NERC

Named Entity Recognition and Classification

Team

- Bizdadea Dan
- Bratiloveanu Florentina
- Ciobanu Catalin
- Sorostinean Mihaela

Outline

- Team
- Project Description
- Motivation
- State of the art
- Related work
- Tools
- QA

What is NERC ?

- NERC – Named Entity Recognition and Classification (NERC) involves identification of proper names in texts, and classification into a set of pre-defined categories of interest as:
  - Person names (names of people)
  - Organization names (companies, government organizations, committees, etc.)
  - Location names (cities, countries etc)

Motivation

- Search engines, question answering systems can be enhanced by allowing searches and questions for particular persons, companies or locations.
- In text mining, accurate NER will allow the construction of databases with information extracted about particular entities.
- Applications of multilingual NER are cross-language information retrieval, and business intelligence applications, where information about a particular person or company has to be extracted from textual sources in different languages.

Project Description (1)

- Approaches:
  - Rule based NERC (grammar-based)
  - Machine learning (ML) based NERC
    - Supervised ML technique
    - Semi-supervised ML technique
    - Unsupervised ML technique

- Applications
  - Machine Translation
  - Information Retrieval
  - Question-Answering system
  - Automatic Summarization

Project Description (2)

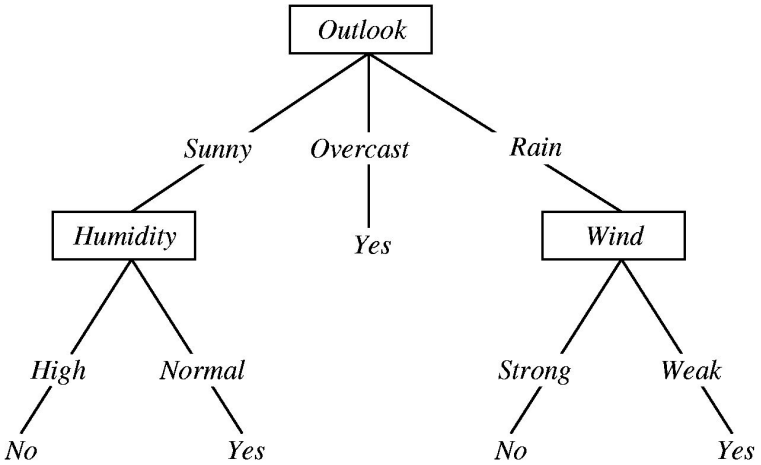
- Supervised learning
  - Hidden Markov Model (HMM)
  - Decision Trees
  - Maximum Entropy Models (ME)
  - Support Vector Machines (SVM)
  - Conditional Random Fields(CRF)

Project Description (3.1) - Hidden Markov Model

- HMM is the earliest model applied for solving NER problem by Bikel et al. (1999) for English. Bikel introduced a system, IdentiFinder, to detect NER
- the model assigns to every word, either one of the desired classes or the label NOT-A-NAME to represent ”none of the desired classes”
- there are two special states, the START-OF-SENTENCE and END-OF-SENTENCE states

Project Description (3.2) - Decision Trees

- Decision Tree is a classifier in the form of a tree structure where each node represent a leaf node, indicates the value of the output attributes of expressions
- A decision is made based on the combination of the features given to the decision tree



Project Description (3.3) - Maximum Entropy

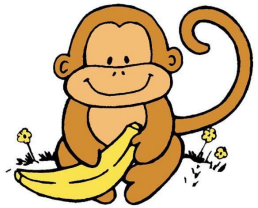
- according to MUC-7 the test corpus contains 7 tags
- 4 new states: x_start, x_continue, x_end, x_unique resulting in 28 tags that can be assigned to entities
- “other” tag for tokens that are not part of the named entity
- Example:
  - [Jerry Lee Lewis flew to Paris] is tagged as:
  - [person_start, person_continue, person_end, other, other, location_unique]
- Use Maximum Entropy model to determine the best way to select the tags.
- The model states that: the probability distribution that should be selected should be one that leaves the highest uncertainty consistent with the constraints.

Project Description (3.4) - Support Vector Machines

- are well-known for their good generalization performance
- SVMs are applied to text categorization
  - high accuracy
  - avoid over-fitting
- a SVM learns a linear hyperplane that separates the set of positive examples from the set of negative examples with maximal margin (the margin is defined as the distance of the hyperplane to the nearest of the positive and negative examples).

Project Description (3.5) - Conditional Random Fields

- The majority of classifiers try to predict an output based on a single input without taking in consideration the previous inputs.
- Imagine that we want to predict a life of a monkey in pictures.
- Firstly, the monkey is climbing in a banana tree and we can predict the climbing action
- Secondly, we see the monkey staying with the hand closely to the mouth
- Using the first picture we can say that the monkey is probably eating a banana after climbing in the tree and this is how CRF works.



State of the art

- Hirschman and Chinchor (1997) considered three classes: person, location and organization.
- Doddington et al. introduced geo-political entities, weapons, vehicles and facilities.
- Others introduces 18 classes categorization, but it fails to distinguish between classes.

Related work

- Named Entity Recognition in Tweets: An Experimental Study
- TwiNER: Named Entity Recognition in Targeted Twitter Stream
- Named Entity Recognition System for Urdu
- Named Entity Recognition on Turkish Tweets

Tools

- General Architecture for Text Engineering (GATE)
- OpenNLP
- **Stanford CoreNLP**
- NETagger
- **MUC Data Sets**
- Language-Independent Named Entity Recognition at CoNLL-2003
- **https://wordnet.princeton.edu/**
-

Questions ?

Thank you

for your attention!