CS 6320: Natural Language Processing

Project 2: Named Entity Recognition (60 + 25 points)

Due date: 11/15/2019 04:00 pm

1 Introduction

For Project 2, you will perform named entity recognition (NER) using a feature-driven machine learning approach; and deep learning (optional).

2 Problem Definition and Dataset

Named Entity Recognition (NER) refers to the task of extracting entities from text and tagging them with labels. For example, a typical NER system would find and tag entities in the sentence "Jim bought 300 shares of Acme Corp. in 2006." as:

Jim → Person; 300 → Number; Acme Corp. → Organization; 2006 → Date

For this project, you will use the CoNLL 2003 dataset [1] to perform NER tagging. This dataset defines 4 tags: Person, Location, Organization and Miscellaneous. Your task is two-fold: given a sentence, extract all relevant named entities from the task; and assign each named entity with a relevant tag. To perform both tasks together, NER is typically modeled as a BIO tagging problem: where B indicates the beginning of a new entity span, I indicates the continuation of an entity span and O indicates that the current word or token does not belong to an entity span. So for the previous example, the NER tags would be represented as:

```
Jim     B-Person
bought  O
300     B-Number
shares  O
of      O
Acme    B-Organization
Corp.   I-Organization
in      O
2006    B-Date
```

Subsequent sections will provide a detailed break-down of the tasks to be carried out for this project.

3 Task - 1: Feature extraction and representation (25 points)

In the dataset, sentences are represented in the CoNLL form, where each new line indicates the beginning of a new sentence. Within each sentence, tokens are also new line-separated; each token associated with its gold NER tag. Each line also contains some additional information like the gold POS tag and the syntactic head for each token: you may disregard this information and instead for each token in a sentence, you will extract the following features: the part-of-speech tag, the lemma, all hypernyms, hyponyms, holonyms and meronyms; and any additional features that you can think of (please document these features in your report)

Once you have extracted all features, use the bag-of-words model to represent the features for each sentence.

4 Task - 2: Extracting and tagging named entities (25 points)

Once all relevant features have been extracted (from Task 1), use the Logistic Regression model to extract all named entities from text and tag them as Person, Location, Organization or Miscellaneous. Report the accuracy of classification, precision, recall and F-score for the test set. Also, report the precision, recall and F-scores for each tag.

Please document any/all design decisions made while carrying out this task in the project report.

5 Optional Task - 3 : NER using deep learning (25 points)

NER can also be carried out using deep learning. The steps are same as the ones for sentiment classification (see Project 1):

1. Design a corpus reader and embedding reader. You will use the same pre-trained word embedding file that you used in the previous project.
2. Design a simple deep learning model that contains the following:
    (a) A word embedding layer that takes in integer-encoded representations of each token and finds its word embeddings.
    (b) A bi-directional LSTM that takes in the output of the embedding layer and generates a representation to be fed to the next layer.
    (c) A fully connected layer that takes in the output of the LSTM and outputs a tag for each word (N.B.: This is different from Project

1 where you were outputting one tag for a sentence; here you will output a tag for each word in the sentence)

3. Choose a suitable loss function, optimizer, batch size and learning rate.
4. Report the accuracy of classification, precision, recall and F-score for the test set. Also, report the precision, recall and F-scores for each tag.

6 External links

Here are some external links for you to check out.

1. https://pytorch.org/docs/stable/index.html
2. https://medium.com/@rohit.sharma_7010/a-complete-tutorial-for-named-entity-recognition-and-extraction-in-natural-language-processing-71322b6fb090
3. https://towardsdatascience.com/named-entity-recognition-and-classification-with-scikit-learn-f05372f07ba2
4. https://dkedar7.github.io/Named%20Entity%20Recognition.html

NOTE: All project files are available at: http://www.hlt.utdallas.edu/~takshak/project_2.zip

You are given the training and test data, and a pre-trained word embedding file. You are free to use any third-party tool to perform aforementioned tasks. A complete list of these tools is available on the class webpage.

What to submit:

1. All your source code bundled into a zip file with a README giving clear instructions on how to run the code. If you have used any external packages, please provide instructions on how to install those packages.

2. A PDF report detailing your project, and results that you obtained. This report is worth **10 points**. A printout of the project report with the evaluation sheet is due in class on 11/15/2019 at 4pm.

3. If you are unable to complete any of the mentioned tasks, you can include some details on the things you tried out for that task. You MAY receive some partial credit for trying ☺.

References

[1] Erik F. Tjong Kim Sang and Fien De Meulder. 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, 2003.