6.891: Lecture 18 (November 12th, 2003)

Word Sense Disambiguation

Overview

- A supervised method: decision lists
- A partially supervised method
- A first hierarchical method
- A second hierarchical method

Words in Context

Sense	Examples (keyword in context)
1	. . . used to strain microscopic plant life from the . . .
1	. . . too rapid growth of aquatic plant life in water . . .
2	. . . automated manufacturing plant in Fremont . . .
2	. . . discovered at a St. Louis plant manufacturing . . .

- **The task:** given a word in context, decide on its word sense

Examples

Examples of words used in [Yarowsky, 1995]:

Word	Senses
plant	living/factory
tank	vehicle/container
poach	steal/boil
palm	tree/hand
axes	grind/tools
sake	benefit/drink
bass	fish/music
space	volume/outer
motion	legal/phsyical
crane	bird/machine

Features Used in the Model

- Word found in $+/-k$ word window
- Word immediately to the right (+1 W)
- Word immediately to the left (-1 W)
- Pair of words at offsets -2 and -1
- Pair of words at offsets -1 and +1
- Pair of words at offsets +1 and +2

Features Used in the Model

- Also maps words to parts of speech, and general classes (e.g., WEEKDAY, MONTH etc.)

- Local features including word classes are added:
  - Pair of tags at offsets -2 and -1
  - Tag at position -2, word at position -1
  - etc.

An Example

The ocean reflects the color of the sky, but even on cloudless days the color of the ocean is not a consistent blue. Phytoplankton, microscopic plant life that floats freely in the lighted surface waters, may alter the color of the water. When a great number of organisms are concentrated in an area, the plankton changes the color of the ocean surface. This is called a 'bloom.'

$$\Downarrow$$

w_{-1} = Phytoplankton	t_{-1} = JJ
w_{+1} = life	t_{+1} = NN
w_{-2}, w_{-1} = (Phytoplankton,microscopic)	t_{-2}, t_{-1} = (NN,JJ)
w_{-1}, w_{+1} = (microscopic,life)	...
w_{+1}, w_{+2} = (life,that)	
word-within-k = ocean	
word-within-k = reflects	
word-within-k = color	
...	
word-within-k = bloom	

A Machine-Learning Method: Decision Lists

- For each feature, we can get an estimate of conditional probability of sense 1 and sense 2
- For example, take the feature w_{+1} =`life`
- We might have

$$\begin{aligned} Count(\text{sense 1 of plant}, w_{+1} = \texttt{life}) &= 100 \\ Count(\text{sense 2 of plant}, w_{+1} = \texttt{life}) &= 1 \end{aligned}$$

- Maximum-likelihood estimate

$$P(\text{sense 1 of plant} \mid w_{+1} = \texttt{life}) = \frac{100}{101}$$

Smoothed Estimates

- Usual problem: some counts are sparse

- We might have

$$\begin{aligned} Count(\text{sense 1 of plant}, w_{-1} = \texttt{Phytoplankton}) &= 2 \\ Count(\text{sense 2 of plant}, w_{-1} = \texttt{Phytoplankton}) &= 0 \end{aligned}$$

- α smoothing (empirically, $\alpha \approx 0.1$ works well):

$$\begin{aligned} P(\text{sense 1 of plant} \mid w_{-1} = \texttt{Phytoplankton}) &= \frac{2+\alpha}{2+2\alpha} \\ P(\text{sense 1 of plant} \mid w_{+1} = \texttt{life}) &= \frac{100+\alpha}{101+2\alpha} \end{aligned}$$

  with $\alpha = 0.1$, gives values of 0.95 and 0.99 (unsmoothed gives values of 1 and 0.99)

Creating a Decision List

- For each feature, find

  $$sense(feature) = \text{argmax}_{sense} P(sense \mid feature)$$

  e.g., $sense(w_{+1} = \texttt{life}) = sense1$

- Create a rule `feature` $\rightarrow$ $sense(feature)$ with weight $P(sense(feature) \mid feature)$. e.g.,

Rule			Weight
w_{+1} =`life`	$\rightarrow$	sense 1	0.99
w_{-1} =`Phytoplankton`	$\rightarrow$	sense 1	0.95
$\ldots$			

Creating a Decision List

- Create a list of rules sorted by strength

Rule			Weight
w_{+1} =life	$\rightarrow$	sense 1	0.99
w_{-1} =manufacturing	$\rightarrow$	sense 2	0.985
word-within-k=life	$\rightarrow$	sense 1	0.98
word-within-k=manufacturing	$\rightarrow$	sense 2	0.979
word-within-k=animal	$\rightarrow$	sense 1	0.975
word-within-k=equipment	$\rightarrow$	sense 2	0.97
word-within-k=employee	$\rightarrow$	sense 2	0.968
w_{-1} =assembly	$\rightarrow$	sense 2	0.965
...			

- To apply the decision list: take the first (strongest) rule in the list which applies to an example

The ocean reflects the color of the sky, but even on cloudless days the color of the ocean is not a consistent blue. Phytoplankton, microscopic plant life that floats freely in the lighted surface waters, may alter the color of the water. When a great number of organisms are concentrated in an area, the plankton changes the color of the ocean surface. This is called a 'bloom.'

Feature	Sense	Strength
w_{-1} = Phytoplankton	1	0.95
w_{+1} = life	1	0.99
w_{-2}, w_{-1} = (Phytoplankton,microscopic)	N/A	
w_{-1}, w_{+1} = (microscopic,life)	N/A	
w_{+1}, w_{+2} = (life,that)	1	0.96
word-within-k = ocean	1	0.93
word-within-k = reflects	N/A	
word-within-k = color	2	0.65
t_{-1} = JJ	2	0.56
t_{-2}, t_{-1} = (NN,JJ)	2	0.7
t_{+1} = NN	1	0.64
...		

- N/A $\Rightarrow$ feature has not been seen in training data
- w_{+1} = life $\rightarrow$ `Sense 1` is chosen

Experiments

- [Yarowsky, 1994] applies the method to accent restoration in French, Spanish

De-accented form	Accented form	Percentage
cesse	cesse	53%
	cessé	47%
coute	coûte	53%
	coûté	47%
cote	côté	69%
	côte	28%
	cote	3%
	coté	$< 1\%$

- Task is to recover accents on words
  - Very easy to collect training/test data
  - Very similar task to word-sense disambiguation
  - Useful for restoring accents in de-accented text, or in automatic generation of accents while typing

Overview

- A supervised method: decision lists
- A partially supervised method
- A first hierarchical method
- A second hierarchical method

A Partially Supervised Method

- Collecting labeled data can be **expensive**

- We'll now describe an approach that uses a small amount of labeled data, and a large amount of unlabeled data

A Key Property: Redundancy

The ocean reflects the color of the sky, but even on cloudless days the color of the ocean is not a consistent blue. Phytoplankton, microscopic plant life that floats freely in the lighted surface waters, may alter the color of the water. When a great number of organisms are concentrated in an area, the plankton changes the color of the ocean surface. This is called a 'bloom.'

$\Downarrow$

w_{-1} = Phytoplankton	word-within-k = ocean
w_{+1} = life	word-within-k = reflects
w_{-2}, w_{-1} = (Phytoplankton,microscopic)	word-within-k = bloom
w_{-1}, w_{+1} = (microscopic,life)	word-within-k = color
w_{+1}, w_{+2} = (life,that)	...

There are often many features which indicate the sense of the word

Another Useful Property: “One Sense per Discourse”

- Yarowsky observes that if the same word appears more than once in a document, then it is very likely to have the same sense every time

Step 1 of the Method: Collecting Seed Examples

- Goal: start with a small subset of the training data being labeled

- Various methods for achieving this:
  - Label a number of training examples by hand
  - Pick a single feature for each class by hand
    e.g., `word-within-k=bird` and
    `word-within-k=machinery` for *crane*
  - Look through frequently occurring features, and label a few of them
  - Using words in dictionary definitions
    e.g., Pick words in the two definitions for "plant"

