From Semi-supervised Up to Unsupervised Word Sense Disambiguation (Yarowsky 1995)

Matěj Korvas

Department of Computational Linguistics & Phonetics
Saarland University

April 29, 2011

Outline

1. Motivation

2. Yarowsky's Solution

3. Evaluation

The Problem

Sense	Training Examples (Keyword in Context)
?	. . . company said the *plant* is still operating
?	Although thousands of *plant* and animal species
?	. . . zonal distribution of *plant* life . . . .
?	. . . to strain microscopic *plant* life from the . . .
?	vinyl chloride monomer *plant*, which is . . .
?	and Golgi apparatus of *plant* and animal cells
?	. . . computer disk drive *plant* located in . . .
?	. . . . . .

The Problem

Sense	Training Examples (Keyword in Context)
?	. . . company said the *plant* is still operating
?	Although thousands of *plant* and animal species
?	. . . zonal distribution of *plant* life . . . .
?	. . . to strain microscopic *plant* life from the . . .
?	vinyl chloride monomer *plant*, which is . . .
?	and Golgi apparatus of *plant* and animal cells
?	. . . computer disk drive *plant* located in . . .
?	. . . . . .

The Problem

Sense	Training Examples (Keyword in Context)
?	. . . company said the *plant* is still operating
?	Although thousands of *plant* and animal species
?	. . . zonal distribution of *plant* life . . . .
?	. . . to strain microscopic *plant* life from the . . .
?	vinyl chloride monomer *plant*, which is . . .
?	and Golgi apparatus of *plant* and animal cells
?	. . . computer disk drive *plant* located in . . .
?	. . . . . .

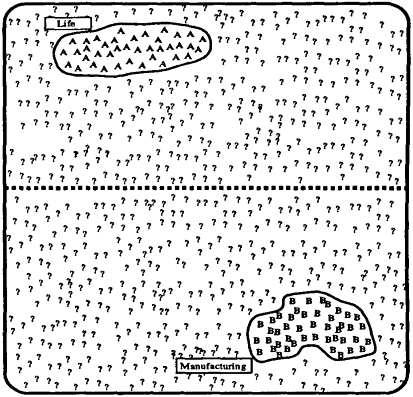
The Problem

Sense	Training Examples (Keyword in Context)
?	... company said the *plant* is still operating
?	Although thousands of *plant* and animal species
?	... zonal distribution of *plant* life . ...
?	... to strain microscopic *plant* life from the ...
?	vinyl chloride monomer *plant*, which is ...
?	and Golgi apparatus of *plant* and animal cells
?	... computer disk drive *plant* located in ...
?	... ...

How to distinguish between the senses??

(We don't want to annotate it all manually. We even want to work as little as possible.)

The Solution



The Initial State

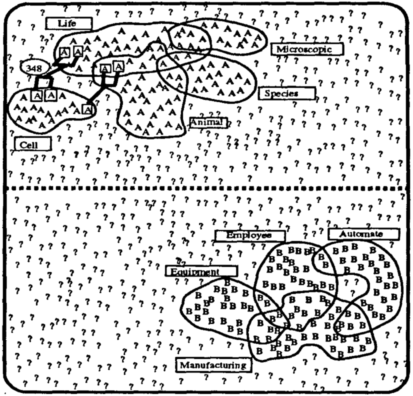
? . . . unclassified occurrence

A . . . occurrence having sense *A*

B . . . occurrence having sense *B*

Life . . . occurrences with “life” in their context (call this a *pattern* “life”)

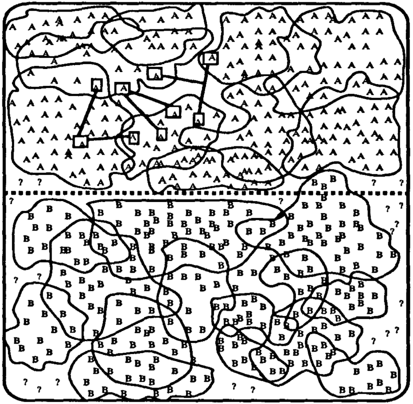
The Solution



An Intermediate State

- changed context words
- changed sense assignment

The Solution



The Final State

- some occurrences not disambiguated – *residual set*
- overlap of patterns:
    - a condition for using SSL
    - ensures cohesion of the output classes

Outline

1 Motivation

2 Yarowsky's Solution

3 Evaluation

Overall Idea

a scientist

Overall Idea

a scientist

→

A:

life

B:

manufacturing

C:

. . .

seed patterns

Overall Idea

a scientist

A:

life

B:

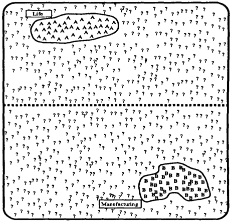
manufacturing

C:

. . .

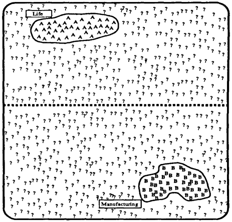
seed patterns

matching



occurrences of A, B, C. . .

Overall Idea



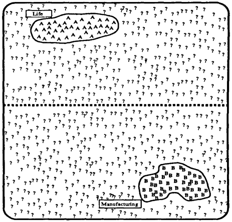
occurrences of A, B, C...

