Machine Learning

Programs "learn" behaviors from labelled examples - *Supervised* learning

Training Data:

Class	Temperature	Sweating?	Chills?	Appetite-Loss?	Post-Nasal-Drip?	Rash?
COLD	98.7	no	yes	no	yes	yes
FLU	100.1	yes	no	yes	no	no
COLD	98.9	no	yes	no	yes	no
FLU	99.4	yes	no	yes	no	yes
FLU	99.1	yes	no	yes	no	yes
COLD	98.4	no	yes	no	yes	no

Test Data:

Class	Temperature	Sweating?	Chills?	Appetite-Loss?	Post-Nasal-Drip?	Rash?
???	98.3	no	yes	no	yes	yes
???	101.2	yes	no	yes	no	no

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 ...

Machine Translation

(English → French)

Problem:

... He wrote the last *sentence* two years later ...

⇒ *peine* (legal sentence)   or

⇒ *phrase* (grammatical sentence)

Training Data:

Translation	Context
(1) peine	... 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) phrase	... 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 ...

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 ...

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)

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 ...

Capitalization Restoration

Problem:

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

$\Rightarrow$ *turkey* (the *bird*) or

$\Rightarrow$ *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 …

Spelling Correction

Problem:

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

$\Rightarrow$ *aid* or

$\Rightarrow$ *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 ...

Other Applications

- **Vowel Restoration in Hebrew and Arabic**
- **Capitalization Restoration** (e.g. TURKEY ⇒ Turkey/turkey)
- **Spelling Correction** (e.g. principal/principle)
- **Proper Noun Classification** (e.g. Washington ⇒ PERSON/PLACE)
- **Speech Recognition** (e.g. /eid/ ⇒ aid/aide)

Machine Learning Algorithms

- Neural Nets
- Decision Trees
- Decision Lists
- Bayesian Classifiers
- Genetic Algorithms

Machine Learning

Programs "learn" behaviors from labelled examples - *Supervised* learning

Training Data:

Class	Temperature	Sweating?	Chills?	Appetite-Loss?	Post-Nasal-Drip?	Rash?
COLD	98.7	no	yes	no	yes	yes
FLU	100.1	yes	no	yes	no	no
COLD	98.9	no	yes	no	yes	no
FLU	99.4	yes	no	yes	no	yes
FLU	99.1	yes	no	yes	no	yes
COLD	98.4	no	yes	no	yes	no

Test Data:

Class	Temperature	Sweating?	Chills?	Appetite-Loss?	Post-Nasal-Drip?	Rash?
???	98.3	no	yes	no	yes	yes
???	101.2	yes	no	yes	no	no

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

$$\frac{P(optimally|Student A)}{P(optimally|Student B)} = \frac{97}{1}$$

$$\frac{P(the|Student A)}{P(the|Student B)} = \frac{1.1}{1}$$

Combining Evidence - One (Bayesian) Approach

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

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

$$\frac{P(StudentA)}{P(StudentB)} = \frac{P(w_{-3}|StudentA)}{P(w_{-3}|StudentB)} \times \frac{P(w_{-2}|StudentA)}{P(w_{-2}|StudentB)} \times \ldots$$

Sources of Evidence - Words in Context

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

Complex Features - Linguistic Patterns

	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

Algorithm 1: Decision Lists

LogL	Evidence	Pronunciation
11.40	*follow/V* + lead	$\Rightarrow$ li:d
11.20	*zinc* (in $\pm k$ words)	$\Rightarrow$ lϵd
11.10	lead *level/N*	$\Rightarrow$ lϵd
10.66	*of* lead *in*	$\Rightarrow$ lϵd
10.59	*the* lead *in*	$\Rightarrow$ li:d
10.51	lead *role*	$\Rightarrow$ li:d
10.35	*copper* (in $\pm k$ words)	$\Rightarrow$ lϵd
10.28	lead *time*	$\Rightarrow$ li:d
10.24	lead *levels*	$\Rightarrow$ lϵd
10.16	lead *poisoning*	$\Rightarrow$ lϵd
	○ ○ ○	

New Sentence:

Studies identified slightly elevated copper and **lead** *levels*.

Classification:

$\Rightarrow$ lϵd

Combining vs. Not Combining Probabilities

- Use all matching patterns in target context

$$Score = \sum_i \left(Log\left(\frac{Pr(Accent\_Pattern_1|Collocation_i)}{Pr(Accent\_Pattern_2|Collocation_i)}\right)\right)$$

- Use only the highest scoring pattern

Agree -	Both classifications correct	92%
	Both classifications incorrect	6%
Disagree -	Single best evidence correct	1.3%
	Combined evidence correct	0.7%
Total -		100%

Smoothing and Interpolation

- **Smoothing of likelihood ratios** sensitive to variables including
  - Collocational distance
  - Type of word (noun, verb, content word, function word)
  - Nature of syntactic relationship

- Improve probability estimates by **interpolating** between *global* and *residual* probabilities

Evaluation

Word	Pron1	Pron2	Sample Size	Prior Prob.	% Correct
lives	laɪvz	lɪvz	33186	69	98
wound	waænd	wund	4483	55	98
Nice	naɪs	nis	573	56	94
Begin	bɪægɪn	beɪgɪn	1143	75	97
Chi	tæi	kaɪ	1288	53	98
Colon	koææloæn	ækoælæn	1984	69	98
lead (N)	lid	læd	12165	66	98
tear (N)	tææ*	tɪæ*	2271	88	97
axes (N)	ææksiz	ææksɪz	1344	72	96
IV	aɪ vi	fæææ	1442	76	98
Jan	dææn	jæn	1327	90	98
routed	æutɪd	æaætɪd	589	60	94
bass	beɪs	bæs	1865	57	99
AVERAGE			63660	67	**97**

Comparative Evaluation

- Accent restoration task in Spanish

N-gram Tagger	93.8%
Bayesian Classifier	89.4%
Decision List	**96.8%**

Advantages of Algorithm

- Successfully integrates non-independent features
- Combines strengths of Bayesian classifiers and N-gram taggers
- Models local sequence and wide context
- Returns probability values with all classifications
- Efficient
- Resulting decision lists are easy to interpret and modify

Problem: Lexical Ambiguity Resolution

- Word sense disambiguation
- Lexical choice in machine translation
- Homograph disambiguation in speech synthesis
- Accent restoration in Spanish and French
- Other applications

Three Algorithms:

- Decision lists (supervised)

⇒ Bayesian word-class discriminators (unsupervised)

- Modulated bootstrapping from seed words (unsupervised)

Need for Unsupervised Algorithms

- Hand-tagged training data are expensive and generally unavailable
- WordNet sense-tagged corpus: small, under development
- Parallel aligned bilingual corpora
  - Source of automatically tagged data for translation distinctions
  - Currently limited availability and coverage

Goal: Methods for training on untagged, monolingual text

Bayesian Word-class Discrimination

Roget's Thesaurus Categories (1042 word classes):

MACHINE - tractor, bulldozer, crane, jackhammer, drill, forklift ...

ANIMAL - alligator, lizard, bat, flamingo, heron, crane, stork ...

MINERAL - strontium, zinc, magnesium, lead, copper, cobalt ...

Statistical word-class detectors:

... *the engine of the* XXX *was damaged* ...

$$p(\text{MACHINE}|\mathit{context}) = .650$$
$$p(\text{ANIMAL}|\mathit{context}) = .007$$
$$p(\text{MINERAL}|\mathit{context}) = .005$$

...

Class Discriminators ⇒ Word-sense Discriminators

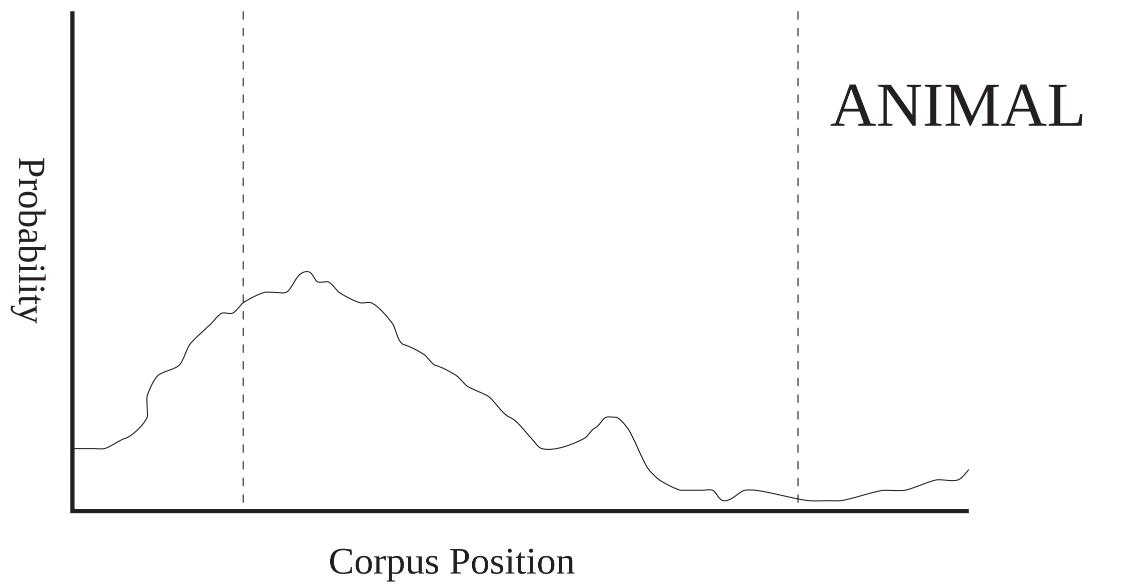
crane ⇒ ANIMAL or
⇒ MACHINE

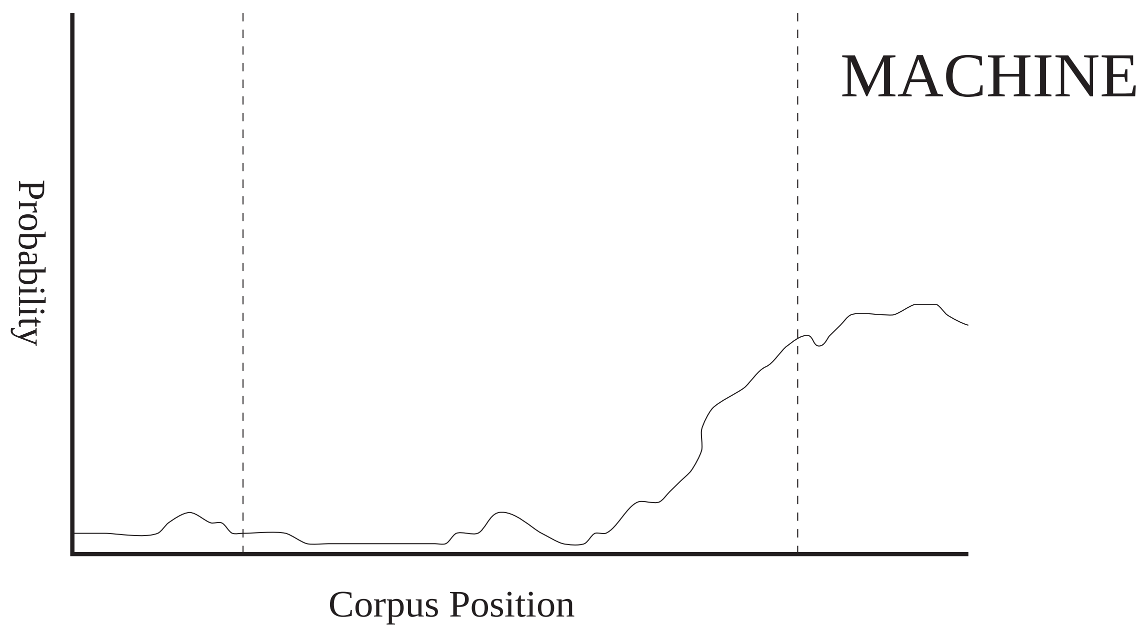
... *the engine of the* **crane** *was damaged* ...

$$p(\text{MACHINE}|\textit{context}) = \mathbf{.650}$$
$$p(\text{ANIMAL}|\textit{context}) = .007$$
$$p(\text{MINERAL}|\textit{context}) = .005$$
...

... *flocks of* **cranes** *nested in the swamp* ...

$$p(\text{MACHINE}|\textit{context}) = .002$$
$$p(\text{ANIMAL}|\textit{context}) = \mathbf{.370}$$
$$p(\text{MINERAL}|\textit{context}) = .004$$
...


ANIMAL
Probability
Corpus Position


MACHINE
Probability
Corpus Position

Training of Class Models

Word Class	Context
MACHINE	... power for the **crane** , hoist and derrick assembly ...
MACHINE	... been manufacturing **forklift** parts for 30 years ...
MACHINE	... found valves for **generator** , refinery turbines ...
MACHINE	... the fumes of the **tractor** began to bother my eyes ...
MACHINE	... the carbon-tipped **drill** forced manufacturers ...
MACHINE	... the noise of a **bulldozer** disturbed the peace of ...
MACHINE	... began a fire **drill** just after the lunch break ...
MACHINE	... while the crowned **crane** often nests in marshy ...
MACHINE	... bought a fleet of **tractor** plows for maintenance ...

Hand-labelled training data are unnecessary

- The majority of words (by type) have only one sense
- Secondary senses are widely distributed across categories
- The noise introduced by the secondary senses is tolerable

$\Rightarrow$ focused signal / diffuse noise

Training of Parameters

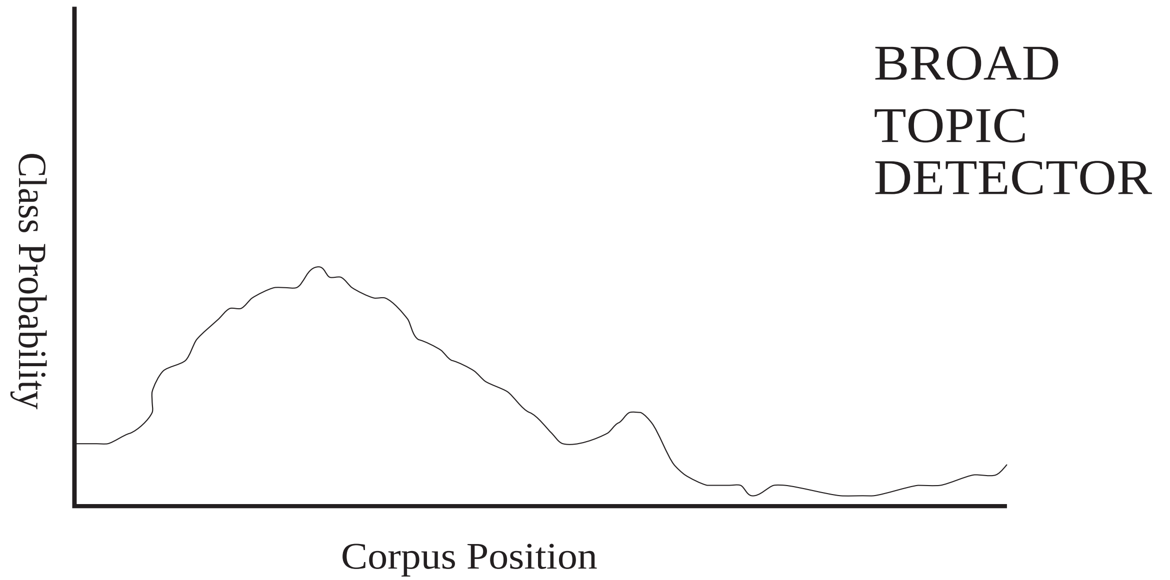
- Weight each class member equally (dog vs. wildebeest)

