UNIVERSITY OF ILLINOIS
URBANA-CHAMPAIGN

Unsupervised Word Sense Disambiguation Rivaling Supervised Methods

David Yarowsky (1995)

By Vincent Medenilla

2/3/2022

Table of Content

Definitions

Sense Disambiguation - is the problem of determining which "sense" (meaning) of a word is activated using the word in a particular context

Collocation – two or more words that tend to appear frequently together

Discourse – any document or a piece of writing

Accuracy – how often a target word (that appeared multiple times) contains one sense (the majority sense) in a document

Applicability – how often a target word appears more than once in a discourse

Polysemous – words having multiple meanings

Seed collocations – initial main/major collocations of a word in a discourse

Background

- Word sense disambiguation had been a major problem in NLP for over forty years (early 1990s)
- Major problem was “sense” vagueness
- Gale, Church, and Yarowsky (1992) utilized parallel text such as Canadian Hansards
- Decision list based on Supervised algorithm (Yarowsky, 1994)

Main Ideas

- Unsupervised algorithm that can disambiguate word senses in a large untagged corpus
- Avoid tedious and time-consuming hand-tagging data training
- Two properties of human language/algorithm constraints:
  1. One sense per collocation
  2. One sense per discourse

One sense per discourse

- words tend to exhibit only one sense in a given discourse (Gale, Church, and Yarowsky)
- tested on a set of 37,232 examples, hand-tagged over 3 years

The one-sense-per-discourse hypothesis:

Word	Senses	Accuracy	Applicblty
plant	living/factory	99.8 %	72.8 %
tank	vehicle/contnr	99.6 %	50.5 %
poach	steal/boil	100.0 %	44.4 %
palm	tree/hand	99.8 %	38.5 %
axes	grid/tools	100.0 %	35.5 %
sake	benefit/drink	100.0 %	33.7 %
bass	fish/music	100.0 %	58.8 %
space	volume/outer	99.2 %	67.7 %
motion	legal/physical	99.9 %	49.8 %
crane	bird/machine	100.0 %	49.1 %
Average		**99.8 %**	**50.1 %**

Literature Review/Theory

One sense per collocation

- observed and quantified by Yarowsky (1993)
- strongest for immediately adjacent collocations and weakens with distance
- stronger with content words than function words
- reliability of 97% for adjacent content words
- Four types of collocation:
    1. the word which collocates with the target word appears in a left window of 2-10 words relatively to the target word
    2. it is the previous word
    3. it is the next word
    4. it appears in a right window of 2-10 words

Unsupervised Learning Algorithm

The algorithm was illustrated by the disambiguation of 7538 instances of polysemous words:

STEP 1: Identify all the polysemous words in a large corpus, storing their contexts as lines in the original (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 possible sense of the word, group a small number of training examples that showcase the sense
- Done by identifying a small number of seed collocations representative of each sense then tagging all training examples containing the seed collocates with the seed's sense label
- The words "life" and "manufacturing" are used as seed collocates for the example shown
- "?" represents untagged residual
- Resulted in 82 examples of living plants (1%), 106 examples of manufacturing (1%), and 7350 residual/unsure (98%)

