Named Entities: Background
Outline

• Introduction and Motivation
• Learning Methods and Features
• Evaluation
Named Entities

• Entity – something that exists in the real world
  Dog, president, company, country, month, currency, day, year

• Named entity – a reference to the entity in the real world
  Racer, Barack Obama, Bavarian Motor Works, Germany, March, Euro, Wednesday, 2003
Named Entity Types

• How many types are there?
• Different hierarchies with different granularities for different domains exist
Named Entity Recognition

- Identify and classify occurrences of named entities

Jim bought 300 shares of Acme Corp. in 2006.

\[ [Jim]_{\text{person}} \text{ bought } [300]_{\text{quantity}} \text{ shares of } [\text{Acme Corp.}]_{\text{organization}} \text{ in } [2006]_{\text{date}}. \]
Applications

• Looking for organizations, products, people in your neighbourhood
• Monitoring news in the huge amount of textual media produced every day
• Simplification of more advanced tasks in NLP (Relation Extraction)
Learning Methods

• Studying of negative and positive examples allows to recognize previously unseen entities

• Hand-crafted rules
  (if the spelling contains “Mr.” then it is a Person;
  if the spelling is all capitalized then it is an organization)

• Automatically induced probabilistic models
  (…Mr. Lane…, …Mr. Douglas…, …Mr. Gomez…, …Mr. Reich… – high probability; and also Mr. de Castro)

• There are different types of learning methods depending on the amount of training data
Supervised Learning

• Requires a lot of annotated training data
• Models (sets of rules) don’t change after the training is finished

... [Rolls-Royce Motor Cars Inc.] organization said it expects its [U.S.] nationality sales to remain steady at about [1,200] quantity cars in [1990] date.
[ITEL CORP.] organization reported third-quarter earnings, which were shown in the [Quarterly Earnings Surprises] organization table in [Monday] date’s edition.

... [The Procter & Gamble Company] organization said it expects earnings growth for the June quarter.
Daimler AG sees its [Indian] nationality sales rising 20 percent to around [3,000] quantity cars in [2008] date on sustained demand for its luxury Mercedes-Benz cars.
Semi-supervised Learning

• Requires a small amount of data for starting the learning process
• Learning process is repeated several times

[Guatemala]_location

Embassy of [Guatemala]_location
List of [Guatemala]_location newspapers
Travel information about [Guatemala]_location

Embassy of [India]_location
Embassy of the [Republic of the Philippines]_location
List of [Czech]_location newspapers
List of [American]_location newspapers
Travel information about [Switzerland]_location
Travel information about [Belgrade]_location
Unsupervised Learning

- No annotated training data available
- Different approaches possible

- Following observation: “(Hypernym of X) such as X“

- Query “such as chair“
  furniture such as chair
  Individual such as chair

- Query “such as a car“
  Vehicle such as a car
  Investment such as a car
  Property such as a car

- Query “such as Belgrade“
  places such as Belgrade
  cities such as Belgrade

- Query “such as Austria“
  states such as Austria
  countries such as Austria
  cultural nation such as Austria
Features

• Features are descriptors or characteristic attributes of words
• Choosing discriminating and independent features is key to any pattern recognition algorithm being successful in classification
• Feature types:
  **Boolean** (Word is capitalized, Word has length $\leq 3$)
  **Numeric** (Word’s length, number of capital letters)
  **Nominal** (Lowercased version of the word)
Feature Levels

- Word-level features
  - Case, punctuation, digit, character, morphology, part-of-speech, function
- List lookup features
  - General list, entities list, list of entity cues
- Document features
  - Multiple occurrences, local syntax, meta information, corpus frequency
Evaluation

• Let’s now hypothesize a system producing the following output for this input:

• Unlike [Robert]_{person}, [John Briggs Jr]_{person} contacted [Wonderful Stockbrockers Inc]_{organization} in [New York]_{location} and instructed them to sell all his shares in [Acme]_{organization}.

• [Unlike]_{location} Robert, [John Briggs Jr]_{organization} contacted Wonderful [Stockbrockers]_{organization} Inc [in New York]_{person} and instructed them to sell all his shares in [Acme]_{organization}.
Evaluation

• In order to be able to create a good system, one requires a method to measure its quality

• What score should be assigned to the system?

• Unlike \[\text{location}\]

• [Robert]_{\text{person}} \rightarrow \text{Robert}

• [John Briggs Jr]_{\text{person}} \rightarrow [John Briggs Jr]_{\text{organization}}

• [Wonderful Stockbrokers Inc]_{\text{organization}} \rightarrow [Stockbrokers]_{\text{organization}}

• [New York]_{\text{location}} \rightarrow [\text{in New York}]_{\text{person}}
Simple Evaluation Metrics

- **Correct Type (TYPE)** – type is predicted correctly and there is an overlap of text boundaries
- **Correct Boundaries (TEXT)** – boundaries are predicted exactly, regardless of the type
- **Correct Answers (COR)** – number of TYPE and TEXT
- **Actual System Guesses (ACT)**
- **Possible Entities in the Solution (POS)**
- **Recall (R)** – COR / POS
- **Precision (P)** – COR / ACT
- **F-Score (F)** – $2 \times P \times R / P + R$
Powerful Evaluation Metrics

• Start with 100% (perfect system) and penalize system for every mistake (type mistakes, missed entities) according to the importance of the classes (e.g. person more important than organization)

• Hypothetical final score: 1 – 30.77 – 24.54 – 5.77 – 7.58 = 31.3%
NER for Historical Texts

• Large amount of training data is necessary, but data for other domains is usually not applicable (Different text types, sentence structures, lexicon)
• For historical tasks rich type hierarchies are required (more than 10 PLACE (natural, region, specific location) types or SITE (industrial, monument, religious) types)

• E.g. Lighthouse of Alexandria, University of Heidelberg(university team, university researchers…)
• Great Plains, Rocky Mountains (great play, great interest, great demand, rocky stock market, rocky opening, rocky path…)
Summary

- Named Entity Recognition is an important task in NLP
  - Per se
  - For other applications
- There are different learning methods
  - Supervised learning
  - Semi-supervised learning
  - Unsupervised learning
- Features of various levels are essential for any classification task
- Evaluation of models is necessary in order to improve their quality
- NER for historical tasks is a very challenging task