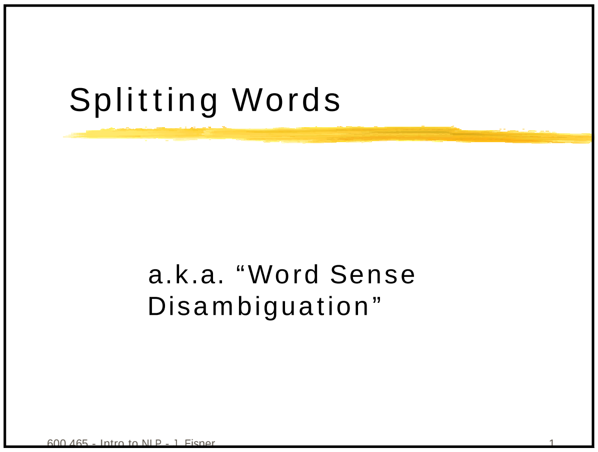
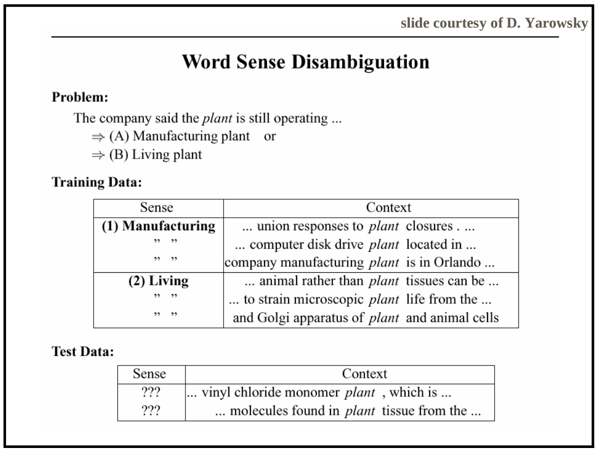


Splitting Words

a.k.a. “Word Sense Disambiguation”



slide courtesy of D. Yarowsky

Word Sense Disambiguation

Problem:

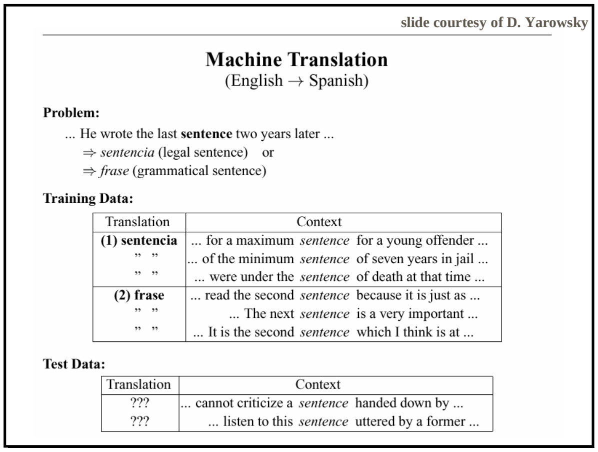
The company said the *plant* is still operating ...
⇒ (A) Manufacturing plant or
⇒ (B) Living plant

Training Data:

Sense	Context
(1) Manufacturing	... union responses to *plant* closures . ...
” ”	... computer disk drive *plant* located in ...
” ”	company manufacturing *plant* is in Orlando ...
(2) Living	... animal rather than *plant* tissues can be ...
” ”	... to strain microscopic *plant* life from the ...
” ”	and Golgi apparatus of *plant* and animal cells

Test Data:

Sense	Context
???	... vinyl chloride monomer *plant* , which is ...
???	... molecules found in *plant* tissue from the ...



slide courtesy of D. Yarowsky

Machine Translation

(English → Spanish)

Problem:

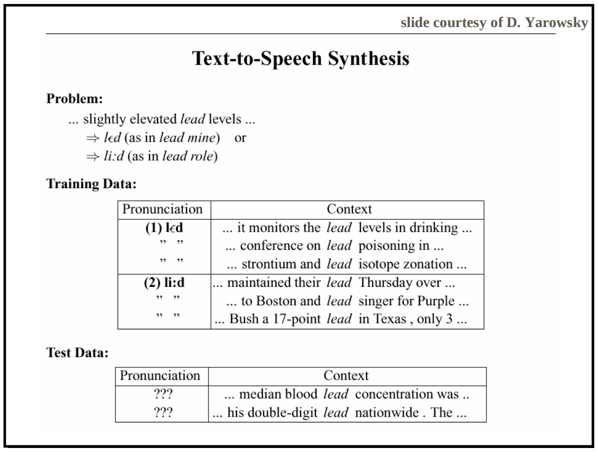
... He wrote the last **sentence** two years later ...
⇒ *sentencia* (legal sentence) or
⇒ *frase* (grammatical sentence)

Training Data:

Translation	Context
(1) sentencia	... for a maximum *sentence* for a young offender ...
” ”	... of the minimum *sentence* of seven years in jail ...
” ”	... were under the *sentence* of death at that time ...
(2) frase	... read the second *sentence* because it is just as ...
” ”	... The next *sentence* is a very important ...
” ”	... It is the second *sentence* which I think is at ...

Test Data:

Translation	Context
???	... cannot criticize a *sentence* handed down by ...
???	... listen to this *sentence* uttered by a former ...



slide courtesy of D. Yarowsky

Text-to-Speech Synthesis

Problem:

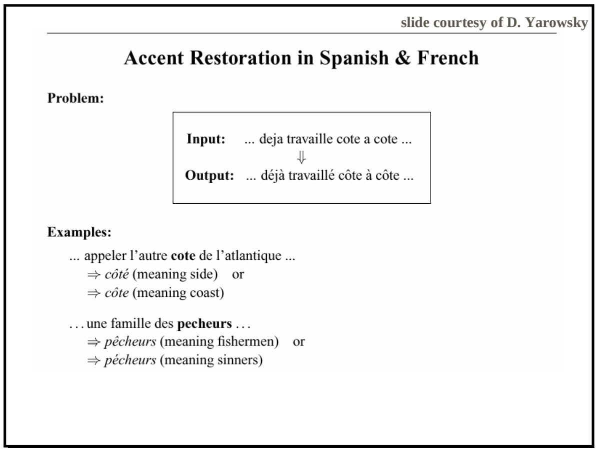
... slightly elevated *lead* levels ...
⇒ *lɛd* (as in *lead mine*) or
⇒ *li:d* (as in *lead role*)

Training Data:

Pronunciation	Context
(1) lɛd	... it monitors the *lead* levels in drinking ...
” ”	... conference on *lead* poisoning in ...
” ”	... strontium and *lead* isotope zonation ...
(2) li:d	... maintained their *lead* Thursday over ...
” ”	... to Boston and *lead* singer for Purple ...
” ”	... Bush a 17-point *lead* in Texas , only 3 ...