Sense	Training Examples (Keyword in Context)
A	used to strain microscopic *plant* **life** from the ...
A	... zonal distribution of *plant* **life** . ...
A	close-up studies of *plant* **life** and natural ...
A	too rapid growth of aquatic *plant* **life** in water ...
A	... the proliferation of *plant* and animal **life** ...
A	establishment phase of the *plant* virus **life** cycle ...
A	... that divide **life** into *plant* and animal kingdom
A	... many dangers to *plant* and animal **life** ...
A	mammals . Animal and *plant* **life** are delicately
A	beds too salty to support *plant* **life** . River ...
A	heavy seas, damage , and *plant* **life** growing on ...
A	... ...
?	... vinyl chloride monomer *plant* , which is ...
?	... molecules found in *plant* and animal tissue
?	... Nissan car and truck *plant* in Japan is ...
?	... and Golgi apparatus of *plant* and animal cells ...
?	... union responses to *plant* closures . ...
?	... ...
?	... ...
?	... cell types found in the *plant* kingdom are ...
?	... company said the *plant* is still operating ...
?	... Although thousands of *plant* and animal species
?	... animal rather than *plant* tissues can be ...
?	... computer disk drive *plant* located in ...
B	... ...
B	automated **manufacturing** *plant* in Fremont ...
B	... vast **manufacturing** *plant* and distribution ...
B	chemical **manufacturing** *plant* , producing viscose
B	... keep a **manufacturing** *plant* profitable without
B	computer **manufacturing** *plant* and adjacent ...
B	discovered at a St. Louis *plant* **manufacturing**
B	... copper **manufacturing** *plant* found that they
B	copper wire **manufacturing** *plant* , for example ...
B	's cement **manufacturing** *plant* in Alpena ...
B	polystyrene **manufacturing** *plant* at its Dow ...
B	company **manufacturing** *plant* is in Orlando ...

Unsupervised Learning Algorithm

STEP 3:

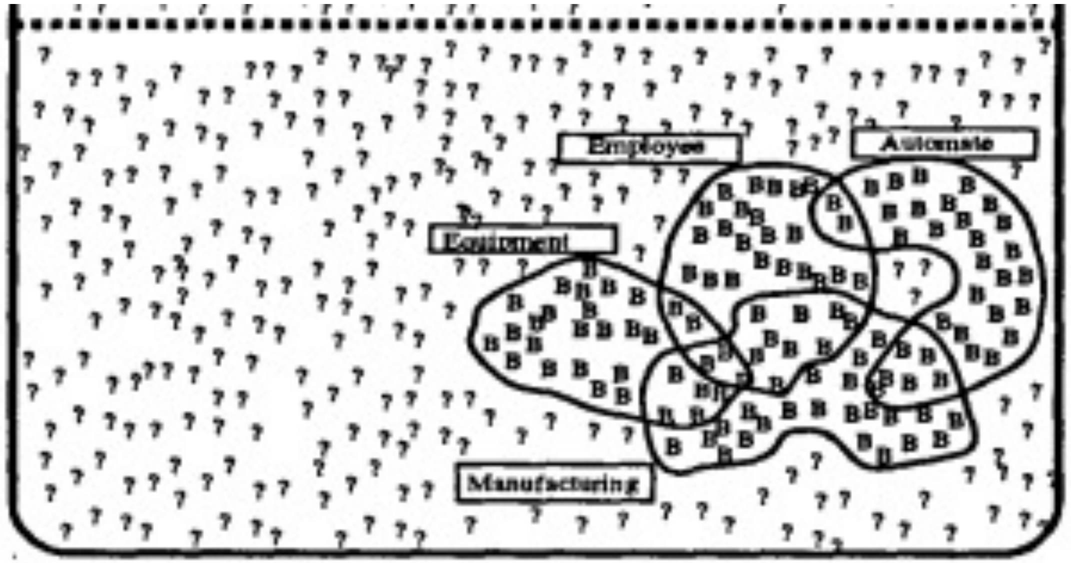
- train the supervised classification algorithm on the Sense A/ Sense B seed sets
- devise a decision list by identifying other collocations that reliably partition the seed training data, ranked by the purity of the distribution
- the purity of distribution is computed for each collocation x and sense A as the log-likelihood ratio for that sense given that collocation: $\log\frac{P(sense-A \mid collocation-x)}{P(sense-B \mid collocation-x)}$, then apply smoothing to avoid 0 values

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
	...	