Overall Idea

LogL	Collocation	Sense
8.10	*plant* life	⇒ A
7.58	manufacturing *plant*	⇒ B
7.39	life (within *k* words)	⇒ A
7.20	manufacturing (in *k* words)	⇒ B
6.27	animal (within *k* words)	⇒ A
	. . .	

decision list

extracting collocations



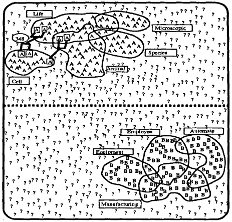
occurrences of A, B, C. . .

Overall Idea

LogL	Collocation	Sense
8.10	*plant* life	⇒ A
7.58	manufacturing *plant*	⇒ B
7.39	life (within *k* words)	⇒ A
7.20	manufacturing (in *k* words)	⇒ B
6.27	animal (within *k* words)	⇒ A
	. . .	

decision list

matching



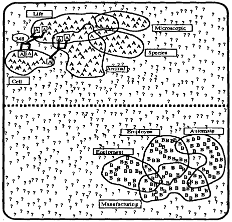
occurrences of A, B, C. . .

Overall Idea

LogL	Collocation	Sense
10.12	*plant* growth	⇒ A
9.68	car (within *k* words)	⇒ B
9.64	*plant* height	⇒ A
9.61	union (within *k* words)	⇒ B
9.54	equipment (within *k* words)	⇒ B
	. . .	

decision list

extracting collocations



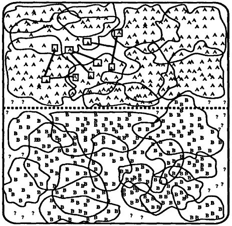
occurrences of A, B, C. . .

Overall Idea

LogL	Collocation	Sense
10.12	*plant* growth	⇒ A
9.68	car (within *k* words)	⇒ B
9.64	*plant* height	⇒ A
9.61	union (within *k* words)	⇒ B
9.54	equipment (within *k* words)	⇒ B
	...	

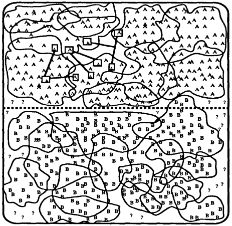
decision list

matching



occurrences of A, B, C...

Overall Idea



occurrences of A, B, C. . .

The Decision List

patterns:

- Collocate + type of the collocation (adjacent × in wider context).
- Weighted with their indicativeness:

$$\log\left(\frac{\Pr(\mathit{Sense}_A \mid \mathit{Collocation}_i)}{\Pr(\mathit{Sense}_B \mid \mathit{Collocation}_i)}\right).$$

If the quantity is above a threshold, the pattern enters the decision list.

decision list: Only the first matching pattern is considered.

- Supports hard classification.
- No probabilistic weighting of patterns – simple, efficient.

The Other Rule

- So far, we only considered

$$\text{meaning} \sim \text{collocational pattern.}$$

- BUT, there is a strong tendency for retaining the same sense also per discourse:

Word	Senses	Accuracy	Applicability
plant	living/factory	99.8 %	72.8 %
tank	vehicle/container	99.6 %	50.5 %
palm	tree/hand	99.8 %	38.5 %
crane	bird/machine	100.0 %	49.1 %
...	...	...	...
Average		99.8 %	50.1 %

SAARLAND UNIVERSITY

The Resulting Algorithm

1. Retrieve contexts of all occurrences of *w*.
2. Identify a few training examples (*seeds*).
3.
   a. Extract patterns (the decision list).
   b. Classify all examples.
   c. Impose the one-sense-per-discourse (OSPD) constraint.
4. Repeat step 3 until stable.

5. Use the decision list as a classifier.
   Optionally, impose the the OSPD constraint also here.

Obtaining Training Examples

Seed patterns can be obtained:

- manually: from the researcher's intuition;

Obtaining Training Examples

Seed patterns can be obtained:

- manually: from the researcher's intuition;
- automatically:

Obtaining Training Examples

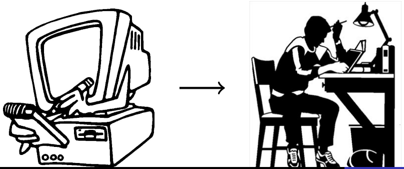
Seed patterns can be obtained:

- manually: from the researcher's intuition;
- automatically:
  - from dictionary definitions (take most indicative words from the definition)
  - from an ontology (WordNet) – the defining hypernym (crane: “bird”, or “machine”)

Obtaining Training Examples

Seed patterns can be obtained:

- manually: from the researcher's intuition;
- automatically:
    - from dictionary definitions (take most indicative words from the definition)
    - from an ontology (WordNet) – the defining hypernym (crane: "bird", or "machine")
- half-automatically:



Obtaining Training Examples

Seed patterns can be obtained:

- manually: from the researcher's intuition;
- automatically:
  - from dictionary definitions (take most indicative words from the definition)
  - from an ontology (WordNet) – the defining hypernym (crane: "bird", or "machine")
- half-automatically:
  1. find indicative collocates in the corpus – automatically
  2. select the valid ones – by a human

Outline

1. Motivation
2. Yarowsky's Solution
3. Evaluation

Results

The algorithm achieves impressively good results on a set of 12 words, for which it could be compared to earlier algorithms.

corpus size: 460 M words

comparison of training options (the accuracy):

two words: 90.6 %
dictionary def.: 94.8 %
top collocations: 95.5 %

accuracy using OSPD constraint:

only after training: 96.1 %
after each iteration: **96.5 %**

Baseline *supervised* algorithm (Schütze, 1992): 92.2 % acc-cy.

Summary

- Collocates and discourse disambiguate word sense so strongly, that the simple *decision list* suffices as the central structure.
- The problem can be solved even without supervision with a great accuracy.