    A vegetable organism, or part of one, ready for planting or lately planted.

    equipment, machinery, apparatus, for an industrial activity

An example: for the “plant” sense distinction, initial seeds are `word-within-k=life` and `word-within-k=manufacturing`

Partitions the unlabeled data into three sets:

- 82 examples labelled with “life” sense
- 106 examples labelled with “manufacturing” sense
- 7350 unlabeled examples

Training New Rules

1. From the seed data, learn a decision list of all rules with weight above some threshold (e.g., all rules with weight > 0.97)

2. Using the new rules, relabel the data (usually we will now end up with more data being labeled)

3. Induce a new set of rules with weight above the threshold from the labeled data

4. If some examples are still not labeled, return to step 2

Experiments

- Yarowsky describes several experiments:
  - A baseline score for just picking the most frequent sense for each word
  - Score for a fully supervised method
  - Partially supervised method with "two words" as a seed
  - Partially supervised method with dictionary defn. as a seed
  - Partially supervised method with hand-chosen rules as a seed
  - Dictionary defn. method combined with one-sense-per-discourse constraint

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
			%		Seed Training Options			(7) + OSPD		
Word	Senses	Samp. Size	Major Sense	Supvsd Algrtm	Two Words	Dict. Defn.	Top Colls.	End only	Each Iter.	Schütze Algrthm
plant	living/factory	7538	53.1	97.7	97.1	97.3	97.6	98.3	98.6	92
space	volume/outer	5745	50.7	93.9	89.1	92.3	93.5	93.3	93.6	90
tank	vehicle/container	11420	58.2	97.1	94.2	94.6	95.8	96.1	96.5	95
motion	legal/physical	11968	57.5	98.0	93.5	97.4	97.4	97.8	97.9	92
bass	fish/music	1859	56.1	97.8	96.6	97.2	97.7	98.5	98.8	–
palm	tree/hand	1572	74.9	96.5	93.9	94.7	95.8	95.5	95.9	–
poach	steal/boil	585	84.6	97.1	96.6	97.2	97.7	98.4	98.5	–
axes	grid/tools	1344	71.8	95.5	94.0	94.3	94.7	96.8	97.0	–
duty	tax/obligation	1280	50.0	93.7	90.4	92.1	93.2	93.9	94.1	–
drug	medicine/narcotic	1380	50.0	93.0	90.4	91.4	92.6	93.3	93.9	–
sake	benefit/drink	407	82.8	96.3	59.6	95.8	96.1	96.1	97.5	–
crane	bird/machine	2145	78.0	96.6	92.3	93.6	94.2	95.4	95.5	–
AVG		3936	63.9	96.1	90.6	94.8	95.5	96.1	96.5	92.2

4 after the algorithm has converged, or in Step 3c after each iteration.

At the end of Step 4, this property is used for error correction. When a polysemous word such as *plant* occurs multiple times in a discourse, tokens that were tagged by the algorithm with low confidence using local collocation information may be overridden by the dominant tag for the discourse. The probability differentials necessary for such a reclassification were determined empirically in an early pilot study. The variables in this decision are the total number of occurrences of *plant* in the discourse (n), the number of occurrences assigned to the majority and minor senses for the discourse, and the cumulative scores for both (a sum of log-likelihood ratios). If cumulative evidence for the majority sense exceeds that of the minority by a threshold (conditional on n), the minority cases are relabeled. The case $n = 2$ does not admit much reclassification because it is unclear which sense is dominant. But for $n \geq 4$, all but the most confident local classifications tend to be overridden by the dominant tag, because of the overwhelming strength of the one-sense-per-discourse tendency.

The use of this property after each iteration is similar to the final post-hoc application, but helps prevent initially mistagged collocates from gaining a foothold. The major difference is that in discourses where there is substantial disagreement concerning which is the dominant sense, *all* instances in the discourse are returned to the residual rather than merely leaving their current tags unchanged. This helps improve the purity of the training data.

The fundamental limitation of this property is coverage. As noted in Section 2, half of the examples occur in a discourse where there are no other instances of the same word to provide corroborating evidence for a sense or to protect against misclassification. There is additional hope for these cases, however, as such isolated tokens tend to strongly favor a particular sense (the less "bursty" one). We have yet to use this additional information.

8 Evaluation

The words used in this evaluation were randomly selected from those previously studied in the literature. They include words where sense differences are realized as differences in French translation (drug → drogue/médicament, and duty → devoir/droit), a verb (poach) and words used in Schütze's 1992 disambiguation experiments (tank, space, motion, plant).[10]

The data were extracted from a 460 million word corpus containing news articles, scientific abstracts, spoken transcripts, and novels, and almost certainly constitute the largest training/testing sets used in the sense-disambiguation literature.

Columns 6-8 illustrate differences in seed training options. Using only two words as seeds does surprisingly well (90.6 %). This approach is least successful for senses with a complex concept space, which cannot be adequately represented by single words. Using the salient words of a dictionary definition as seeds increases the coverage of the concept space, improving accuracy (94.8%). However, spurious words in example sentences can be a source of noise. Quick hand tagging of a list of algorithmically-identified salient collocates appears to be worth the effort, due to the increased accuracy (95.5%) and minimal cost.

Columns 9 and 10 illustrate the effect of adding the probabilistic one-sense-per-discourse constraint to collocation-based models using dictionary entries as training seeds. Column 9 shows its effectiveness

[10] The number of words studied has been limited here by the highly time-consuming constraint that full hand tagging is necessary for direct comparison with supervised training.

Some Comments

- Very impressive results using relatively little supervision

- How well would this perform on words with "weaker" sense distinctions? (e.g., *interest*)

- Can we give formal guarantees for when this method will/won't work?
  (how to give a formal characterization of redundancy, and show that this implies guarantees concerning the utility of unlabeled data?)

- There are several "tweakable" parameters of the method
  (e.g., the weight threshold used to filter the rules)

- Another issue: the method as described may not ever label all examples

Overview

- A supervised method: decision lists
- A partially supervised method
- A first hierarchical method
- A second hierarchical method

The Structure of Wordnet

- Each sense for a word is associated with a different *synset*
- For example, *chair* might be associated with the synset:

  chair: president, chairman, chairwoman, chair, chairperson – (the officer who presides at the meetings of an organization); “address your remarks to the chairperson”

  definition is in parantheses, **example** is in quotes
- A second synset covers the “furniture” sense of chair

The Structure of Wordnet

- The synsets are organized into an is-a hierarchy
  (i.e. each synset has links to one or more parent synsets)

- At the top of the hierarchy are 26 categories

  person, communication, artifact, act, group, food, cognition, possession, location, substance, state, time, attribute, object, process, Tops, phenomenon, event, quantity, motive, animal, body, feeling, shape, plant, relation

[Ciaramita and Johnson, 2003]

- **The task:** for new/unknown words (not in wordnet, for example), classify them into 1 of the 26 "supersenses"

- Example words: *irises*, *exane*

- Motivation: using WordNet 1.6, 1 in 8 sentences have a noun not seen in Wordnet (Wordnet 1.6 lists 95,000 noun types)

- Note also: WordNet has around 65,000-75,000 different synsets at the most specific level

Creating Training Data

- Used a 40 million word corpus (parsed with Charniak's parser)
- From WordNet 1.6, tagged all nouns which are *unambiguous* (by doing this, create a labelled dataset)
- Features used:
  - part of speech of the neighbouring words, single words in the surrounding context, bigrams and trigrams located around the word, syntactic dependencies, spelling/morphological features (prefixes, suffixes etc.)

Creating Test Data

- Test data 1: all nouns in WordNet 1.71 but not in WordNet 1.6, and which are *unambiguous*
  (over 90% of new nouns are unambiguous, giving 744 new noun types, and 9,537 test occurrences)

- Test data 2: 755 noun types removed from training set at random (i.e., taken from WordNet 1.6)

- Learning method: perceptron

- Voting multiple occurrences: for each test noun, for each instance in the data, run the classifier. Choose the *supersense* which is returned by the classifier most frequently

Creating Extra Training Data

chair: president, chairman, chairwoman, chair, chairperson – (the officer who presides at the meetings of an organization); “address your remarks to the chairperson”

Can use examples, e.g., “address your remarks to the chairperson”, as extra training data for the supersense

gives 66,841 extra training instances
(787,186 training instances in the original data)

Results

Method	Token	Type	Test set
Baseline	20.0	27.8	Test 1
AP-B-55	35.9	50.7	
AP-B-65	35.9	50.8	
AP-B-55+WN	36.9	52.9	
Baseline	24.1	21.0	Test 2
AP-B-55	47.4	47.7	
AP-B-65	47.9	48.3	
AP-B-55+WN	52.3	53.4	

- Baseline result is to pick "person" every time
- AP-B-55 is original training data
- AP-B-55+WN is original training data + "extra" wordnet data
- AP-B-65 is original training data + extra training data

Overview

- A supervised method: decision lists
- A partially supervised method
- A first hierarchical method
- A second hierarchical method

The Senseval Data

Senseval 2/ACL 01 data:

- 8.611 paragraphs annotated: each contain an ambiguous word whose synset sense is annotated

- Test set is 4,328 examples

[Ciarimata, Hofmnn and Johnson, 2003]

- Method applied to noun data:
  3,512 training, 1,754 test instances

- Same feature set used as before

- They again use a “two-level” hierarchy:
  synset of the word, and its *supersense*

The Task as a Global Linear Model

- Defining the set of possible labels, $\mathcal{Y}$:
  each $y \in \mathcal{Y}$ is a (synset,supersense) pair

- e.g. for *chair*, one label could be
  `(synset=presidential-chair,supersense=person)`

- $\mathbf{GEN}(x)$ is set of all possible (synset,supersense) pairs for a word

  e.g. $\mathbf{GEN}(chair) =$
  $\{$`(synset=presidential-chair,supersense=person),`
  `(synset=furniture-chair,supersense=artifact)`$\}$

The Task as a Global Linear Model

- How to define $\mathbf{\Phi}(x, y)$?: Features look at either the synset or supersense

$x =$ Here the quality of the finest chair components is merged with art

$y =$ (synset=furniture-chair,supersense=artifact)

$$\Downarrow$$

$\mathbf{\Phi}(x, y) =$

word-within-k=components;synset=furniture-chair
word-within-k=components;supersense=artifact
w_{-1}=finest;synset=furniture-chair
w_{-1}=finest;supersense=artifact
$\ldots$

Sharing Supersense Data Between Different Words

$x =$ Here the quality of the finest chair components is merged with art
$y =$ `(synset=furniture-chair,supersense=artifact)`

$$\Downarrow$$

```
word-within-k=components;synset=furniture-chair
word-within-k=components;supersense=artifact
```
w_{-1}`=finest;synset=furniture-chair`
w_{-1}`=finest;supersense=artifact`
...

$x =$ components for manufacturing the bass guitar are shipped across the state
$y =$ `(synset=musical-instrument,supersense=artifact)`

$$\Downarrow$$

```
word-within-k=components;synset=musical-instrument
word-within-k=components;supersense=artifact
```
...

Additional Training Data

- Can once again use the examples in WordNet itself:

  $chair_1$ – he put his coat over the back of the chair and sat down

  $chair_2$ – address your remarks to the chairperson

- Not much use at the synset level: only one example per word

- **But:** potentially useful if propagated up to the supersense level e.g., examples for chair (furniture), bass (guitar), car etc. are all used for the "artifact" supersense distinction

Experiments

- Two sets of training data: examples within WordNet (supersense only); Senseval data (synset and supersense annotated)

- Perceptron method used to train the global linear model:
  - At each iteration first pass over the WordNet data (supersense only)
  - Then pass over the Senseval data (synset and supersense)

Summary

- Supervised methods for decision lists (Yarowsky 94)
- A partially supervised method (Yarowsky 95)
- A first hierarchical method: finding supersenses for new words
- A second hierarchical method: using supersenses to improve synset discovery