STEP 3 (cont.):

- Apply the resulting classifier to the whole data set
- Classify the residual/tagged "?" as sense A or sense B with a probability above a certain threshold
- Results in an augmented collocation sets

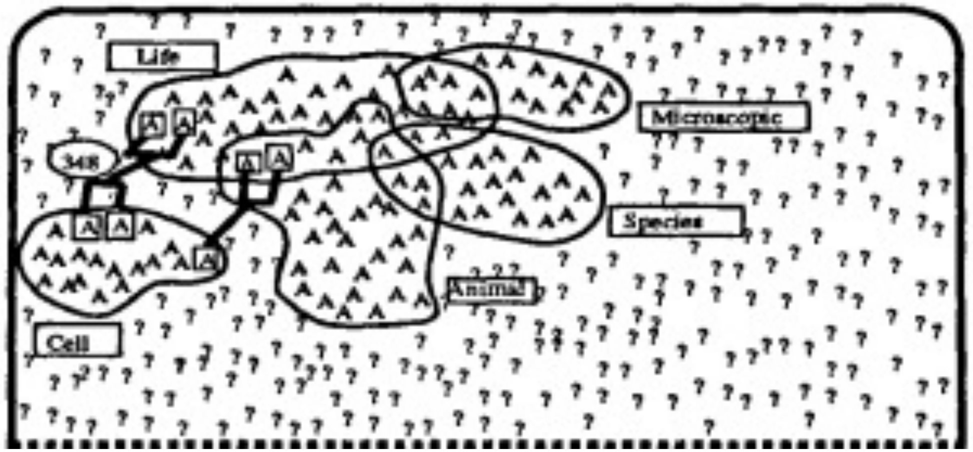


STEP 3 (cont.):

- following the one sense per discourse principle, label previously untagged contexts:

Change in tag	Disc. Numb.	Training Examples (from same discourse)
A → A	724	... the existence of *plant* and animal life ...
A → A	724	... classified as either *plant* or animal ...
? → A	724	Although bacterial and *plant* cells are enclosed
A → A	348	... the life of the *plant* , producing stem
A → A	348	... an aspect of *plant* life , for example
? → A	348	... tissues ; because *plant* egg cells have
? → A	348	photosynthesis, and so *plant* growth is attuned

- may lead to new collocations that might be related to already identified collocations

