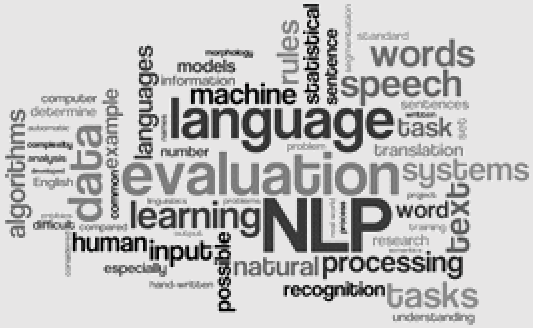
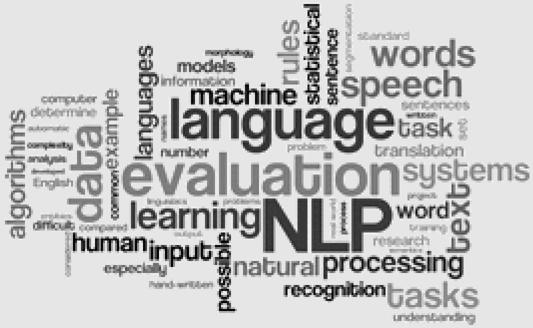


Unsupervised word sense disambiguation rivalling supervised methods – D. Yarowsky

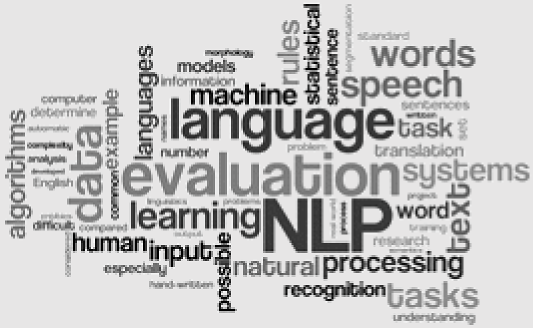
Presentation prepared by Nicholas Mifsud

Contents



- Background
  - Supervised vs. Unsupervised Learning
  - Word sense disambiguation
  - Language properties
- Unsupervised Model
  - Steps involved in the training
- Evaluation
- Conclusion & Future Work

Supervised vs. Unsupervised Learning



- Both seek to infer a classification function from data
- Able to map new examples based on inferred function

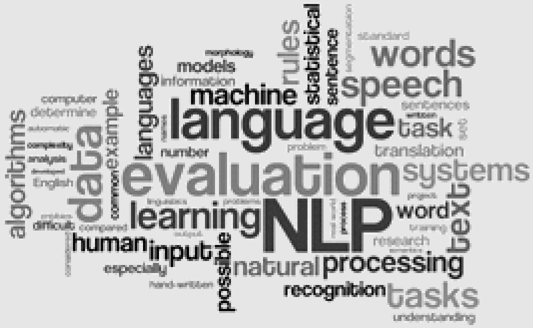
- Supervised
  - Have tagged data consisting of pairs
  - Have an error function

- Unsupervised
  - No tags found in data
  - Bases decision on trends found in the data itself

Word sense disambiguation

- One word – multiple meanings
- Determine meaning through context

Word sense disambiguation

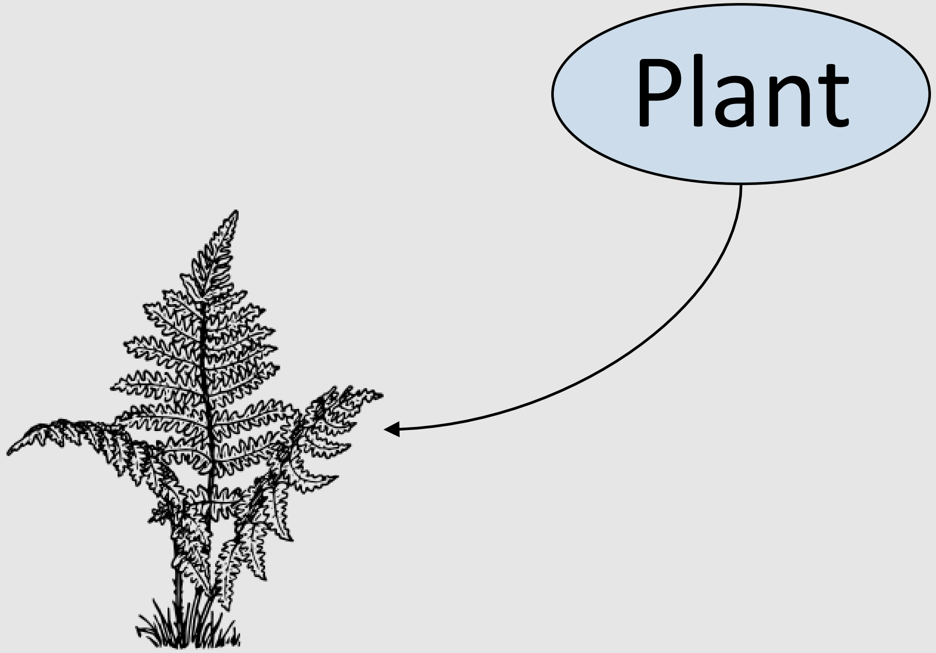


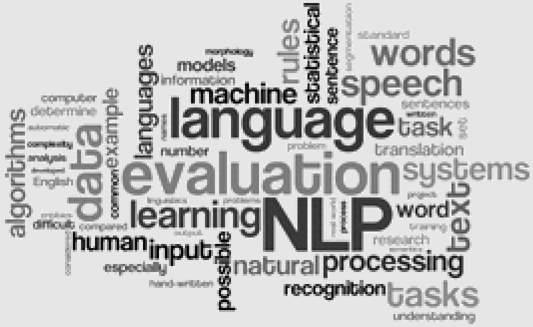
- One word – multiple meanings
- Determine meaning through context

