CS 181: Natural Language Processing

Lecture 20: Word Sense Disambiguation

Kim Bruce
Pomona College
Spring 2008

Disclaimer: Slide contents borrowed from many sources on web!

Final Project

- Progress Report due on Thursday
  - Written report
  - Oral report (< 5 minutes)
- Guest lecture on Information Retrieval next Tuesday by Professor Sood.

Word Disambiguation

- Used Thesaurus and relations
  - hyponym, hypernym, meronym, ...
  - Look for sense definition overlap w/context (Lesk)
  - Use similarity measures to determine similarity w/neighboring words to get senses of all.
- Talked about bootstrapping when minimally supervised.

Unsupervised Disambiguation

Unsupervised Disambiguation

- No dictionaries, labeled training text, etc.
- Don't label senses.
- Instead cluster contexts to discriminate between groups
- "You shall know a word by the company it keeps" -- Firth
- Warning: If remove sense tags may not rediscover same classes!

Unsupervised Disambiguation

- Hypothesis: same sense of words will have similar words in context
- Algorithm:
  - Identify context vectors for all occurrences of the word.
  - Partition into regions of high density
  - Assign a sense to each region

Unsupervised Disambiguation

- Example:
  - Sit on a chair.
  - Take a seat on this chair.
  - The chair of the CS department
  - The chair of the committee

The Problem

- Large corpora of data
- Typically one targeted word per context
- Does not attempt to assign senses to clusters
- Find the targeted words that occur in most similar contexts and place in cluster

Agglomerative Clustering

- Represent context by feature vector.
- Create similarity matrix where entry (i,j) is the similarity score between contexts i & j
- Start w/ each instance in its own cluster
- Form cluster from most similar instances
- Continue until have desired # clusters
- *Expensive to look at all pairs!*

Example

	arts	boil	data	function	large	sugar	summarized	water
apricot	0	1	0	0	1	1	0	1
pineapple	0	1	0	0	1	1	0	1
digital	0	0	1	1	1	0	1	0
information	0	0	1	1	1	0	1	0

Feature Vectors

- Find small number (<30) features
  - Morphological form of target word
  - POS of 2 words to left and right of target
  - co-occurrences w/most frequent content word
  - Most frequent content words to left or right of target
  - Ignore stopwords
  - Parsing can help find better neighbors:
    - direct objects, subjects, indirect objects, etc.

Measuring Similarity

- Distance between feature vectors:
  - Euclidean: $d_{euclid}(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}$
  - Manhattan: $d_{manh}(\vec{x}, \vec{y}) = \sum_{i=1}^{N}|x_i - y_i|$
  - Don't work well in practice

Measuring Similarity

- Count up # matching entries
- Measure angle between vectors:
  - $sim_{cos}(\vec{v}, \vec{w}) = \frac{\vec{v} \cdot \vec{w}}{|\vec{v}||\vec{w}|}$
  - Answer between -1 and 1, but normally between 0 (orthogonal) and 1 (same).

More Similarity

- Jaccard similarity:

$$sim_{Jaccard}(\vec{v}, \vec{w}) = \frac{\sum_{i=1}^{n} min(v_i, w_i)}{\sum_{i=1}^{n} max(v_i, w_i)}$$

- Dice similarity:

$$sim_{Dice}(\vec{v}, \vec{w}) = \frac{2\sum_{i=1}^{n} min(v_i, w_i)}{\sum_{i=1}^{n} v_i + w_i}$$

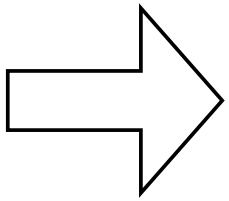
Simple Example

	P-2	P-1	P+1	P+2	fish	check	river	interest
S1	adv	det	prep	det	Y	N	Y	N
S2		det	prep	det	N	Y	N	Y
S3	det	adj	verb	det	Y	N	N	N
S4	det	noun	noun	noun	N	N	N	N

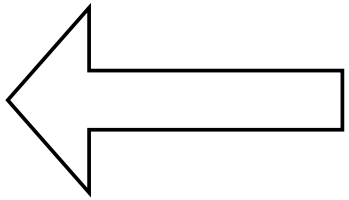
	S1	S2	S3	S4
S1		3	4	2
S2	3		2	0
S3	4	2		1
S4	2	0	1	

Average Link Clustering

	S1	S2	S3	S4
S1		3	4	2
S2	3		2	0
S3	4	2		1
S4	2	0	1	



	S13	S2	S4
S13		2.5	1.5
S2	2.5		0
S4	1.5	0	



	S123	S4
S123		1.5
S4	1.5	

Computational Discourse

What is Discourse?

- Consider coherent groups of sentences.
- Stick w/monologues for now
- Cover dialogs in Chapter 24

Discourse Segmentation

Discourse Segmentation

- Useful in summarizing documents
  - News broadcast into separate stories
  - Pronominal resolution
  - Help with information retrieval
- Cohesion: use of linguistic devices to link together textual units.
  - Lexical cohesion: based on words
- Skip here

Coherence

Coherence

- Different sentences of discourse must relate to each other.
  - *John didn't come to class today. He was sick.*
    - Explanation
  - *John didn't come to class today. He wasn't there yesterday either. (or Neither did Alex.)*
    - Parallel or elaboration
  - *John didn't come to class today. The teacher sent him e-mail.*
    - Result

Coherence

- Can parse discourse into tree based on relations between sentences.
- Subtrees form locally coherent clauses/ sentences called discourse segment.
- Rhetorical structures similar.

Automatic Coherence Assignment

- Can use *cue phrases*
  - John went home *because* he felt sick.
  - Identify cue phrases in text.
  - Break into discourse segments, using cue phrases.
  - Classify relationship between consecutive phrases, using cue phrases.

Automatic Coherence Assignment

- Finding cue phrases a bit tricky.
  - With his last test completed, he was ready to go home.
  - He took his test with his calculator.
- Break into discourse segments, using cue phrases.
  - Use hand-written rules based on punctuation & sentence boundaries.
- Unfortunately many coherence relations not signaled by cue phrases:
  - I don't want to study; I want to sleep!
- Try bootstrapping!

Reference Resolution

Coreference Resolution

- Input:
  - Today, Secretary of State Colin Powell met with ... he ... Condoleeza Rice ... Mr. Powell ... she ... Powell ... President Bush ... Rice ... Bush ...
- Output: (3 entities)
  - Secretary of State Colin Powell, he, Mr. Powell, Powell.
  - Condoleeza Rice, she, Rice
  - President Bush, Bush

Noun Phrase Coreference

- Identify all noun phrases that refer to the same entity.
- Object being referred to is *referent*.
- Natural language expression is *referring expression*.
- Two referring expressions that refer to the same entity are said to *corefer*.

Pronouns

- Reference to an entity already introduced called *anaphora*.
- Pronoun is *licensed* by previous mention of an *antecedent*.
- Pronoun resolution subset of general reference resolution.

Discourse Model

- Need to keep track of conversational context, esp. hearer's mental model of the discourse.
- Changes over time.
- When referent introduced, say it is *evoked*.
- When it is mentioned again, say *accessed*.

Coreference Resolution

- Look for set of coreferring expressions
  - Coreference chain
- *A boy was hit by a car. The poor kid broke his arm. The driver was arrested when he had no license.*
  - {A boy, the poor kid, his}
  - {The driver, he}

Pronominal Anaphor Resolution

- Coreference resolution: find all referring expressions in discourse and group into coreference chains.
- Anaphora resolution: find antecedent for single pronoun. *Subtask of coreference resolution.*

Referring Expressions

- Indefinite Noun Phrases
  - Introduce entities into discourse context
    - John is going to buy *a new car*. *specific or non-specific*
    - *Three boys* knocked at her door.
    - *Some flowers* blew in the wind.
- Definite Noun Phrases
  - Refers to entity that is identifiable to hearer
    - I'm sure that *his car* will be very cool!
    - *Her mother* turned *the boys* away.
    - *The President of Pomona* is giving a speech today.

Referring Expressions

- Pronouns
  - Another form of definite reference
    - *They* went home sadly.
    - *It* will need to provide him with reliable transportation
    - Jane was sad her mother turned *them* away.
- Demonstratives (this, that, these, those)
  - Can appear alone or as determiners
    - *That* boy is quite tall.
    - *This* is not a good situation.

Referring Expressions

- Names
  - proper names
    - *Lee* went to the store
    - *General Motors* had a bad year.

Information Status/ Structure

- Givenness scale:
  - in focus > activated > familiar > uniquely identifiable
    {it} *{this, that}* *{that N}* *{the N}*

    > referential > type identifiable
    {indef, this N} *{a N}*
- Accessibility scale
  - Full name > long def. descrip. > short def. descr. > last name > first name > distal demonstrative > proximate demonstrative > NP > stressed pronoun > unstressed pronoun

Information Status/ Structure

- Hearer status
  - Whether previously known to the hearer or new
- Discourse status
  - Whether previously mentioned in discourse or new

Complicating Factors

- Inferrables:
  - I wanted to take CS 181, but *the time* didn't work. *Time not previously introduced!*
  - The class was a disaster because *a student* fell asleep and snored. *Doesn't introduce a new student*
- Generic:
  - Computer Science graduates must work hard. *They* must keep learning or become obsolete. *Generic, refers to class of all CS grads*
  - In California, *you* must be prepared for earthquakes. *Generic "you"*

Complicating Factors

- Non-referential uses:
  - It's hailing.
  - It is smart to go to bed on time.
  - *What is "it"?*

Antecedent Game

- Constraints on antecedents:
  - Number agreement.
    - John his a ball. He threw *them* far.
    - *but:*
      - Microsoft released a new version of Windows today.
        They hope it will be more successful than Vista.
  - Person agreement
    - 1st, 2nd, 3rd person match
  - Gender agreement
    - he/she/it

Antecedent Game

- Binding theory constraints:
  - John bought himself an ice cream.
  - John bought him an ice cream
  - John said that Bill bought him an ice cream
  - John said that Bill bought himself an ice cream
  - He said that he bought Bill an ice cream
  - *Constraints on meaning of him, himself, he.*

Antecedent Game

- Selectional restrictions:
  - John ate his sandwich in his office.
    - It was made with roast beef.
    - It was quieter than eating in the snack bar.
- Recency:
  - Lee met Mary for lunch. They saw Sue at the restaurant. She gave Lee a hug.
- Grammatical role: *Subject > object*
  - Jane saw Sally at the market. She went over to say hello.

Antecedent Game

- Repeated mention:
  - John had a long day. He had not gotten much sleep the night before. He and Fred went to the movies that night. He had a hard time staying awake.
- Parallelism
  - Jane helped Mary with her Physics homework. Ellen helped her with her English.
- Verb Semantics:
  - Jane gave Mary the letter.
    - She was excited to receive it.
    - She had received it yesterday.

Algorithms for Pronominal Anaphora Resolution

Hobbs Algorithm

Any Questions?