

Sid #	Name	Year	GPA
1	Smith	3	3.0
2	Jones	2	3.5
3	Doe	1	1.2
4	Varda	4	4.0
5	Carey	4	0.5

Student Relation

Fid #	Name	Position	Dept
9	Henry	Prof.	Math
2	Jackson	Assist. Prof	Hist
14	Schuh	Assoc. Prof	Chem
21	Lerner	Assist. Prof	CS

Faculty Relation

Course #	Course Name	Cr	Dept
223	Calculus	5	Math
302	Intro Prog	3	CS
302	Organic Chem	3	Chem
542	Asian Hist	2	Hist
222	Calculus	5	Math

Course Relation

# Extracting Semantics from the Web

Peter Norvig



**Willie Sutton**



# What to Extract

## FROM

- Unstructured text
- Semi-structured
- Structured databases
- Link patterns
- Usage patterns
- ...

## TO

- Probability distributions
- Info. retrieval models
- Language models
- Relational databases
- Semantic networks
- ...

# Effect of large corpus size

Banko & Brill, 2001

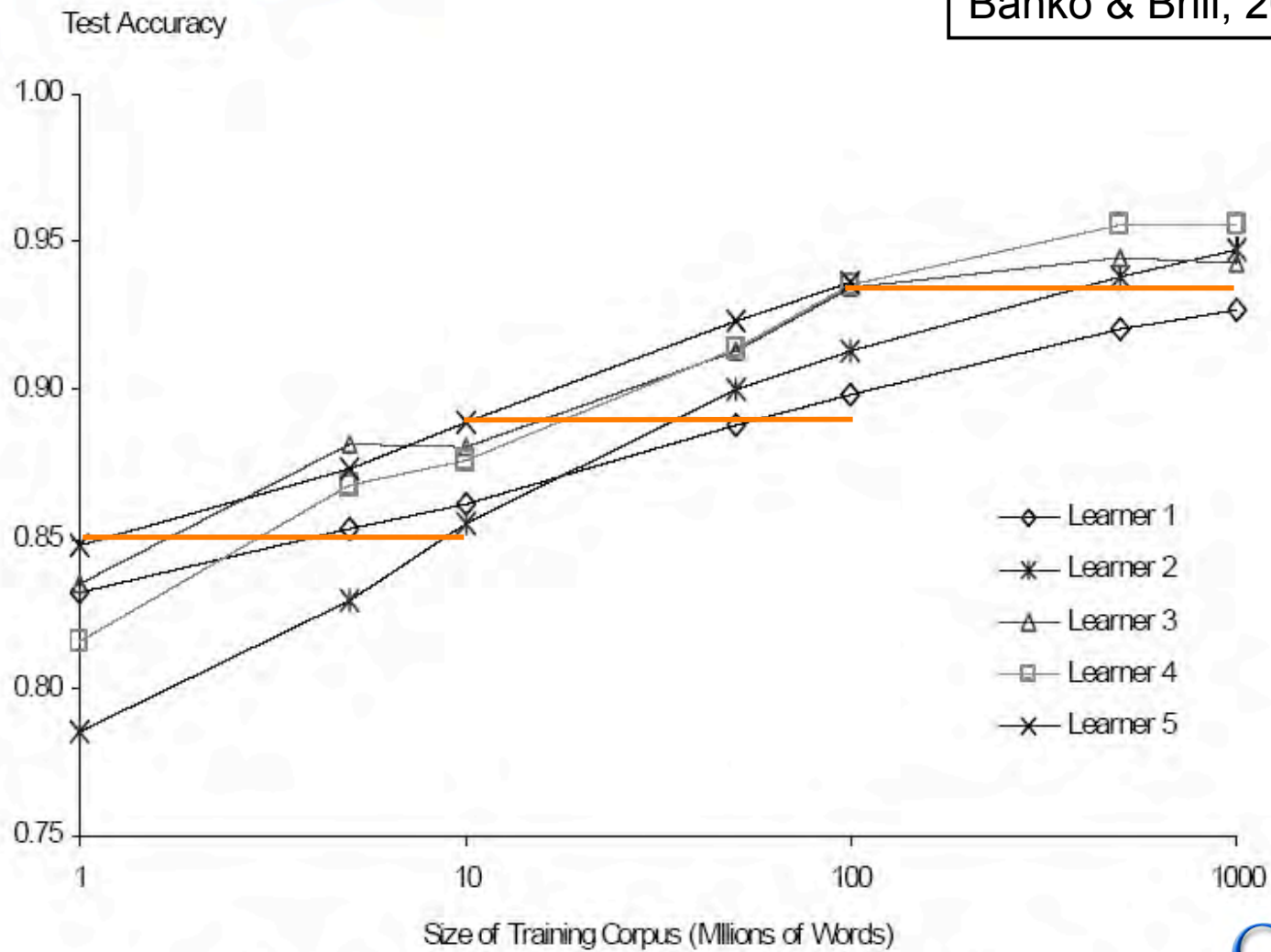
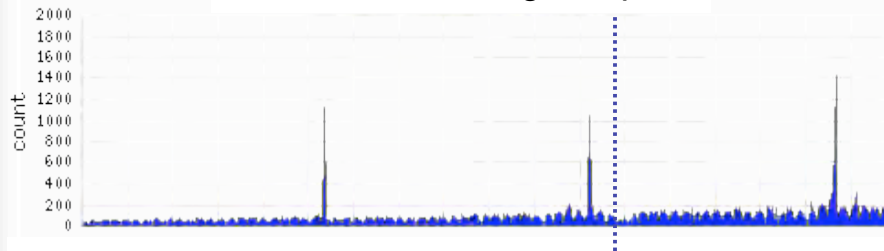


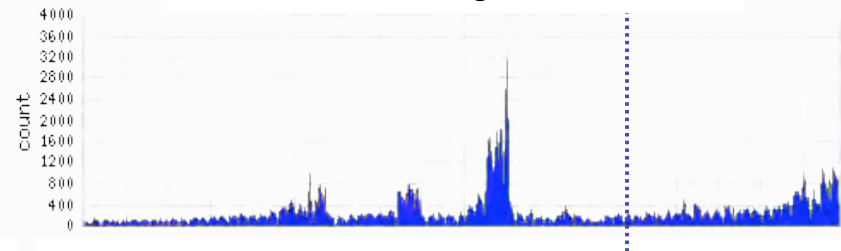
Figure 2. Learning Curves for Confusable Disambiguation

# Extraction from usage patterns

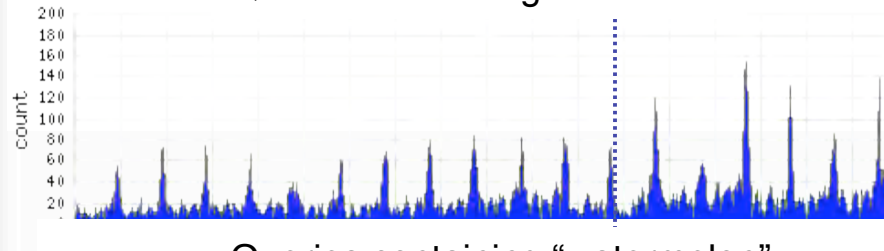
Queries containing “eclipse”



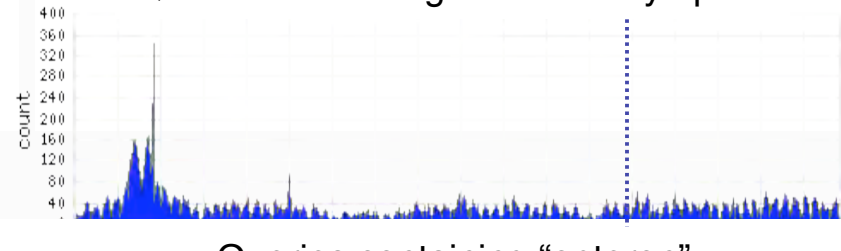
Queries containing “world series”



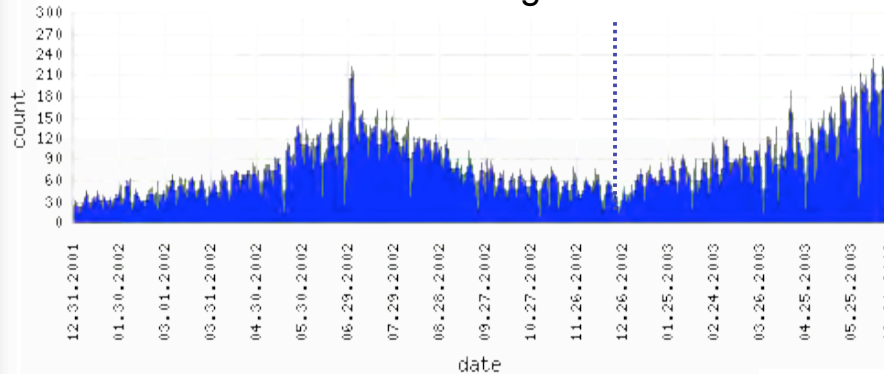
Queries containing “full moon”



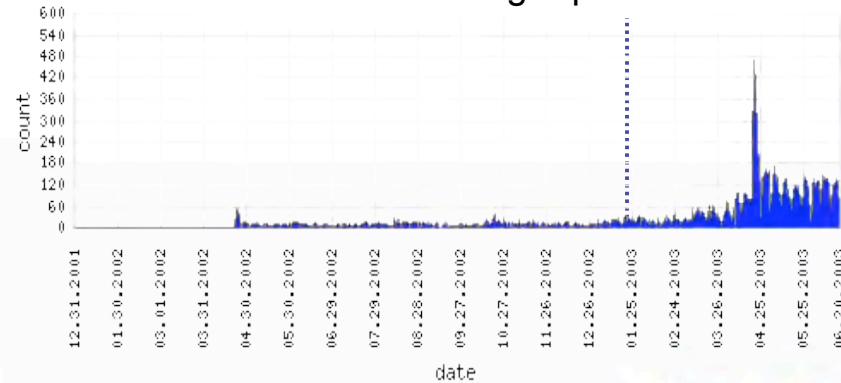
Queries containing “summer olympics”



Queries containing “watermelon”



Queries containing “opteron”





# Spelling model from usage patterns

## Searching for Britney Spears...

488941 britney spears	20 britney spears	9 britnany spears	5 bney spears	3 britiy spears	2 brrrney spears
40134 britvany spea	britany spears	9 britanay spears	5 broitney spears	3 britmeny spears	2 britvany spears
36315 brittney spea	britny spears	9 britinay spears	5 brobny spears	3 britneey spears	2 britvany spears
24342 britany spears	britny spears	9 britn spears	5 brubny spears	3 britnehy spears	2 britvney spears
7331 britny spears	briteny spears	9 britnew spears	5 bruiyney spears	3 britnely spears	2 britain spears
6533 briteny spears	briteny spears	9 britneyn spears	5 briritney spears	3 britnesy spears	2 britane spears
2595 britney spea	britteny spears	9 britney spears	5 gritney spears	3 britnetty spears	2 britaneny spears
1807 briney spears	briney spears	9 brtiny spears	5 spritney spears	3 britnek spears	2 britania spears
1635 britny spears	britny spears	9 brtittney spears	4 bittny spears	3 britneyxxx spears	2 britanu spears
1479 britney spears	britny spears	9 brtny spears	4 britny spears	3 britnity spears	2 britanna spears
1479 britney spea	brintey spears	9 brytny spears	4 brandy spears	3 britney spears	2 britannie spears
1338 britny spears	britanny spears	9 rbitney spears	4 bibrithney spears	3 britneyey spears	2 britannt spears
1211 britnet spears	britanny spears	8 birtiny spears	4 breatiny spears	3 britterny spears	2 britannu spears
1096 britney spea	britiny spears	8 bithney spears	4 breetney spears	3 britoneey spears	2 britanyl spears
991 britney spea	britiny spears	8 bractany spears	4 brectany spears	3 brittonney spears	2 britany spears
991 britney spears	britnet spears	8 braitny spears	4 brifitney spears	3 brittonyey spears	2 briteeny spears
811 britney spea	britney spears	8 bretny spears	4 briaatany spears	3 brityen spears	2 britenany spears
811 britney spears	britney spears	8 brightny spears	4 brieveny spears	3 briyney spears	2 britenet spears
654 britney spears	britaney spears	8 brinty spears	4 brieety spears	3 brlney spears	2 briteniy spears
654 britney spea	britney spears	8 brittney spears	4 briitny spears	3 brobeny spears	2 britenys spears
654 britney spea	britney spears	8 britotney spears	4 briitany spears	3 brtaney spears	2 britianey spears
601 britney spears	brithney spears	8 britany spears	4 brinie spears	3 brtliany spears	2 britin spears
601 britny spears	brtiney spears	8 britley spears	4 brintehy spears	3 brtliny spears	2 britinary spears
544 britney spe	brtiney spears	8 britneyb spears	4 brintne spears	3 brtiney spears	2 britny spears
544 britney spea	birtney spears	8 britneyey spears	4 britaby spears	3 brititany spears	2 britnany spears
364 britny spears	brintney spears	8 britnby spears	4 britacy spears	3 brititany spears	2 britnat spears
364 britny spea	brintney spears	8 brittnez spears	4 britainey spears	3 brtnet spears	2 britnby spears
329 britney spears	briteney spears	8 broctany spears	4 britinie spears	3 brytyny spears	2 britndy spears
269 britney spears	bitney spears	7 baritney spears	4 britinney spears	3 btney spears	2 britnek spears
269 britney spea	brinty spears	7 birtney spears	4 britmney spears	3 dritney spears	2 britneey spears
244 britny spears	brinty spears	7 bitney spears	4 britnear spears	3 preoney spears	2 britney6 spears
244 britny spears	brittaney spears	7 bitny spears	4 britnel spears	3 rbitney spears	2 britneye spears
220 bzeatney spea	brittany spears	7 brianey spears	4 britnew spears	2 baritbany spears	2 britneyh spears
220 britiany spea	brittany spears	7 bzianty spears	4 britnewy spears	2 bbbritney spears	2 britneym spears
199 britney spea	brity spears	7 brintye spears	4 britneyey spears	2 bbitney spears	2 britneyyy spears
163 britny spears	brittiny spears	7 britianny spears	4 brittaby spears	2 bbritny spears	2 britneyj spears
147 breatny spears	brittiny spears	7 britly spears	4 brittety spears	2 bbritbany spears	2 britneyk spears
147 britiny spea	brtney spears	7 britnej spears	4 britthey spears	2 beitany spears	2 britne spears
147 brity spears	brtney spears	7 britneyu spears	4 brittoney spears	2 beitny spears	2 britnu spears
147 brooney spears	brtney spears	7 britniy spears	4 brittnat spears	2 bertney spears	2 britoney spears
147 brunny spears	brtney spears	7 britnny spears	4 brittneny spears	2 bertny spears	2 britiany spears
133 britteney spe	britneys spears	7 brittian spears	4 brittnye spears	2 betney spears	2 britreny spears
133 brinyey spears	britne spears	7 brityny spears	4 brittteny spears	2 betny spears	2 britry spears
121 britany spears		7 brzittany spears	4 brittuney spears	2 bhriney spears	2 britsany spears
121 britney spears	17 brittanie spears	7 brttiney spears	4 brityny spears	2 biney spears	2 brittanay spears
121 britney spears	15 brinney spears	7 britny spears	4 brityy spears	2 binbey spears	2 brittang spears
121 britney spears	15 briten spears	7 brittany spears	4 brityny spears	2 birecty spears	2 britteans spears
109 britney spears	15 britterney spears	6 beritny spears	4 brobany spears	2 biritany spears	2 brittanyh spears
109 britny spears	15 britheny spears	6 bhritney spears	4 bryney spears	2 biritbany spears	2 brittanym spears



# Suggestions

Google News - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

Google a Search

amazon	214,000,000 results
ask jeeves	1,240,000 results
argos	4,340,000 results
aol	59,300,000 results
aim express	5,390,000 results
adaware	1,410,000 results

Caribou peter.norvis

News Search and browse 4,500

Google News - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

Google aaa Search

aaa	19,100,000 results
aaa.com	1 result
aaa travel	4,470,000 results
aaaa	3,270,000 results
aaa insurance	2,180,000 results
aaas	1,570,000 results

Caribou peter.norvis

News Search and browse 4,500



# Extraction and remixing

- For Rent
- For Sale

City:

Price Range:

[Refresh](#)

Powered by [craigslist](#) and [Google Maps](#)  
(this site is in no way affiliated with craigslist or Google)

[Send Feedback](#)

**\$1650**  
**Studio Apt For ReNT**  
 E 61st St & 1st Ave  
 New York

917-892-8431 / [email](#)  
 Directions: [To here](#) - [From here](#)

Map interface ©2005 Google  
 Map data ©2005 NAVTEQ™, Tele Atlas

pics	price	bd	description	city	date
	\$1500	1bd	<a href="#">Gorgeous Furnished Penthouse with Boatslip</a>	Aventura	4/12
	\$1800	3bd	<a href="#">Rent Brand New House With Option To Buy</a>	Brookhaven	4/12
	\$1700	1bd	<a href="#">Garden City Village Apartment</a>	Garden City	4/13
	\$1500		<a href="#">Huge 2 room studio! Must see!</a>	Greenwich	4/12
	\$1500	2bd	<a href="#">Brand New 2 Bedroom</a>	Hempstead	4/13
	\$1500	2bd	<a href="#">2 level, 2 bedroom sunny, cute apt for rent</a>	Hempstead	4/13
	\$1500	2bd	<a href="#">Beautiful 2 bedroom apt available. Definately a Must see!</a>	Hoboken	4/12
	\$1550	3bd	<a href="#">House for rent</a>	Islip	4/12
	\$1700	3bd	<a href="#">Relaxed, Laid back living!</a>	Jersey City	4/12
	\$1795	1bd	<a href="#">Luxury Park Foundry Loft</a>	Jersey City	4/12
	\$2000	2bd	<a href="#">Spectacular 2 Bedroom Duplex! 10 minutes from Manhattan</a>	Jersey City	4/12
	\$1500	3bd	<a href="#">2BA, HWY floors, W/D hookup, private balcony, close to public transp.</a>	Jersey City	4/12
	\$1700	2bd	<a href="#">Beautiful Apartment In Historic Downtown</a>	Jersey City	4/12
	\$1600	3bd	<a href="#">Jersey City, New Construction 3 Bed room, 2 Full Bathrooms, Garage</a>	Jersey City	4/12
	\$1500	2bd	<a href="#">Gorgeous Space, Unbeatable Location</a>	Montclair Twp	4/13
	\$1675	3bd	<a href="#">3 Bdrm 1 1/2 bath</a>	New Paltz	4/12
	\$1900	1bd	<a href="#">Parkslope Floor Through + Laundry</a>	New York	4/13
	\$1800	1bd	<a href="#">800 sq ft plus Garden in Park Slope</a>	New York	4/13
	\$1750	2bd	<a href="#">Real 2 Bd apt*No Broker Fee</a>	New York	4/13
	\$1525		<a href="#">Beautiful Spacious Artist Lofts *see photo*</a>	New York	4/13
	\$1500	3bd	<a href="#">Three Bedroom Apt Available</a>	New York	4/13



# Extraction from semi-structured text



## Web

### Japan

**Population: 127,333,002 (July 2004 est.)**

According to <http://www.cia.gov/cia/publications/factbook/fields/2119.html>

### Statistical Handbook of JAPAN 2004

... In 2003, **Japan** had a total **population** of 128 million. **Japan's population** in 2001 was the ninth largest in the world, equivalent to 2.1 percent of the ...

[www.stat.go.jp/english/data/handbook/c02cont.htm](http://www.stat.go.jp/english/data/handbook/c02cont.htm) - 20k - [Cached](#) - [Similar pages](#)

### POPULATION OF JAPAN

... 2000 **Population** of **Japan** provides statistical data on the current states of ... and Ratio of Daytime **Population** to Nighttime **Population - Japan** and ...

[www.stat.go.jp/english/data/kokusei/2000/final/hyodai.htm](http://www.stat.go.jp/english/data/kokusei/2000/final/hyodai.htm) - 69k - Apr 9, 2005 - [Cached](#) -  
[ [More results from www.stat.go.jp](#) ]

Researches on Population Ecology



# Extraction from semi-structured text



Web Images Groups News Froogle Local<sup>New!</sup> Desktop [more »](#)

who is buried in grant's tomb

Search

[Advanced Search](#)  
[Preferences](#)

## Web

### Who Is Buried in Grant's Tomb?

Property: **No one actually. Mr. and Mrs. Grant are not buried. Their sarcophagi lie above ground.**

According to <http://www.jrn.columbia.edu/studentwork/cns/2004-03-15/643.asp>

### [CNS: March 15, 2004: Who's buried in Grant's Tomb?](#)

... You may think that you know the answer to that old question: "Who's **buried** in **Grant's Tomb**?" But even visitors who are standing inside the magnificent ...

