Assignment #1: Named Entity Recognition

Dr. Zornitsa Kozareva
USC Information Sciences Institute
Spring 2013

Task Description: You will be given three data sets total. First you will receive the train and development data sets, which will have on each line the word, its automatically generated part-of-speech tag and the correct named entity class. The human annotators have used 9 classes (B-PER, I-PER, B-LOC, I-LOC, B-ORG, I-ORG, B-MISC, I-MISC and 0) to indicate the **B**eginning of a named entity, the **I**nside of a named entity and the **O**utside of a named entity. You will see that each major BIO tag is followed by the corresponding named entity category. For instance, the tag B-PER indicates the beginning of a person name, I-PER indicates inside a person name, and so forth. LOC stands for Location, ORG stands for organization, MISC stands for miscellaneous.

Your objective is to build a machine learning named entity recognition system, which when given a new previously unseen text can identify and classify the named entities in the text. This means that your system should annotate each word in the text with one of the nine possible classes.

To achieve this goal, you must use the train data to design the features you believe are relevant for the identification of the named entities. You can design and test different sets of features, you can experiment with 10-fold cross validation and feature selection until you find the feature set that gives you optimal performance. You can also use the development data to track the performance of your system on data different than the training one. You can repeat this process multiple times. You will have 10 days to work on the feature development and system engineering part.

On 7th of February 2013, we will provide you with the final test data set, which will contain on each line only the word and its automatically generated part-of-speech tag. **Note that this data set will not contain the correct named entity classes** because this is what your system must generate.

What to turn in:

- The predicted classes from your system on the test data.
  Important: the final output on the test data must contain for each word (i.e. each row) only the named entity tag predicted by your system. The empty lines in the data indicate sentence and document boundaries. These must be preserved in order for the evaluation scorer to run correctly.
- The official train and test feature files used in the final run
- Your source code with a readme explanation on how we can run your code to check that it is working. Make sure there are comments in your code.
- The gazetteer and trigger lists, which you have generated/collected
- A brief description of your system explaining:
  - used external tools including PoS tagger, parser, chunker etc.
  - designed features
  - used machine learning algorithm
  - results on 10-fold cross validation
  - results on train data
  - results on development data
  - information on the feature selection algorithm (if used)

Timeline:

	Release
Train&Development Data	**January 29th 2013**
Test Data	**February 7th 2013**
Result Submission Deadline	**February 8th 2013 (11:59 pm California time)** **later submissions will not be accepted**
Technical Report Deadline	**February 8th 2013**

Evaluation is based on the:

- Ranking of your system against the rest of the systems in the class
- Designed features:
  - **novel** and previously unexplained features will be favored
  - system's pre or post processing
  - a study on feature selection and data analysis
- **Generated resources**:
  - size, methods and sources for gazetteer extraction
  - trigger lists
- Quality of the technical report:
  - **error analysis**

IMPORTANT: The use of existing named entity system(s) or libraries either as a feature generator, gazetteers, output generator, pre-/post processors among others IS NOT ALLOWED and STRICTLY FORBIDDEN.

Where should I start?

- Use the train and development data to design, build and tune your named entity system
- Decide on the features you would like to use, gather all the resources you need for their generation
- Choose a machine learning classifier from Weka
  - http://www.cs.waikato.ac.nz/ml/weka/
  - Intro by Marti Hearst http://courses.ischool.berkeley.edu/i256/f06/lectures/lecture16.ppt
- For each experiment, evaluate the performance of the system by running the evaluation script in the following way:

  perl eval.txt < input_file

  where the input_file has to have the following format

  word1 gold_standard_named_entity_tag predicted_named_entity_tag
  word2 gold_standard_named_entity_tag predicted_named_entity_tag
  …
  wordn gold_standard_named_entity_tag predicted_named_entity_tag

You will obtain a summary of the results indicating how well your system performed for each one of the PER, LOC, ORG and MISC classes

processed 46666 tokens with 5648 phrases; found: 5620 phrases; correct: 5001.

accuracy: 97.63%; precision: 88.99%; recall: 88.54%; FB1: 88.76

LOC: precision: 90.59%; recall: 91.73%; FB1: 91.15 1689

MISC: precision: 83.46%; recall: 77.64%; FB1: 80.44 653

ORG: precision: 85.93%; recall: 83.44%; FB1: 84.67 1613

PER: precision: 92.49%; recall: 95.24%; FB1: 93.85 1665

Then you can design new features or tune the already existing ones in order to improve the current performance of your system.

This is a big assignment start early!

Generating/Collecting Your Own Resources (brings you extra points):

- Gazetteer entries from Wikipedia:
  - extract names of people like singers, teachers, scientists
  - extract names of locations like cities, countries
  - extract names of organizations like universities, IT companies
- Trigger words

Analysis (brings you extra points):

- How much data do we need to reach stable performance?
- Extract and rank the patterns in which the NEs occurred in the train and development data. Show what percentages of these were found in the final test data.
- Extract lists of verbs found next to the NEs. Do you find any similarity/regularity of the verbs associated with each one of the NE categories?

Available Resources:

- WordNet http://wordnet.princeton.edu/
- Part-of-speech taggers:
  - TreeTagger http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
  - Stanford PoS Tagger http://nlp.stanford.edu/software/tagger.shtml
- NP chunkers:
  - http://www.dcs.shef.ac.uk/~mark/index.html?http://www.dcs.shef.ac.uk/~mark/phd/software/chunker.html
- Parsers:
  - Stanford Parser http://nlp.stanford.edu/software/lex-parser.shtml
- Other
  - http://nlp.stanford.edu/links/statnlp.html

Papers that might help you

- Jenny Rose Finkel and Christopher D. Manning. 2009. Nested Named Entity Recognition. *Proceedings of EMNLP-2009* http://www.stanford.edu/~jrfinkel/papers/nested-ner.pdf

- Xavier Carreras, Lluís Márques and Lluís Padró, Named Entity Extraction using AdaBoost. In: *Proceedings of CoNLL-2002* http://www.cnts.ua.ac.be/conll2002/pdf/16770car.pdf

- Radu Florian, Named Entity Recognition as a House of Cards: Classifier Stacking. In: *Proceedings of CoNLL-2002* http://www.cnts.ua.ac.be/conll2002/pdf/17578flo.pdf

- Silviu Cucerzan and David Yarowsky, Language Independent NER using a Unified Model of Internal and Contextual Evidence. In: *Proceedings of CoNLL-2002* http://www.cnts.ua.ac.be/conll2002/pdf/17174cuc.pdf

SAMPLE FORMAT TECHNICAL REPORT FOR ASSIGNMENT #1

Name of system:
Name of student:
Student ID:
Date:

1. System Description:
    1.1. Used tools, available resources and/or those newly generated by you like gazetteers, trigger words
    1.2. Feature set description

    feel free to show a graphic with the system architecture

2. Results with:
    2.1. Machine Learning Algorithm 1 on
        - train data
        (on 10-fold or feature selection if any of these are used)
        - development data
    2.2. Machine Learning Algorithm 2 on
        - train data
        (on 10-fold or feature selection if any of these are used)
        - development data

3. Conclusions and Observations: