1) Identify unambiguous types of each class

CITY - AP datelines
PERSON - AT&T employee database

2) Collect training contexts

3) Measure the distribution of word associations at various positic

- word to left
- word to right
- words in +/- 5 context window

...	from	Aberdeen	...
	in	Boston	
	visited	Sydney	
	suburban	Akron	
	near	Albany	
	SENT	Austin	
	in	Amman	
	of	Philadephia	
	leave	Detroit	

...	said	Baker	...
	condemn	Fonda	
	opposing	Gephardt	
	predicted	Clinton	
	SENT	Smith	
	with	Hurd	
	said	Thatcher	
	Ms	Davidson	
	but	Shamir	

$$I(x;y) = \frac{P(y|x)}{P(y)}$$

$$I(in;PLACE) = \frac{P(in|PLACE)}{P(in)} = \frac{\text{CONDITIONAL PR}}{\text{GLOBAL PRO}}$$

$$\frac{\;\frac{P(in|PLACE)}{\cancel{P(in)}}\;}{\;\frac{P(in|PERSON)}{\cancel{P(in)}}\;} \Rightarrow \frac{P(in|PLACE)}{P(in|PERSON)}$$

4) **Compute log likelihoods**

	f(city)	f(person)	Majority Class	Context	
3.83	3	48	PERSON	said	PERSON/CITY
4.43	5	112	PERSON	"	PERSON/CITY
3.69	2	26	PERSON	with	PERSON/CITY
3.17	31	248	PERSON	#SENT#	PERSON/CITY
2.45	1	8	PERSON	by	PERSON/CITY
4.25	189	8	CITY	in	PERSON/CITY
2.95	17	2	CITY	near	PERSON/CITY
2.94	29	3	CITY	from	PERSON/CITY
2.70	148	20	CITY	of	PERSON/CITY
2.45	3	0	CITY	outside	PERSON/CITY
2.03	30	6	CITY	at	PERSON/CITY
1.48	5	1	CITY	nearby	PERSON/CITY
1.17	26	10	CITY	to	PERSON/CITY

5) **Score new contexts using combination of the models for various positions**

$$Prob\_Ratio = \prod_{tok\ in\ context} \frac{Pr(tok_i|PERSON)}{Pr(tok_i|CITY)}$$

$$Log\_Prob = \sum_{tok\ in\ context} \log \frac{Pr(tok_i|PERSON)}{Pr(tok_i|CITY)}$$

CITY LEFT CONTEXT

felt	in	Aberdeen	and	other
time	,	Aberdeen	was	in
Crawford	of	Aberdeen	were	fishing
board	into	Aberdeen	harbor	,
,	which	Aberdeen	police	said
northeast	of	Aberdeen	climbed	sharply
her	suburban	Akron	home	in
,	near	Akron	,	took
accident	in	Akron	,	where
Capitol	in	Albany	.	.PP
south	of	Albany	,	said
back	toward	Albany	,	a
fire	threatened	Albany	,	the
Shultz	at	Amman	's	military
,	visited	Amman	on	Saturday
capital	of	Amman	.	.PP
``	the	Amman	decision	will

PERSON LEFT CONTEXT

to	protest	Fonda	's	visit
behalf	of	Fonda	.	.PP
not	ask	Fonda	publicist	Jerry
spokesman	for	Fonda	,	called
,	predicted	Fonda	's	apology
``	typical	Fonda	hogwash	''
meeting	with	Hurd	.	.SE
Preston	,	Hurd	was	to
,	''	Hurd	said	in
Cabinet	member	Hurd	's	Home
offered	by	Pell	aide	William
,	said	Pell	.	The
.SB	The	Pell	resolution	declares
,	''	Pell	told	reporters

```
0 CITY    -1458    miles northeast of >Aberdeen< climbed sharply as
1 PERS     1124    the Lexington and >Aberdeen< weapons to Tooele
0 CITY    -6324    major city of >Aberdeen< to its collection
0 CITY    -2906    load arriving at >Aberdeen< in two weeks
1 PERS     1124    Rapid City and >Aberdeen< setting records for
0 CITY    -4558    port city of >Aberdeen< said in a
0 CITY    -3735    and police in >Aberdeen< at 4:15 a.m
1 PERS      388    leave immediately for >Aberdeen< to oversee care
0 CITY    -5453    Elder arrived in >Aberdeen< by helicopter ,
0 CITY    -1943    company heads in >Aberdeen< before visiting injured
0 CITY      -63    also flew to >Aberdeen< to console families
0 CITY    -4770    married life in >Aberdeen< when her 24-year-old
0 CITY    -2711    was felt in >Aberdeen< and other parts
0 CITY    -5739    annual passengers through >Aberdeen< airport rose 1,370
0 CITY    -2129    , said in >Aberdeen< he had ``
0 CITY    -1096    coast guard in >Aberdeen< said the Sikorsky
0 CITY    -1351    and east of >Aberdeen< and spotty elsewhere
0 CITY    -4064    news conference in >Aberdeen< total settlements in
0 CITY    -1986    Angie Crawford of >Aberdeen< were fishing in
1 PERS     1520    '' Chapla said >Aberdeen< is used to
0 CITY     -955    being produced at >Aberdeen< proving ground .
0 CITY    -1410    will return to >Aberdeen< this weekend if
0 CITY    -6705    on board into >Aberdeen< harbor , police
0 CITY    -2924    accident , which >Aberdeen< police said occurred
0 CITY    -5582    told reporters in >Aberdeen< that two explosions
1 PERS      113    be flown to >Aberdeen< for identification .
1 PERS     1045    police spokesman in >Aberdeen< said .
0 CITY    -1025    - Folks in >Aberdeen< are building tall
1 PERS     1045    guard spokesman in >Aberdeen< said .
```

```
MLE Pass 0:  P: 0.724138  (21/29)      c1: 0.000000
MLE Pass 1:  P: 0.827586  (24/29)      c1: 965.080750
MLE Pass 2:  P: 0.965517  (28/29)      c1: 1568.616089
MLE Pass 3:  P: 1.000000  (29/29)      c1: 3332.203857
```

Next row

$$\frac{P\,(\text{Place} \mid w)}{P\,(\text{Person} \mid w)} = \frac{P\,(\text{Place})}{P\,(\text{Person})} * \frac{P\,(w \mid \text{Place})}{P\,(w \mid \text{Person})}$$

<u>Posterior</u>
Probability

<u>Prior</u>
Probability

What if we don't know initial prior ratio (initial odds) ?

Ans:

Start off with ratio of 1,
then iteratively reestimate

EM ITERATION

- Begin with uninformative prior probability
- $Final\_Prob = Prior\_Prob \times Model\_Prob$
- Score all instances of a name with above
- Recompute Prior_Prob

Old Prior		New Classification		
.50	$\Rightarrow$	21/29	(.72)	
.72	$\Rightarrow$	24/29	(.84)	
.84	$\Rightarrow$	28/29	(.97)	
.97	$\Rightarrow$	29/29	(1.00)	
1.00	$\Rightarrow$	29/29	(1.00)	$\Rightarrow$ **Convergence**

Old Prior		New Classification		
.50	⇒	21/29	(.72)	
.72	⇒	24/29	(.84)	
.84	⇒	28/29	(.97)	
.97	⇒	29/29	(1.00)	
1.00	⇒	29/29	(1.00)	⇒ **Convergence**

Old Prior		New Classification		
.50	⇒	20/100	(.20)	
.20	⇒	7/100	(.07)	
.07	⇒	3/100	(.03)	
.03	⇒	1/100	(.01)	
.01	⇒	0/100	(.00)	
.00	⇒	0/100	(.00)	⇒ **Convergence**

Old Prior		New Classification		
.50	⇒	53/89	(.59)	
.59	⇒	57/89	(.64)	
.64	⇒	62/89	(.70)	
.70	⇒	64/89	(.72)	
.72	⇒	65/89	(.73)	
.73	⇒	65/89	(.73)	⇒ **Convergence**

0.000000	0.000000	0.004950	100	Anderson
0.000000	0.000000	0.004950	100	Baker
0.000000	0.000000	0.004950	100	Burns
0.000000	0.000000	0.008065	61	Walker
0.000000	0.000000	0.008333	59	Tucker
0.000000	0.000000	0.009259	53	Campbell
0.000000	0.000000	0.009804	50	Richardson
0.000000	0.000000	0.011364	43	Martinez
0.000000	0.020000	0.024752	100	Taylor
0.000000	0.024390	0.035714	41	Hinckley
0.000000	0.027027	0.039474	37	Roosevelt
0.000000	0.030303	0.044118	33	Hayes
0.126471	0.070000	0.074257	100	Williams
0.165132	0.153846	0.166667	26	Perry
0.186924	0.181818	0.195652	22	Stanley
0.347209	0.311111	0.315217	45	Carson
0.367992	0.357143	0.366667	14	Greenfield
0.371823	0.363636	0.375000	11	Hershey
0.435153	0.428571	0.431818	21	Greenwood
0.462830	0.458333	0.460000	24	Medina
0.500000	0.500000	0.500000	12	Chatham
0.500000	0.500000	0.500000	26	Dixon
0.500000	0.500000	0.500000	40	Rhodes
0.528220	0.534884	0.534091	43	Florence

0.347209	0.311111	0.315217	45	Carson
0.367992	0.357143	0.366667	14	Greenfiel(
0.371823	0.363636	0.375000	11	Hershey
0.435153	0.428571	0.431818	21	Greenwood
0.462830	0.458333	0.460000	24	Medina
0.500000	0.500000	0.500000	12	Chatham
0.500000	0.500000	0.500000	26	Dixon
0.500000	0.500000	0.500000	40	Rhodes
0.528220	0.534884	0.534091	43	Florence
0.540823	0.553571	0.552632	56	Pyongyang
0.544586	0.570000	0.569307	100	Baghdad
0.647015	0.720000	0.717822	100	Islamabad
0.654478	0.730000	0.727723	100	Beijing
0.654478	0.730000	0.727723	100	Berlin
0.684328	0.770000	0.757426	100	Austin
0.790246	0.785714	0.766667	14	Warwick
0.898864	0.929825	0.922414	57	Madison
1.000000	0.990000	0.985148	100	Budapest
1.000000	0.990000	0.985148	100	Kabul
1.000000	0.990000	0.985148	100	Khartou
1.000000	0.990000	0.985148	100	Sacrament
1.000000	1.000000	0.995049	100	Tampa
1.000000	1.000000	0.995049	100	Zurich

OUTPUT OF ALGORITHM

1. A model for classifying an instance of a word as PERSON or PLACE based on context

2. The probability that a given name is either a person or a place based on a collective analysis of all its instances

OTHER EVIDENCE

1) WORD-INTERNAL EVIDENCE

Krulovich	==>	PERSON
Yarowsky	==>	PERSON
Smitterson	==>	PERSON
Endlersberg	==>	PERSON
Endlersburg	==>	PLACE
Kotterston	==>	PLACE
Siouxport	==>	PLACE
Causville	==>	PLACE

---

2) MORE REFINED MODELS OF CONTEXT

- syntactic relations (subj/verb, verb/obj)

Hanoi was invaded	==>	invade/V Hanoi
Dole was married	==>	marry/V Dole

- trigrams (in PLACE said)
- wide context window

---

3) CLASS-MODELS

- Part of speech

```
PERSON bought
PERSON listened  ==>   PERSON <VBD>
PERSON ran
```

- Lemmas (say/V = said/say/saying/says...)

- Semantic (thesaurus) classes

---

4) BURST MODELLING
DISCOURSE MODELLING
TOPIC MODELLING

==> How to combine these non-independent sources of evidence?