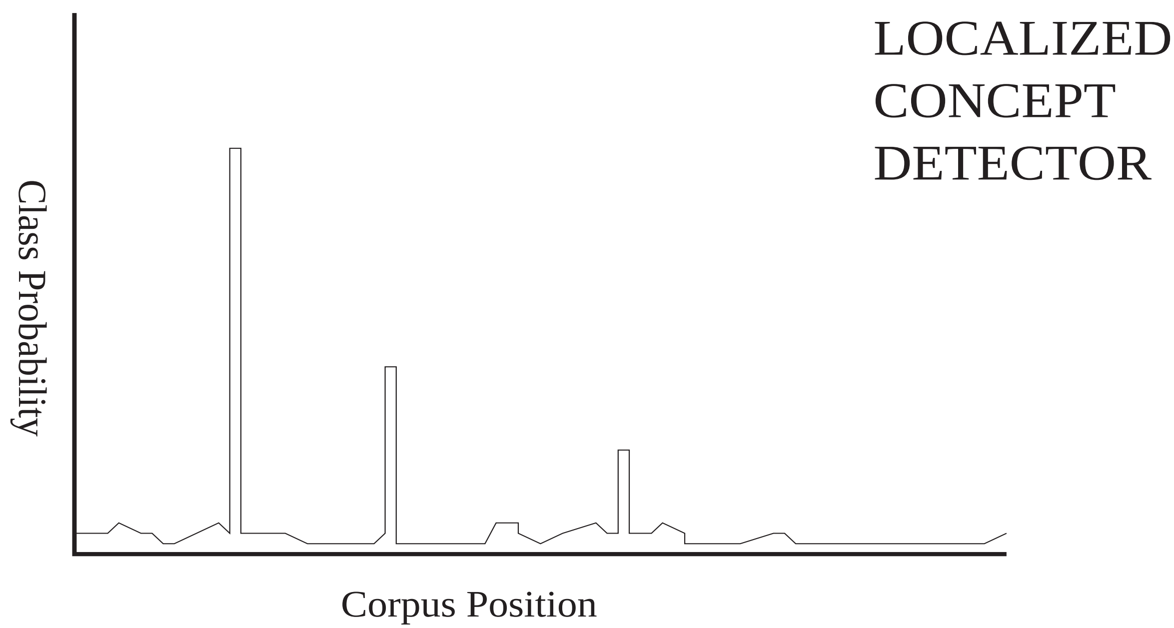
  ⇒ model *typical* members of the class, not most *frequent*

- Bag-of-words Bayesian models (topic detectors)

$$p(Rcat_j|context) = p(Rcat_j) \prod_{i=-50}^{50} \frac{p(word_i|Rcat_j)}{p(word_i)}$$

- Add richer set of collocation models (from decision list work)


BROAD
TOPIC
DETECTOR
Class Probability
Corpus Position


LOCALIZED
CONCEPT
DETECTOR
Class Probability
Corpus Position

Application: Language modeling for speech recognition

....	he	consumed	an	enormous	/steɪk/	with	wine	....

$$/\text{steɪk}/ \rightarrow p(steak|\text{FOOD}) \times p(\text{FOOD}|context)$$
$$\rightarrow p(stake|\text{FOOD}) \times p(\text{FOOD}|context)$$

N-gram Language Models:

1)	$p(steak\|\ an\ \ enormous)$	trigram
2)	$p(steak\|\ enormous)$	bigram
3a)	$p(steak)$	unigram (static)
3b)	$p(steak\|\text{TOPIC})$ $= \sum_{i=1}^{1024} p(steak\|Rcat_i) \times p(Rcat_i\|context)$	topic sensitive unigram

- Sensitive to long distance dependencies
- Successful in face of sparse n-grams
- Improves smoothed probability estimates

Performance

Word	Sense	Roget Category	Accuracy
sentence	punishment	LEGAL_ACTION	98%
	set of words	GRAMMAR	
mole	quantity	CHEMICALS	99%
	mammal	ANIMAL	
	skin blemish	DISEASE	
taste	preference	PARTICULARITY	93%
	flavor	SENSATION	
duty	obligation	DUTY	96%
	tax	PRICE,FEE	

[Yarowsky, 1992]

$\Rightarrow$ 92% mean accuracy

Problem: Lexical Ambiguity Resolution

- Word sense disambiguation
- Lexical choice in machine translation
- Homograph disambiguation in speech synthesis
- Accent restoration in Spanish and French
- Other applications

Three Algorithms:

- Decision lists (supervised)
- Bayesian word-class discriminators (unsupervised)

⇒ Modulated bootstrapping from seed words (unsupervised)

Motivating Phenomena

- **One sense per collocation**
- **One sense per discourse:**

Word	Senses	Accuracy	Applicability
tank	vehicle/contnr	99.6 %	50.5 %
motion	legal/physical	99.9 %	49.8 %
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 %
plant	living/factory	99.8 %	72.8 %
crane	bird/machine	100.0 %	49.1 %
Average		**99.8 %**	**50.1 %**

$\Rightarrow$ Algorithm driven by the joint exploitation of these properties

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 $\Rightarrow$ (A) manufacturing plant or

$\Rightarrow$ (B) living plant

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.