Test Data:

Pronunciation	Context
???	... median blood *lead* concentration was ..
???	... his double-digit *lead* nationwide . The ...



slide courtesy of D. Yarowsky

Accent Restoration in Spanish & French

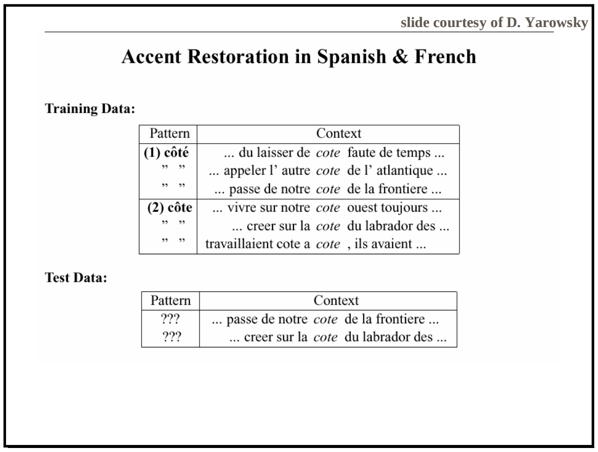
Problem:

Input: ... deja travaille cote a cote ...
⇓
Output: ... déjà travaillé côte à côte ...

Examples:

... appeler l’autre **cote** de l’atlantique ...
⇒ *côté* (meaning side) or
⇒ *côte* (meaning coast)

. . . une famille des **pecheurs** . . .
⇒ *pêcheurs* (meaning fishermen) or
⇒ *pécheurs* (meaning sinners)



slide courtesy of D. Yarowsky

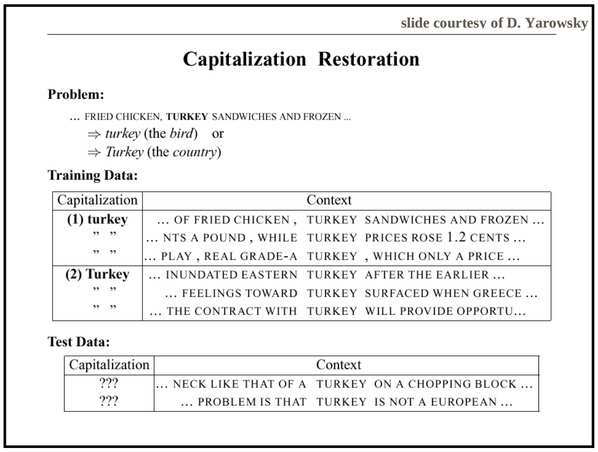
Accent Restoration in Spanish & French

Training Data:

Pattern	Context
(1) côté	... du laisser de *cote* faute de temps ...
” ”	... appeler l’ autre *cote* de l’ atlantique ...
” ”	... passe de notre *cote* de la frontiere ...
(2) côte	... vivre sur notre *cote* ouest toujours ...
” ”	... creer sur la *cote* du labrador des ...
” ”	travaillaient cote a *cote* , ils avaient ...

Test Data:

Pattern	Context
???	... passe de notre *cote* de la frontiere ...
???	... creer sur la *cote* du labrador des ...



slide courtesy of D. Yarowsky

Capitalization Restoration

Problem:

… FRIED CHICKEN, **TURKEY** SANDWICHES AND FROZEN ...

⇒ *turkey* (the *bird*) or

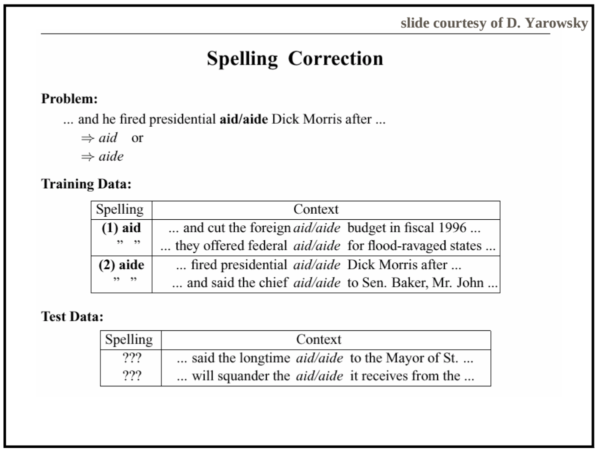
⇒ *Turkey* (the *country*)

Training Data:

Capitalization	Context
(1) turkey	… OF FRIED CHICKEN , TURKEY SANDWICHES AND FROZEN …
” ”	… NTS A POUND , WHILE TURKEY PRICES ROSE 1.2 CENTS …
” ”	… PLAY , REAL GRADE-A TURKEY , WHICH ONLY A PRICE …
(2) Turkey	… INUNDATED EASTERN TURKEY AFTER THE EARLIER …
” ”	… FEELINGS TOWARD TURKEY SURFACED WHEN GREECE …
” ”	… THE CONTRACT WITH TURKEY WILL PROVIDE OPPORTU…

Test Data:

Capitalization	Context
???	… NECK LIKE THAT OF A TURKEY ON A CHOPPING BLOCK …
???	… PROBLEM IS THAT TURKEY IS NOT A EUROPEAN …



slide courtesy of D. Yarowsky

Spelling Correction

Problem:

... and he fired presidential **aid/aide** Dick Morris after ...

