Unsupervised Word Sense Disambiguation

Yarowsky (1995)

Introduction

- Word Sense Disambiguation: Word Sense Disambiguation (WSD) is the process of identifying the sense of a polysemic word.
- Polysemic word: *Clear* (to brighten? unclutter?…)
- Aim: To use an unsupervised algorithm for WSD that will rival supervised techniques.
- Why: Supervised techniques require hand/human annotation; therefore are time-consuming.

Background

- Premise: Exploits two properties of human language
- One sense per collocation and One sense per discourse
- One sense per collocation: Nearby words give strong (and consistent) clues to the sense of the target word
- One sense per discourse: The sense of the target word is consistent within any given document

One sense per discourse

- Gale, Church and Yarowsky(1992): Words strongly exhibit only one sense in a given discourse/ document.
- Does not use this as a hard constraint. If local evidence is stronger, it can be overridden.

One sense per collocation

- Collocation: Words appearing in the same location. Has not considered idiomatic sense etc..
- Yarowsky(1993) observed and quantified that words exhibit one sense in a given collocation.
- Strongest with adjacent collocations, and weakens with distance. Strong with words in a predicate-argument relationship, content words.
- Properties are highly reliable, therefore useful for WSD

One sense per collocation

- Yarowsky(1994): Supervised algorithm based on the 'One sense per collocation' property

- Training procedure: Calculates the probability *Pr(Sense | Collocation),* and orders them by log likelihood ratio:

$$Log \frac{Pr(Sense_a \mid Collocation)}{Pr(Sense_b \mid Collocation)}$$

- Integrates evidence sources (POS, inflected forms..) with positional relationships (trigram sequences, predicate argument association..) using a decision list algorithm

- **sense different from collocation!

Unsupervised learning algorithm

- Seed collocations: Accurately represent SenseA and SenseB of a word.

- For example: Present - noun sense: day(SenseA) OR gift (SenseB).

- Words occur not only in collocations that indicate sense, but multiple such collocations.

- For example: For Present, ‘time’ and ‘day’ are both collocations that could indicate the same sense (day).

- Demonstrated on 7538 instances of plant, a polysemous word in an untagged corpus.

Unsupervised learning algorithm: Step1

- Given a large corpus, identify all polysemous words
- Store contexts as lines in an untagged training set.

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 ...
?	... divide life into *plant* and animal kingdom
?	... close-up studies of *plant* life and natural
?	... Nissan car and truck *plant* in Japan is ...
?	... keep a manufacturing *plant* profitable without
?	... molecules found in *plant* and animal tissue
?	... union responses to *plant* closures . ...
?	... animal rather than *plant* tissues can be
?	... many dangers to *plant* and animal life
?	company manufacturing *plant* is in Orlando ...
?	... growth of aquatic *plant* life in water ...
?	automated manufacturing *plant* in Fremont ,
?	... Animal and *plant* life are delicately
?	discovered at a St. Louis *plant* manufacturing
?	computer manufacturing *plant* and adjacent ...
?	... the proliferation of *plant* and animal life
?	... ...

Unsupervised learning algorithm: Step 2

- For each sense of the word, identify a small number of training samples representative of that sense.

- How: Dictionary definitions, Single collocate for each class (such as taken from *WordNet*)

- For example: *life* and *manufacturing* are used as seed collocations for two major senses of a plant.

- The remaining examples (85-98%) constitute untagged residual.

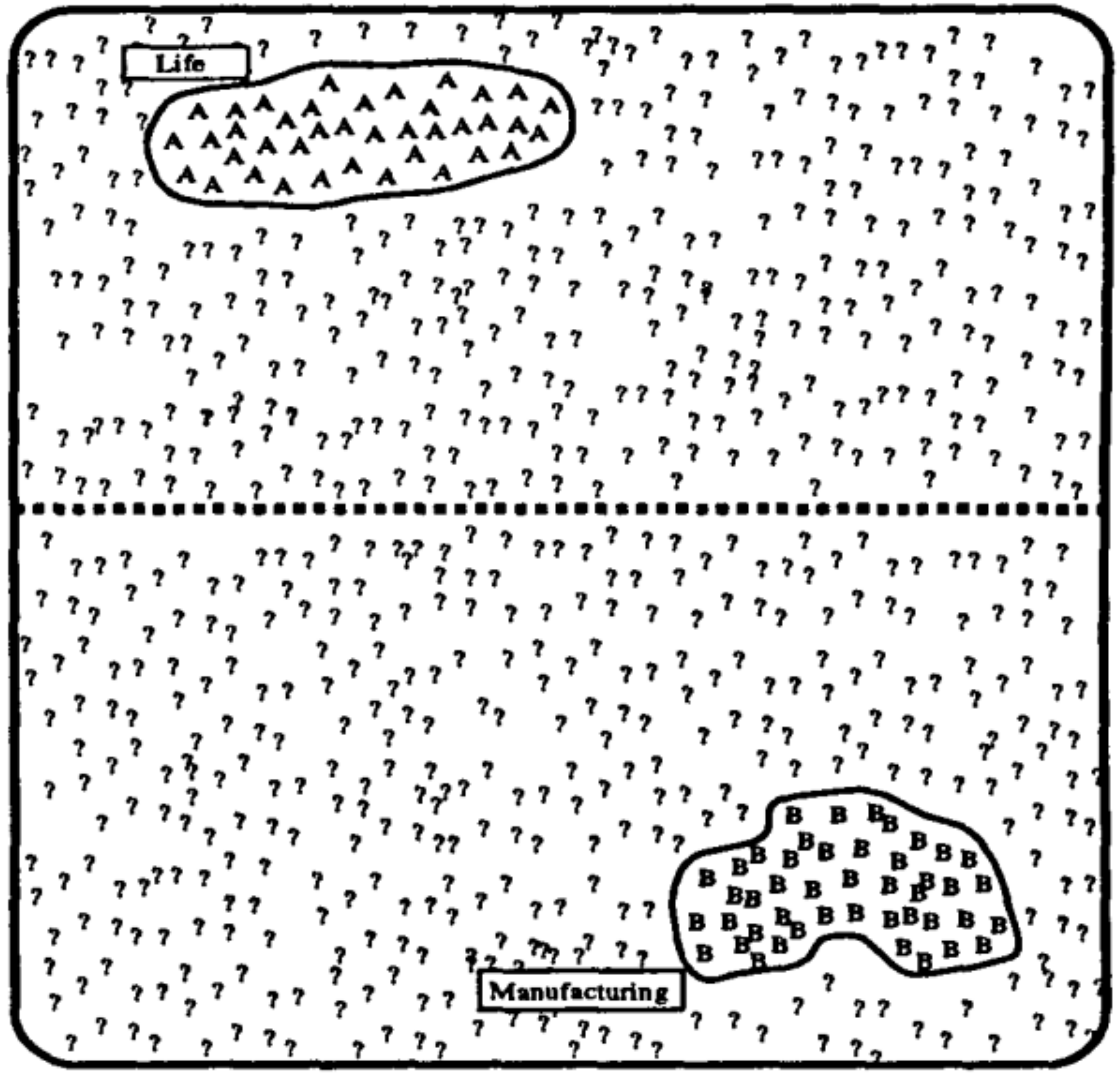


Figure 1: Sample Initial State

A = SENSE-A training example
B = SENSE-B training example
? = currently unclassified training example
$\boxed{\text{Life}}$ = Set of training examples containing the collocation "life".

Unsupervised learning algorithm: Step 3a

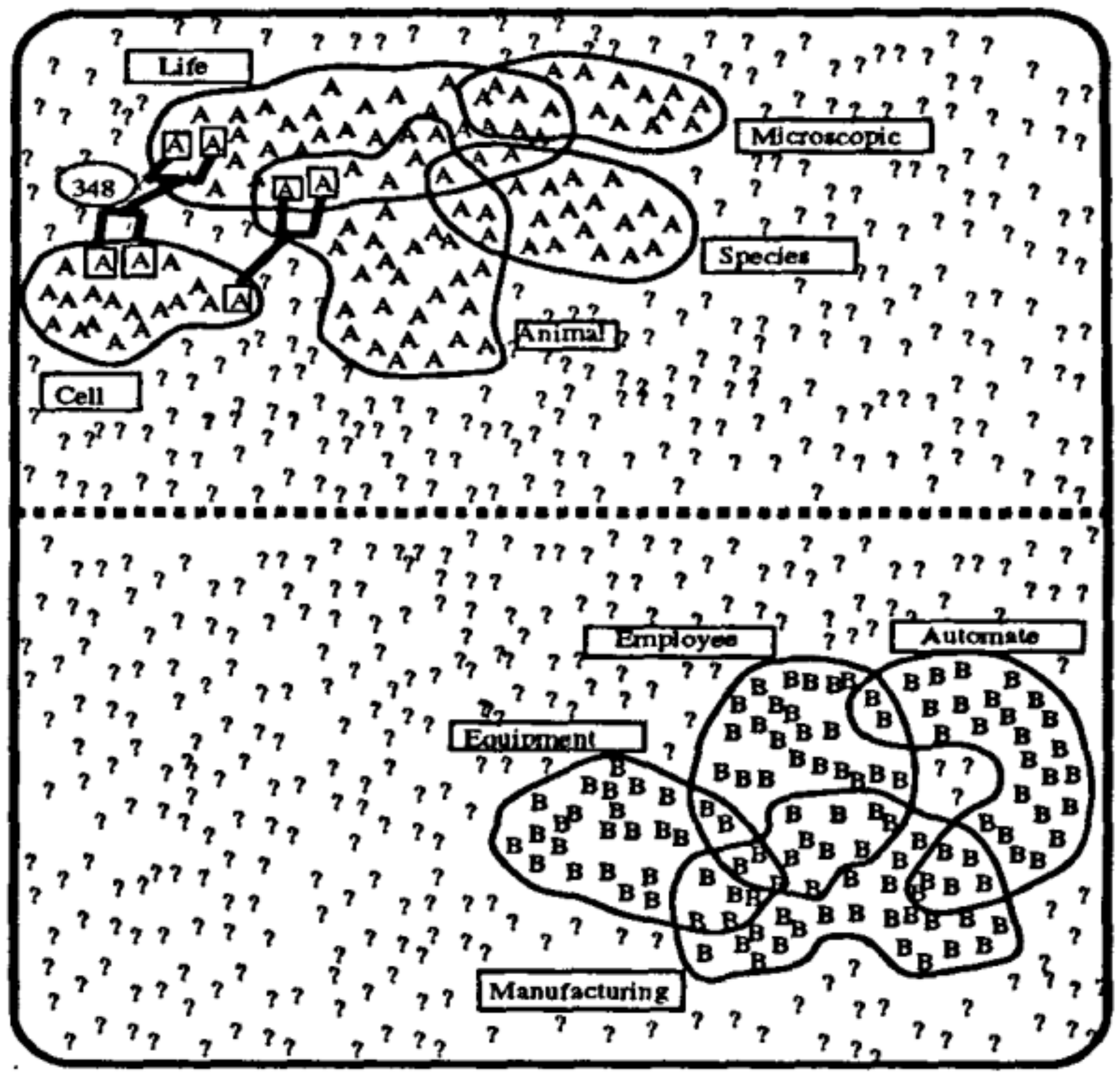
- Train the classification algorithm on SenseA/ SenseB seed sets - Yarowsky(1994).

