Machine Learning for NLP Assignment 1

Deadline: November 11th 2022

This first assignment consists of a small theoretical component and the first preparations for building and evaluating systems for Named Entity Recognition. The recommended time division is that you complete the first two sections in the first week and the third section in the second week. All code needed for this assignment can be found under code/assignment1 in the github repository: `https://github.com/cltl/ma-ml4nlp-labs`

Throughout this course, you will be working on three different components:

- Theoretical components
- Experimental report: please use the Overleaf template [tba]
- Practical component: Your code

Each template has a field for specifying how much time you spent on each component in your submission. We will use this information to monitor your progress and provide support when needed. For each section of the Assignment, we specify what you are supposed to submit. You will work on all three components throughout the course. For each Module, you will submit an updated version.

1 What is Machine Learning

Provide an explanation of your own view in a short paragraph. If you take a definition from someone else, explain why you think it is a good definition or how you would alter it to make it a good definition. You do not need to do research to answer this question: your current understanding of machine learning is sufficient.

Submission: Please provide your answer in the theoretical component of the assignment.

2 Preparation Experimental Setup

Each NLP experiment or NLP system development starts by getting a good understanding of the task and ensuring you have the right materials for developing your system (or for doing your research). This means you need to get an understanding of the goal of the task and the data and

you need to make sure you have the means to evaluate your outcome. This exercise will guide you through an exploration of the named entity recognition task, a few (basic) data exploration steps and a basic setup for a standard evaluation.

2.1 Understanding the task and the data

In Natural Language Processing, it is often worthwhile to explore the data, i.e. look at some examples and snippets. We generally use data structures that are intuitive to understand: you can either figure out how they work by looking at them or by finding information online. We will look at the training and evaluation data provided through the conll shared task of 2003 (Sang and De Meulder, 2003) (Section 1 and Section 2).

In this case, there two things you need to figure out:

1. What is the exact task of Named Entity Recognition in the specific dataset you are working with? To give you some guidance, please focus on the following aspects:
    - What categories are used? What do they represent?
    - What annotations are used? What conventions do they follow?
2. How does the representation format (conll) work? To give you a bit of guidance, please consider the following aspects:
    - What do the lines in the files represent?
    - What do empty lines mean?
    - What do the four fields in one line mean?

You can probably figure both of these questions out by carefully studying the development data and having a look at Sang and De Meulder (2003). You can use python scripts to explore that data format, extract the classes that are used, etc.

Submission: Please provide the answers to the question in Section 3.1 of the **report component**.

2.2 Evaluation

We can now complete our experimental setup with basic evaluations (for now). You will program functions that can provide precision, recall and f-score as well as a confusion matrix. For this specific assignment, you should provide these metrics and the confusion matrix **without making use of external modules**: i.e. your code should include the actual calculations.

1. Go through the notebook basic_evaluation.ipynb. The notebook provides examples of how to read data in using Pandas, how to output tables and how to test code using assert statements. It also contains tests using the mini-datafiles provided on github.

2. Complete this code adding functions that can compare system and gold results and provide precision, recall, f-score and a confusion matrix. For now, please simply use the mini-datafiles.

3. Run the tests to see if your code works properly.

You can change the structure of the code or the provided functions in any way you see fit. Just remember that you must program the code that carries out the calculations yourself and you should not rely on external modules for this (e.g. by scikit-learn). The results will be included in your final report.
Submission:

- Practical component: code
- Report: briefly describe the evaluation you carried out (brief explanation of the metrics and how you apply them to the NER data). (Section 3.4 Evaluation)

2.3 Data Exploration

Before you start training a machine learning model, it is always a good idea to understand the distribution of your data in your training and development set.

Please write code to answer the following questions:

1. What is the distribution of the NER labels? How many instances do you have per class? What class is best represented? For which class do you have the least amount of data?

2. What features could be informative for the task? Formulate hypotheses about linguistic (or orthographic) features and test them by exploring their distribution in the data. Please explore at least three different aspects. For example, you can look into the following:

    - Use the linguistic information already included in the dataset
    - Look into orthographic features
    - Explore the most common words per category

You can use NLP tools (e.g. NLTK, Spacy) in your analysis.

Submission:

- Code:
    - A script analyzing the label distribution in the data (`analyze_distributin.py`:
        * Input argument: path to the respective data split (only use training and development)
        * Output: A csv file representing the distribution of the labels
    - A script or notebook analyzing the potentially informative features.

- Report:
  - Please summarize your findings in the following sections of the report:
    * Section 3.2 Dataset and Distribution
    * Section 3.3 Preprocessing (if you used tools)
    * Section 4.2 Features: Your exploration of potentially informative features

3 A Basic System (first results)

In this final component, you will train a basic named entity recognition system and evaluate it on the development data. We will start with a simple logistic regression system that makes use of three simple features only. One feature (the token itself) has already been implemented for you.

1. Examine the code provided in basic_system.ipynb and check out the documentation mentioned in the notebook.
2. Provide documentation to the code. Make sure to ask questions on Piazza if there are things you do not understand: you will need to extend this code next week.
3. Implement two additional features (ideally based on your analysis).
4. Provide the correct arguments and train your system.
5. Evaluate your system using your evaluation code.

Please submit the following:

- Code
- Report:
  - Section 4.1: A brief description or the logistic regression model (not detailed, can be a couple of sentences)
  - Section 5.1 Results of the evaluation

4 Submission Assignment 1

4.1 Recommended pre-submission week 1

We highly recommend to submit the following components of Assignment 1 by the end of week 1:

- Section 1: What is Machine Learning)
- Section 2: Preparation Experimental Setup

  - finish 2.1
  - finish 2.2
  - start 2.3

Please submit:

- Theoretical component
- Report
- Practical component

4.2 Compulsory submission week 2

Submit all components of Assignment 1. Your submission should include:

- Theoretical component
- Report
- Practical component

4.3 Format of the practical submission (code)

A .zip or .tar.gz file of a folder with your student id as its name. This folder should contain any code you have worked with and written for this first assignment, with a readme explaining how to run things, if necessary (this is currently not necessarily if you simply completed the notebooks provided).

References

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.