Ambiguity and Word Sense Disambiguation

D. Thenmozhi
Associate Professor
SSNCE

Ambiguity

- Having more than one meaning
- Can occur at four levels
  - Lexical
  - Syntactic
  - Semantic
  - pragmatic

Lexical Ambiguity

- Ambiguity of a single word
- Example (ambiguity viewed as POS tagging)
  - She bagged two silver medals
  - His worries had silvered his hair
- Example (same syntactic category with different meaning)
  - There is a hike in price of gold
  - She has a heart of gold
- Soln: WordNet sense information
  - Gold has 5 senses in WordNet

Syntactic Ambiguity

- There are different ways in which a sequence of words can be grammatically structured
- Example
  - The man saw the girl with the telescope
  - Stolen painting found by tree
- Soln: Selectional restriction
  - ‘find’ takes an agent with the property, animate

Semantic Ambiguity

- Meaning of the words themselves can be misinterpreted
- Example : Iraqi head seeks arms
  - Chief seeks weapons
  - Anatomical head of a body seeks a body part
- Soln: WSD algorithms

Pragmatic Ambiguity

- Context of a phrase give multiple interpretation
- Give it to the kids
  - it may refer to anything
- Cake is on the table. I have prepared some snacks. Give it to kids.
  - it may refer to cake or snacks or both
- Cake is on the table. I have prepared some snacks. Give it to kids. Kids enjoy cake and snacks
  - it may refer to both

Word Sense Disambiguation

- Selectional Restriction-based WSD
- Context-based WSD
  - Knowledge-based
  - Corpus-based
    - Supervised
      - Bayesian
      - K-nearest neighbour
      - Bootstrap
      - Decision list
      - Bilingual corpora
    - Unsupervised

Selectional Restriction-based WSD

- The institute will employ new employees (to hire)
  - Subject - human/organization, object - human
- The committee employed her proposal (to accept)
  - Subject – human/organization, object- idea
- Given employee as object the sense to hire is selected as the interpretation of employee and the sense to accept is ruled out

Context-based WSD

- Knowledge-based
  - Utilize information from an explicit lexicon (dictionary/thesaurus (WordNet)/ontology) or knowledge base
- Corpus-based
  - Extract word sense from a large sense-tagged corpus
  - Data acquisition bottleneck
  - Creating sense tagged corpus is a labour intensive task/difficult
  - Soln: bootstrapping

Corpus-based WSD

- Supervised
  - Rely on a sense tagged training corpus
  - Considered as a classification task
- Unsupervised
  - Make use of raw or unlabelled text corpora for training and annotated data for evaluation
  - Considered as a clustering task

Knowledge-based WSD

- Knowledge-based WSD = class of WSD methods relying (mainly) on knowledge drawn from dictionaries and/or raw text
- Resources
  - Yes
    - Machine Readable Dictionaries
    - Raw corpora
  - No
    - Manually annotated corpora

Machine Readable Dictionaries

- In recent years, most dictionaries made available in Machine Readable format (MRD)
  - Oxford English Dictionary
  - Collins
  - Longman Dictionary of Ordinary Contemporary English (LDOCE)
- Thesauruses – add synonymy information
  - Roget Thesaurus
- Semantic networks – add more semantic relations
  - WordNet
  - EuroWordNet

MRD – A Resource for Knowledge-based WSD

- For each word in the language vocabulary, an MRD provides:
  - A list of meanings
  - Definitions (for all word meanings)
  - Typical usage examples (for most word meanings)

WordNet definitions/examples for the noun *plant*

1. buildings for carrying on industrial labor; "they built a large plant to manufacture automobiles"
2. a living organism lacking the power of locomotion
3. something planted secretly for discovery by another; "the police used a plant to trick the thieves"; "he claimed that the evidence against him was a plant"
4. an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience

MRD – A Resource for Knowledge-based WSD

- A thesaurus adds:
  - An explicit synonymy relation between word meanings

WordNet synsets for the noun“plant”
1. plant, works, industrial plant
2. plant, flora, plant life

- A semantic network adds:
  - Hypernymy/hyponymy (IS-A), meronymy/holonymy (PART-OF), antonymy, etc.

WordNet related concepts for the meaning “plant life”
{plant, flora, plant life}
hypernym: {organism, being}
hypomym: {house plant}, {fungus}, …
meronym: {plant tissue}, {plant part}
member holonym: {Plantae, kingdom Plantae, plant kingdom}

Lesk Algorithm

- (Michael Lesk 1986): Identify senses of words in context using definition overlap. That is, disambiguate more than one word.
- Algorithm:
  - Retrieve from MRD all sense definitions of the words to be disambiguated
  - Determine the definition overlap for all possible sense combinations
  - Choose senses that lead to highest overlap

(1) for each sense *i* of *W*

(2) determine *Overlap(i)*, the number of words in common
between the definition of sense *i* and current sentential context

(3) find sense *i* for which *Overlap(i)* is maximized

(4) assign sense *i* to *W*

Lesk Algorithm

Example: disambiguate PINE in

"Pine cones hanging in a tree"

- PINE
  1. kinds of evergreen tree with needle-shaped leaves
  2. waste away through sorrow or illness

Pine#1 ∩ Sentence = 1
Pine#2 ∩ Sentence = 0

Example: disambiguate PINE CONE

- PINE
  1. kinds of evergreen tree with needle-shaped leaves
  2. waste away through sorrow or illness
- CONE
  1. solid body which narrows to a point
  2. something of this shape whether solid or hollow
  3. fruit of certain evergreen trees

Pine#1 ∩ Cone#1 = 0
Pine#2 ∩ Cone#1 = 0
Pine#1 ∩ Cone#2 = 1
Pine#2 ∩ Cone#2 = 0
Pine#1 ∩ Cone#3 = 2
Pine#2 ∩ Cone#3 = 0

Lesk Algorithm for More than Two Words?

- *I saw a man who is 98 years old and can still walk and tell jokes*
  - nine open class words: *see(26), man(11), year(4), old(8), can(5), still(4), walk(10), tell(8), joke(3)*

- 43,929,600 sense combinations! How to find the optimal sense combination?
- Simulated annealing (Cowie, Guthrie, Guthrie 1992)
  - Let's review (from CS1571)

Lesk Algorithm-Problems

- Few dictionary entries are relatively short.
- The words used in the context and their definitions must have **direct overlap with the words contained in the appropriate sense definition in order to be useful.**

- **Solution:**
- Expand the list of words used in the classifier to include words related to, but not contained in their individual sense definitions.
- **Ex. The word *deposit does not occur in the definition of bank . However, bank* does occur in the definition of *deposit. Therefore, the classifier for bank can* be expanded to include *deposit as a relevant feature.***

- ***Walkers and Wilks algorithms works on this principle of expanding dictionary definition***

Resnik Algorithm

- Parse the sentence in the training corpus to extract syntactic relations such as subject-verb, verb-object and adj-noun.
- Let n has k senses $s_1,...s_k$
- For each of these k senses, Resnik method computes a class C_i as
- $C_i=\{c \mid c \text{ is an ancestor of } s_i\}$
- $a_i=\max(A_R(v, c)) == \max(\Sigma\ count_R(v,w)/|classes(w)|)$
  - Where $count_R(v,w)$ – no of times the word w occurs in syntactic relation R with v
  - Class(w) – no of classes to which the word w belongs

Resnik Algorithm - Example

- I would like to drink coffee (disambiguate coffee)
- Syntactic relation between drink and coffee : verb-object
- The noun coffee has 4 senses in Wordnet
  - Beverage, tree, seed and colour
- Parse training corpus to get all occurrences of drink involved in a verb-object relation
- Let the corpus contains 3 occurrences drink tea, drink milk and drink cocoa
- These 4 nouns coffee, tea, milk and cocoa have multiple senses

Resnik Algorithm - Example

- Coffee
  - Beverage, seed, tree, colour (WordNet Senses)
- Tea
  - Beverage, meal, leaf, party, shrub
- Milk
  - Beverage, body-fluid, river, nutrient
- Cocoa
  - Beverage, powder

- For each sense of coffee find associations
  - a1= ¼ + 1/5 + ¼ + ½ = 1.2 (highest value for the sense beverage of coffee)
  - a2 = ¼ = 0.25
  - a3 = ¼ = 0.25
  - a4 = ¼ = 0.25