Word-Sense Disambiguation, and Semi-Supervised Learning

Overview

- A supervised method for word-sense disambiguation: decision lists

- A semi-supervised method for word-sense disambiguation

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
w_{+1} = life
w_{-2}, w_{-1} = (Phytoplankton,microscopic)
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

t_{-1} = JJ
t_{+1} = NN
t_{-2}, t_{-1} = (NN,JJ)
...

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 $w_{+1} = \texttt{life}$

- 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 $\texttt{feature} \rightarrow sense(feature)$ with weight $P(sense(feature) \mid feature)$. e.g.,

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

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
$\dots$			

- 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 for word-sense disambiguation: decision lists

- A semi-supervised method for word-sense disambiguation

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		
		Samp.	Major	Supvsd	Two	Dict.	Top	End	Each	Schütze
Word	Senses	Size	Sense	Algrtm	Words	Defn.	Colls.	only	Iter.	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. 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

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