Example Initial State

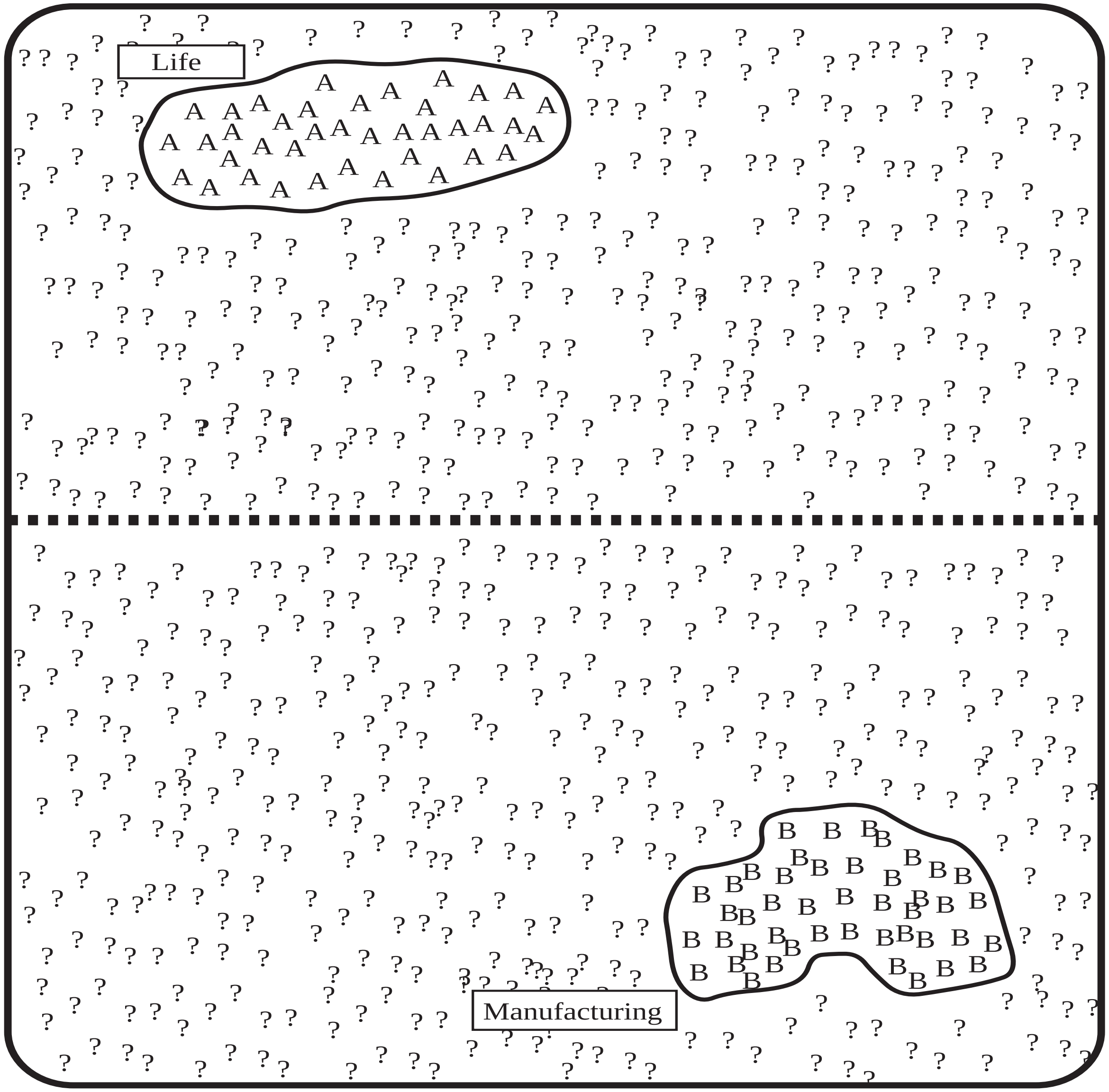
Sense	Training Examples	(Keyword in Context)
A	used to strain microscopic	*plant* **life** from the ...
A	... rapid growth of aquatic	*plant* **life** in water ...
A	... that divide **life** into	*plant* and animal kingdom
A	beds too salty to support	*plant* **life** . River ...
A	...	...
?	... company 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 ...
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

A = Seed contexts containing the collocation *life* (1%)

B = Seed contexts containing the collocation *manufacturing* (1%)

? = Untagged residual (98%)

Example Initial State



Iteration Step

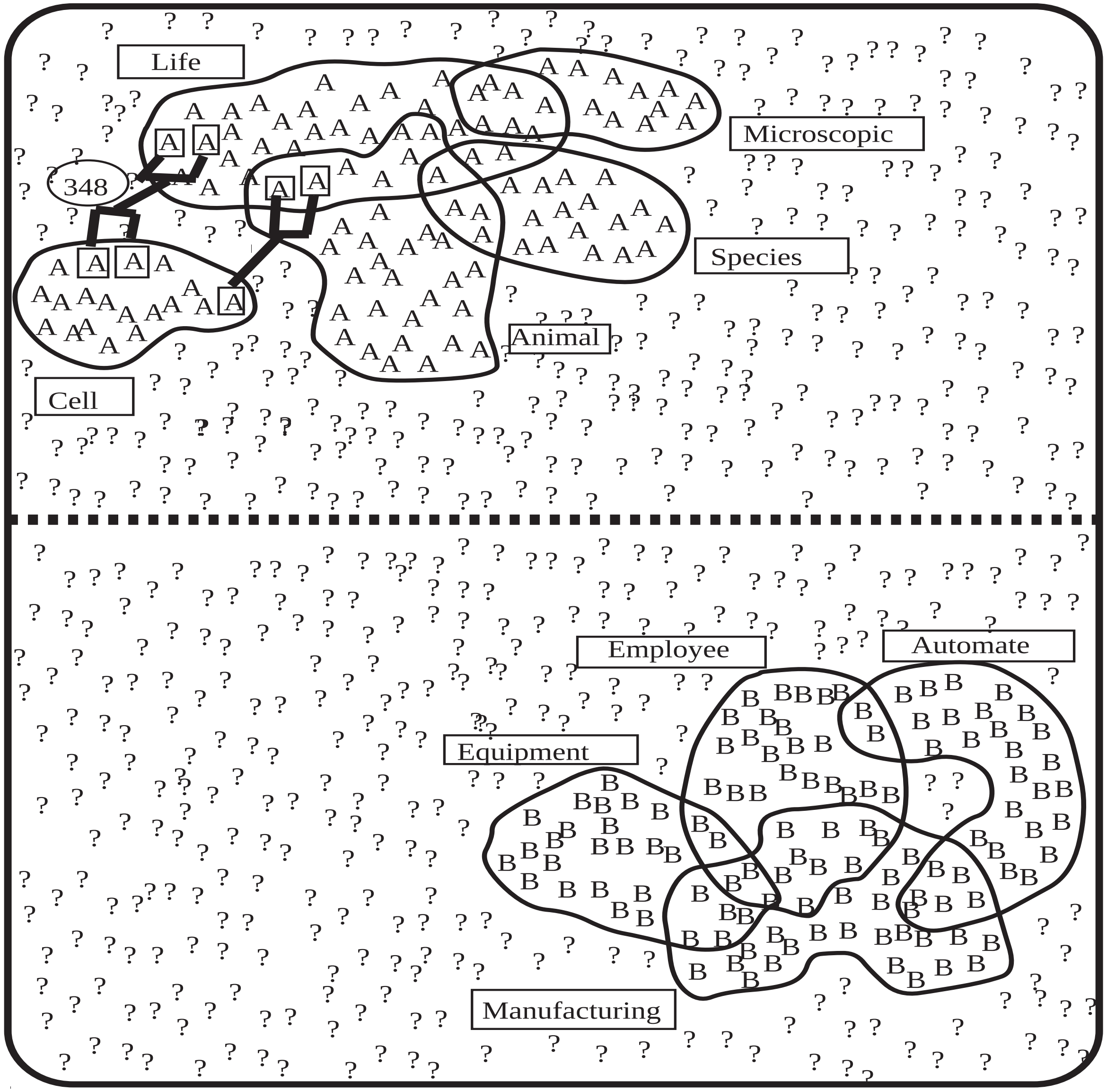
- Train a supervised sense tagger on the current seed sets

Initial decision list for *plant* (abbreviated)

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

- Apply the resulting tagger to the residual examples
- Add the examples exceeding threshold to the growing seed sets

