Machine Learning for Named Entity Recognition Report

Ankur Satya

Abstract

Fill in a short abstract for your final submission.

Referring to literature

1 Introduction

Write a short introduction to your approach for the final submission. The introduction should include:

- A brief description of the task (not detailed)
- An outline of the ML approaches included in the report
- A brief summary of your results

Final submission

2 Related work

Assignment 2

3 Task and Data

3.1 Task

For the given dataset and the problem, Named Entity recognition(NER) involves the locating and classifying the following entities:

- persons(PER): name of people like Ankur
- organisations(ORG): name of organisations like Netflix
- locations(LOC): locations like France
- miscellaneous names(MISC): other names that don't fit in the first 3 categories.

The IOB tagging scheme is used which was put forward by Ramshaw and Marcus (Lance A. Ramshaw (1995)). This scheme tags entities using the following format: I-XXX where XXX refers to the entity tag like I-ORG, I-PER.

The data is represented in the *conll* format. Every line in the data file represents information about a single word i.e a single entity. There are four fields in every line: the word, the associated part-of-speech tag, associated chunk tag and the named entity tag according to the IOB tagging scheme Sang and De Meulder (2003). Since the data collected is in the form of articles etc, it contains sentences so to show the sentence boundary, an empty line is used in the data files.

3.2 Dataset and Distribution

Assignment 1 Data description and distribution

Assignment 2 update if necessary

Assignment 3 update if necessary

3.3 Preprocessing

Assignment 1 Preprocessing you did for first analysis

Assignment 2 update (if applicable)

Assignment 3 update (if applicable)

3.4 Evaluation Metrics

Three metrics were used for evaluation, namely: precision, recall and f-score. Mathematically, they are defined as:

$$precision = \frac{true\_positive}{true\_positive + false\_positive}$$

$$recall = \frac{true\_positive}{true\_positive + false\_negative}$$

$$f - score = \frac{2 * precision * recall}{precision + recall}$$

We evaluated the above 3 metrics for all the classes present in the used dataset.

4 Models and Features

4.1 Models

Assignment 1 Logistic Regression

Assignment 2 Alternative Methods (SVM, NB)

Assignment 3 More advanced models (fine-tuning BERT, CRF for ReMA students only)

4.2 Features

Assignment 1 Feature exploration

Assignment 2 More features

Assignment 3 More advanced features

5 Experiments and Results

5.1 Evaluation

Assignment 1: first results

Assignment 2: update

Assignment 3: update

5.2 Feature Ablation

Assignment 3: feature ablation for one system

6 Error Analysis

Assignment 3: beyond the confusion matrix

7 Discussion

Assignment 3

8 Conclusion

Assignment 3

Acknowledgements

References

Mitchell P. Marcus Lance A. Ramshaw. 1995. Text chunking using transformation-based learning. *Third ACL Work- shop on Very Large Corpora*, pages 82–94.

Erik Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

A Time spent

Please use Table 1 to give an overview of the time you spent on each submission.

Week	Task	Time
1	watch videos	1 hour 30 mins
2	understand the problem and data(assignment 1)	40 minutes
2	working on section 2 of assignment 1	2 hours 30 minutes
Total	4 hours 40 minutes	

Table 1: Time overview.