Plant

Word sense disambiguation

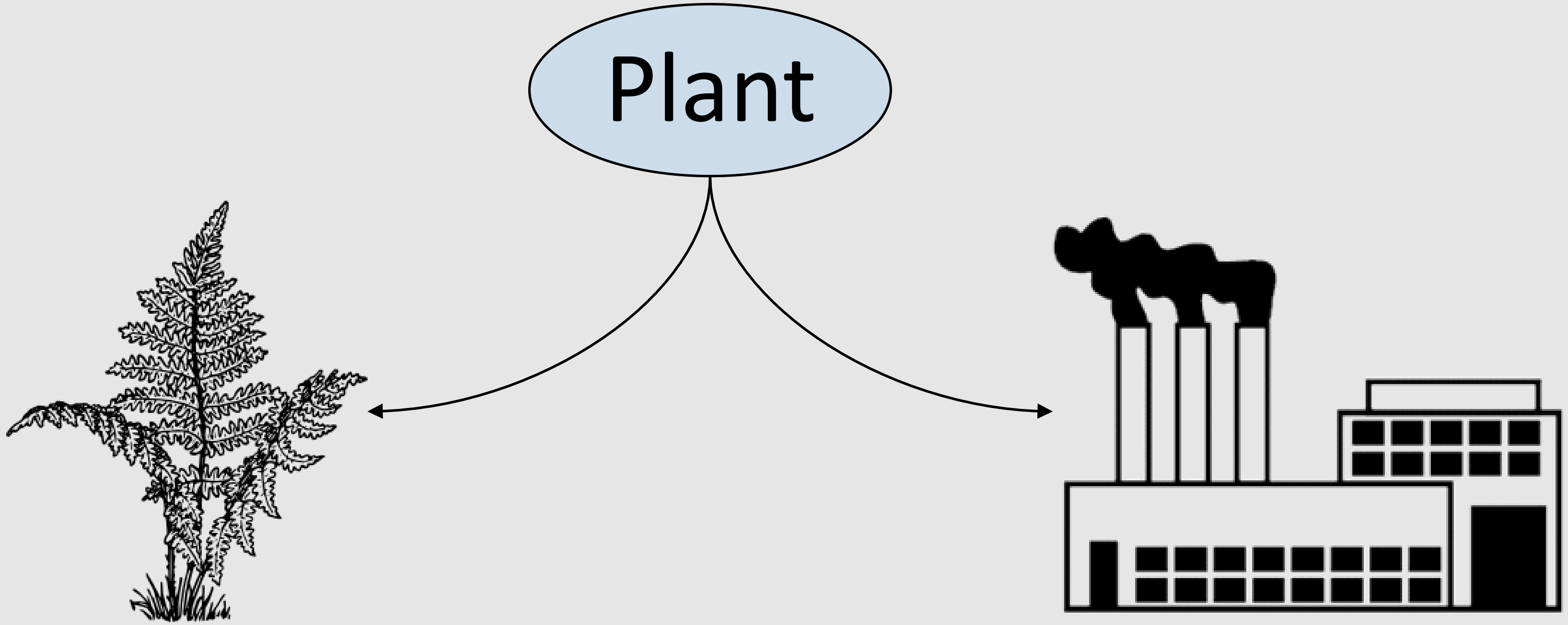
- One word – multiple meanings
- Determine meaning through context



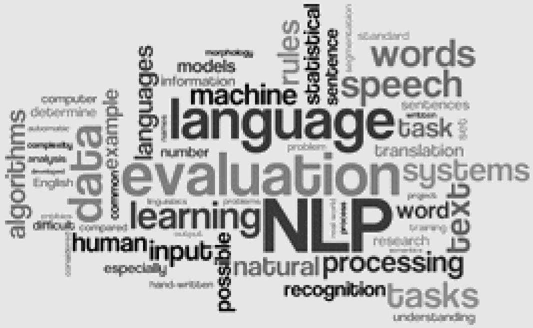


Word sense disambiguation

- One word – multiple meanings
- Determine meaning through context



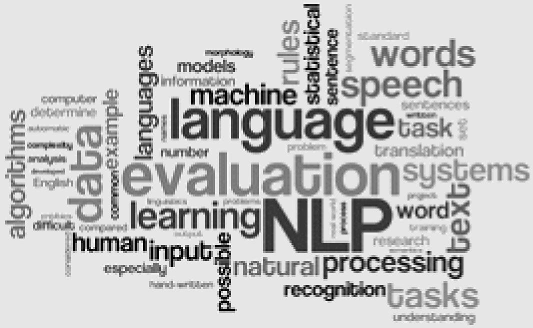
One Sense per Discourse



- Words exhibit one sense in a given context
- The sense of a word is the same throughout a document
- Can be used as a source of evidence in sense tagging

