# Word Sense Disambiguation (Sense Tagging)

Presenter:
Omar Salman Manzoor

23rd December, 2014 | Center for Language Engineering (CLE)

# Definition and Basic Idea

- **Word Sense Disambiguation** refers to the task of identifying the correct meaning and sense of a word according to the context.
- It is quite useful and vital in many natural language processing applications like machine translation.
- Statistic data extracted from sense tagged corpus can be implemented in
  - Information Retrieval (IR)
  - Information Extraction
  - Text Summarization

# Need for WSD

- An Urdu Sense Tagged Corpus has been developed.
- The need for developing WSD is to use this corpus to develop a training model which can assign senses to various words.
- WSD for Urdu is important because it can be used to enhance the Urdu Word Net by adding more senses and also adding relationship between various senses

# Example

- He deposited money in the bank.
- He likes to go visit the river bank every Sunday.

- The task here is to provide the correct meaning of the word bank in each case.

# Techniques

- Supervised Learning methods
- Dictionary Methods
- Bootstrapping Approach
- Unsupervised Learning

# Supervised Learning

- Collocation Features
- Collocation is a word or phrase in a position specific relationship to a target word.
- These features encode information about specific words or phrases located at specific positions to the left or right of the target word.

# Supervised Learning

- Bag of Words Features
- These features include an unordered set of words.
- A specific window size is chosen with the target word at the center so that words to the right and left of the target word are checked.

# Supervised Learning

- Naïve Bayes Classifier
- $P(f|s) \approx \prod_{j=1}^{n} P(f_j|s)$
- Probability of feature vector given a sense estimated by the probabilities of its individual features given that sense.
- Training the classifier first requires estimate for prior probability of each sense.
- Also needed are individual feature probabilities given a sense.
- Smoothing is essential in this approach.

# Supervised Learning

- Decision List Classifiers..
- A sequence of tests applied to each target word feature vector.
- A test indicates a particular sense.
- If a test succeeds that sense is applied.
- Otherwise next test is applied and process continues.
- In case of no test succeeding majority test retuned as default.

# Dictionary Methods

- Lesk Algorithm
- Chooses the sense whose dictionary gloss or meaning shares the most words with the target word's neighborhood.
- Example : The **bank** can guarantee deposits will cover future tuition costs because it invests in adjustable-rate mortgage securities.

| | | |
|---|---|---|
| bank[1] | Gloss: | a financial institution that accepts deposits and channels the money into lending activities |
| | Examples: | "he cashed a check at the bank", "that bank holds the mortgage on my home" |
| bank[2] | Gloss: | sloping land (especially the slope beside a body of water) |
| | Examples: | "they pulled the canoe up on the bank", "he sat on the bank of the river and watched the currents" |

# Bootstapping

- Semi or Minimally Supervised Learning.
- Need only a small set of hand labeled data.
- Small seed set of labeled instances $\Lambda_0$ of each sense. A larger unlabeled corpus $V_0$.
- Algorithm first trains initial classifier on $\Lambda_0$ and then labels the corpus $V_0$.
- Then examples in $V_0$ that are most convincing are added to training set now becomes $\Lambda_1$. This is repeated.

# Unsupervised Learning

- Clustering
- Similar senses occur in similar contexts and are found by clustering based on similarity in context referred to as word sense induction.
- New instances classified into closet induced clusters.

# Urdu Sense Tagged Corpus

- Total Number of Sentences is 5611
- Total Number of Words is 100,000
- Tagged total word types 2225
- Tagged total sense types 2285
- Tagged total word tokens 17006
- 559 words which have more than 2 senses tagged. 1522 words with one sense.

# Urdu Sense Tagged Corpus

- Challenges include ambiguity in tagging non standardized translations of some English Words.
- For some foreign language words no sense tagging found. E.g. test match, basket ball
- There are complex predicates in Urdu.
- Normalization is required.
- This corpus can act as a seed corpus.

# Challenges in Development

- There are a number of pre processing considerations like stemming and removal of stop words.
- The data has a number of senses which have not been tagged sufficiently.
- Many of the words in the data have not been tagged or have no specific sense tags.

# How to Proceed

- We plan on using the words which have **at least 20** tagged instances .
- Using these instances the idea is to develop a semi supervised learning algorithm using **Naïve Bayes Classification** as the base method.
- Then labeling of the untagged data will be done automatically by choosing only the most confident output instances through **clustering**.

# Thank You!!