CS460/626 : Natural Language Processing/Speech, NLP and the Web (Lecture 24– WSD)

Pushpak Bhattacharyya
CSE Dept.,
IIT Bombay
5th March, 2012
Layers of NLP

- Increased Complexity Of Processing
  - Discourse and Coreference
    - Semantics
      - Parsing
        - Chunking
          - POS tagging
            - Morphology

- Problem
  - Parsing
    - Part of Speech Tagging
      - Morph Analysis
        - CRF
          - HMM
            - MEMM

- Language
  - Marathi
  - French
  - Hindi
  - English

NLP Trinity
Motivation

WSD: At the Heart of NLP

SRL : Semantic Role Labeling
TE  : Text Entailment
CLIR: Cross Lingual Information Retrieval
NER : Named Entity Recognition
MT  : Machine Translation
SP  : Shallow Parsing
SA  : Sentiment Analysis
WSD : Word Sense Disambiguation
LEARNING BASED v/ S HYBRID APPROACHES

- Knowledge Based Approaches
  - Rely on knowledge resources like WordNet, Thesaurus etc.
  - May use grammar rules for disambiguation.
  - May use hand coded rules for disambiguation.

- Machine Learning Based Approaches
  - Rely on corpus evidence.
  - Train a model using tagged or untagged corpus.
  - Probabilistic/Statistical models.

- Hybrid Approaches
  - Use corpus evidence as well as semantic relations form WordNet.
Bird’s eye view

WSD Approaches

Machine Learning

- Supervised
- Unsupervised

Knowledge Based

- Semi-supervised
- Hybrid

CFILT - IITB
KNOWLEDGE BASED APPROACHES
Sense 1

- This airlines \textit{serves} dinner in the evening flight.
- \textit{serve} (Verb)
  - agent
  - object – edible

Sense 2

- This airlines \textit{serves} the sector between Agra & Delhi.
- \textit{serve} (Verb)
  - agent
  - object – sector

Requires exhaustive enumeration of:

- Argument-structure of verbs.
- Selectional preferences of arguments.
- Description of properties of words such that meeting the selectional preference criteria can be decided.

E.g. This flight serves the “\textit{region}” between Mumbai and Delhi

How do you decide if “\textit{region}” is compatible with “\textit{sector}”?
SELECTIONAL PREFERENCES (INDIAN TRADITION)

- “Desire” of some words in the sentence (“aakaangksha”).
  - *I saw* the boy with long hair.
  - *The verb “saw” and the noun “boy” desire an object here.*

- “Appropriateness” of some other words in the sentence to fulfil that desire (“yogyataa”).
  - *I saw the boy with long hair.*
  - *The PP “with long hair” can be appropriately connected only to “boy” and not “saw”.*

- In case, the ambiguity is still present, “proximity” (“sannidhi”) can determine the meaning.
  - *E.g. I saw the boy with a telescope.*
  - *The PP “with a telescope” can be attached to both “boy” and “saw”, so ambiguity still present. It is then attached to “boy” using the proximity check.*
SELECTIONAL PREFERENCES (RECENT LINGUISTIC THEORY)

- There are words which demand arguments, like, verbs, prepositions, adjectives and sometimes nouns. These arguments are typically nouns.
- Arguments must have the property to fulfil the demand. They must satisfy selectional preferences.
  - Example
    - Give (verb)
      - agent – animate
      - obj – direct
      - obj – indirect
    - *I gave* him the *book*
    - *I gave* him the *book* (yesterday in the school) -> adjunct
- How does this help in WSD?
  - One type of contextual information is the information about the type of arguments that a word takes.
Verb Argument frame

- Structure expressing the desire of a word is called the *Argument Frame*
- Selectional Preference
  - Properties of the “Supply Words” meeting the desire of the previous set
Argument frame (example)

Sentence: I am fond of X

Fond
{
  Arg1: Prepositional Phrase (PP)
    {
      PP: of NP
        {
          N: somebody/something
        }
    }
}
Verb Argument frame (example)

Verb: give

Give

\[
\begin{array}{l}
\text{agent: } <\text{the give}>_{\text{animate}} \\
\text{direct object: } <\text{the thing given}> \\
\text{indirect object: } <\text{beneficiary}>_{\text{animate/organization}}
\end{array}
\]

[I]_{\text{agent}} \text{ gave a } [\text{book}]_{\text{dobj}} \text{ to } [\text{Ram}]_{\text{obj}}.
Resources for Verbs

- VerbNet
  (http://verbs.colorado.edu/~mpalmer/projects/verbnet.html)
- Propbank (http://en.wikipedia.org/wiki/PropBank)
- VerbOcean
  (http://demo.patrickpantel.com/demos/verbocean/)
CRITIQUE

- Requires exhaustive enumeration in machine-readable form of:
  - Argument-structure of verbs.
  - Selectional preferences of arguments.
  - Description of properties of words such that meeting the selectional preference criteria can be decided.
    - E.g. This flight serves the "region" between Mumbai and Delhi
    - How do you decide if “region” is compatible with “sector”

- **Accuracy**
  - 44% on Brown corpus.
OVERLAP BASED APPROACHES


- Find the overlap between the features of different senses of an ambiguous word (sense bag) and the features of the words in its context (context bag).

- These features could be sense definitions, example sentences, hypernyms etc.

- The features could also be given weights.

- The sense which has the maximum overlap is selected as the contextually appropriate sense.
LESK’S ALGORITHM

**Sense Bag:** contains the words in the definition of a candidate sense of the ambiguous word.

**Context Bag:** contains the words in the definition of each sense of each context word.

E.g. “On burning *coal* we get *ash*.”

From Wordnet

- The noun *ash* has 3 senses (first 2 from tagged texts)
  - 1. (2) *ash* -- (the residue that remains when something is burned)
  - 2. (1) *ash*, *ash tree* -- (any of various deciduous pinnate-leaved ornamental or timber trees of the genus *Fraxinus*)
  - 3. *ash* -- (strong elastic wood of any of various ash trees; used for furniture and tool handles and sporting goods such as baseball bats)
- The verb *ash* has 1 sense (no senses from tagged texts)
  - 1. *ash* -- (convert into ashes)
CRITIQUE

- Proper nouns in the context of an ambiguous word can act as strong disambiguators.

  E.g. “Sachin Tendulkar” will be a strong indicator of the category “sports”.

  Sachin Tendulkar plays **cricket**.

- Proper nouns are not present in the thesaurus. Hence this approach fails to capture the strong clues provided by proper nouns.

- **Accuracy**
  - 50% when tested on 10 highly polysemous English words.
Extended Lesk’s algorithm

- Original algorithm is sensitive towards exact words in the definition.

- Extension includes glosses of semantically related senses from WordNet (e.g. hypernyms, hyponyms, etc.).

- The scoring function becomes:

  \[
  score_{ext}(S) = \sum_{s' \in rel(s) \text{ or } s = s'} |\text{context}(w) \cap \text{gloss}(s')| 
  \]

- where,
  - \(\text{gloss}(S)\) is the gloss of sense \(S\) from the lexical resource.
  - \(\text{Context}(W)\) is the gloss of each sense of each context word.
  - \(rel(s)\) gives the senses related to \(s\) in WordNet under some relations.
A place that serves as the living quarters of one or more families

WordNet Sub-Graph
Example: Extended Lesk

“On combustion of coal we get ash”

From Wordnet

- The noun ash has 3 senses (first 2 from tagged texts)
  - 1. (2) ash -- (the residue that remains when something is burned)
  - 2. (1) ash, ash tree -- (any of various deciduous pinnate-leaved ornamental or timber trees of the genus Fraxinus)
  - 3. ash -- (strong elastic wood of any of various ash trees; used for furniture and tool handles and sporting goods such as baseball bats)
- The verb ash has 1 sense (no senses from tagged texts)
- 1. ash -- (convert into ashes)
Example: Extended Lesk (cntd)

“On combustion of coal we get ash”

From Wordnet (through hyponymy)
- ash -- (the residue that remains when something is burned)
  => fly ash -- (fine solid particles of ash that are carried into the air when fuel is combusted)
  => bone ash -- (ash left when bones burn; high in calcium phosphate; used as fertilizer and in bone china)
Critique of Extended Lesk

- Larger region of matching in WordNet
  - Increased chance of Matching
  - BUT
  - Increased chance of Topic Drift