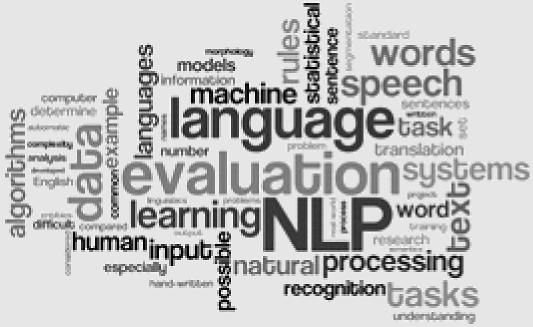
Word	Senses	Accuracy	Applicblty
plant	living/factory	99.8 %	72.8 %
tank	vehicle/contnr	99.6 %	50.5 %
poach	steal/boil	100.0 %	44.4 %
palm	tree/hand	99.8 %	38.5 %
axes	grid/tools	100.0 %	35.5 %
sake	benefit/drink	100.0 %	33.7 %
bass	fish/music	100.0 %	58.8 %
space	volume/outer	99.2 %	67.7 %
motion	legal/physical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

One Sense per Collocation



- Words close to the ambiguous word give strong evidence to the sense of the word
- Strongest indication by words that are immediately adjacent
- Same collocations may appear in different documents

One Sense per Collocation

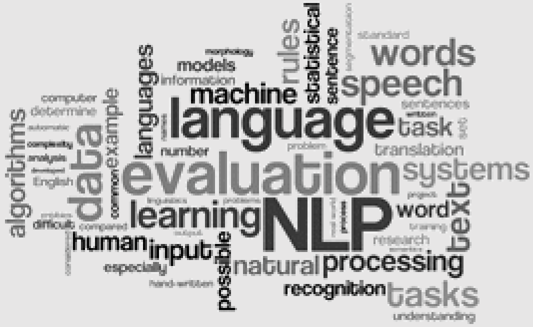


- Words close to the ambiguous word give strong evidence to the sense of the word
- Strongest indication by words that are immediately adjacent
- Same collocations may appear in different documents

…. manufacturing *plant* ….

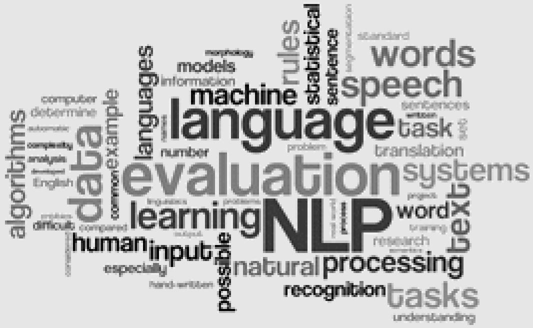
…. *plant* life ….

Unsupervised Model



- Exploits these linguistic properties
- Begin with a small set of seed examples
- Unsupervised model expands on unseen examples
- Updates seeds as new data is analysed
- No requirement for large amounts of hand-tagged training data
- Disambiguation of 7538 instances of *plant*

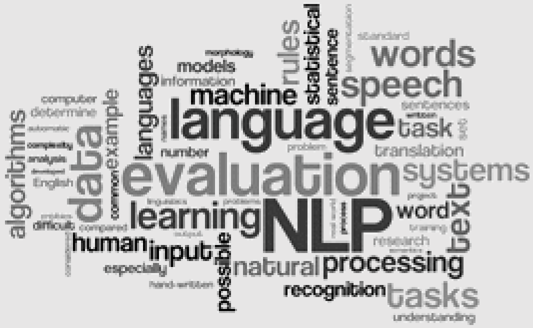
Step 1 – Preparing data



- All examples of plant are listed
- Lines included as context
- Untagged training set

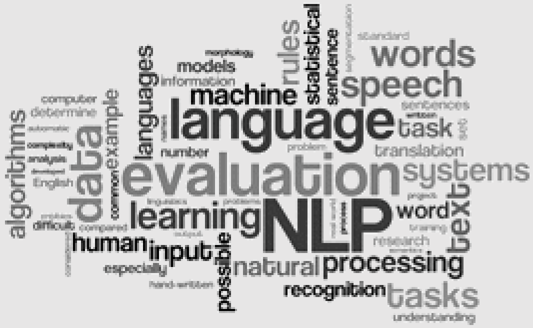
Sense	Training Examples (Keyword in Context)
?	... company said the *plant* is still operating
?	Although thousands of *plant* and animal species
?	... zonal distribution of *plant* life . ...
?	... to strain microscopic *plant* life from the ...
?	vinyl chloride monomer *plant* , which is ...
?	and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

Step 2 - Seed Collocations



- Identify small number of seed collocations
  - Words in dictionary definitions
  - Single defining collocate for each class
  - Label salient corpus collocates
- Give an indication of the sense
- Tag all training data that contains the seed collocations with the seed’s sense label

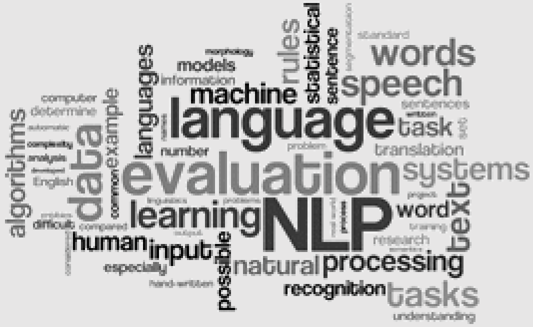
Step 2 - Seed Collocations



- Identify small number of seed collocations
  - Words in dictionary definitions
  - Single defining collocate for each class
  - Label salient corpus collocates
- Give an indication of the sense
- Tag all training data that contains the seed collocations with the seed's sense label

plant life

manufacturing *plant*



Step 2 - Seed Collocations

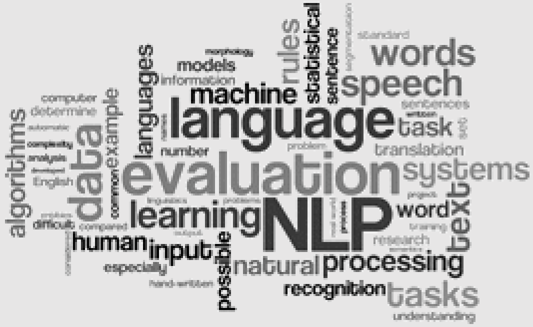
- Identify small number of seed collocations
  - Words in dictionary definitions
  - Single defining collocate for each class
  - Label salient corpus collocates
- Give an indication of the sense
- Tag all training data that contains the seed collocations with the seed's sense label

plant life

manufacturing *plant*

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	... ...

Step 2 - Seed Collocations



- Identify small number of seed collocations
  - Words in dictionary definitions
  - Single defining collocate for each class
  - Label salient corpus collocates
- Give an indication of the sense
- Tag all training data that contains the seed collocations with the seed's sense label

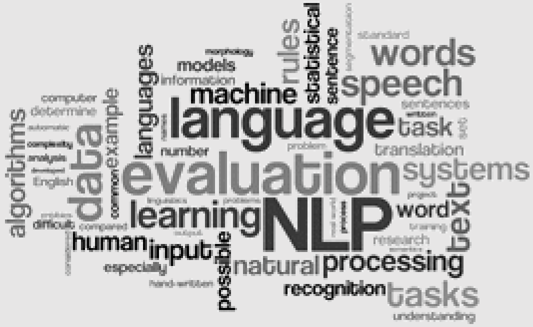
plant life

manufacturing *plant*

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	... ...

Sense	Training Examples (Keyword in Context)
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 ...

Step 2 - Seed Collocations



- Identify small number of seed collocations
  - Words in dictionary definitions
  - Single defining collocate for each class
  - Label salient corpus collocates
- Give an indication of the sense
- Tag all training data that contains the seed collocations with the seed's sense label

plant life

manufacturing *plant*

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	... ...

Sense	Training Examples (Keyword in Context)
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 ...