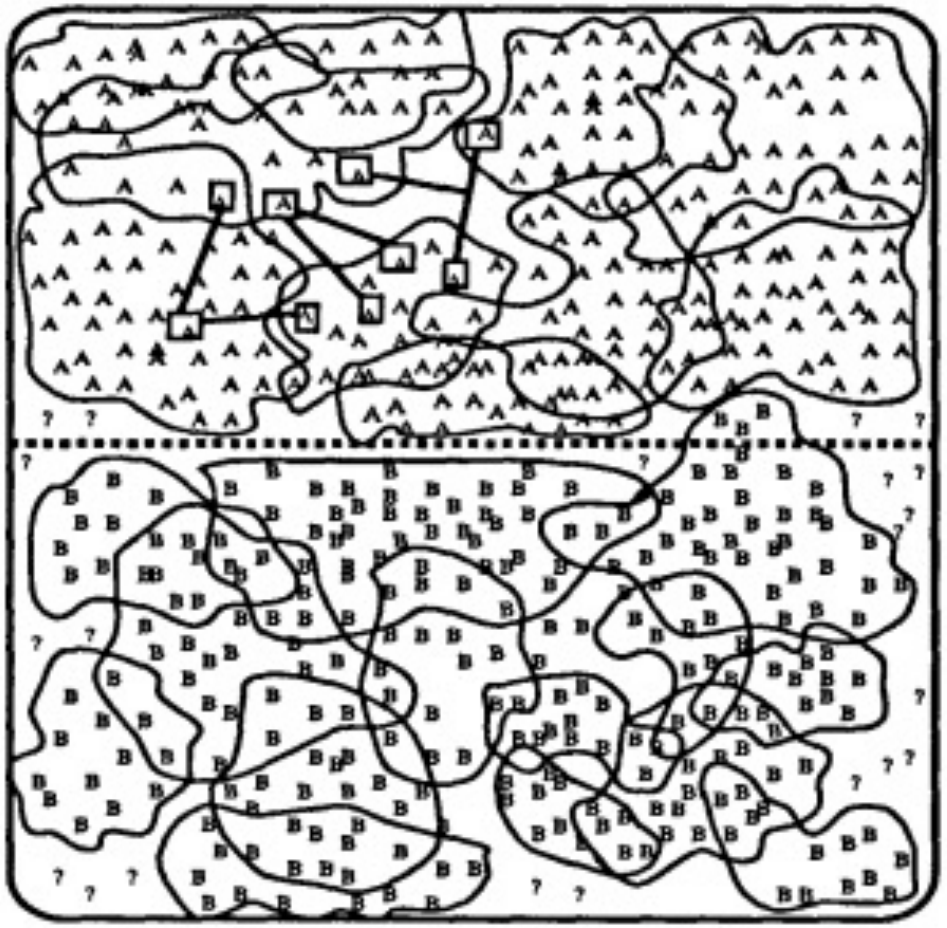


- repeat Step 3 iteratively

Unsupervised Learning Algorithm

STEP 4:

- algorithm converges on a stable residual set
- resolves conflicts by using only the single most reliable piece of evidence, not a combination of related collocations



STEP 5:

- original untagged corpus is then tagged with sense labels and probabilities
- the new model can now be applied to new data
- notice that the original seeds are replaced

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
	...	

⇒

Final decision list for *plant* (abbreviated)

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

Unsupervised Learning Algorithm

Example:

…”the loss of animal and *plant* species through extinction…,”

Based on the final decision list, the collocation “plant species” has a LogL value of 9.02, which means it refers to sense-A which is life or living plant

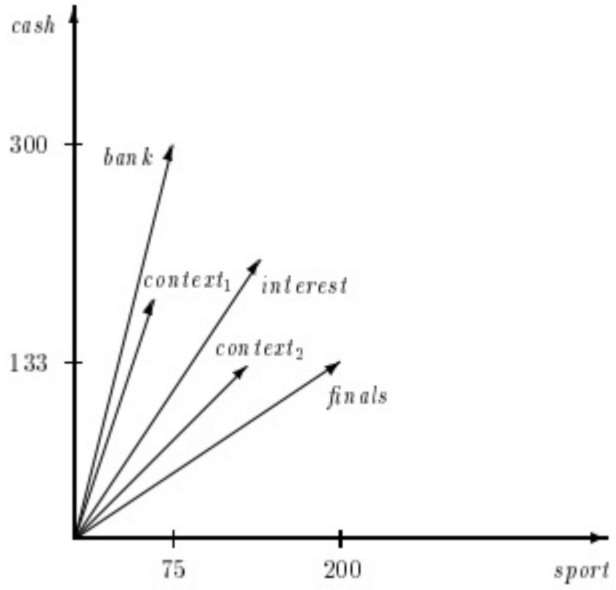
Evaluation

- **Words used were randomly selected from previous literature (Yarowski) like drug = drogue/medicament**
- **Schultze's 1992 disambiguation experiments (tank, space, motion, and plant)**
- **460 million word corpus containing news articles, scientific abstracts, spoken transcripts, and novels**

Schutze's "Dimension of Meaning" Paper (1992):

- unsupervised algorithm, trained on a New York Times corpus
- represented the semantics of words and contexts as vectors
- applied SVD to reduce dimensionality

