CS544: Overview Assignment 2

February 21, 2012

Zornitsa Kozareva
USC/ISI
Marina del Rey, CA
kozareva@isi.edu
www.isi.edu/~kozareva

Assignment #2

• Build an automated system that can teach computers how to identify named entities from written texts:
  – Locations (Paris, California)
  – People (George Bush, Dr. Smith)
  – Organization (University of Southern California)
  – Miscellaneous
• **Given:** a train and development set of English sentences tagged with the following classes:

- B-PER (Begin-Person)
- I-PER (Inside-Person)
- B-ORG (Begin-Organization)
- I-ORG (Inside-Organization)
- B-LOC (Begin-Location)
- I-LOC (Inside-Location)
- B-MISC (Begin-Miscellaneous)
- I-MISC (Inside-Miscellaneous)
- O (Outside entity classes)

• **Your objective is:** to program the features for a machine learning classifier, build a model and use this model to identify and classify the same classes for a new previously unseen text (i.e. test set)

---

**Data Description**

• The data consists of three columns separated by a single space. Each word has been put on a separate line and there is an empty line after each sentence.

```
U.N. NNP B-ORG
official NN O
Ekere NNP B-PER
heads VBZ O
for IN O
Baghdad NNP B-LOC
... O
```

Make sure to preserve the empty lines in the output of the test data
Given Training Data Examples

<table>
<thead>
<tr>
<th>Example</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft</td>
<td>PERSON</td>
</tr>
<tr>
<td></td>
<td>ORGANIZATION</td>
</tr>
<tr>
<td></td>
<td>PERSON</td>
</tr>
<tr>
<td></td>
<td>LOCATION</td>
</tr>
<tr>
<td></td>
<td>ORGANIZATION</td>
</tr>
<tr>
<td></td>
<td>LOCATION</td>
</tr>
<tr>
<td></td>
<td>OTHER</td>
</tr>
</tbody>
</table>
What is a Features?

- A feature is a quantitative measurement of some data element
  - numerical
  - binary
  - categorical
  - dependency tree

Features for NER

Adam
Smith
Works
for
IBM
in
London
.

• **Contextual**
  - current word \( W_0 \)
  - words around \( W_0 \) in \([-3,\ldots,+3]\) window
Features for NER

Adam Smith Works for IBM in London

- **Contextual**
  - current word \( W_0 \)
  - words around \( W_0 \) in \([-3,\ldots,3]\) window

Features for NER

Adam Smith Works for IBM in London

- **Contextual**
  - current word \( W_0 \)
  - words around \( W_0 \) in \([-3,\ldots,3]\) window
Features for NER

- **Ortographic**
  - *initial-caps*
  - *all-caps*

```
Adam, null, null, null, Smith, works, for, 1, 0
Smith, Adam, null, null, works, for, IBM, 1, 0

fp, London, in, IBM, null, null, null, 0, 0
```

- **Word-Type Patterns:**
  - *functional*
  - *capitalized*
  - *lowercased*
  - *punctuation mark*
  - *quote*
  - *other*

- **Left Predictions**
  - the tag predicted in the current classification for W-3, W-2, W-1

*functional-word is preposition, conjunction, article*
Features for NER

- **Part-of-speech tag**
  - look mainly for proper names

<table>
<thead>
<tr>
<th>Tag</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN</td>
<td>preposition</td>
</tr>
<tr>
<td>JJ</td>
<td>adjective</td>
</tr>
<tr>
<td>NN</td>
<td>singular or mass noun</td>
</tr>
<tr>
<td>NNS</td>
<td>plural noun</td>
</tr>
<tr>
<td>NP</td>
<td>proper noun or part of name phrase</td>
</tr>
<tr>
<td>RB</td>
<td>adverb</td>
</tr>
<tr>
<td>VB</td>
<td>verb, base form</td>
</tr>
<tr>
<td>VBD</td>
<td>verb, past tense</td>
</tr>
<tr>
<td>VBG</td>
<td>verb, present participle/gerund</td>
</tr>
</tbody>
</table>

Features for NER

- **Trigger words**
  - for person (*Mr, Miss, Dr, PhD*)
  - for location (*city, street*)
  - for organization (*Ltd., Co.*)
Features for NER