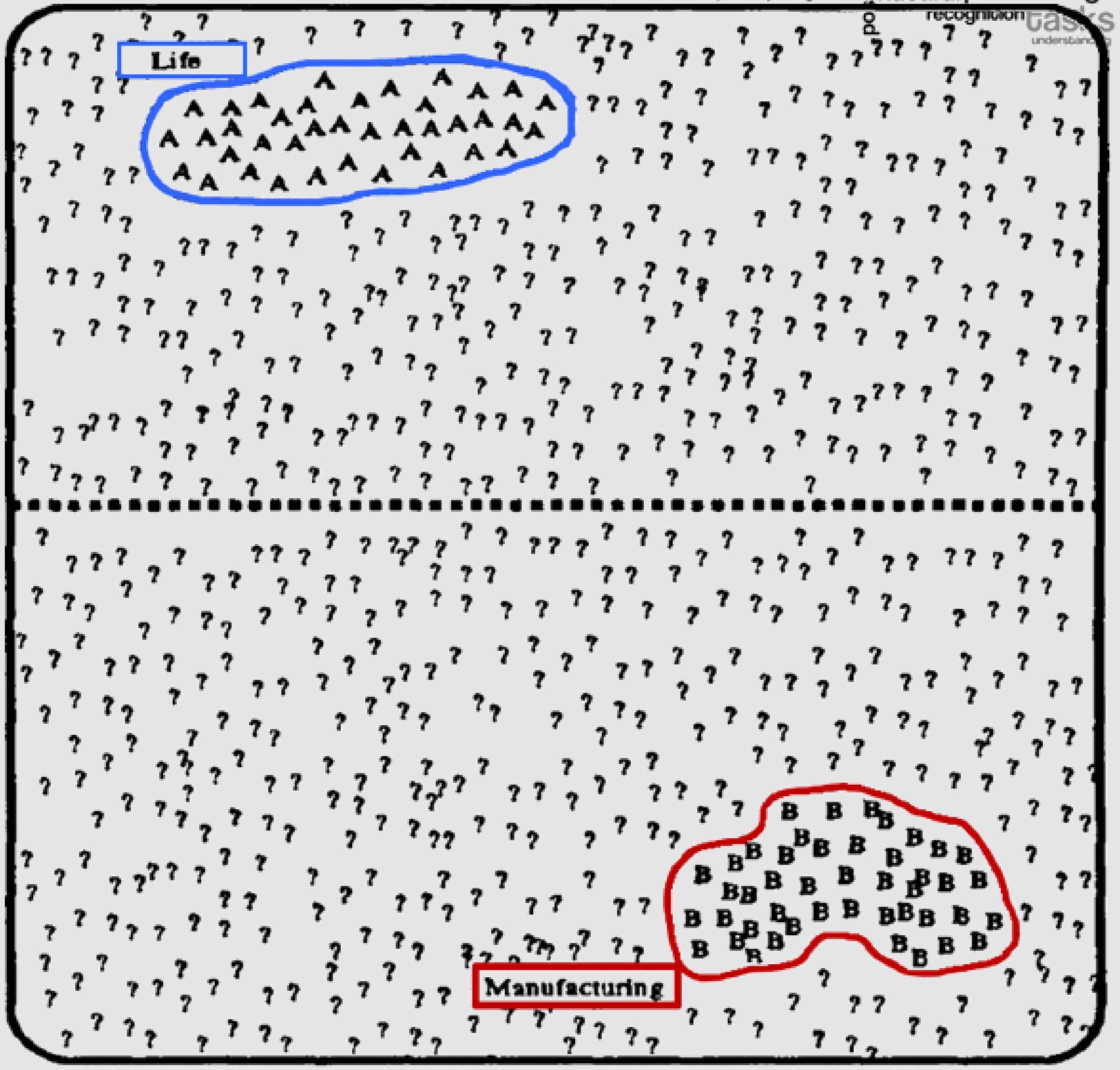
Sense	Training Examples (Keyword in Context)
?	... vinyl chloride monomer *plant* , which is ...
?	... molecules found in *plant* and animal tissue
?	... Nissan car and truck *plant* in Japan is ...
?	... and Golgi apparatus of *plant* and animal cells ...
?	... union responses to *plant* closures . ...
?	... ...
?	... ...
?	... 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 ...

Step 2 – Seed Collocations

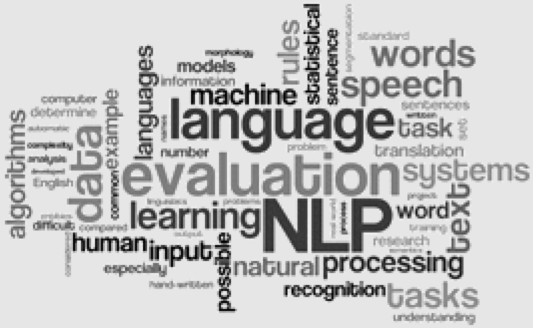
Sense	Training Examples (Keyword in Context)
A	used to strain microscopic *plant* **life** from the ...
A	... zonal distribution of *plant* **life** . ...
A	close-up studies of *plant* **life** and natural ...
A	too rapid growth of aquatic *plant* **life** in water ...
A	... the proliferation of *plant* 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	... ...

Sense	Training Examples (Keyword in Context)
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 ...

Sense	Training Examples (Keyword in Context)
?	... 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 ...



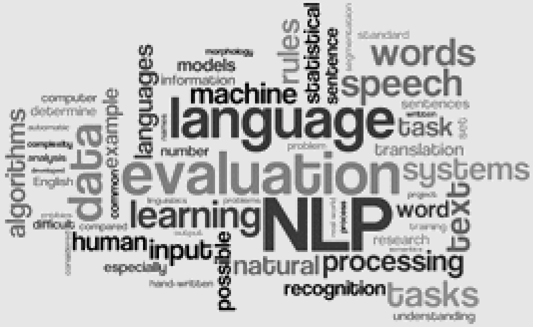
Step 3A



- Supervised algorithm trains on partly partitioned data set (seed sets)
- Identifies other collocations that partition training data.
- Computes word-sense probability distribution for each collocation
  - Collects ±k words of context around all occurrences in data set
  - K increments with each iteration
  - Compute the collocation probability distribution for each context

Collocation	A	B
plant closure	174	536
animal (within ±2-10 words)	288	23
life (within ±2-10 words)	216	1
employee (within ±2-10 words)	10	70

Step 3A



- Ranks collocations in a decision list based on log likelihood ratio

$$Log\left(\frac{Pr(Sense_A|Collocation_i)}{Pr(Sense_B|Collocation_i)}\right)$$

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	... ...

Sense	Training Examples (Keyword in Context)
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 ...

Initial decision list for *plant* (abbreviated)

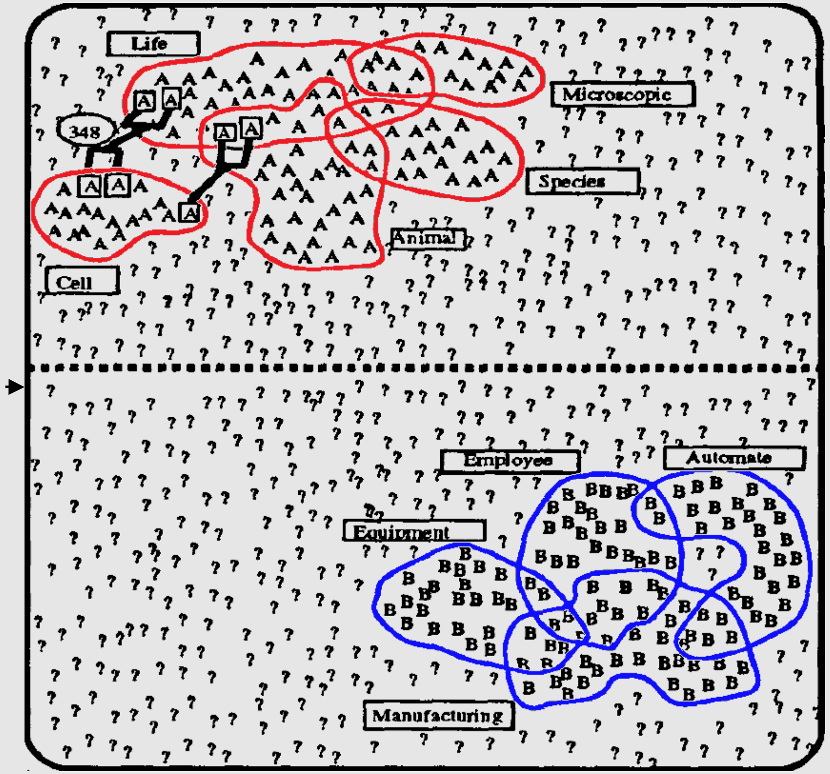
LogL	Collocation	Sense
8.10	*plant* life	⇒ A
7.58	manufacturing *plant*	⇒ B
7.39	life (within ±2-10 words)	⇒ A
7.20	manufacturing (in ±2-10 words)	⇒ B
6.27	animal (within ±2-10 words)	⇒ A
4.70	equipment (within ±2-10 words)	⇒ B
4.39	employee (within ±2-10 words)	⇒ B
4.30	assembly *plant*	⇒ B
4.10	*plant*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

Step 3B

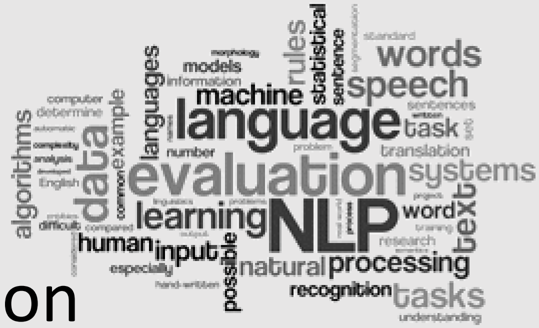
- Apply members of classifier decision list that have probability greater than a threshold to untagged data set
- Threshold follows a simulated annealing technique
- Extends seed set with these new collocations and tags new data

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
	...	

Threshold



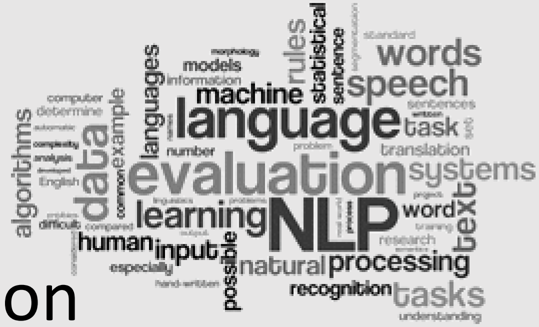
Step 3C



- One sense per discourse constraint used to augment addition
- If several instances of the ambiguous word has already been assigned a tag then the tag can extend to all examples in that discourse

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

Step 3C

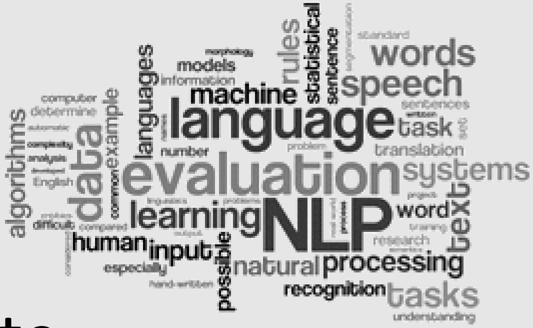


- One sense per discourse constraint used to augment addition
- If several instances of the ambiguous word has already been assigned a tag then the tag can extend to all examples in that discourse