	bank	*interest*	*finals*
cash	300	210	133
sport	75	140	200



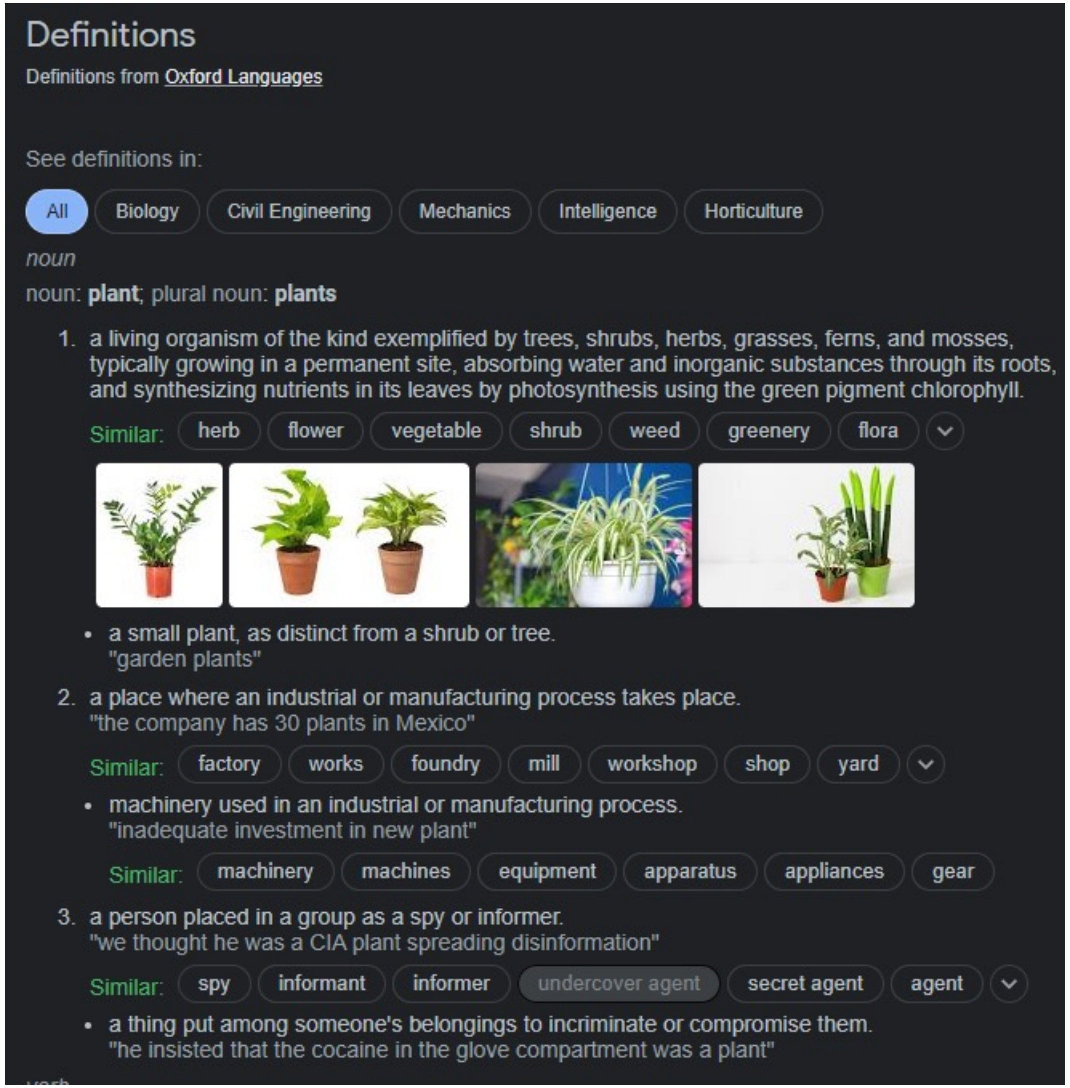
Results

(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

Evaluation

Seed Training Options

1. **Two words: hand-tagged like "plant life" and "manufacturing plant"**
   - **easy to implement but not so robust**
2. **Dictionary definitions: find significantly frequent words w.r.t the most reliable collocational relationships (decision list)**
3. **Top collocates label salient corpus collocates**



One Sense Per Discourse constraint

- **Instead of treating tokens of target word independently, we assume (put bias) that they likely exhibit the same sense**
- **Error correction in step 4**
- **Example: "[discourse is plant life]…<u>sell plants </u>especially locally grown ones…"**

Conclusions

- Utilized one sense per discourse and one sense per collocation properties of language
- Outperformed Schultze's unsupervised algorithm (96.7% to 92.2%) on 4 words
- Achieved relatively same performance as Supervised model (95.5% to 96.1%)
- Shown better results with one sense per discourse restraint (96.5% to 96.1%)
- The model successfully shown improvement from supervised word-sense disambiguation's tedious hand-tagging

Final Thoughts/Discussions

- Were One-sense-per-collocations and one-sense-per-discourse fair assumptions/properties?
- What about small corpus?

The Grainger College of Engineering

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN