Introduction to Natural Language Processing

Chapter 17

Outline

Papers

Project Evaluation

Review Lexical Semantics

- Meaning of Words, Lexical Relations, WordNet
- Thematic Roles, FrameNet
- Selectional Restrictions
- Conceptual Dependency

Word Sense Disambiguation

- constraint satisfaction approaches
- dictionary approaches
- supervised, semi-supervised, and unsupervised machine learning

Word Sense Disambiguation

For any given lexeme, how can its sense be reliably distinguished?

The problem of *word sense disambiguation* is to determine the correct sense of each word in a sentence.

Is word sense disambiguation like part of speech disambiguation (tagging)?

WSD vs. POS Tagging

Similarities

- many words have only one sense
- the most frequent words have many senses
- the frequency distribution of senses within a word is heavily skewed
- assigning the most frequent sense is not horrible

WSD vs. POS Tagging

Differences

POS

- with POS tagging there is a small fixed set of tags that are candidates for all the input words
- there is wide consensus as to what a POS tagset should look like
- evaluation is easy, if you get it wrong you're wrong

WSD

- with WSD there may be a large open-ended set of tags specific to each word (i.e., 100K vocabulary may have 10 times as many tags!)
- there is less consensus as to what a word sense should be
- evaluation is harder

WSD: The Problem

Associating tokens with meanings

Potentially extremely useful problem, but . . .

Very difficult to formulate and evaluate

Semantic parsers need this information

Approaches

There are currently many approaches to this problem, with no clear winner yet

- frequency-based
- constraint satisfaction approaches
- rule-based approaches
- dictionary approaches
- supervised machine learning
- semi-supervised machine learning
- unsupervised machine learning
- parallel corpora

Representational Preliminaries

We will refer to the word being disambiguated as the *target word*.

Almost all of the approaches rely on the use of a simple feature modeling of the context in which the target word appears.

We'll see that these feature encodings make use of simple things

- the identify of the lexemes at specific positions to the left and right of the target
- the parts of speech of the words in these positions

Frequency-Based WSD

If you have a corpus in which each word is annotated with its sense, you can collect unigram statistics (count the number of times each sense occurs in the corpus)

E.g., if you have

- 5845 uses of the word bridge,
- 5641 cases in which it is tagged with the sense STRUCTURE
- 194 instances with the sense DENTAL-DEVICE

As in the case of POS tagging, frequency-based WSD gets about 70% correct

To improve upon these results, need context

Constraint Satisfication

As we have seen, our meaning representations typically assume a predicate / argument structure, with related constraints.

This suggests three cases

- disambiguate an argument by using the constraints from an unambiguous predicate
- disambiguation a predicate by using the constraints from an ambiguous argument
- mutual disambiguation of an argument and a predicate

Selectional Restriction WSD

Example: disambiguate *arguments* using predicates

"In our house, everybody has a career and none of them includes washing *dishes*," he says.

In her tiny kitchen at home, Ms. Chen works efficiently, stir-frying several simple *dishes*, including braised pig's ears and chicken livers with green peppers.

- *wash*, *stir-fry* have different selectional restrictions on their arguments

WordNet Entry for "dish"

The noun "dish" has 6 senses in WordNet.

1. dish – a piece of dishware normally used for holding or serving food; "we gave them a set of dishes for a wedding present"
2. dish – a particular item of prepared food; "she prepared a special dish for dinner"
3. dish, dishful – the quantity that a dish will hold; "they served me a dish of rice"
4. smasher, stunner, knockout, beauty, sweetheart peach, lulu, looker, mantrap, dish – a very attractive or seductive looking woman
5. dish, dish aerial, dish antenna, saucer - a parabolic reflector for microwave or radio frequency radiation that is used as a directional antenna
6. cup of tea, bag, dish - an activity that you like or at which you are superior; "chemistry is not my cup of tea"; "his bag now is learning to play golf"; "marriage was scarcely his dish"

Another Example

Disambiguate *predicates* using arguments

- Well, there was the time they *served* green-lipped mussels from New Zealand.
- Which Airlines *serve* Denver?
- Which ones *serve* breakfast?

In this example, the object determines the meaning of sense, i.e., we can disambiguate *serve* by its thematic role

- first sense requires some kind of food
- second sense requires some kind of geographical or political entity
- third sense requires some kind of meal designator

WordNet Entry for "serve"

The verb "serve" has 13 senses in WordNet.

1. serve, function – serve a purpose, role, or function; "The tree stump serves as a table"; "The female students served as a control group"; "This table would serve very well"; "His freedom served him well"; "The table functions as a desk"
2. serve – do duty or hold offices; serve in a specific function; "He served as head of the department for three years"; "She served in Congress for two terms"; "They served as medics in Vietnam"
3. serve - contribute or conduce to; "The scandal served to increase his popularity."
4. service, serve - be used by; as of a utility; "The sewage plant served the neighboring communities"; "The garage served to shelter his horses"
5. serve, help - help to some food; help with food or drink; "I served him three times, after that he helped himself"
6. serve, serve up, dish out, dish up, dish – provide food; "We serve meals for the homeless"; "She dished out the soup at 8 P.M."

Final Example

In some other cases, both the predicate and the argument can have multiple senses (mutual disambiguation)

- I am looking for a restaurant that *serves* vegetarian *dishes*.
- look at word net listings again

What about

- What kind of *dishes* do you recommend?

Problems

Problems with the selectional restriction approach to WSD

- scaling up to large numbers of words (WordNet helps with this)
- getting the details of the restrictions correct
- dealing with metaphorical uses that violate the constraints
- the need to parse to get the verb-argument information needed to make it work

Modified Earley Algorithm

Need to modify the ENQUEUE function to enforce selectional restrictions.

This is accomplished by comparing the selection restrictions associated with the arguments to typed information associated with the actual argument.

Needed: hierarchical type information about the arguments and selection restrictions on the argument roles of predicates.

First: hypernym information from WordNet.

Second; associated synsets with the arguments to each predicate-bearing lexical item.

Problems (continued)

But it fell apart in 1931, perhaps because people realized that you can't *eat* gold for lunch if you're hungry.

In his two championship trials, Mr. Kulkarni *ate* glass on an empty stomach, accompanied only by water and tea.

If you want to *kill* the Soviet Union, get it to try to *eat* Afghanistan.

Dictionary-Based Approaches

Another approach to dealing with scale is to use a dictionary

Current electronic dictionaries provide target set of senses

Algorithm

- retrieve all the target word definitions from the dictionary
- compare these definitions to the definitions of all the other words in the context.
- pick the sense of the target word with the highest overlap

Using Machine-Readable Dictionaries

What sense of *cone* is in *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

Answer: sense 3 (due to *evergreen*, *tree*)

Dictionaries: Problems

The dictionary approach is quite brittle. It depends entirely on the identical overlap in the immediate definitions of the context words.

Robust WS Disambiguation

In *supervised machine learning* approaches, we make use of a sense-tagged corpus and train a system to make the same judgements as are reflected in the training set.

The training data consists of a feature encoding of the context found with the target word.

This can consist of collocational information, or co-occurrence information, or both.

Example

The noun “bass” has 8 senses in WordNet.
1. bass - (the lowest part of the musical range)
2. bass, bass part - (the lowest part in polyphonic music)
3. bass, basso - (an adult male singer with the lowest voice)
4. sea bass, bass - (flesh of lean-fleshed saltwater fish of the family Serranidae)
5. freshwater bass, bass - (any of various North American lean-fleshed freshwater fishes especially of the genus Micropterus)
6. bass, bass voice, basso - (the lowest adult male singing voice)
7. bass - (the member with the lowest range of a family of musical instruments)
8. bass - (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Inputs: Feature Vectors

An electric guitar and *bass* player stand off to one side ...

Collocation

- $\{guitar, NN1, and, CJC, player, NN1, stand, VVB\}$

Co-occurance

- The 12 most frequent content words in the *bass* sentences in the corpus are *fishing, big, sound, player, fly, rod, pound, double, runs, playing, guitar, band*
- Assuming a window size of 10, $\{0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0\}$

In general, the feature vectors that have been used consider a context, and information such as POS, morphology, parsing.

Disambiguators as Classifiers

Typically take a vector representing a sense use in context and assign it to the class representing the right use of that word.

Once we cast the WSD problem as a classification problem, then all sorts of well-studied methods become available

- Naive Bayes
- decision trees (e.g., C5)
- decision rules (e.g., RIPPER)
- neural networks
- MBR (Timbl)
- ...

Naive Bayes

Ideas from before

- $ArgmaxP(s \mid V)$: choosing the most probable sense s given vector V.
- rewriting this using $Bayes\ Rule$ yields $argmax\frac{P(V|s)P(s)}{P(V)}$

Where do the numbers come from?

- a tagged corpus
- e.g., the probability of *guitar* occuring one position to the right of a each sense of the word *bass* is computed from the corpus

Tree/List Supervised Algorithms

Inputs

- class and feature (symbolic, continuous, sets) definitions
- labeled training data

Output: a classification model

Decision Lists

Like naive Bayes, decision list classifiers are simple to learn/train and are often effective.