- Gazetteers (big dictionaries of names)
  - geographical
  - first name
  - surname
  - company names

Gazetteer Collection Method 1

- Yago contains over 2 million entities (like persons, organizations, cities among others)

- Download Yago from:
  

- Extract from the relevant relations all named entities

  Ex.
  - $X$ born in $Y$, where $X$ is a person and $Y$ is a location
  - $X$ works for $Y$, where $X$ is a person and $Y$ is a person or an organization
Gazetteer Collection Method 2

Madonna (entertainer)

Madonna (born Madonna Louise Ciccone; August 16, 1958) is an American recording artist, actress and entrepreneur. Born in Bay City, Michigan, she moved to New York City in 1977, for a career in dance. After performing as a member of the pop musical groups Breakfast Club and Emmy, she released her first studio album, Madonna, in 1983 on Sire Records.

A series of hit singles from her next studio album, Like a Virgin (1984) and True Blue (1986), gained her global recognition. They established her as a pop icon, for pushing the boundaries of lyrical content in mainstream popular music and imagery in her music videos, which became a fixture on MTV. Her recognition was augmented by the film Desperately Seeking Susan (1985) which widely became known as a Madonna vehicle, despite her not playing the lead. Expanding on the use of religious imagery with Like a Prayer (1989), Madonna received positive critical reception for her diverse musical productions, while at the same time was lambasted by religious conservatives and the Vatican. In 1985, Madonna founded the Maverick corporation, a joint venture between herself and Timm Werner. The same year, she expanded the use of sexually explicit material in her work, beginning with the release of the studio album Like a Virgin, followed by the publishing of the coffee table book Sex, and starring in the erotic thriller Body of Evidence, all of which received negative responses from conservatives and liberals alike.

In 1996, Madonna played the starring role in the film Evita, for which she won a Golden Globe Award for Best Actress in a Motion Picture Musical or Comedy. Madonna's seventh studio album, Ray of Light (1998), became one of her most critically acclaimed, recognized for its lyrical depth. During the 2000s, Madonna released four studio albums – namely Music (2000), American Life (2003), Confessions on a Dance Floor (2005) and Hard Candy (2008) – all of which debuted at number one on the Billboard 200. Departing from Warner Bros. Records, Madonna signed an unprecedented $120 million dollar contract with Live Nation in 2008.

According to the International Federation of the Phonographic Industry, Madonna has sold more than 300 million albums worldwide.12 She is ranked by the Recording Industry Association of America as the best-selling female rock artist of the 20th century, and the second top-selling female artist in the United States, behind Barbra Streisand, with 40 million certified albums under her name. From 2008, Billboard magazine ranked Madonna at number two, behind only The Beatles, on the "Billboard Hot 100 All Time Top Artist", making her the most successful solo artist in the history of the chart. She was also inducted into the Rock and Roll Hall of Fame in the same year. Considered to be one of the most influential women in contemporary music, Madonna has been known for continually reinventing both her music and image, and for retaining a standard of anonymity within the recording industry. She is recognized as an influence among numerous music artists.

Gazetteer Collection Method 3

- contains structured information from Wikipedia

http://wiki.dbpedia.org/Datasets#h18-11
For Extra Points Generate Statistics

- Extract and rank the patterns in which 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?

Pattern Extraction

- Collect statistics for patterns containing NEs
  
  Ex.
  - Jenny_PER works_O for_O IBM_ORG ._O
  - Sam_PER works_O for_O Microsoft_ORG ._O
  - Paul_PER Adams_PER worked_O for_O George_PER ._O

  - Jenny_PER bought_O an_O organge_O ._O
  - Yahoo!_ORG bought_O Overtue_ORG ._O

- Extract verbs to the left and to the right of the NE
  
  Ex.
  - London_LOC is_O located_O in_O
  - John_PER drinks_O juice_O
Choose a machine learning classifier from Weka

nxm matrix, where n is number of examples, m is number of features+class label
Weka

• Classification algorithms:
  – decision trees, linear classifiers, SVM, Naive-bayes, kNN