⇒ *aid* or

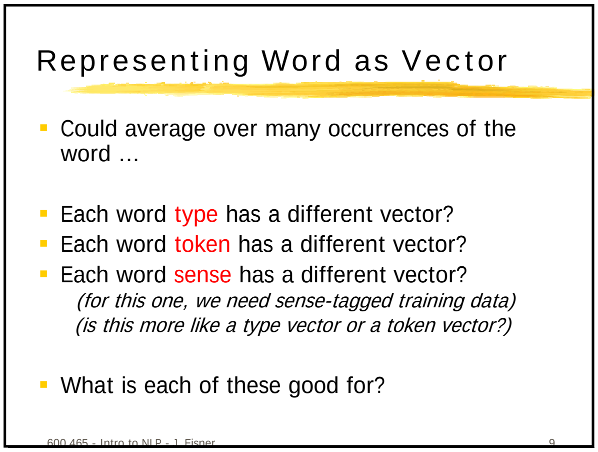
⇒ *aide*

Training Data:

Spelling	Context
(1) aid	... and cut the foreign *aid/aide* budget in fiscal 1996 ...
” ”	... they offered federal *aid/aide* for flood-ravaged states ...
(2) aide	... fired presidential *aid/aide* Dick Morris after ...
” ”	... and said the chief *aid/aide* to Sen. Baker, Mr. John ...

Test Data:

Spelling	Context
???	... said the longtime *aid/aide* to the Mayor of St. ...
???	... will squander the *aid/aide* it receives from the ...

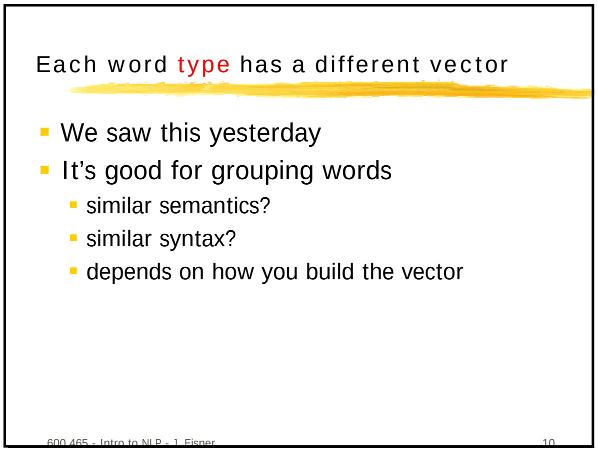


Representing Word as Vector

- Could average over many occurrences of the word …

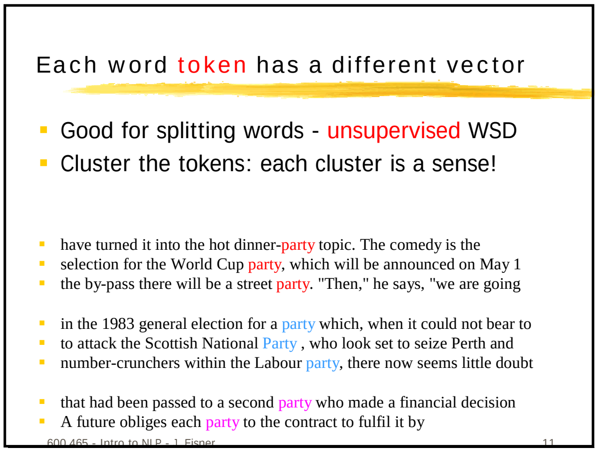
- Each word type has a different vector?
- Each word token has a different vector?
- Each word sense has a different vector?
  (for this one, we need sense-tagged training data)
  (is this more like a type vector or a token vector?)

- What is each of these good for?



Each word type has a different vector

- We saw this yesterday
- It's good for grouping words
  - similar semantics?
  - similar syntax?
  - depends on how you build the vector



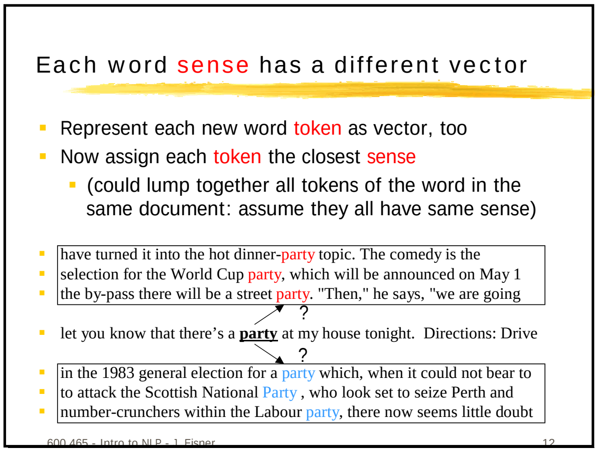
Each word token has a different vector

- Good for splitting words - unsupervised WSD
- Cluster the tokens: each cluster is a sense!

- have turned it into the hot dinner-party topic. The comedy is the
- selection for the World Cup party, which will be announced on May 1
- the by-pass there will be a street party. "Then," he says, "we are going

- in the 1983 general election for a party which, when it could not bear to
- to attack the Scottish National Party , who look set to seize Perth and
- number-crunchers within the Labour party, there now seems little doubt

- that had been passed to a second party who made a financial decision
- A future obliges each party to the contract to fulfil it by



Each word sense has a different vector

- Represent each new word token as vector, too
- Now assign each token the closest sense
  - (could lump together all tokens of the word in the same document: assume they all have same sense)

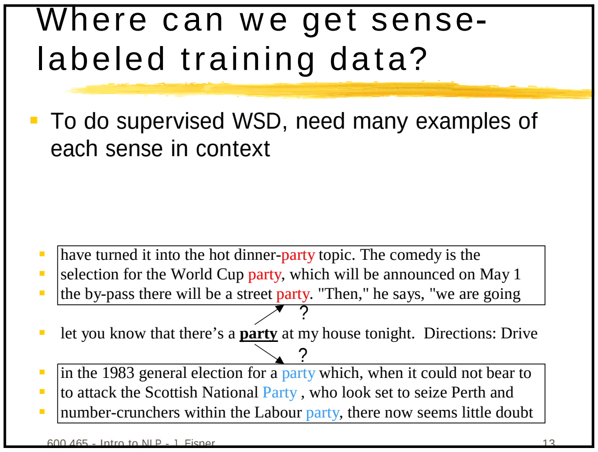
- have turned it into the hot dinner-party topic. The comedy is the
- selection for the World Cup party, which will be announced on May 1
- the by-pass there will be a street party. "Then," he says, "we are going

?

- let you know that there's a **party** at my house tonight. Directions: Drive

?

- in the 1983 general election for a party which, when it could not bear to
- to attack the Scottish National Party , who look set to seize Perth and
- number-crunchers within the Labour party, there now seems little doubt



Where can we get sense-labeled training data?

- To do supervised WSD, need many examples of each sense in context

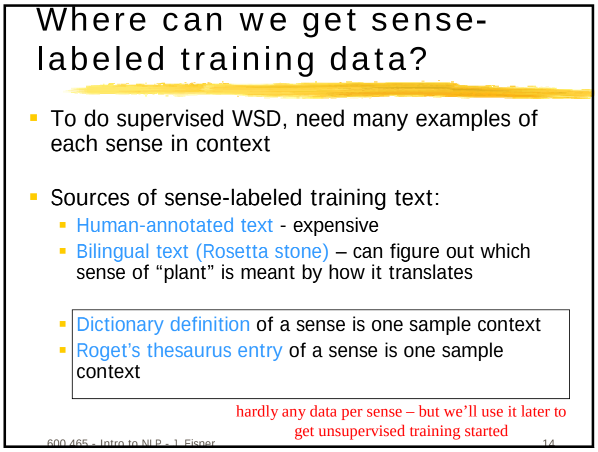
- have turned it into the hot dinner-party topic. The comedy is the
- selection for the World Cup party, which will be announced on May 1
- the by-pass there will be a street party. "Then," he says, "we are going

?

- let you know that there’s a **party** at my house tonight. Directions: Drive

?

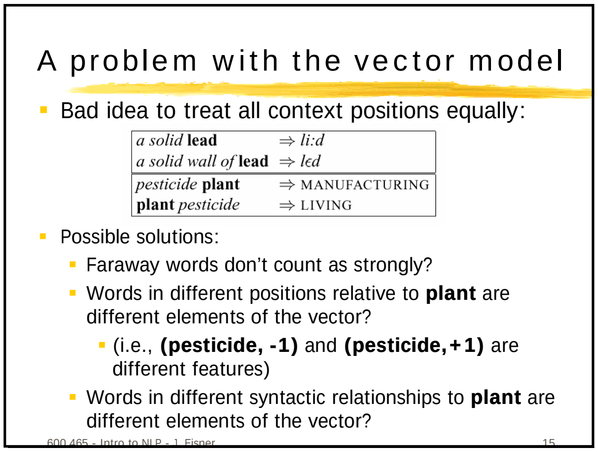
- in the 1983 general election for a party which, when it could not bear to
- to attack the Scottish National Party , who look set to seize Perth and
- number-crunchers within the Labour party, there now seems little doubt



Where can we get sense-labeled training data?

- To do supervised WSD, need many examples of each sense in context
- Sources of sense-labeled training text:
  - Human-annotated text - expensive
  - Bilingual text (Rosetta stone) – can figure out which sense of “plant” is meant by how it translates
  - Dictionary definition of a sense is one sample context
  - Roget’s thesaurus entry of a sense is one sample context

hardly any data per sense – but we’ll use it later to get unsupervised training started

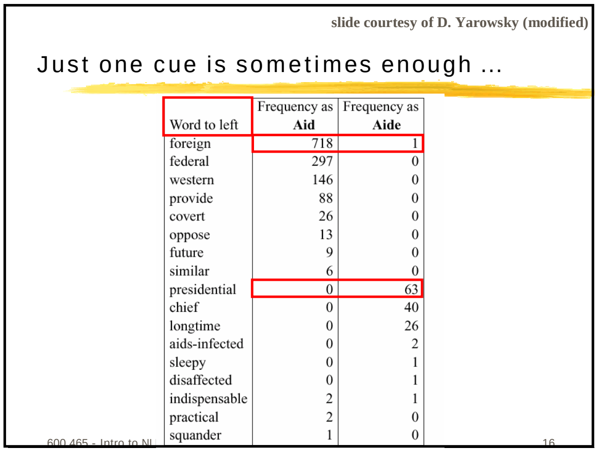


A problem with the vector model

- Bad idea to treat all context positions equally:

a solid **lead**	⇒ *li:d*
a solid wall of **lead**	⇒ *lɛd*
pesticide **plant**	⇒ MANUFACTURING
plant *pesticide*	⇒ LIVING

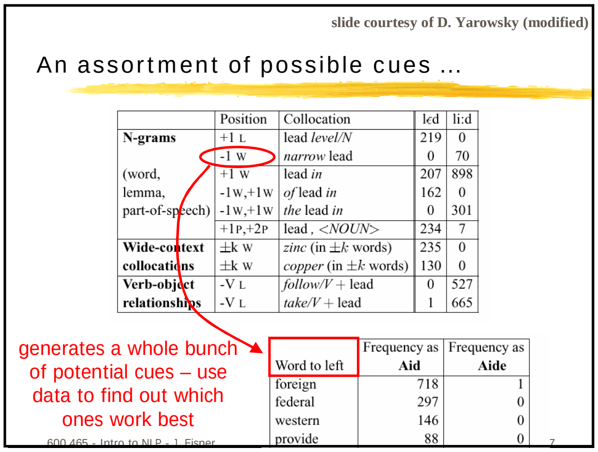
- Possible solutions:
  - Faraway words don’t count as strongly?
  - Words in different positions relative to **plant** are different elements of the vector?
    - (i.e., **(pesticide, -1)** and **(pesticide,+1)** are different features)
  - Words in different syntactic relationships to **plant** are different elements of the vector?



slide courtesy of D. Yarowsky (modified)

Just one cue is sometimes enough ...