They are similar to most case statements in programming languages, i.e., they consist of an ordered set of conditions with simple conclusions.

Rule		Sense
fish within window	$\Rightarrow$	**bass**^{1}
striped bass	$\Rightarrow$	**bass**^{1}
guitar within window	$\Rightarrow$	**bass**^{2}
bass player	$\Rightarrow$	**bass**^{2}
piano within window	$\Rightarrow$	**bass**^{2}
tenor within window	$\Rightarrow$	**bass**^{2}
sea bass	$\Rightarrow$	**bass**^{1}
play/V bass	$\Rightarrow$	**bass**^{2}
river within window	$\Rightarrow$	**bass**^{1}
violin within window	$\Rightarrow$	**bass**^{2}
salmon within window	$\Rightarrow$	**bass**^{1}
on bass	$\Rightarrow$	**bass**^{2}
bass are	$\Rightarrow$	**bass**^{1}

Learning Decision Lists

Informally, learning decision lists is similar to decision tree learning, and uses a hill-climbing search through the sapce of possible lists.

Search space is defined by complexity of the tests you allow and the length of the list you are willing to consider.

The search is guided by some metric such as information gain.

Learning (continued)

How do you pick a classification method?

Most classifiers perform similarly for this task.

They differ in their end product (how inspectable), how much data they need, how long they take to run ...

Evaluation

Evaluation

- error rate on unseen labeled test set (or, cross-validated estimate)
- baseline: most frequent sense
- partial credit?
  - tagging to a dictionary sense
  - binary tagging
  - how do you choose?

Evaluation (continued)

How do you choose how to evaluate?

As we have seen before, it is often based on your application. A text to speech system needs to know if it is a [baes] or a [beys]. It doesn't need to know if the sense is a singer or a bass fiddle. Other applications might though.

Other Approaches

Bootstrapping (semi-supervised algorithms)

- provide seeds (recall Roark and Charniak paper)

Unsupervised Algorithms

- *discover* word senses
- clustering, self-organizing maps, ...

Parallel Corpora

Bootstrapping

In a lot of domains like this one, there is an easier way to get some training data.

This data can also be used as fodder for previously discussed ML algorithms.

For example, think of two good words that you think should occur with each of two senses of $bass$.

One Sense Per Collocation

Extracted *bass* sentences from the WSJ using *play* and *fish*

Klucevsek **play**s Giulietti or Titano piano accordions with the more flexible, more difficult free **bass** rather than the traditional Stradella **bass** with its preset chords designed mainly for accompaniment.

We need more good teachers – right now, there are only a half a dozen who can **play** the free **bass** with ease.

An electric guitar and **bass play**er stand off to one side, not really part of the scene, just as a sort of nod to gringo expectations perhaps.

When the New Jersey Jazz Society, in a fund-raiser for the American Jazz Hall of Fame, honors this historic night next Saturday, Harry Goodman, Mr. Goodman's brother and **bass play**er at the original concert, will be in the audience with other family members.

The researchers said the worms spend part of their life cycle in such **fish** as Pacific salmon and striped **bass** and Pacific rockfish or snapper.

Associates describe Mr. Whitacre as a quiet, disciplined and assertive manager whose favorite form of escape is **bass fish**ing.

And it all started when **fish**ermen decided the striped **bass** in Lake Mead were too skinny.

Though still a far cry from the lake's record 52-pound **bass** of a decade ago, "you could fillet these **fish** again, and that made people very, very happy," Mr. Paulson says.

Saturday morning I arise at 8:30 and click on "America's best-known **fish**erman," giving advice on catching **bass** in cold weather from the seat of a bass boat in Louisiana.

Unsupervised Methods

The problem with the approaches just discussed is that they require a large amount of training data.

Unsupervised learning methods attempt to both learn the senses, and learn to classify senses from an untagged corpus.

Typical methods essentially cluster the feature vectors and call the clusters sense.

How to evaluate?

Cue Phrase Sense Disambiguation

Sentential sense

- **Harry:** You know I see more coupons *now* than I've ever seen before and I'll bet you have too.

Discourse sense

- **Doris:** I have a couple quick questions about the income tax. The first one is my husband is retired and on social security and in '81 he ... few odd jobs for a friend uh around the property and uh he was reimbursed for that to the tune of about 640 dollars. *Now* where would he where would we put that on the form?

Both senses

- *Now now* that we have all been welcomed here it's time to get on with the business of the conference.

ML, using prosodic and orthographic features, can disambiguate (Litman 96)

For Next Time

Chapter 18

Homework 3 Assigned