CS544: NER with Weka

March 26, 2010

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

Named Entity Recognition and Classification

• Identify mentions in text and classify them into a predefined set of categories of interest:
  – Person Names: Prof. Jerry Hobbs, Jerry Hobbs
  – Organizations: Hobbs corporation, FbK
  – Locations: Ohio
  – Date and time expressions: February 2010
  – E-mail: mkg@gmail.com
  – Web address: www.usc.edu
  – Names of drugs: paracetamol
  – Names of ships: Queen Marry
  – Bibliographic references:
  – ...

<PER>Prof. Jerry Hobbs</PER> taught CS544 during <DATE>February 2010</DATE>.<PER>Jerry Hobbs</PER> killed his daughter in <LOC>Ohio</LOC>.<ORG>Hobbs corporation</ORG> bought <ORG>FbK</ORG>. 
NE System Overview
Given

<table>
<thead>
<tr>
<th>example</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Microsoft" /></td>
<td>PERSON</td>
</tr>
<tr>
<td><img src="image" alt="University of LiU" /></td>
<td>ORGANIZATION</td>
</tr>
<tr>
<td><img src="image" alt="Perspective" /></td>
<td>PERSON</td>
</tr>
<tr>
<td><img src="image" alt="ASIT" /></td>
<td>LOCATION</td>
</tr>
<tr>
<td><img src="image" alt="ASIT" /></td>
<td>ORGANIZATION</td>
</tr>
<tr>
<td><img src="image" alt="Perspective" /></td>
<td>LOCATION</td>
</tr>
<tr>
<td><img src="image" alt="Perspective" /></td>
<td>OTHER</td>
</tr>
</tbody>
</table>

<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><img src="image" alt="Microsoft" /></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>PERSON</td>
</tr>
<tr>
<td><img src="image" alt="University of LiU" /></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>ORGANIZATION</td>
</tr>
<tr>
<td><img src="image" alt="Perspective" /></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>PERSON</td>
</tr>
<tr>
<td><img src="image" alt="ASIT" /></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>LOCATION</td>
</tr>
<tr>
<td><img src="image" alt="ASIT" /></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>ORGANIZATION</td>
</tr>
<tr>
<td><img src="image" alt="Perspective" /></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>LOCATION</td>
</tr>
<tr>
<td><img src="image" alt="Perspective" /></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>OTHER</td>
</tr>
</tbody>
</table>
Choose a machine learning classifier from Weka
<table>
<thead>
<tr>
<th>Example</th>
<th>Cap</th>
<th>inDicPer</th>
<th>inDicOrg</th>
<th>inDicLoc</th>
<th>NP</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="example1.png" alt="Image" /></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><img src="example2.png" alt="Image" /></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><img src="example3.png" alt="Image" /></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><img src="example4.png" alt="Image" /></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><img src="example5.png" alt="Image" /></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

### Evaluation

<table>
<thead>
<tr>
<th>True Answer</th>
<th>Predicted Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOCATION</td>
<td>LOCATION</td>
</tr>
<tr>
<td>OTHER</td>
<td>OTHER</td>
</tr>
<tr>
<td>PERSON</td>
<td>PERSON</td>
</tr>
<tr>
<td>ORGANIZATION</td>
<td>ORGANIZATION</td>
</tr>
<tr>
<td>ORGANIZATION</td>
<td>OTHER</td>
</tr>
</tbody>
</table>

**Precision** = \# correct identified NEs / \# identified NEs

**Recall** = \# correct identified NEs / \# gold standard data
NE Feature Generation

Features (1)

- **Contextual**
  - current word Wo
  - words around Wo in [-3,...,+3] window

- **Part-of-speech tag** (when available)

- **Orthographic (binary and not mutually exclusive)**

  - initial-caps
  - roman-number
  - acronym
  - single-char
  - all-caps
  - contains-dots
  - lonely-initial
  - functional-word*
  - all-digits
  - contains-hyphen
  - punctuation-mark
  - URL

- **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 (2)

- **Bag-of-Words**
  - words in [-5,...,+5] window

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

- **Gazetteers**
  - names of cities, countries, villages, streets
  - names of organizations
  - person first name
  - person surname

*put each type of trigger words and gazetteers in separate files, because you can treat them as separate features*

Features (3)

- Length in words of the entity being classified

- Pattern of the entity with regard to the type of constituent words

  - **For each class**
    - whole NE is in gazetteer
    - any component of the NE appears in gazetteer

- **Suffixes** (length 1 to 4)

- Previous word is an article

- Previous word is a noun
Collecting External Resources

Gazetteer Collection Method 1

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

- Download 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 organization
Gazetteer Collection Method 2

• Step 1: Check if identified NE exists in Wikipedia
• Step 2: Extract the first 2-3 sentences
• Step 3: Pull the nouns matching the expression
  
  X is Y, Z
  X is Y and Z

• Step 4: Extract the information from the infobox
• Step 5: Verify in WordNet whether the found concepts are hyponyms of person, location, organization
Gazetteer Collection Method 3

- Use Stanford Named Entity Recognizer
  http://nlp.stanford.edu/software/CRF-NER.shtml
  to identify the named entities in the current data sets.

- Use the predicted output as features

Patterns
Capturing Simple Patterns

• Extract patterns in which the NEs occurred
  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 Overtrue_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

WEKA
Waikato Environment for Knowledge Analysis
Weka: Data Mining Software

- Collection of machine learning algorithms
  - open-source package written in Java
- Used for research, education and application
- Main features:
  - data pre-processing tools
  - learning algorithms
  - evaluation methods
  - graphical inference
  - environment for comparing learning algorithms

Weka: Data Mining Software

- Classification algorithms:
  - decision trees, linear classifiers, SVM, Naive-bayes, kNN
- Prediction algorithms:
  - regression (linear/SVM) , perceptron
- Meta-algorithms:
  - bagging, boosting (AdaBoost)

among others
Getting Started

• Install Weka software (on Linux):
  – Download link:
    • http://prdownloads.sourceforge.net/weka/weka-3-6-2.zip
    • Unzip the software
  – Requirement: Java 1.5 (or higher)
  – Invoke Weka command:
    • java -cp weka.jar <weka-command>

Weka GUI Chooser

java -Xmx1000M -jar weka.jar
Data file format (.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

The Preprocessing Tab

- Classification
- Preprocessing
- Statistical attribute selection
- Filter selection
- Manual attribute selection
- List of attributes (last: class variable)

Frequency and categories for the selected attribute

Statistics about the values of the selected attribute
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.

Choosing a classifier
all other numbers can be obtained from it

accuracy
different/easy class

Running on Test Set
Weka specifications

- Train classifier on training data and output model
  - java -cp weka.jar <classifier-function> -t <train-file> -d <trained-model>
- Run trained classifier model on test data
  - java -cp weka.jar <classifier-function> -T <test-file> -I <trained-model>
- Specifying parameters:
  -t: training file (.arff)
  -T: test file (.arff)
  -d: output filename (trained classifier model)
  -I: input model (for testing)
  -K: number of nearest neighbors for kNN algorithm
  -h: help (check out other parameter options, etc.)
Example: kNN in Weka

- Train a classifier using 2NN algorithm
  - java -cp weka.jar
    weka.classifiers.lazy.IBk
    -t data/weather.arff
    -K 2
    -d model.2nn

- Run the trained classifier on test data
  - java -cp weka.jar
    weka.classifiers.lazy.IBk
    -T data/weather.arff
    -I model.2nn

Sample Weka output

--- Error on test data ---

<table>
<thead>
<tr>
<th>Correctly Classified Instances</th>
<th>13</th>
<th>92.8571 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrectly Classified Instances</td>
<td>1</td>
<td>7.1429 %</td>
</tr>
<tr>
<td>Kappa statistic</td>
<td>0.8372</td>
<td></td>
</tr>
<tr>
<td>Mean absolute error</td>
<td>0.1333</td>
<td></td>
</tr>
<tr>
<td>Root mean squared error</td>
<td>0.2333</td>
<td></td>
</tr>
<tr>
<td>Total Number of Instances</td>
<td>14</td>
<td></td>
</tr>
</tbody>
</table>
More detailed output

- Classification labels for each instance (use “-p 1” option)
  - java -cp weka.jar weka.classifiers.lazy.lbk -T data/weather.arff -l model.2nn -p 1

### Predictions on test data ###

<table>
<thead>
<tr>
<th>Inst#</th>
<th>actual</th>
<th>predicted</th>
<th>error</th>
<th>prediction (outlook)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>no</td>
<td>no</td>
<td>0.967</td>
<td>sunny</td>
</tr>
<tr>
<td>2</td>
<td>no</td>
<td>yes</td>
<td>0.5</td>
<td>overcast</td>
</tr>
<tr>
<td>3</td>
<td>yes</td>
<td>yes</td>
<td>0.967</td>
<td>overcast</td>
</tr>
<tr>
<td>4</td>
<td>yes</td>
<td>yes</td>
<td>0.967</td>
<td>overcast</td>
</tr>
<tr>
<td>5</td>
<td>yes</td>
<td>yes</td>
<td>0.967</td>
<td>overcast</td>
</tr>
<tr>
<td>6</td>
<td>no</td>
<td>no</td>
<td>0.967</td>
<td>rainy</td>
</tr>
<tr>
<td>7</td>
<td>yes</td>
<td>yes</td>
<td>0.967</td>
<td>overcast</td>
</tr>
<tr>
<td>8</td>
<td>no</td>
<td>no</td>
<td>0.967</td>
<td>sunny</td>
</tr>
<tr>
<td>9</td>
<td>yes</td>
<td>yes</td>
<td>0.5</td>
<td>sunny</td>
</tr>
<tr>
<td>10</td>
<td>yes</td>
<td>yes</td>
<td>0.967</td>
<td>rainy</td>
</tr>
<tr>
<td>11</td>
<td>yes</td>
<td>yes</td>
<td>0.5</td>
<td>sunny</td>
</tr>
<tr>
<td>12</td>
<td>yes</td>
<td>yes</td>
<td>0.967</td>
<td>overcast</td>
</tr>
<tr>
<td>13</td>
<td>yes</td>
<td>yes</td>
<td>0.967</td>
<td>overcast</td>
</tr>
<tr>
<td>14</td>
<td>no</td>
<td>no</td>
<td>0.967</td>
<td>rainy</td>
</tr>
</tbody>
</table>

Weka classification functions

- kNN:  [weka.classifiers.lazy.lbk](weka.classifiers.lazy.lbk)
- Naïve Bayes: [weka.classifiers.bayes.NaiveBayes](weka.classifiers.bayes.NaiveBayes)
- AdaBoost: [weka.classifiers.meta.AdaBoostM1](weka.classifiers.meta.AdaBoostM1)
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