Word to left	Frequency as **Aid**	Frequency as **Aide**
foreign	718	1
federal	297	0
western	146	0
provide	88	0
covert	26	0
oppose	13	0
future	9	0
similar	6	0
presidential	0	63
chief	0	40
longtime	0	26
aids-infected	0	2
sleepy	0	1
disaffected	0	1
indispensable	2	1
practical	2	0
squander	1	0



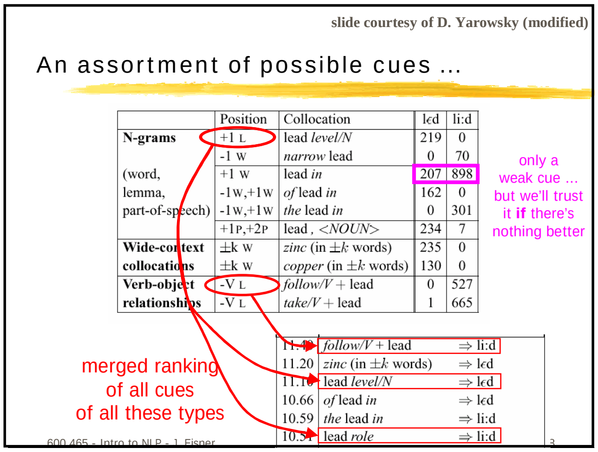
slide courtesy of D. Yarowsky (modified)

An assortment of possible cues ...

	Position	Collocation	lɛd	li:d
N-grams	+1 L	lead *level/N*	219	0
	-1 W	*narrow* lead	0	70
(word,	+1 W	lead *in*	207	898
lemma,	-1W,+1W	*of* lead *in*	162	0
part-of-speech)	-1W,+1W	*the* lead *in*	0	301
	+1P,+2P	lead , *<NOUN>*	234	7
Wide-context	±k W	*zinc* (in $\pm k$ words)	235	0
collocations	±k W	*copper* (in $\pm k$ words)	130	0
Verb-object	-V L	*follow/V* + lead	0	527
relationships	-V L	*take/V* + lead	1	665

generates a whole bunch of potential cues – use data to find out which ones work best

Word to left	Frequency as **Aid**	Frequency as **Aide**
foreign	718	1
federal	297	0
western	146	0
provide	88	0



slide courtesy of D. Yarowsky (modified)

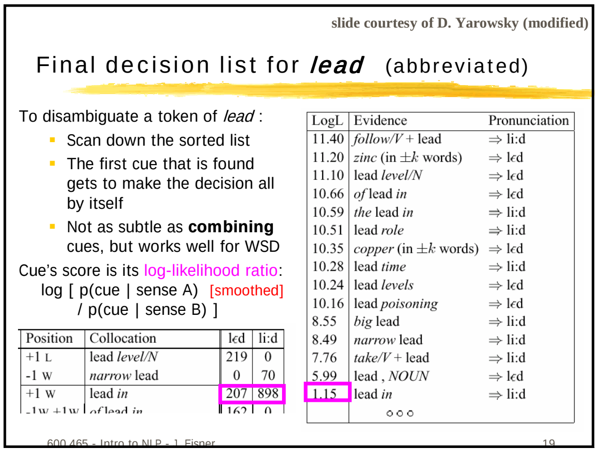
An assortment of possible cues ...

	Position	Collocation	lɛd	li:d
N-grams	+1 L	lead *level/N*	219	0
	-1 W	*narrow* lead	0	70
(word,	+1 W	lead *in*	207	898
lemma,	-1W,+1W	*of* lead *in*	162	0
part-of-speech)	-1W,+1W	*the* lead *in*	0	301
	+1P,+2P	lead , *<NOUN>*	234	7
Wide-context	±k W	*zinc* (in $\pm k$ words)	235	0
collocations	±k W	*copper* (in $\pm k$ words)	130	0
Verb-object	-V L	*follow/V* + lead	0	527
relationships	-V L	*take/V* + lead	1	665

only a weak cue ... but we’ll trust it **if** there’s nothing better

merged ranking of all cues of all these types

11.40	*follow/V* + lead	⇒ li:d
11.20	*zinc* (in $\pm k$ words)	⇒ lɛd
11.10	lead *level/N*	⇒ lɛd
10.66	*of* lead *in*	⇒ lɛd
10.59	*the* lead *in*	⇒ li:d
10.51	lead *role*	⇒ li:d



slide courtesy of D. Yarowsky (modified)

Final decision list for ***lead*** (abbreviated)

To disambiguate a token of *lead* :

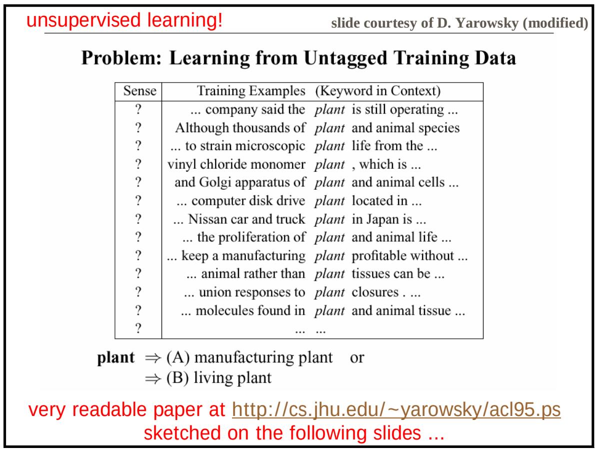
- Scan down the sorted list
- The first cue that is found gets to make the decision all by itself
- Not as subtle as **combining** cues, but works well for WSD

Cue's score is its log-likelihood ratio:
log [p(cue | sense A) [smoothed]
/ p(cue | sense B)]

Position	Collocation	lɛd	li:d
+1 L	lead *level/N*	219	0
-1 W	*narrow* lead	0	70
+1 W	lead *in*	207	898
-1W +1W	*of* lead *in*	162	0

LogL	Evidence	Pronunciation
11.40	*follow/V* + lead	⇒ li:d
11.20	*zinc* (in $\pm k$ words)	⇒ lɛd
11.10	lead *level/N*	⇒ lɛd
10.66	*of* lead *in*	⇒ lɛd
10.59	*the* lead *in*	⇒ li:d
10.51	lead *role*	⇒ li:d
10.35	*copper* (in $\pm k$ words)	⇒ lɛd
10.28	lead *time*	⇒ li:d
10.24	lead *levels*	⇒ lɛd
10.16	lead *poisoning*	⇒ lɛd
8.55	*big* lead	⇒ li:d
8.49	*narrow* lead	⇒ li:d
7.76	*take/V* + lead	⇒ li:d
5.99	lead , *NOUN*	⇒ lɛd
1.15	lead *in*	⇒ li:d
	○○○	