Initial decision list for *plant* (abbreviated)		
LogL	Collocation	Sense
8.10	*plant* **life**	⇒ A
7.58	**manufacturing** *plant*	⇒ B
7.39	**life** (within ±2-10 words)	⇒ A
7.20	**manufacturing** (in ±2-10 words)	⇒ B
6.27	animal (within ±2-10 words)	⇒ A
4.70	equipment (within ±2-10 words)	⇒ B
4.39	employee (within ±2-10 words)	⇒ B
4.30	assembly *plant*	⇒ B
4.10	*plant* closure	⇒ B
3.52	*plant* species	⇒ A
3.48	automate (within ±2-10 words)	⇒ B
3.45	microscopic *plant*	⇒ A
	...	

Unsupervised learning algorithm: Step 3b

- Apply the resulting classifier to the entire sample set.
- <u>Threshold</u>: Words in the residuals that are tagged as SenseA or Sense B with probability above a certain threshold can be added to the growing seed sets.

- <u>Result</u>: Newly learned collocations that are reliably indicative as the previous trained seed sets.



Unsupervised learning algorithm: Step 3c|3d

- Step 3c: Optionally, use the one sense per discourse constraint.

- If several instances of a polysemous word have been tagged SenseA, then the tag can extend to all examples in the discourse

- Step 3d: Repeat Step 3 iteratively.

- The training set will continue to grow, and the residual will diminish.

Unsupervised learning algorithm: Step 4

- Stop when the training parameters are held constant.
- Most training examples will show multiple collocations indicative of the same sense.
- The one sense per discourse property can also be utilized here, for error correction.

Unsupervised learning algorithm: Step 5

- The classification procedure learned can be applied to new data
- Can be used to annotate the untagged corpus with sense tags and probabilities.

- Original seed words may not remain on top of the list for the final classification.
- They may be displaced by more broadly applicable collocations

Final decision list for *plant* (abbreviated)		
LogL	Collocation	Sense
10.12	*plant* growth	$\Rightarrow$ A
9.68	car (within $\pm k$ words)	$\Rightarrow$ B
9.64	*plant* height	$\Rightarrow$ A
9.61	union (within $\pm k$ words)	$\Rightarrow$ B
9.54	equipment (within $\pm k$ words)	$\Rightarrow$ B
9.51	assembly *plant*	$\Rightarrow$ B
9.50	nuclear *plant*	$\Rightarrow$ B
9.31	flower (within $\pm k$ words)	$\Rightarrow$ A
9.24	job (within $\pm k$ words)	$\Rightarrow$ B
9.03	fruit (within $\pm k$ words)	$\Rightarrow$ A
9.02	*plant* species	$\Rightarrow$ A
...	...	

Evaluation

- 460 million word corpus consisting of news articles, scientific abstracts, spoken transcripts.
- Using 2 seed collocations overall gives the best accuracy for words (avg 90.6%)
- Dictionary definitions as seeds increase accuracy
- OSPD - one sense per discourse constraint

Results

(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

Summary

- An unsupervised algorithm that can accurately disambiguate words in a large untagged corpus.
- Avoids hand-tagging data.
- Self correcting; hence exhibiting strengths of supervised approaches.
- Operates on the assumption that human language has one sense per collocation and one sense per discourse.

Thank you!