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

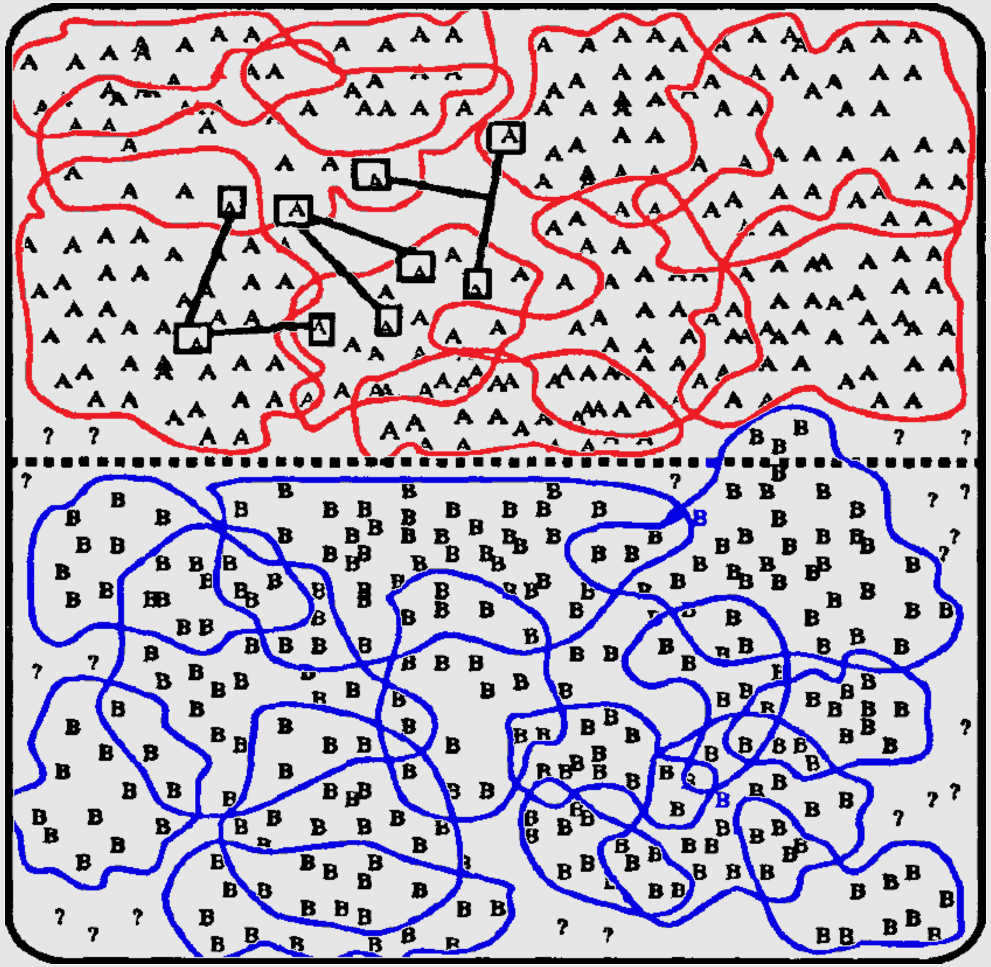
- This could form a bridge to new collocations and can perform error correction

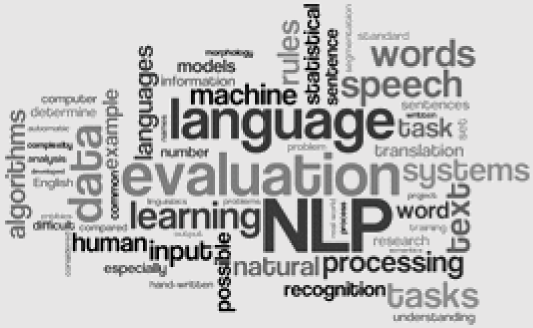
Change in tag	Disc. Numb.	Training Examples (from same discourse)
A → A	525	contains a varied *plant* and animal life
A → A	525	the most common *plant* life , the ...
A → A	525	slight within Arctic *plant* species ...
B → A	525	are protected by *plant* parts remaining from



Step 3D & 4

- Step 3 is repeated until algorithm converges on a stable state

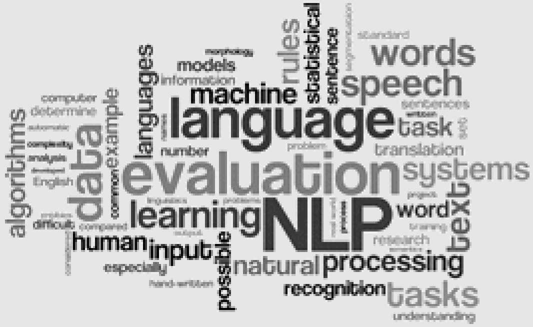




Step 5

- Classification list can then be applied to unseen data

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
...	...	



Step 5

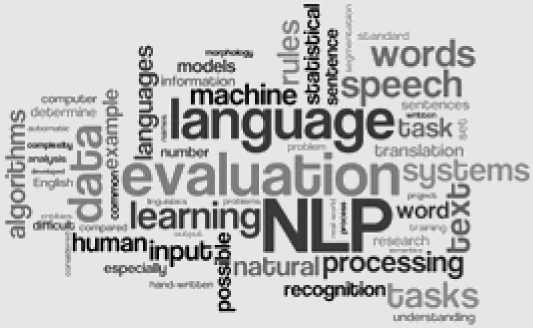
- Classification list can then be applied to unseen data

Final decision list for *plant* (abbreviated)