600.465 - Intro to NLP - J. Eisner 19



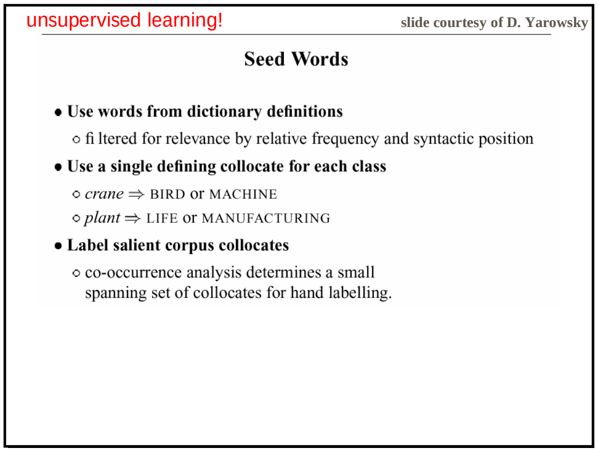
unsupervised learning! slide courtesy of D. Yarowsky (modified)

Problem: Learning from Untagged Training Data

Sense	Training Examples	(Keyword in Context)	
?	... company said the	*plant*	is still operating ...
?	Although thousands of	*plant*	and animal species
?	... 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 ...
?	... Nissan car and truck	*plant*	in Japan is ...
?	... the proliferation of	*plant*	and animal life ...
?	... keep a manufacturing	*plant*	profitable without ...
?	... animal rather than	*plant*	tissues can be ...
?	... union responses to	*plant*	closures . ...
?	... molecules found in	*plant*	and animal tissue ...
?	...	...	

plant ⇒ (A) manufacturing plant or
⇒ (B) living plant

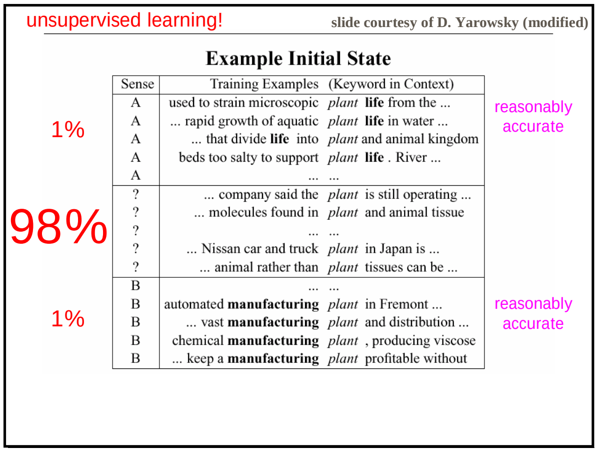
very readable paper at http://cs.jhu.edu/~yarowsky/acl95.ps
sketched on the following slides ...



unsupervised learning! slide courtesy of D. Yarowsky

Seed Words

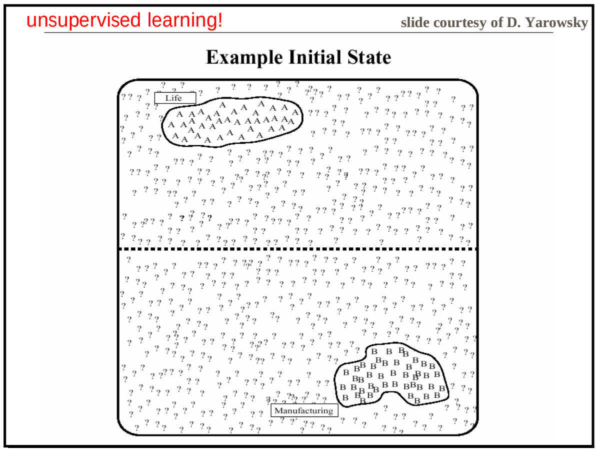
- **Use words from dictionary definitions**
  - ○ filtered for relevance by relative frequency and syntactic position
- **Use a single defining collocate for each class**
  - ○ *crane* ⇒ BIRD or MACHINE
  - ○ *plant* ⇒ LIFE or MANUFACTURING
- **Label salient corpus collocates**
  - ○ co-occurrence analysis determines a small spanning set of collocates for hand labelling.



unsupervised learning! slide courtesy of D. Yarowsky (modified)

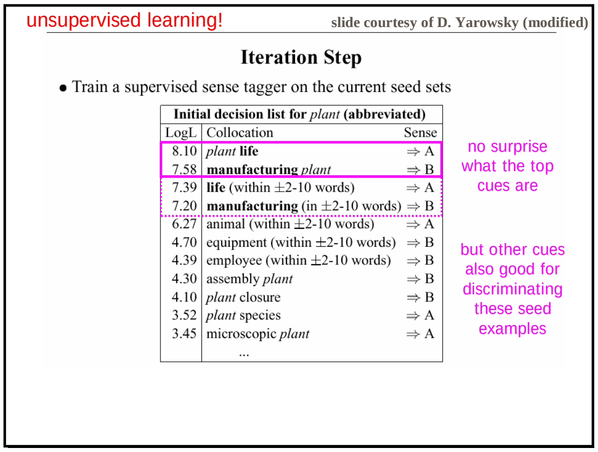
Example Initial State

	Sense	Training Examples	(Keyword in Context)		
1%	A	used to strain microscopic	*plant*	**life** from the ...	reasonably accurate
	A	... rapid growth of aquatic	*plant*	**life** in water ...	
	A	... that divide **life** into	*plant*	and animal kingdom	
	A	beds too salty to support	*plant*	**life** . River ...	
	A	...	...		
98%	?	... company said the	*plant*	is still operating ...	
	?	... molecules found in	*plant*	and animal tissue	
	?	...	...		
	?	... Nissan car and truck	*plant*	in Japan is ...	
	?	... animal rather than	*plant*	tissues can be ...	
1%	B	...	...		reasonably accurate
	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	



unsupervised learning! slide courtesy of D. Yarowsky

Example Initial State



unsupervised learning! slide courtesy of D. Yarowsky (modified)