Example Intermediate State



Use of the one-sense-per-discourse constraint

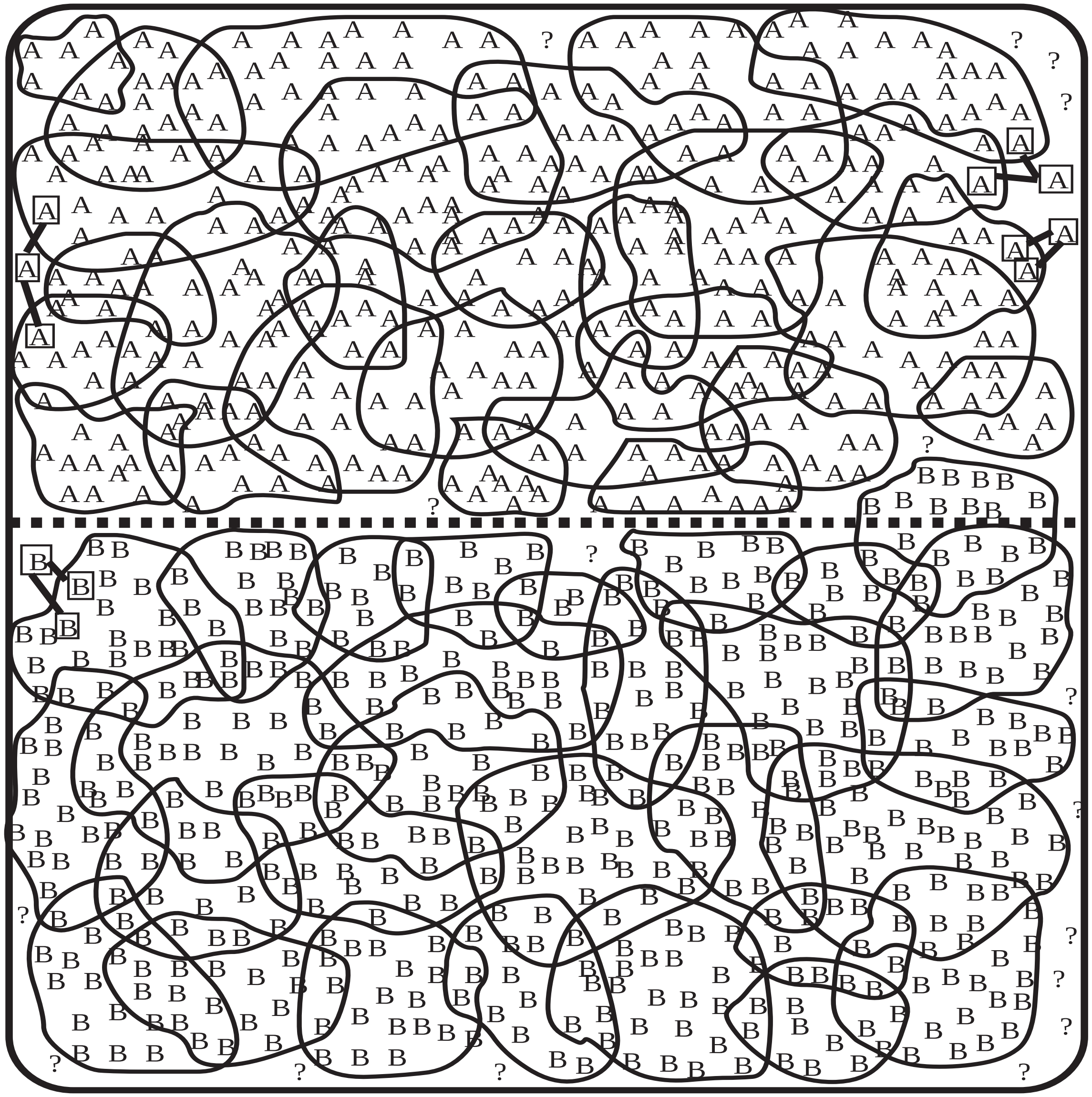
- **Error correction**

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

- **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

Final Training Iteration



Final Decision List

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
...	...	

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

Escaping from Initial Misclassification

- **Discourse consistency** can override local collocational evidence
- **Redundancy of language** makes the process self correcting
- **Change in training parameters**
  - incremental increase in context width after intermediate convergence
  - perturbation of the class-inclusion threshold (similar to simulated annealing)

Performance

Word	Senses	Samp. Size	% Major Sense	Supvsd Algrtm	Seed Training Options: Two Words	Dict. Defn.	Top Colls.	With OSPD	Schütze Algrthm
plant	living/factory	7538	53.1	97.7	97.1	97.3	97.6	98.6	92
space	volume/outer	5745	50.7	93.9	89.1	92.3	93.5	93.6	90
tank	vehicle/container	11420	58.2	97.1	94.2	94.6	95.8	96.5	95
motion	legal/physical	11968	57.5	98.0	93.5	97.4	97.4	97.9	92
bass	fish/music	1859	56.1	97.8	96.6	97.2	97.7	98.8	–
palm	tree/hand	1572	74.9	96.5	93.9	94.7	95.8	95.9	–
poach	steal/boil	585	84.6	97.1	96.6	97.2	97.7	98.5	–
axes	grid/tools	1344	71.8	95.5	94.0	94.3	94.7	97.0	–
duty	tax/obligation	1280	50.0	93.7	90.4	92.1	93.2	94.1	–
drug	medicine/narcotic	1380	50.0	93.0	90.4	91.4	92.6	93.9	–
sake	benefit/drink	407	82.8	96.3	59.6	95.8	96.1	97.5	–
crane	bird/machine	2145	78.0	96.6	92.3	93.6	94.2	95.5	–
AVG		**3936**	**63.9**	**96.1**	**90.6**	**94.8**	**95.5**	**96.5**	**92.2**

Baseline (% major sense)	63.9%
Two defining words	90.6%
Dictionary definitions	94.8%
Top collocations (2 minutes work)	95.5%
Dictionary defns. (with OSPD)	96.5%
Fully supervised algorithm	96.1%

Conclusion

- Unavailability of hand-tagged training data has been a bottleneck for progress in sense disambiguation
- This algorithm, trained on raw text and an on-line dictionary without any human supervision, rivals the performance of fully supervised methods
- Thus, costly hand-tagged training data may be unnecessary to achieve accurate lexical ambiguity resolution.

Gender Classification

Problem:

... company president *Burak* Chopra announced his plan ...

⇒ MALE or

⇒ FEMALE

Training Data:

Gender	Context
(1) male	... company president *Burak* Chopra announced his plan ...
” ”	... and they hired Mr. *Walter* Brill as an accountant ...
” ”	... the young actor *Keanu* Reeves was paid over 5 ...
(2) female	... the noted author *Ardinia* Lospel listed her favorite ...
” ”	... and his sister *Susan* Miller was also found ...
” ”	... members included Dr. *Livonia* Dey who said she would ...

Test Data:

Gender	Context
???	... the retired General *Fidel* Ramos died last night ...
???	... was visited by *Altonette* Smith, a doctor from St. ...

Problem: Is an unusual name male of female?

Alditha	$\Rightarrow$	FEMALE
Ardinia	$\Rightarrow$	FEMALE
Altonnette	$\Rightarrow$	FEMALE
Burak	$\Rightarrow$	MALE
Deryk	$\Rightarrow$	MALE

Solution: Look at Final Characters (Suffix) of Word

–	–	a	⇒	99+% FEMALE
t	t	e	⇒	97% FEMALE
–	–	k	⇒	98% MALE
–	–	d	⇒	96% MALE
–	–	p	⇒	97% MALE

Application: Gender Classfication

Problem: Where to obtain training data?

- Available name databases not labelled with gender
- How to identify gender in a large employee name database?

Application: Gender Classfication

Problem: Where to obtain training data?

- Available name databases not labelled with gender
- How to identify gender in a large employee name database?

Solution: Gender for a name is very closely correlated with the mean *salary* of persons with the name

Name	Mean Salary Grade
Bernard	6.92
Phillip	6.47
Arthur	6.39
Sandra	4.64
Carolyn	4.47
Dorthy	4.11

$SG > 5.3 \Rightarrow$ Male, $SG < 5.3 \Rightarrow$ Female

Problem: What about *Adam* **and** *Todd*

Name	Mean Salary Grade
Bernard	6.92
Phillip	6.47
Arthur	6.39
David	5.91
Robert	5.87
John	5.47
Susan	5.24
Adam	5.13
Todd	5.09
Sandra	4.64
Carolyn	4.47
Dorthy	4.11

Problem: Age is a Factor

Name	Mean Salary Grade
Bernard	6.92
Phillip	6.47
Arthur	6.39
David	5.91
Robert	5.87
John	5.47
Susan	5.24
Adam	5.13
Todd	5.09
Sandra	4.64
Carolyn	4.47
Dorthy	4.11

Problem: Age is a Factor

Solution:

- Compute mean age for a name from references in AP Newswire

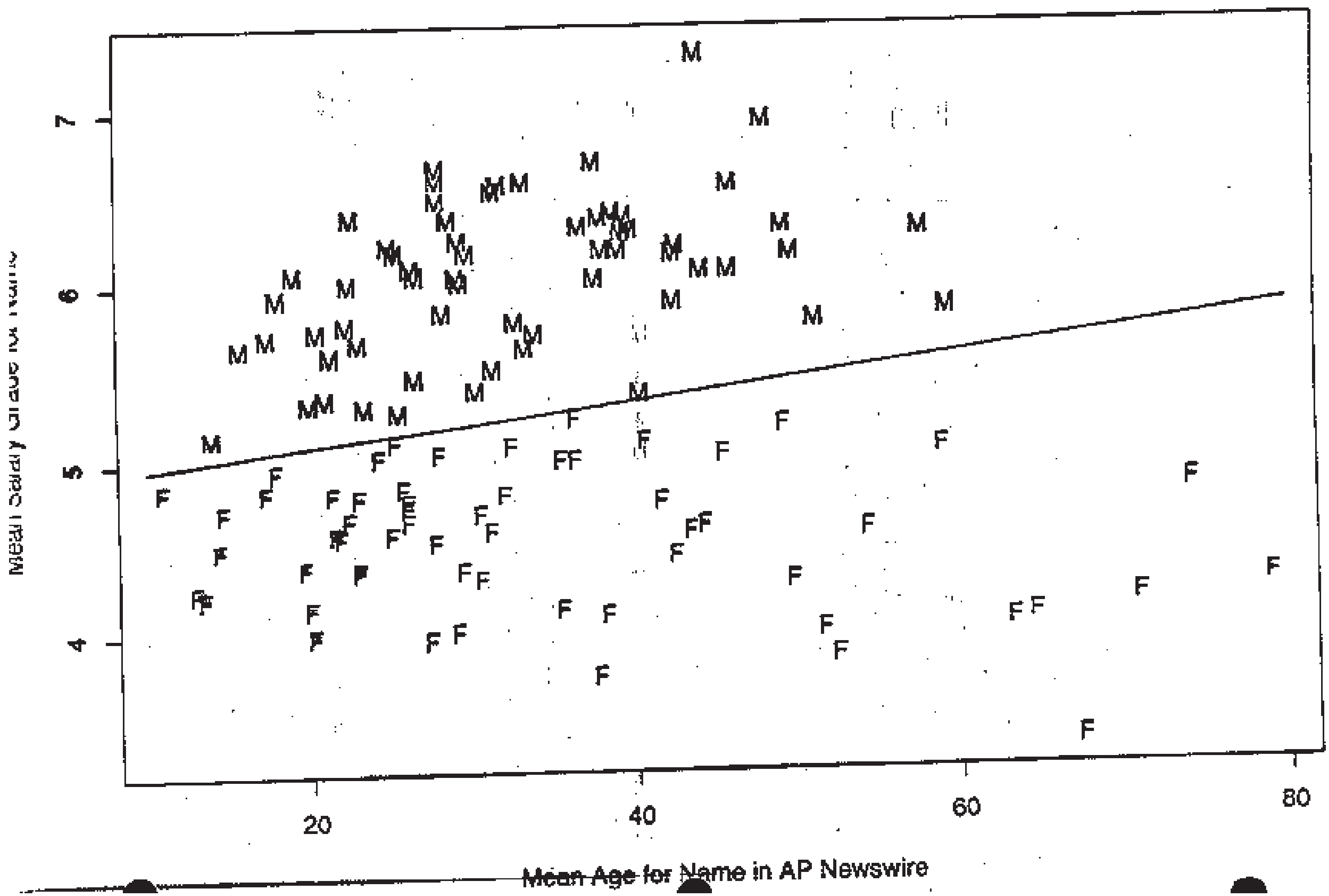
  Matthew Stuart , 23 , said he was not aware ...

  Mildred Jones , 87 , died yesterday in Boston ...

  Todd Wilson , 11 , was abducted from outside ...

Name	Mean Age in AP
Ethel	64.7
Mildred	63.3
Elmer	60.0
Todd	23.1
Heather	22.4
Tammy	20.0

Correlation between Gender and Mean Salary for Name



Moral

- Training data is often difficult to obtain

  (Especially finding automatic sources for annotation)
- However, doing so can be half the fun