receive a good precision unit, and prevent the concept of overfitting which would be a major issue if we did have to manually train a dataset for a huge corpus.

Also if one can find the interlink between languages, the concept of language dependency can be easily avoided. Annotating a billion words from one language to another would be difficult if using a manual annotation method. One particular that might be faced, is as Nothman[1] discussed upon using other features like the POS Tagger tokenization, lemmatization and other preprocessing techniques to handle.NER, this plays an important role in identifying patterns as sentence structure plays a key role in identifying the difference between say person and a month for example. Considering an example of June, this could be a person’s name under a specific context of sentence structure, but this could also be the case of a month, like the month of June.

All these factors must be considered while playing an important role in evaluating the measure of performance of the system.

# CONCLUSION:

On critically analyzing many of the features we can conclude that Distant Learning proves to be a much better approach in classifying NER for Wikipedia dataset. Making use of an efficient algorithm, can also help with boosting the evaluation factor of the algorithm. Performance measure does not depend on the size of the dataset.

# REFERENCES

[1] J. Nothman, T. Murphy, J.R. Curran, Analysing Wikipedia and gold-standard corpora for NER training, in: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, 2009, pp. 612–620.

[2] Andrei Mikheev, Marc Moens, and Claire Grover.1999. Named entity recognition without gazetteers.In Proceedings of the 9th Conference of the European Chapter of the Association for Computational

Linguistics, pages 1–8, Bergen, Norway.

[3]Distant Supervised Learning(Stanford).[Online]. Available: <http://deepdive.stanford.edu/distant_supervision> [Accessed:February 19th 2019].

[4] Named Entity Recognizer).[Online]. Available: <https://en.wikipedia.org/wiki/NER>[Accessed:February 19th 2019].