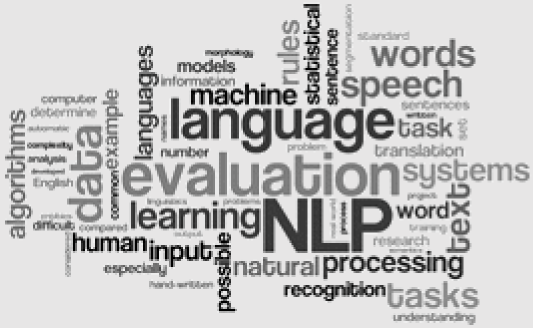
LogL	Collocation	Sense
10.12	*plant* growth	⇒ A
9.68	car (within $\pm k$ words)	⇒ B
9.64	*plant* height	⇒ A
9.61	union (within $\pm k$ words)	⇒ B
9.54	equipment (within $\pm k$ words)	⇒ B
9.51	assembly *plant*	⇒ B
9.50	nuclear *plant*	⇒ B
9.31	flower (within $\pm k$ words)	⇒ A
9.24	job (within $\pm k$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

Original seed collocations not at the top anymore

Noise & Initial Misclassifications



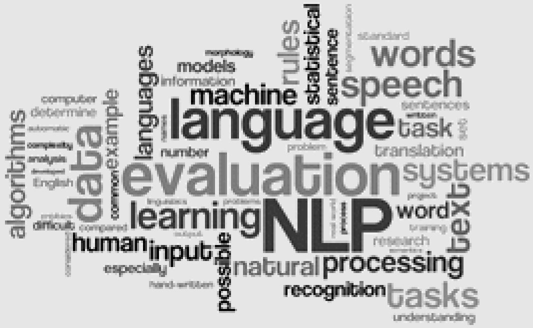
- As algorithm progresses and analyses more data, seeds may change their associated sense, making it resistant to noise
- If collocations are previously in the seed set but are then dropped from the set as their probability goes below threshold due to new data, their associated data is untagged
- This data is then re-tagged in the following iterations, overcoming potential misclassifications



Evaluation

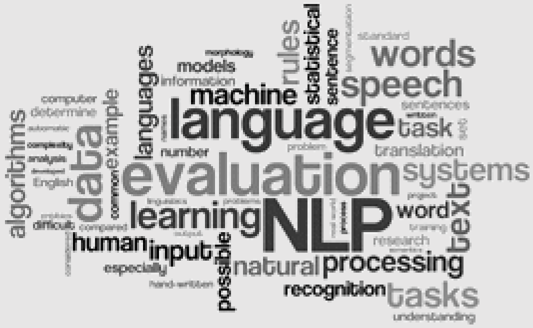
- 460 million word corpus
  - News articles
  - Scientific abstracts
  - Spoken transcripts
  - Novels
- Varied seed selection strategies
- Varied location of one word per discourse constraint
- Compared against Schütze algorithm – hierarchical clustering
- Compared against full supervised training using decision list algorithm

Results



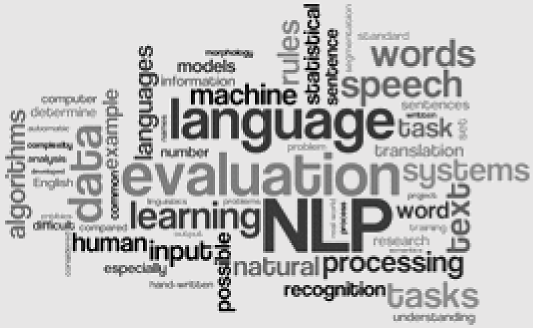
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
					Seed Training Options			(7) + OSPD		
Word	Senses	Samp. Size	% Major Sense	Supvsd Algrtm	Two Words	Dict. Defn.	Top Colls.	End only	Each Iter.	Schütze Algrthm
plant	living/factory	7538	53.1	97.7	97.1	97.3	97.6	98.3	98.6	92
space	volume/outer	5745	50.7	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

Conclusion



- Makes use of strong properties from language
  - One sense per collocation
  - One sense per discourse
- Strong discriminating information as context is used
- Builds on top of a supervised model – Bootstrapping technique
- Outperforms Schütze's algorithm and supervised model
- Achieves these rates without laborious task of tagging data!

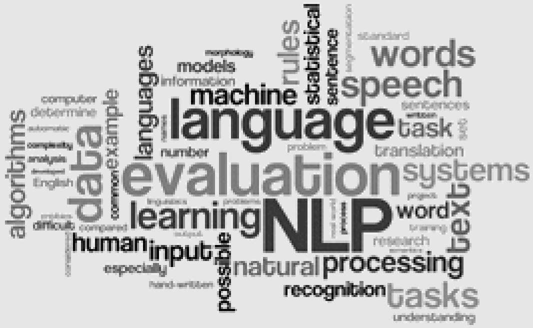
Future Work



- Limited to binary sense partition – easily extended to K partitions, but what about higher dimensionality of data?

- This approach builds up on a supervised mechanism to avoid the cold start problem, what if a fully unsupervised approach is used from the initial stage?

Reference



- Yarowsky, David. "Unsupervised word sense disambiguation rivalling supervised methods." *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1995.

Questions?

Thank you for your attention!