Homework: Named Entity Challenge

Zornitsa Kozareva
Information Sciences Institute/University of Southern California
Spring 2010

The task: You will be given a train and development sets of English sentences that are tagged with four named entity categories: PER, ORG, LOC and MISC indicating people, organization, location and miscellaneous names respectively. Your objective is to build a machine learning named entity system which when given a new previously unseen text (i.e. test set) will identify and classify the named entities correctly.

	Release
Train and Development data	March 24th, 2010
Test data	April 9th , 2010
Result submission deadline	April 10th, 2010 (11:59 GMT) **later submissions will not be accepted**
Presentation submission deadline	April 13th, 2010

What to turn in:

- Source code for the feature generation (make sure it will run under Linux)
- Official train and test feature files used in the final run, and the final output of your system on the test data set.
  The final output on the test data set must contain for each word (i.e. each row) the named entity tag predicted by your system. Note, the empty lines in the data indicate sentence boundaries. These must be preserved in order for the evaluation scorer to run correctly.
- Additionally generated resources (if any)
- A brief 1-2 page description of your system explaining the:
  - used NLP tools such as PoS tagger, parser, chunker etc.
  - designed features
  - employed machine learning algorithm
- Make a short power point presentation, which you will present in 3 minutes to the class on April 15th . Please be prompt so I can include your slides in the set to be presented. Your presentation must be to the point given the short time window in which you will have to present

Evaluation is based on the:

- Ranking of your system against the rest of the systems
- Designed features:
  - novel and previously unexplained features will be favored
  - system's pre or post processing
  - a study on feature selection
- Generated resources:
  - size, methods and sources for gazetteer extraction

- trigger lists

Hints

Where should I start?

- Use the train and development data to design and tune your NE system
- Decide on the features you would like to incorporate, 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, you can 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

  $word_1$ gold_standard_named_entity_tag predicted_named_entity_tag
  $word_2$ gold_standard_named_entity_tag predicted_named_entity_tag
  …
  $word_n$ 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 so start early!

Generate/Collect 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 from WordNet:
  - pull hyponyms related to people, location, organization
- 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
- Named Entity Recognizer
  - Stanford NER http://nlp.stanford.edu/software/CRF-NER.shtml
  - LingPipe http://alias-i.com/lingpipe/
  - ANNIE http://www.aktors.org/technologies/annie/
- 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

- Jenny Rose Finkel and Christopher D. Manning. 2009. Joint Parsing and Named Entity Recognition. *Proceedings of NAACL-2009*
  http://www.stanford.edu/~jrfinkel/papers/joint-parse-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

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

1. Problem definition:

2. System description:
    2.1. Used tools, available resources and/or those newly generated by you like gazetteers, trigger words
    2.2. Feature set used for NED
    2.3. Feature set used for NEC
    2.4. Description of machine learning classifier

feel free to show a graphic with the system architecture

3. Performance on:
    3.1. Development set
        - Feature selection (if any)
        - 10 fold cross validation (if any)
    3.2. Test set

4. Conclusion: