

Build Intelligence from the Physical World

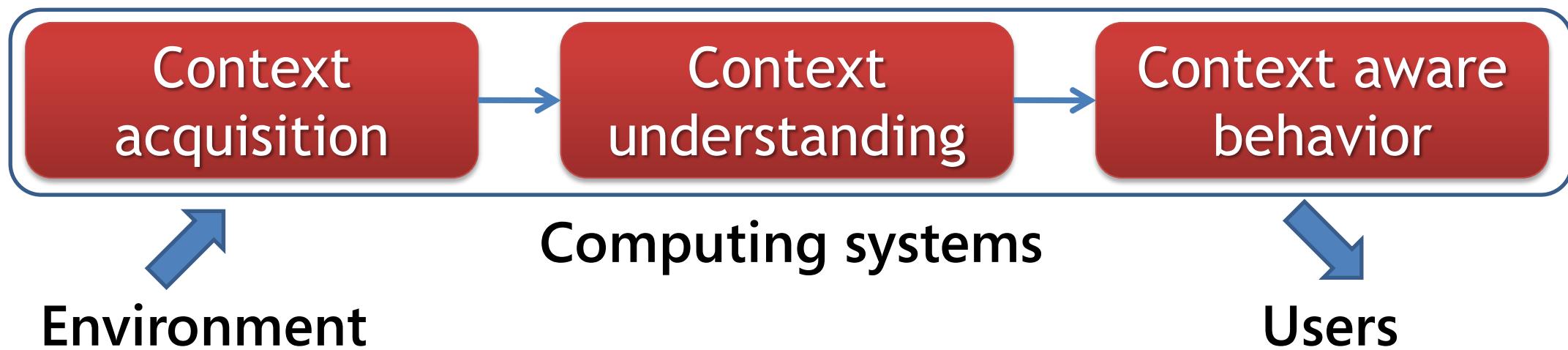
Xing Xie

Microsoft Research Asia

Aug. 30, 2011

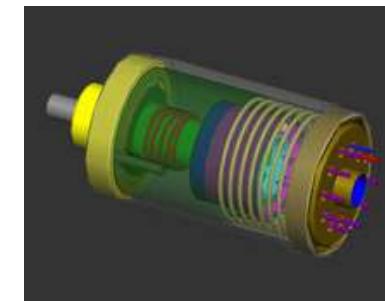
Context Awareness

- A key concept in Ubicomp: deal with linking changes in the environment (**physical world**) with computing systems
 - Acquisition of context
 - Abstraction and understanding of context
 - Application behavior based on the recognized context
- Build **intelligence about physical world** in computing systems

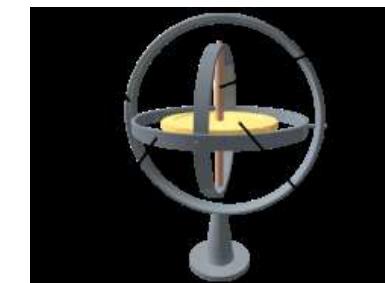


Context and Sensors

- Sensor: a device that measures a **physical** quantity and converts it into a **signal** which can be read by an observer or by an instrument (from Wiki)

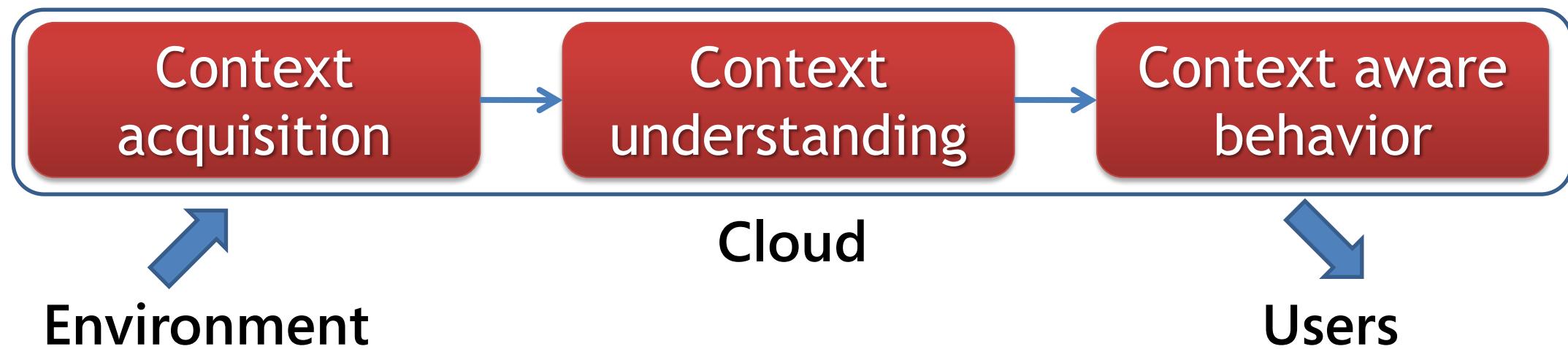


- Device time
- Device location
 - GPS, Wi-Fi, cell-tower, Bluetooth
- Device movement
 - Accelerometer, gyroscope
 - Digital compass
- Environment
 - Microphone
 - Camera, ambient light sensor
 - Proximity sensor
 - Barometer, humidity sensor, thermometer



Make the Cloud Intelligent

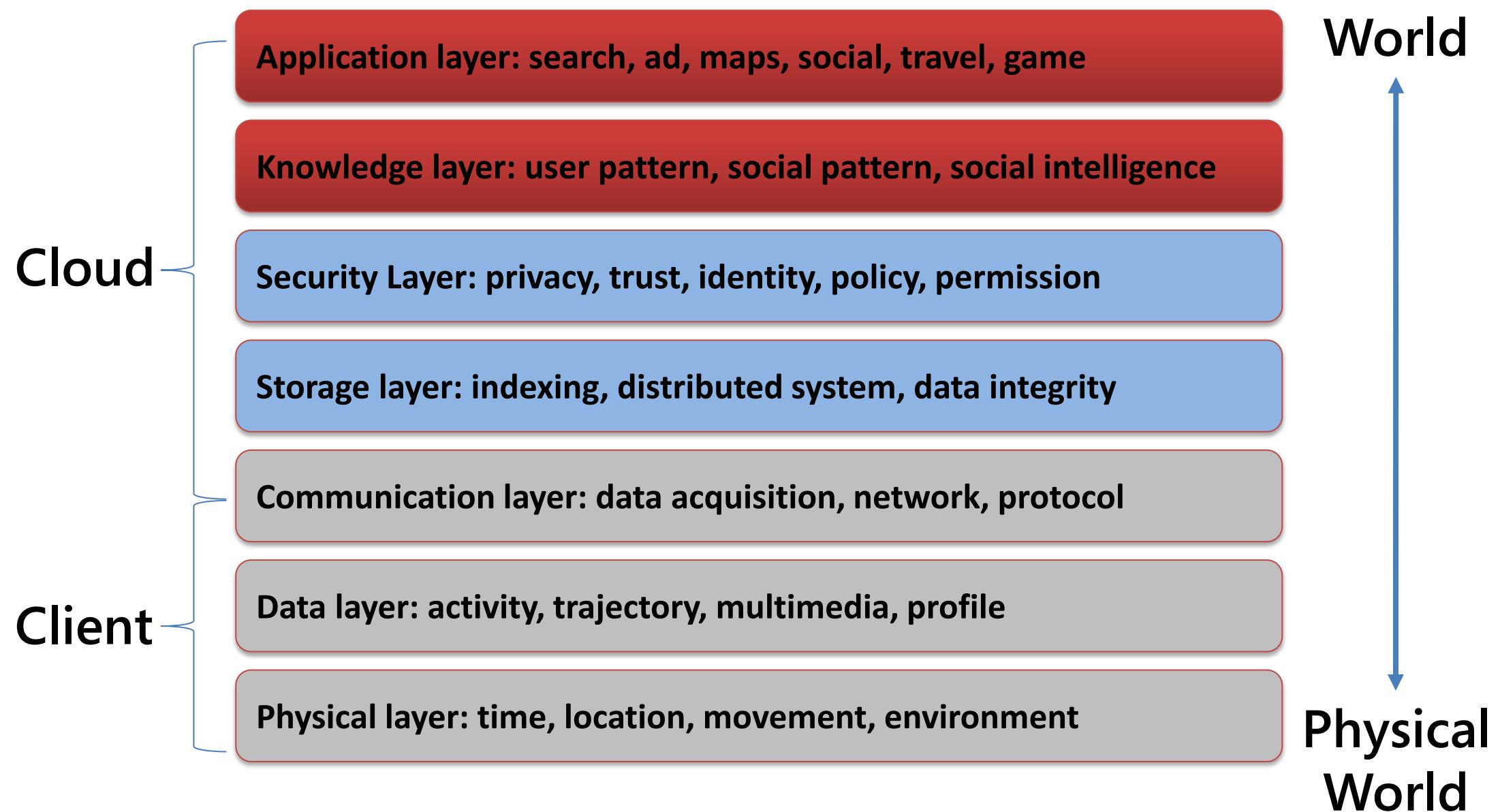
- The coming era of cloud computing brings new opportunities to this long studied research area
- By accumulating and aggregating context from multiple users, multiple devices, and over a long period, we can obtain **collective social intelligence** from them



Future Devices = Universal Sensors

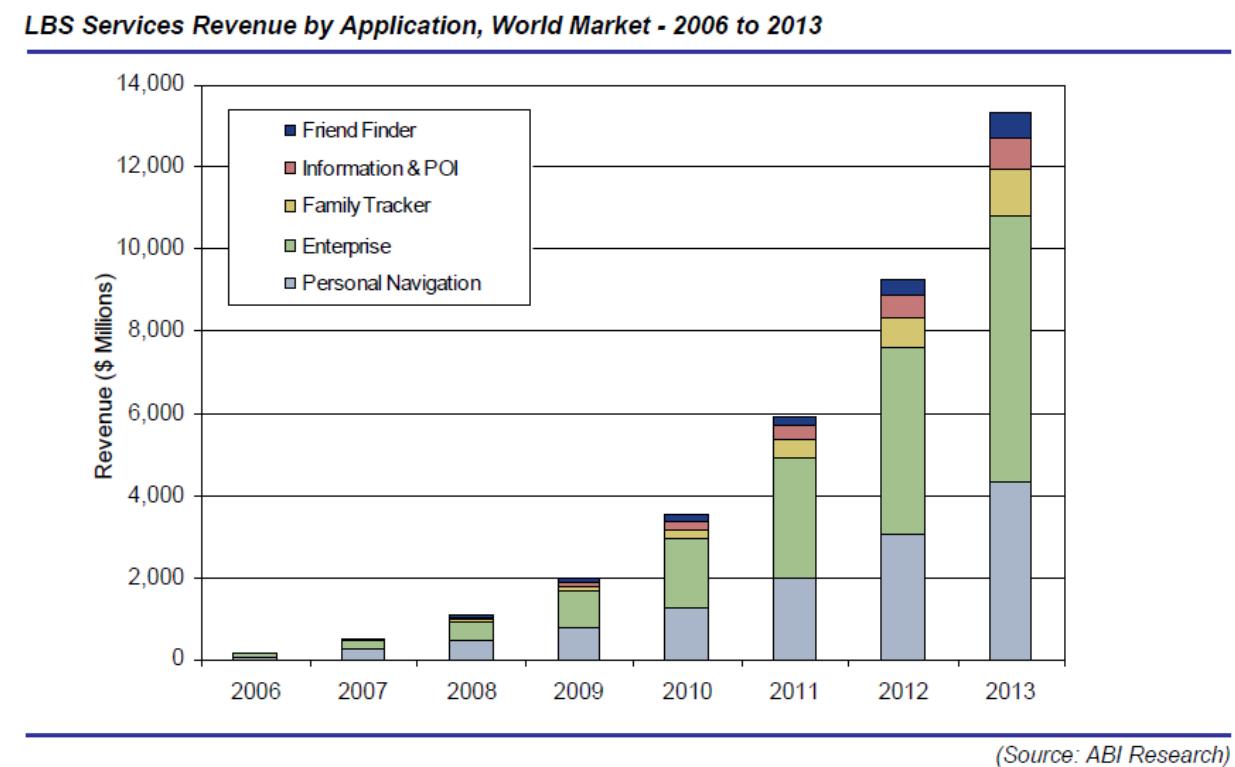
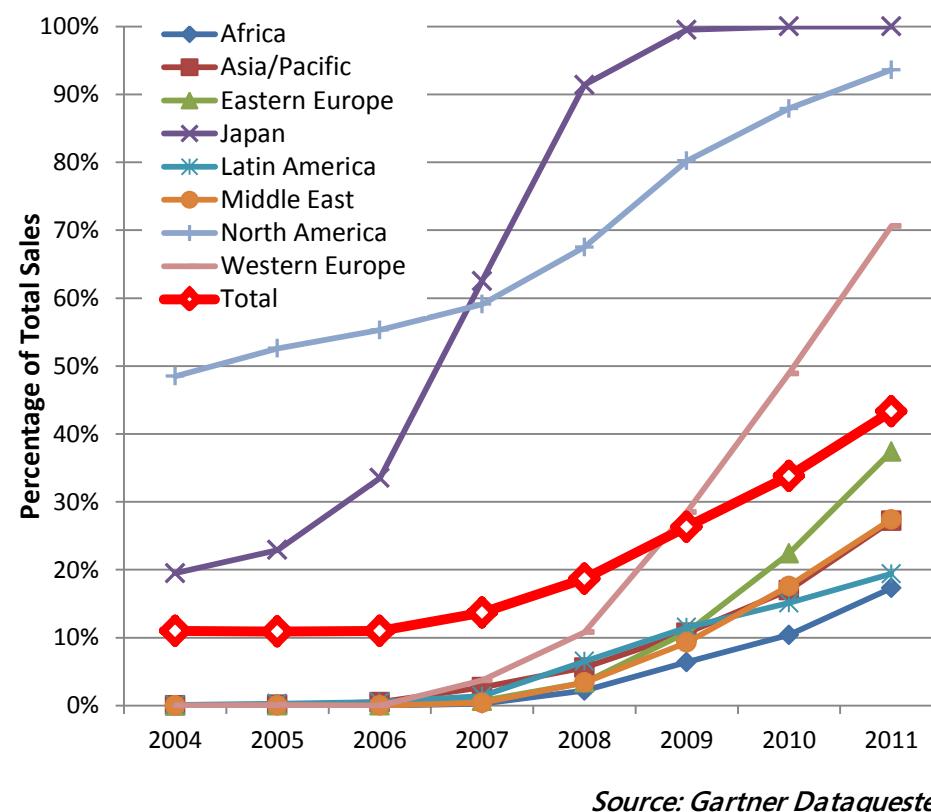


7-Layer Architecture

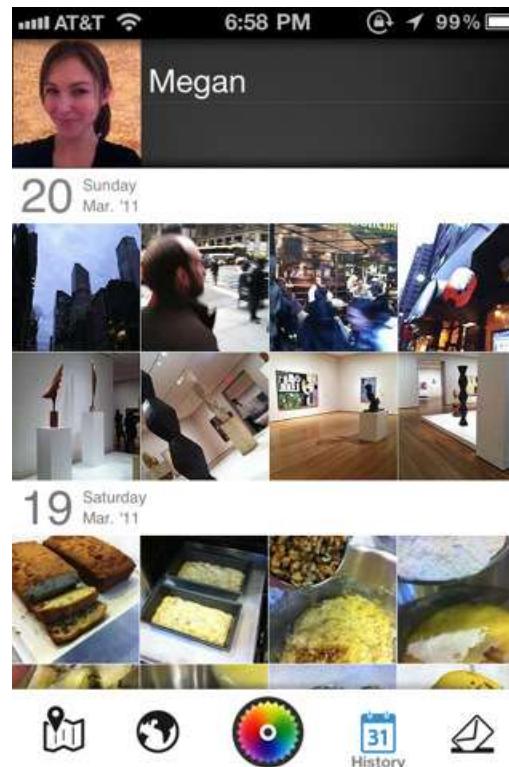


Location: the Most Important Context Data

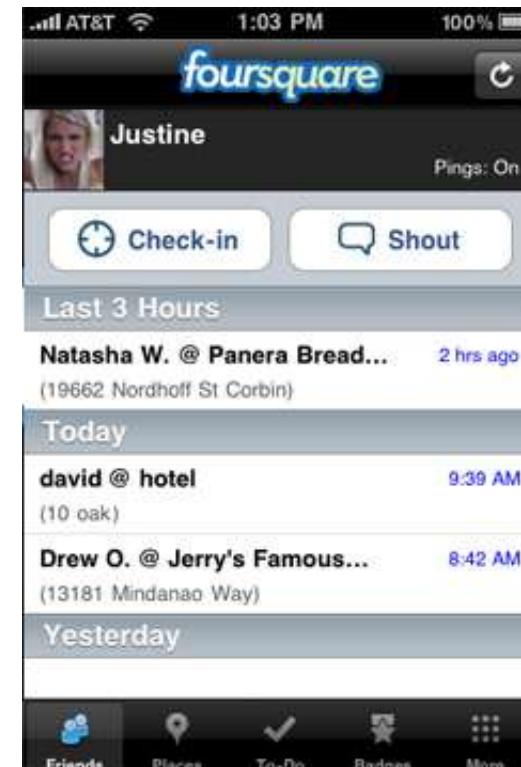
- GPS will be installed on 40+% phones by 2011 worldwide
- Location based service (LBS) will become a 13B business by 2013



Location Based Social Networks



Color



Foursquare



Bedo(贝多)

Projects in MSR Asia

- GeoLife: Building Social Networks Using Human Location History (WWW 2010/2009, AAAI 2010, SIGMOD 2010)

Knowledge from General People → Social Network Service

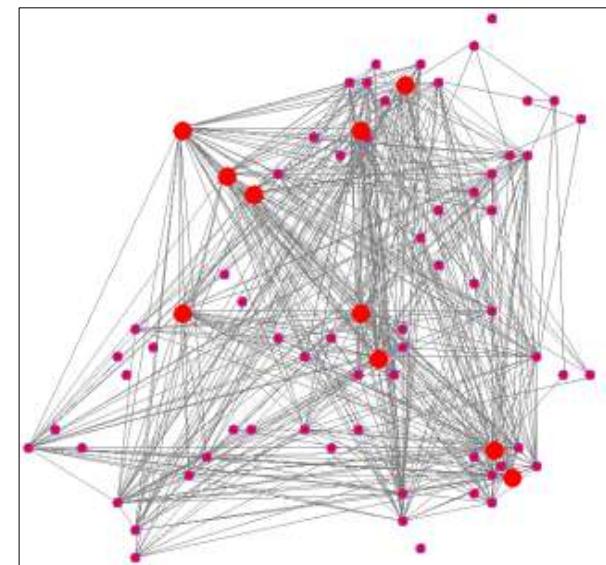
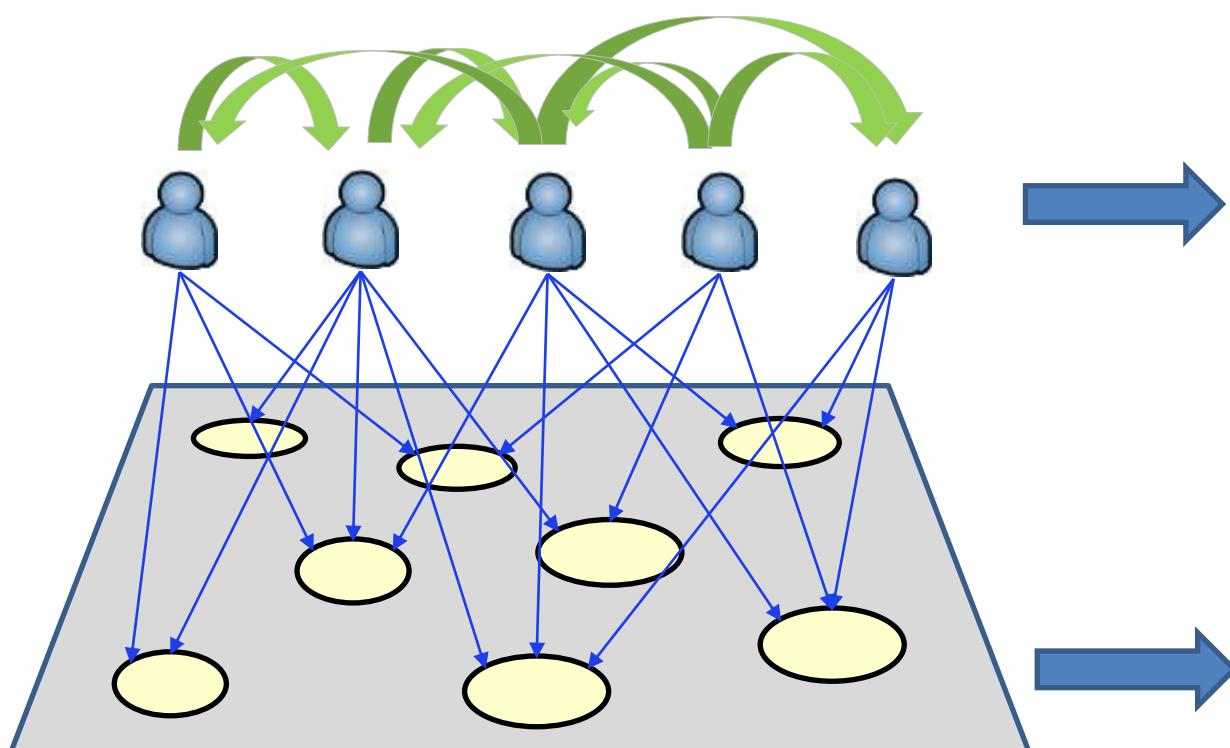
- Mining Geo-Tagged Photos for Travel Recommendation (ACM MM 2010/2009)

Knowledge from Photographers → Travel Service

- T-Drive: Driving Directions Based on Taxi Traces (ACM GIS 2010/2009)

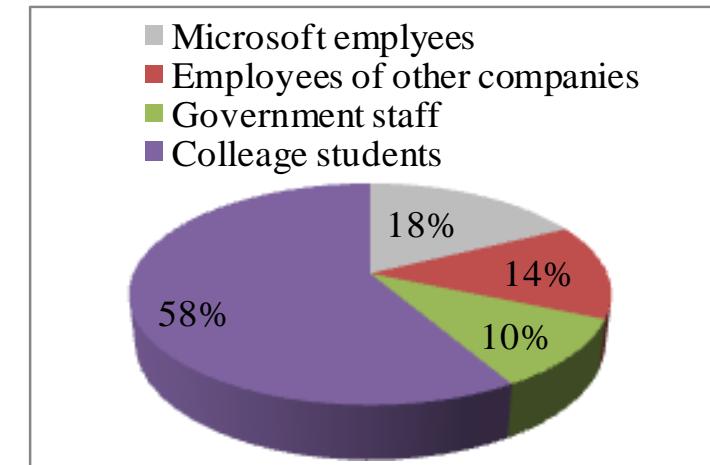
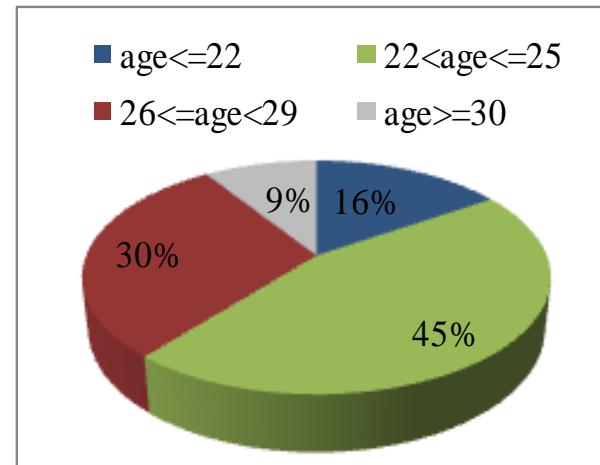
Knowledge from Taxi Drivers → Map and Navigation Service

GeoLife: Building Social Networks Using Human Location History



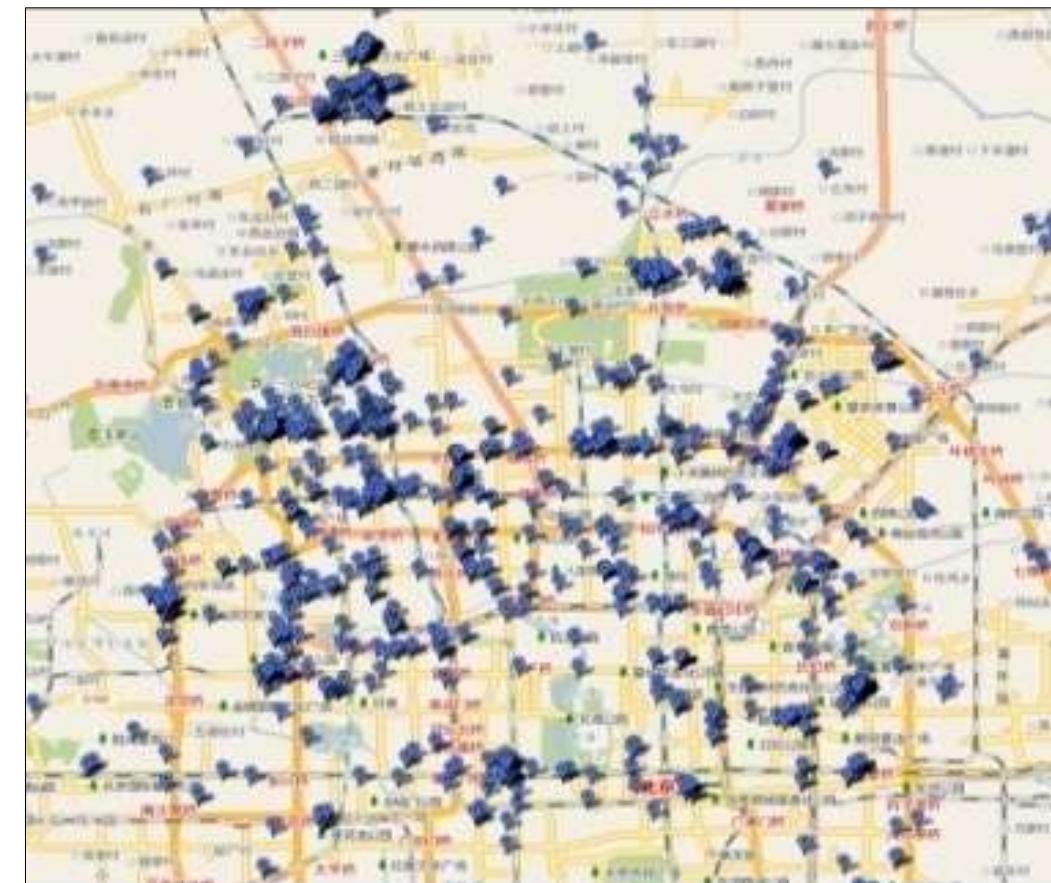
GPS Devices and Users

- 165 users, Apr. 2007 ~ Aug. 2009



A Free Large-Scale GPS Dataset

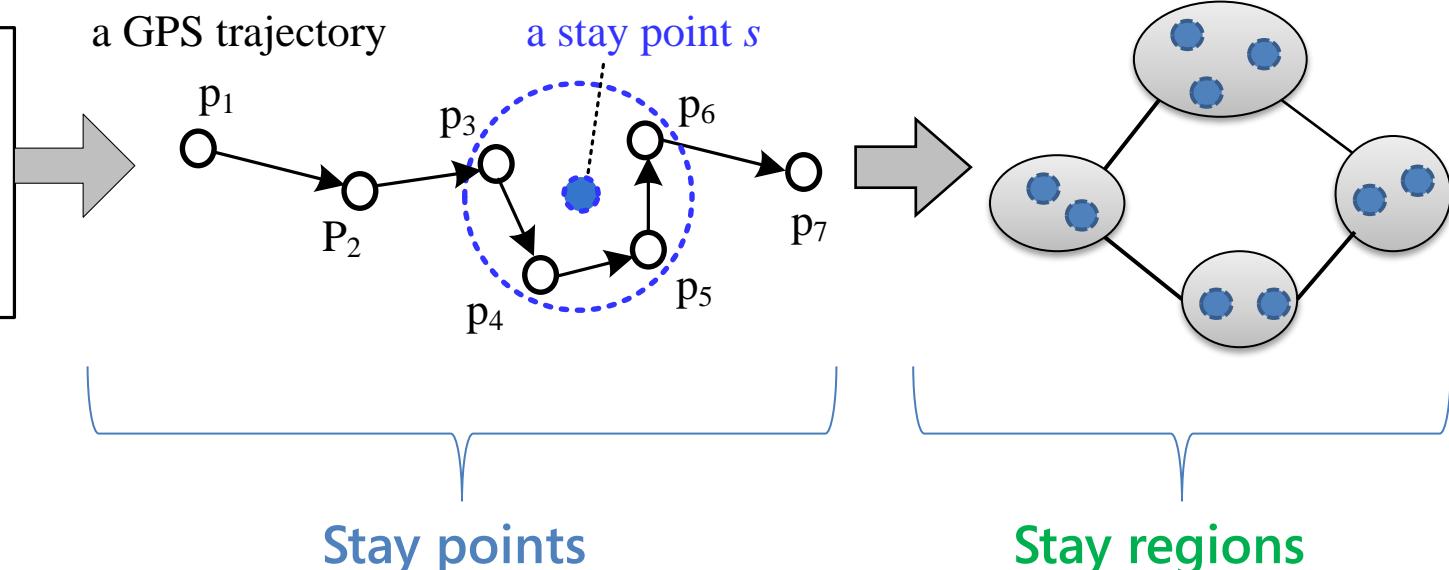
- Shared at my home page (search for “Xing Xie”)
- <http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/default.aspx>



GPS Log Processing

GPS trajectories*

Latitude, Longitude, Arrival Timestamp
 $p_1: 39.975, 116.331, 9/9/2009 17:54$
 $p_2: 39.978, 116.308, 9/9/2009 18:08$
...
 $p_K: 39.992, 116.333, 9/12/2009 13:56$



Raw GPS points

Stay points

Stay regions

- Stand for a geo-spot where a user has stayed for a while
- Preserve the sequence and vicinity info

- Stand for a geo-region that we may recommend
- Discover the meaningful locations

* In GPS logs, we have some user comments associated with the trajectories.

Collaborative Activity and Location Recommendation

- Location Recommendation
 - Question: *I want to find nice food, where should I go?*
- Activity Recommendation
 - Question: *I will visit the downtown, what can I do there?*



Data Modeling

User <-> Location <-> Activity



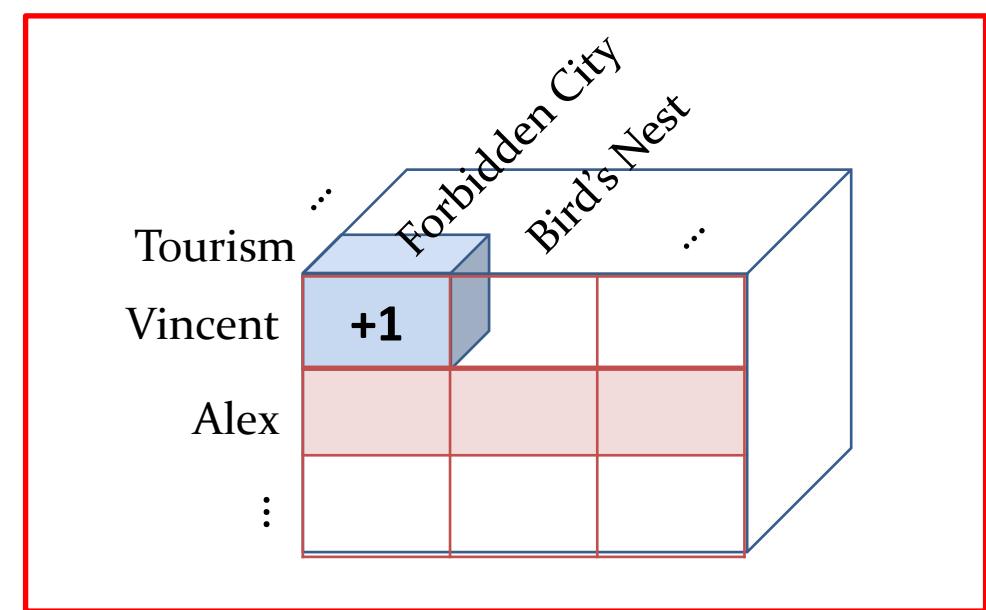
GPS: "39.903, 116.391, 14/9/2009 15:25"

Stay Region: "39.910, 116.400 (Forbidden City)"

"User Vincent: We took a tour bus to see around along the forbidden city moat ..."

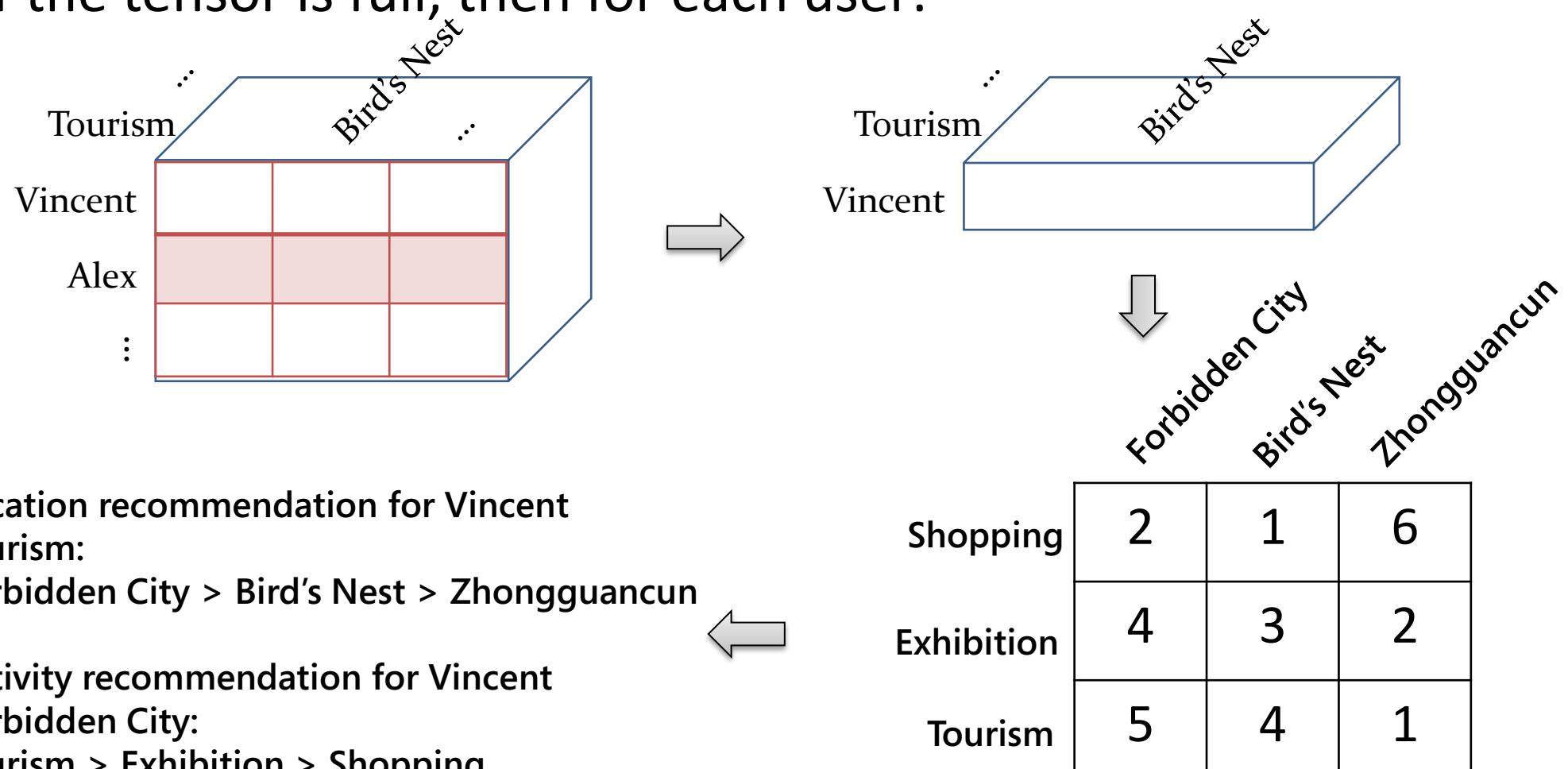
Activity: tourism

Activities	Descriptions
Food and Drink	Dinning/drinking at restaurants/bars, etc.
Shopping	Supermarkets, department stores, etc.
Movie and Shows	Movie/shows in theaters and exhibition in museums, etc.
Sports and Exercise	Doing exercises at stadiums, parks, etc.
Tourism and Amusement	Tourism, amusement park, etc.



How to Do Recommendation?

- If the tensor is full, then for each user:



Location recommendation for Vincent

Tourism:

Forbidden City > Bird's Nest > Zhongguancun

Activity recommendation for Vincent

Forbidden City:

Tourism > Exhibition > Shopping

Shopping

Exhibition

Tourism

2	1	6
4	3	2
5	4	1

Unfortunately, in practice, the tensor is usually sparse!

Our First Solution (WWW 2010)

Tourism Exhibition Shopping

	Tourism	Exhibition	Shopping
Forbidden City	5	?	?
Bird's Nest	?	1	?
Zhongguancun	1	?	6



User not explicitly modeled!

1. Not modeling each single user's Loc-Act history
2. = a sum compression of our tensor

Features

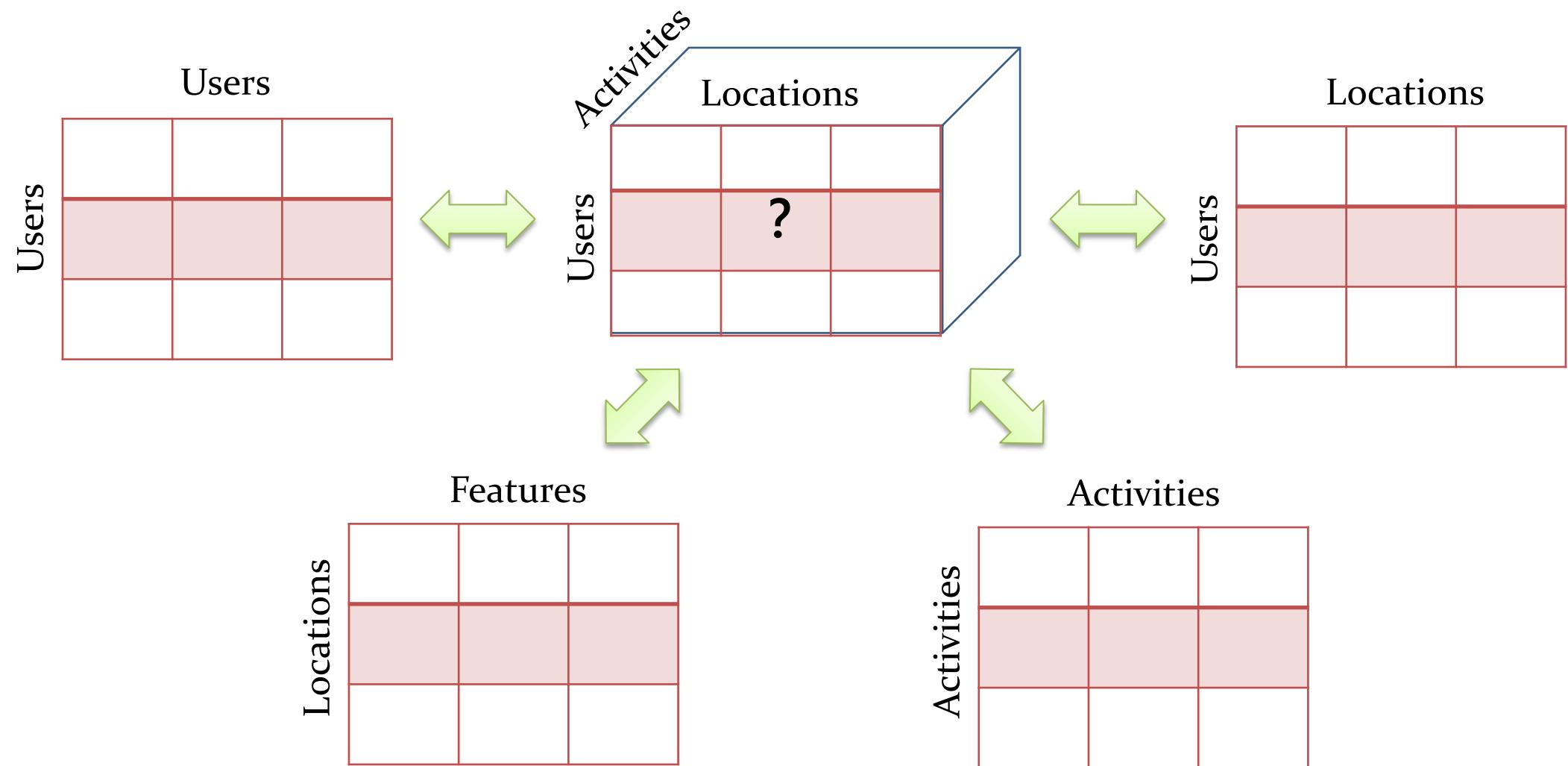


Activities

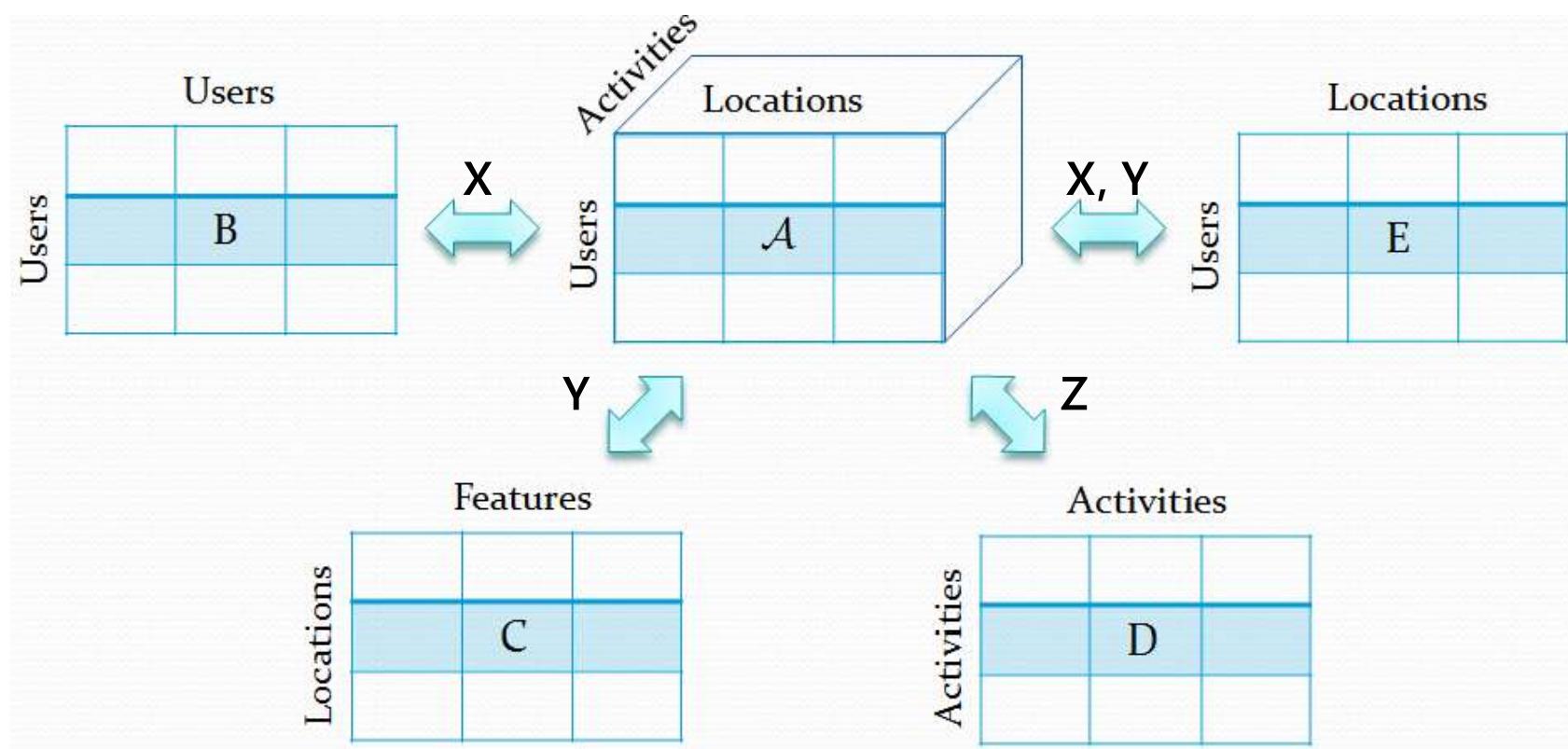
		?

Our Second Solution

- Regularized Tensor and Matrix Decomposition



Our Model



$$\begin{aligned}
 \mathcal{L}(X, Y, Z, U) = & \frac{1}{2} \| \mathcal{A} - [X, Y, Z] \|_F^2 \\
 & + \frac{\lambda_1}{2} \text{tr}(X^T L_B X) + \frac{\lambda_2}{2} \| C - YU^T \|_F^2 + \frac{\lambda_3}{2} \text{tr}(Z^T L_D Z) + \frac{\lambda_4}{2} \| E - XY^T \|_F^2 \\
 & + \frac{\lambda_5}{2} (\| X \|^2 + \| Y \|^2 + \| Z \|^2 + \| U \|^2)
 \end{aligned}$$

Location Feature Extraction

- Location features: Points of Interests (POIs)



Stay Region: "39.980, 116.306 (Zhongguancun)"

[restaurant, bank, shop] = [3, 1, 1]

TF-IDF style normalization*: feature = [0.13, 0.32, 0.18]



	restaurant	bank	...
Forbidden City			
Zhongguancun	0.13	0.32	
:			
Location-Feature Matrix			

TF-IDF (Term-Frequency Inverse Document Frequency):

$$tf-idf_{i,t} = \frac{n_{i,t}}{\sum_l n_{i,l}} \cdot \log \frac{|\{d_i\}|}{|\{d_i : t \in d_i\}|}$$

Example:

Assume in 10 locations, 8 have restaurants (less distinguishing), while 2 have banks and 4 have shops:

$$tf-idf(\text{restaurant}) = (3/5) * \log(10/8) = 0.13$$

$$tf-idf(\text{bank}) = (1/5) * \log(10/2) = 0.32$$

$$tf-idf(\text{shop}) = (1/5) * \log(10/4) = 0.18$$

Activity Correlation Extraction

- How possible for one activity to happen, if another activity happens?
 - Automatically mined from the Web, potentially useful when #(act) is large

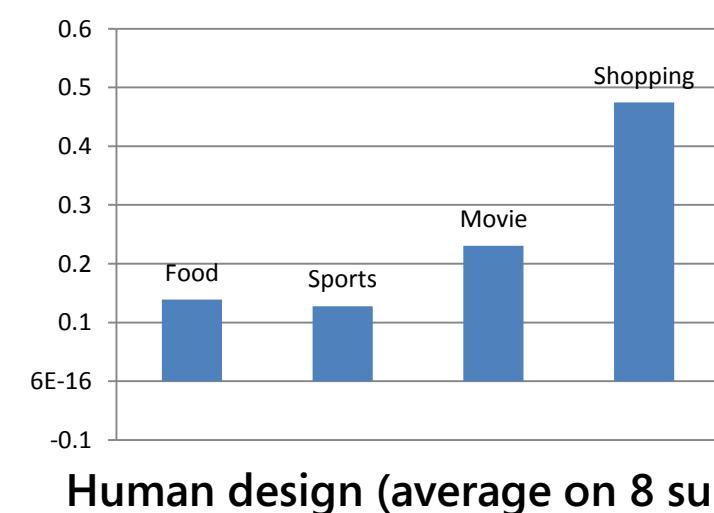
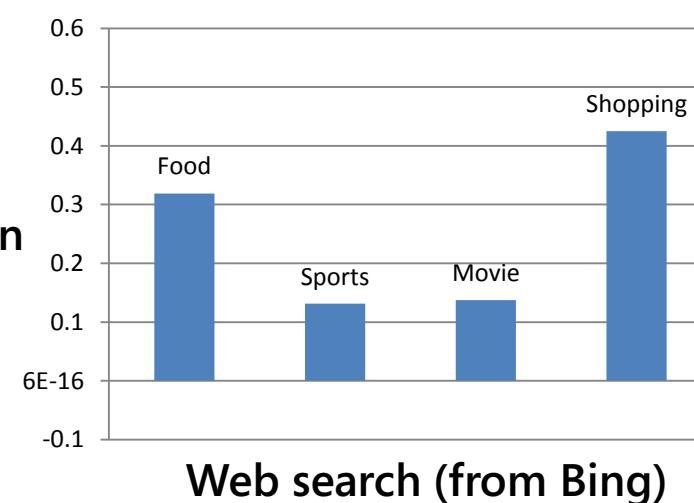
A screenshot of a Bing search results page. The search bar at the top contains the query "Tourism and Amusement, Food and Drink". Below the search bar, there are two sections: "ALL RESULTS" and "RELATED SEARCHES". The "ALL RESULTS" section shows a snippet for "Amusement Park | ScreenScape Networks" with the text: "Food & Drink; General Purpose; Healthcare; Hospitality ... Rushing Waters Am Park. Rushing Waters Amusement Park ... Tourism operators are benefiting fr... ScreenScape ... screenscape.net/park - Cached page". To the right of the search bar, a red dashed oval highlights the text "1,100 of 1,160,000 results".

"Tourism and Amusement"
and
"Food and Drink"

Correlation = $h(1.16M)$,
where h is a normalization func.

Most mined correlations are reasonable. Example: "Tourism" with other activities.

Tourism-Shopping
more likely to happen
together than
Tourism-Sports



Optimization

- Minimize the object function $L(X, Y, Z, U)$

- Gradient descent

$$X_{t+1} = X_t - \gamma \nabla_X, Y_{t+1} = Y_t - \gamma \nabla_Y, Z_{t+1} = Z_t - \gamma \nabla_Z, U_{t+1} = U_t - \gamma \nabla_U$$

where $\nabla_X \mathcal{L} = -A^{(1)}(Z * Y) + X [(Z^T Z) \odot (Y^T Y)] + \lambda_1 L_B X + \lambda_4 (XY^T - E)Y + \lambda_5 X,$

$$\nabla_Y \mathcal{L} = -A^{(2)}(Z * X) + Y [(Z^T Z) \odot (X^T X)] + \lambda_2 (YU^T - C)U + \lambda_4 (XY^T - E)^T X + \lambda_5 Y,$$

$$\nabla_Z \mathcal{L} = -A^{(3)}(Y * X) + Z [(Y^T Y) \odot (X^T X)] + \lambda_3 L_D Z + \lambda_5 Z,$$

$$\nabla_U \mathcal{L} = \lambda_2 (YU^T - C)^T Y + \lambda_5 U,$$

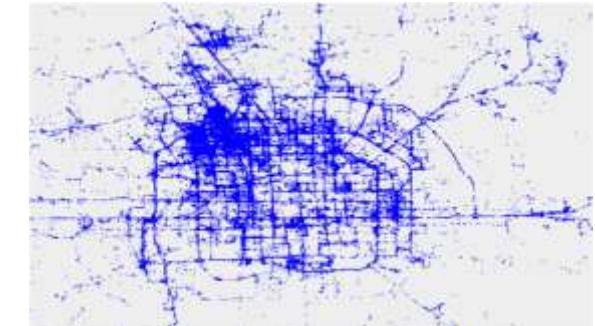
- Complexity: $O(T \times (mnr + m^2 + r^2))$

- T is #(iteration), m is #(user), n is #(location), r is #(activity)

Experiments

• Data

- GeoLife data set
- 13K GPS trajectories, 140K km long
- 530 comments
- After clustering, #(loc) = 168; #(user) = 164, #(act) = 5, #(locfea) = 14
- The user-loc-act tensor has 1.04% of the entries with values

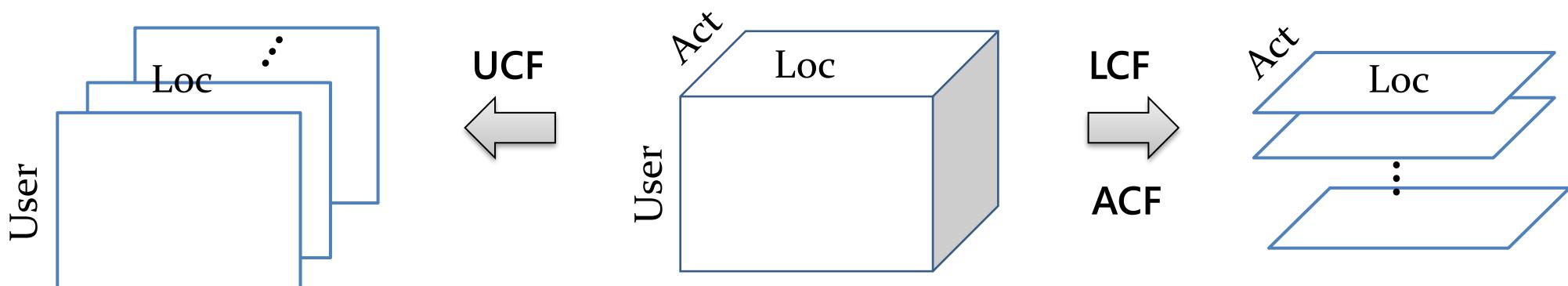


• Evaluation

- Ranking over the hold-out test dataset
- Metrics:
 - Root Mean Square Error (RMSE)
 - Normalized discounted cumulative gain ($nDCG$)

Baselines – Category I

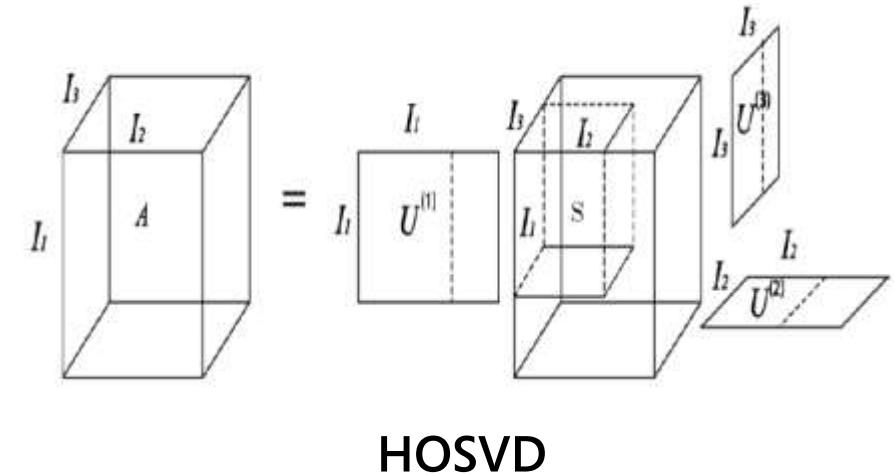
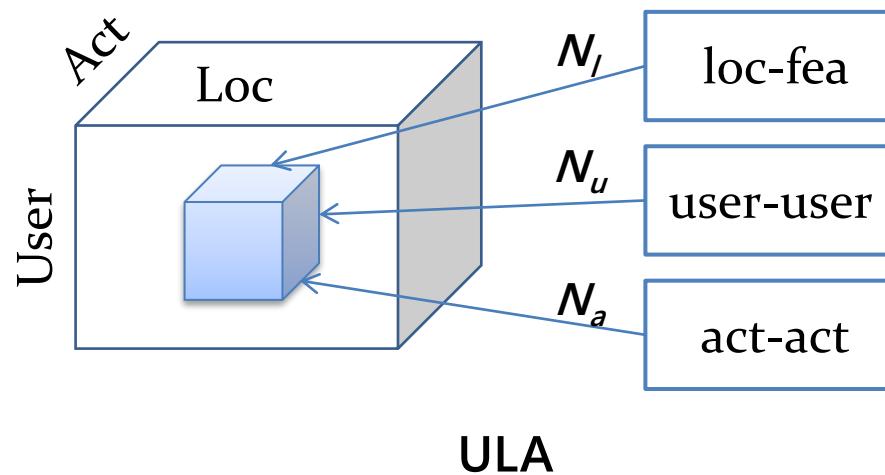
- Tensor \rightarrow Independent matrices [Herlocker et al. 1999]
 - Baseline 1: UCF (user-based CF)
 - CF on each user-loc matrix + Top N similar users for weighted average
 - Baseline 2: LCF (location-based CF)
 - CF on each loc-act matrix + Top N similar locations for weighted average
 - Baseline 3: ACF (activity-based CF)
 - CF on each loc-act matrix + Top N similar activities for weighted average



Baselines – Category II

Tensor-based CF

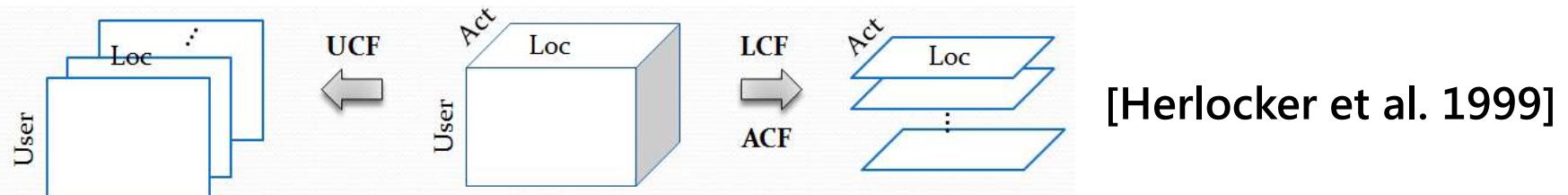
- Baseline 4: ULA (unifying user-loc-act CF) [Wang et al. 2006]
 - Top N_u similar users, top N_l similar loc's, top N_a similar act's
 - Similarities from additional matrices + Small cube for weight average
- Baseline 5: HOSVD (high order SVD) [Symeonidis et al. 2008]
 - Singular value decomposition with matrix unfolding



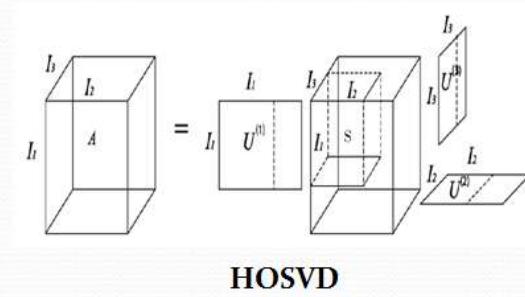
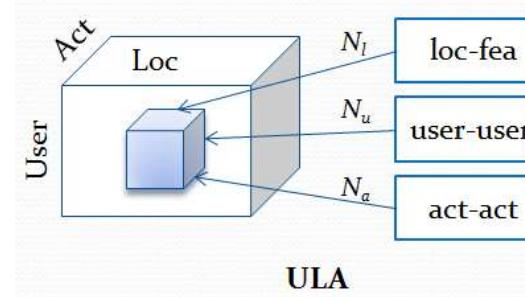
Comparison with Baselines

- Reported in “mean \pm std”

	RMSE	$nDCG_{loc}$	$nDCG_{act}$
UCF	0.027 ± 0.006	0.297 ± 0.024	0.807 ± 0.007
LCF	0.009 ± 0.000	0.532 ± 0.021	0.614 ± 0.019
ACF	0.022 ± 0.005	0.408 ± 0.012	0.785 ± 0.006
ULA	0.015 ± 0.003	0.291 ± 0.022	0.799 ± 0.012
HOSVD	0.006 ± 0.001	0.390 ± 0.021	0.913 ± 0.004
UCLAF	0.006 ± 0.001	0.599 ± 0.036	0.959 ± 0.009



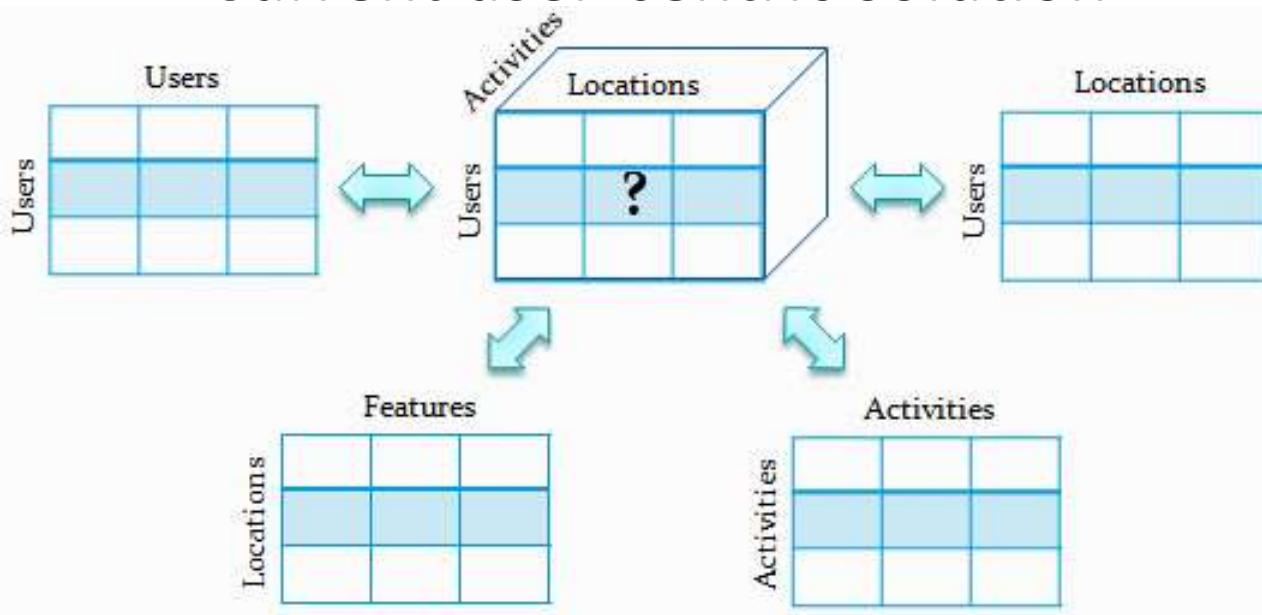
[Wang et al. 2006]



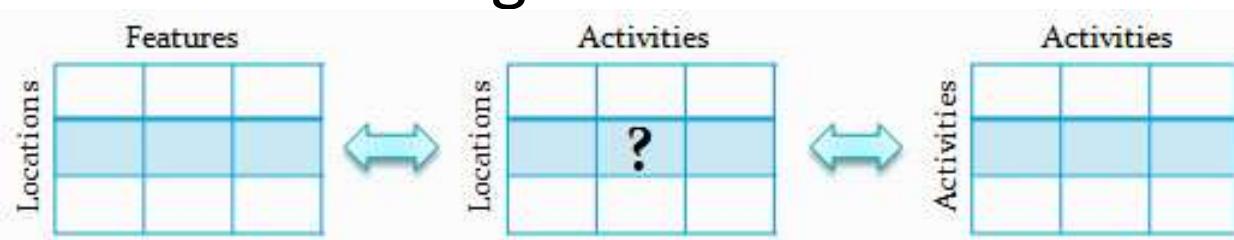
[Symeonidis et al. 2008]

Comparison with Our First Solution

- Current user-centric solution



- Previous generic solution



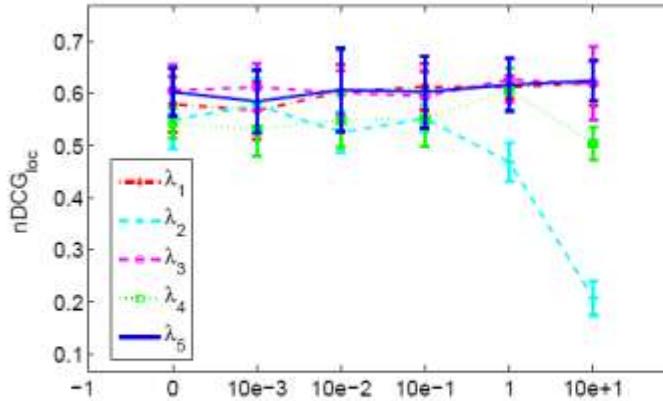
Performance

	Current Solution	Previous Solution
RMSE	0.006 ±0.001	0.041 ±0.006
nDCG _{loc}	0.576 ±0.043	0.552 ±0.027
nDCG _{act}	0.931 ±0.009	0.885 ±0.019

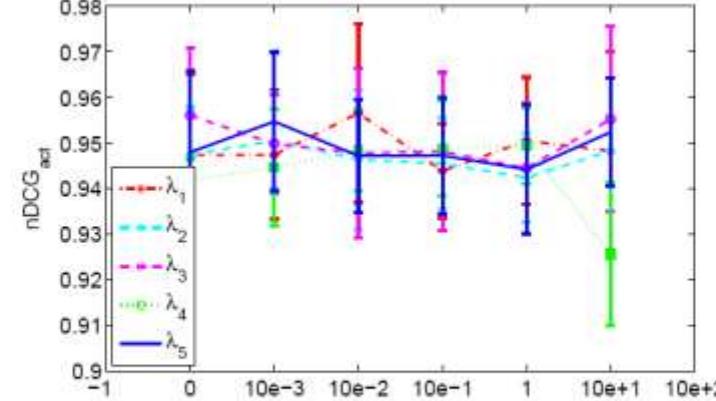
Impacts of the Model Parameters

- Some observations

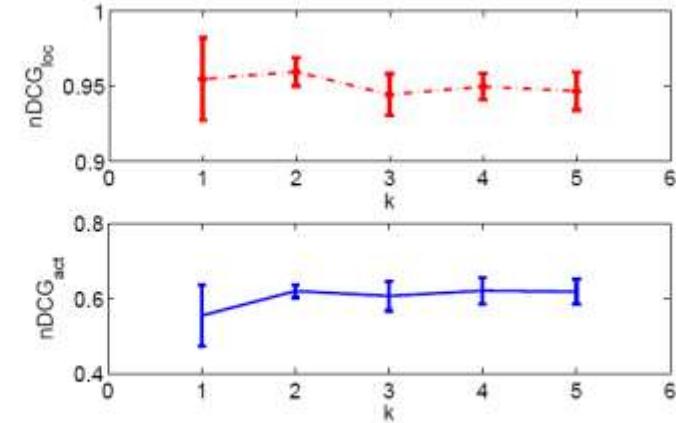
- Using additional info (i.e. $\lambda_i > 0$) is better than not (i.e. $\lambda_i = 0$)
- Not very sensitive to most parameters
 - Model is robust + Contribution from additional info is limited
- As λ_2 increases, nDCG for loc recommendation greatly decreases
 - Maybe because the loc-feature matrix is noisy in extracting the POIs
 - Not directly related to act, so no similar observation for act recommendation



(a) Impact of λ_i 's to location recommend.



(b) Impact of λ_i 's to activity recommend.



(c) Impact of the low dimension k .

Collaborative Activity and Location Recommendation

- We showed how to mine knowledge from GPS data to answer
 - If I want to do something, where should I go?
 - If I will visit some place, what can I do there?
- We evaluated our system on a large GPS dataset
 - 19% improvement on location recommendation
 - 22% improvement on activity recommendation over the simple memory-based CF baseline (i.e. UCF, LCF, ACF)

Mining City Landmarks from Photos

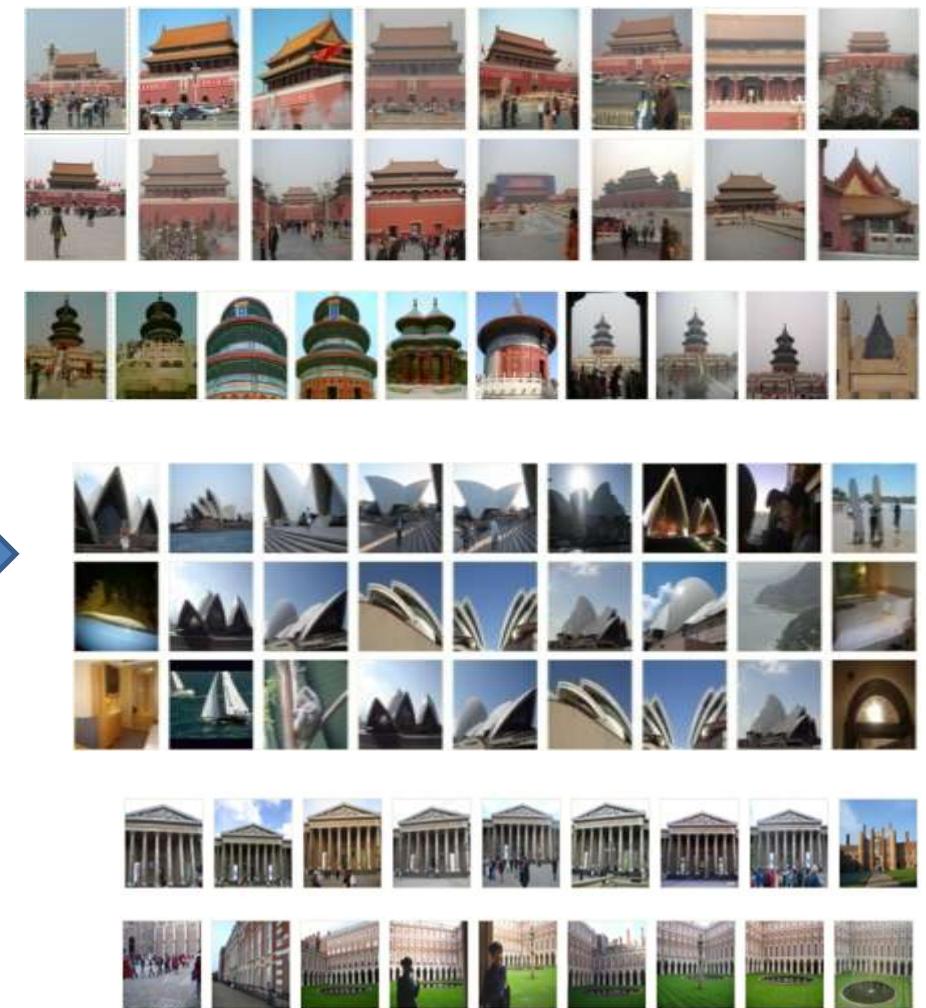
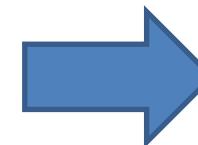
July 28
Shopping at Wangfujing Street, with a Digression on Urban Renewal
The Friday after a round of talks at PECO including both giving brief talks on increasing adoption of the third metric, a small group of us hopped in a car and headed out to experience a little more of Beijing. We'd thought we'd figured out what Beijing felt like: modern-garish apartment blocks and shopping plazas, overhanging party temples. Wrong! Preparation: BAA roads lined with big steps. Perhaps this is the reason the end of the city we've seen so far is in the suburbs—not because the building movement shopping areas, which were entirely different.

Beijing, June: technological. The fastest road train used by an over-worked around.

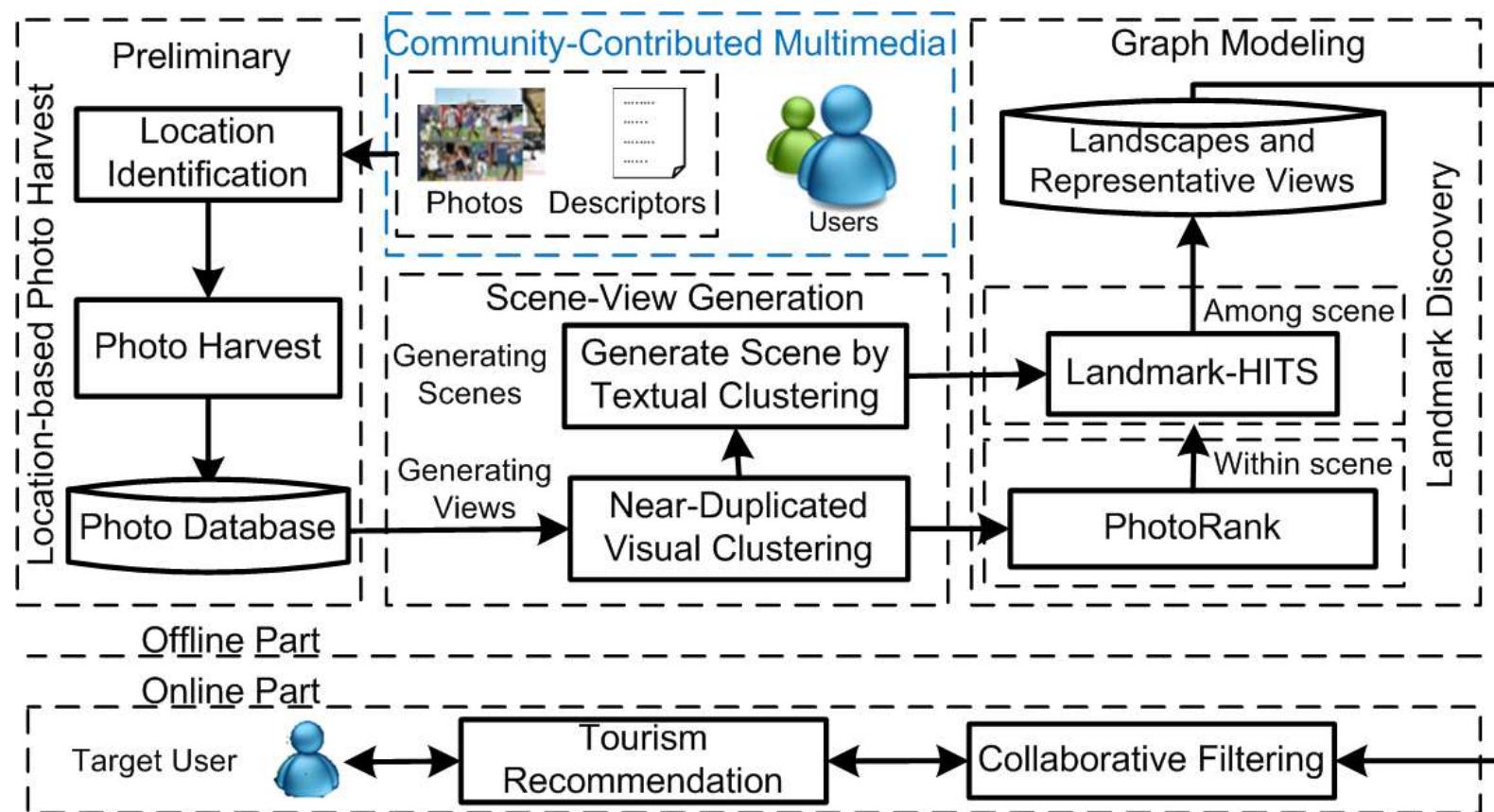
Next stop was inside the local bookshop, which was selling Harry Potter (not yet translated, apparently, but well-preserved), stories of Mao, Marx & Lenin, and copies of "100 Best's Ten Best Books for Young People", about which I know nothing except the title, which, needless to say, I love.

April 13, 2008

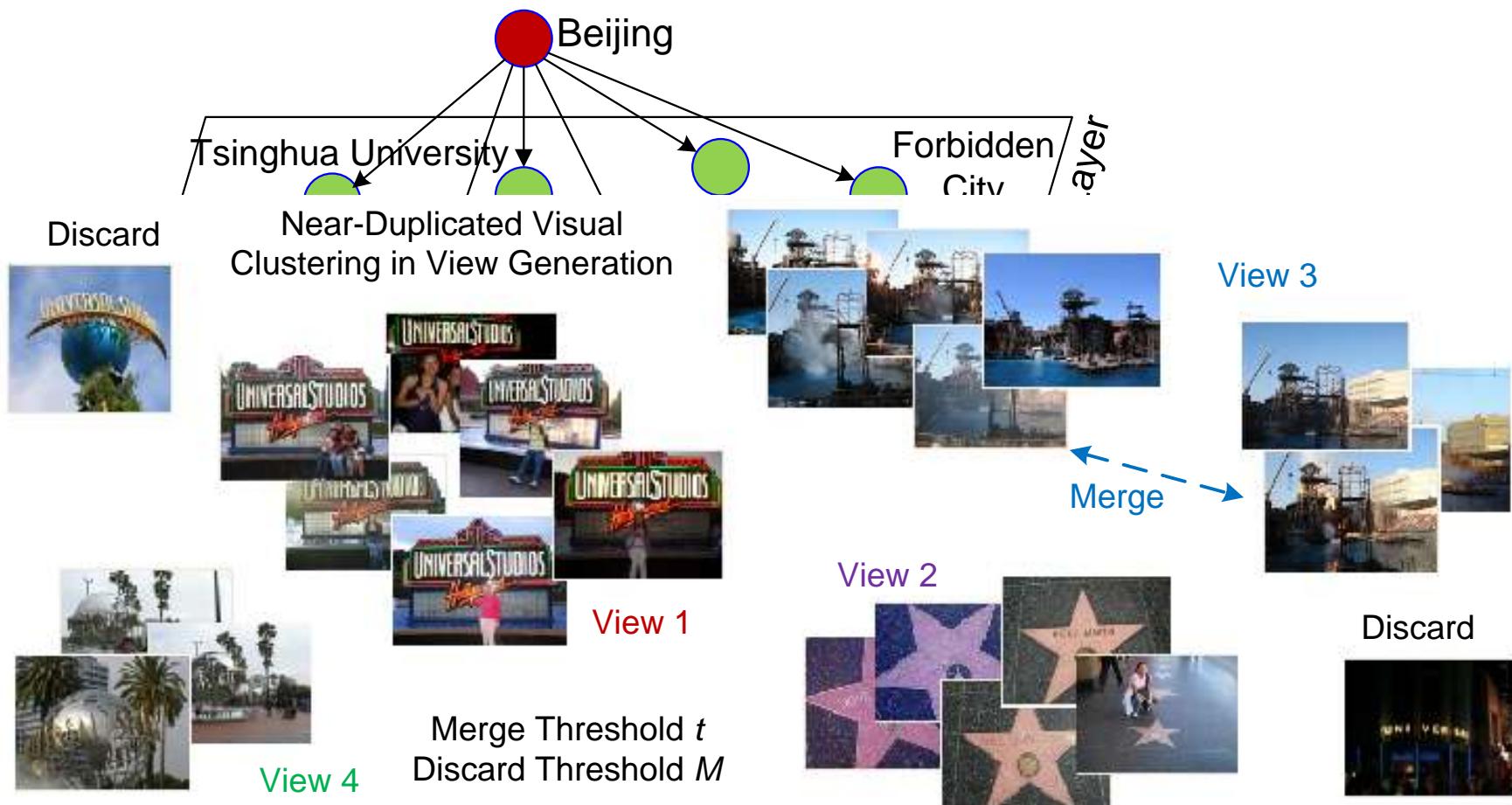
Originally known as the "Garden of Clear Ripples", the Summer Palace is a magnificent imperial gardens located 15 km west of the old Beijing City. It is the largest ancient gardens well-preserved in China and a famous summer resort for Emperors and the Chinese emperors. The site includes more than 150 ancient style pavilions, mansions, towers, hills, temples, bridges and an enormous clear water lake.



System Framework

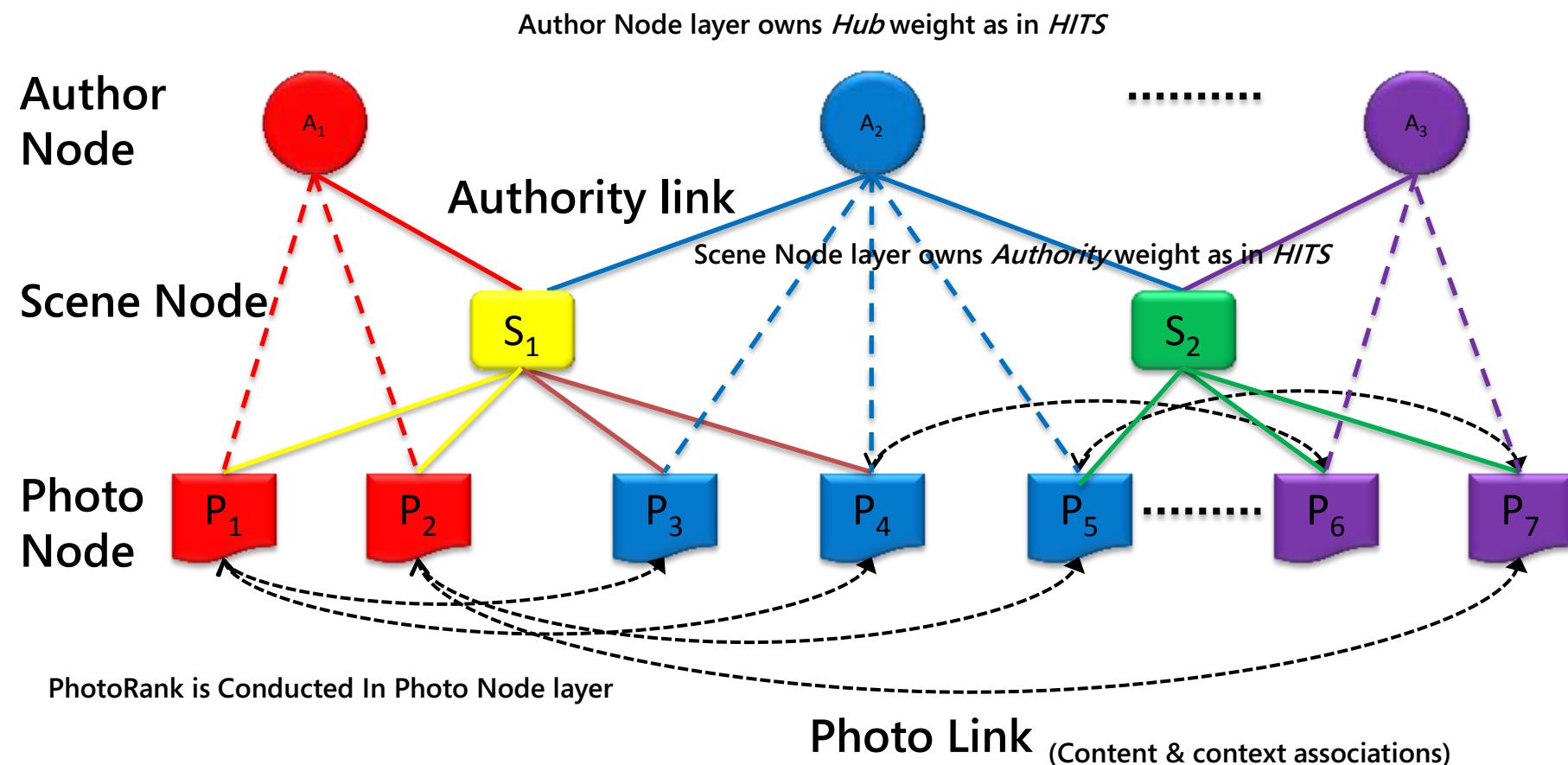


View Generation by Visual Clustering

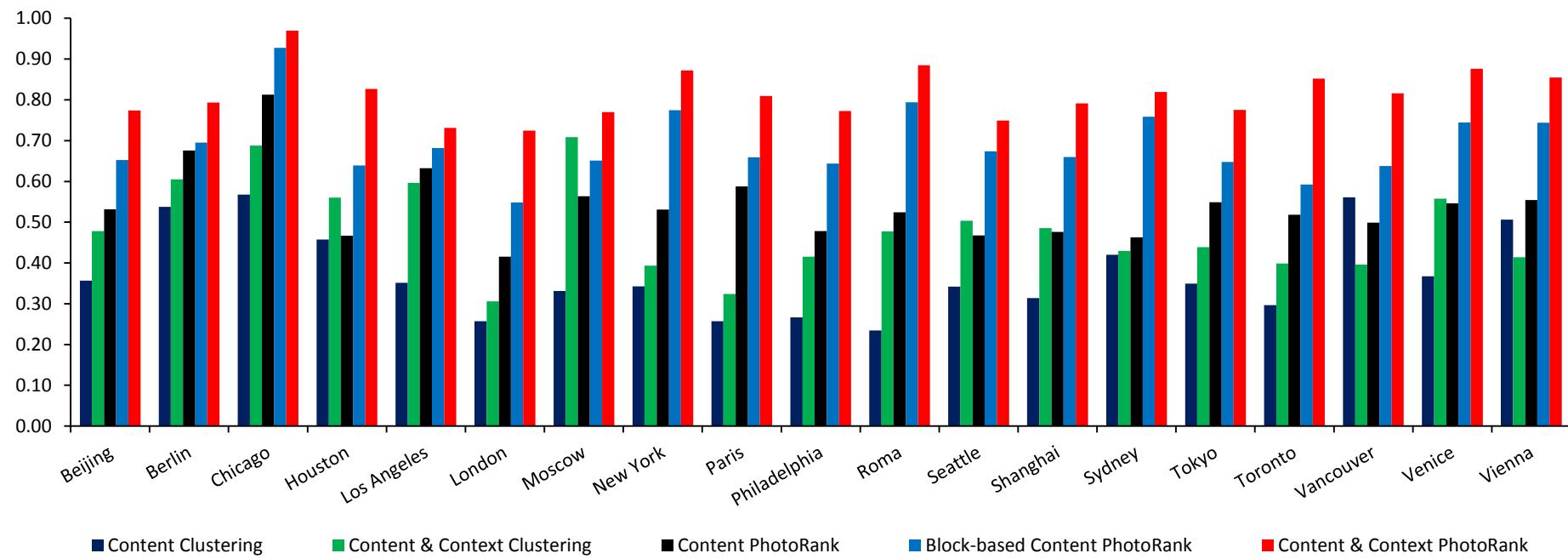


Mining Landmarks by Graph Modeling

HITS-like process is conducted in Author and Scene layers, and affects the PhotoRank iteratively



Experimental Results



Blog Users	Top Ranked Landmarks at Worldwide Scale by Landmark-HITS
Asian	1. Summer Palace (Beijing), 2. Sydney Opera House (Sydney), 3. Louvre Museum (Paris), 4. Tiananmen (Beijing), 5. Tokyo Tower (Tokyo), 6. Universal studios (L.A.), 7. Oriental Pearl (Shanghai), 8. Tower of London (London), 9. Empire State Building (New York), 10. Statue of Liberty (New York)
European	1. Sydney Opera House (Sydney), 2. Louvre Museum (Paris), 3. London Museum (London), 4. Summer Palace (Beijing), 5. Tower of London (London), 6. Empire State Building (New York), 7. Statue of Liberty (New York), 8. Oriental Pearl (Shanghai), 9. Tokyo Tower (Tokyo), 10. Universal studios (L.A.)
American	1. Statue of Liberty (New York), 2. Universal studios (L.A.), 3. Sydney Opera House (Sydney), 4. Empire State Building (New York), 5. Louvre Museum (Paris), 6. Space Needle (Seattle), 7. Summer Palace (Beijing), 8. Cn Tower (Toronto), 9. Tokyo Tower (Tokyo), 10. Oriental Pearl (Shanghai)

Mining Trip Knowledge from Geo-tagged Photos

- Trace people's trips from geo-tagged photo collections
- Photo trip patterns:
 - Sequence of visited cities and durations of stay
 - Typical description of trips represented by tags
- Classify photo trip patterns based on their trip themes

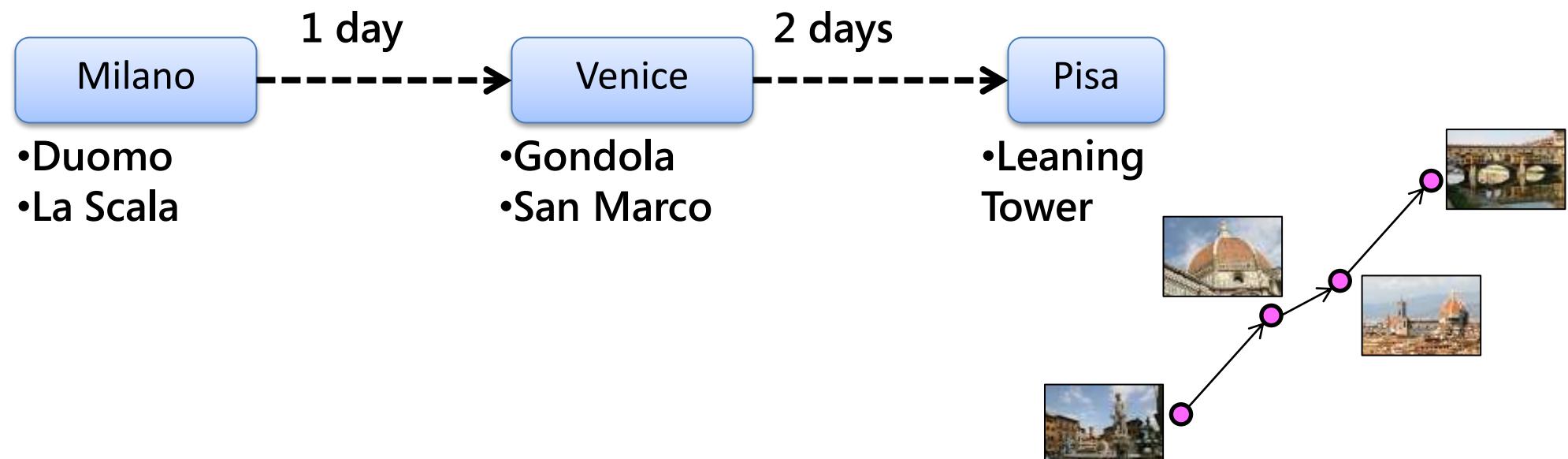


Photo Trip Pattern Mining: Segmentation

- Detect changes of trips based on captured time gaps, distance between photos and tags

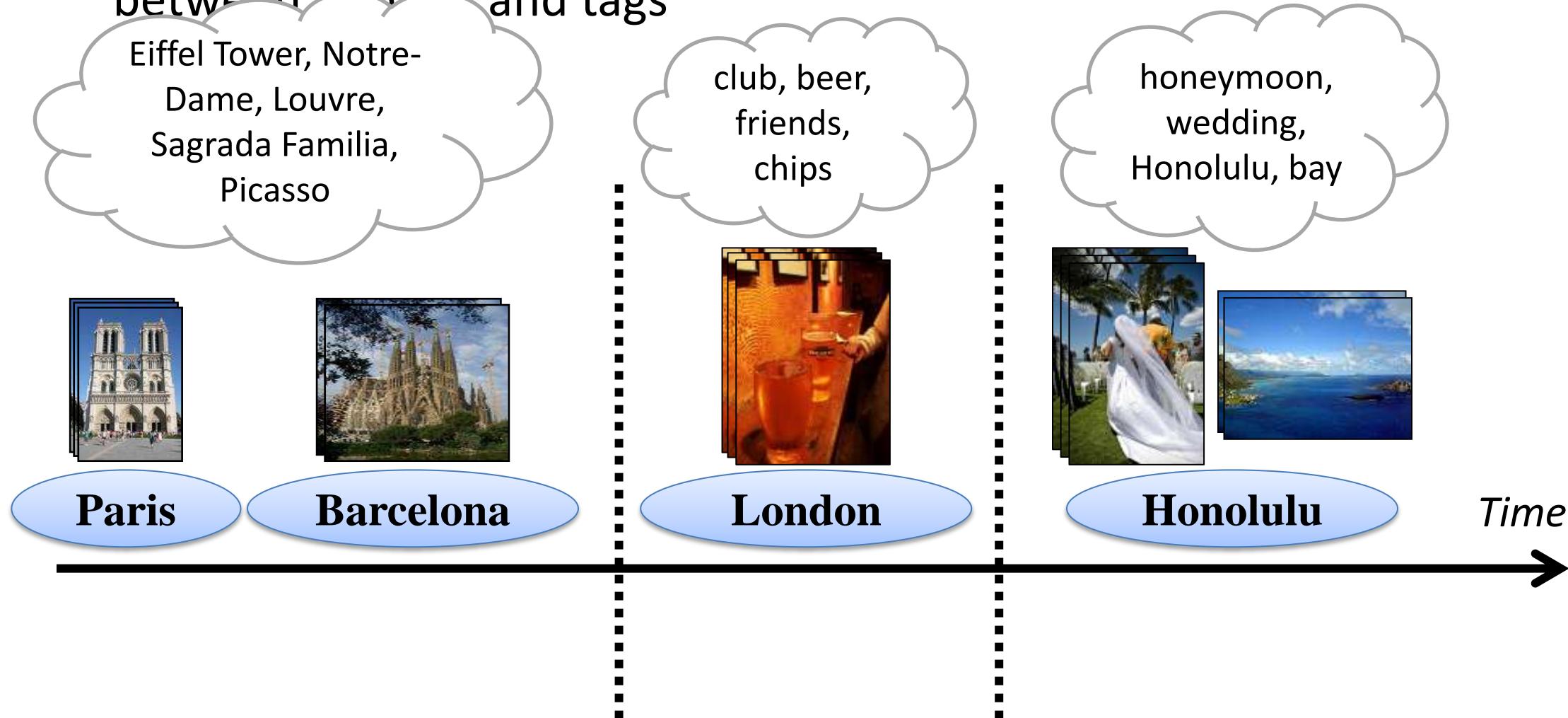
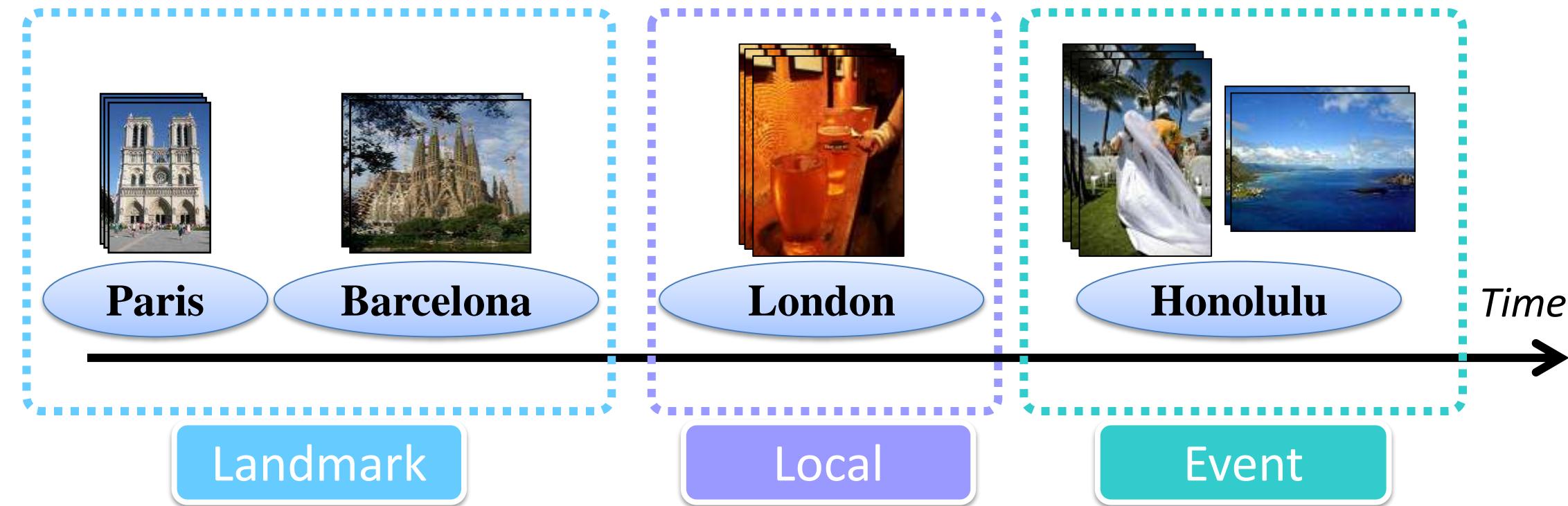


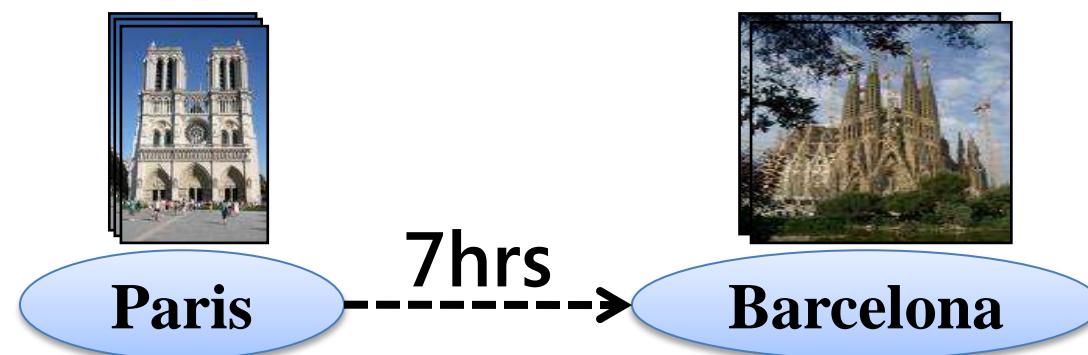
Photo Trip Pattern Mining: Classification

- Classify photo trips into categories by SVMs
 - Landmark/Nature/Gourmet/Event/Business/Local
 - Features: tags and locations

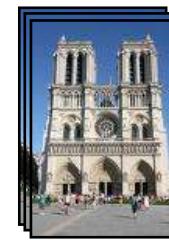


Trip Pattern Mining for Trip Classes

- Apply TAS (Temporary Annotated Sequence) mining algorithm
 - Input: Set of trips extracted from all users
 - Output: Frequent trip patterns, e.g., a set of visited cities and typical transition times.



Trip Semantic Identification



Trip semantics

Trip Semantic Identification

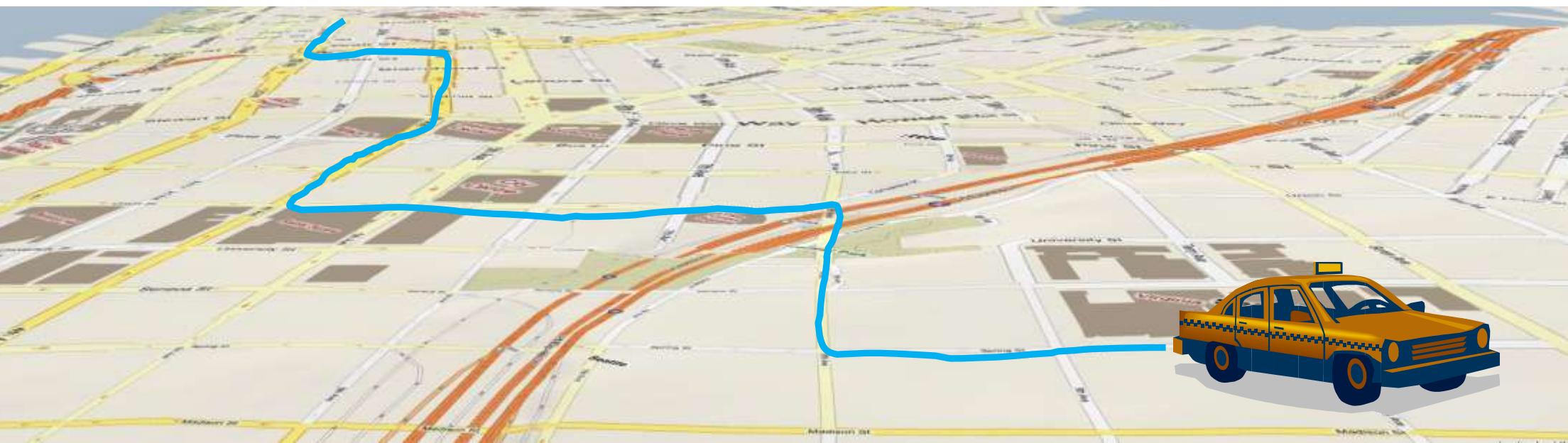
- Detect descriptive tags for each trip pattern
- TF/IDF based method
 - Tag frequency, inverse tag frequency
 - User frequency
- Consider geographical scale of tags to exclude locally/globally common tags
 - “shop”: globally common tags
 - “Beijing,” “BJ”: locally common tags

Evaluation

- Collected 5.7 million geo-tagged photos and conducted evaluation
- 72% precision and 85% recall for segmentation detection
- 79% accuracy for trip classification
 - Tags are most dominant feature
 - Combination of tags and locations performed best
 - Locations can compensate photos without tags

Examples

Trip class	Trip pattern	Trip semantics
Landmark	{Paradise, Las Vegas}	casinos, VMA, Bellagio, The Strip, WYNN
Nature	{Sydney, Randwick}	blue sky, barbed wire, inner, bay, Manly
Gourmet	{Camberwell, Melbourne}	cookie, spoon, rice, Colonial hotel, DJ
Event	{Washington D.C, Arlington}	mountain biking, WW, Wednesdays at Wakefield, mountain bike race, racing
Business	{Jersey City, New York, Jersey City}	comedians, MSN, live.com, Steve Kelley, Yahoo
Local	{Boston, Cambridge}	ants, mall, hospital, highway, living room



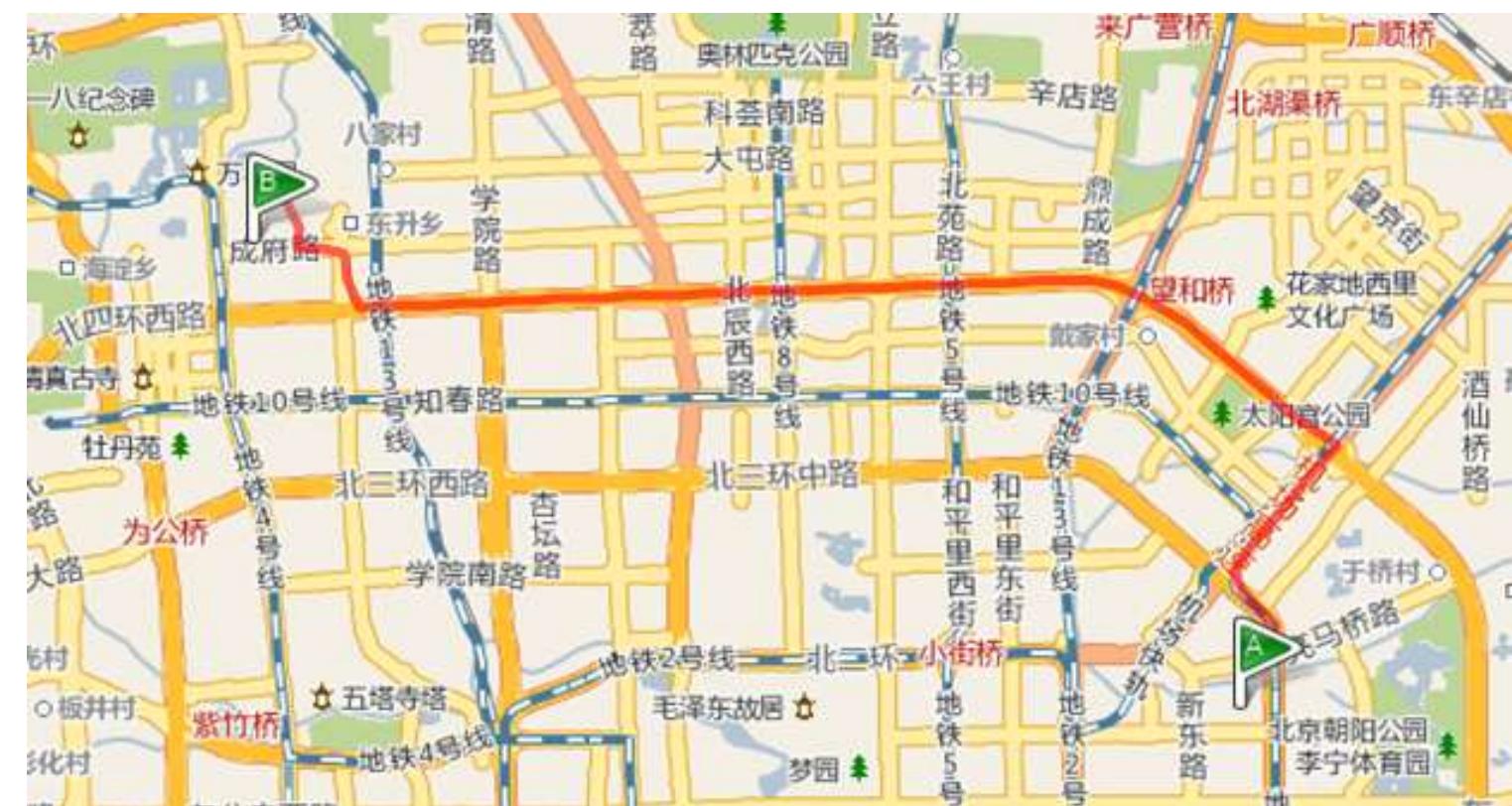
T-Drive: Driving Directions Based on Taxi Traces

Q=(q_s , q_d and t)

t =7:00am

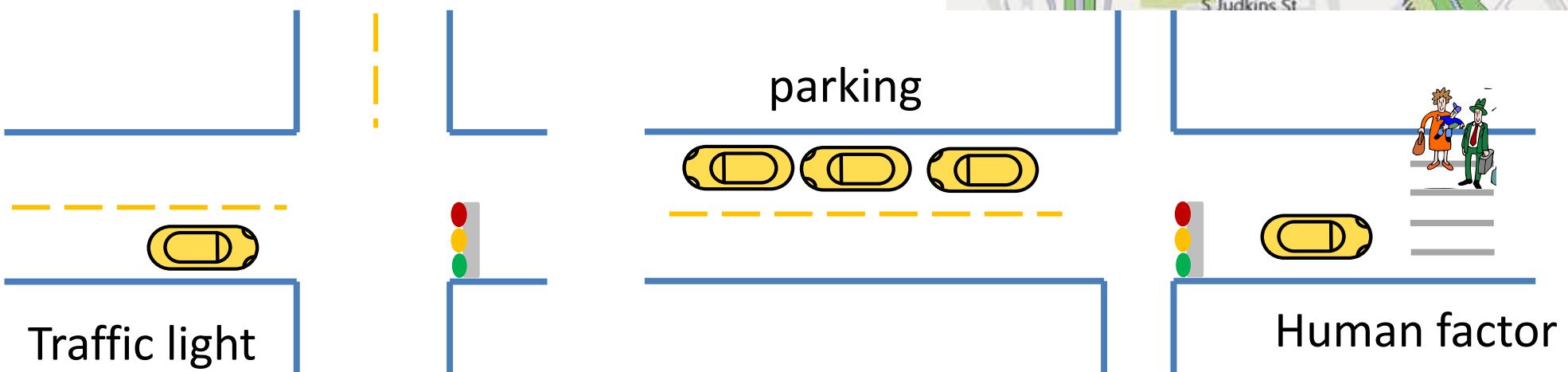
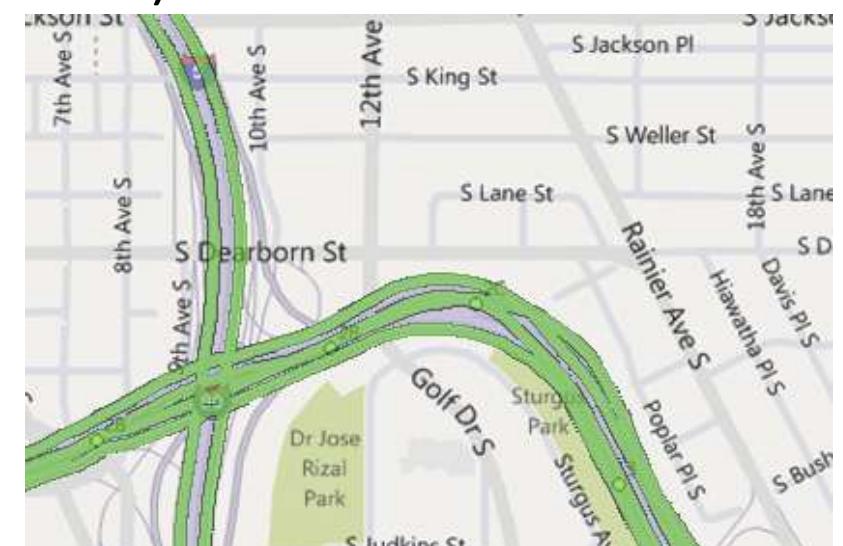


t = 8:30am



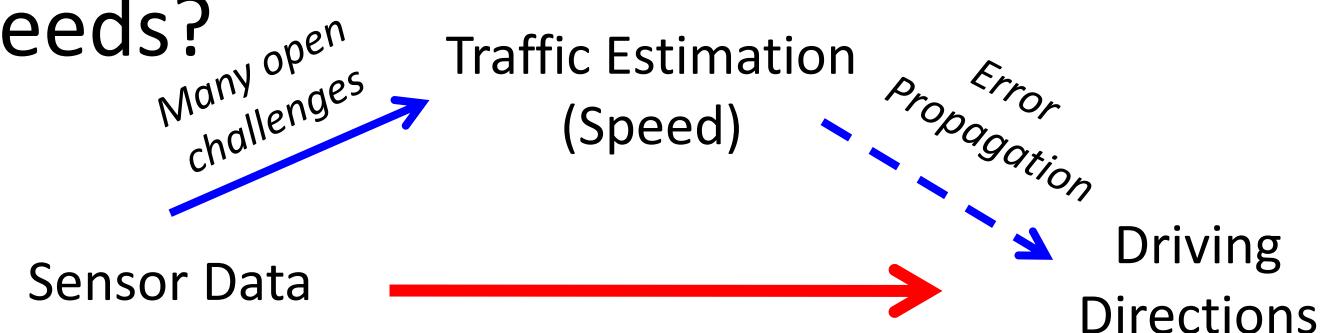
Background

- Shortest path and Fastest path (speed constraints)
- Real-time traffic analysis
 - Methods
 - Road sensors
 - Visual-based (camera)
 - Floating car data
 - Open challenges: coverage, accuracy,...
 - Have not been integrated into routing



Background

- What a drive really needs?



- Finding driving direction >> Traffic analysis



Physical
Routes



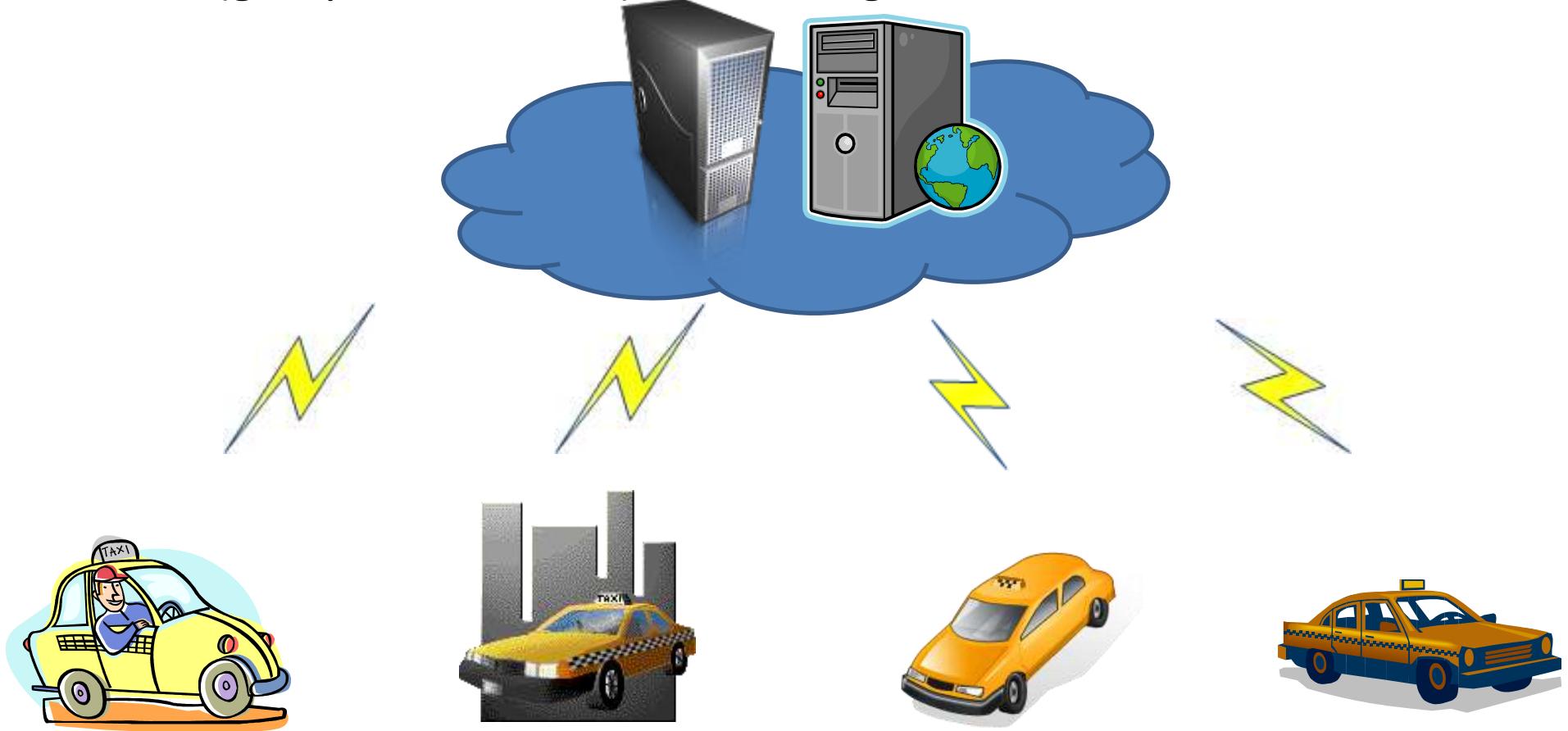
Traffic flows



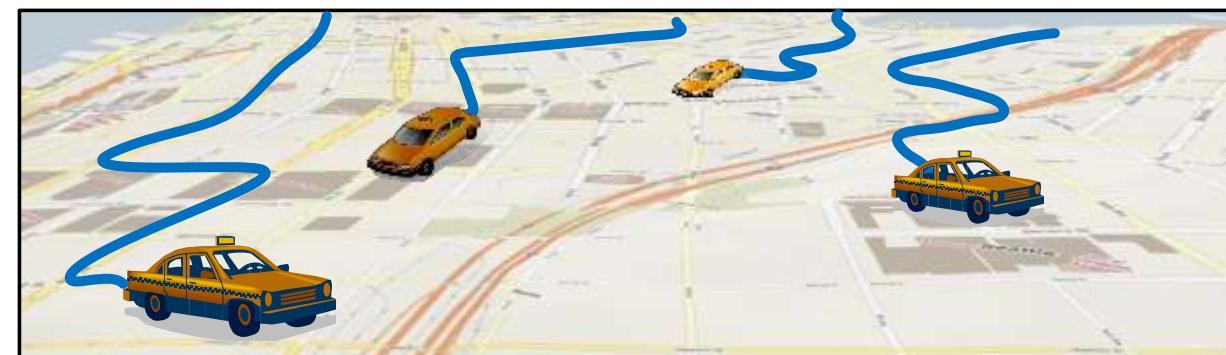
Drivers

Observations

- A big city with traffic problem usually has many taxis
 - Beijing has 70,000+ taxis with a GPS sensor
 - Send (geo-position, time) to a management center



Motivation



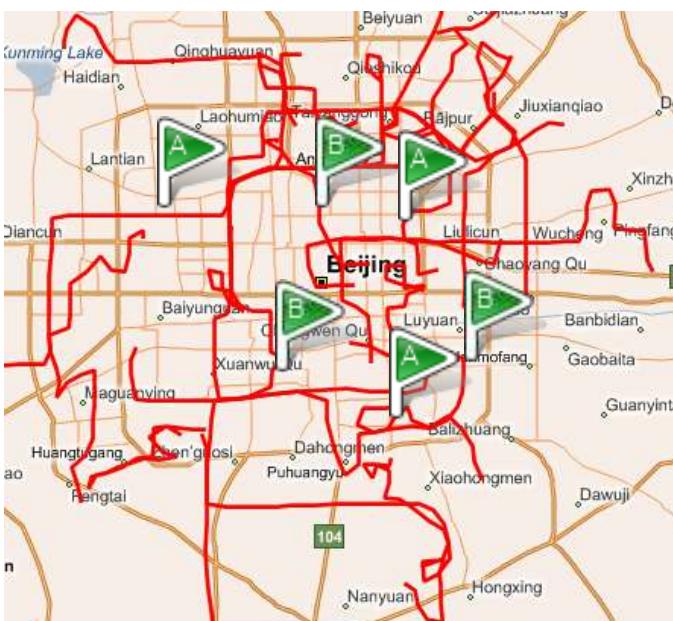
Human Intelligence



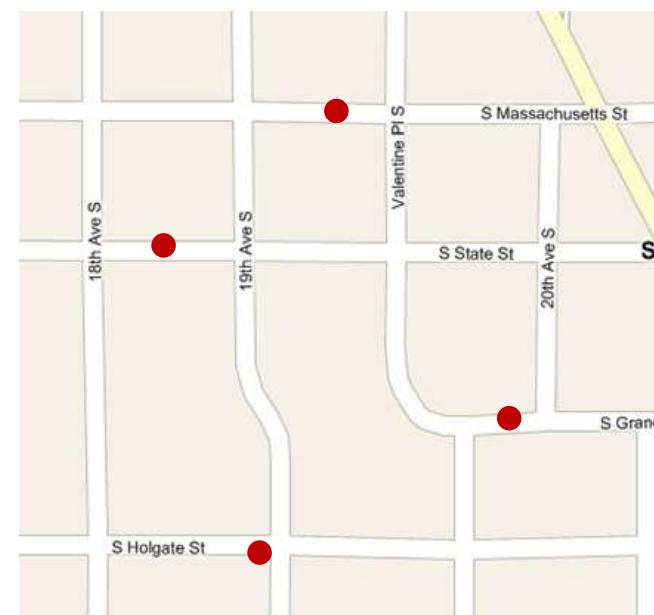
Traffic patterns

Challenges we are faced

- Intelligence modeling



- Data sparseness

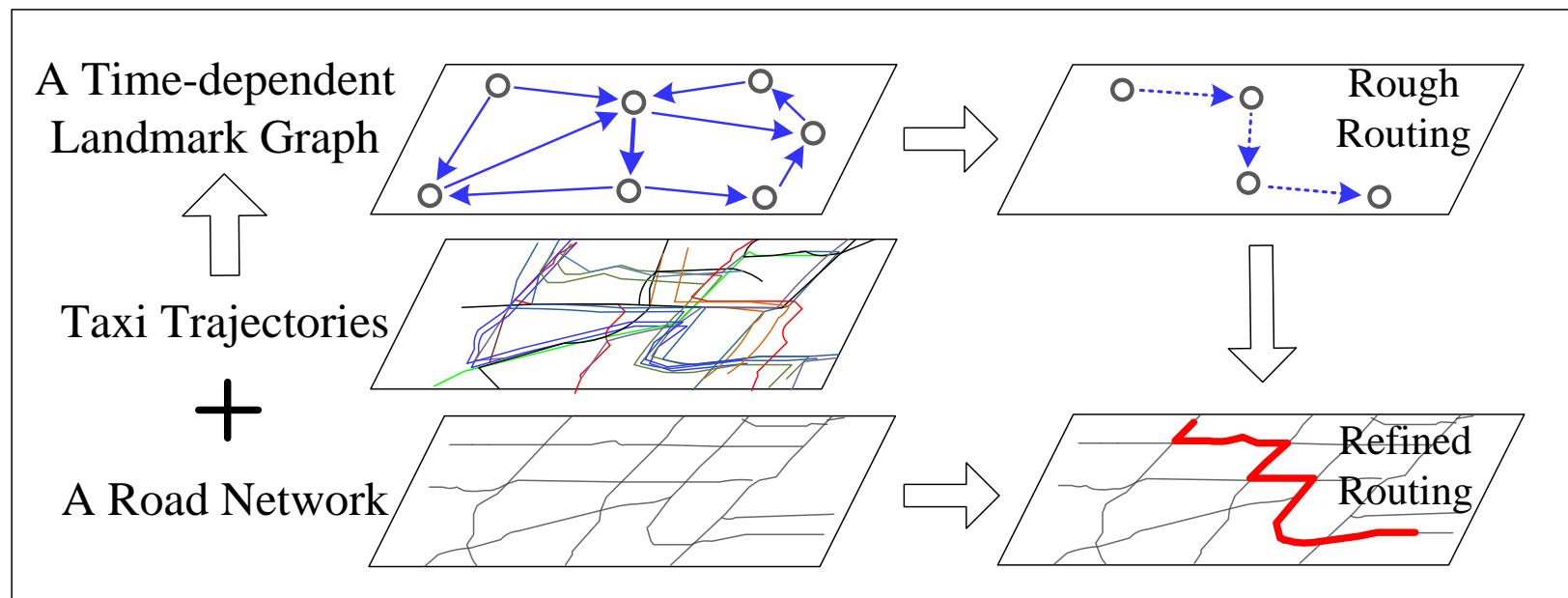


- Low-sampling-rate



Methodology

- Pre-processing
- Building landmark graph
- Estimate travel time
- Time-dependent two-stage routing



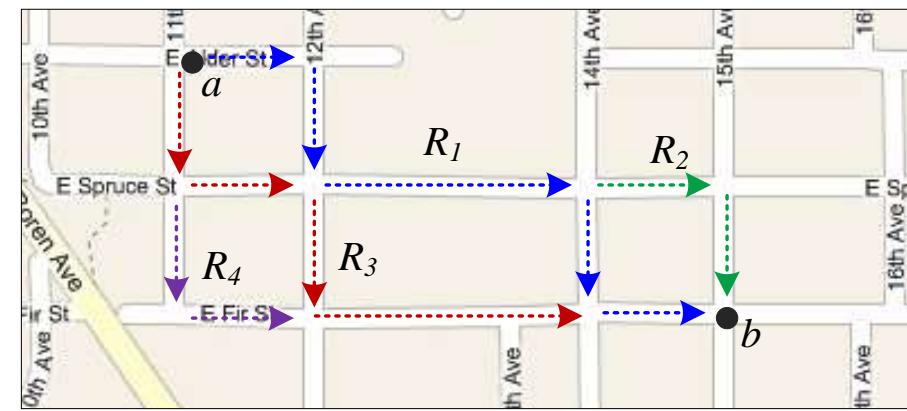
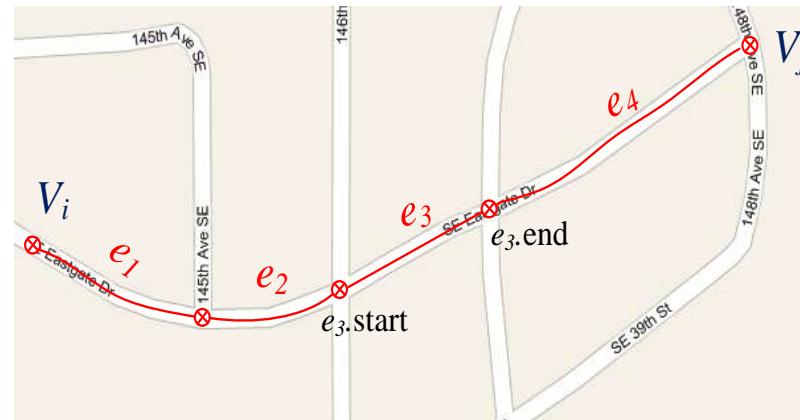
Step 1: Pre-processing

● Trajectory segmentation

- Find out effective trips with passengers inside a taxi
- A tag generated by a taxi meter

● Map-matching

- map a GPS point to a road segment
- IVMM method (accuracy 0.8, <3min)



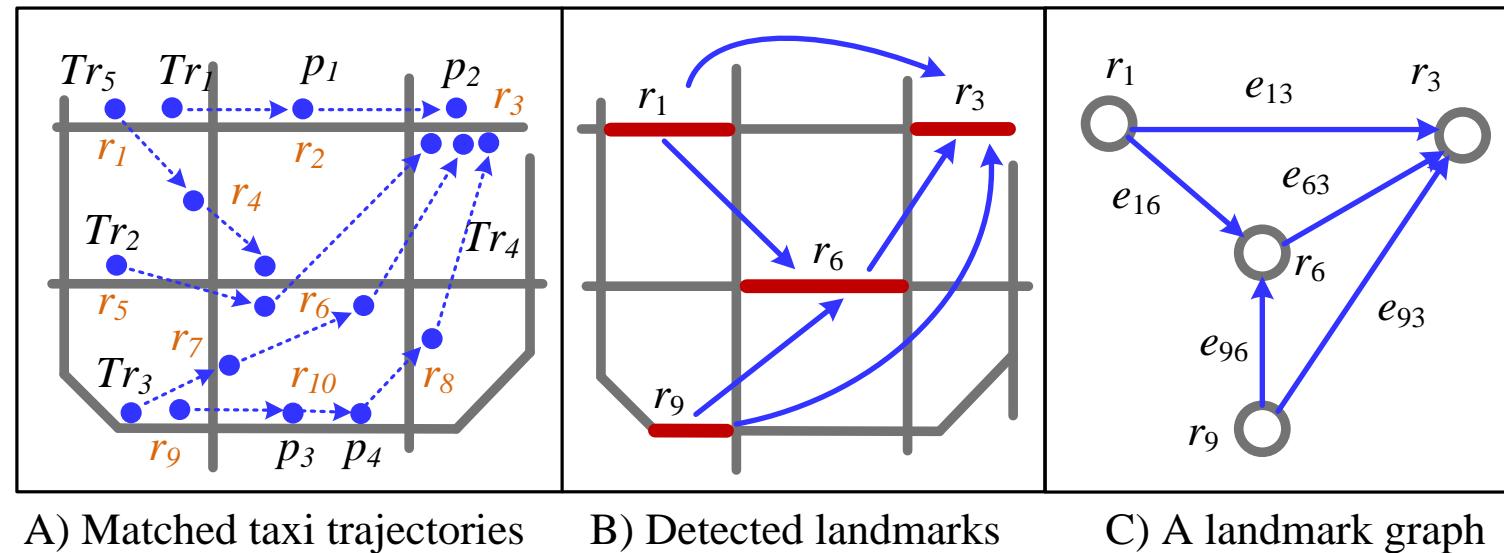
Step 2: Building landmark graphs

• Detecting landmarks

- A landmark is a frequently-traversed road segment
- Top k road segments, e.g. k=4

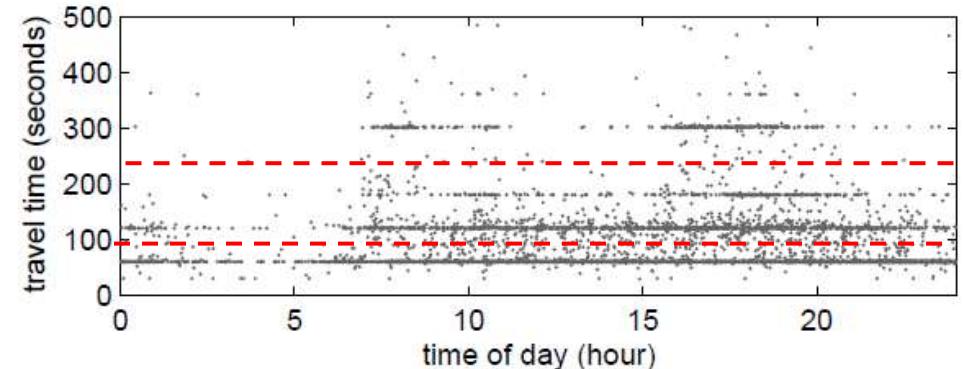
• Establishing landmark edges

- Number of transitions between two landmark edges $> \delta$
- E.g., $\delta = 1$

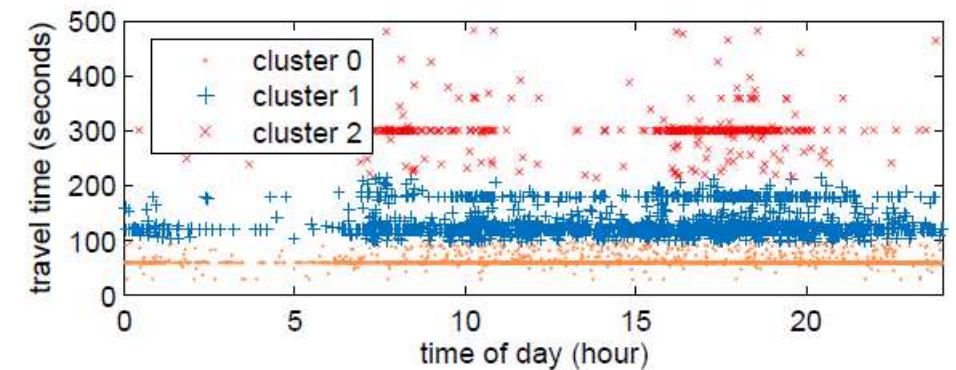


Step 3: Travel time estimation

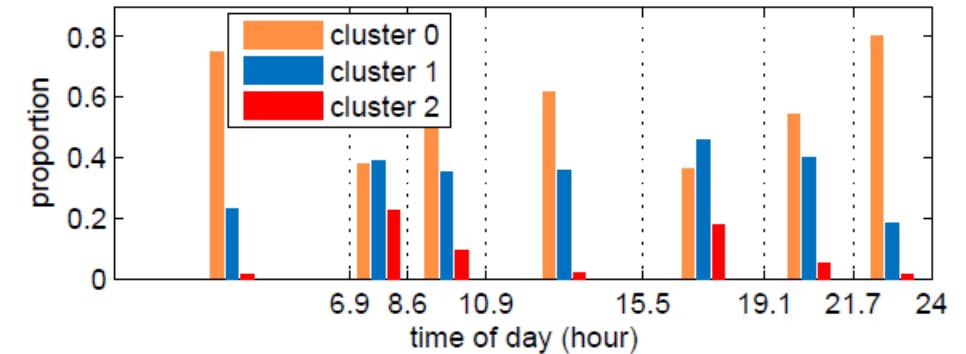
- The travel time of an landmark edge
 - Varies in time of day
 - is not a Gaussian distribution
 - Looks like a set of clusters
- A time-based single valued function is not a good choice
 - Data sparseness
 - Loss information related to drivers
 - Different landmark edges have different time-variant patterns
 - Cannot use a predefined time splits
- VE-Clustering
 - Clustering samples according to variance
 - Split the time line in terms of entropy



(a) Transitions of a landmark Edge



(b) V-Clustering result



(c) VE-Clustering result

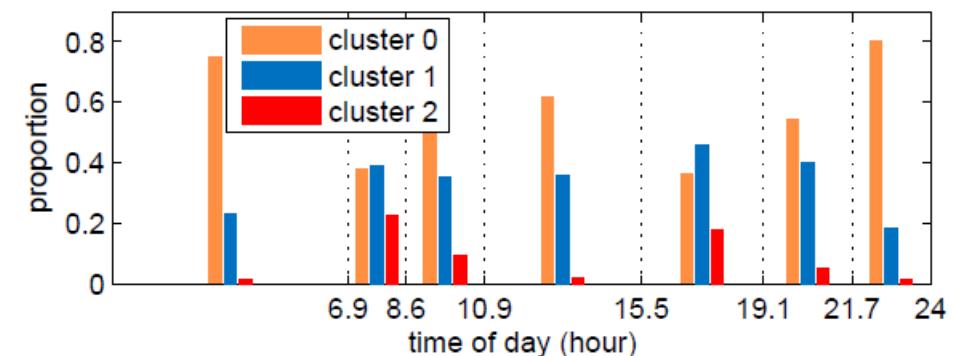
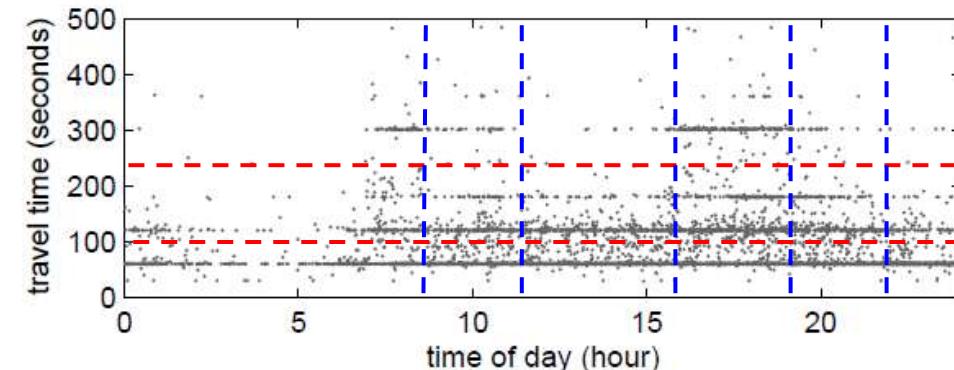
Step 3: Travel time estimation

• V-Clustering

- Sort the transitions by their travel times
- Find the best split points on Y axis in a binary-recursive way

$$\text{WAV}(i; L) = \frac{|L_1(i)|}{|L|} \text{Var}(L_1(i)) + \frac{|L_2(i)|}{|L|} \text{Var}(L_2(i))$$

$$\Delta V(i) = \text{Var}(L) - \text{WAV}(i; L).$$



(c) VE-Clustering result

• E-clustering

- Represent a transition with a cluster ID
- Find the best split points on X axis

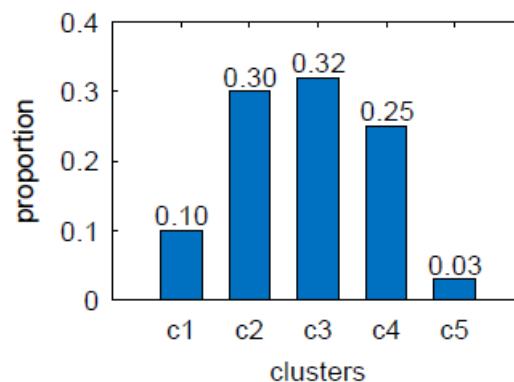
$$\text{WAE}(i; S^{xc}) = \frac{|S_1^{xc}(i)|}{|S^{xc}|} \text{Ent}(S_1^{xc}(i)) + \frac{|S_2^{xc}(i)|}{|S^{xc}|} \text{Ent}(S_2^{xc}(i))$$

$$\Delta E(i) = \text{Ent}(S^{xc}) - \text{WAE}(i; S^{xc}). \quad \text{Ent}(S^{xc}) = - \sum_{i=1}^m p_i \log(p_i)$$

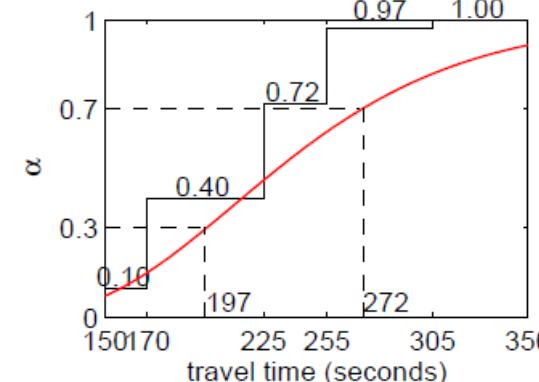
Step 4: Two-stage routing

• Rough routing

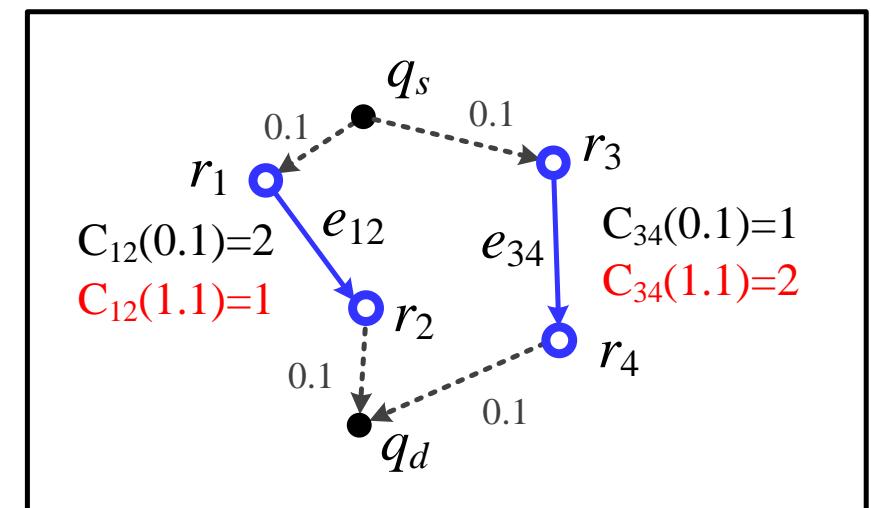
- Search a landmark graph for
- A rough route: a sequence of landmarks
- Based on a user query (q_s, q_d, t, α)
- Using a time-dependent routing algorithm



(a) Travel time distribution



(b) Cumulative frequency



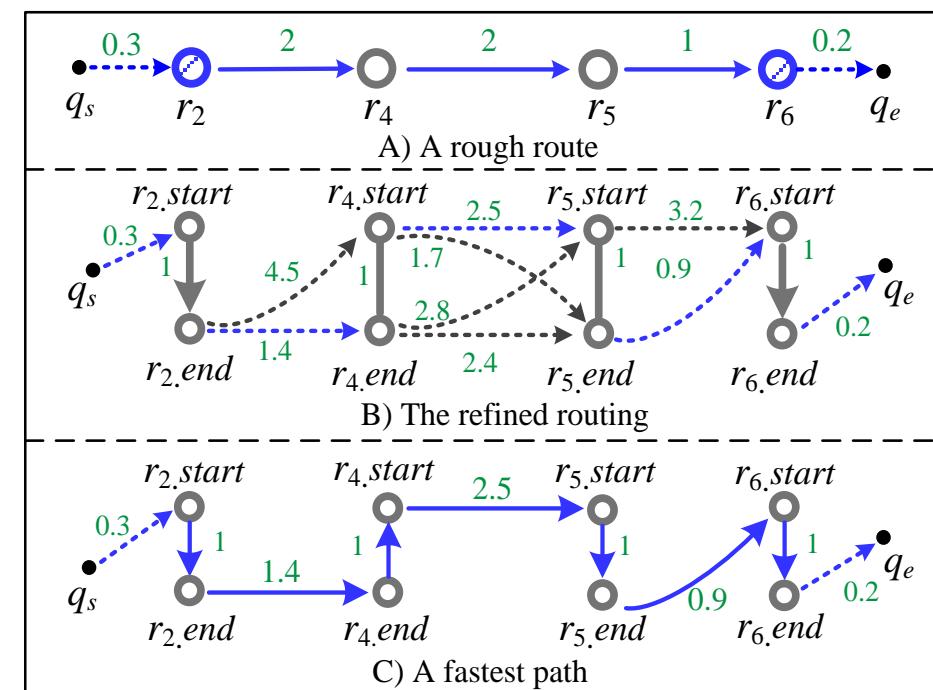
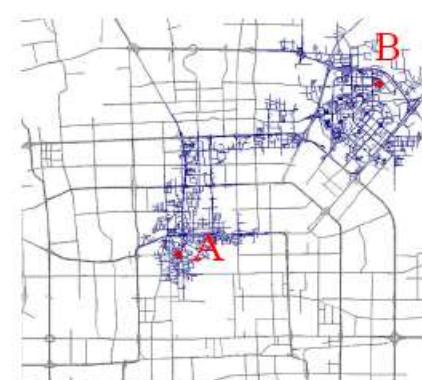
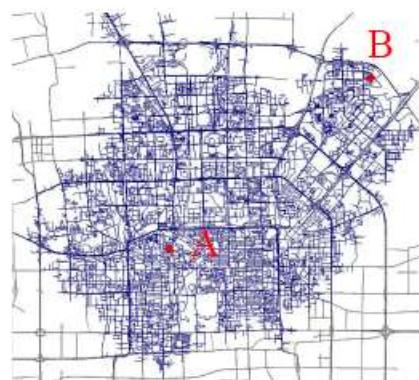
Step 4: Two-stage routing

Refined routing

- Find out the fastest path connecting the consecutive landmarks
- Can use speed constraints
- Dynamic programming

Very efficient

- Smaller search spaces
- Computed in parallel



Implementation & Evaluation

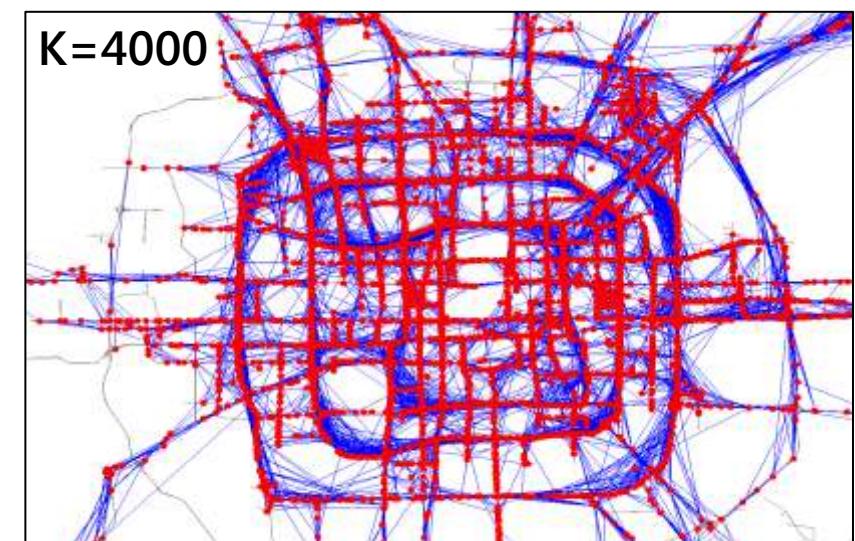
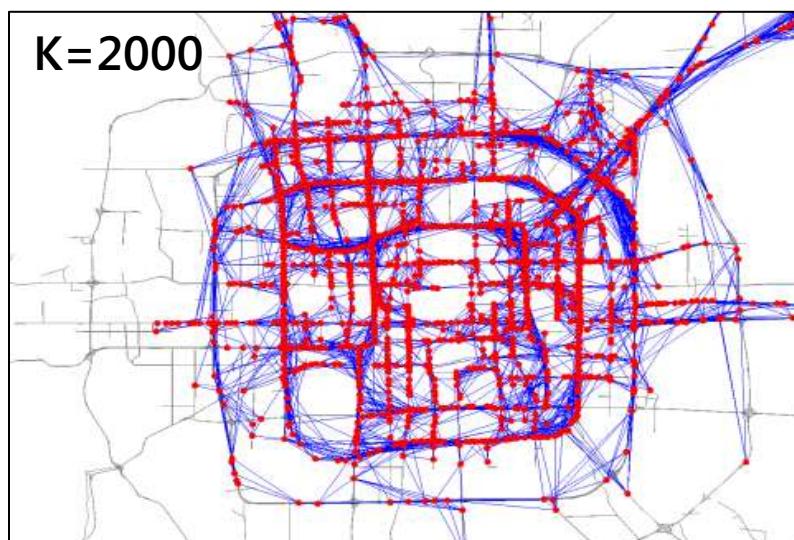
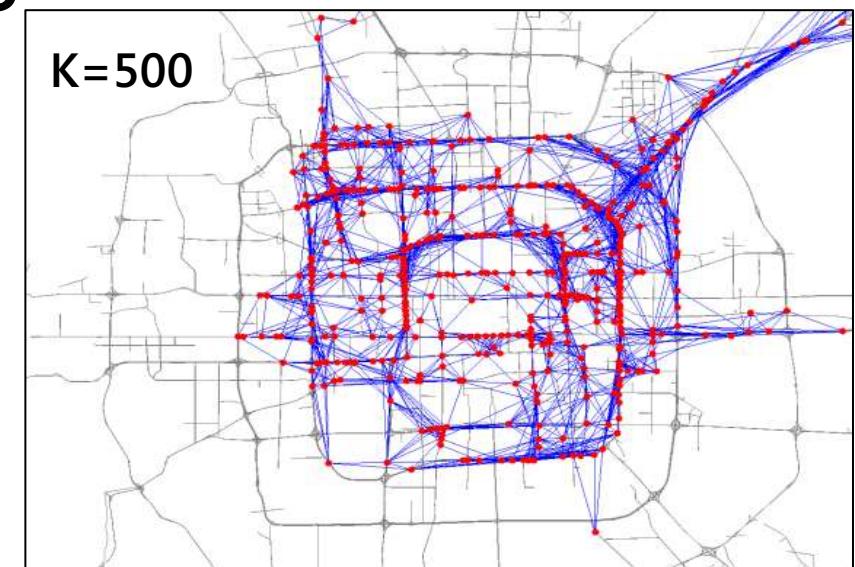
- **6-month real dataset of 30,000 taxis in Beijing**

- Total distance: almost 0.5 billion (446 million) KM
- Number of GPS points: almost 1 billion (855 million)
- Average time interval between two points is **2 minutes**
- Average distance between two GPS points is **600 meters**

- Evaluating landmark graphs
- Evaluating the suggested routes by
 - Using synthetic queries
 - In the field studies

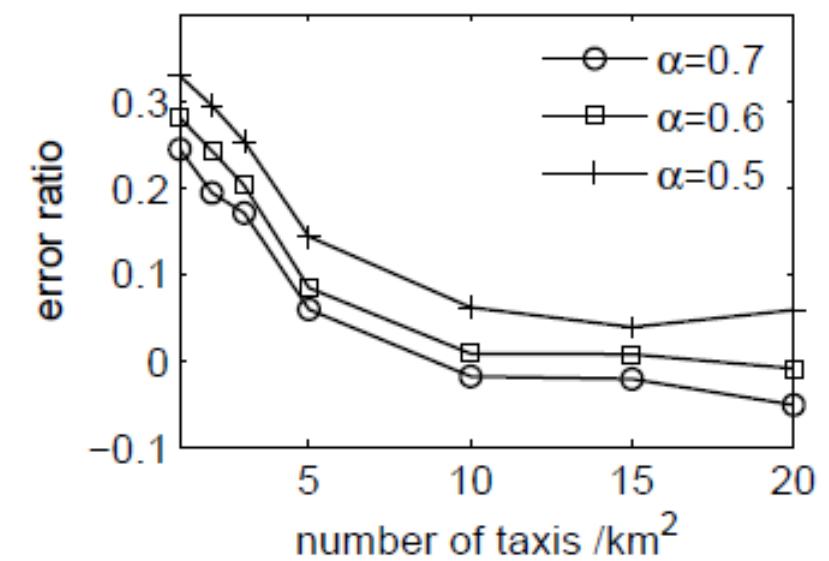
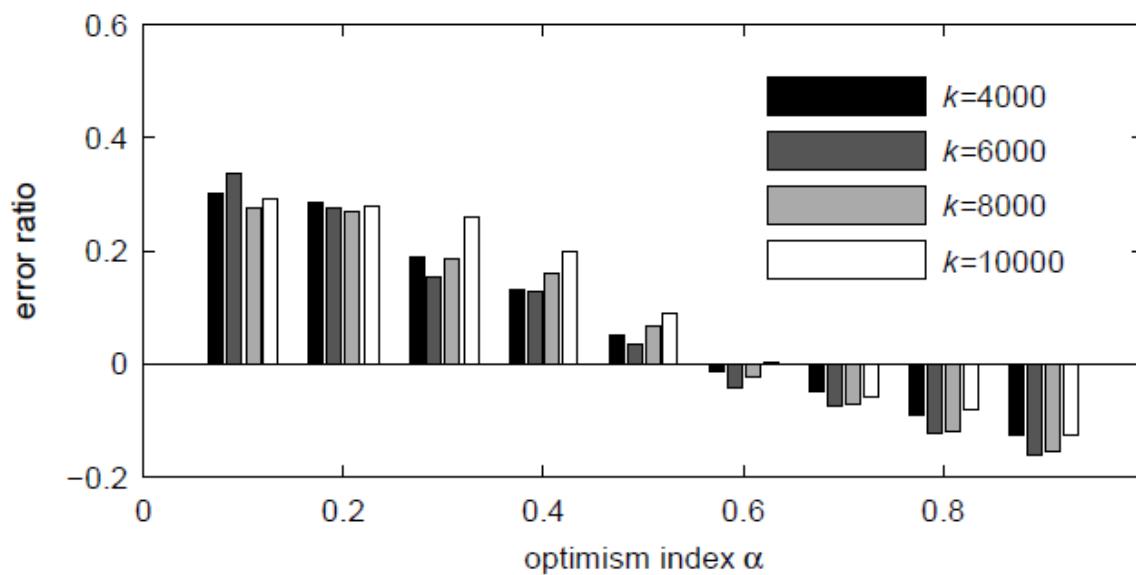
Evaluating landmark graphs

- Estimate travel time with a landmark graph
- Using real-user trajectories
 - 30 users' driving paths in 2 months
 - GeoLife GPS trajectories (released)



Evaluating landmark graphs

- Accurately estimate the travel time of a route
- 10 taxis/ km^2 is enough



(b) $k=10000$

Synthetic queries

- Baselines

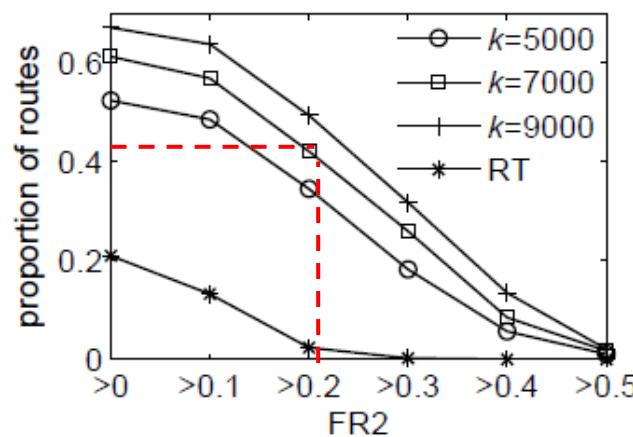
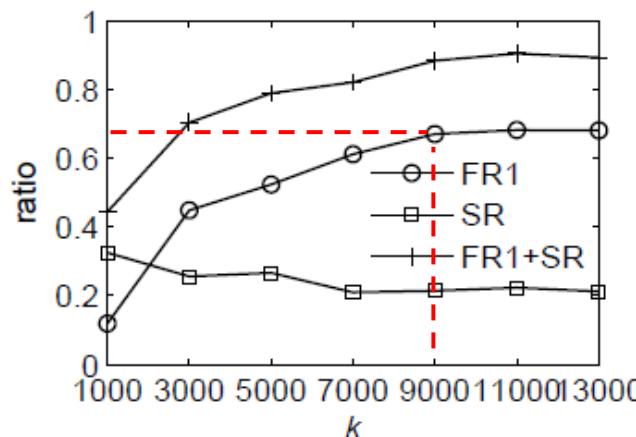
- Speed-constraints-based method (SC)
- Real-time traffic-based method (RT)

- Measurements

- FR1, FR2 and SR
- Using SC method as a basis

$$FR1 = \frac{\text{Number}(A's \text{ travel time} < B's \text{ travel time})}{\text{Number}(\text{queries})}$$

$$FR2 = \frac{B's \text{ travel time} - A's \text{ travel time}}{B's \text{ travel time}}.$$



	α	k	FR1	SR
TDrive	0.4	6,000	0.509	0.281
	0.4	9,000	0.647	0.222
	0.6	6,000	0.511	0.272
	0.6	9,000	0.653	0.216
	0.7	6,000	0.544	0.227
	0.7	9,000	0.672	0.214
RT approach			0.206	0.671

In the field study

- Evaluation 1
 - Same drivers traverse different routes at different times
- Evaluation 2
 - Different two users with similar driving skills
 - Traverses two routes simultaneously

Table 1: Trajectories of the In-the-field Study

	Evaluation 1	Evaluation 2
Num. Trajectories	360	60
Num. Users	30	2
Total Distance (km)	5304	814
Total Duration (hour)	165.24	25.09
Evaluation Days	10	6

Table 5: In-the-field Evaluation 1

	T-Drive	Google	△	R1	R2
Distance	13.91km	15.56km	1.65km	0.517	0.106
Duration	25.80min	29.28min	3.48min	0.808	0.119

Table 6: In-the-field Evaluation 2

	T-Drive	Google	△	R1	R2
Distance	13.58km	13.55km	-0.03km	0.367	-0.002
Duration	23.18min	27.00min	3.82min	0.750	0.141
WaitTime	4.77min	6.50min	1.73min	0.633	0.267

Results

- **More effective**
 - **60-70%** of the routes suggested by our method are faster than Bing and Google Maps.
 - Over **50%** of the routes are **20+%** faster than Bing and Google.
 - On average, we save **5** minutes per 30 minutes driving trip.
- More efficient
- More functional



Conclusions

- Build intelligence from the physical world
 - Activity/location recommendation based on GPS trajectories
 - Mining geo-tagged photos for travel recommendation
 - Driving directions based on taxi traces
- Challenges and future directions
 - How to protect privacy?
 - How to support real-time information sharing and search?
 - How to reduce energy consumption?

UbiComp 2011 in Beijing: weibo.com/ubicomp2011

Date: Sep. 17-21, 2011

Venue: Tsinghua University

Chairs: Yuanchun Shi (Tsinghua), James Landay (UW/MSR)

Program Chairs: Don Patterson (UCI), Yvonne Rogers (OU), Xing Xie (MSR)



Thanks!

Xing Xie

Microsoft Research Asia

Aug. 30, 2011

Generating Chinese Couplets using a Statistical MT Approach

Ming Zhou

Microsoft Research Asia

Outline

- Introduction
- Couplet generation model
- Experimental results
- Conclusions

Outline

- Introduction
- Couplet generation model
- Experimental results
- Conclusions

Chinese Couplet (对联)

- A special type of poetry
 - ❖ Composed of two sentences with same length and in a regular form
- ❖ One of the most important Chinese cultural heritages
 - ❖ Originates as early as the Five Dynasties (907 AD - 960 AD)
- An example:
 - ❖ First sentence (FS): “海阔凭鱼跃” (sea wide allow fish jump)
 - ❖ Second sentence (SS): “天高任鸟飞” (sky high permit bird fly)
 - ❖ The sea is wide enough so that fish can jump unrestrictedly; and the sky is high enough so that bird can fly at their pleasure.

Problem Definition

- Given the FS, to generate the SS so that the two sentences can form a qualified Chinese couplet
 - For example
 - Input: “海 阔 凭 鱼 跃” (sea wide allow fish jump)
 - Output: “天 高 任 鸟 飞” (sky high permit bird fly)
- A popular language game since one thousand years ago
- The difficulty of this problem comes from the principles of Chinese couplet

Chinese Couplets (<http://duilian.msra.cn>)



<http://video.sina.com.cn/v/b/10937201-1452530713.html>

FS and SS Share the Same Style

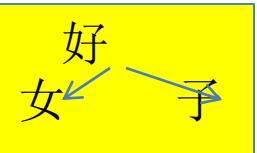
Repetition of
pronunciations(音韵联)

- 
- 风 (wind) ----- 水 (water)
 - 吹 (blow) ----- 使 (make)
 - 荞(buckwheat) ----- 舟 (ship)
 - 动(wave) ----- 流 (go)
 - 桥 (bridge) ----- 洲 (island)
 - 未 (not) ----- 不 (not)
 - 动(wave) ----- 流(go)

FS and SS Share the Same Style

Decomposition of
characters (拆字联)

有 (have)	-----	缺 (lack)
子 (son)	-----	鱼 (fish)
有 (have)	-----	缺 (lack)
女 (daughter)	-----	羊 (mutton)
方 (so)	-----	敢 (dare)
称 (call)	-----	叫 (call)
好 (good)	-----	鲜 (fresh)



FS and SS Share the Same Style

Person
name
(人名联)

Palindrome
(回文联)

板桥(Banqiao)----- 东坡 (Dongpo)
造(produce) ----- 居 (live)
桥(bridge) ----- 坡 (mountain)
板(board)----- 东(east)

- Banqiao(板桥) and Dongpo(东坡) are famous litterateurs
- Reading from top to down is identical to down to top

Principle 1

- The FS and SS should agree in length and word segmentation
 - FS: 知识 能 致 富 (knowledge can bring richness)
 - SS: 勤劳 可 兴 家 (work can raise family)
- English translation: knowledge can make one rich and hard work can make one's family live better.

Principle 2

- Corresponding words in FS and SS should agree in their part of speech

海	阔	凭	鱼	跃
sea	wide	allow	fish	jump
天	高	任	鸟	飞
sky	high	permit	bird	fly
noun	adjective	conjunction	noun	verb

Principle 3

- The contents of the FS and SS should be related but cannot be duplicated
 - FS: 海阔凭鱼跃 (sea wide allow fish jump)
 - SS: 天高任鸟飞 (sky high permit bird fly)
 - Examples in different situations: fish in the sea and bird in the sky
 - Same truth that only in broad space one can use all one's talents

Principle 4

- The last character of the FS should be pronounced in “仄”(Ze) tone
- The last character of the SS should be in “平”(Ping) tone
 - FS: 海 阔 凭 鱼 跃 (sea wide allow fish jump)
 - SS: 天 高 任 鸟 飞 (sky high permit bird fly)
 - 跃-> “仄”(Ze)
 - 飞-> “平”(Ping)

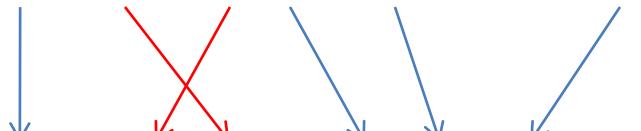
Principle 5

- The writing styles of the FS and SS should be identical
 - Character repetition
 - Pronunciation repetition
 - Character decomposition
 - FS: “有 女 有 子 方 称 好”(have daughter have son so call good)
 - SS: “缺 鱼 缺 羊 敢 叫 鲜”(lack fish lack mutton dare call delicious)

MT vs. SS Generation

- Machine translation
 - He sent her a bunch **of** flowers .

– 他 **给** 她 **送** **了** 一 束 花 。



- SS generation
 - FS: 海 阔 凭 鱼 跃 (sea wide allow fish jump)

– SS: 天 高 任 鸟 飞 (sky high permit bird fly)

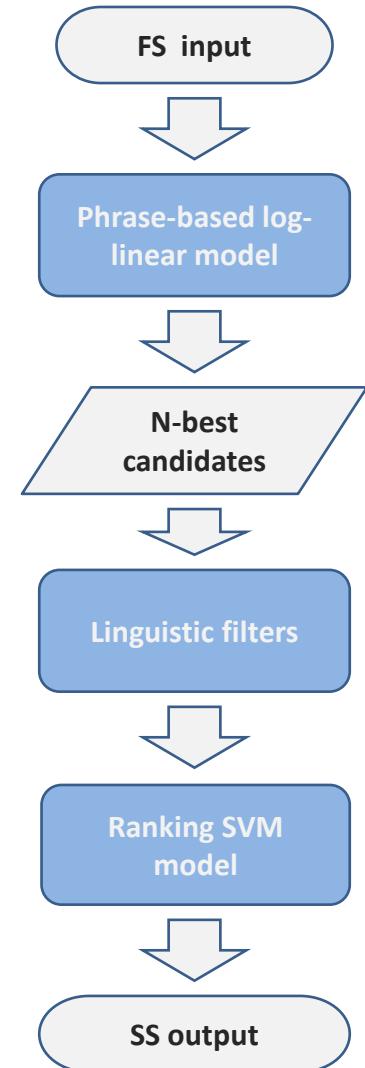
- No word insertion, deletion and reordering

Outline

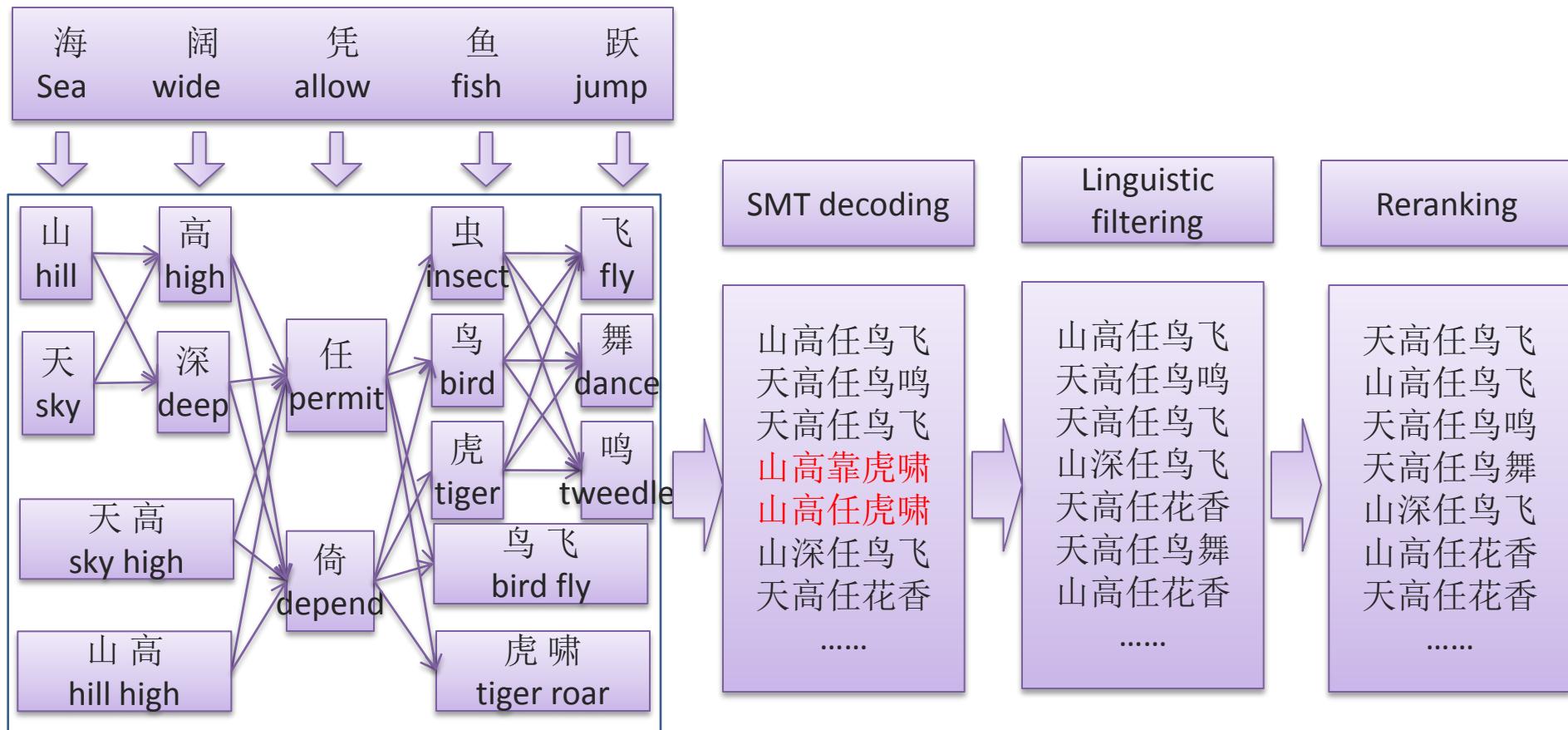
- Introduction
- Couplet generation model
- Experimental results
- Conclusions

SS Generation Approach

- A multi-phase SMT approach
 - Phase1: a phrase-based log-linear model
 - Phase2: some linguistic filters
 - Phase3: a Ranking SVM



SS Generation Process



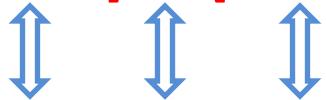
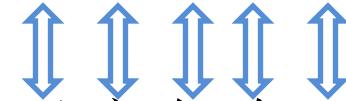
Phrase-based Log-linear Model

- Given a FS denoted as $F=\{f_1, f_2, \dots, f_n\}$, to seek a SS denoted as $S^*=\{s_1, s_2, \dots, s_n\}$ that satisfies

$$S^* = \arg \max_S \sum_{i=1}^M \lambda_i \log h_i(S, F)$$

- Where f_i and s_i are Chinese characters
- Five feature functions
 - Phrase translation model (PTM)
 - Inverted PTM
 - Character translation model (CTM)
 - Inverted CTM
 - Language model

Feature Functions

- Phrase translation model and inverted PTM
 - FS: [海 阔] 凭 [鱼 跃] ([sea wide] allow [fish jump])

 - SS: [天 高] 任 [鸟 飞] ([sky high] permit [bird fly])
- Character translation model and inverted CTM
 - FS: 海 阔 凭 鱼 跃 (sea wide allow fish jump)

 - SS: 天 高 任 鸟 飞 (sky high permit bird fly)
- Language model
 - Character-based trigram model
 - $p(\text{海 阔 凭 鱼 跃}) = p(\text{海} | \text{START}) * p(\text{阔} | \text{START 海}) * p(\text{凭} | \text{海 阔}) * p(\text{鱼} | \text{阔 凭}) * p(\text{跃} | \text{凭 鱼})$

Training Data

- Couplet Data
 - Classic Chinese couplets
 - From books
 - From the web
 - Extract sentence pairs from ancient Chinese poems
 - From couplet forums
 - 970,000 couplets obtained finally
- LM training
 - Chinese poems besides the couplet data for LM training
 - 1,600,000 sentences in total

Linguistic Filters

- Only SMT model can not guarantee the SS has
 - The same writing styles as the FS
 - Correct tone for the last character
- Four filters
 - Character repetition filter
 - Pronunciation repetition filter
 - Character decomposition filter
 - Phonetic harmony filter

Linguistic Filters 1

- Character repetition filter
 - The FS
 - “有女有子方称好” (have daughter have son so call good)
 - SS candidates
 - “缺鱼缺羊敢叫鲜” (lack fish lack mutton dare call delicious)
✓
 - “缺鱼少羊敢叫鲜” (lack fish miss mutton dare call delicious)
✗

Linguistic Filters 2

- Pronunciation repetition filter
 - The FS
 - “风 吹 莽 动 桥 未 动” (wind blow buckwheat wave bridge not wave)
 - SS candidates
 - “水 使 舟 流 洲 不 流” (water make ship move island not move)
✓
 - “水 使 舟 流 岛 不 流” (water make ship move island not move)
✗

Linguistic Filters 3

- Character decomposition filter
 - The FS
 - “有 女 有 子 方 称 好” (have daughter have son so call good)
 - SS candidates
 - “缺 鱼 缺 羊 敢 叫 鲜” (lack fish lack mutton dare call delicious)
✓
 - “缺 鱼 缺 牛 敢 叫 鲜” (lack fish lack beef dare call delicious)
✗

Linguistic Filters 4

- Phonetic harmony filter
 - The FS
 - “海 阔 凭 鱼 跃” (sea wide allow fish jump)
 - 跃 : Ze
 - SS candidates
 - “天 高 任 鸟 飞” (sky high permit bird fly)
 - 飞: Ping
 - ✓
 - “山 高 任 虎 啸” (mountain high permit tiger roar)
 - 啸: Ze
 - ✗

Candidate Re-ranking

- Ranking SVM for re-ranking SS candidate
 - To leverage long-distance features

$$f_{\vec{w}}(\vec{x}) = \langle \vec{w}, \vec{x} \rangle$$

- Two more features
 - Mutual information (MI)
 - MI-based structural similarity (MISS)
- Parameter estimation
 - Tool: SVM Light
 - Training data: 200 FSs and each of them has 50 SSs labeled as positive or negative by human

Candidate Re-ranking (con't)

- Mutual information (MI)
 - Motivation
 - Candidate 1: “天 高 任 鸟 飞” (sky high permit bird fly)
 - Candidate 2: “天 高 任 狗 叫” (sky high permit dog bark)
 - Candidate 1 is better
 - $MI(\text{天,鸟}) > MI(\text{天,狗})$ ($MI(\text{sky, bird}) > MI(\text{sky, dog})$)
 - To measure the semantic consistency of words in a candidate SS

$$MI(S) = \sum_{i=1}^{n-1} \sum_{j=i+1}^n I(s_i, s_j) = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \log \frac{p(s_i, s_j)}{p(s_i)p(s_j)}$$

Candidate Re-ranking (con't)

- MI-based structural similarity (MISS)
 - Motivation: structural similarity in Chinese couplets
 - 海阔凭鱼跃 (sea wide allow fish jump)
 - 天高任鸟飞 (sky high permit bird fly)
 - To measure the structural similarity

$$MISS(F, S) = \cos(V_f, V_s) = \frac{V_f \bullet V_s}{|V_f| \times |V_s|}$$

- Given the FS $F = \{f_1, f_2, \dots, f_n\}$, we first build its MI vector

$$V_f = \{MI_{12}, MI_{13}, \dots, MI_{1n}, MI_{23}, \dots, MI_{n-1n}\}$$

Outline

- Introduction
- Couplet generation model
- Experimental results
- Conclusions

Automatic Evaluation of SS

- BLEU for SS evaluation

$$BLEU = BP \bullet \exp\left(\sum_{n=1}^N w_n \log p_n\right)$$

- N = 3; BP = 1
- P_n is position-sensitive
- Data set
 - 1051 FSs and their SSs mined from couplet forums
 - 24 [references](#) for each FS on average
 - 600 as development set and 451 as test set

Translation Unit Setting

- Experiment setting
 - Only translation model and language model are used
 - Same training data, same linguistic filters and no re-ranking

Translation Unit setting	BLEU
character-based	0.236
word-based	0.261
phrase-based	0.276

Feature Evaluation

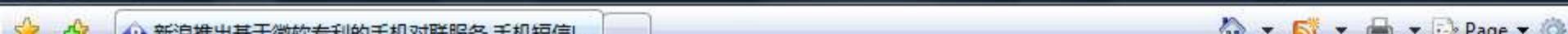
- ❖ Incrementally adding new features
- ❖ Same linguistic filters for all settings

	Features	BLEU
Baseline	Phrase TM(PTM) + LM	0.276
	+ Inverted PTM	0.282
	+ Character TM (CTM)	0.315
	+ Inverted CTM	0.348
Ranking SVM	+ Mutual information (MI)	0.356
	+ MI-based structural similarity	0.361

Overall Performance

- By human evaluation
 - 100 FSs
 - Output 10 best SSs by our best system for each FS
 - Human labeling: acceptable or not
 - Metric
 - *top-n inclusion rate* is defined as the percentage of the test sentences whose top-n outputs contain at least one acceptable SS.

	Top-1	Top-10
<i>Top-n inclusion rate</i>	0.21	0.73



新浪推出基于微软专利的手机对联服务

CNET中国 · PChome.net · 编译 作者： 责编:江海明 时间:2009-01-08 标签： 手机短信 MSRA 春节短信

2009年1月7日，美国雷德蒙及中国北京——微软公司今天与新浪公司共同宣布，双方签署了一项有关“中文对联生成器”技术的专利许可协议，它是微软亚洲研究院（MSRA）在自然语言处理领域的尖端创新之一。此项专利授权将增强新浪在中国市场提供创新性移动增值服务的能力。

根据这项许可协议，微软的专利技术将用于新浪的全新移动电话对联服务，手机用户可以将自己编写的上联以短信形式发送至新浪的服务器，服务器上运行的微软中文对联生成引擎将自动构造下联以及横批，并以彩信或短信的形式发送给用户。这项服务将于2009年1月6日起投入运作，迎接新春佳节（2009年1月26号），届时人们将发送数十亿短消息互致新年问候。新浪还计划利用微软的核心技术自行开发其他创新产品，为这项技术寻找新用途，并且使移动增值服务更加具有个性、互动性和娱乐性。

“对联的分享与展示是中国千百年来的一项春节传统。我们非常高兴，因为此次与微软的全新合作将帮助我们把这一习俗搬到电脑和移动电话上。”新浪公司副总裁，新浪无线总经理王高飞表示。“通过直接与微软研究院合作，我们实现了先进的机器学习和语言技术与我们自有创新的成功结合，扩展了这项基于Web的技术，用于手机和短信，进而提供一项对于中国乃至全球华人客户十分有意义的服务。”

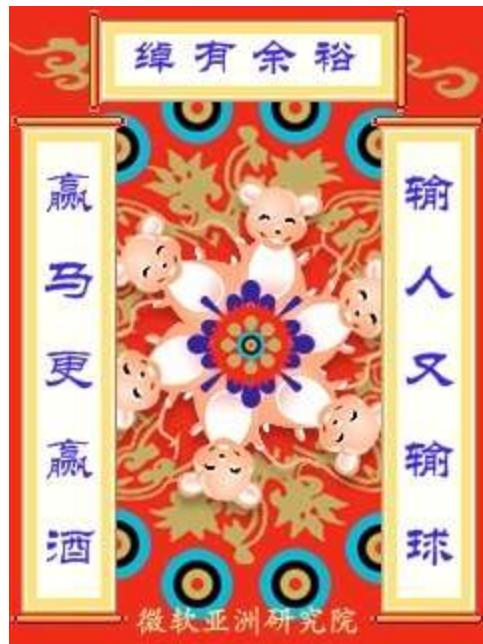
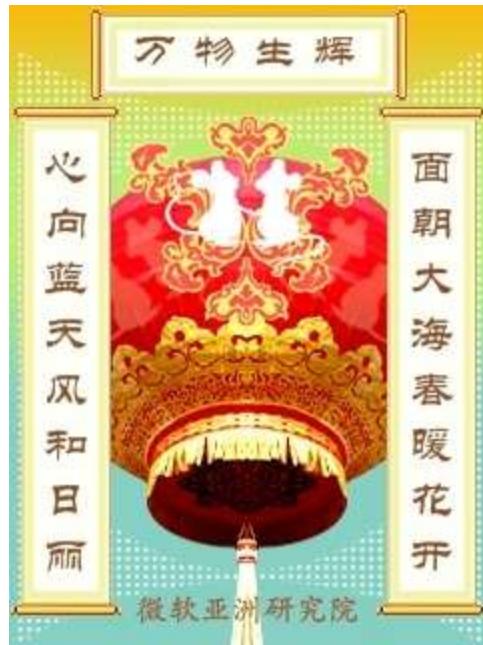
“我们感到高兴的是，新浪选择了微软研究院自然语言处理技术的使用许可证。”位于北京的微软亚洲研究院院长洪小文博士说：“这项协议是一个很好的例子，我们向来致力于通过知识产权的合作提升消费者体验。这也是我们与中国本地企业展开更密切合作的重要步骤之一，让中国IT产业更具活力。”

热门搜索: 智能 诺基亚 索爱 索尼 单反
iPod 高清 7300GT 笔记本 Dell 直销

EPSON
EXCEED YOUR IMAGINATION

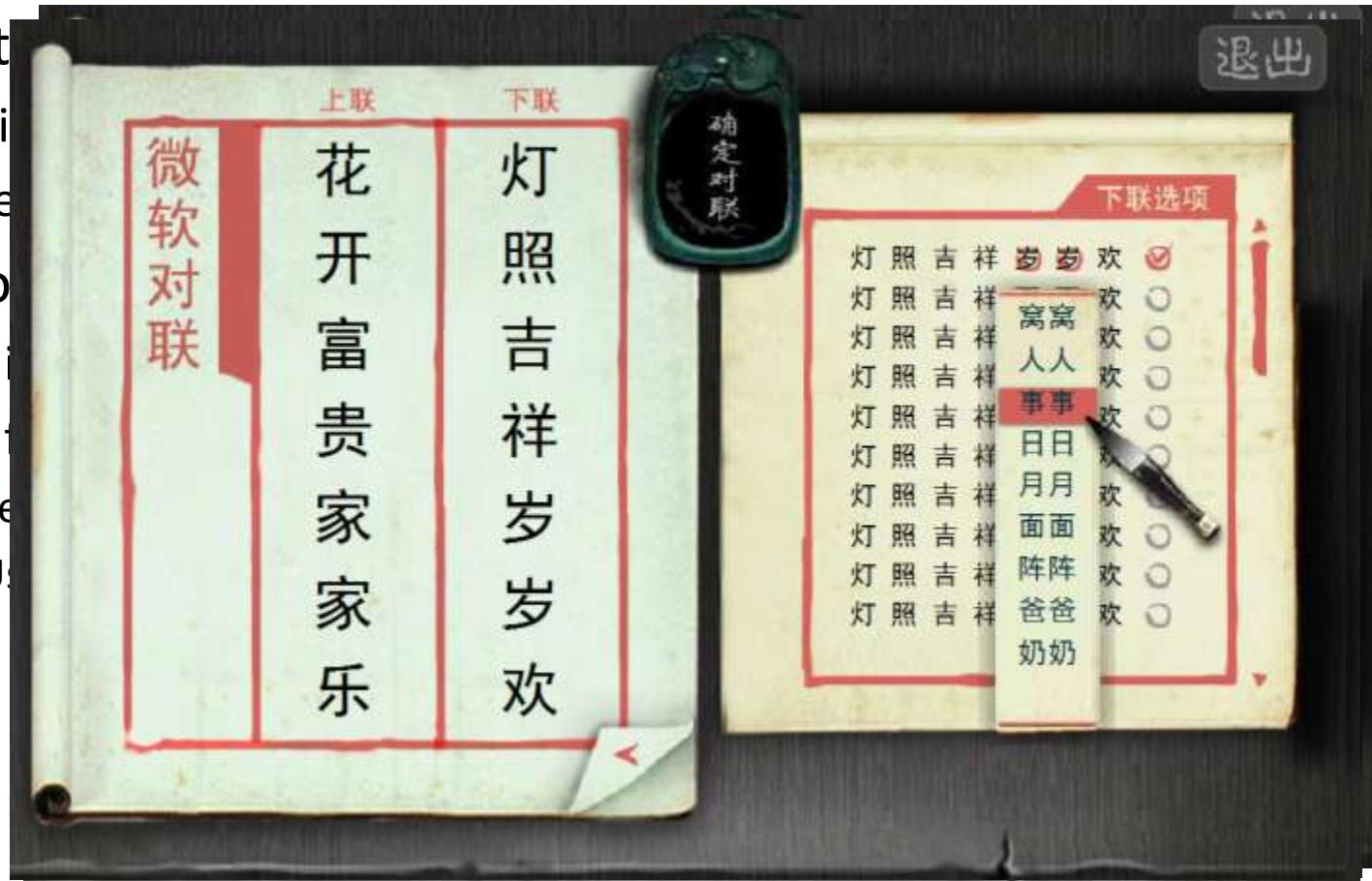
金牛送福
带 ME 回家
火热促销中！
现在购买ME系列打印机及商
务打印机即可获赠精美礼品！

看图说新闻



User log for Model Enhancement

- Motivation
 - Training
 - While
- What logs
 - User input
 - User feedback
 - Selective
 - Using

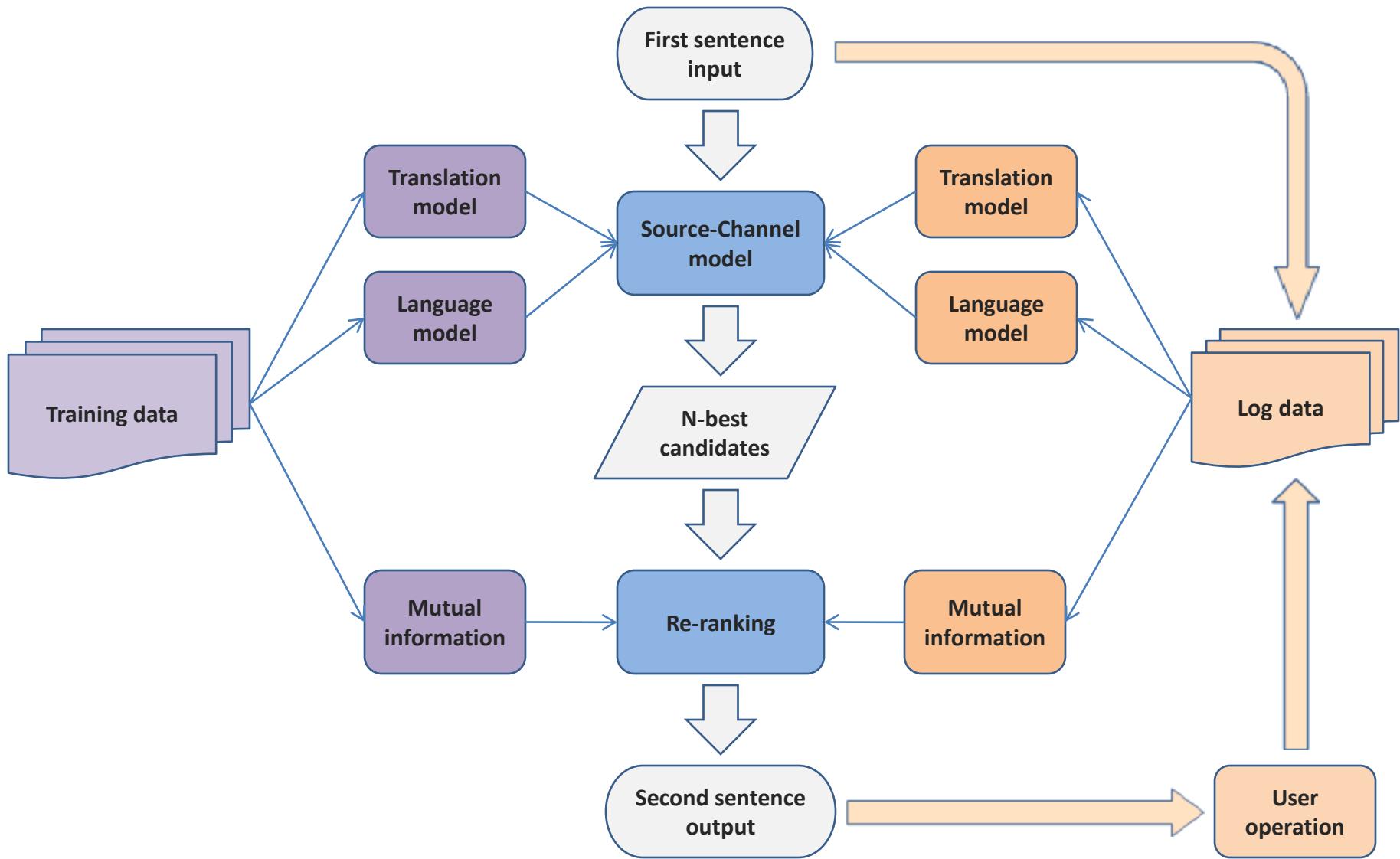


User's Log Analysis

Number of input sentences	12,322
Number of unique input sentences	6,698
Users directly select from system output	3,459
User manual modify system output	606
Save as favorite couplets	109
Invalid user input	618
No second sentence generated	2,211
Banner generation	2,687
Select the generated banner as favorite	428
No banner output	265

- Data Source
 - Log from
<http://couplet.msra.cn>
- Time period
 - Aug. 31-Oct. 9, 2006

New Framework with Log Data



Extended to Quatrains(绝句)

• 感归

零落泣鬼神
秋来愁人心
肝肠断挥洒
不思归日吟

• 春兴

残花飘黄叶
细雨落青山
蝶飞红杏里
燕舞绿杨湾

• 从军北征

雁字风月一时清
天书云山千里远
锦字凭谁寄笔力
人生何日归来晚

• 望洞庭

移舟雨逐行云水
一路风随日月天
云破长江万里船
风来一水千山川

Outline

- Introduction
- Couplet generation model
- Experimental results
- Conclusions

Conclusions

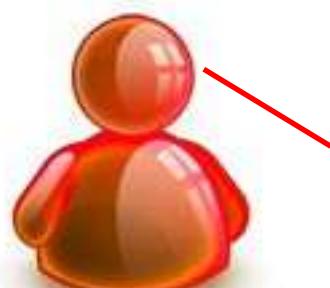
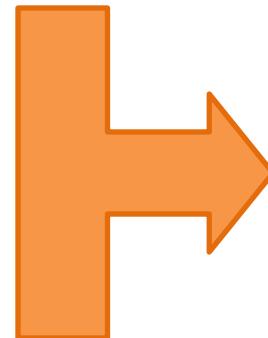
- Conclusions
 - A multi-phase SMT approach for generating the second sentence given a first sentence of a Chinese couplet
 - Promising experimental results obtained
 - A popular website (<http://couplet.msra.cn>)
 - 50,000 visitors / day during the peak time
- Future work
 - Word clustering for smoothing TM
 - As a extension, work on Chinese poetry generation

Thanks!
Questions?

The Game of Second Sentence Generation



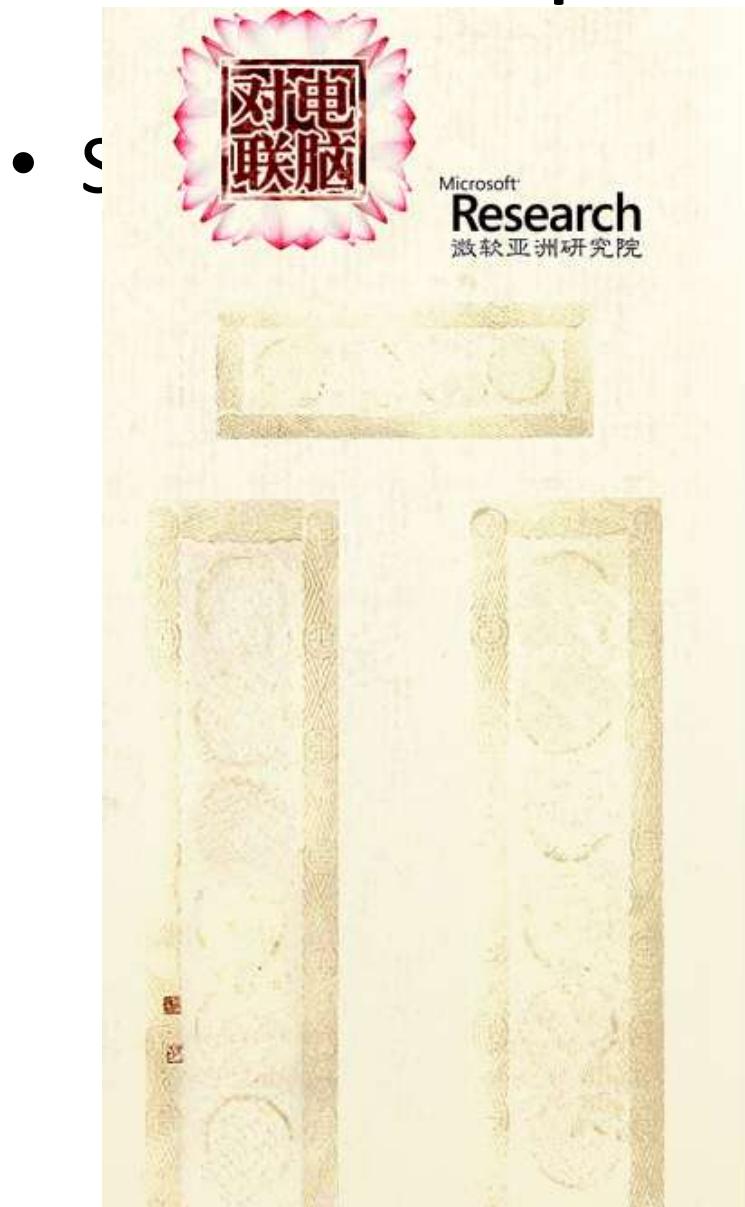
创大业一帆风顺



展宏图万事胜意



Couplet Web Service



第一步 拟上联

海阔凭鱼跃

下联示例

海阔凭鱼跃

爆竹一声辞旧岁

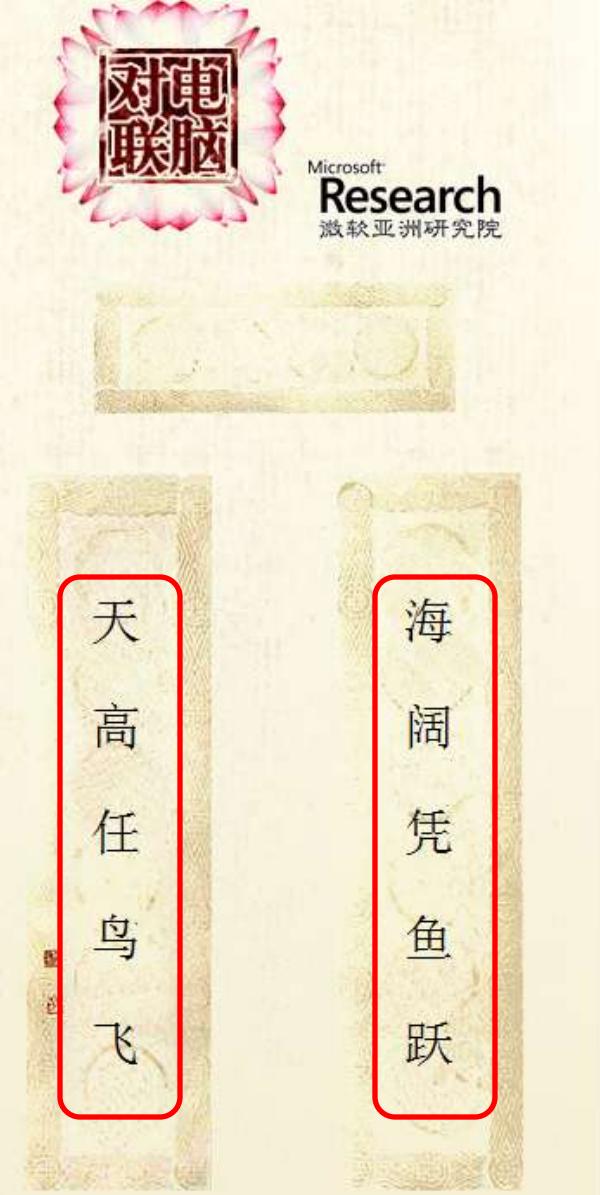
灵鼠迎春春色美

海南南海出海观景

鸿是江边鸟

Couplet Web Service (con't)

- Step 1
a)



This screenshot shows the second step of the couplet generation process. It features a traditional Chinese background with a red seal at the top right labeled '横批' (couplet title).

第一步 拟上联 (Step 1: Propose Upper Couplet):
上联: 海 | 阔 | 凭 | 鱼 | 跃

第二步 对下联 (Step 2: Oppose Lower Couplet):
下联: 天 | 高 | 任 | 鸟 | 飞

在输入框内输入部分下联，点击刷新候选
In the input box, enter part of the lower couplet, click to refresh suggestions

刷新候选 (Refresh Suggestions) button

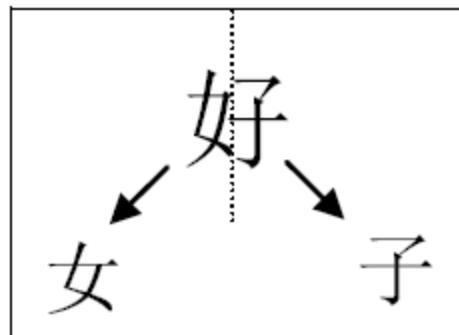
A list of suggested lower couplets is shown in a red-bordered box:

- 天高任鸟飞
- 天空任鸟飞
- 天高任你飞
- 天高任我飞
- 天高任鸟翔
- 路遥任马驰
- 天高任花香
- 天高任鸟语
- 天高任流行
- 山高与人行

如果对结果不满意，推荐您 换一种方式
If you are not satisfied with the results, we recommend you try another way

Character Decomposition

- Character decomposition is an interesting language phenomenon in Chinese: some Chinese characters can be decomposed into other characters.
- For example, “好” (good) can be decomposed into “女” (daughter) and “子” (son).



Some Reference Examples

- The FS
 - 品茶不为渴 (degust tea not because thirstiness)
- References:
 - 弹曲却因情 (play zither but for feeling)
 - 踏雪只因梅 (trample snow only for plum)
 - 醉酒总关情 (drink wine always relate feeling)
 - ...



Couplet in Poems

- Eight-line Poem

春望

国破山河在，
城春草木深。
感时花溅泪，
恨别鸟惊心。
烽火连三月，
家书抵万金。
白头搔更短，
浑欲不胜簪。



Engkoo

Parallel data mining from the web

Video:

<http://video.sina.com.cn/v/b/37417609-1286528122.html>

Rapidly Changing Language

- Approximately 1.5 billion people speak English as a primary, secondary or business language
- China: The largest “English speaking” country with 250 million English learners and USD 60 billion annual expenses
- Problem: Live language: new words, new meanings

Key Insight:

With billions of translated web pages and sharable repositories of language data growing every day, the Internet holds the sum of human language knowledge

Engkoo 英库

Major Features:

首页 | 潜规则

Endless Lexicon with Native Definitions

词形变化 (无相关结果)

网络释义:

1. [hidden rules](#)
<http://61.152.116.30/bbs/read.php?tid=128369>

例句 类别: 全部 ▾ 来源: 全部 ▾ 难度: 全部 ▾

4. [Prev](#)
对潜规则的定义 - 专业术语与名词 / 网络用语
<http://www.ilib.cn/RSS-zndxkb-shlxkb.xml>
5. [Opening the Hidden Regulation of Officialism](#)
破解官场潜规则
<http://www.ilib.cn/l-hebswdxgb.2006.02.html>

State-of-the-Art Machine Translation (NIST OpenMT Winner)

If we walk slow
如果我们比其

Real-time Interactive Alignment

其

www.engkoo.com

Microsoft Products:

A screenshot of the Bing search results page. The search term 'trigger' is highlighted in blue. Below it, there's a snippet of text from a dictionary entry: '英汉词典中 trigger 的中文解释'. The page also shows other search results and navigation links.

Office **Jan** | 

1 微软 2 微 3 喂 4 为 5 魏  

Massive Dictionary Mined from the Web

The screenshot shows a Microsoft Internet Explorer browser window displaying the Engkoo dictionary website. The search term '小心碰头' is entered in the search bar. The results page shows the following content:

词形变化 (共 2 件中)

1. 网络
Café’ 一行大字，当你来到门口你会注意到几个醒目标语：
Pull; 小心碰头 Mind your head 等。走进饭堂，你会看到堂
餐后必须清理餐具并放到饭筐 Clean and return your tr
Wash hands before meal! 小心地滑，请勿乱跑 Beware of
此洗碗、抹布等其它东西 For washing hands only, please
里还有“意见箱 (Suggestions Box)呢。再走走看看：什么
· 山西外企资源网 · 山西外企资源网入口 · 会议厅宣

Fresh and Diverse Examples

The screenshot shows the Engkoo website interface. At the top left is the Engkoo logo with a sun icon and the word 'BETA'. To the right is a search bar containing the Chinese characters '与时俱进'. Below the search bar is a navigation menu with tabs for '首页' (Home) and '例句:与时俱进' (Example Sentences: Keeping Pace with the Times). Underneath the search bar are dropdown menus for '类别' (Category), '来源' (Source), and '难度' (Difficulty), all set to '全部' (All). The main content area displays four numbered examples.

1. Go ahead with **times** and open up the new phase of community health ser
与时俱进开创社区卫生服务新局面
<http://www.ilib.cn/l-zgfybj.2003.11.html>
2. Develop with the **advance** of the **times**
与时俱进
<http://218.22.70.71/printtext.asp?id=62917>
3. **Keeping pace** with the **times** and the Party mind line
与时俱进与党的思想路线
<http://www.ilib.cn/l-hbdxxb-zxsh.2003.01.html>
4. Marx and Engels are the models of **keeping pace** with the **times**
马克思恩格斯是与时俱进的典范

Advanced Search with Sentence Analysis

The screenshot shows the Microsoft Engkoo English Corpus (BETA) search interface. The search query "She is a adj. lady" is entered in the search bar, with the adjective "adj." highlighted by a red box. The search results page displays four examples of sentences containing this phrase, each with its Chinese translation and a link to the original source.

Microsoft
Engkoo BETA 英库

She is a adj. lady

首页 小心碰头 She is a adj. l... X

例句 类别: 全部 来源: 全部 难度: 全部 逐词释义

2. **She is a charming young lady.** 她是一个迷人的年轻女士。
<http://news.sogou.com/news?query=she%27s+a+lady>

3. **She is a perfect lady.** 她是一位十足的淑女。
<http://dict.bitunion.org/xdict.php?word=lady>

4. **She is a devout old lady.** 她是个虔诚的老太太。
<http://gzs1.tougao.com/UserWork/alluser/901382/index/Article.asp?ArticleID=121507>

Microsoft



微软Bing在美国的份额继续上升



首页 潜规则 微软Bing在美国... X

微软Bing在美国的份额继续上升

报告问题或瑕疵



计算机翻译:

Microsoft Bing shares continue to rise in the United States

微软Bing在美国的份额继续上升

Sentences Classification

例句 类别: **书面语** ▾ 来源: 全部 ▾ 难度: ▾

1. The **mouse** sped down the oak-tree. 耗子刷地一下便下了橡树。

2. The **mouse** squeaked in the corner. 墙角处有老鼠的吱吱声。

3. They can grow in all **mouse** strain. 它们可以在所有的小鼠株内生长。

例句 类别: **技术** ▾ 来源: 全部 ▾ 难度: ▾

1. That represents the **mouse** cursor. 它代表鼠标光标。
<http://msdn2.microsoft.com/en-us/library/system.windows.forms.cursor.aspx>

2. The user presses the **mouse** button. 用户按鼠标按钮。
<http://msdn2.microsoft.com/en-us/library/ms171542.aspx>

3. The user releases the **mouse** button. 用户释放鼠标按钮。



首页

mouse

例句

类别

- 全部
全部
口语
书面语
标题
技术

来源

全部

难度

全部

逐词释义

1. The **mouse** down the oak-tree.
耗子刷地下了橡树。
2. The **mouse** squeaked in the corner.
墙角处有**老鼠**的吱吱声。
3. The **mouse** whisked into its hole.
这只**老鼠**急速地跑进洞去。
4. They can grow in all **mouse** strain.
它们可以在所有的小**鼠**株内生长。
5. The **mouse** looked at her rather inquisitively.
那只**耗子**用疑问的眼光看看她。
6. The girl shrieked when she saw the **mouse**.
那个姑娘一见**耗子**就吓得尖叫起来。
7. The **mouse** ran nimbly up the horse's leg.
那**耗子**灵活地顺着马腿爬上去。
8. "Ahem!" said the **Mouse** with an important air.



首页

mouse

例句

类别: 技术

来源: 全部

难度: 全部

逐词释义

1. That represents the **mouse** cursor.

它代表**鼠标**光标。

<http://msdn2.microsoft.com/en-us/library/system.windows.forms.cursor.current.aspx>

2. The user presses the **mouse** button.

用户按**鼠标**按钮。

<http://msdn2.microsoft.com/en-us/library/ms171542.aspx>

3. The user releases the **mouse** button.

用户释放**鼠标**按钮。

<http://msdn2.microsoft.com/en-us/library/ms171542.aspx>

4. Release the **mouse** button to drop the text.

释放**鼠标**按钮以放下文本。

[http://msdn2.microsoft.com/en-us/library/kddy153a\(VS.80\).aspx](http://msdn2.microsoft.com/en-us/library/kddy153a(VS.80).aspx)

5. A Microsoft **mouse** or compatible pointing device is required.

需要Microsoft**鼠标**或兼容的指点设备。

<http://msdn2.microsoft.com/en-us/library/ms143506.aspx>

6. The controllable **mouse** carries on the image programming demonstration.

可控**鼠标**进行图象编程示例

<http://www.88pi.com/soft/codelist.asp?page=324>

Learn Contextual Usage with Word Alignment

每日一句

If we walk slower than the others we shall lag behind them. 

如果我们比其他人走得慢，我们就会落在他们后面。



Learn Contextual Usage with Word Alignment

每日一句

If we walk slower than the others we shall lag behind them. 

如果我们比其他人走得慢，我们就会落在他们后面。



Learn Contextual Usage with Word Alignment

每日一句

If we walk slower than the others we shall lag behind them. 

如果我们比其他人走得慢，我们就会落在他们后面。



Hints of Easy-Confused Words

The screenshot shows a Microsoft Internet Explorer browser window. The address bar contains 'fiziks'. The main content area displays the search results for 'fiziks' on the Engkoo website. The results are categorized into '音近词' (Homophones) and '形近词' (Homographs). The homophone list includes physics, physical, physique, phoenix, and Felix. The homograph list includes fizzes, filix, faikes, finis, and fitios.

您要找的是不是：

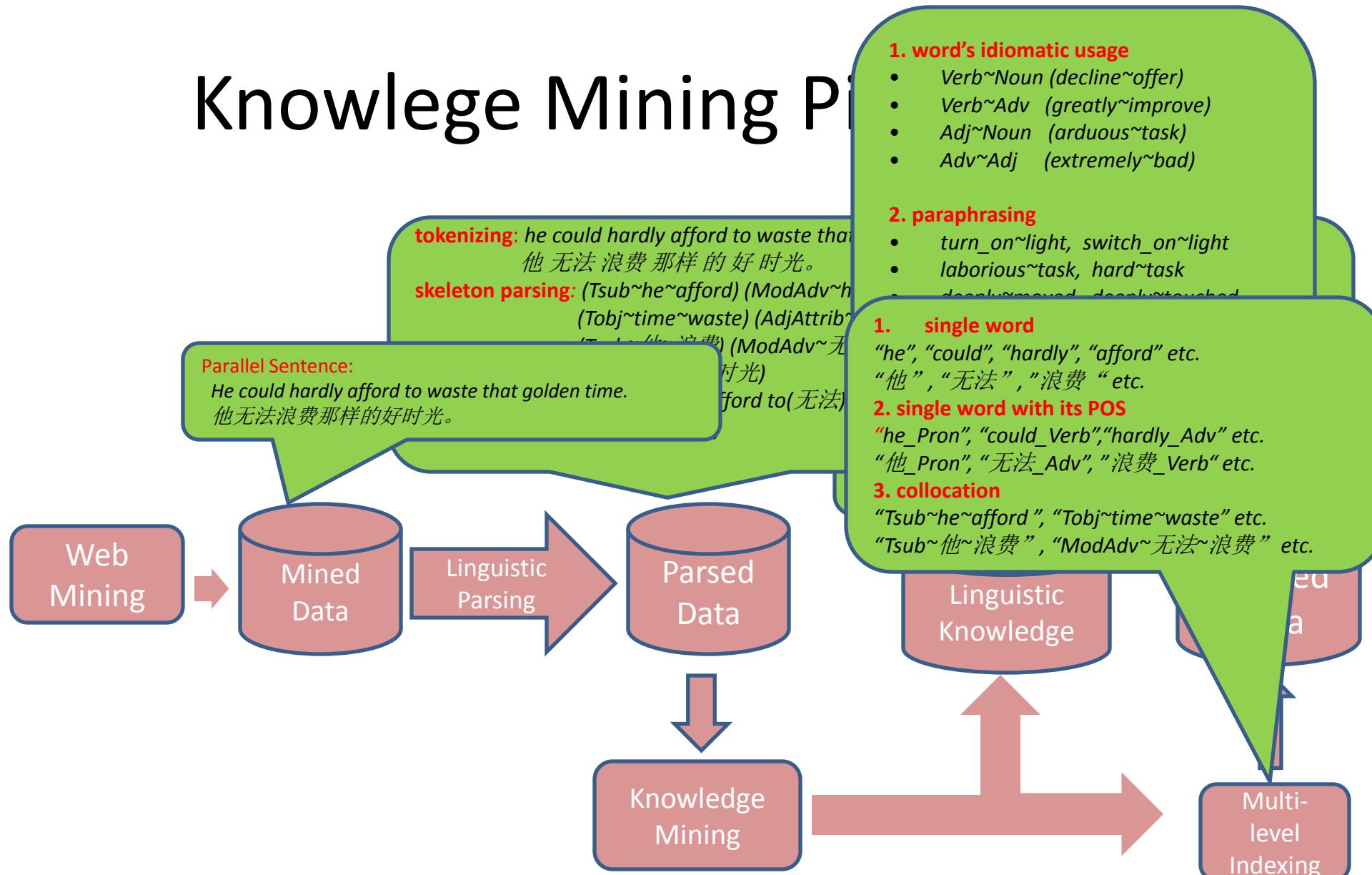
♪ 音近词

- ♫: physics
- ♫: physical
- ♫: physique
- ♫: phoenix
- ♫: Felix

★ 形近词

- fizzes
- filix
- faikes
- finis
- fitios

Knowledge Mining Process



WSJ Asian Innovation Award: Reader's Choice Award



Bilingual Data (BD) in Engkoo

The screenshot shows the Microsoft Engkoo BETA interface. The search bar at the top contains the Chinese character '奥巴马'. Below the search bar, the results page displays the search term '奥巴马' again, followed by the message '词形变化 (无相关结果)'. A blue callout bubble labeled 'Bilingual terms' points to the search bar area. In the middle section, there's a red box highlighting a network definition for 'barack obama' with a link to <http://bbs.haozhai.com/viewthread.php?tid=94027>. Another blue callout bubble labeled 'Bilingual sentences' points to this definition. At the bottom, a large red box encloses two examples of bilingual sentences. The first example is: 'The man handed me a pamphlet. "Mr. Obama, I want you to know that I agree with a lot of what you have to say." 那个男人递给了我一本小册子并说道, "奥巴马先生,我想告诉你,我并不完全同意你要说的。"' with a link to http://groups.google.com/group/mimiqiao/browse_thread/thread/19080e4523a27d15/4d7a471.... The second example is: 'Obama, Romney to Formally Enter US Presidential Race 奥巴马和罗姆尼正式宣布参加下届总统竞选' with a link to <http://board.verycd.com/t460479.html>.

Microsoft Engkoo BETA 奥巴马

首页 奥巴马 奥巴马

奥巴马

词形变化 (无相关结果)

网络释义:

1. barack obama
<http://bbs.haozhai.com/viewthread.php?tid=94027>

例句 类别: 全部 来源: 全部 难度: 全部 逐词释义

1. The man handed me a pamphlet. "Mr. Obama, I want you to know that I agree with a lot of what you have to say." 那个男人递给了我一本小册子并说道, "奥巴马先生,我想告诉你,我并不完全同意你要说的。"
http://groups.google.com/group/mimiqiao/browse_thread/thread/19080e4523a27d15/4d7a471...
2. Obama, Romney to Formally Enter US Presidential Race 奥巴马和罗姆尼正式宣布参加下届总统竞选
<http://board.verycd.com/t460479.html>

Mining BD from the Web

- From parallel pages
 - (Shi et al., ACL-06)
- From search results
 - (Jiang et al., IJCAI-07)
- Using parenthesis pattern
- From collective bilingual pages
 - (Jiang et al., ACL-09)

Parallel Page

<http://support.microsoft.com/kb/914962/en-us>

<http://support.microsoft.com/kb/914962/zh-cn>

This article lists problems that are fixed in Microsoft Windows Server 2003 Service Pack 2 (SP2). Service packs are cumulative. This means that the problems that are fixed in a service pack are also fixed in later service packs.

[↑ Back to the top](#)

本文列出了 Microsoft Windows Server 2003 Service Pack 2 (SP 2) 中修复的问题。Service Pack 具有累积性。 这意味着在一个 Service Pack 中修复该问题还会在以后的 Service Pack 中修复。

[↑ 回到顶端](#)

This article contains a list of Microsoft Knowledge Base (KB) updates that are contained in Microsoft Windows Server 2003 Service Pack 2 (SP2).

This article is primarily intended to help IT professionals and support and maintenance personnel help and maintaining a company's computer systems. If you have questions about Windows Server 2003 SP2, the following links provide answers:

<http://support.microsoft.com/ph/3198>

本文描述的修复程序和 Microsoft Windows Server 2003 Service Pack 2 (SP 2) 中包含的更新的 Microsoft 知识库 (KB) 文章列表。

本文主要被为了帮助 IT 专业人员和支持和维护公司已经计算机系统中的公司 helpdesks。 所有其他人对有疑问 Windows Server 2003 SP 2，下面的 Microsoft Web 站点可能是一个更好的站点，以查找答案：

<http://support.microsoft.com/ph/3198>

Mining BD from Parallel Pages

Source



新渡輪 FIRST FERRY 
Member of Chow Tai Fook Enterprises & NWS Holdings

Login name: Password: Login

中文 

Ferry Schedule and Fare Table 

Our Story

New World First Ferry Services Limited is jointly owned by Chow Tai Fook Enterprises Limited and [NWS Holdings Limited](#) ("NWS Holdings"). First Ferry was established in November 1999, and commenced its service on 15 January 2000, and operates a total of eight outlying and inner harbour ferry routes at present.

Under the same umbrella, New World First Ferry Servicos Maritimos (Macau) Limitada [First Ferry (Macau)] has operated ferry service between Macau and Tsim Sha Tsui, Hong Kong since January 2000.

"Customer First and Foremost" is our corporate mission, First Ferry strives to provide a safe and comfortable journey to the commuting public. Quality people and a reliable vessel fleet are essences in our quest for quality ferry service.



Our Story

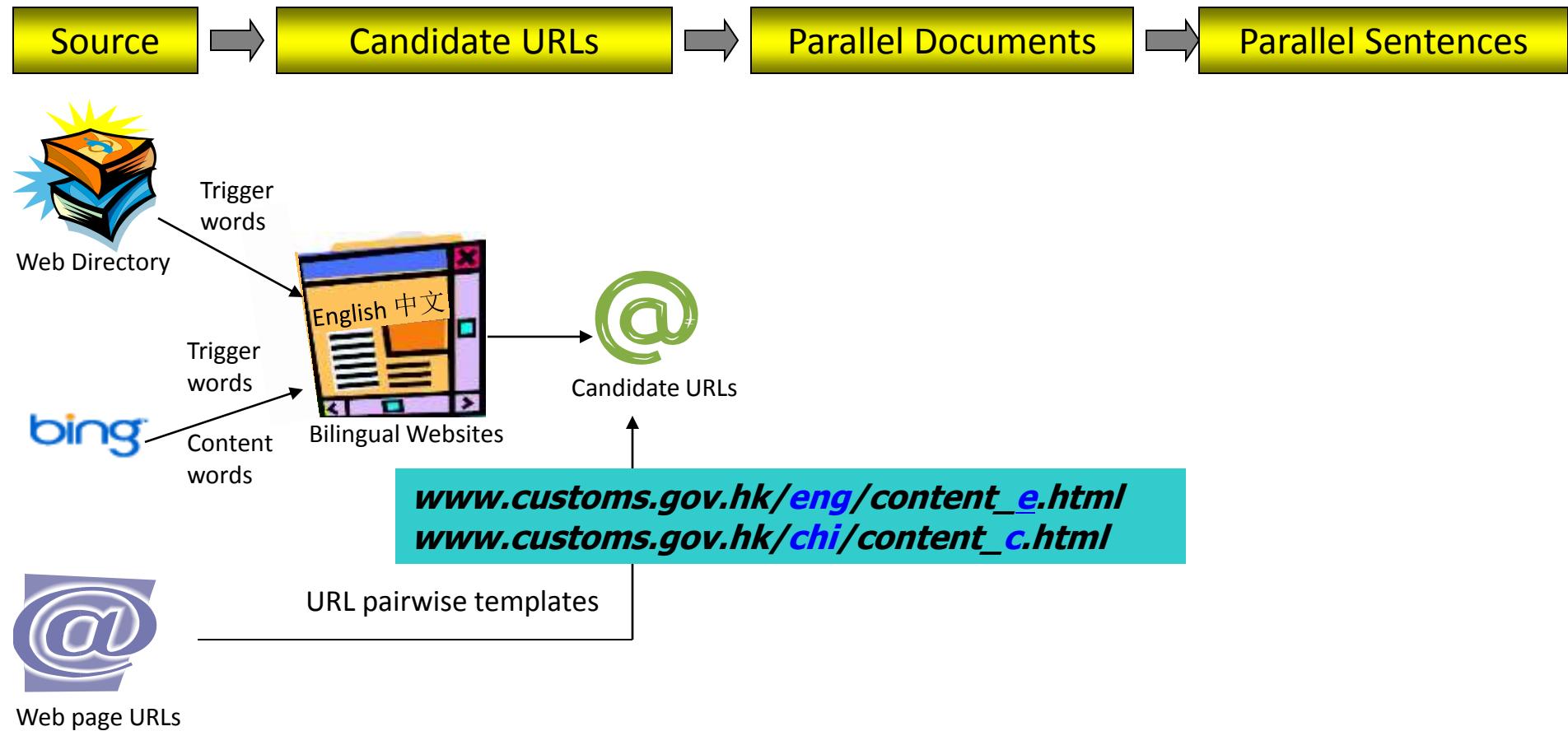
New World First Ferry Services Limited is jointly owned by Chow Tai Fook Enterprises Limited and [NWS Holdings Limited](#) ("NWS Holdings"). First Ferry was established in November 1999, and commenced its service on 15 January 2000, and operates a total of eight outlying and inner harbour ferry routes at present.

Under the same umbrella, New World First Ferry Servicos Maritimos (Macau) Limitada [First Ferry (Macau)] has operated ferry service between Macau and Tsim Sha Tsui, Hong Kong since January 2000.

"Customer First and Foremost" is our corporate mission, First Ferry strives to provide a safe and comfortable journey to the commuting public. Quality people and a reliable vessel fleet are essences in our quest for quality ferry service.

... Airport Codes | Consolidated Airfare Quotes | Travel ... to find your way back as it may be quite different from the Chinese version.
www.travel360degrees.com/article0022.htm Cached page

Mining BD from Parallel Pages



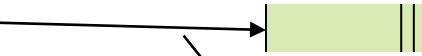
Mining BD from Parallel Pages

COURT OF FINAL APPEAL

The Court of Final Appeal is the highest appellate court in the Hong Kong Special Administrative Region. It has jurisdiction in respect of matters conferred on it by the [Hong Kong Court of Final Appeal Ordinance](#), Cap. 484 and by any other law.

It hears appeals on civil and criminal matters from the High Court (the Court of Appeal and the First Instance).

Trigger



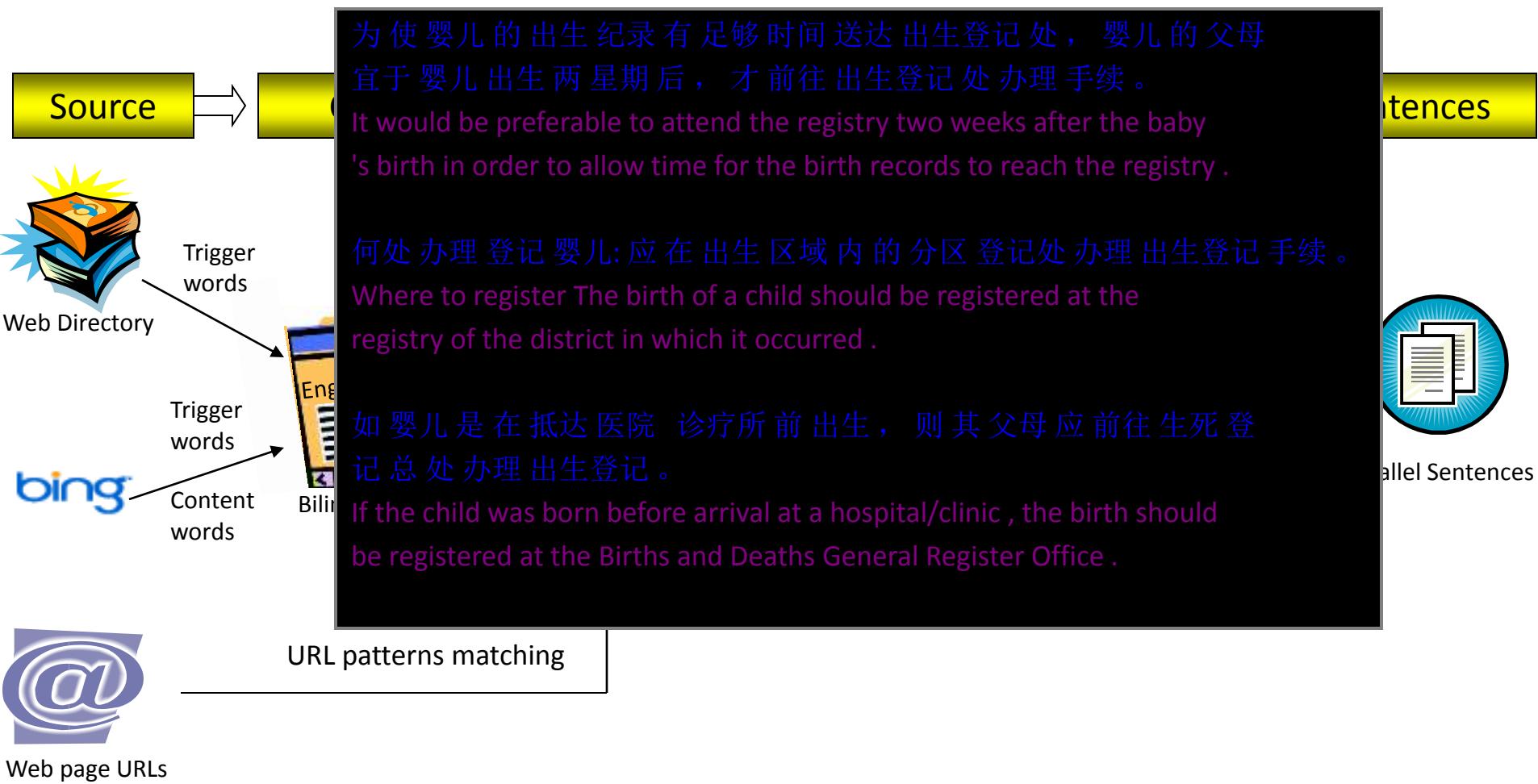
Illegal Sentences

终审法院

终审法院是香港特别行政区最高的上诉法院，根据香港法例 第484章 [《香港终审法院条例》](#) 赋予的权力，处理针对高等法院（上诉法庭及原讼法庭）的判决而作出的上诉及有关事项。

Parallel hyperlinks

Mining BD from Parallel Pages



Mining BD from Search Results

The screenshot shows a Bing search results page for the query "Microsoft Research Asia". A blue callout bubble points to the "Search for Chinese pages" button in the top right corner of the search bar.

Search Bar: Microsoft Research Asia

Filter Options: Show all, Only English, Your settings

Results Count: 1-10 of 26,800 results · Advanced

Result 1: [Microsoft Research Asia](#) · [Translate this page](#)
igroup.msra.cn · [Mark as spam](#)

Result 2: [微软亚洲研究院](#) · [Translate this page](#)
Microsoft Research Asia English Website | 网站地图 最新发布 创新成果 热门下载
www.msra.cn · [Cached page](#) · [Mark as spam](#)

Result 3: [微软亚洲研究院](#) · [Translate this page](#)
微软亚洲研究院 Microsoft Research Asia
www.msra.cn/rmc · [Cached page](#) · [Mark as spam](#)

Result 4: [Microsoft Research Asia Theory Workshop](#) · [Translate this page](#)
Organizers Andrej Bogdanov (Tsinghua University, andrejb@tsinghua.edu.cn) Dr. Wei Chen (MSR Asia, weic@microsoft.com) Cynthia Dwork (MSR Silicon Valley, dwork@microsoft.com) Shanghua Teng (Boston University / MSR Asia, steng@cs.bu.edu) The goal of the workshop is to enhance the communication between the nascent ...
[research.microsoft.com/en-us/um/beijing/events/theory...](#) · [Cached page](#) · [Mark as spam](#)

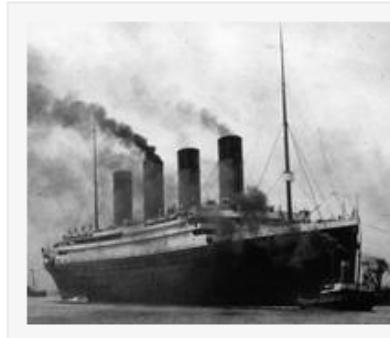
Result 5: [Microsoft Research Asia Family Day - Windows Live](#) · [Translate this page](#)
June 04 Microsoft Research Asia Family Day 上个月的wine down,公司把它推迟到今天，作为一次儿童节聚会，邀请公司的成员带上自己的家属，特别是那些有孩子的员工，让孩子们和爸爸妈妈一起开心一下：）记个流水账//
[cloudsun.spaces.live.com/blog/cns!927F11B12223353A!260.entry](#) · [Cached page](#) · [Mark as spam](#)

Mining BD using Parenthesis Pattern

泰坦尼克号邮船简介

[编辑本段]

英国皇家邮船泰坦尼克号(RMS Titanic)是奥林匹克级邮轮的第二艘邮船，20世纪初，由英国白星航运公司(White Star Line)制造的一艘巨大豪华客轮。由位于爱尔兰**贝尔法斯特**(Belfast)的哈兰德与沃尔夫(Harland and Wolff)造船厂兴建。泰坦尼克号是当时世界上最大的豪华客轮，被称为是“永不沉没的船”或是“梦幻之船”。泰坦尼克号共耗资7500万**英镑**，**吨位**46328吨，长882.9**英尺**，宽92.5英尺，从龙骨到四个大烟囱的顶端有175英尺，高度相当于11层楼。是当时一流的超级豪华巨轮。计划与姐妹船**奥林匹克号**(RMS Olympic)和**不列颠尼克号**(RMS Britannic)一道为英国白星航运公司的乘客们提供快速且舒适的跨**大西洋**旅行。

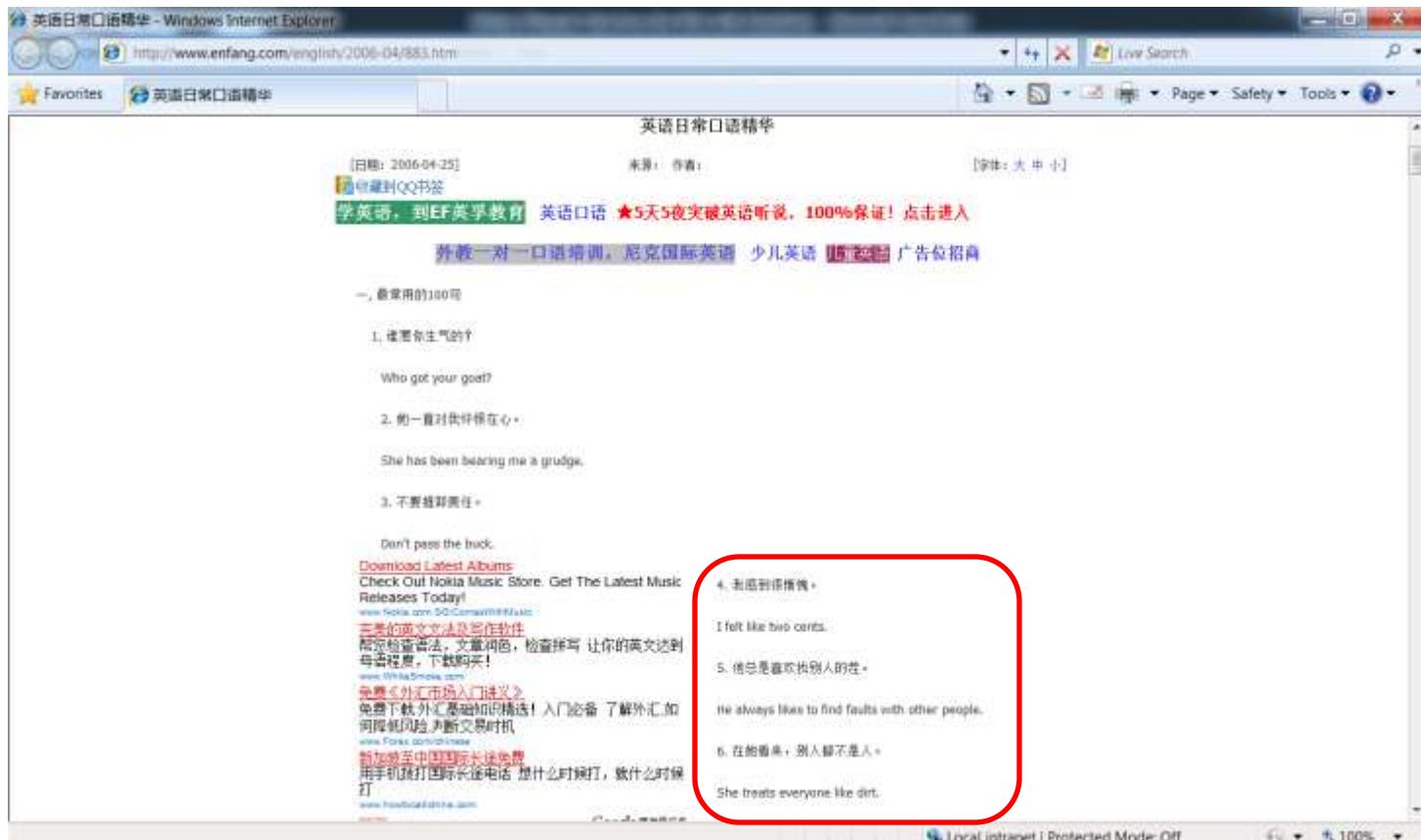


1912年4月10日，泰坦尼克号从英国**南安普敦**(Southampton)出发，途经**法国**瑟堡-奥克特维尔(Cherbourg-Octeville)以及**爱尔兰**昆士敦(Queenstown)，计划中的目的地为**美国**(New York)，开始了这艘“梦幻客轮”的**处女航**。4月14日晚11点40分，泰坦尼克号在北大西洋撞上**冰山**(大约在41°43'55.66"N 49°56'45.02"W附近)，两小时四十分钟后，4月15日凌晨2点20分沉没，由于缺少足够的救生艇，1500人葬生海底，造成了当时在和平时期最严重的一次航海事故，也是迄今为止最为人所知的一次海难。电影《泰坦尼克号》就是根据这一真实海

难而改编。

Mining BD from Collective Bilingual Pages

- Collective Bilingual Page (CBP): pages in which bilingual data **appear collectively** and are **formatted consistently**



Mining BD from Collective Bilingual Pages

- Collective I
collectively

4. 我感到很惭愧。

I felt like two cents.

5. 他总是喜欢找别人的茬。

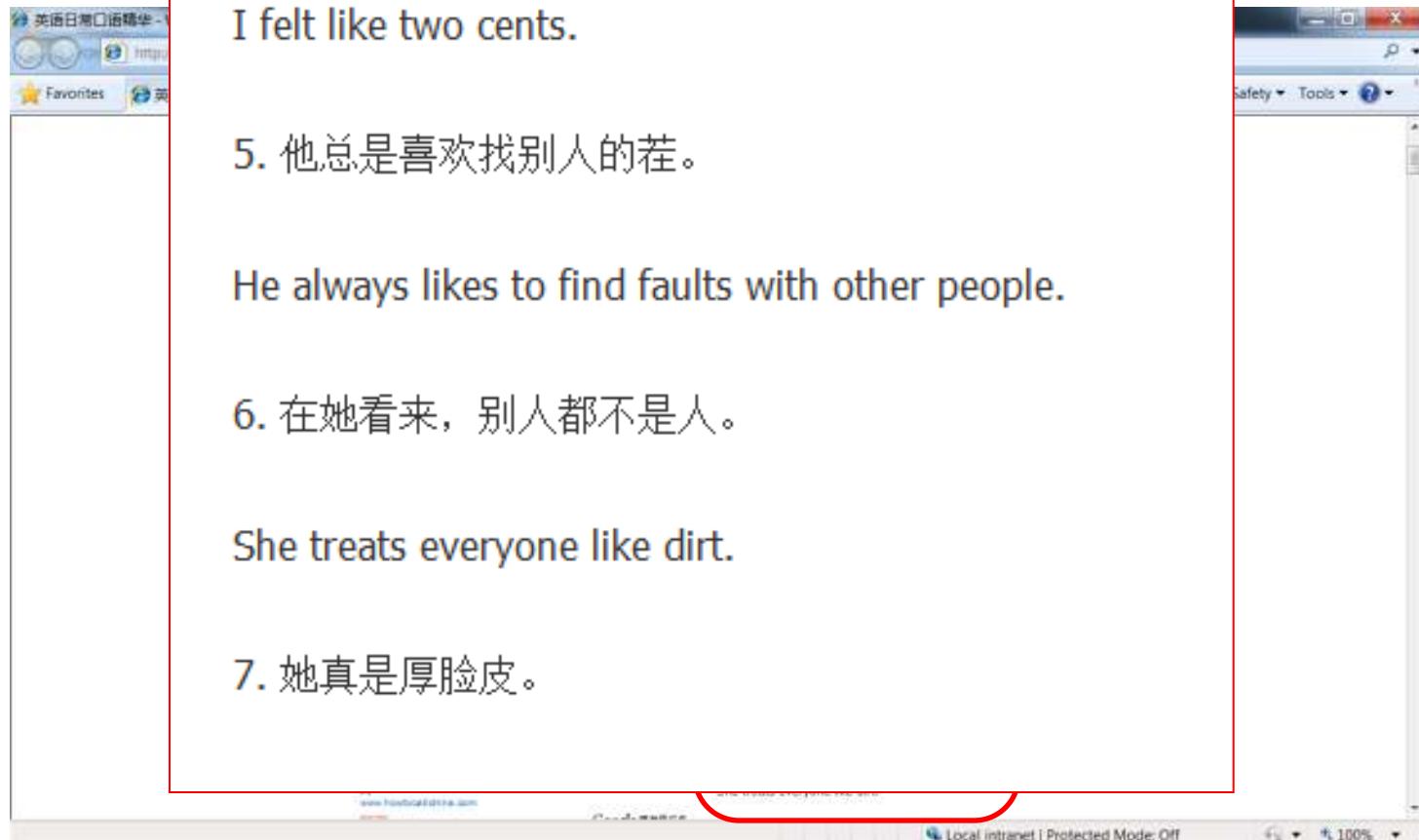
He always likes to find faults with other people.

6. 在她看来，别人都不是人。

She treats everyone like dirt.

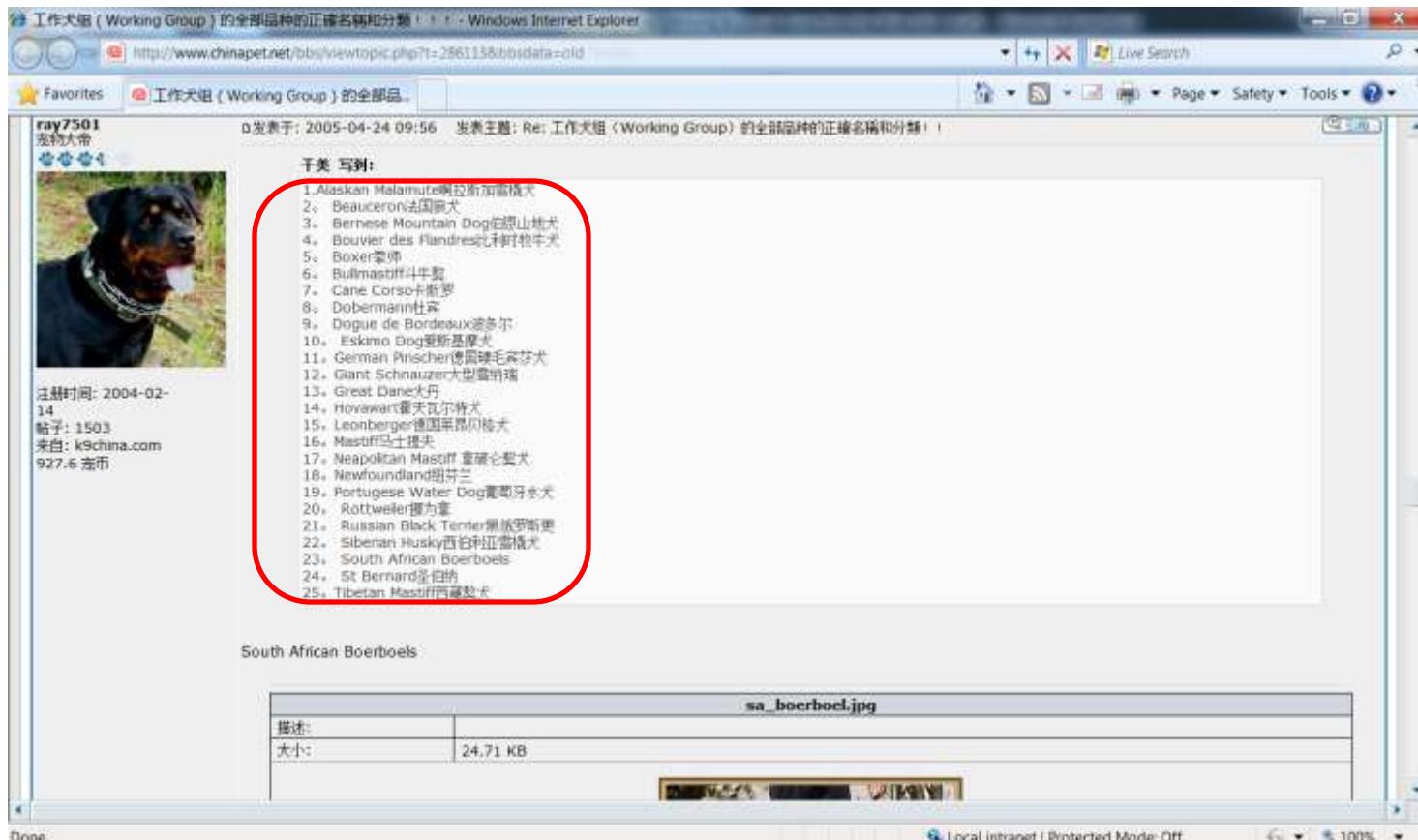
7. 她真是厚脸皮。

appear



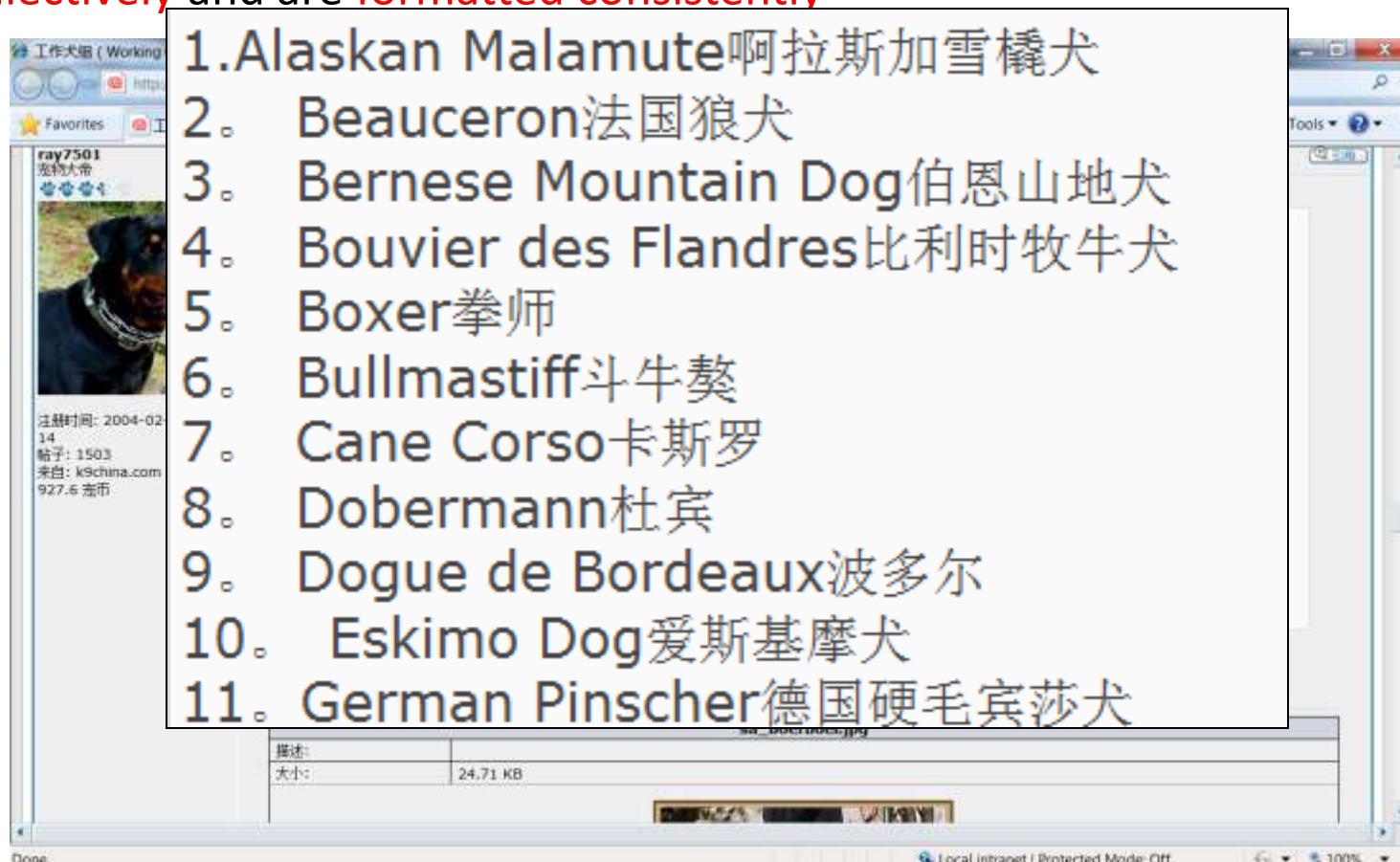
Mining BD from Collective Bilingual Pages

- Collective Bilingual Page (CBP): pages in which bilingual data **appear collectively** and are **formatted consistently**



Mining BD from Collective Bilingual Pages

- Collective Bilingual Page (CBP): pages in which bilingual data **appear collectively** and are **formatted consistently**



Mining Bilingual Data from CBPs

- Identification of seed translation pairs

1. Alaskan Malamute 哈拉斯加雪橇犬

2. Beauceron 法国狼犬

3. Bernese Mountain Dog 伯恩山地犬

4. Bouvier des Flandres 比利时牧牛犬

5. Boxer 拳师

6. Bullmastiff 牛头梗

7. Cane Corso

8. Dobermann

9. Dogue de Bordeaux

10. Eskimo Dog

11. German Pinscher

Bernese Mountain Dog

伯恩 山地 犬

Mining Bilingual Data from CBPs

- Pattern: <number> <period> <English-term> <Chinese-term>

1. Alaskan Malamute 哥拉斯加雪橇犬
2. Beauceron 法国狼犬
3. Bernese Mountain Dog 伯恩山地犬
4. Bouvier des Flandres 比利时牧牛犬
5. Boxer 拳师
6. Bullmastiff 斗牛獒
7. Cane Corso 卡斯罗
8. Dobermann 杜宾
9. Dogue de Bordeaux 波多尔
10. Eskimo Dog 爱斯基摩犬
11. German Pinscher 德国硬毛宾莎犬

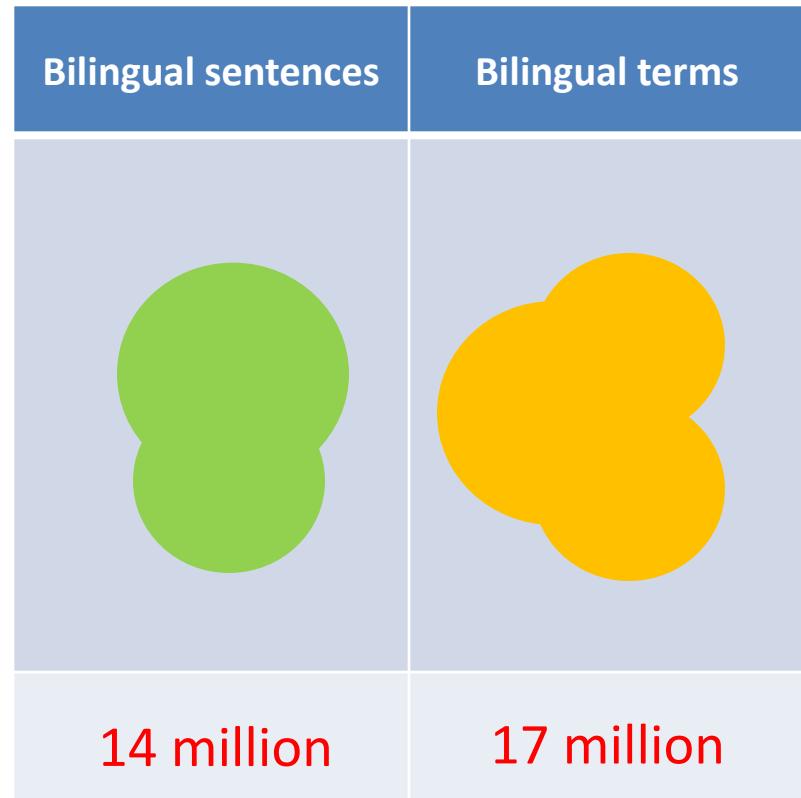
Mining Bilingual Data from CBPs

- Pattern: <number> <period> <English-term> <Chinese-term>

1.	Alaskan Malamute	阿拉斯加雪橇犬
2.	Beauceron	法国狼犬
3.	Bernese Mountain Dog	伯恩山地犬
4.	Bouvier des Flandres	比利时牧牛犬
5.	Boxer	拳师
6.	Bullmastiff	斗牛獒
7.	Cane Corso	卡斯罗
8.	Dobermann	杜宾
9.	Dogue de Bordeaux	波多尔
10.	Eskimo Dog	爱斯基摩犬
11.	German Pinscher	德国硬毛宾莎犬

Results of BD Mining

- From parallel pages
 - (Shi et al., ACL-06)
 - About 6 million bilingual sentences
- From search results
 - (Jiang et al., IJCAI-07)
 - About 6 million bilingual terms
- Using parenthesis pattern
 - About 6 million bilingual terms
- From collective bilingual pages
 - (Jiang et al., ACL-09)
 - About 10.5 million bilingual sentences and 10.1 million bilingual terms



Summary

- Mine massive translation knowledge from the web to build a huge lexicon
- Build a search engine to provide dictionary explanation and example sentences
- Learn a SMT engine with the mined example sentences for translation references
- Apply NLP based search to enable advanced search
- Learn pronunciation with TTS
- Language-independent technologies extensible to other language pairs



Internet Advertising

Online Advertising

*A cross-discipline of
information retrieval, machine learning, and economics*

Microsoft
Research



The Speaker

- Tie-Yan Liu
 - Lead Researcher, Microsoft Research Asia
 - Co-author of 70+ papers in SIGIR, WWW, NIPS, ICML, KDD, etc; co-author of SIGIR Best Student Paper (2008) and JVCIR Most Cited Paper Award (2004~2006); author of two books (Now publisher and Springer).
 - Associate editor, ACM Transactions on Information System; Editorial board member, IR Journal and ISRN Artificial Intelligence.
 - PC Chair of RIAO 2010, Area Chair of SIGIR (2008~2011), Track Chair of WWW (2011).
 - Senior member of the IEEE and the CCF.

Advertising

- Advertising is a form of communication intended to persuade an audience (viewers, readers, or listeners) to purchase or take action upon products, ideals, or services.
- Venues
 - Magazines,
 - Billboards,
 - Newspapers,
 - Handbills,
 - TV, etc.



Advertising

- Advertisers are buying attention!
 - Advertisers want the attention of certain people
 - People are only open to certain ads
(Herbert Simon, 1971)
- "Half the money I spend on advertising is wasted; the trouble is I don't know which half."
(John Wanamaker)



Advertising is Not Easy!



Online Advertising

- Advertising in the context of online services
 - Users: online users with search and browsing behaviors
 - Publishers: search engines and websites
 - Advertisements: ads copies and landing pages

Online Advertising

- First online advertisement
 - In 1994. 10. 24.
 - Place on the website of HotWired.
 - Created for Modem Media/AT&T by TANGENT Design/Communications of Westport, CT.
 - A staggering CTR of 42%!



Why Online

- Classical:
 - High cost per venue (\$3Mil for a Super Bowl TV ad)
 - No personalization
 - Targeting by the wisdom of people
 - Hard to measure ROI
- Online (almost the exact opposite):
 - Billions of opportunities and tiny cost per opportunity
 - Personalizable
 - Targeting by algorithms
 - Much more quantifiable (Impression, CTR, CPC, ...)

Online Advertising vs. Offline Advertising

Advertising Spending Worldwide, by Media, 2007-2011 (millions)

	2007	2008	2009	2010	2011
TV	\$180,460	\$185,788	\$172,320	\$174,836	\$183,177
Newspapers	\$130,178	\$123,109	\$102,136	\$97,703	\$97,228
Internet	\$40,242	\$49,544	\$54,087	\$60,253	\$68,557
Magazines	\$59,196	\$56,588	\$45,415	\$42,762	\$42,573
Radio	\$38,583	\$37,630	\$33,647	\$33,280	\$34,216
Outdoor	\$31,752	\$31,888	\$29,112	\$29,828	\$31,430
Cinema	\$2,268	\$2,377	\$2,180	\$2,274	\$2,422
Total	\$482,680	\$486,924	\$438,896	\$440,936	\$459,603

Note: currency conversion at 2008 average rates

Source: ZenithOptimedia as cited in press release, October 19, 2009

Types of Online Advertising

- Paid Search
 - Ads driven by search query issued by user
- Display Ads
 - Graphical ads primarily driven by behavioral targeting
- Contextual Ads
 - Ads driven by content of a page
- Mobile Ads
 - Mix of above types with focus on user profile and location
- Gaming Ads
 - Currently not focused deeply on matching

Paid Search

Web Images Videos Shopping News Maps More | MSN Hotmail



Web

new york hotels



Web Visual Search Local Listings News Hotels Wikipedia

RELATED SEARCHES

[Manhattan Hotels](#)
[Cheap Hotels New York](#)
[W Hotel New York](#)
[Roosevelt Hotel New York](#)

[New York City Hotel Guide](#)

[New York Cheap Hotels Directory](#)

[Waldorf Astoria](#)
[Sheraton New York Hotel Towers](#)

SEARCH HISTORY

Search more to see your history

See all

Clear all · Turn off

ALL RESULTS

1-10 of 126.000.000 results - [Advanced](#)

540 New York Hotels - [www.Expedia.com/NewYorkCity](#)

Book your Hotels in New York on Expedia® - The #1 Site in Travel.

Marriott New York Hotels - [Marriott.com/NewYork](#)

Book Marriott's Best Rate Guarantee & Earn Rewards at Official Site.

250 Hotels in New York - [Booking.com/Hotels-New-York-NY](#)

Hotels online in New York. Book online now. Pay at the hotel.

Hotels in New York City - [www.CandlewoodSuites.com](#)

Pet Friendly Hotel In Times Square Unbeatable Location & Spacious Room

Sponsored sites

Sponsored sites

New York City Hotel

A Botique Hotel With Personality.
Unique Features - Weekends From \$239

[HotelRogerWilliams.com](#)

Discount New York Hotels

Book online or call 1 800 522 9991 and save \$\$

[www.hotelconxions.com](#)

Over 540 New York Hotels

Good availability and great rates. Save up to 50% on your reservation!

[www.Hotels.com](#)

New York Hotels

Free Internet & Hot Breakfast Bar.
Book Lowest Rates Direct!

[www.HolidayInnExpress.com](#)

New York Hyatt Hotels

Book Online for Hyatt's Best Hotel Rates on the Internet, Guaranteed.

[www.NewYork.Hyatt.com](#)

See your message here

[Find hotels in New York City \(Manhattan\)](#)



Hotel Gansevoort



The Peninsula New York



The Plaza Hotel



The Benjamin Hotel New York



See Them All

New York Hotels - The Official New York Hotels .com Website

Book hotels at the OFFICIAL New York Hotels .com Website. With the help of thousands of user reviews, BEST PRICE guarantee and SECURE online reservations.

[www.newyorkhotels.com](#) · Cached page

Display Ads

New York, NY CHANGE | SAVE

Click "Save" above and set up to three favorite cities.

Local News & Info

[Headlines](#) [Twitter](#) [Blogs](#)

Top Stories

[Find more national news](#)

Long Island Man Accused of Killing Toddler

An apparently grisly murder case on the Shinnecock Indian Reservation, where police have charged a man with killing a 17-month-old boy.

Sealed With a Kiss: Aquarium Home to New Romance

Shes young, hes old. Shes active, while he likes to relax. Maybe opposites really do attract – at least thats what Bernie the harbor seal, 23, is discovering. The...

Suspect Arrested in Latest SI Hate Crime; Victim Speaks Out

Police have arrested a 15-year-old Staten Island boy for attacking and robbing a Mexican teen early Saturday morning, the latest in a streak of violent hate-crimes plaguing the borough. The boy...

Iraqi Boy Thanks U.S. Doctors Who Saved Him

He is, said one observer, the "poster child" for the innocent victims of the Iraq war. And like so many others, Waad Burkan, 9, still wears the scars from that war. Waad and two friends were...

Commissioner Kelly Heads to Haiti to Help Rebuilding Efforts

New York City Police Commissioner Raymond W. Kelly is soaking up the sun – but he's not on vacation. Kelly headed to Port au Prince, Haiti today where he scheduled to meet with Haitian...

 Wednesday
87°F/74°F

 Thursday
88°F/76°F

Mobile | Make MSN your home page

An advertisement for Ford's Model Year End Sales Event. It features the Ford logo at the top left, followed by the text "model year end" and "SALES EVENT" in large, bold letters. Below this, a central message reads: "During our Model Year End Sales Event, get all that and a great deal." A blue rectangular button below the text contains the white text "ROLL OVER FOR SAVINGS". At the bottom of the ad, there are three images of Ford vehicles: a blue Ford Edge SUV on the left, a black Ford Taurus sedan in the center, and a red Ford Fusion sedan on the right.

Advertisement

LOCAL DEALS

YOUR SHOPPING EXPERIENCE JUST GOT EASIER.

A collage of various promotional images from a Target catalog, including a large central image of a Target bullseye with the text "YOUR DAILY NEEDS OUR LOW PRICES!", and smaller images of products like orange juice, toothpaste, and laundry detergent.

Contextual Ads

advertisement

Refi rates are still low

Lock in a low fixed rate
Free home loan check up
Call now: 1.800.641.2987

Bank of America  Home Loans

The best of MSN Money

Readers' Choice		Editors' Choice
Rating	Top 5 Articles	▪ On Bing: Statutes of limitations on debt
4.52	Is this America's worst investment?	▪ 3 ways you can thwart identity thieves
4.24	Fannie and Freddie must die	▪ The world's 10 most reputable companies
4.14	Is there a statute of limitations on debt?	▪ Bundle.com: Take an inventory of your home
4.08	Game on: Number nonsense continues	▪ Video: 5 tips for keeping your car cool
4.03	Chinese lesson: Better red than Fed	▪ Find a broker and start trading today

View all Top-Rated articles

Digg This Board

advertisement

Sponsored Links

Hot Stock Alert - EHSI
Profit From Healthcare Explosion. New Millionaires Created Today.
EmergingHealthcareSolutionsInc.com

Don't Be Stock Stupid
Follow Former Hedge Fund Manager, Jim Cramer. Start Now For Only \$1.
www.TheStreet.com

\$10,000 Penny Stock Rise?
Secret Penny Stock Alerts FREE: Make up to 3000% in 1 Day From Home
www.KillerPennyStocks.com

Invest in Gold: Goldline®
Gold Delivered to Your Door. Free Investor Kit. Since 1960.
Goldline.com/Gold

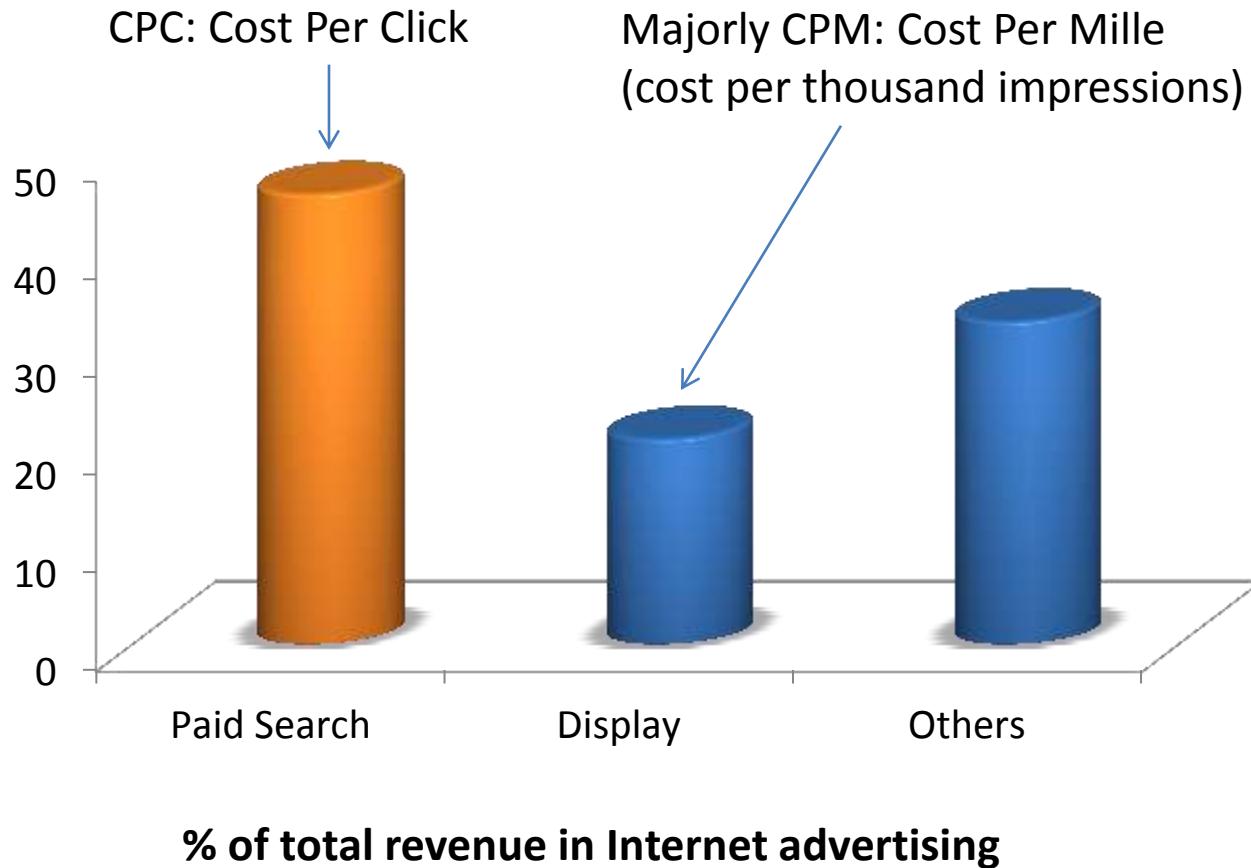
Mobile Ads



Gaming Ads

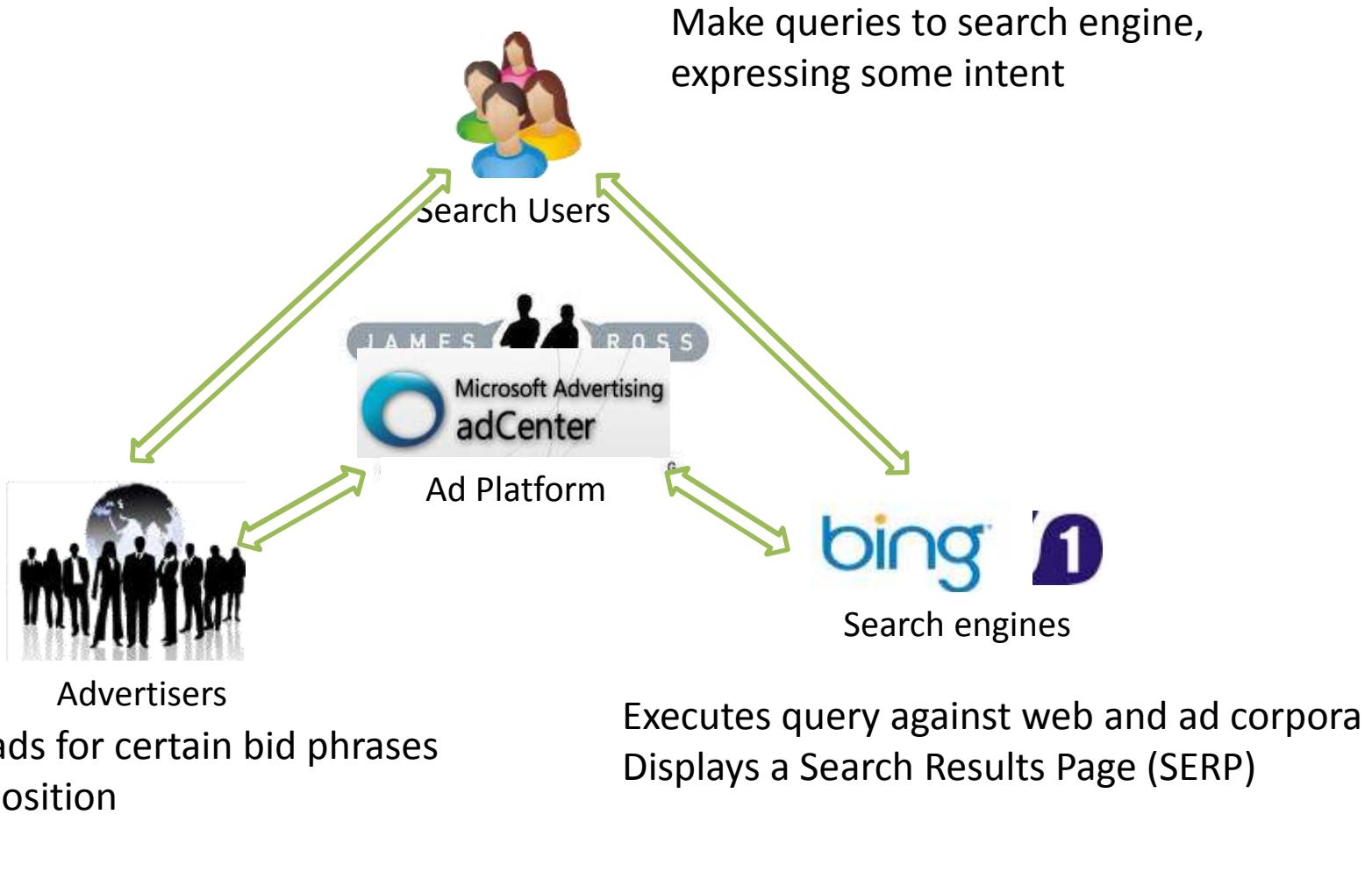


Types of Online Advertising



Paid Search: A Deep Dive

Eco-System in Paid Search



Overall Eco-system Utility

- Advertiser utility
 - Impressions, Clicks, CPC, Average position, Conversions
- Search engine utility
 - Revenue
- User utility
 - Relevance, usefulness

Facts about Paid Search

- Number of Advertisers: 1~2 Million
- Number of Ads: 500 Million ~ 1 Billion
- Number of Bid Keywords: ~1 Billion
- Search RPM: tens of USD
- Average CTR: 2~5%
- Ad Coverage in Search: ~50%
- Google's daily revenue in US: ~45 Million

User's View

Ad Copy

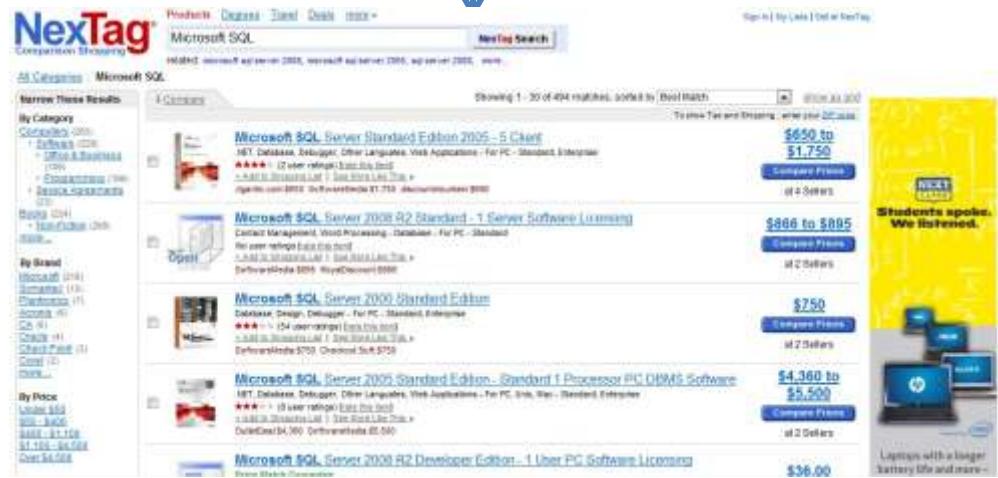


Bid Phrase: Microsoft SQL
Bid: \$1.2

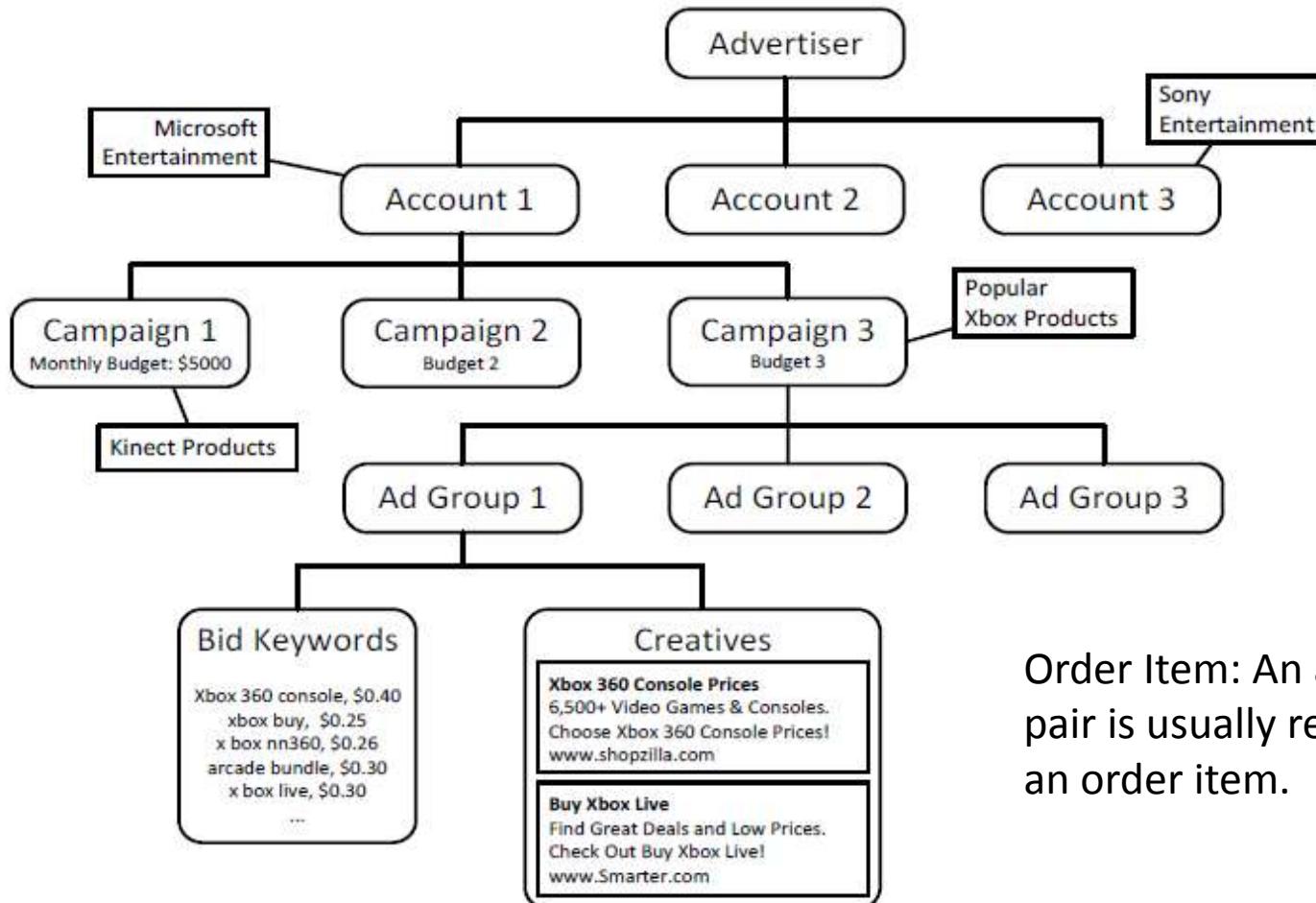
<http://www.nextag.com/Microsoft-SQL/products-html?nxtg=4fdc0a28050c-9BCACF0210B560E9>

Landing
URL

Landing
Page



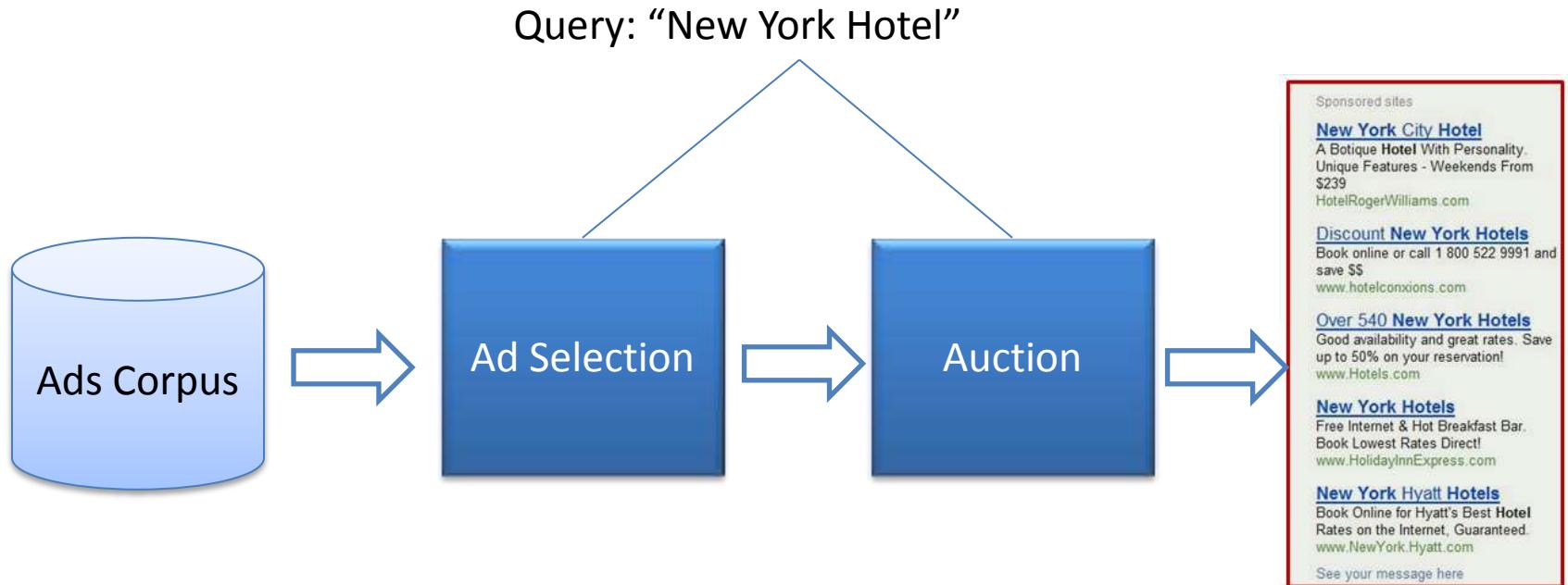
Advertiser's Campaign



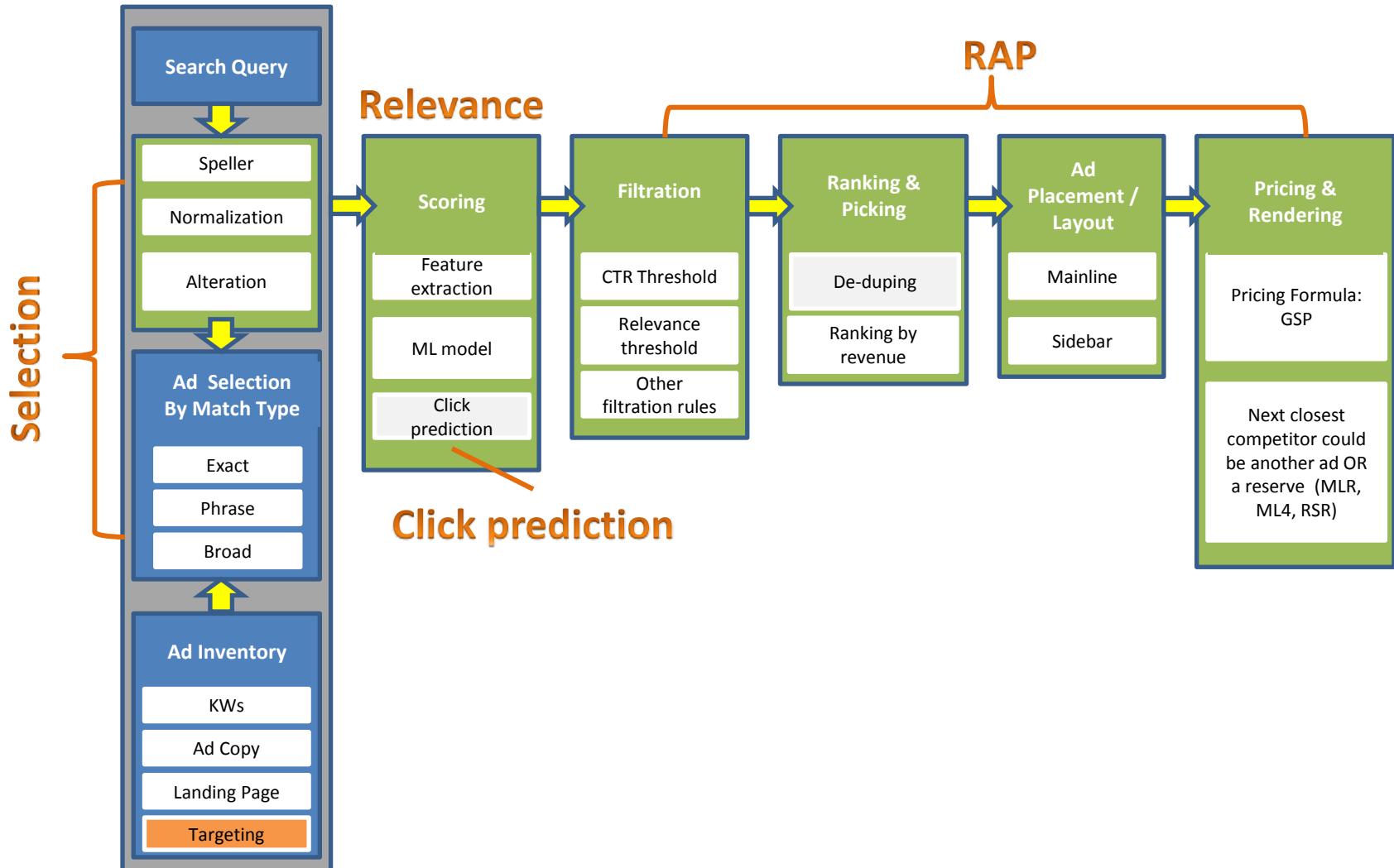
Order Item: An ad-keyword pair is usually referred to as an order item.

Ad Platform's Mechanism

- High-level pipeline



The Architecture



More Details

- Selection
- Relevance
- Click prediction
- RAP
- Evaluation

Selection

- The goal is to find those candidate ads that are potentially related to the query.

Match Type	Definition
Broad	Broad match includes words that are closely related to your keywords. For example, a search query for <i>red carnation</i> might result in your ad being displayed, because adCenter automatically identifies carnation as a type of flower. Use broad match to expose your ads to a wider audience.
Phrase	triggers the display of your ad if the word or words in your keyword appear in a customer's search query—even if other words are present in the typed query. Your keyword <i>red flower</i> would match searches for <i>big red flower</i> and <i>red flower</i> , but not <i>yellow flower</i> or <i>flower red</i> .
Exact	triggers the display of your ad only when the exact word or words in your keyword, in <i>exactly</i> the same order, appear in a customer's query. Your keyword <i>red flower</i> would <i>only</i> match searches for <i>red flower</i> , with no spelling variations. With exact match you might see fewer impressions (An ad that is served to and displayed on a user's browser.) but a higher click-through rate (The ratio of the number of times an ad is clicked to the number of times the ad is displayed.) , because your ad is shown to people who might be more interested in your product.

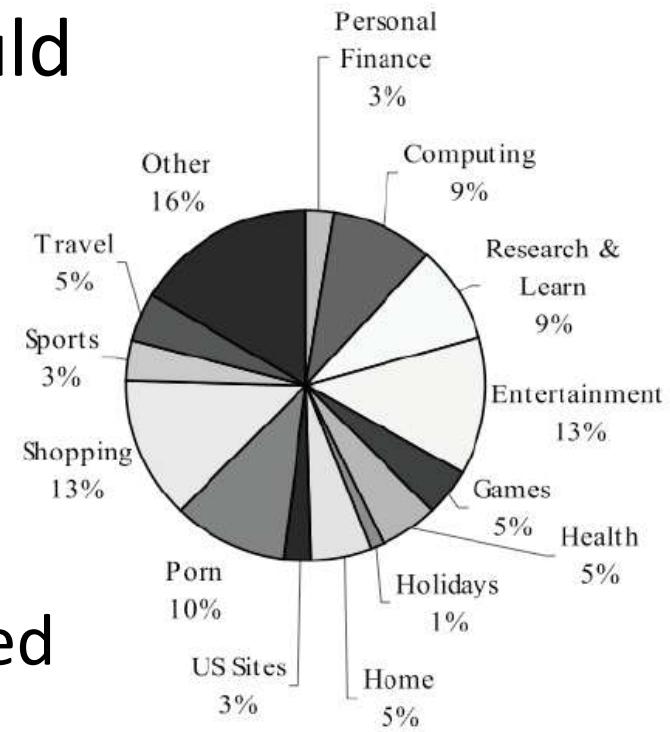
Broad Match

- Query rewriting / alteration: mostly a data-driven approach

Query rewriting technique		Data source
1.	Generating Query Substitutions: Jones et al, in Proc of WWW 2006	query logs (query sessions)
	Using the Wisdom of the Crowds for Keyword Generation: Fuxman et al., In proc of WWW 2004	co-clicks on web search results
2.	Simrank++: Query Rewriting through Link Analysis of the Click Graph: Atoanellis et al., In proc of VLDB 2008	co-clicks on ads
3.	Learning Query Substitutions for Online Advertising: Broder et al. in Proc of ACM SIGIR 2008	query-to-ad similarity
4.	Online Expansion of Rare Queries for Sponsored Search: Broder et al, In Proc. of WWW 2009	query-to-query similarity
5.	Query Word Deletion Prediction: Jones et al., in Proc of ACM SIGIR 2003	query logs

Challenge: Select or Not?

- Showing ads is not always appropriate
- For some queries, we should even not select any ads.
 - To Swing or not to Swing: Learning when (not) to Advertise, CIKM 2008.
 - Estimating Advertisability of Tail Queries for Sponsored Search, SIGIR 2010.



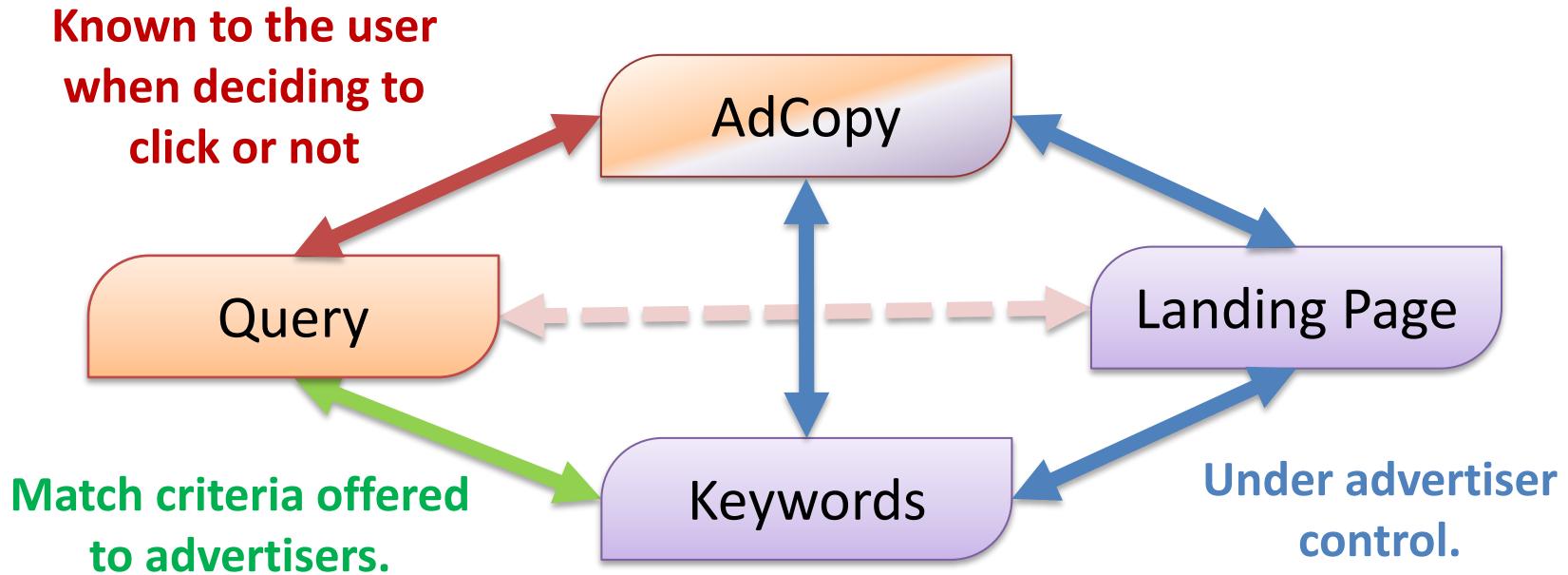
Challenges: Recall of Selection

- Recall is very important in the first step of ad retrieval due to economic considerations
- To guarantee recall, broad match is preferred
 - However, broad match does not work very well for tail queries, since usually tail queries do not have many log / session data.
 - For tail queries, semantic / syntactic analysis may work better.

Relevance

- **Relevance:** is a proxy for total utility:
 - Users – better experience
 - Advertisers – better (more qualified) traffic but possible volume reduction
 - SE gets revenue gain through more clicks but possible revenue loss through lower coverage

Relevance



Ad Copy Relevance: Pre Click User Satisfaction

how to lose a guy in 10 days

Web Videos Images

ALL RESULTS 1-11 of 8,620,000 results · Advanced

Cheap Movie DVDs Sponsored sites

HotMovieSale.com · How To Lose A Guy In 10 Days DVD. Yours For Only \$2.29 + Free S&H!

How to Lose a Guy in 10 Days
www.howtoloseaguymovie.com · Cached page · Mark as spam

How to Lose a Guy in 10 Days (2003) - IMDb
User rating: 6.1/10 · Comedy/Romance · 116 min
Andie Anderson covers the "How To" beat for "Composure" magazine and is assigned to write an article on "How to Lose a Guy in 10 days." They meet in a bar shortly after the bet is ...
www.imdb.com/title/tt0251127 · Cached page · Mark as spam

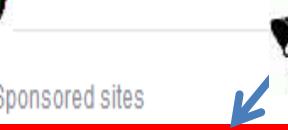
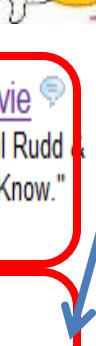
Jack Nicholson New Movie Sponsored sites

With Reese Witherspoon, Paul Rudd & Owen Wilson - "How Do You Know."
HowDoYouKnow-Movie.com

Lose Guy 10 Days

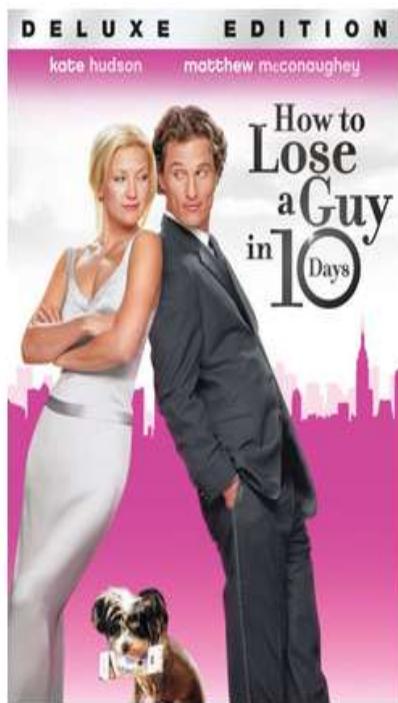
Save up to 35% on top movies.
Qualified orders over \$25 ship free
www.Amazon.com/dvd

See your message here



Ad LP Relevance: Post Click User Satisfaction

[Homepage](#) > [Comedy](#) > [Romance](#) > [How To Lose A Guy In 10 Days](#) »



[Compare](#) [Tell a Friend](#) [Print](#)

How To Lose A Guy In 10 Days [DVD]

Movie Credits: [Kate Hudson](#), [Matthew McConaughey](#), [Michael Michele](#), [Adam Goldberg](#), [Kathryn Hahn](#), [Rebecca Harris](#), [Bill Kotsaris](#), [Thomas Lennon](#), [David Marnivan](#)

Showing 1 - 12 of 23 Results

1.



How to Lose a Guy in 10 Days Starring Kate Hudson, Matthew McConaughey,

Buy new: \$12.99 **\$8.99**

27 new from \$5.34 17 used from \$4.64

Watch It Now: \$9.99 to buy

Get it by **Monday, Dec 13** if you order in the next **19 hours** and choose one-day shipping.

Eligible for **FREE** Super Saver Shipping.

(255)

2.



How to Lose a Guy in 10 Days (Widescreen Edition) Starring Kate Hudson, I

Buy new: \$12.98 **\$9.99**

18 new from \$7.30 242 used from \$0.16

Watch It Now: \$9.99 to buy

Eligible for **FREE** Super Saver Shipping.

Only 1 left in stock - order soon.



Computing Relevance

- Machine learning approach
 - Extract features from query and ad copy/landing pages
 - Use a ML model (e.g., logistic regression) to learn from human relevance judgment
- Information retrieval approach
 - Index all ads copies and their landing pages
 - Use search system to retrieve ads

Challenges: Ad Copy Relevance

- Short text
- Term frequency cannot differentiate ads any more
- No anchor (which is one of the most useful signals in search)
- Both visible and invisible parts

Challenges: LP Relevance

- Classify landing page types for all the ads for 200 queries from the 2005 KDD Cup labeled query set.
 - **I. Category (37.5%):** Landing page captures the broad category of the query
 - **II. Search Transfer (26%):** Land on dynamically generated search results (same q) on the advertiser's web page
 - Product List – search within advertiser's web site
 - Search Aggregation – search over other web sites
 - **III. Home page (25%):** Land on advertiser's home page
 - **IV. Other (11.5%):** Land on promotions and forms

Challenges: Is Relevance Crucial?

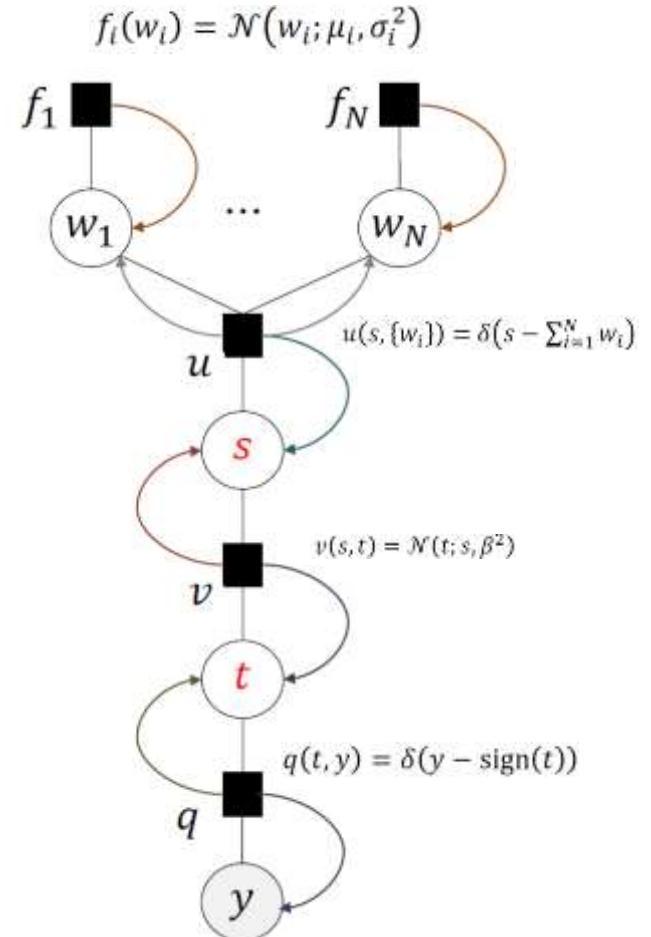
- Sometimes, the user satisfaction on an ad is not just determined by its relevance.
 - Informativeness
 - Attractiveness
 - Real discount
 - Spam...

Click Prediction

- Since advertisers are charged per click, an accurate prediction of click will be very important for ranking and pricing.
- Representative works
 - [ICML10] Web-Scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine
 - [ADKDD09] Data-driven text features for sponsored search click prediction
 - [ADKDD10] Missing Click History in Sponsored Search A Generative Modeling Solution
 - [WWW07] Predicting clicks estimating the click through rate for new ads
 -

Examples

- adPredictor in Bing
 - Bayesian approach
 - A Sparse Linear Probit Regression
- pClick model in Yahoo!
 - A logistic regression based model
 - Leveraging the hierarchical structure of advertiser data to handle sparseness
 - Use GMM and EM to handle missing data



Challenges: Exploration

We only have **training data** for building models valid for these ad impressions.

not shown

No examples available because users never get a chance to click on such impressions.

shown

Cheap training examples are available.

head

So many examples that we can simply memorize the answer

But we need a model that works for all potential impressions (because we need to know which ones should not be shown.)

Challenges: Second-order Effect

- Second-order effect: Result -> Cause -> Result
 - The pClick model is learned from historical data, reflecting user's click behaviors with respect to the previous ranking algorithm
 - After a new model is learned, users will see different ads and their behaviors may change correspondingly.
 - The gradual change of user behaviors will make the offline evaluation on the learned pClick model not valid when put online.

The RAP Problem

Ranking

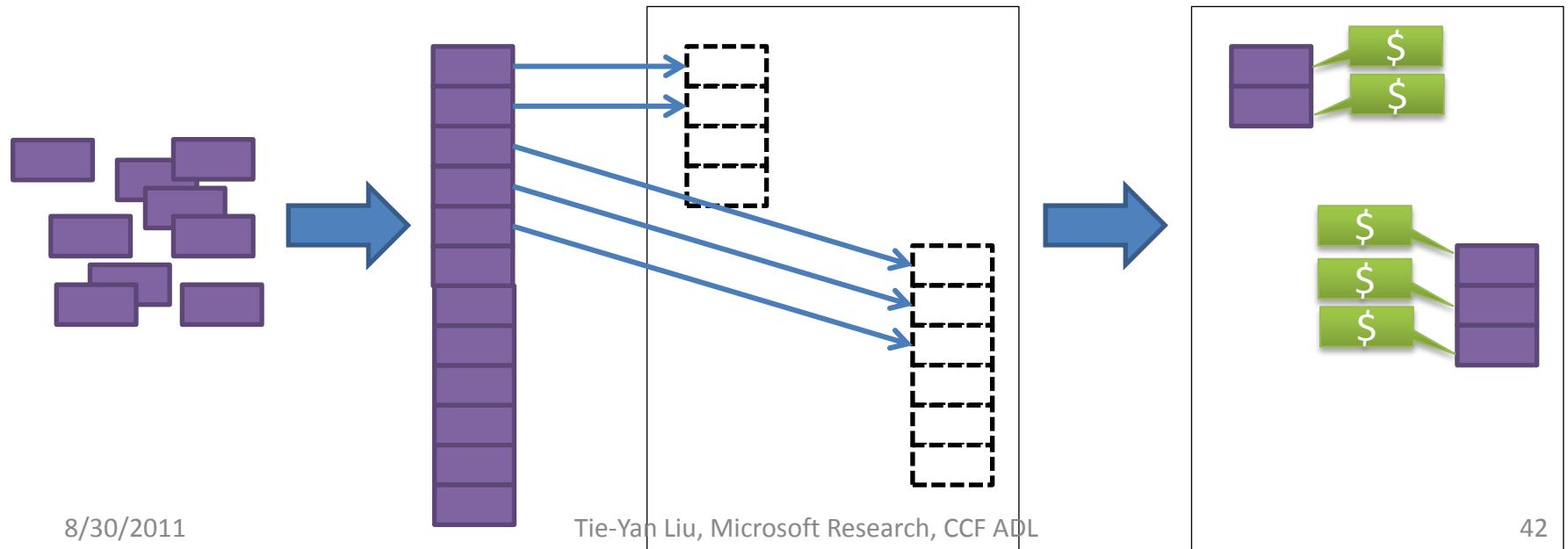
- Sort the candidate ads

Allocation

- Say where on the page each ad goes

Pricing

- Set CPC prices (cost per click)



RAP: Ranking, Allocation, Pricing

- **Ranking:** sort the candidate ads
 - RS = Bid x pClick
- **Allocation:** say where on the page each ad goes.
 - Mainline Reserve
 - Overall Reserve
- **Pricing:** Generalized Second Price

$$CPC_i = Bid_{i+1} \times \frac{pClick_{i+1}}{pClick_i}$$

RAP Mechanism

- Second price auction
 - All advertisers submit their bids privately
 - Ads are ranked according to their bid prices
 - The ad ranked on the first place will win the auction; the owner of the ad will pay a price slightly higher than the price of the second highest bid.
- Generalized second price auction
 - “Pay-per-click” pricing: buy placement but not pay for placement!
 - Ads are ranked in the order of $\text{CTR} * \text{bid}$
 - The winner pays the cost of maintaining his/her position, if clicked

Factors to Consider for a Mechanism

- Incentive compatibility (truthfulness)
- Equilibrium
- Social welfare
- Auctioneer revenue
- Approximation ratio

Incentive Compatibility

- A process is said to be *incentive compatible* if all of the players fare best when they truthfully reveal any private information asked for by the mechanism
- In particular, an auction mechanism is **truthful**, if the dominant strategy for every bidder is to truthfully bid their own value on the item.
- SP is truthful, FP and GSP are not truthful.

Equilibrium

- **Nash equilibrium** = choice of strategies in which each player is assumed to know the equilibrium strategies of the other players, and no player has anything to gain by changing his/her own strategy unilaterally.
- For GSP, there exist many Nash equilibrium strategies.

Social Welfare

- $\sum_i U_i(x_i)$
- The utilities can be
 - Advertiser ROI (true value – cost)
 - Clicks /CTR
 - Impressions
 - Conversion rate
- In most previous work, auction mechanisms are designed to maximize social welfare.

Auctioneer Revenue

- How much does the auctioneer actually get from the auctions?
 - E.g., $\sum_i CTR(x_i)Prc(x_i)$
- In some previous work, mechanisms are designed to optimize auctioneer revenue.
- GSP is not revenue optimal.
 - We can increase RSP by finding a revenue-optimal mechanism.

Approximation Ratio

- The problem of finding the optimal allocation is usually NP-hard.
- An approximation algorithm with polynomial time complexity is desirable.
- Approximation ratio is related to the reduction of the objective function (social welfare and/or auctioneer revenue), introduced by the approximation algorithm.

Challenges: Simple vs Optimal

- We can design an optimal mechanism by considering many different requirements, however, the optimal mechanism might be complex.
- Advertisers prefer simple mechanism that they can understand and can play with.

Challenges: Assumptions

- Many researches on incentive compatibility and equilibrium assume
 - Advertisers are rational
 - Advertisers know their true values
 - Advertisers have good access to other advertisers' information.
- However, in reality, it is often not the case.
 - When these assumptions do not hold, the theoretical results might not be applicable to advertising practices.

Challenges: Beyond Traditional Auctions

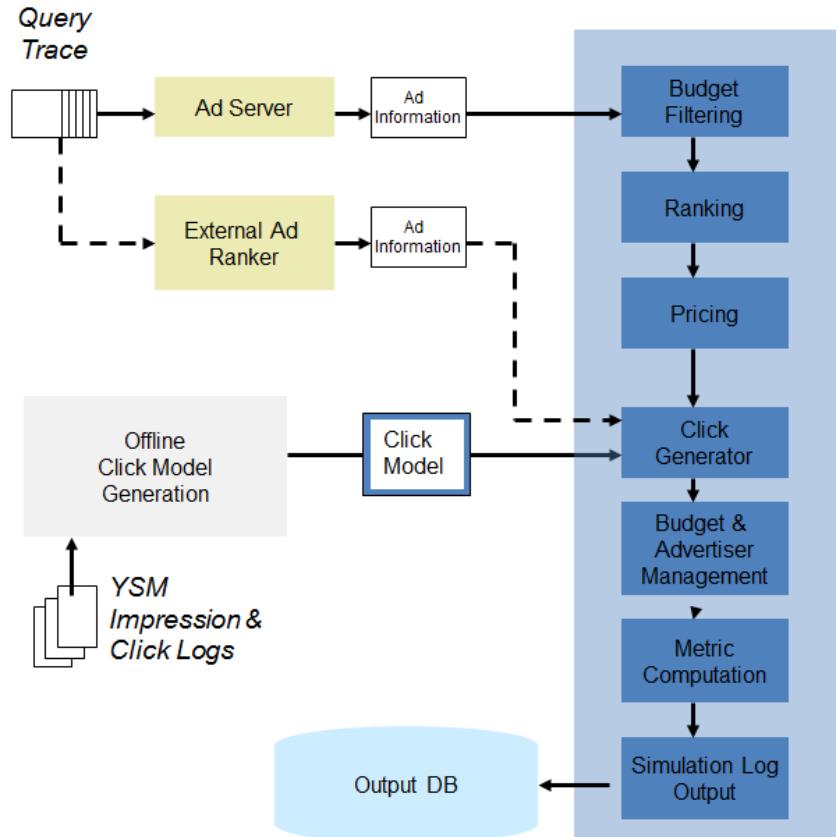
- Real paid search auctions are
 - Repeated auctions
 - Broad-match auctions
- Paid search auctions can also be
 - CPA auctions
 - Hybrid auctions (CPM + CPC + CPA)
 - References
 - Hybrid Keyword Search Auctions, WWW 2009
 - Dynamic Cost-Per-Action Mechanisms and Applications to Online Advertising, WWW 2008

Evaluation

- Online evaluation
 - Live test itself is risky
 - A failed configuration will decrease our revenue
 - May damage our brand impression
 - The fraction of live traffic is typically small to mitigate risk
 - The collected data/observations may be insufficient
 - The results may be not reliable
 - Cannot run many configurations simultaneously
 - Long time period is needed to accumulate data
 - Isolating all design factors in a production system is usually not possible
 - Difficult to draw conclusions for the investigated factor

Evaluation

- Offline evaluation
 - Entire traffic
 - Implement faster
 - Easy to isolate design factors
 - Create scenarios to model advertiser consequences



Cassini in Yahoo!

Evaluation

- Challenges of offline experiments
 - Efficiency
 - Evaluation measures
 - Dynamics in users and advertisers
 - Second-order effects

Summary

Online Advertising is Promising!

- It is about a business model, but not just a component technology!
- It is cross-discipline, with a lot of unexploited directions!
 - Information retrieval
 - Machine learning
 - Game theory
 - Economics
 -

Online Advertising is Hard!

- It is cross-discipline, and therefore requires a wide range of knowledge and skills
- It can be highly theoretical: needs deep understanding of game theory and computational theory
- It can be highly practical: needs a lot of data support and needs to be verified by real systems.

Let's Do It!

- As researchers, we like to work on promising directions.
- As researchers, we like to solve challenging problems.

Thanks!

tyliu@microsoft.com

<http://research.microsoft.com/users/tyliu/>

<http://weibo.com/tieyanliu>

Introduction to Machine Translation

Mu Li
Microsoft Research Asia

Introduction to Statistical Machine Translation

Mu Li
Microsoft Research Asia

Outline

- Machine translation overview
- Fundamental of SMT
- SMT Models
- SMT model training
- MT evaluation

An Example of Machine Translation

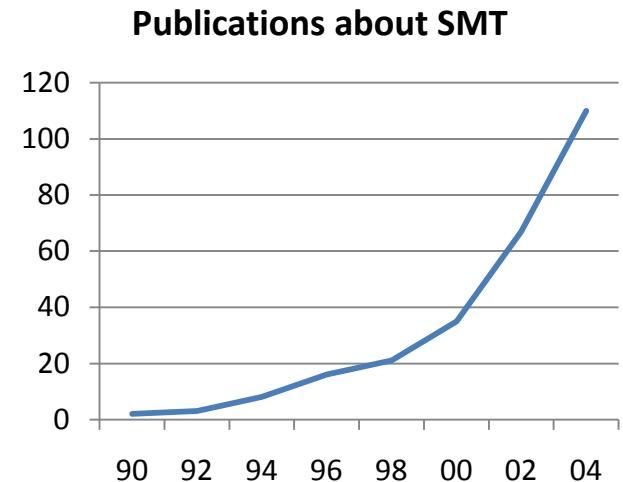
The screenshot shows a web browser window for the Microsoft Bing Dictionary Beta (<http://dict.bing.com.cn/>). The search query is "6日上午,S&P下调美国主权信用评级,由AAA降到AA,这在近百年来尚属首次....". The results page displays the query in both Chinese and English. A yellow gear icon indicates that the text has been translated by a computer. The English translation reads: "The morning of 6th, S&P cut United States sovereign credit ratings from AAA down to AA, which in the past century is the first time. Experts believe that this will once again raise fears of debt crisis on the United States in the world, increase the uncertainty in the global market, investor confidence will face a huge challenge." Below the translation, a summary in Chinese states: "6日上午,s&p下调美国主权信用评级,由aaa降到aa,这在近百年来尚属首次. 专家认为,此举将再次引发世界对美债危机的担忧,增加全球市场的不确定性,投资者信心将面临巨大考验." Filter options at the bottom include categories (全部), sources (全部), difficulty (全部), and a checked checkbox for "逐词释义" (word-by-word interpretation).

Definition of Machine Translation

- Machine Translation (MT)
 - Translate one language to another with computers
 - Mostly working at sentence level
 - Cheap way of access cross-lingual information
 - Classic problem of natural language processing research
- Application scenarios
 - Fully Automatic Machine Translation (全自动机器翻译)
 - Human Assisted Machine Translation (人助机译)
 - Computer Aided Translation (机助人译)

History of Machine Translation Research

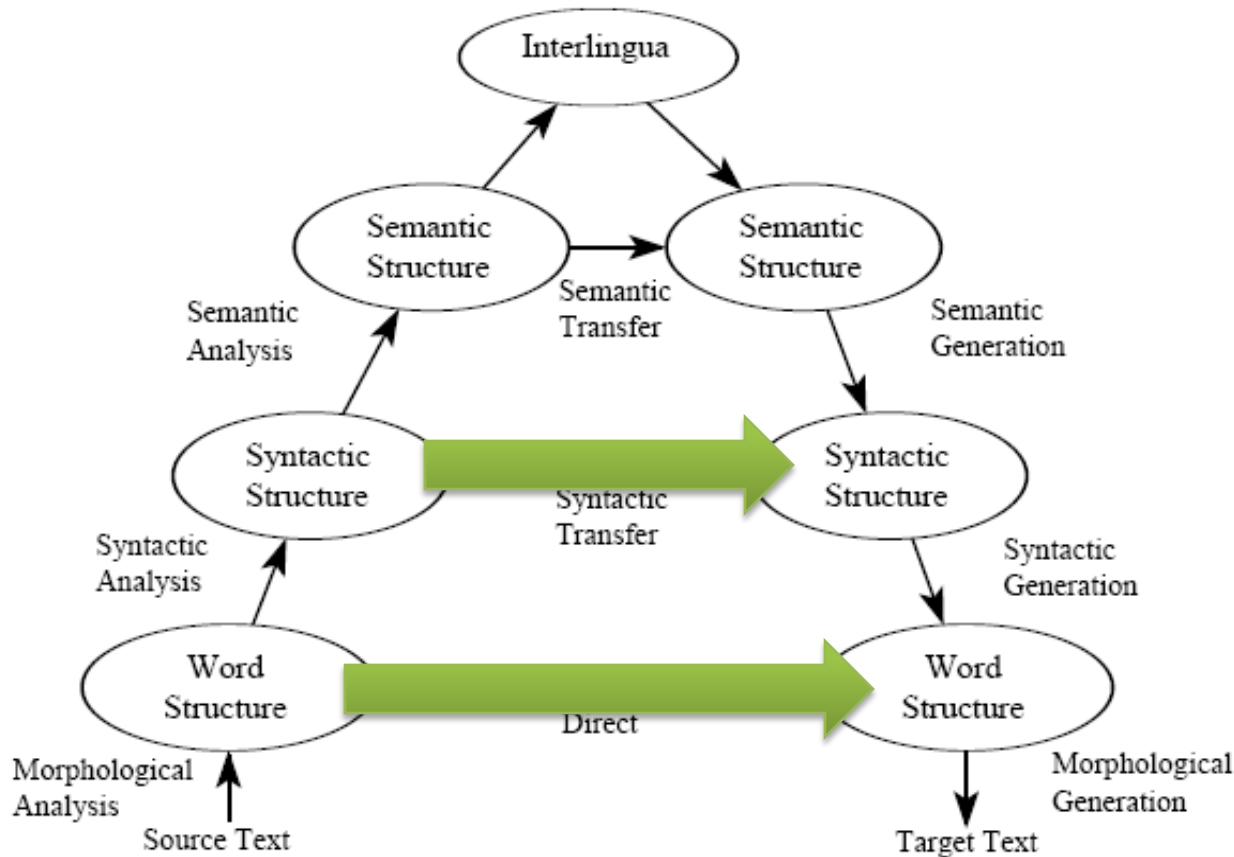
- 1946 – 1954
 - The first MT system in Georgetown
 - Russian-English, 6 rules, 250 words, 50 sentences
- 1966
 - ALPAC Report
- 1970 – 1980
 - Fundamental research on natural language theory
 - Unification grammar, semantic network
- 1980 – 1990
 - Commercialized of rule-based MT systems
 - SYSTRAN
- 1988
 - Candie system @ IBM
- 1990 – 2000
 - Pioneer work on statistical machine translation
- 2000 – 2011
 - World-wide interest in statistical method in machine translation
 - Web service for machine translation using large scale data



Machine Translation Technologies

- RBMT (Rule-Based MT)
 - Word-by-word translation
 - Grammar-based direct transfer
 - Interlingua-based method
- EBMT (Example-Based MT)
 - Example as skeleton, top-down translation
- SMT (Statistical MT)
 - Assemble of translation units, bottom-up translation

Machine Translation Pyramid

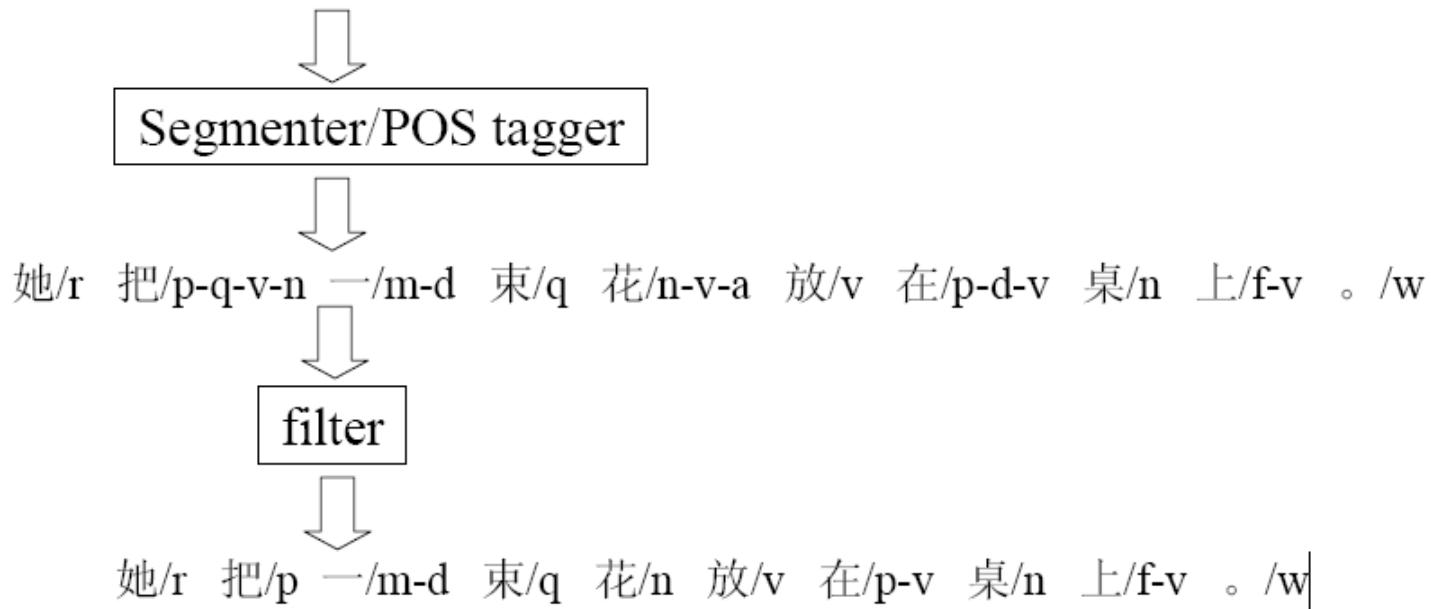


Machine Translation Technologies

- RBMT (Rule-Based MT)
 - Word-by-word translation
 - Grammar-based direct transfer
 - Interlingua-based method
- EBMT (Example-Based MT)
 - Example as skeleton, top-down translation
- SMT (Statistical MT)
 - Assemble of translation units, bottom-up translation

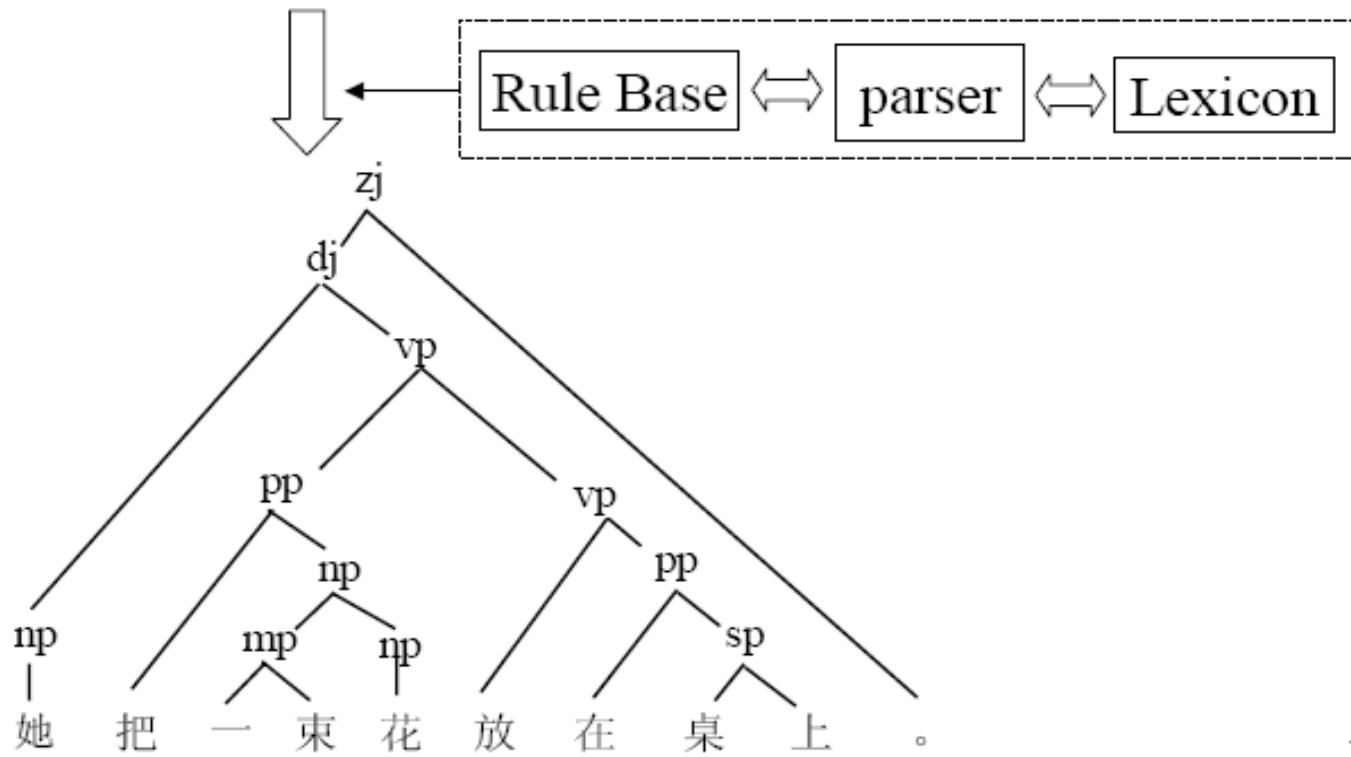
RBMT: Source Language Analysis

她把一束花放在桌上。 \rightarrow She put a bunch of flowers on the table.

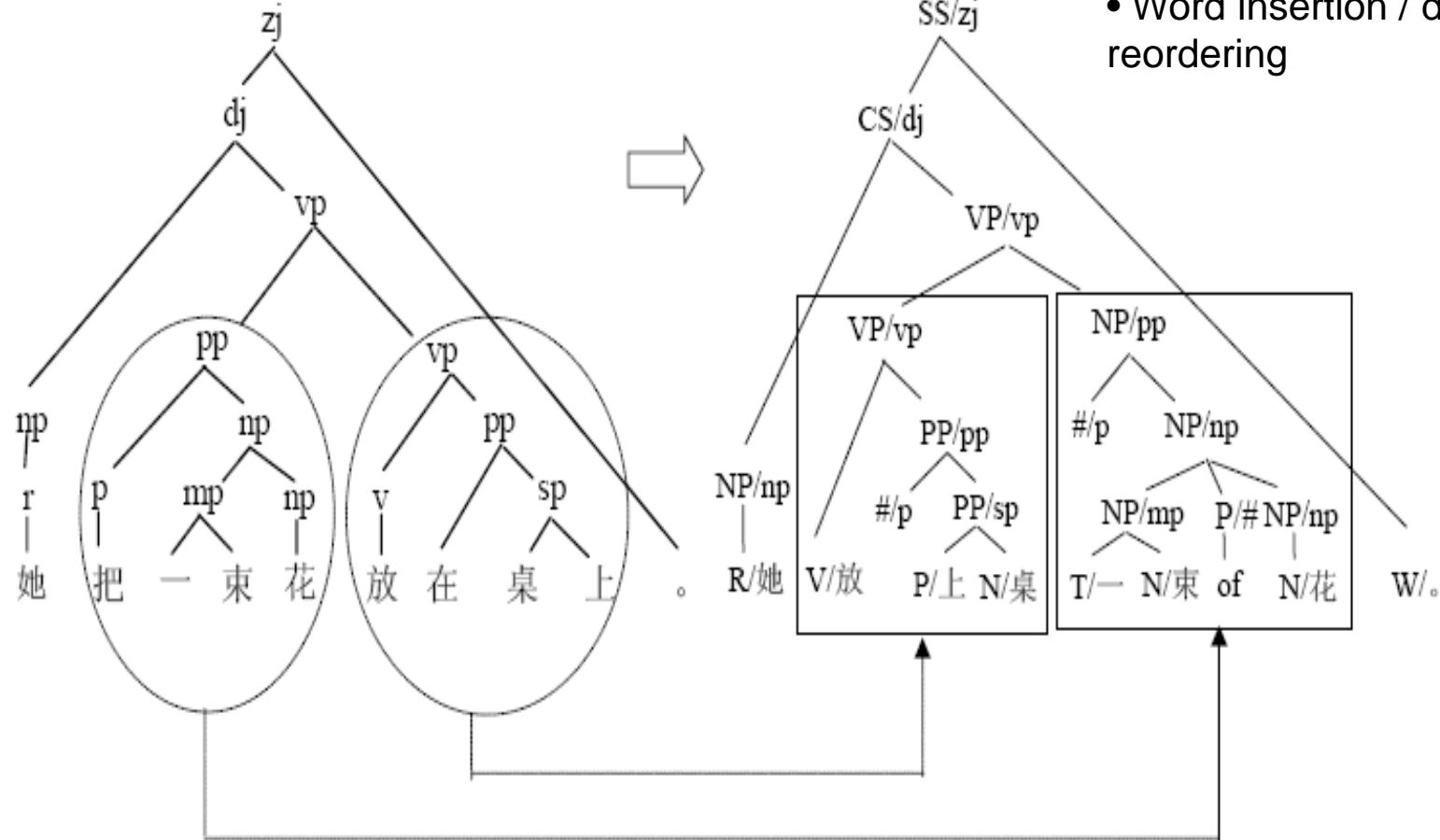


RBMT: Source Language Parse Tree

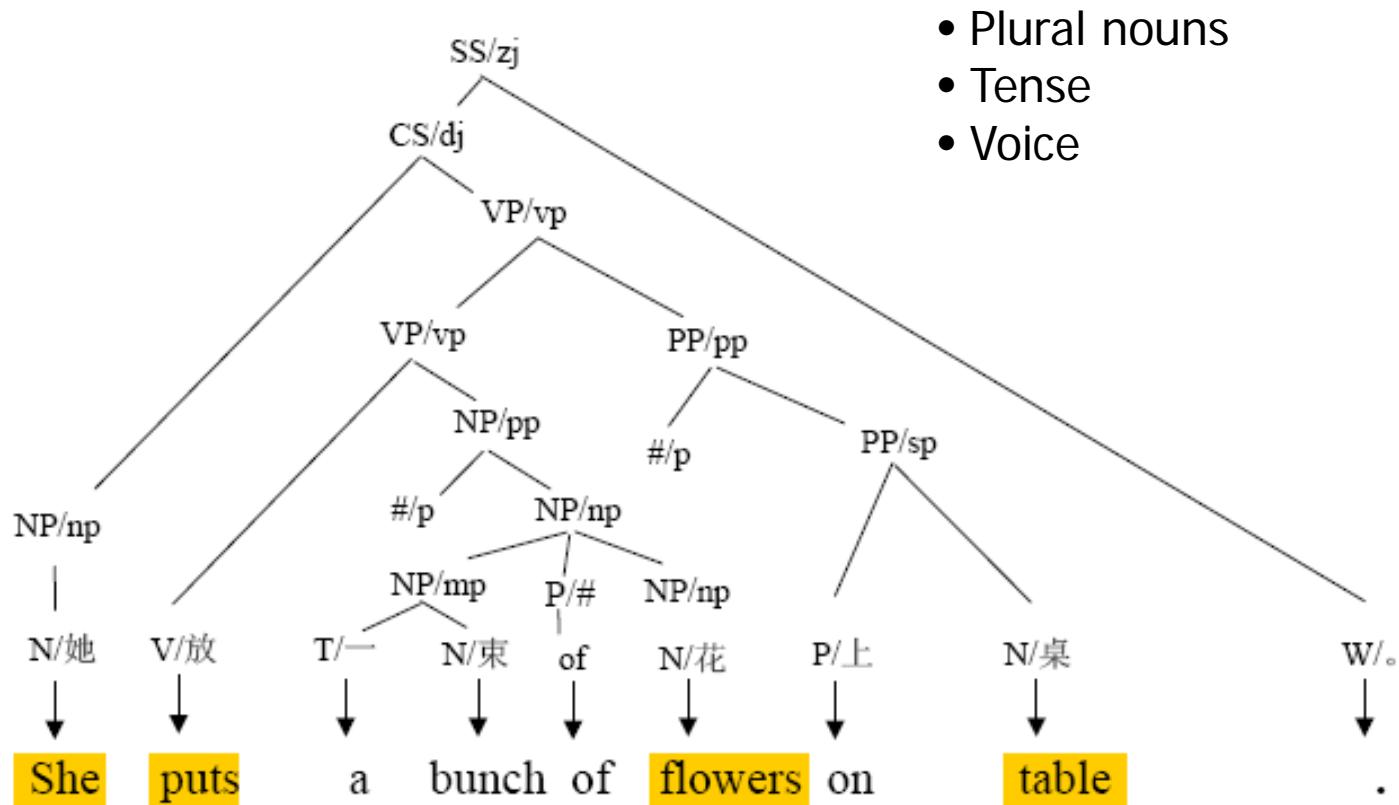
她/r 把/p 一/m-d 束/q 花/n 放/v 在/p-v 桌/n 上/f-v 。 /w



RBMT: Tree Transformation



Target Language Generation

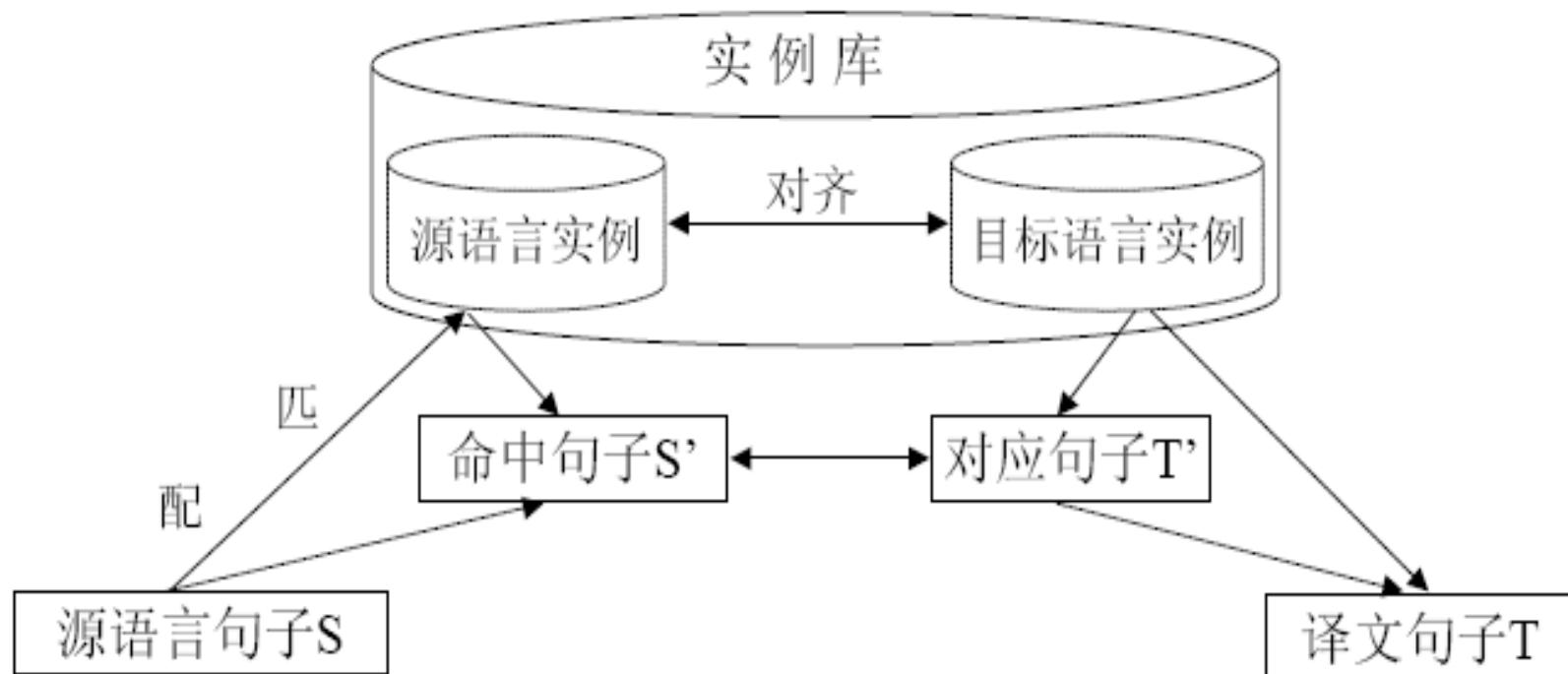


Comments on RBMT

- Pros: Easy to start
 - Intuitive
- Cons: Hard to improve (AI complete problem)
 - Require world knowledge
 - Knowledge-based maintenance

Example-based Machine Translation

- Makoto Nagao (1984)



Example-based Machine Translation

照猫画虎的机器翻译

英语实例	汉语实例
He eats vegetable	他吃蔬菜
Acid eats metal	酸腐蚀金属

输入:

I eat potatoes

输出:

我吃土豆

More about EBMT

- A data driven approach
- Learning from examples
 - Usually in a top-down manner
 - No principled and quantified method to choose examples
 - No principled way to find best translation

Statistical Machine Translation (SMT)

- MT as a statistical decision problem
 - Doing MT with statistical methods
 - Given a source sentence, use statistical data and metrics to decide what is the best translation
 - Learning from data
 - Quantized computation
 - Probabilistic predication
 - Herman Ney
 - SMT = linguistic modeling + statistical decision theory

SMT Chronicle

- 1988
 - IBM models
- 1999
 - JHU summer workshop
- 2002
 - Log-linear framework
 - SMT won in NIST evaluation
- 2003
 - Max BLEU training
- 2005
 - Power of web scale language model
- 2006
 - First large-scale success of syntax-based model in SMT

Why SMT?

- 知识获取瓶颈
 - 基于规则的翻译面临知识获取的困难。统计机器翻译从双语对照文本中自动学习翻译知识
 - 虽然在建立统计机器翻译模型时要花费很大的人力，但是在开拓一个新语言对的时候，代价相对基于规则的方法要小很多。
- 知识表达的颗粒度
 - 由于统计机器翻译是数据驱动，可获细小颗粒度的知识并且可以获得上下文有关的约束，因此译文质量要好于粗颗粒度的基于规则的方法。
- 系统的可维护性和可扩展性
 - 规则系统利用专家手工知识比较困难。而统计方法利用数据驱动易于维护和扩展。
- 但是，如果双语的数据少，比如对某些语言对来说，双语的数据很难获得，则统计翻译方法会变得无效。那时，基于规则的方法要好很多。

名人名言

- A word is a world.
 - **Douglas Lenat, founder of CYC project**
- It must be recognized that the notion “probability of a sentence” is an entirely useless one, under any known interpretation of this term.
 - **Noam Chomsky, 1969**
- Whenever I fire a linguist our system performance improves.
 - **Frederick Jelinek, 1988**

Translation as Decoding



人们会自然地认为翻译的问题实际上可以看作是一个密码破译的问题。当我看到一本用俄语写的书，我认为它实际上是英语写的，只不过是用一些奇快的符号编码。我只要想想如何破解即可。

Warren Weaver, 1947

Fundamental Problems of SMT

- Modeling
 - What translation to find
- Searching / decoding
 - How to find translation
- Training / learning
 - Model parameter estimation



An Old Story – Source-Channel Modeling

$$p(e|f) \sim P(\cdot|\cdot)$$

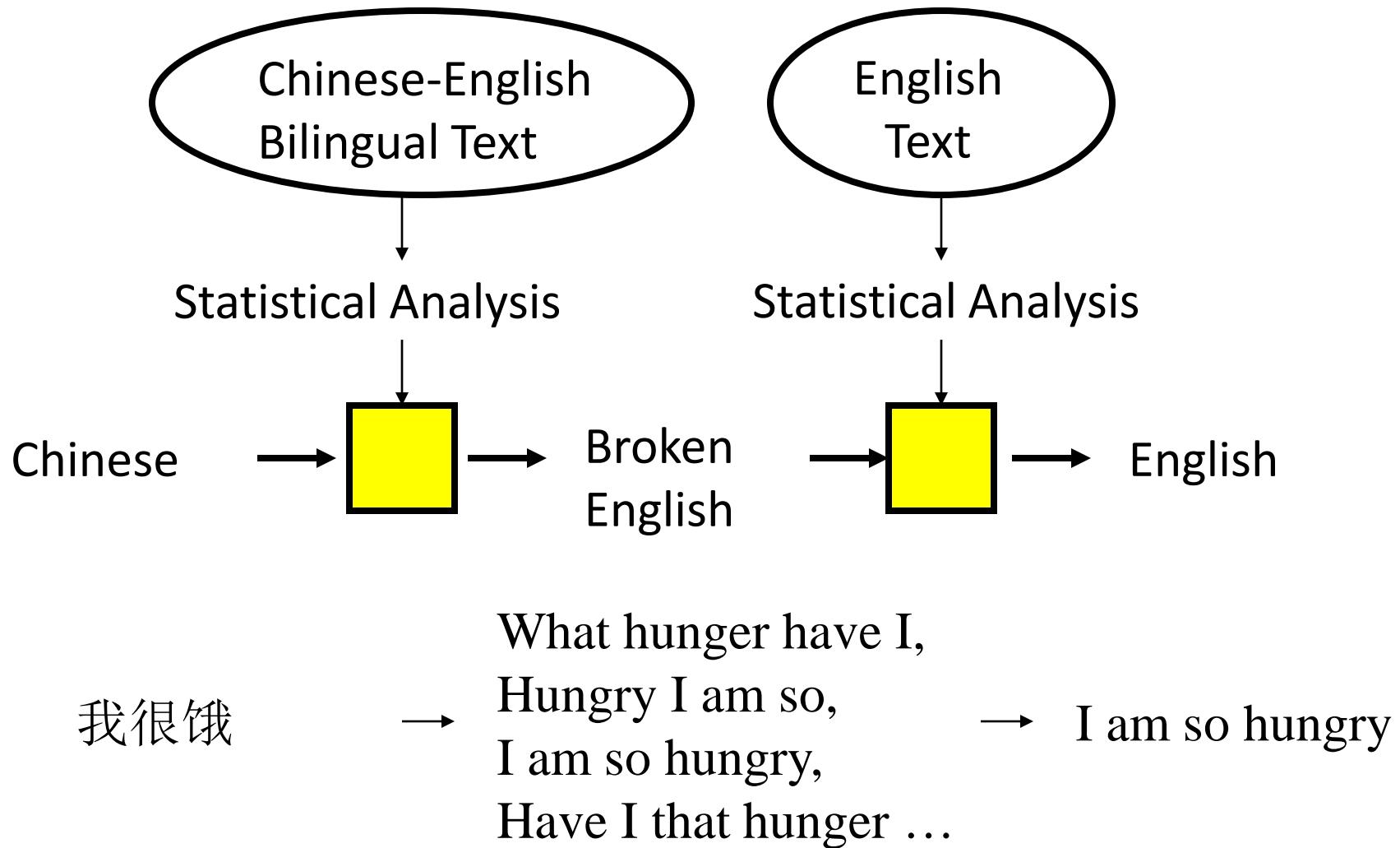
$$e^* = \operatorname{argmax}_e P(e|f)$$

$$P(e|f) = \frac{P(e)P(f|e)}{P(f)}$$

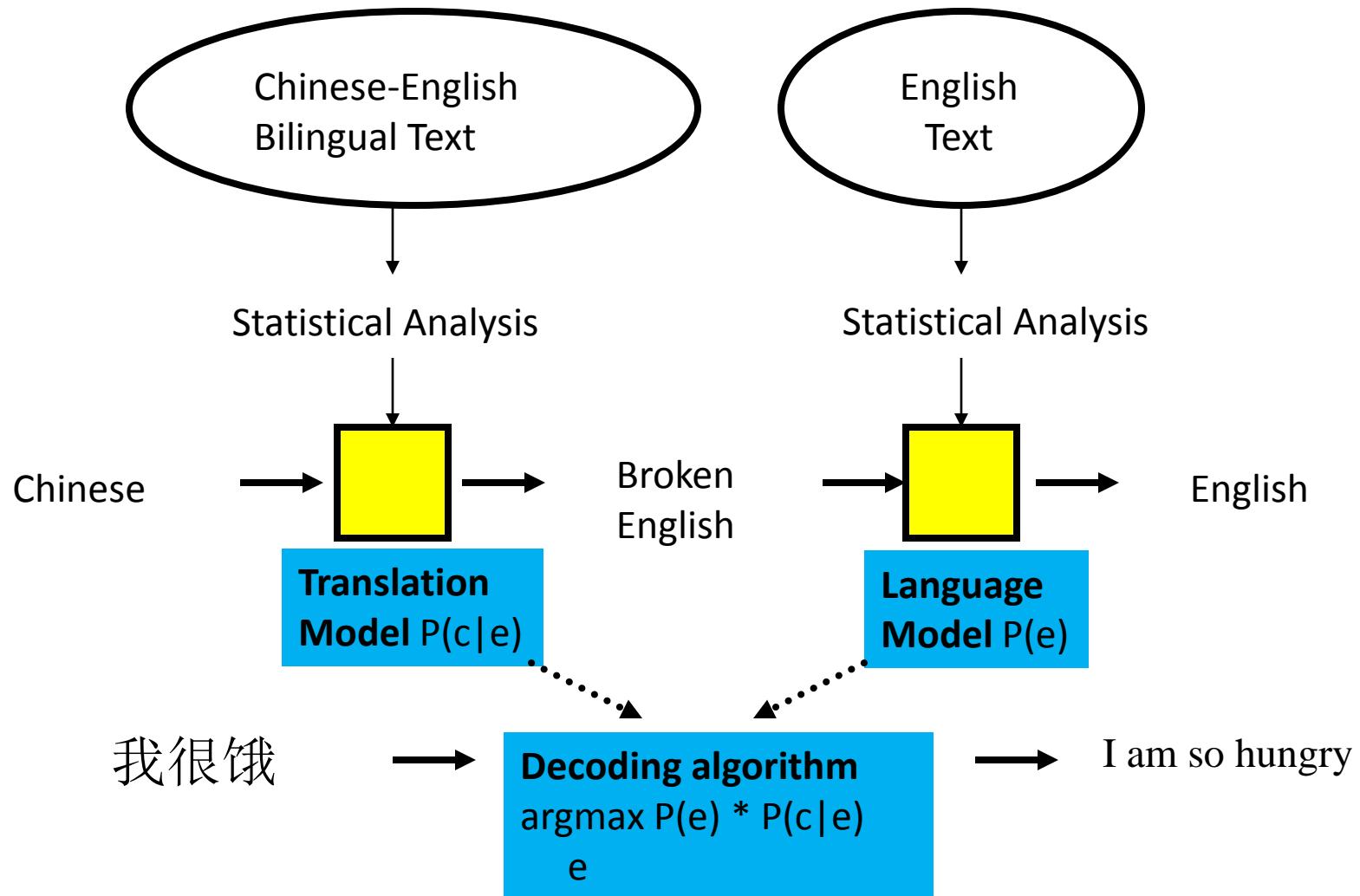
$$e^* = \operatorname{argmax}_e P(e) P(f|e)$$

Language Model (Source Model) Translation Model (Channel Model)

Source-Channel SMT System



Source-Channel SMT System



More on Modeling

- N-gram language model

- $e = e_1^m = e_1 \dots e_m$

$$P(e) = P(e_1)P(e_2|e_1) \dots P(e_{n-1}|e_1 \dots e_{n-2}) \prod_{i=n}^m P(e_i|e_{i-n+1} \dots e_{i-1})$$

- Translation model

- $P(f|e) = \prod P(f_j|e_{a_j})$

More on Language Model

Site	BLEU
Google	0.3531
ISI	0.3073

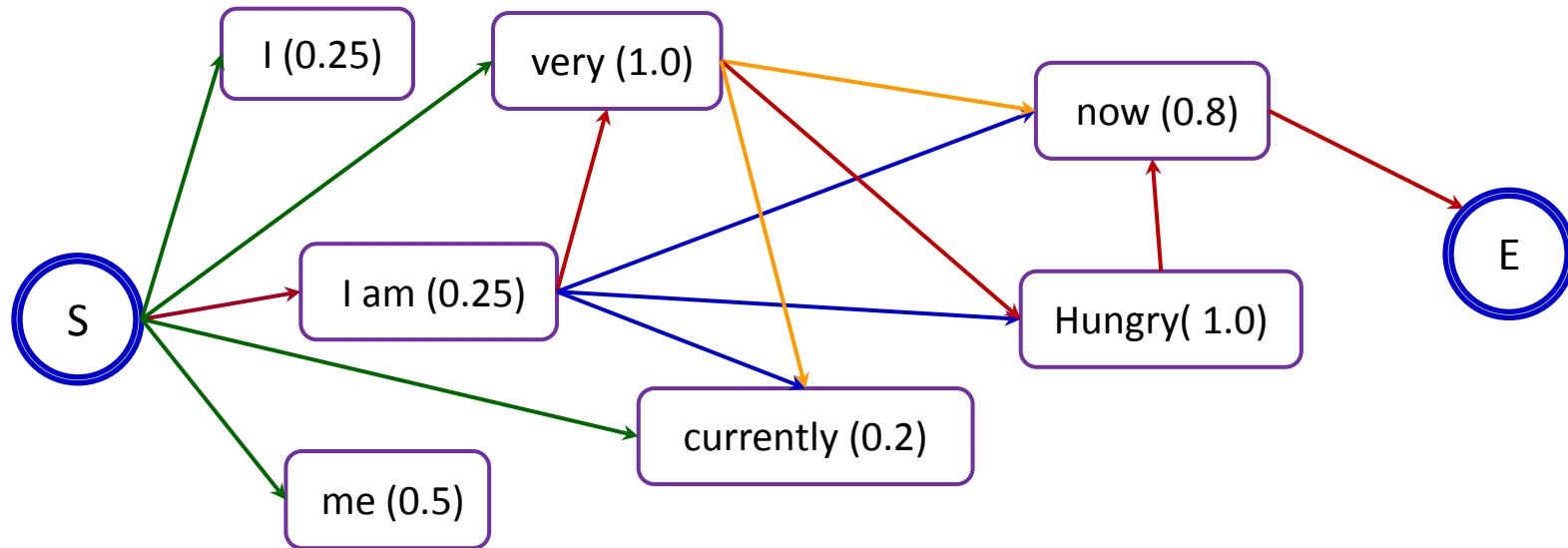
NIST 2005 Chinese-English MT Evaluation results

ngram	BLEU
2	31.9
3	36.7
4	38.4
5	38.8
6	38.8

# words	BLEU
30 M	35.58
60 M	36.51
120 M	37.43
250 M	38.80
400 M	39.39

Search in Word Graph / Lattice

Input: 我 现在 很 饿

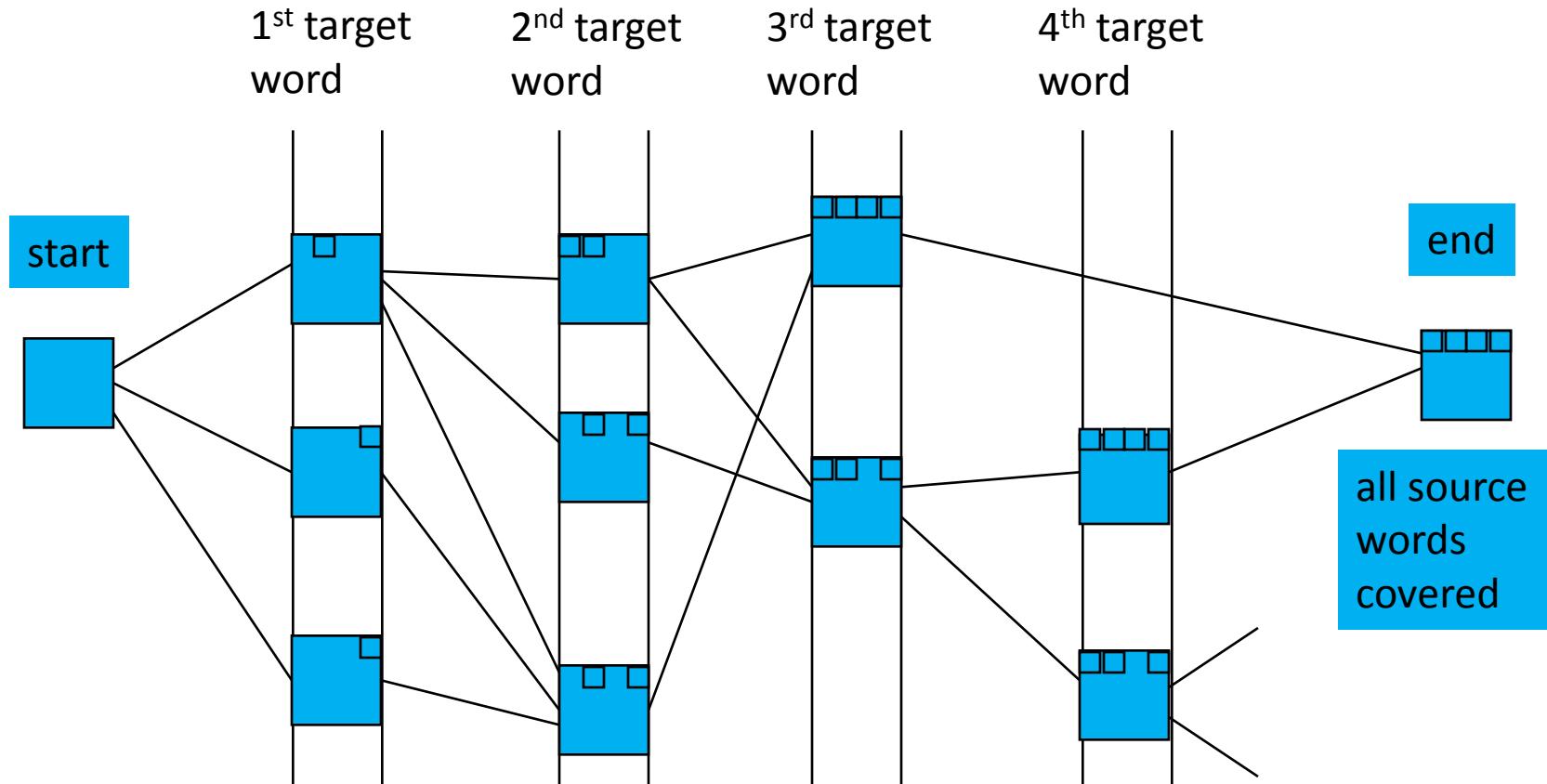


- Translation for each word as nodes
- A link exists between every two nodes not from the same source word
- Find best path from start to end node
- Cost of path determined by translation and language models

Decoding for Classic SMT Models

- Of all conceivable English word strings, find the one maximizing $P(e)P(e|f)$
- Decoding is an NP-complete challenge (Knight, 1999)
 - $n!$ permutations for an English sentence with n words
 - Each potential English output is called a *hypothesis*.
- Several speed-up strategies are available
 - Dynamic programming
 - Histogram pruning
 - Limited number of candidates in each bucket/stack
 - Threshold pruning
 - Abandon candidates with low score

Dynamic Programming Beam Search



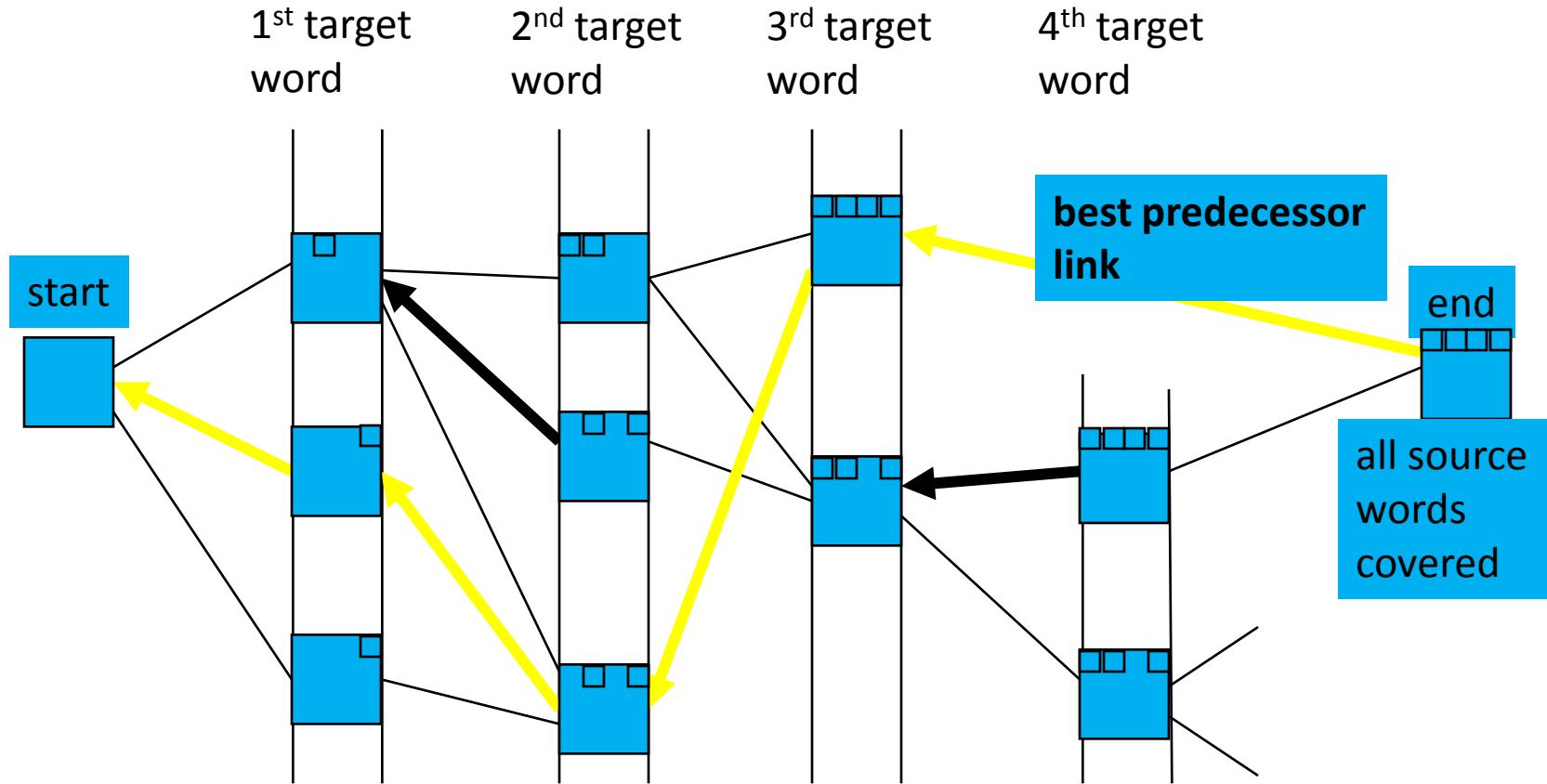
Each partial translation hypothesis contains:

- Last English word chosen + source words covered by it
- Next-to-last English word chosen
- Entire coverage vector (so far) of source sentence
- Language model and translation model scores (so far)



[Jelinek, 1969;
Brown et al, 1996 US Patent;
(Och, Ueffing, and Ney, 2001)]

Dynamic Programming Beam Search



Each partial translation hypothesis contains:

- Last English word chosen + source words covered by it
- Next-to-last English word chosen
- Entire coverage vector (so far) of source sentence
- Language model and translation model scores (so far)



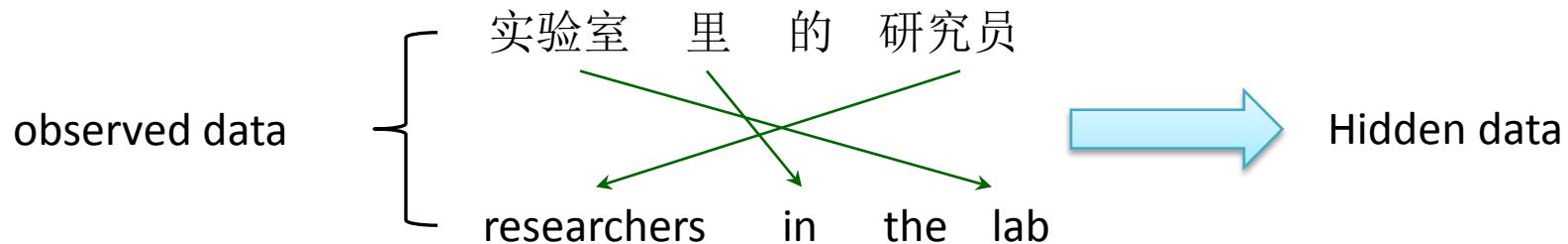
[Jelinek, 1969;
Brown et al, 1996 US Patent;
(Och, Ueffing, and Ney, 2001)]

Translation Models and Word Alignment

- IBM Models
 - Model 1 ~ 5 dealing with estimating $P(f|e)$
- HMM model
 - Improvement over IBM Model 2

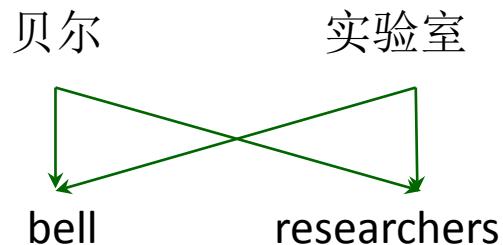
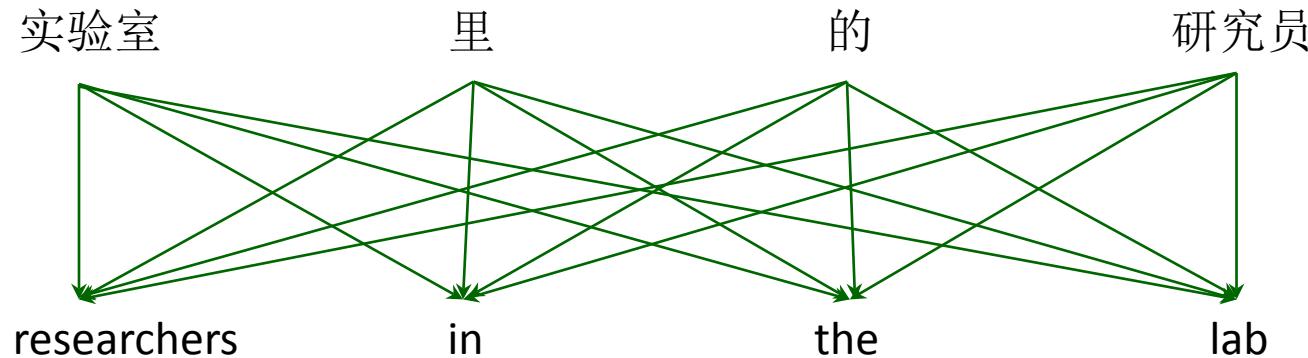
EM Training for Word Alignment

- EM in a couple of slides
 - Expectation Maximization, one optimization method
 - Unsupervised method working on incomplete data
 - Interactively optimize the objective function



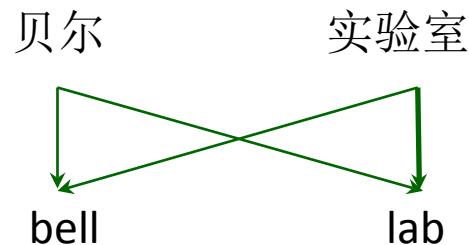
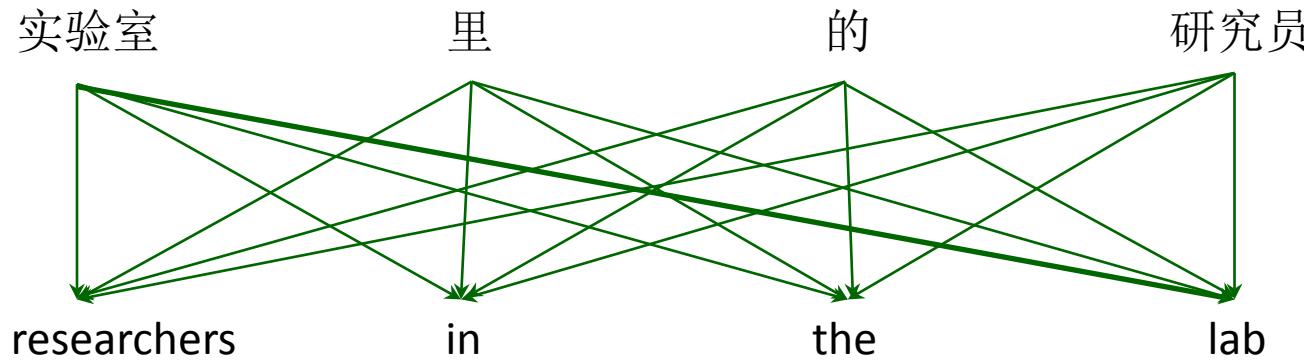
EM Training for Word Alignment

- Initial step: all alignment links are equally likely



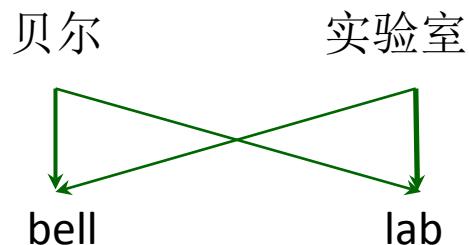
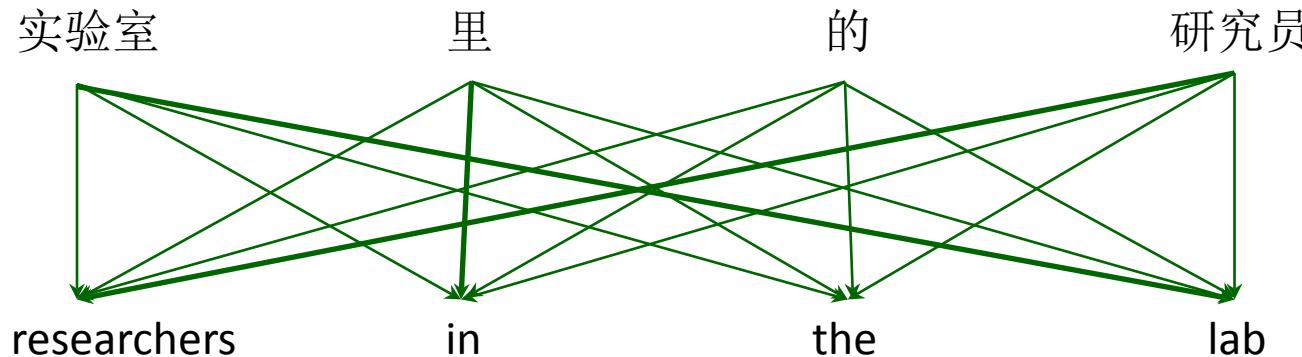
EM Training for Word Alignment

- After one iteration, link between 实验室 and lab becomes stronger



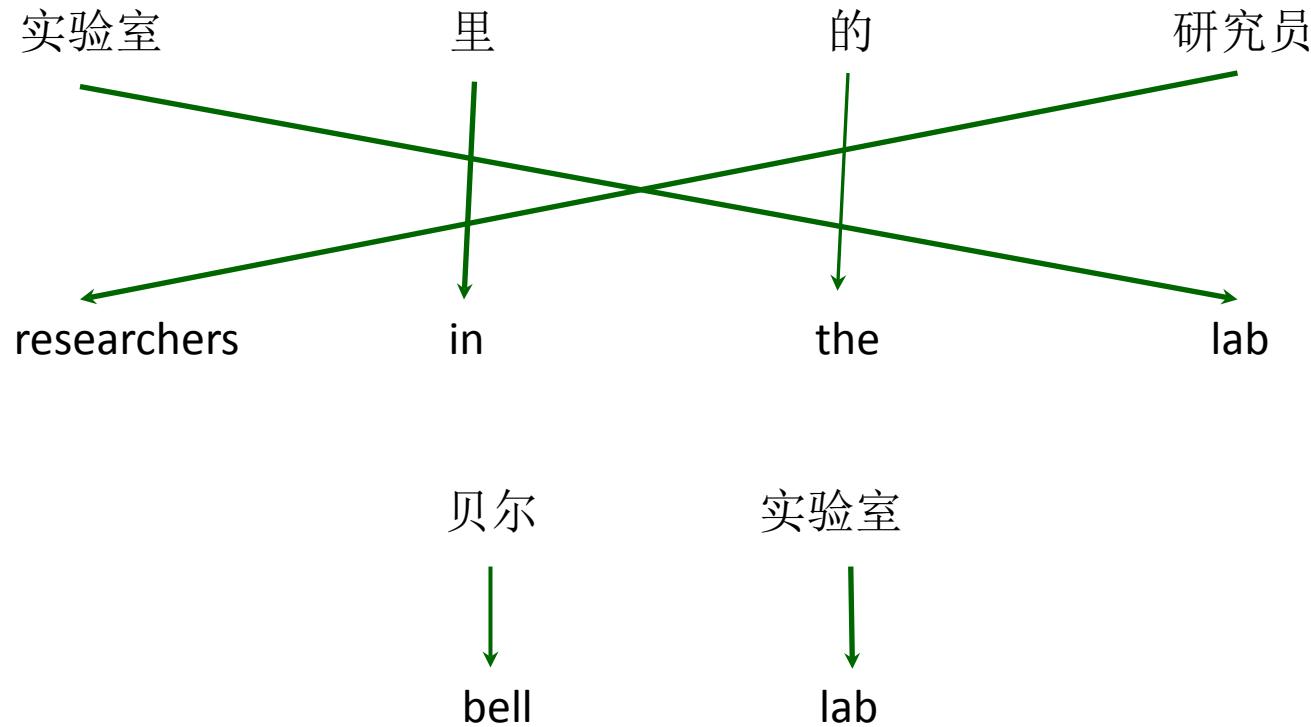
EM Training for Word Alignment

- After two iterations, more links between words become stronger



EM Training for Word Alignment

- Finally the algorithm will converge with some hidden structure



Notations

$$\boldsymbol{e} = e_1^l = e_1 \dots e_l \qquad \boldsymbol{f} = f_1^m = f_1 \dots f_m$$

$$\boldsymbol{a} = a_1^m = a_1 \dots a_m \quad \left(0 \leq a_i \leq l\right)$$

$$a_j=i \Rightarrow (f_j,e_i)$$

$$P(\boldsymbol{f}|\boldsymbol{e})=\sum_{\boldsymbol{a}} P(\boldsymbol{f},\boldsymbol{a}|\boldsymbol{e})$$

$$P(\boldsymbol{f},\boldsymbol{a}|\boldsymbol{e})=P(m|\boldsymbol{e})\prod_{j=1}^m P(a_j|a_1^{j-1},f_1^{j-1},m,\boldsymbol{e})P(f_j|a_1^j,f_1^{j-1},m,\boldsymbol{e})$$

IBM Model 1

$$P(\mathbf{f}, \mathbf{a}|\mathbf{e}) = P(m|\mathbf{e}) \prod_{j=1}^m P(a_j|a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) P(f_j|a_1^j, f_1^{j-1}, m, \mathbf{e})$$

\downarrow \downarrow \downarrow
 ϵ $\frac{1}{l+1}$ $t(f_j|e_{a_j})$

$$P(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^m t(f_j|e_{a_j})$$
$$P(\mathbf{f}|\mathbf{e}) = \frac{\epsilon}{(l+1)^m} \sum_{a_1=0}^l \cdots \sum_{a_m=0}^l \prod_{j=1}^m t(f_j|e_{a_j})$$

Goal: maximize $P(\mathbf{f}|\mathbf{e})$ subject to $\sum_f t(f|\mathbf{e}) = 1$ for all \mathbf{e}

IBM Model 1

$$h(t, \lambda) = \frac{\epsilon}{(l+1)^m} \sum_{a_1=0}^l \cdots \sum_{a_m=0}^l \prod_{j=1}^m t(f_j | e_{a_j}) - \sum_e \lambda_e \left(\sum_f t(f | e) - 1 \right)$$

$$\frac{\partial h}{\partial t(f|e)} = \frac{\epsilon}{(l+1)^m} \sum_{a_1=0}^l \cdots \sum_{a_m=0}^l \sum_{j=1}^m \delta(f, f_j) \delta(e, e_{a_j}) t(f|e)^{-1} \prod_{k=1}^m t(f_k | e_{a_k}) - \lambda_e$$

$$t(f|e) = \lambda_e^{-1} \frac{\epsilon}{(l+1)^m} \sum_{a_1=0}^l \cdots \sum_{a_m=0}^l \sum_{j=1}^m \delta(f, f_j) \delta(e, e_{a_j}) \prod_{k=1}^m t(f_k | e_{a_k})$$

$$t(f|e) = \lambda_e^{-1} \sum_{\mathbf{a}} P(\mathbf{f}, \mathbf{a}|e) \boxed{\sum_{j=1}^m \delta(f, f_j) \delta(e, e_{a_j})}$$

number of times e connects to f in \mathbf{a}

IBM Model 1

$$c(f|e; \mathbf{f}, \mathbf{e}) = \sum_{\mathbf{a}} P(\mathbf{a}|\mathbf{e}, \mathbf{f}) \sum_{j=1}^m \delta(f, f_j) \delta(e, e_{a_j})$$

$$\sum_{a_1=0}^l \cdots \sum_{a_m=0}^l \prod_{j=1}^m t(f_j | e_{a_j}) = \prod_{j=1}^m \sum_{i=0}^l t(f_j | e_i)$$

$$P(\mathbf{f}|\mathbf{e}) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^m \sum_{i=0}^l t(f_j | e_i)$$

$$c(f|e; \mathbf{f}, \mathbf{e}) = \frac{t(f|e)}{t(f|e_0) + \cdots + t(f|e_l)} \boxed{\sum_{j=1}^m \delta(f, f_j)} \boxed{\sum_{i=0}^l \delta(e, e_i)}$$

IBM Model 2

$$P(\mathbf{f}, \mathbf{a}|\mathbf{e}) = P(m|\mathbf{e}) \prod_{j=1}^m P(a_j|a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) P(f_j|a_1^j, f_1^{j-1}, m, \mathbf{e})$$
$$\downarrow \quad \quad \quad \downarrow \quad \quad \quad \downarrow$$
$$\epsilon \quad \quad \quad a(i | j, m, l) \quad \quad \quad t(f_j | e_{a_j})$$

Goal: maximize $P(\mathbf{f}|\mathbf{e})$ subject to $\sum_{i=0}^l a(i | j, m, l) = 1$ for each (j, m, l)

IBM Model 3, 4 and 5

- Model 3
 - Adds fertility model
- Model 4
 - Dealing with distortion (relative reordering model)
- Model 5
 - Removing deficiency

IBM Model 4



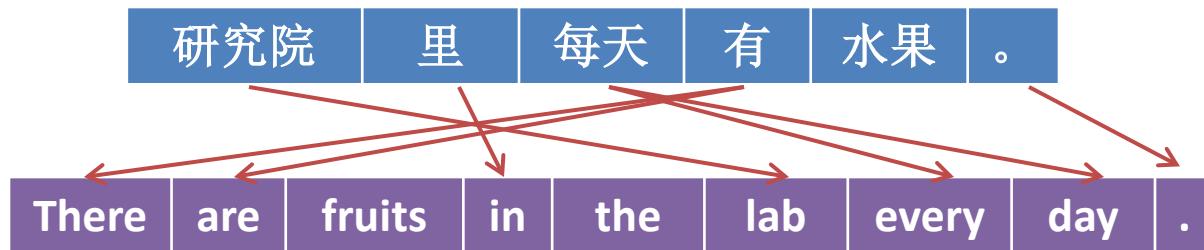
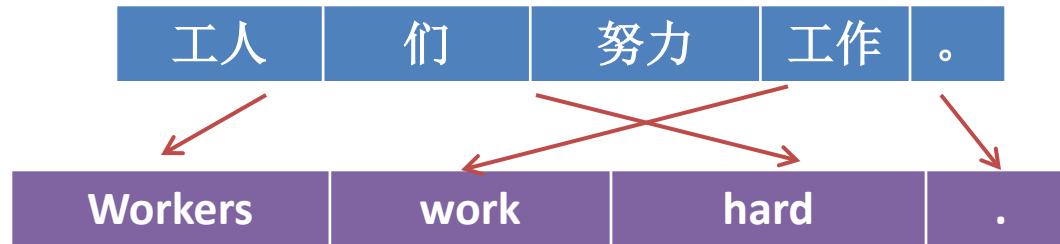
HMM

$$P(\mathbf{f}, \mathbf{a} | \mathbf{e}) = P(m | \mathbf{e}) \prod_{j=1}^m P(a_j | a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) P(f_j | a_1^j, f_1^{j-1}, m, \mathbf{e})$$
$$\downarrow \qquad \qquad \qquad \downarrow \qquad \qquad \qquad \downarrow$$
$$\epsilon \qquad \qquad P(a_j | a_{j-1}, l) \qquad \qquad t(f_j | e_{a_j})$$
$$\downarrow$$
$$\frac{s(a_i - a_{i-1})}{\sum_{j=1}^l s(l - a_{i-1})}$$

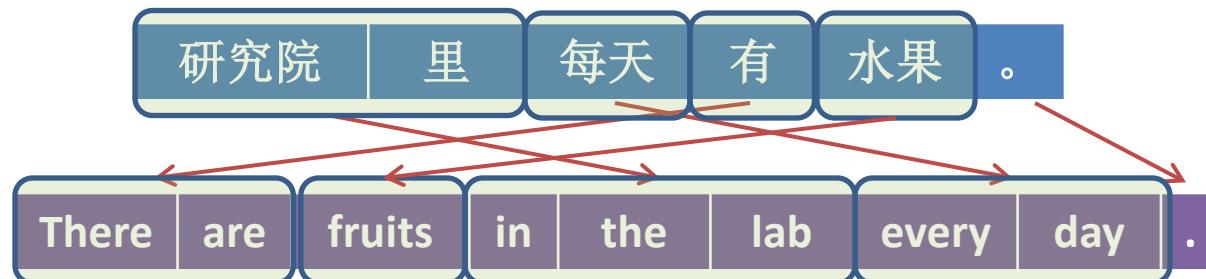
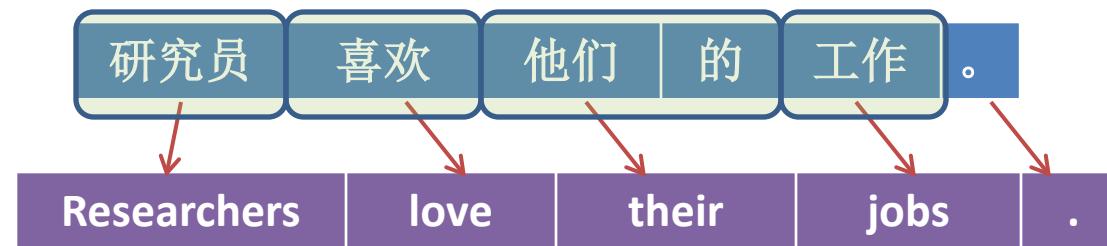
From Word To Phrase

- Words are not always natural units for translation
 - with all due respect ==> 恕我直言
 - you are welcome ==> 不必客气
- Language model is powerful
 - But not as powerful as imagined
- Solution
 - Remember more beyond word translations

Word Alignment

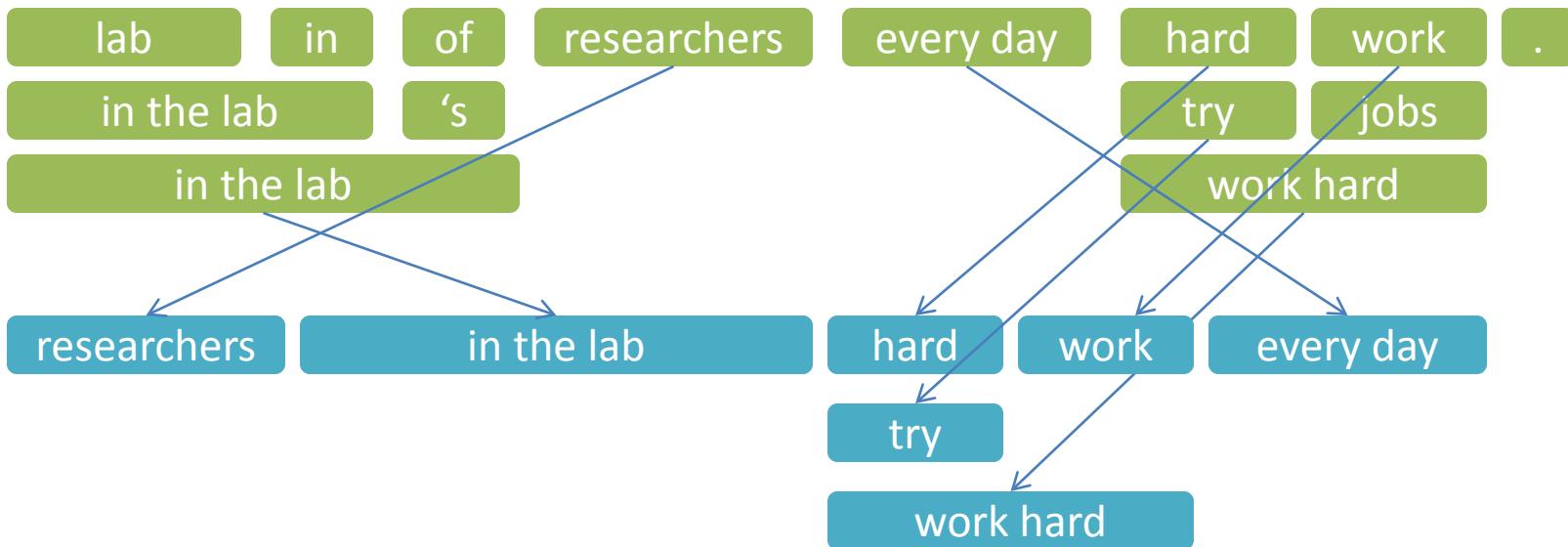


Phrase Extraction



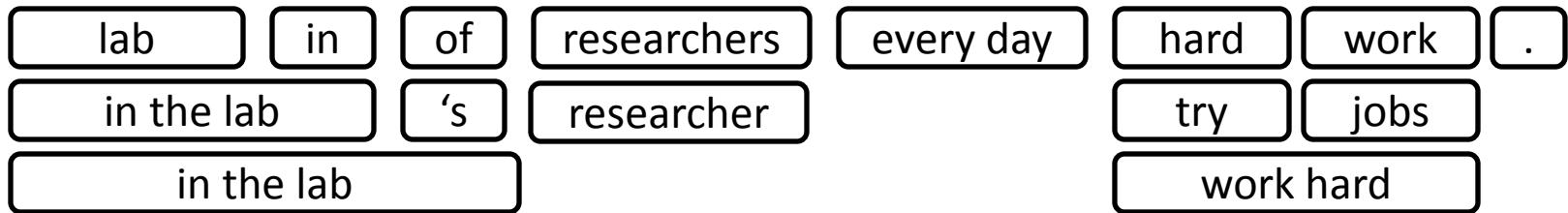
Decoding for Phrase Graph

研究院 | 里 | 的 | 研究员 | 每天 | 努力 | 工作 | 。



Decoding for Phrase Graph

研究院 | 里 | 的 | 研究员 | 每天 | 努力 | 工作 | 。

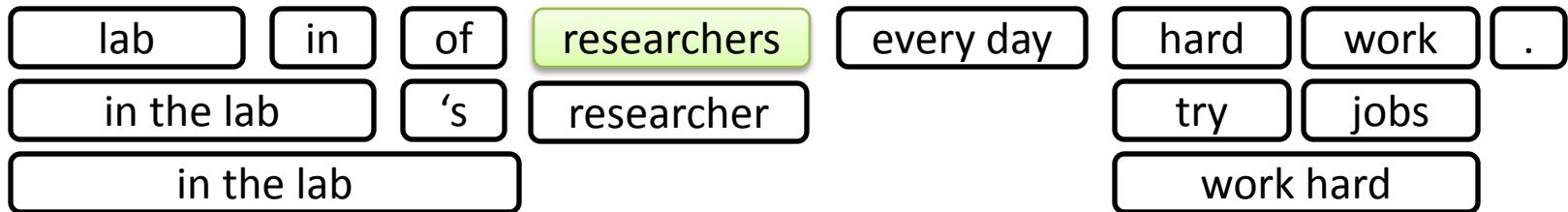


e:
f: -----
p: 1.0

- Start with empty hypothesis

Decoding for Phrase Graph

研究院 | 里 | 的 | 研究员 | 每天 | 努力 | 工作 | 。



e:
f: -----
p: 1.0

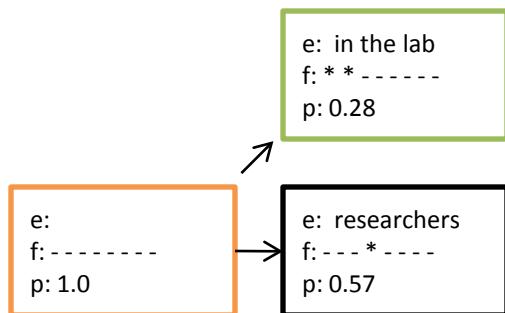
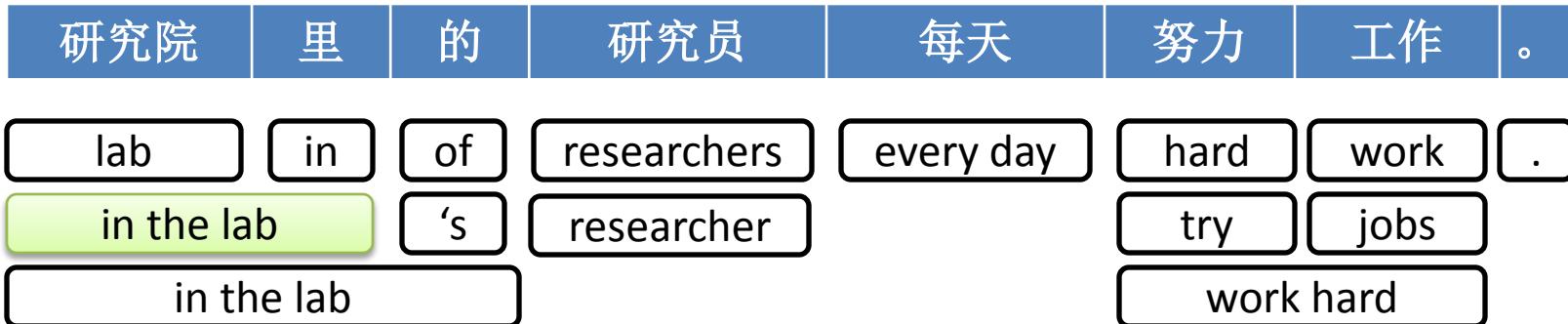
→

e: researchers
f: --- * ----
p: 0.57

- Pick translation option
- Create hypothesis
 - ✓ add target phrase researchers
 - ✓ cover the forth foreign word
 - ✓ assign probability 0.57

Adapted from Philip Koehn's tutorial on SMT

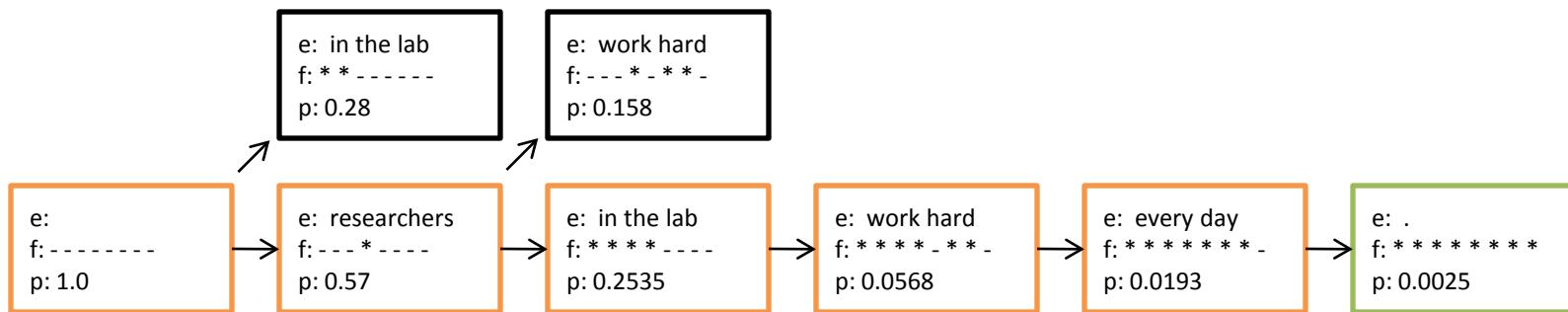
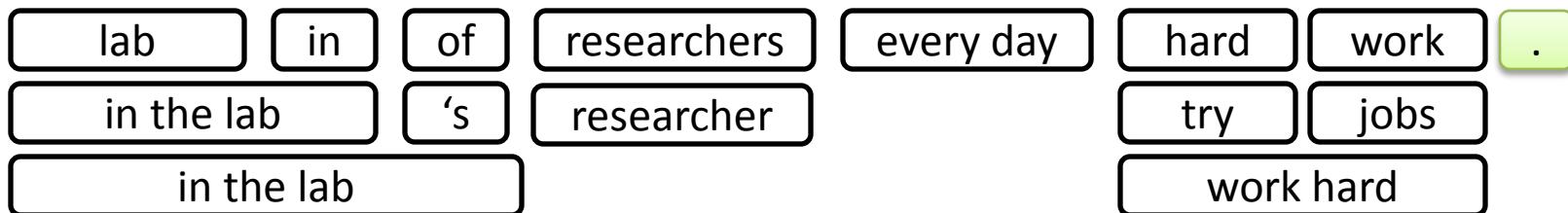
Decoding for Phrase Graph



- Add another hypothesis

Decoding for Phrase Graph

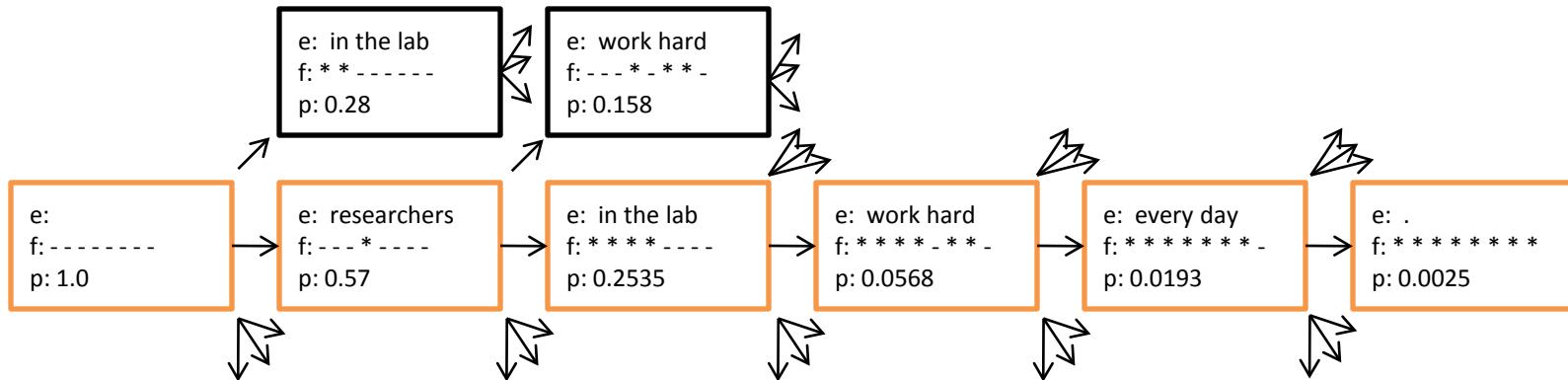
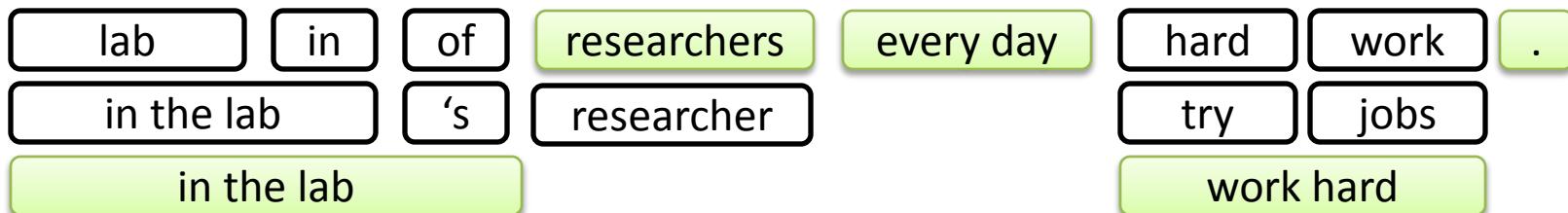
研究院 | 里 | 的 | 研究员 | 每天 | 努力 | 工作 | 。



- Further hypothesis expansion

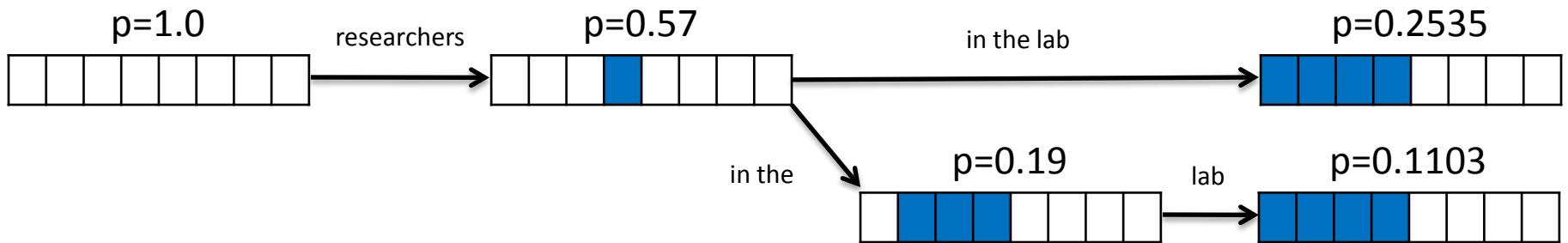
Decoding for Phrase Graph

研究院 里 的 研究员 每天 努力 工作 。



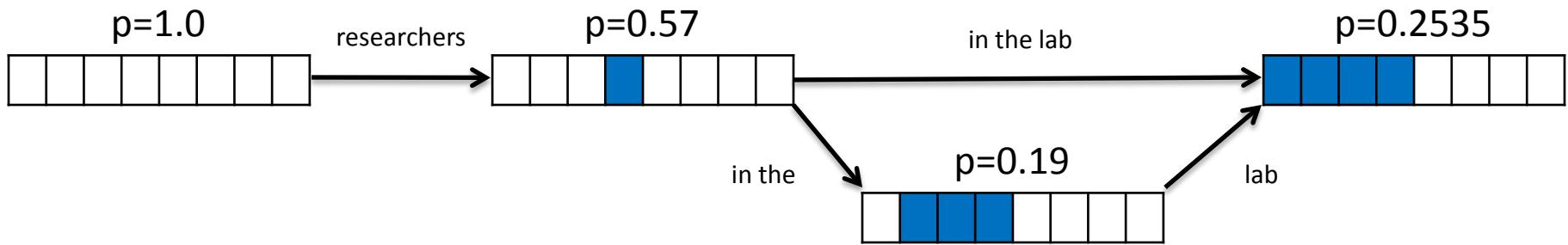
- ... until all foreign words covered
 - ✓ find best hypothesis that covers all foreign words
 - ✓ backtrack to read off the translation

Hypothesis Recombination



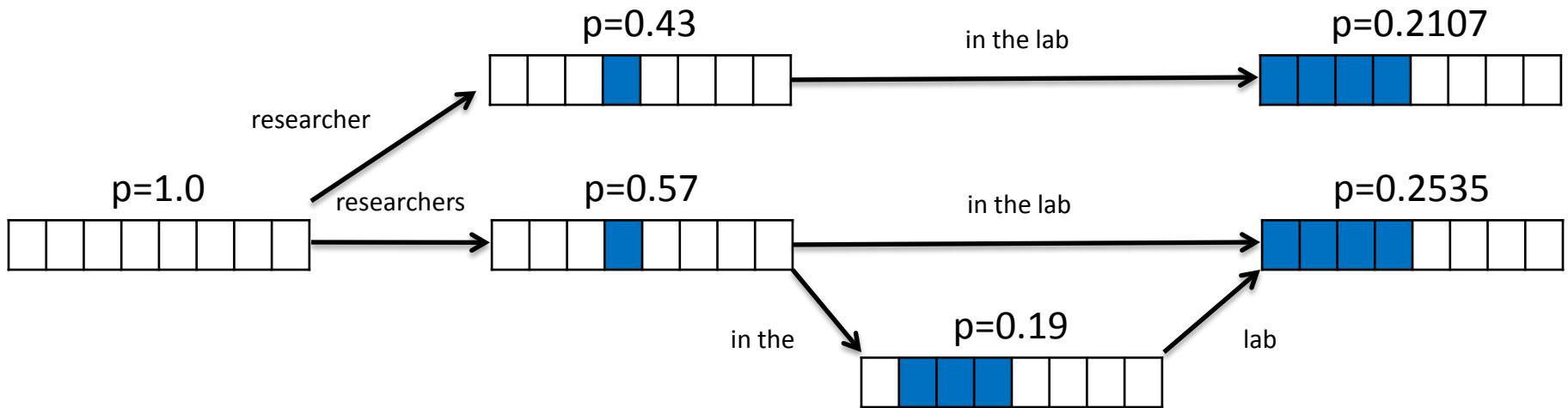
- Different paths to the *same* partial hypothesis

Hypothesis Recombination



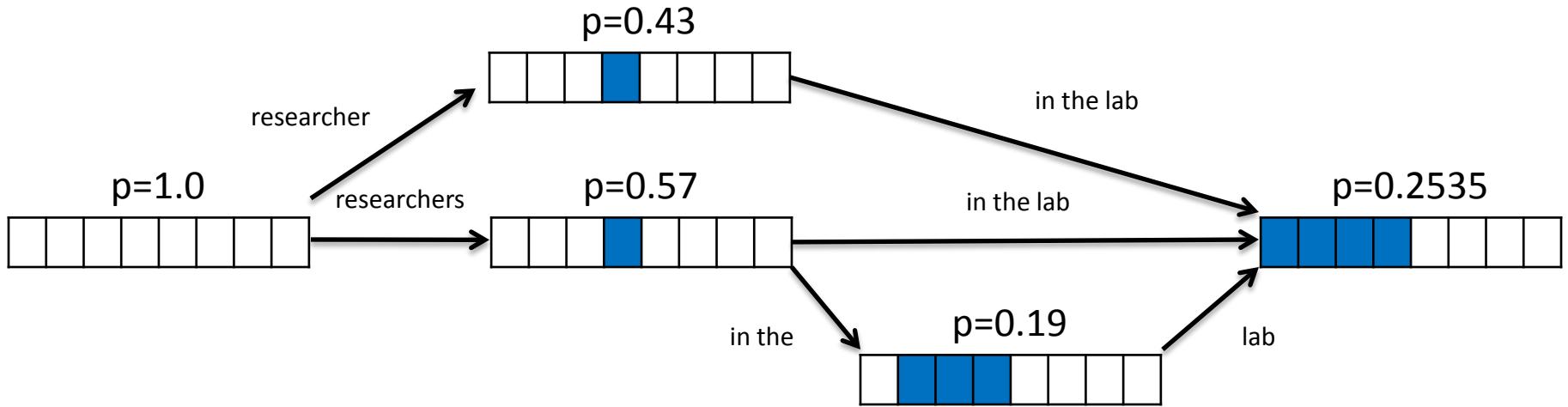
- Different paths to the **same** partial hypothesis
- Hypothesis recombination **combines** the path
 - ✓ **drop weaker** path
 - ✓ keep pointer from weaker (for lattice generation)

Hypothesis Recombination



- Recombined hypotheses do **not** have to match completely
- No matter what is added, weaker path can be dropped, if:
 - ✓ **last ($n-1$) target words** match (for language model computation)
 - ✓ **foreign word coverage vectors** match (effects future path)

Hypothesis Recombination

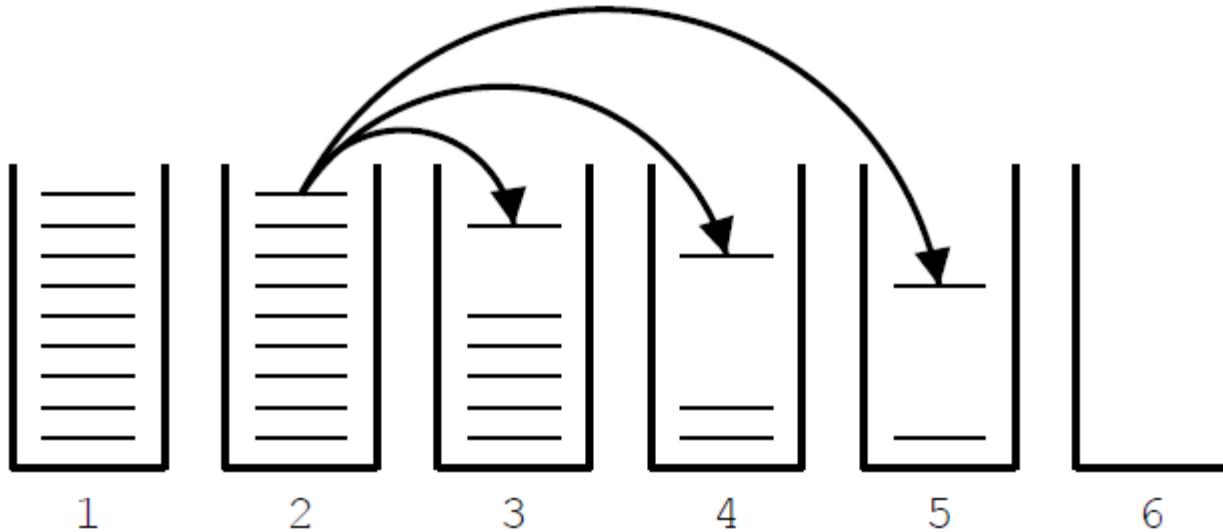


- Recombined hypotheses do **not** have to match completely
- No matter what is added, weaker path can be dropped, if:
 - ✓ **last ($n-1$) target words** match (for language model computation)
 - ✓ **foreign word coverage vectors** match (effects future path)

Pruning

- Heuristically discard weak hypothesis early
- Organize hypothesis in stacks (buckets), e.g. by
 - *same* foreign words covered
 - *same number* of source words covered
 - *same number* of target words covered
- Compare hypotheses in stacks, discard bad ones
 - *histogram pruning*: keep top n hypotheses in each stack
 - *threshold pruning*: keep hypotheses that are at most α times the cost of the best hypothesis in stack

Hypothesis Stack

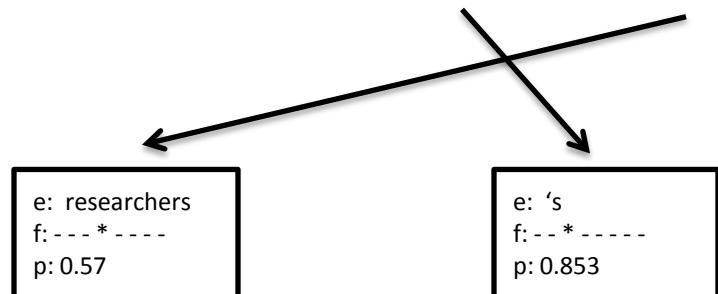


- Organization of hypothesis into stacks
 - in this example, based on ***number of foreign words*** translated
 - during translation, all hypotheses from one stack are expanded
 - expanded hypotheses are placed into stacks

Comparing Hypothesis

- Comparing hypotheses with same number of foreign words covered

研究院 | 里 | 的 | 研究员 | 每天 | 努力 | 工作 | 。



Better partial
translation

Covers easier part, lower
cost but bad translation

- Hypothesis that covers easy part of sentence is preferred
 - Solution: to consider **future cost** of uncovered parts

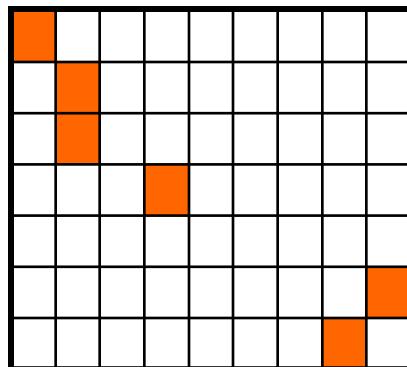
Advantages of Phrase-Based SMT

- Still simple enough
 - but much better than word-based models
- n-to-n mappings can handle non-compositional phrases
 - with all due respect ==> 恕我直言
 - as far as I know ==> 据我所知
- Local context is very useful for disambiguating
 - interest rate ==> 利率
 - interest in ==> ... 方面的兴趣
- The more data, the longer the learned phrases
 - even whole sentences

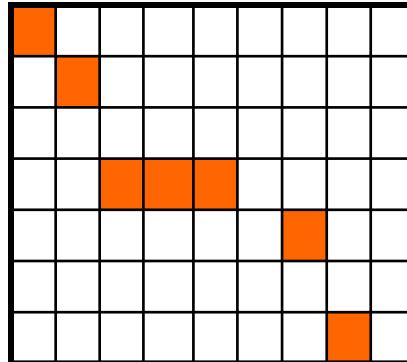
IBM Models are 1-to-Many

- Run IBM-style aligner both directions, then merge:

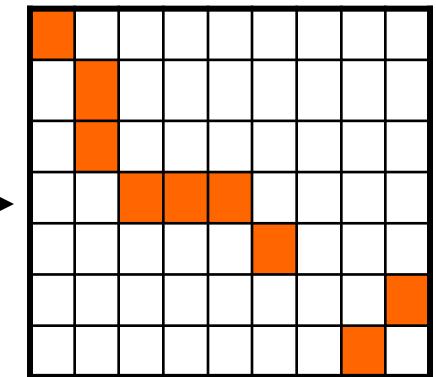
$E \rightarrow F$ best alignment



$F \rightarrow E$ best alignment



MERGE



Union or Intersection

Merge Heuristics

- GDF (intersection-Grow-Diag-Finalization)
 - Better precision
- Union-Reduce
 - Better recall
- Works better than using EM at phrase level

Symmetrizing Word Alignments

English to Chinese

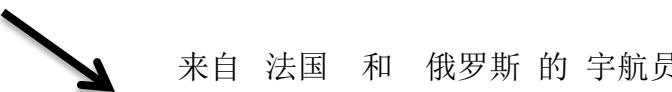
来自 法国 和 俄罗斯 的 宇航员

astronauts						blue
coming						
from	blue					
France		blue				
and			blue			
Russia				blue		

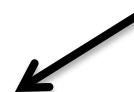
Chinese to English

来自 法国 和 俄罗斯 的 宇航员

astronauts						red
coming	red					
from	red					
France		red				
and			red			
Russia				red		

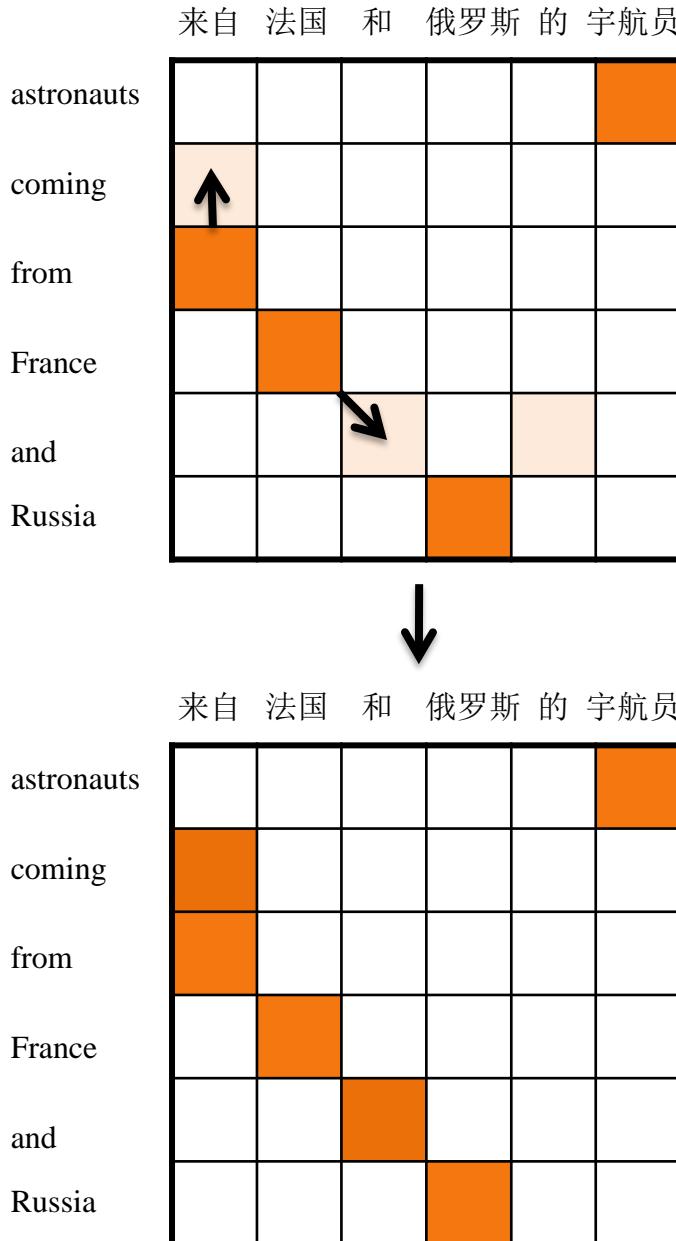


astronauts						orange
coming						
from	orange					
France		orange				
and			orange			
Russia				orange		



(1)
Intersection of bilingual
alignments from GIZA++

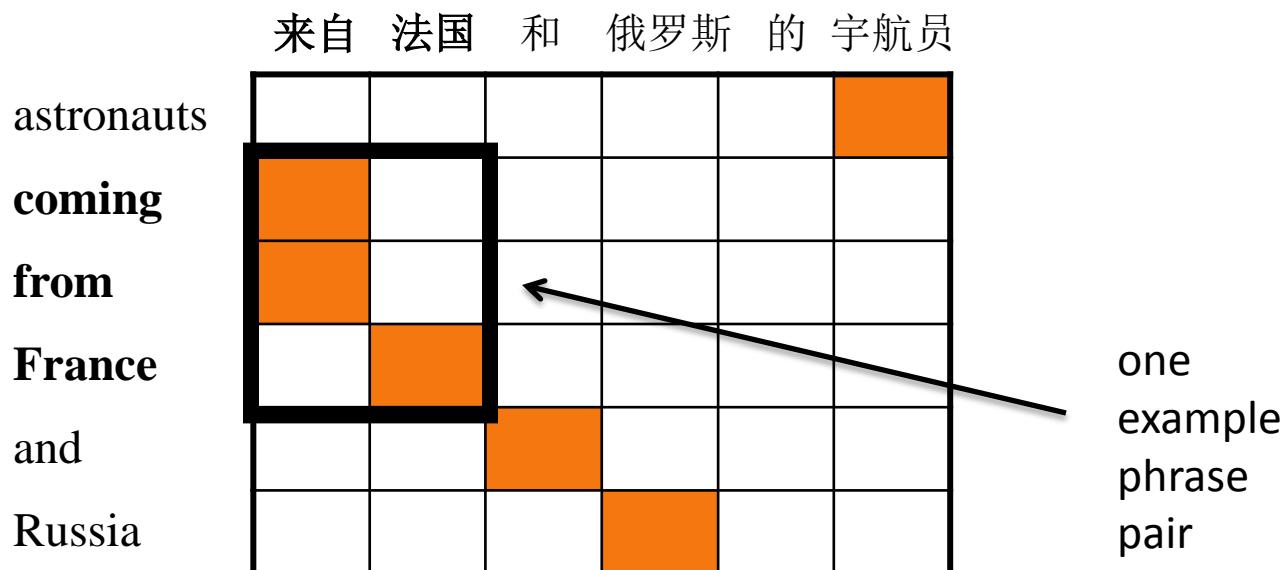
Symmetrizing Word Alignments



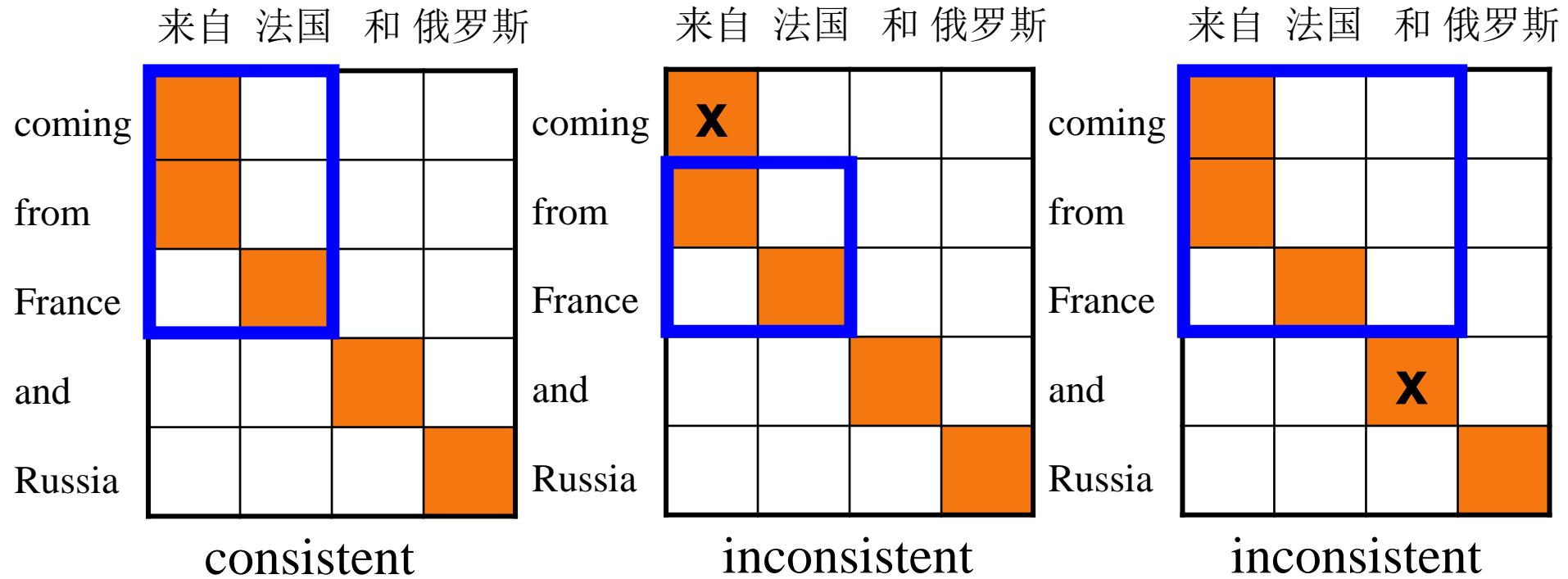
(2)
Grow additional alignment points

How to Learn the Phrase Translation Table?

- Collect all phrase pairs *that are consistent with the word alignment*



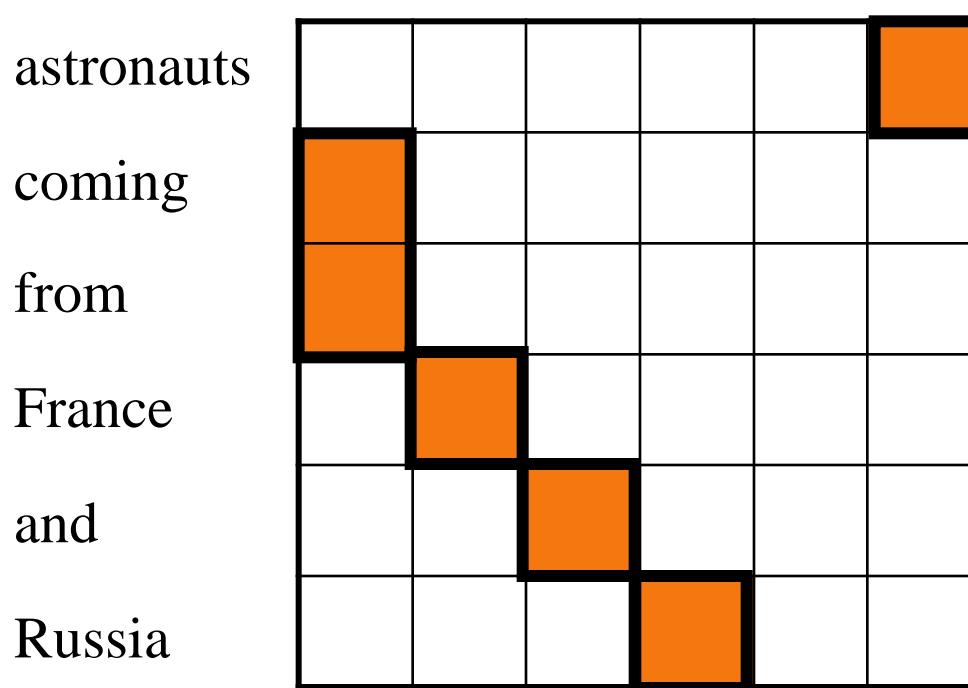
Consistent with Word Alignment



Phrase alignment must contain all alignment points for all the words in both phrases!

Word Alignment Induced Phrases

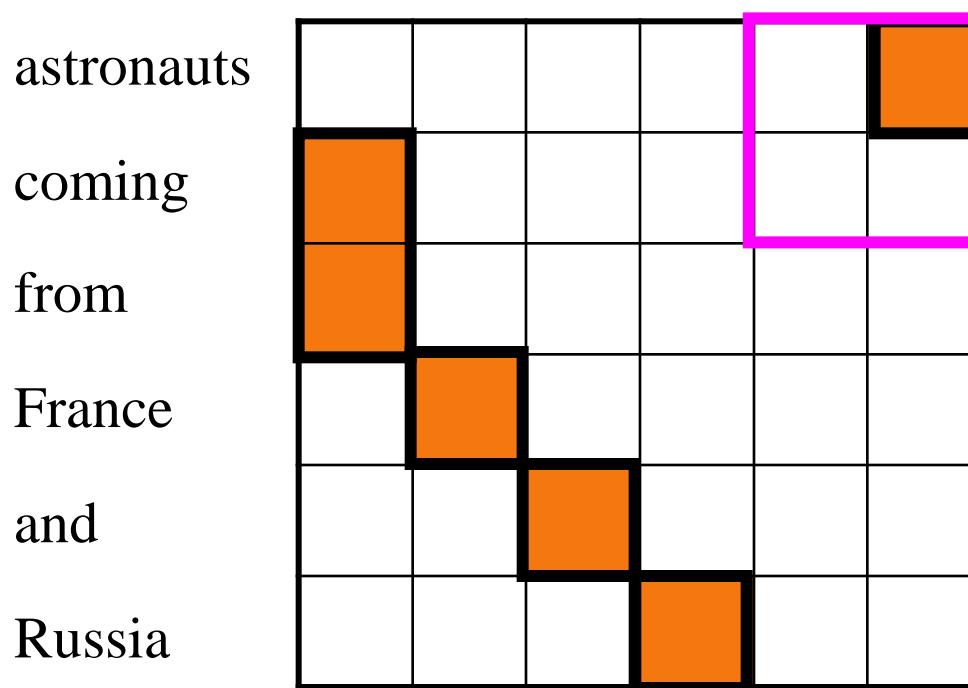
来自 法国 和 俄罗斯 的 宇航员



(宇航员, astronauts) (来自, coming from) (法国, France) (和, and) (俄罗斯, Russia)

Word Alignment Induced Phrases

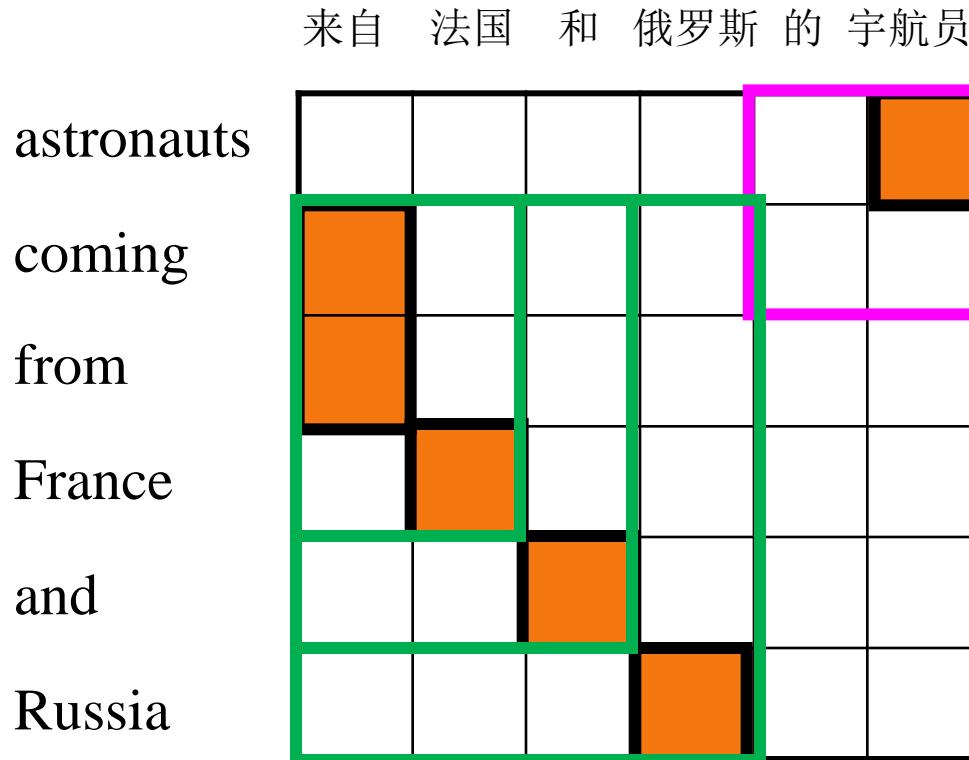
来自 法国 和 俄罗斯 的 宇航员



(宇航员, astronauts) (来自, coming from) (法国, France) (和, and) (俄罗斯, Russia)

(的 宇航员, astronauts) ...

Word Alignment Induced Phrases



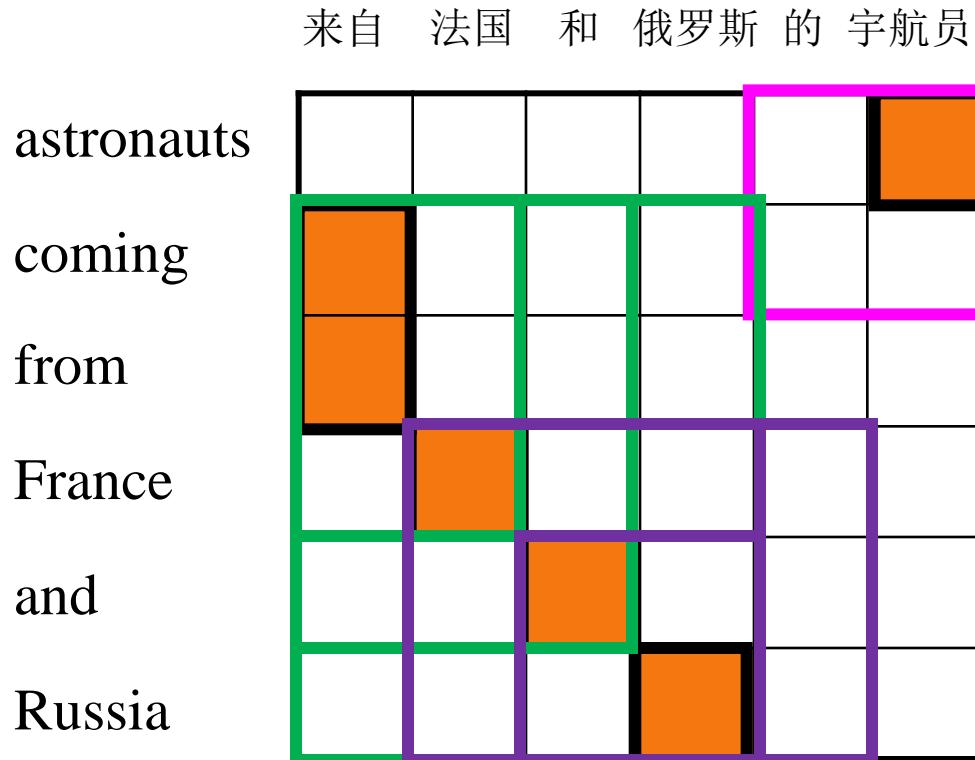
(宇航员, astronauts) (来自, coming from) (法国, France) (和, and) (俄罗斯, Russia)

(的 宇航员, astronauts) ...

