Named Entity Recognition
Lecture 12: October 18, 2013

CS886-2 Natural Language Understanding

University of Waterloo

Entities and Relations

- The essence of a document can often be captured by the entities and relations that are mentioned
- Entity: object, person, organization, date, etc.
  - Most things denoted by a noun phrase or pronoun
- Relation: property that links one or several entities
  - Most things denoted by an adjective, verb or adverb

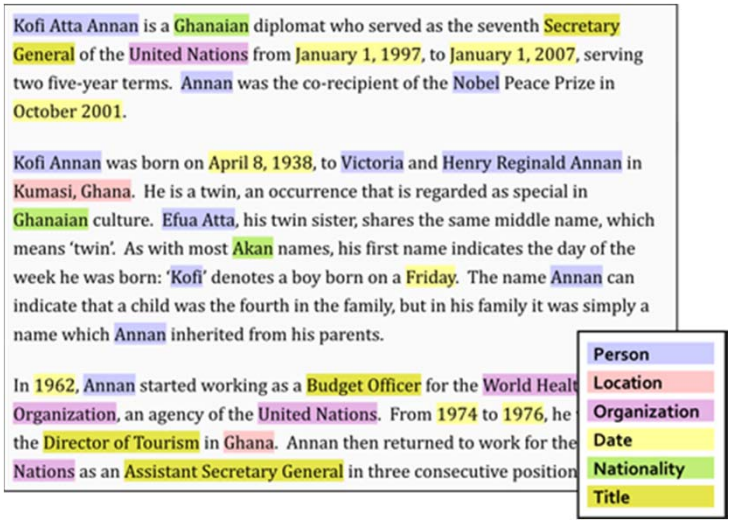
Named Entities

- Among all entities, named entities are often the most important ones for
  - Text summarization
  - Question answering
  - Information retrieval
  - Sentiment analysis
- Definition: subset of entities referred by a "rigid designator"
- Rigid designator: expression that always refers to the same thing in all possible worlds

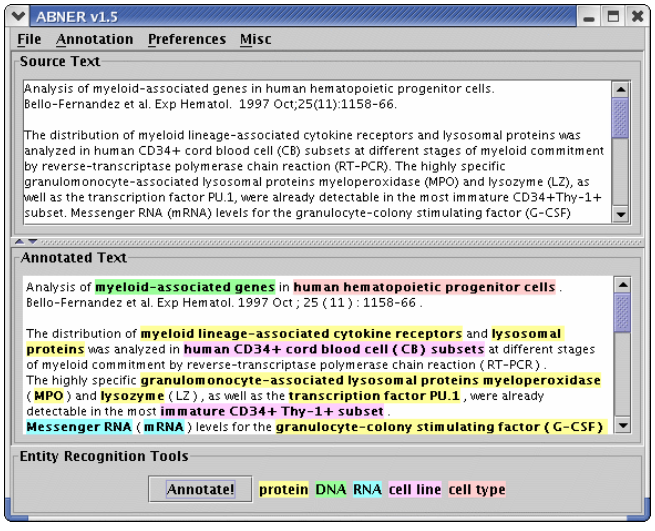
Named Entity Recognition (NER)

- Task:
  - Identify named entities
  - Classify named entities
- Classes:
  - Common: Person, organization, location, quantity, time, money, percentage, etc.
  - Biology: genes, proteins, molecules, etc.
  - Fine grained: all Wikipedia concepts (one concept per Wikipedia page)

News NER example



Biomedical NER example



Classification

- Approach: classify each word (phrase) with an entity type
- Supervised learning:
  - Train with corpus of labeled text (labels are entity types)
- Semi-supervised learning:
  - Train with some labeled texts and large corpus of unlabeled texts

Independent Classifiers

- Classify each word in isolation
  - Naïve Bayes model
  - Logistic regression
  - Decision tree
  - Support vector machine

Correlated Classifiers

- Jointly classify all words while taking into account correlations between some labels
  - Hidden Markov Model
  - Conditional Random Field
- Adjacent words (phrases) often have correlated labels
- Identical words often have the same label

Naïve Bayes Model

- Picture

Features

- Features are more important than the model itself
  - Results: very sensitive to the choice of features
- Feature: anything that can be computed by a program based on the text
- Common features:
  - Word, previous word, next word (more words do not seem to help)
  - Prefixes and suffixes
  - Shape
  - Combination of features
  - Part-of-speech tags
  - Gazetteer

Common Features

- Word, previous word, next word
- Prefixes and suffixes
- Shape

Common Features

- Part-of-speech tags
- Gazetteer
- Combination of features

Training

- Generative training: maximum likelihood
  - $\theta^* = argmax \Pr(class, features|\theta)$
  - Closed form solution: relative frequency counts
  - Fast, but inaccurate
- Discriminative training: conditional maximum likelihood
  - $\theta^* = argmax \Pr(class|features, \theta)$
  - No closed form solution: iterative technique such as gradient ascent
  - Slow but more accurate (optimize the right objective)

Example

Derivation

Logistic Regression

- Alternative to Naïve Bayes model
  - Different parameterization, but often equivalent to discriminative naïve Bayes learning
- Idea: joint distribution is proportional to the exponential of a weighted sum of the features

$$\Pr(class_i|Features, w) \propto e^{\sum_j w_{ij}\phi_j(words)}$$

Example

Discriminative Training

- Maximize conditional likelihood

$$w^* = argmax_w \Pr(class|Features, w)$$

- No closed form solution: iterative technique
  - E.g. Gradient ascent

Joint Classification

- Joint classification allows us to take into account correlations between some labels
  - Adjacent words often have correlated entity types
  - Identical words often have the same entity type
- Approaches:
  - Naïve Bayes extension: Hidden Markov Model
  - Logistic regression extension: conditional random field