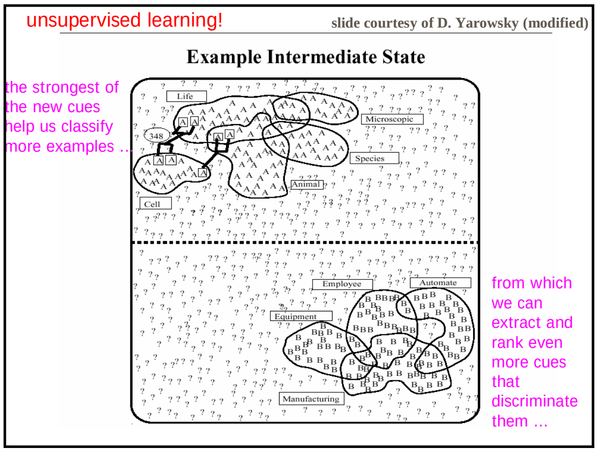
Iteration Step

- Train a supervised sense tagger on the current seed sets

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
3.45	microscopic *plant*	⇒ A
	...	

no surprise what the top cues are

but other cues also good for discriminating these seed examples



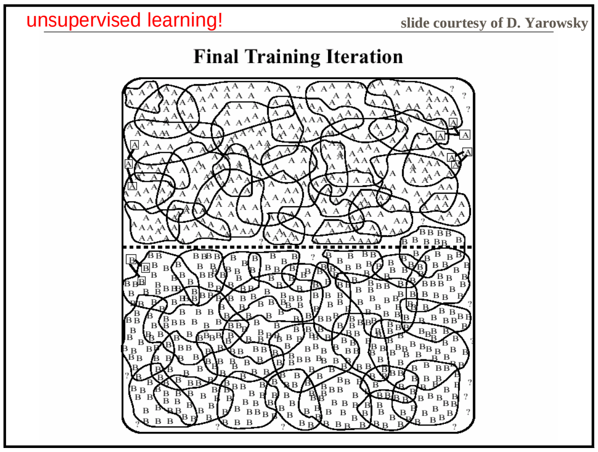
unsupervised learning!

slide courtesy of D. Yarowsky (modified)

Example Intermediate State

the strongest of the new cues help us classify more examples ...

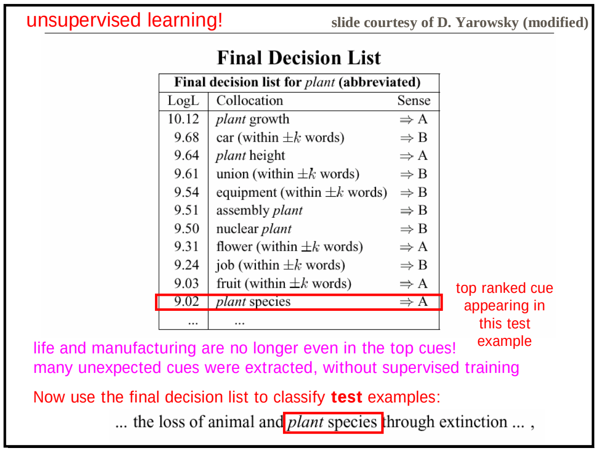
from which we can extract and rank even more cues that discriminate them ...



unsupervised learning!

slide courtesy of D. Yarowsky

Final Training Iteration



unsupervised learning!

slide courtesy of D. Yarowsky (modified)

Final Decision List

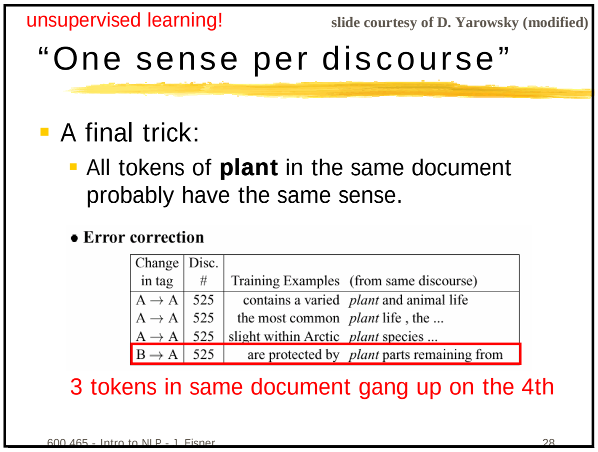
Final decision list for *plant* (abbreviated)		
LogL	Collocation	Sense
10.12	*plant* growth	⇒ A
9.68	car (within $\pm k$ words)	⇒ B
9.64	*plant* height	⇒ A
9.61	union (within $\pm k$ words)	⇒ B
9.54	equipment (within $\pm k$ words)	⇒ B
9.51	assembly *plant*	⇒ B
9.50	nuclear *plant*	⇒ B
9.31	flower (within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

top ranked cue appearing in this test example

life and manufacturing are no longer even in the top cues!
many unexpected cues were extracted, without supervised training

Now use the final decision list to classify **test** examples:

... the loss of animal and *plant* species through extinction ... ,



unsupervised learning!

slide courtesy of D. Yarowsky (modified)

"One sense per discourse"

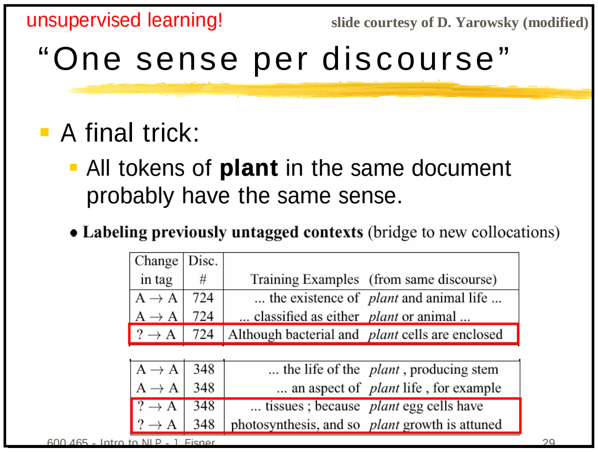
- A final trick:
  - All tokens of **plant** in the same document probably have the same sense.

- **Error correction**

Change in tag	Disc. #	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

3 tokens in same document gang up on the 4th

600.465 - Intro to NLP - J. Eisner 28



unsupervised learning!

slide courtesy of D. Yarowsky (modified)

"One sense per discourse"

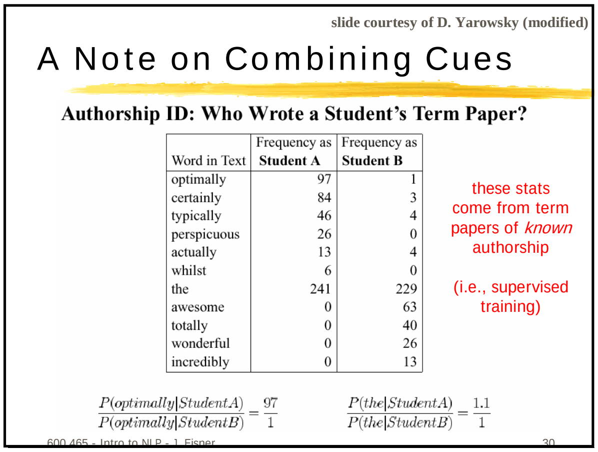
- A final trick:
  - All tokens of **plant** in the same document probably have the same sense.

- **Labeling previously untagged contexts** (bridge to new collocations)

Change in tag	Disc. #	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

600.465 - Intro to NLP - J. Eisner 29



slide courtesy of D. Yarowsky (modified)

A Note on Combining Cues

Authorship ID: Who Wrote a Student's Term Paper?

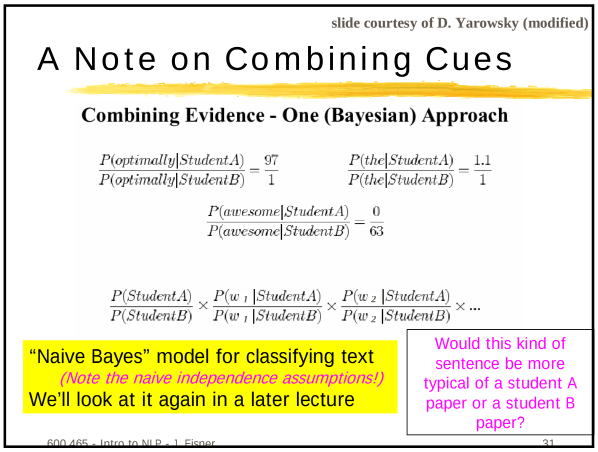
Word in Text	Frequency as **Student A**	Frequency as **Student B**
optimally	97	1
certainly	84	3
typically	46	4
perspicuous	26	0
actually	13	4
whilst	6	0
the	241	229
awesome	0	63
totally	0	40
wonderful	0	26
incredibly	0	13

these stats come from term papers of *known* authorship

(i.e., supervised training)

$$\frac{P(optimally|StudentA)}{P(optimally|StudentB)} = \frac{97}{1} \qquad \frac{P(the|StudentA)}{P(the|StudentB)} = \frac{1.1}{1}$$

600.465 - Intro to NLP - J. Eisner 30



slide courtesy of D. Yarowsky (modified)

A Note on Combining Cues

Combining Evidence - One (Bayesian) Approach

$$\frac{P(optimally|StudentA)}{P(optimally|StudentB)} = \frac{97}{1} \qquad \frac{P(the|StudentA)}{P(the|StudentB)} = \frac{1.1}{1}$$

$$\frac{P(awesome|StudentA)}{P(awesome|StudentB)} = \frac{0}{63}$$

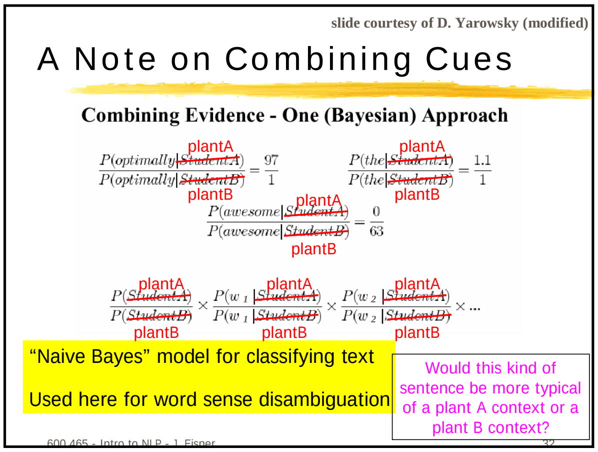
$$\frac{P(StudentA)}{P(StudentB)} \times \frac{P(w_1|StudentA)}{P(w_1|StudentB)} \times \frac{P(w_2|StudentA)}{P(w_2|StudentB)} \times \dots$$

“Naive Bayes” model for classifying text

(Note the naive independence assumptions!)

We’ll look at it again in a later lecture

Would this kind of sentence be more typical of a student A paper or a student B paper?



slide courtesy of D. Yarowsky (modified)

A Note on Combining Cues

Combining Evidence - One (Bayesian) Approach

$$\frac{P(optimally|\overset{\text{plantA}}{\cancel{StudentA}})}{P(optimally|\underset{\text{plantB}}{\cancel{StudentB}})} = \frac{97}{1} \qquad \frac{P(the|\overset{\text{plantA}}{\cancel{StudentA}})}{P(the|\underset{\text{plantB}}{\cancel{StudentB}})} = \frac{1.1}{1}$$

$$\frac{P(awesome|\overset{\text{plantA}}{\cancel{StudentA}})}{P(awesome|\underset{\text{plantB}}{\cancel{StudentB}})} = \frac{0}{63}$$

$$\frac{P(\overset{\text{plantA}}{\cancel{StudentA}})}{P(\underset{\text{plantB}}{\cancel{StudentB}})} \times \frac{P(w_1|\overset{\text{plantA}}{\cancel{StudentA}})}{P(w_1|\underset{\text{plantB}}{\cancel{StudentB}})} \times \frac{P(w_2|\overset{\text{plantA}}{\cancel{StudentA}})}{P(w_2|\underset{\text{plantB}}{\cancel{StudentB}})} \times \dots$$

“Naive Bayes” model for classifying text

Used here for word sense disambiguation

Would this kind of sentence be more typical of a plant A context or a plant B context?