• Prediction algorithms:
  – regression (linear/SVM), perceptron

• Meta-algorithms:
  – bagging, boosting (AdaBoost)

among others

Data file format for Weka (.arff)

@relation english_named_entity
@attribute position numeric
@attribute pos_tag { NN, NP, VB, DT}
@attribute word_length numeric
@attribute in_gazetteer { no, yes}
@attribute class { PER, LOC, ORG, MISC}
@data
  3,DT,3,no,ORG
  4,NP,10,yes,ORG
  15,NP,6,yes,PER
  7, NN,12,?,MISC
  ...
  Other attribute types:
  • String
  • Date
  Missing value
Additional Information

- General documentation:
  http://www.cs.waikato.ac.nz/ml/weka/
  http://prdownloads.sourceforge.net/weka/weka.ppt

- Command line doc:
  http://weka.wikispaces.com/Primer
Choice of classifier

The attribute whose value is to be predicted from the values of the remaining ones. Default is the last attribute.

Cross-validation: split the data into e.g. 10 folds and 10 times train on 9 folds and test on the remaining one.

10-fold Cross Validation

• Data is split into 10 approximately equal partitions
• Each partition is used in turn for testing while the remainder is used for training
  9/10 of data is used for training
  1/10 of the data is used for testing
• Repeat the whole procedure 10 times
• Overall error rate is equal to the average of the error rates on each partition
• Finally generate the final classifier by learning from all of the data.
Choosing a classifier
Given Test Data examples

<table>
<thead>
<tr>
<th>example</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td>unknown</td>
<td></td>
</tr>
<tr>
<td>unknown</td>
<td></td>
</tr>
<tr>
<td>unknown</td>
<td></td>
</tr>
<tr>
<td>unknown</td>
<td></td>
</tr>
<tr>
<td>unknown</td>
<td></td>
</tr>
<tr>
<td>unknown</td>
<td></td>
</tr>
</tbody>
</table>

**accuracy**

**different/easy class**

**all other numbers can be obtained from it**
<table>
<thead>
<tr>
<th>example</th>
<th>Cap.</th>
<th>inDicPer</th>
<th>inDicOrg</th>
<th>inDicLoc</th>
<th>NP</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>unknown</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>unknown</td>
</tr>
<tr>
<td>IBM</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>unknown</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>unknown</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>unknown</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>unknown</td>
</tr>
</tbody>
</table>
Timeline

<table>
<thead>
<tr>
<th>Event</th>
<th>Date/Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Release</td>
<td>February 9th 2012</td>
</tr>
<tr>
<td>Train&amp;Development data</td>
<td>February 9th 2012</td>
</tr>
<tr>
<td>Test data</td>
<td>February 28th 2012</td>
</tr>
<tr>
<td>Result submission deadline</td>
<td>March 1st 2012 (11:59 pm GMT) later submissions will not be accepted</td>
</tr>
<tr>
<td>Paper description deadline</td>
<td>March 1st 2012</td>
</tr>
</tbody>
</table>
Submit

• The source code for the feature generation

• The official train and test feature files used in the final run, together with the final output of your system for the test data

• Additionally generated resources (if any)

• Write 1-2 page brief description of your approach explaining:
  – used NLP tools
  – designed features
  – employed machine learning algorithm & motivation

Submit everything to

kozareva@isi.edu
dirkh@isi.edu

with subject CS544 NER
Evaluation is based on

- ranking of your system against the rest
- designed features
  - *novel*, previously unknown features will be favored
  - system’s pre or post processing
- **generated resources**
  - size, methods and sources for gazetteer extraction
  - trigger lists
- quality of the paper description
  - structure
  - use of literature
  - error analysis

What must I do ...

- Use the train and development data to design and tune your NE system
- Decide on the features you would like to incorporate in your NE system
- Choose a machine learning classifier from Weka
  - Intro by Marti Hearst
    [http://courses.ischool.berkeley.edu/i256/f06/lectures/lecture16.ppt](http://courses.ischool.berkeley.edu/i256/f06/lectures/lecture16.ppt)
- or use other good toolkits like
  - CRF++
  - SVM light, LibSVM
    etc.