[www.jrn.columbia.edu/studentwork/cns/2004-03-15/643.asp](http://www.jrn.columbia.edu/studentwork/cns/2004-03-15/643.asp) - 12k - Apr 9, 2005 - [Cached](#) - [Similar pages](#)

### [Grants Tomb in The NYC Insider: an insider's guide to New York City](#)

... Who's **Buried** in **Grant's Tomb**? This is an old joke to which the answer is no one. Our eighteenth president and his wife are, however, entombed here. ...

[www.theinsider.com/nyc/attractions/?general.htm](http://www.theinsider.com/nyc/attractions/?general.htm) - 48k - Apr 9, 2005 - [Cached](#) - [Similar pages](#)



## Named Entity Extraction (Pasca et al.)

1. Filter HTML, break text into sentences, and mark POS with tagger (Brants' TnT).
2. Match sentences against patterns like  
...  $C$  ["such as" | "including"]  $N$  ...
3. Retain ( $N$  isa  $C$ ) if  $C$  is simple, ends in a plural noun.
4. Discard noisy data; regularize over names.
5. Plug ( $N$  isa  $C$ ) pairs back into text to find new patterns. Iterate.



# Named Entity Extraction

- **Hybrid cars**: Toyota Prius, Honda Insight
- **High-speed networks**: ATM, Gigabit Ethernet, B-ISDN, Myrinet, Frame Relay, Fast Ethernet, ...
- **Rappers**: Eminem, Jay-Z, Nas, Dmx, Snoop Dogg, Dr. Dre, Ja Rule, Mos Def, Nelly, 50-cent ...
- **Anti-depressants**: Prozac, Zoloft, Paxil, Wellbutrin, Effexor, Elavil, Luvox, SSRIs, ...

# Statistical Machine Translation

- Data: aligned parallel corpus
- Model:  $\Pr(e|f) = p_{\lambda_1^M}(e|f) \propto \exp\left[\sum_{m=1}^M \lambda_m h_m(e, f)\right]$
- Training: maximize BLEU scores directly
- Inference: beam search
- Evaluation: BLEU

# Statistical Machine Translation

- Model features ( $h_m$ ):
  - All based on word[i..j] statistics
- **Not** used:
  - Syntax: parser, chunker, POS tagger
  - Semantics: Wordnet, ontologies
  - Annotated data: TreeBank, FrameBank

# Statistical Machine Translation

- ضد عقوبات فرض المخلو الوحيدي هو الامن مجلس ان يذكر  
ال عى ب نكست عتبر ان ها من حذرت التي الشمالية كوري ا  
حرب اعل ان بمثابة
- ايج الح عندما 1974 ال عام من ذ شطرين الى مقسمة وقبرص  
انقل بل على ردا ال جزيرة من الشمالي الثلث التركي الجيش  
ال يونان الى ال جزيرة ضم بمدف قبارصة قوميون نفذه
- ثم اني استمرت ال عراق ضد حربا خاضت التي ايران ان يذكر  
على اميركي عسكري هجوم شن ت عارض (1980-1988) سنوات  
الهل لدمار اسلحة بامتلاك واشن طن تتهمه الذي ال عراق  
بالاره ابمرت بوط بانه



# Statistical Machine Translation

- It is noteworthy that the Security Council is the only authorized to impose sanctions against North Korea, which warned that it would consider sanctions a declaration of war.
- Cyprus has been divided into two parts since the year 1974 when the Turkish army invaded the northern third of the island in response to a coup by Greek nationalists with the aim of annexing the island to Greece.
- It is worth mentioning that Iran, which fought a war against Iraq lasted eight years (1980 - 1988) opposes American military attack on Iraq, which Washington accuses of possessing weapons of mass destruction and that it was linked to terrorism.

# Statistical Machine Translation

- 新华社大马士革4月15日电(记者拱振喜)叙利亚总统巴沙尔·阿萨德15日在此间与来访的美国国务卿鲍威尔举行了会谈,双方讨论了中东局势的最新发展,特别是巴勒斯坦的严重局势以及黎以边界地区的紧张局势等问题。
- 叙利亚通讯社报道,巴沙尔总统在会谈中说:“在巴勒斯坦发生的事件使(中东)和平进程走进了死胡同,如果不能认识到这一点,事情的发展有可能达到无法挽回的程度,那时,我们只能再等待一代人的时间。”
- 他指出,只有在以色列从它占领的巴勒斯坦领土撤军,停止屠杀巴勒斯坦人以后,才可以谈和平进程的问题。

# Statistical Machine Translation

- Xinhua News Agency, Damascus, April 15 (Reporter Gong Zhenxi) Syrian President Bashar Assad 15th here with visiting US Secretary of State Colin Powell held talks, the two sides discussed the latest development of the situation in the Middle East, especially the serious situation in Palestine and the tension in the border region between Lebanon and Israel and other issues.
- According to the Syrian News Agency, President Bashar during the talks, said: "In the incident to the Palestinian (Middle East) peace process into a dead end, if not realize that this is happening may not be able to restore to the extent that time, we can only wait for the generation of time."
- He pointed out that only in the Israeli withdrawal from the occupied Palestinian territories, stop massacre of Palestinians, can talk about the peace process.

# Results, NIST competition

## Arabic-to-English Task, *Large Data Track*

Table 1	
Site	BLEU-4 Score
GOOGLE	0.5131
ISI	0.4657
IBM	0.4646
UMD	0.4497
JHU-CU	0.4348
EDINBURGH	0.3970
SYSTRAN	0.1079
MITRE	0.0772
FSC	0.0037

## Arabic-to-English Task, *Unlimited Data Track*

Table 2	
Site	BLEU-4 Score
GOOGLE	0.5137
SAKHR	0.3403
ARL	0.2257

## Chinese-to-English Task, *Large Data Track*

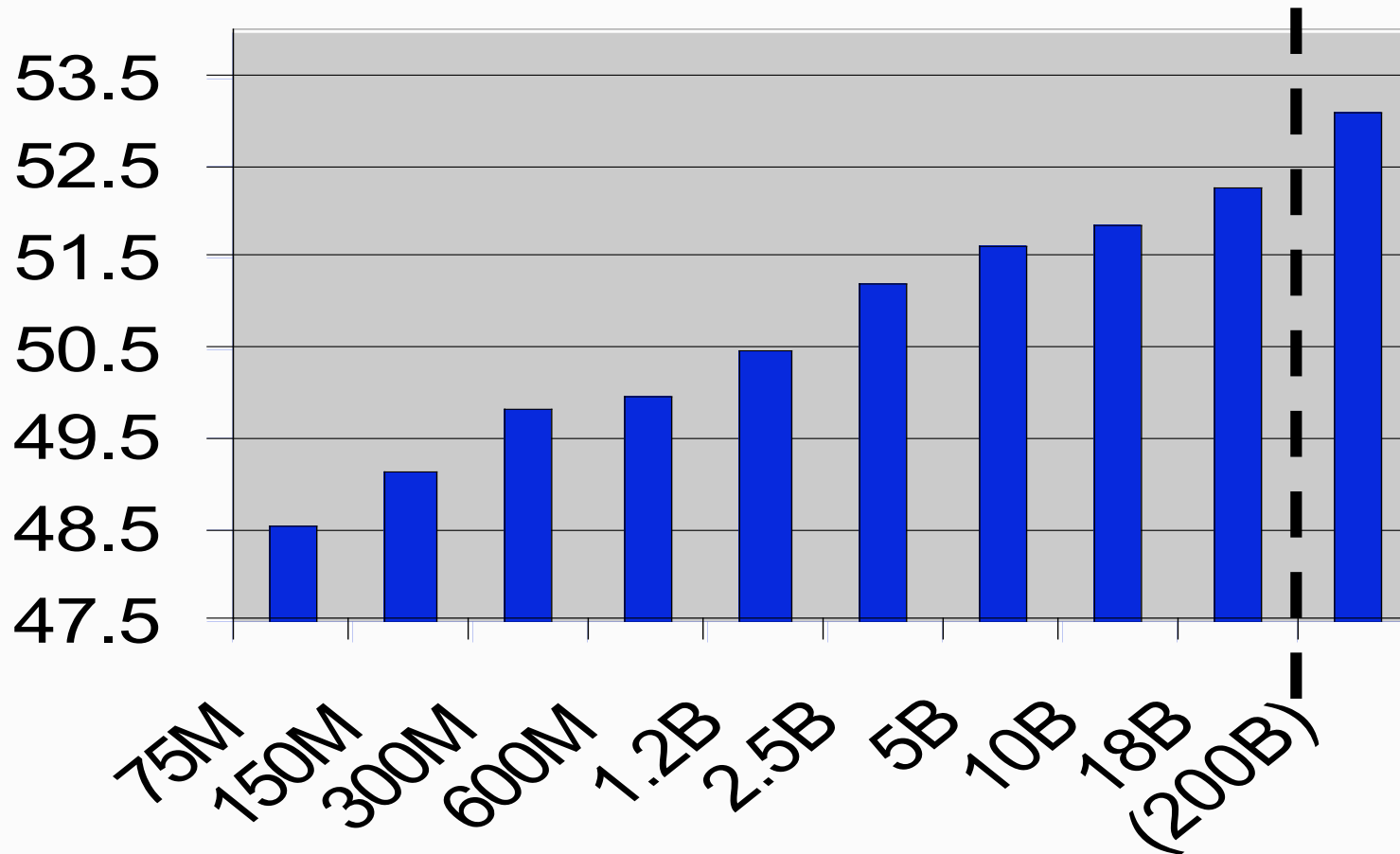
Table 4	
Site	BLEU-4 Score
GOOGLE	0.3531
ISI	0.3073
UMD	0.3000
RWTH	0.2931
JHU-CU	0.2827
IBM	0.2571
EDINBURGH	0.2513
ITCIRST	0.2445
NRC	0.2323
NTT	0.2321
ATR	0.1822
SYSTRAN	0.1471
SAAR	0.1310
MITRE	0.0542

## Chinese-to-English Task, *Unlimited Data Track*

Table 5	
Site	BLEU-4 Score
GOOGLE	0.3516
ICT	0.1293
HIT	0.0797



# More Data Helps



Five-gram language model, no count-cutoff,  
integrated into search



Questions?

# Word Clustering

pepsi

sprite

coke

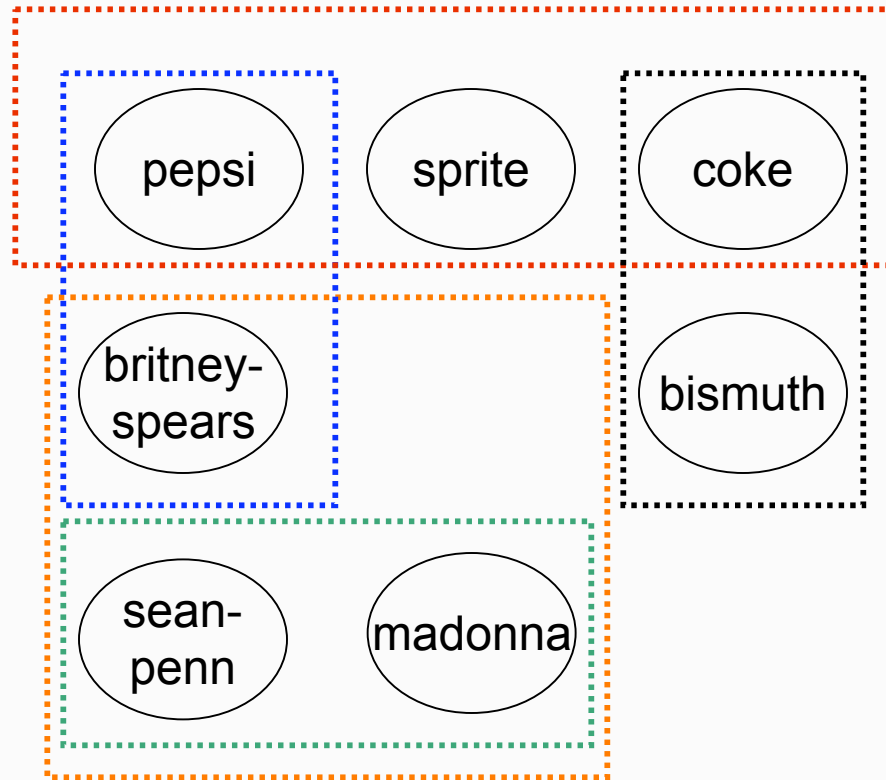
britney-  
spears

bismuth

sean-  
penn

madonna

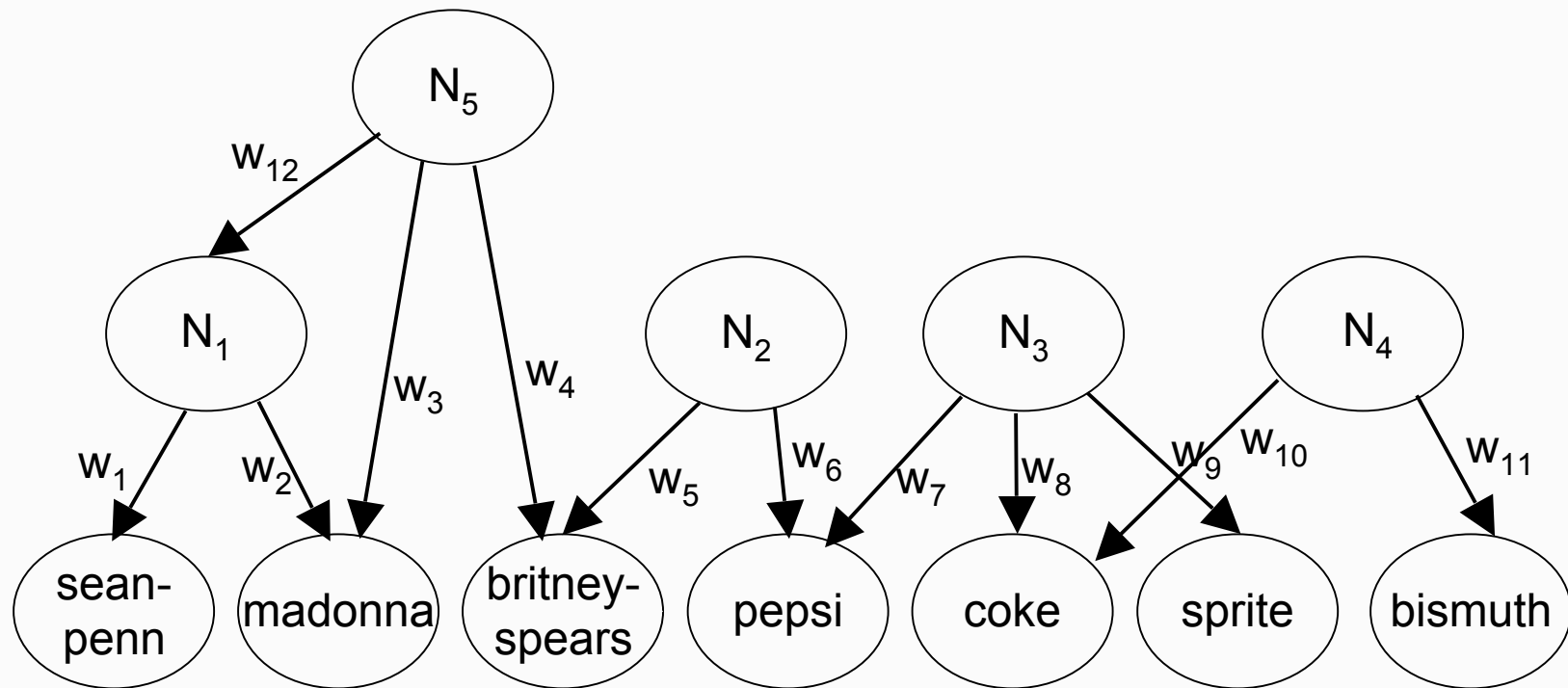
# Word Clustering



# Word Clustering

- Data: Billions of short word subsequences
- Model: Noisy-Or Bayes Net
  - Millions of terminal nodes (words)
  - 100,000s of non-terminals (topics) (Sparse.)
- Training: EM
- Inference: Greedy Deterministic Sampling
- Evaluation: ???

# Word Clustering





/search?q=visa

visa

Search

QUERY: [visa](#)

[visa](#) 15.624945/4.039471/8.826189 [[TOP](#)]

Probability Activation Cluster

0.582235	0.525590	<a href="#">[ visa for visas uk to tourist ]</a>
0.359859	0.214223	<a href="#">[ credit-card capital-one credit-cards american-express mbna visa ]</a>
0.149869	0.124134	<a href="#">[ visa card platinum first first-usa united-mileage-plus ]</a>
0.051598	0.042750	<a href="#">[ visa 三井-住友 カード 住友 visa マスター ]</a>
0.097412	0.012965	<a href="#">[ ins green-card visa us status application ]</a>
0.021811	0.012400	<a href="#">[ visa requirements embassy tourist requirement citizens ]</a>
0.006395	0.005442	<a href="#">[ visa pour formulaire france demande visas ]</a>
0.006069	0.003239	<a href="#">[ visa h1 h4 l1 h1b b1 ]</a>
0.042917	0.001117	<a href="#">[ カード jcb nicos セゾン-カード dc 日本-信販 ]</a>
1.000000	0.000050	<a href="#">TOP</a>
0.582595	0.000000	<a href="#">[ in embassy consulate usa us-embassy us ]</a>
0.040458	0.000000	<a href="#">L.ja</a>

Query matched 13 clusters

/search?q=visa+h1b

visa h1b

Search

QUERY: [visa](#) [h1b](#)

[h1b](#) 19.241566/2.164040/12.102268 [TOP]

[visa](#) 14.624947/1.298397/8.826189 [TOP]

Probability    Activation    Cluster

0.877054    1.613480    [[h1b visa h1 usa h-1b sponsor](#) ]

0.104648    0.281994    [[visa h1 h4 l1 h1b b1](#) ]

0.203830    0.240929    [[h1b h1-b visa h1 transfer h-1b](#) ]

0.018289    0.051459    [[visa h1 sponsorship h1visa h1b sponsoring](#) ]

0.061363    0.008124    [[ins green-card visa us status application](#) ]

0.023886    0.003838    [[visa us-consulate us chennai consulate us-embassy](#) ]

1.000000    0.000021    [TOP](#)

0.023871    0.000000    [[india indian in of delhi bombay](#) ]

0.868206    0.000000    [[jobs job in employment job-search careers](#) ]

0.008232    0.000000    [[yellow-pages white-pages area-codes zip-codes mapquest people-search](#) ]

0.011326    0.000000    [L.en-ca](#)

0.047433    0.000000    [L.es](#)

0.083118    0.000000    [L.de](#)

Query matched 15 clusters

/search?q=visa+card

visa card

Search

QUERY: [visa card](#)

[card](#) 5.355414/2.596364/8.062965 [[credit-card](#) [capital-one](#) [credit-cards](#) [american-express](#) [mbna](#) [visa](#)]

[visa](#) 5.221890/1.154132/8.826189 [[credit-card](#) [capital-one](#) [credit-cards](#) [american-express](#) [mbna](#) [visa](#)]

Probability Activation Cluster

0.931197 1.830076 [[visa card platinum first first-usa united-mileage-plus](#)]

0.050427 0.117189 [[card debit-card cards visa debit-cards gift](#)]

0.952259 0.042121 [[credit-card](#) [capital-one](#) [credit-cards](#) [american-express](#) [mbna](#) [visa](#)]

0.014113 0.030300 [[card visa credit-card master-card card-services nederland](#)]

0.029524 0.000450 [[payment online payments services electronic pay](#)]

1.000000 0.000338 [TOP](#)

0.008299 0.000073 [[miles frequent-flyer frequent-flyer-miles program reward rewards](#)]

0.014806 0.000000 [[abbey-national natwest lloyds-tsb hsbc barclays bank](#)]

0.013837 0.000000 [L.en-ca](#)

0.016587 0.000000 [L.en-au](#)

0.018484 0.000000 [L.it](#)

0.057745 0.000000 [L.es](#)

0.013925 0.000000 [L.en-gb](#)

0.049068 0.000000 [L.ja](#)

0.101071 0.000000 [L.de](#)

Query matched 17 clusters





**/search?q=visa+application**

visa application

Search

**QUERY:** [visa application](#)

[visa](#) 14.624947/2.741010/8.826189 [TOP]

[application](#) 13.903603/5.769062/8.515065 [TOP]

Probability    Activation    Cluster

0.802836    1.679451    [\[ visa for visas uk to tourist \]](#)

0.172430    0.513658    [\[ credit-card capital-one credit-cards american-express mbna visa \]](#)

0.178909    0.091250    [\[ ins green-card visa us status application \]](#)

0.024841    0.084685    [\[ credit-cards credit-card visa secured application canadian \]](#)

0.046922    0.036228    [\[ visa card platinum first first-usa united-mileage-plus \]](#)

0.012266    0.005756    [\[ passport us passport-application passport-renewal passports renew \]](#)

1.000000    0.001210    [TOP](#)

0.024840    0.000000    [\[ credit first peoples chexsystems banks chex-systems \]](#)

0.803012    0.000000    [\[ in embassy consulate usa us-embassy us \]](#)

0.012100    0.000000    [L.en-ca](#)

0.014410    0.000000    [L.en-au](#)

0.055389    0.000000    [L.en-gb](#)

Query matched 14 clusters