(来自 法国, coming from France) (来自 法国 和, coming from France and)

(来自 法国 和 俄罗斯, coming from France and Russia) ...

Word Alignment Induced Phrases



(宇航员, astronauts) (来自, coming from) (法国, France) (和, and) (俄罗斯, Russia)

(的 宇航员, astronauts) ...

(来自 法国, coming from France) (来自 法国 和, coming from France and)

(来自 法国 和 俄罗斯, coming from France and Russia) ...

(和 俄罗斯, and Russia) (法国 和 俄罗斯, France and Russia)

(法国 和 俄罗斯 的, France and Russia) ...

Phrase Pair Probabilities

- A certain phrase pair $(f_1 f_2 f_3, e_1 e_2 e_3)$ may appear many times across the bilingual corpus.
 - And we hope so
- Then there is a vast list of phrase pairs and their frequencies – how to assign probabilities?

Phrase Pair Probabilities

- Basic idea:
 - Relative frequency:
 - $P(f_1f_2f_3, e_1e_2e_3) = \frac{\#(f_1f_2f_3, e_1e_2e_3)}{\#(e_1e_2e_3)}$
- Some important refinements:
 - Smooth using word probs $P(f|e)$ for individual words connected in the word alignment
 - Some low count phrase pairs now have high probability, others have low probability
 - Discount for ambiguity
 - If phrase $(e_1e_2e_3)$ can map to 5 different French phrases, due to the ambiguity of unaligned words, each pair gets a 1/5 count
 - Count **BAD** events too
 - If phrase $(e_1e_2e_3)$ doesn't map onto *any* contiguous French phrase, increment event $\#(\text{BAD}, e_1e_2e_3)$

Log-linear Model for SMT

- Mis-use of translation probability
 - $e^* = \operatorname{argmax}_e P(e) \cdot P(f|e)$
 - $P(f|e) = P_{\text{TM}}(e|f) \cdot P_{\text{LM}}(e)$
- Model scaling in source-channel model
 - $P(f|e) = P_{\text{LM}}(e)^{\lambda_1} \cdot P_{\text{TM}}(f|e)^{\lambda_2}$
 - $\log P(f|e) = \lambda_1 \log P_{\text{LM}}(e) + \lambda_2 \log P_{\text{TM}}(f|e)$
- Generalized log-linear model?
 - $\log P(f|e) = \lambda_1 \log P_{\text{LM}}(e) + \lambda_2 \log P_{\text{TM}}(f|e) + \lambda_3 \log P_{\text{TM}}(e|f)$

Log-linear Model for SMT

- Maximum entropy model for SMT

$$P(e|f) = \frac{1}{Z} \exp \left(\sum_i \lambda_i h_i(f, e) \right)$$

$$e^* = \operatorname{argmax}_e P(e|f) = \operatorname{argmax}_e \sum_i \lambda_i h_i(f, e)$$

- Features

- Log language model probability
- Log translation probability (both directions)
- Number of phrases
- Number of target words
- Distortion cost

Model Training

- GIS algorithms used to learn feature weights
- Problems:
 - Correct translation
 - Oracle translation approximation
 - Normalization factor computation
 - N-best approximation for solution space
 - N-best translations sensitive to model parameter
 - Iterative decoding

Minimum Error Rate Training

- Optimization criteria

- Data likelihood

- $$\bullet \quad \boldsymbol{\lambda}^* = \operatorname{argmax}_{\boldsymbol{\lambda}} \left\{ \sum_{s=1}^S \log p_{\boldsymbol{\lambda}}(e_s | f_s) \right\}$$

- Error count

- $$\bullet \quad \boldsymbol{\lambda}^* = \operatorname{argmin}_{\boldsymbol{\lambda}} \left\{ \sum_{s=1}^S E(r_s, e^*(f_s, \boldsymbol{\lambda})) \right\}$$

- Smoothed error count

- $$- \quad \boldsymbol{\lambda}^* = \operatorname{argmin}_{\boldsymbol{\lambda}} \left\{ \sum_{s,k} E(e_{s,k}) \frac{p(e_{s,k}|f)^\alpha}{\sum_k p(e_{s,k}|f)^\alpha} \right\}$$

Unsmoothed vs. Smoothed Error Count

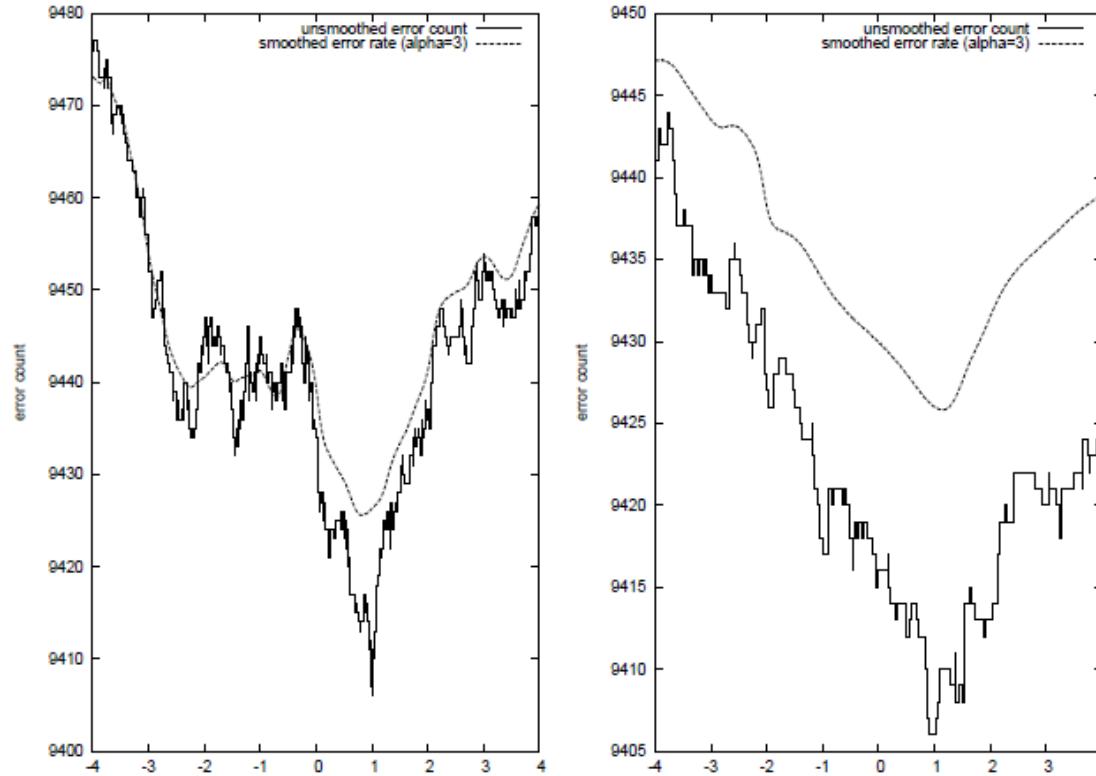
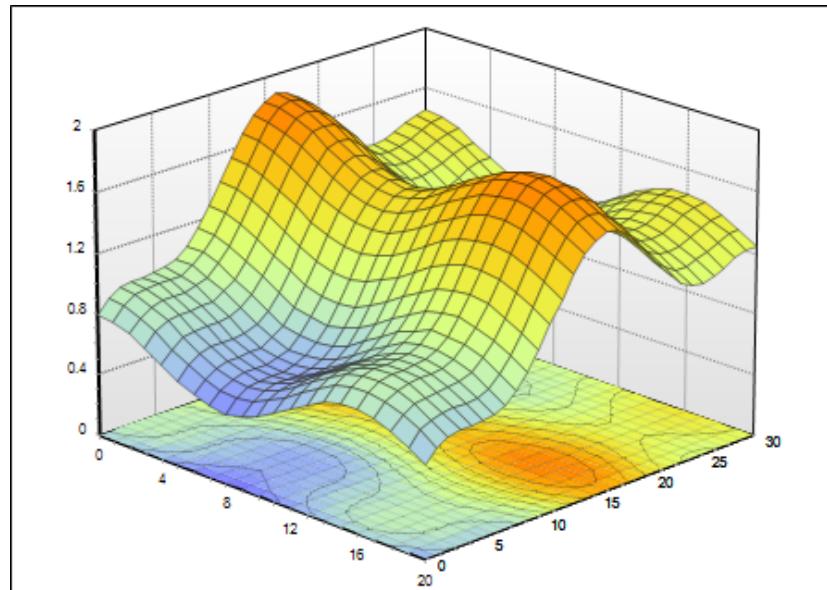


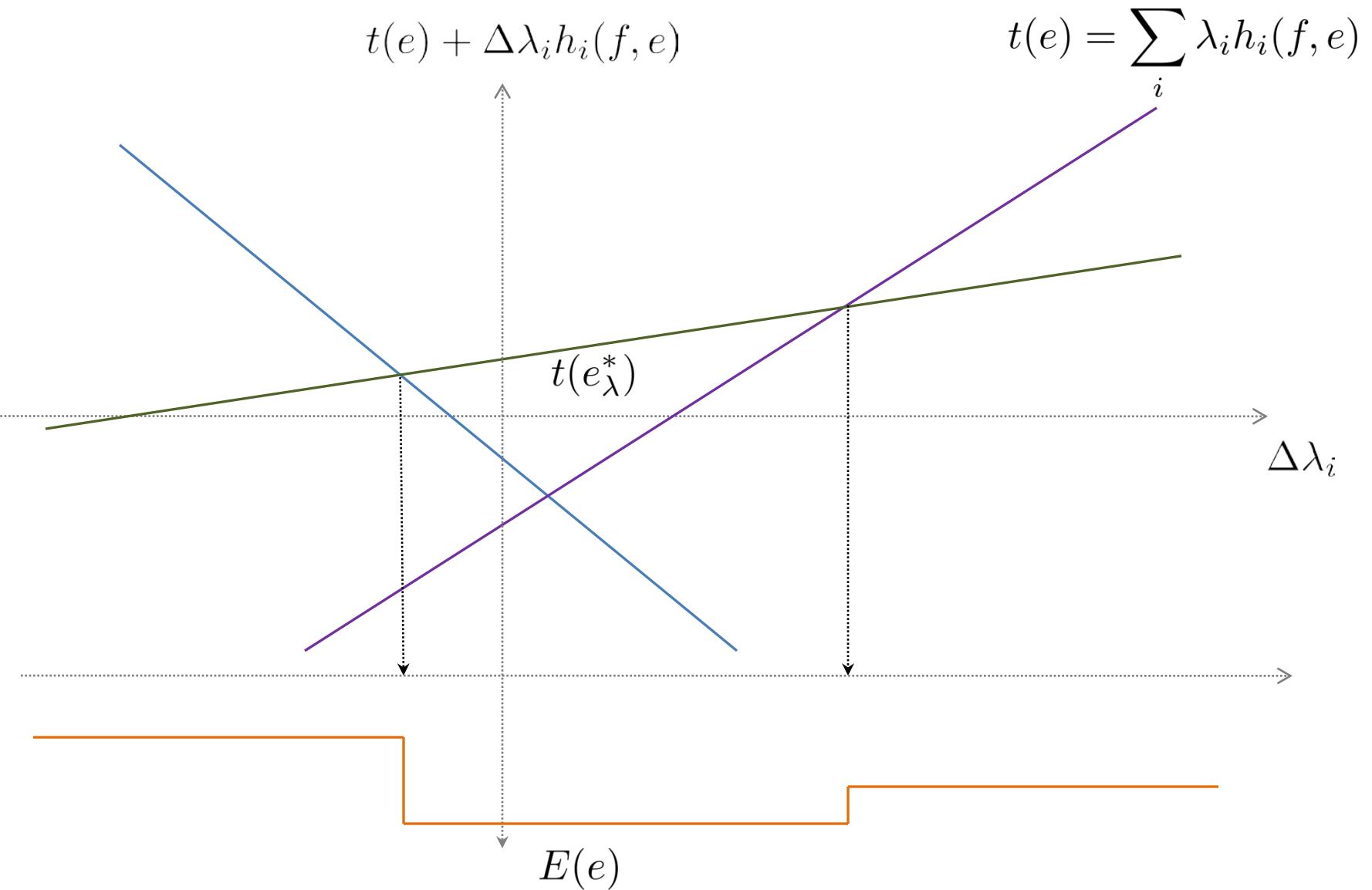
Figure 1: Shape of error count and smoothed error count for two different model parameters. These curves have been computed on the development corpus (see Section 7, Table 1) using 1,600 alternatives per source sentence. The smoothed error count has been computed with a smoothing parameter $\alpha = 3$.

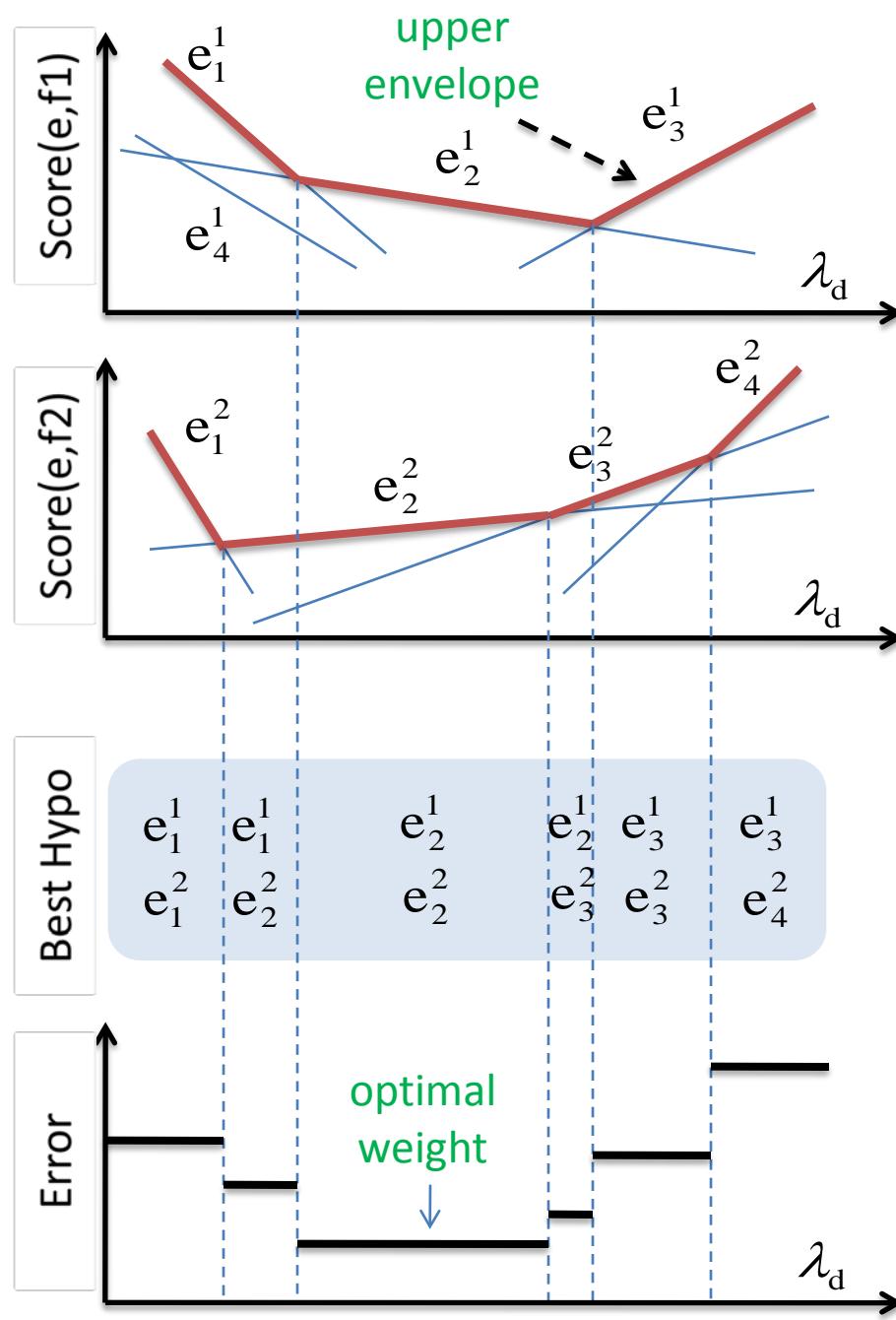
Grid-base Line Search

- Issues
 - Local maximum
 - Efficiency



Error Surface

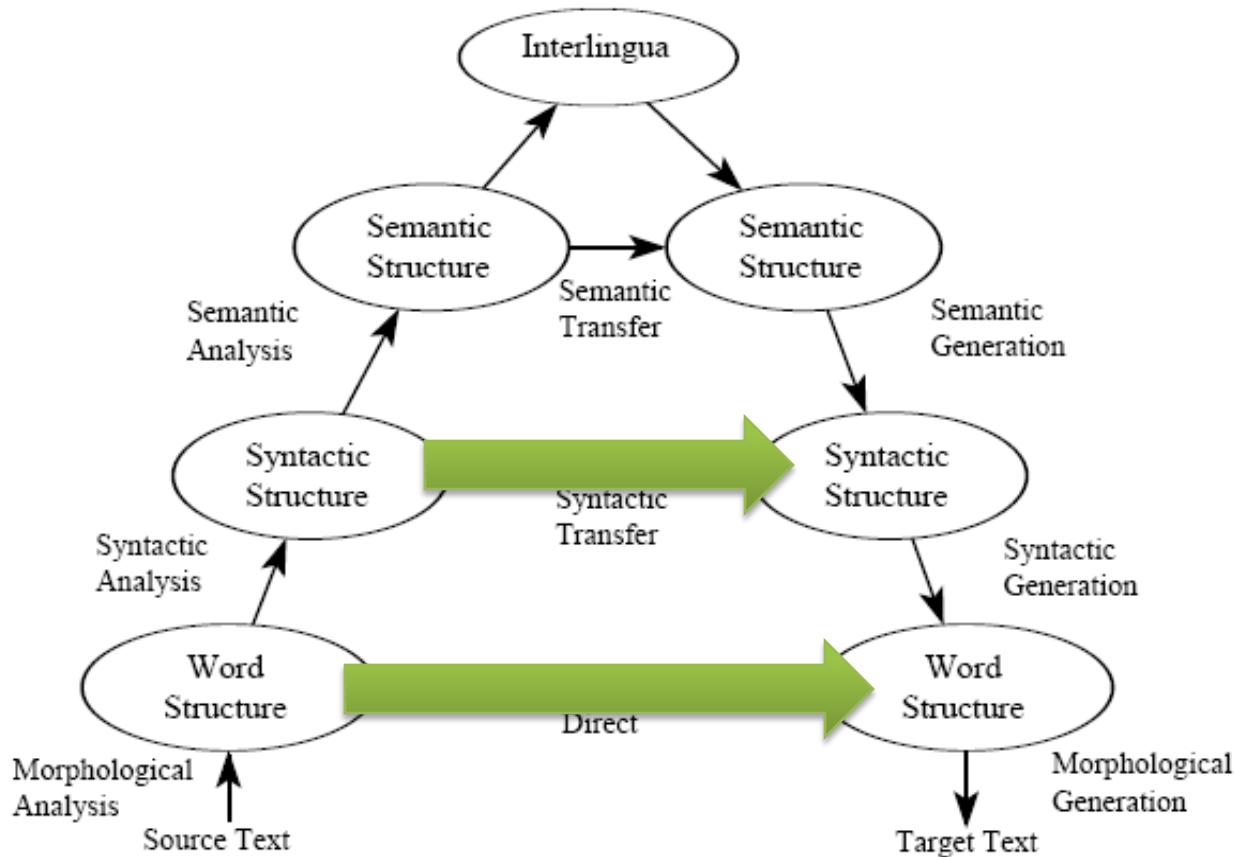




Comments on MERT

- Random start to work around local maxima
- Effective when
 - Solution space is limited
 - Feature space is small (< 20)
- Still need iterative decoding

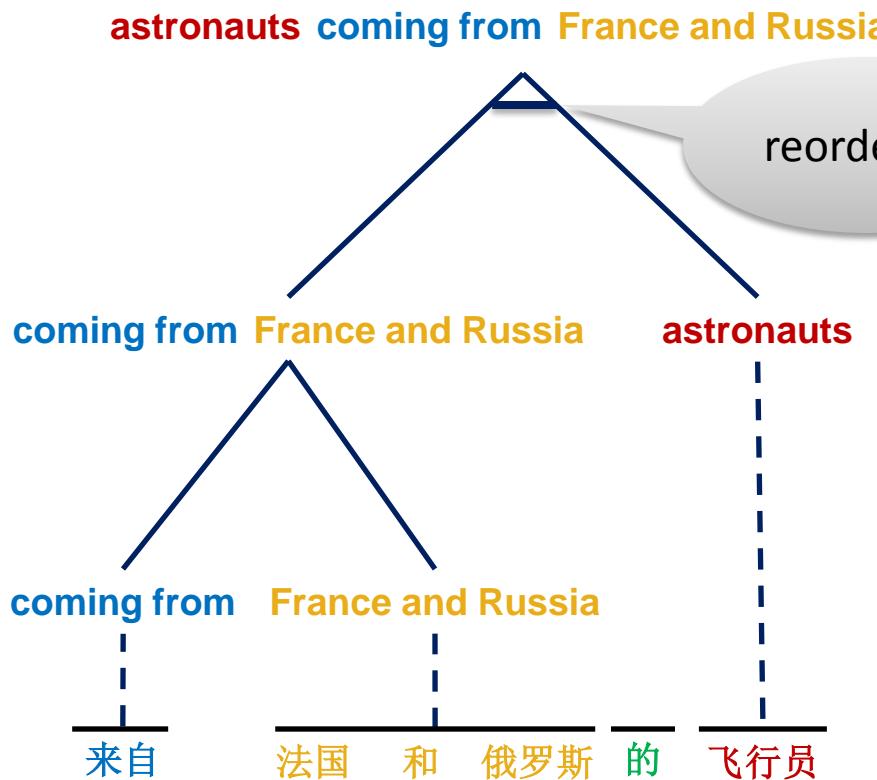
Machine Translation Pyramid



Syntax-based SMT Models

- Formal synchronous syntax
 - BTG (Bracketing Transduction Grammar)
 - $X \rightarrow [XX]$ $X \rightarrow <XX>$ $X \rightarrow \alpha, \gamma$
 - Hiero rules
 - $X \rightarrow <\gamma, \alpha, \sim>$
 - $x \rightarrow$ 与 x_1 有 x_2 , have X_2 with X_1
- Linguistic synchronous syntax
 - $\text{NP} \rightarrow \text{VP}_1 (\text{NP} (\text{NNS} (\textit{fei-xing-yuan})))$, astaurants VP_1

BTG decoding



$X_1 X_2 \parallel\parallel X_2 X_1$

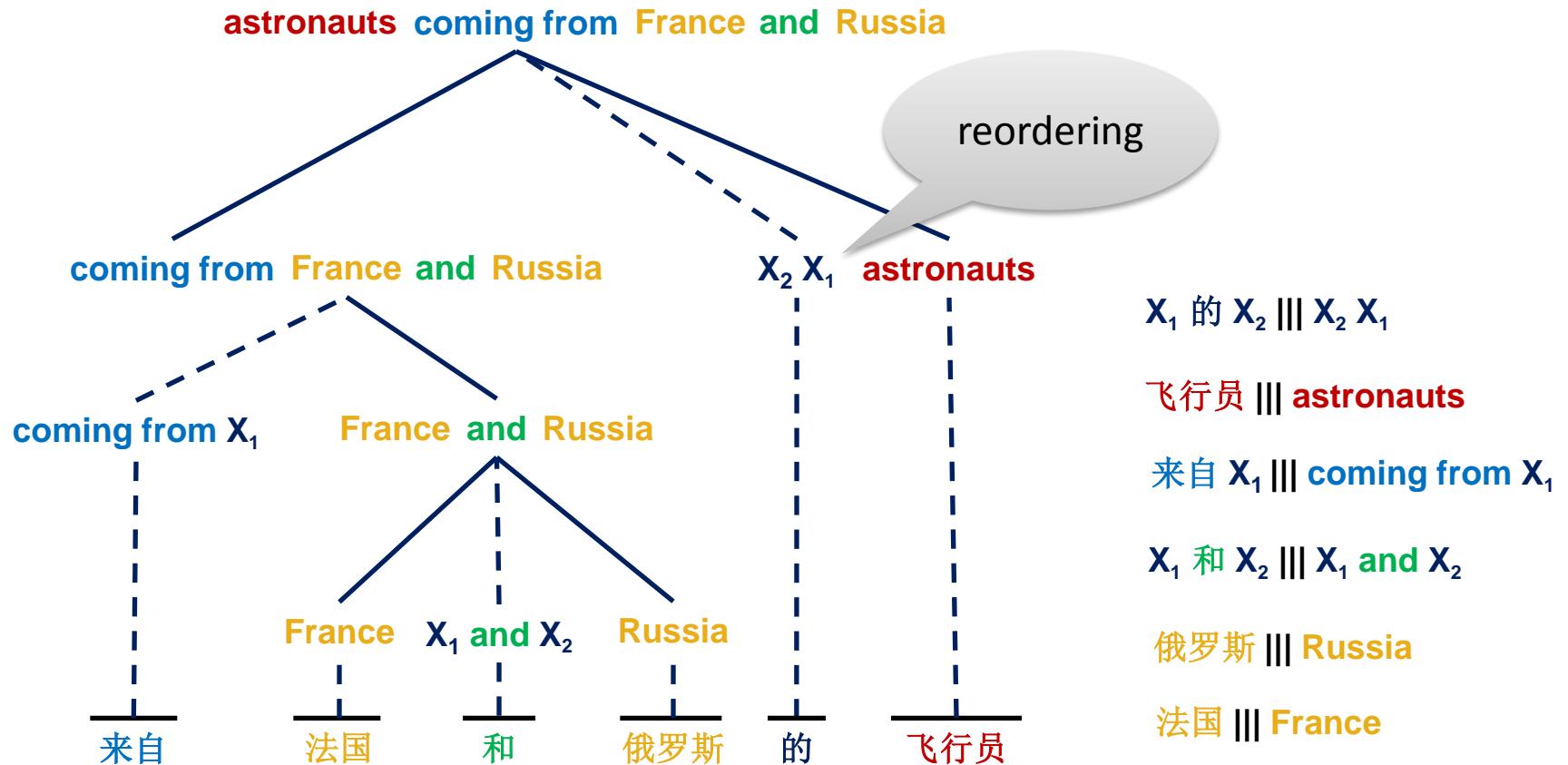
飞行员 ||| astronauts

$X_1 X_2 \parallel\parallel X_1 X_2$

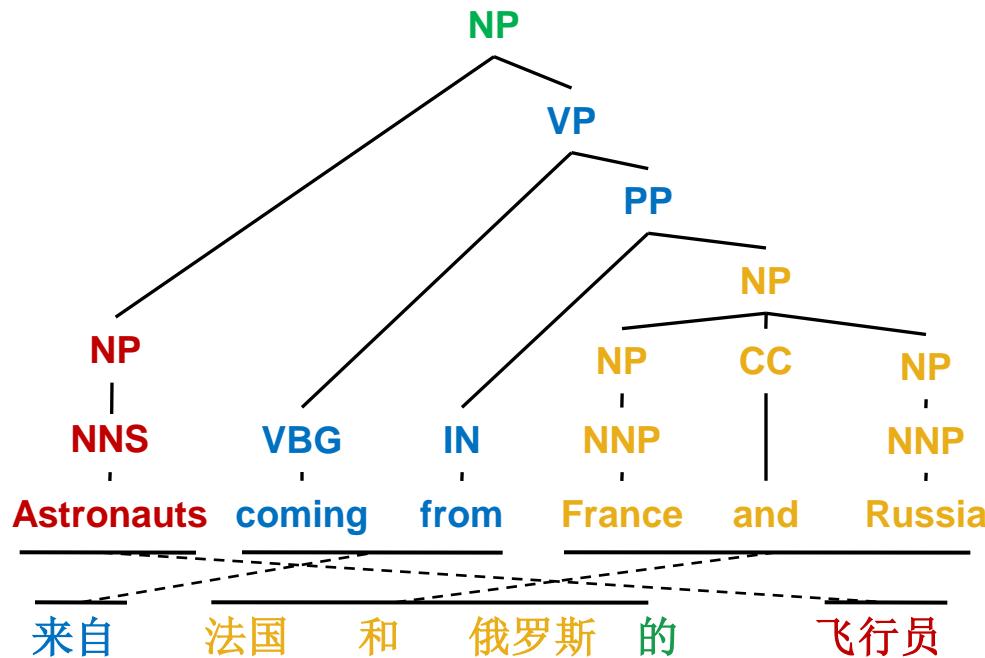
法国 和 俄罗斯 ||| France and Russia

来自 ||| coming from

Hiero decoding



Syntax-based Model



SMT Bet

- String models vs. syntax-based models

	Arabic-English	Chinese-English
Google	42.81	33.16
ISI	39.08	33.93

NIST 2006 MT Evaluation results

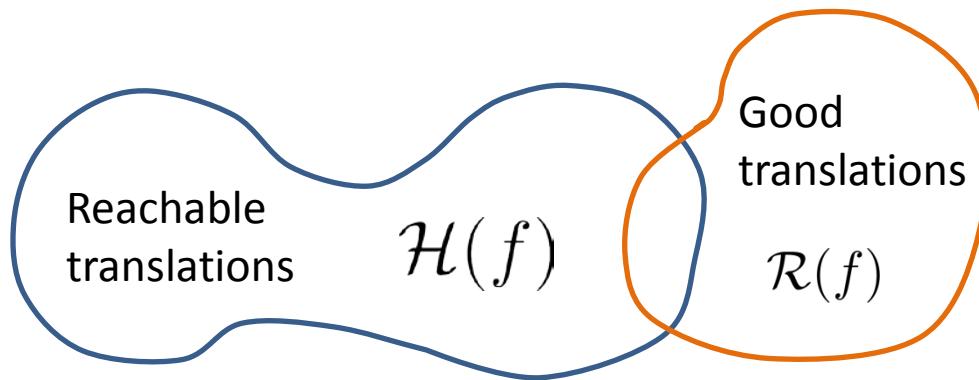
Consensus-based SMT Decoding

- Minimum Bayes Risk (MBR) decoding
 - Consensus decoding using one system
- System combination
 - Consensus decoding using multiple systems

MAP Decoding

- Maximum A Posteriori (MAP) decision rule

$$e^* = \operatorname{argmax}_{e \in \mathcal{H}(f)} P(e|f)$$



Minimum Bayes-Risk Decoding

Finding consensus within one system

- Risk function

- $R(e) = \sum_{r \in \mathcal{R}(f)} L(e, r)P(r|f)$ $\mathcal{R}(f) = (r_1, r_2, \dots, r_n)$

- MBR decision Rule

- Ideal MBR

$$e^* = \operatorname{argmin}_e R(e) = \operatorname{argmin}_{e \in \mathcal{H}(f)} \sum_{r \in \mathcal{R}(f)} L(e, r)P(r|f)$$

- MBR in practice

$$e^* = \operatorname{argmin}_e R(e) = \operatorname{argmin}_{e \in \mathcal{H}(f)} \sum_{e' \in \mathcal{H}(f)} L(e, e')P(e'|f)$$

Minimum Bayes-Risk Decoding

Finding consensus within one system

- Loss vs. Gain

- $L(e, e') = C(f) - G(e, e')$
- $e^* = \operatorname{argmax}_{e \in \mathcal{H}(f)} \sum_{e' \in \mathcal{H}(f)} G(e, e') P(e'|f)$

- Consensus measure

- Unigram overlapping
- N-gram overlapping
- Structure overlapping

Literature of MBR Decoding

- Speech recognition
 - Bickel and Doksum, 1977
- Statistical machine translation
 - Kumar and Byrne, 2004
 - Tromble et al., 2008
 - DeNero et al., 2009
 - Li et al., 2009

System Combination

Finding consensus between systems

- Combining outputs from multiple machine translation systems
 - Rosti et al., NAACL 2007
 - Sentence-level, phrase-level, word-level
- Relation between word-level combination consensus decoding

$$- c(e|f) = \sum_{w_i \in e = w_1, \dots, w_n} c(w_i) + \mu N_{null}(e) \quad c(w_i) = \sum_j^{N_s} \lambda_j c(w_i, j)$$

- More work
 - Confusion network decoding
 - Bangalore et al., 2001
 - Matusov et al., 2006
 - Sim et al., 2007
 - Better word alignment
 - Rosti et al., 2008
 - Xiaodong He et al., 2008
 - Li et al., 2009

Combination Approach

- Word-level system combination

- N-best translation generation
- Skeleton translation selection
- Confusion network construction
- Confusion network decoding



Combination Method

- Word-level system combination

- N-best translation generation
- Skeleton translation selection
- Confusion network construction
- Confusion network decoding

I like eating chocolate icecream.



我 喜欢 巧克力 冰激凌。
我 喜欢 吃 巧克力 。
我 爱 吃 巧克力 冰淇凌 。
我 爱 巧克力 冰激凌 。
.....



我 喜欢 巧克力 冰激凌。

Combination Method

- Word-level system combination

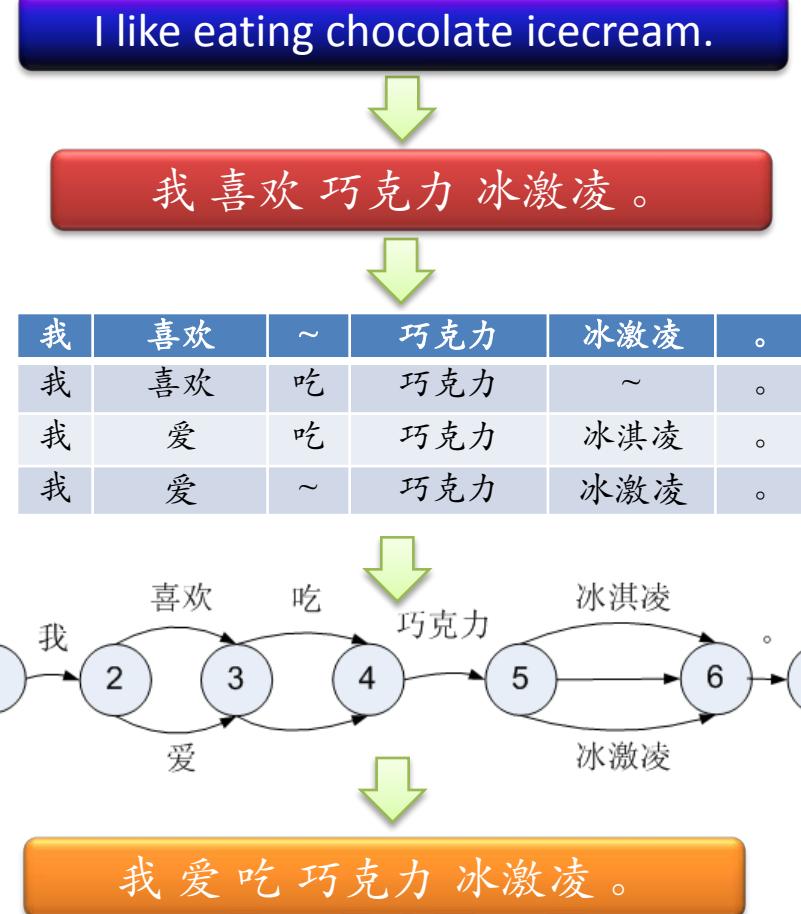
- N-best translation generation
- Skeleton translation selection
- Confusion network construction
- Confusion network decoding



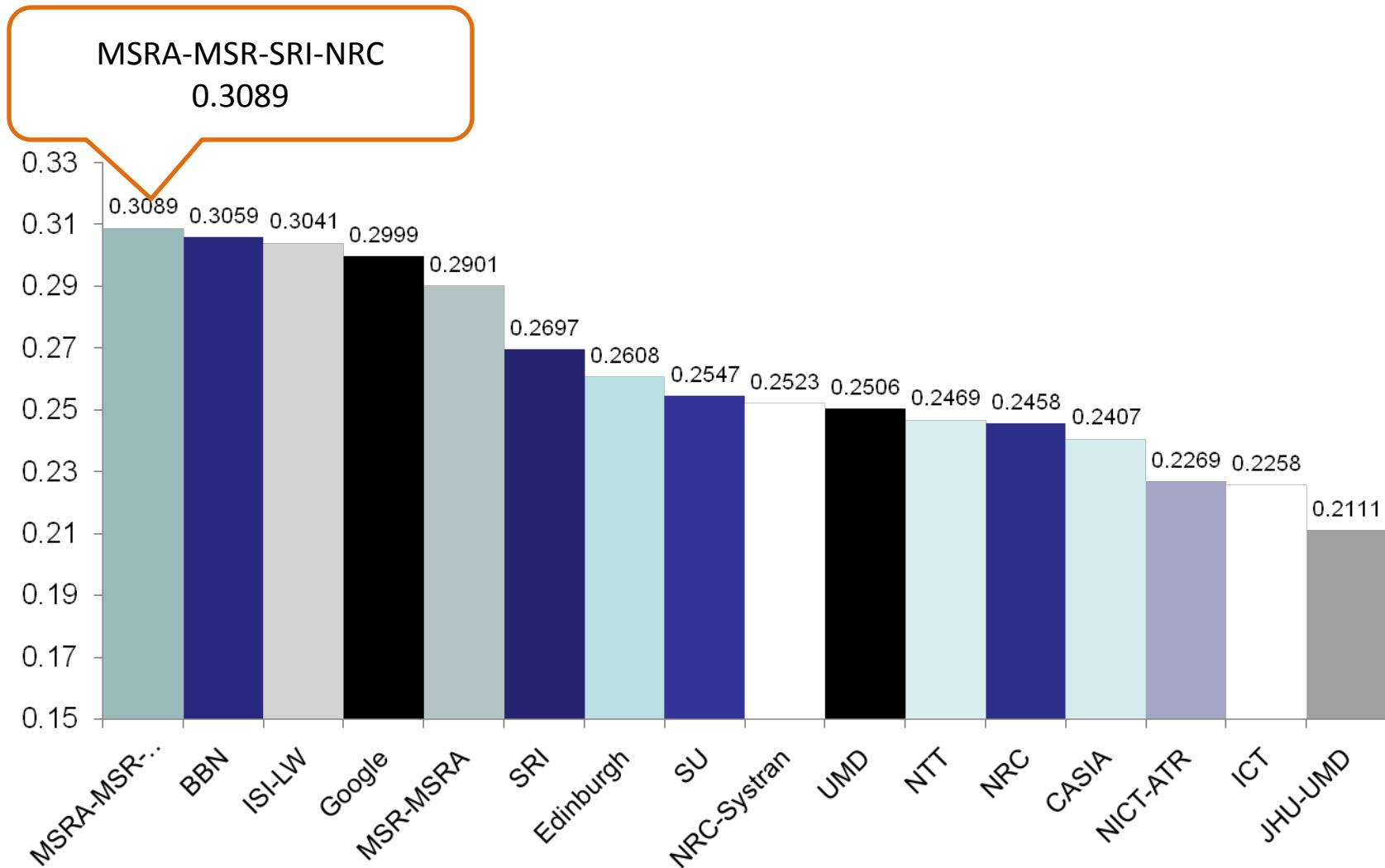
Combination Method

- Word-level system combination

- N-best translation generation
- Skeleton translation selection
- Confusion network construction
- Confusion network decoding

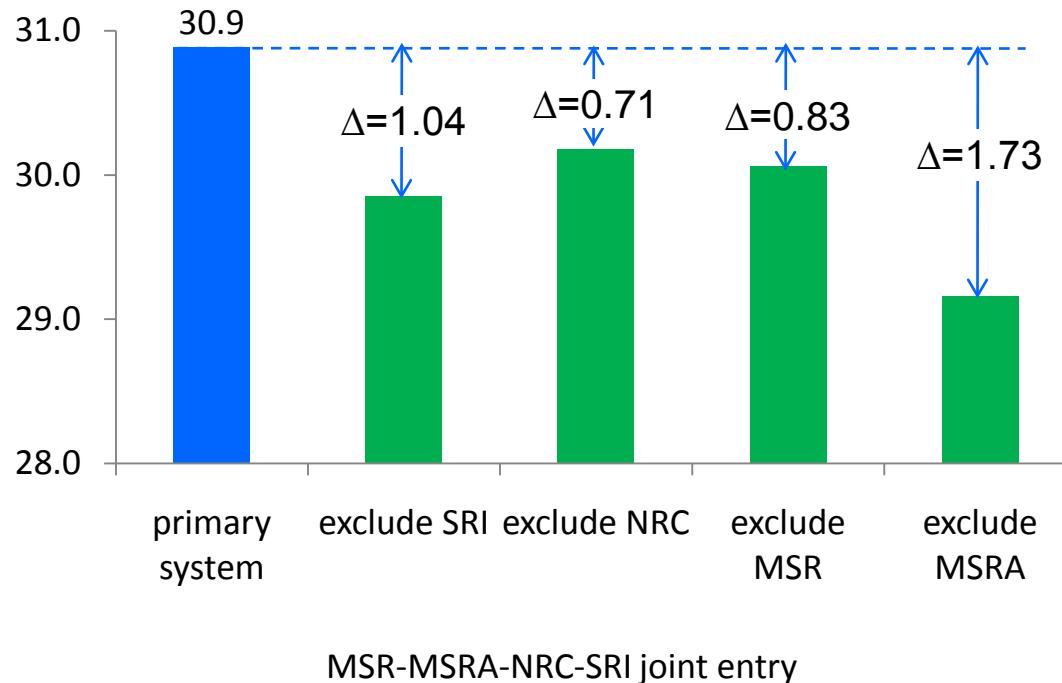


Chinese-English Results



Impact of Systems of Each Site

- Each site provided individual systems for combination
 - MSR: 3 systems, MSRA: 3 systems, NRC:1 system, SRI: 1 system
- The impact is measured by the BLEU score loss due to excluding system(s) of that site



Machine Translation Evaluation

- Human evaluation
- 信达雅
 - Adequacy
 - I cannot agree you more → 我不能同意你更多
 - Fluency
 - How old are you → 怎么老是你
 - High cost
- Automatic evaluation
 - Convenient
 - May not be consistent with human preference

BLEU Evaluation Metric

(Papineni et al, ACL-2002)

Reference (human) translation:

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

Machine translation:

The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

- N-gram precision (score is between 0 & 1)
 - What percentage of machine n-grams can be found in the reference translation?
 - An n-gram is a sequence of n words
 - Not allowed to use same portion of reference translation twice (can't cheat by typing out "the the the the")
- Brevity penalty
 - Can't just type out single word "the" (precision 1.0!)

*** Amazingly hard to "game" the system (i.e., find a way to change machine output so that BLEU goes up, but quality doesn't)

BLEU Metric

$$BLEU = BP \bullet \exp\left(\sum_1^N w_n \log p_n\right)$$

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

$$\log BLEU = \min\left(1 - \frac{r}{c}, 0\right) + \sum_1^N w_n \log p_n$$

$$N = 4, w_n = 1/N$$

BLEU: An Example

- Candidate 1: the book is on the desk

- Reference 1: there is a book on the desk
- Reference 2: the book is on the table

unigram:	bigram:	trigram:
$Count_{clip}(the) = 2$	$Count_{clip}(the, book) = 1$	$Count_{clip}(the, book, is) = 1$
$Count_{clip}(book) = 1$	$Count_{clip}(book, is) = 1$	$Count_{clip}(book, is, on) = 1$
$Count_{clip}(is) = 1$	$Count_{clip}(is, on) = 1$	$Count_{clip}(is, on, the) = 1$
$Count_{clip}(on) = 1$	$Count_{clip}(on, the) = 1$	$Count_{clip}(on, the, desk) = 1$
$Count_{clip}(desk) = 1$	$Count_{clip}(the, desk) = 1$	
$\sum_{unigram \in C} Count(unigram) = 6$	$\sum_{bigram \in C} Count(bigram) = 5$	$\sum_{trigram \in C} Count(trigram) = 4$
$p_1 = 1$	$p_2 = 1$	$p_3 = 1$

$$\left. \begin{array}{l} c=6 \\ r=6 \end{array} \right\} = e^{1-\frac{r}{c}} = e^0 = 1 = BP$$

$$BLEU = BP \bullet \exp \left(\sum_{n=1}^N w_n \log p_n \right)$$

$$= \exp \left[\frac{1}{3} (\log 1 + \log 1 + \log 1) \right] = \sqrt[3]{1 \cdot 1 \cdot 1} = 1$$

BLEU

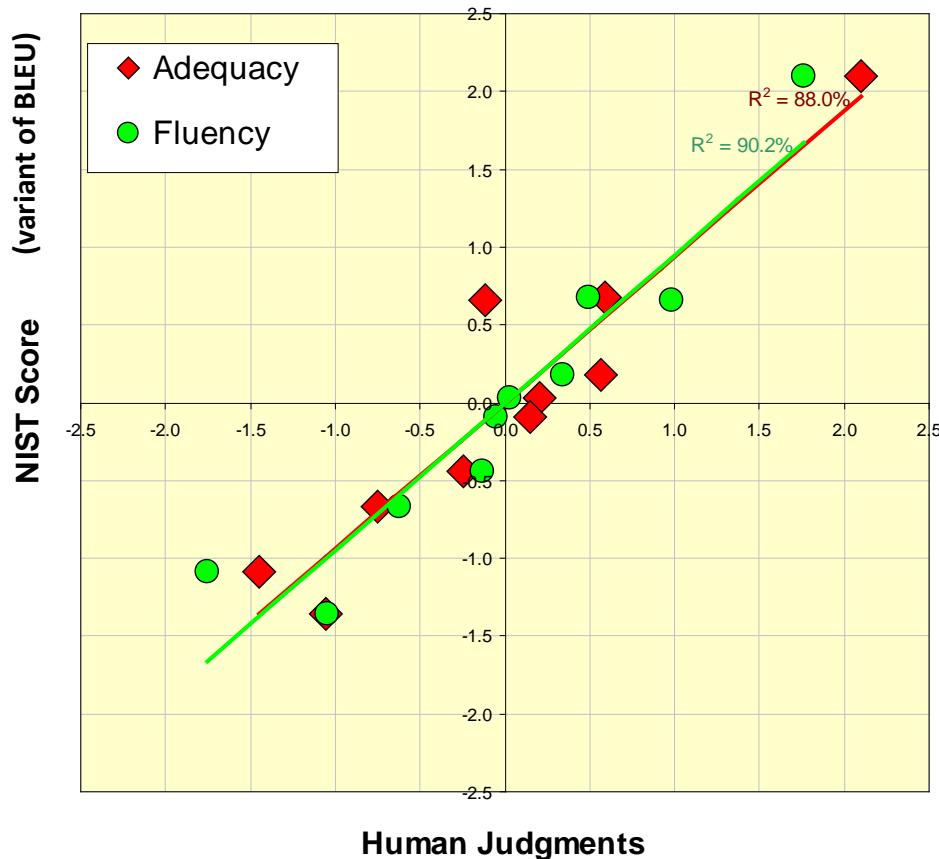
I do not speak French . reference

I not speak French . output

$$\left(\frac{5}{5} \times \frac{3}{4} \times \frac{2}{3} \times \frac{1}{4}\right)^{\frac{1}{4}} \times e^{1-5/6}$$

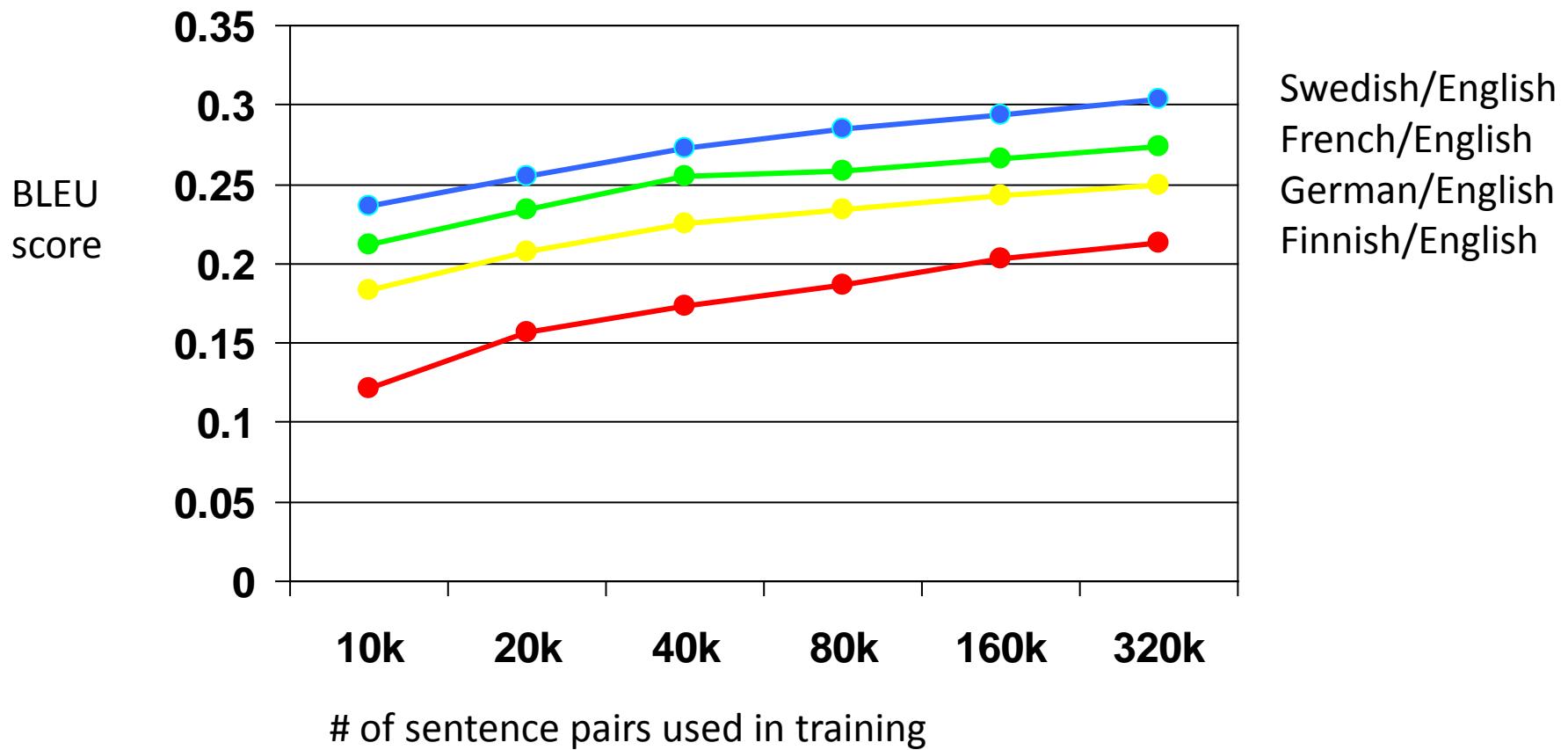
1-grams 2-grams 3-grams 4-grams brevity penalty

BLEU Tends to Predict Human Judgments



slide from G. Doddington (NIST)

Observing Learning Curves using Bleu



Experiments by
Philipp Koehn

More Comments on BLEU

- Cannot be used to evaluate human translators
- Not work well for heterogeneous systems
 - E.g. a statistical and a rule-based system
- Test set should be large enough
- Marginal improvements may not be meaningful

Other Metrics

- Metrics based on lexical similarity
 - (most of the metrics!)
- Edit Distance
 - WER, PER, TER
- Precision
 - BLEU, NIST, WNM
- Recall
 - ROUGE, CDER
- Precision/Recall
 - GTM, METEOR, BLANC, SIA

Recommend Readings

1. [A Statistical MT Tutorial Workbook](#). Kevin Knight. 1999.
Very good introduction to word-based statistical machine translation.
Written in an informal, understandable, tutorial oriented style.
2. [The Mathematics of Statistical Machine Translation: Parameter Estimation](#). P. F. Brown, S. A. Della Pietra,
V. J. Della Pietra and R.L. Mercer. 1993.
3. Phrase based statistical MT:
[Statistical Phrase-Based Translation](#).
Philipp Koehn, Franz Jasof Ock and Daniel Marcu. 2003.
4. [Discriminative Training and Maximum Entropy Models for Statistical Machine Translation](#).
Och and Ney. 2002.
5. [BLEU: A Method for Automatic Evaluation of Machine Translation](#).
Papineni, Roukos, Ward and Zhu. 2001.

CCF ADL 2011
Beijing
Aug. 27, 2011

Learning to Match

Hang Li
Microsoft Research Asia

Talk Outline

- Introduction to Web Search
- Relevance Model (Matching Model)
- Query Term Mismatch
- Learning to Match
- Our Methods
 - Robust Similarity Function Learning Using Kernel Methods
 - Regularized Latent Semantic Indexing
 - Query Generation Using Log Linear Model
 - Query Rewriting Using Conditional Random Fields

Introduction to Web Search

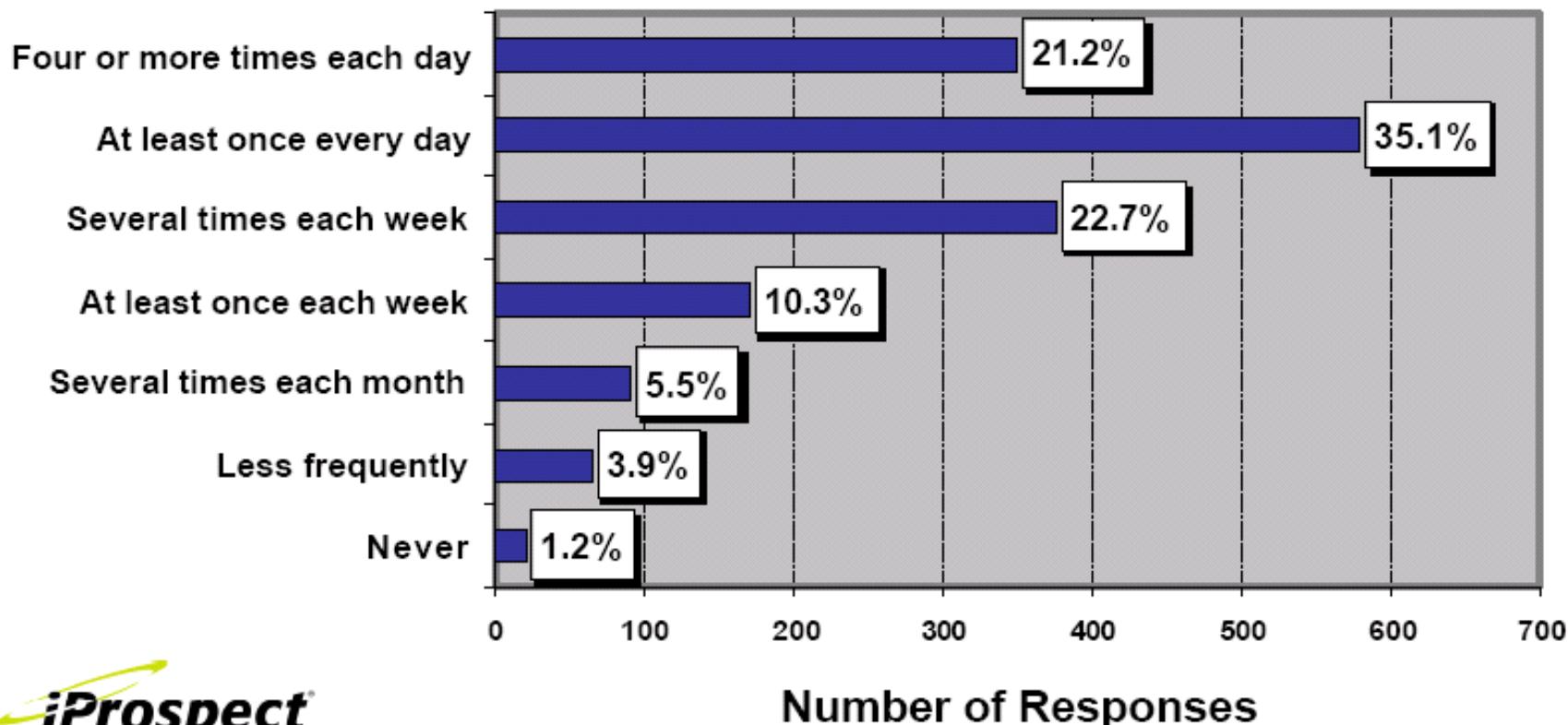
Web Search is Part of Our Life



Web Users Heavily Rely on Search Engines

<http://www.iprospect.com/premiumPDFs/iProspectSurveyComplete.pdf>

How often do you use search engines on the Internet?



Simple UI

The screenshot shows a Microsoft Internet Explorer window displaying search results for the query "web information retrieval". The search bar at the top contains the query. Below the search bar, there is a red box highlighting the search results area. The results include a mix of sponsored ads and organic search results. The first few results are from "Adobe SiteSearch" and "Homestead® Websites". The main organic search results are for "Tutorial: Web Information Retrieval", "Information retrieval - Wikipedia, the free encyclopedia", "Information retrieval: Definition from Answers.com", "Introduction to Information Retrieval", "Web Information Retrieval - Microsoft Academic Search", and "Web Information Retrieval — Universität Koblenz-Landau". A second red box highlights the "Information retrieval - Wikipedia, the free encyclopedia" result, which provides a brief overview of what queries are and how they are processed by search engines.

Search box

Ranking result

Internal privacy Help improve Bing

Internet Protected Mode On

111%

Huge Data Center



Goal of Web Search

Effectiveness	Efficient	Easy to Use
Results are relevant	Response time is short	Good presentation
Results are comprehensive	Results are novel	Friendly user interface

Overview of Web Search Technologies

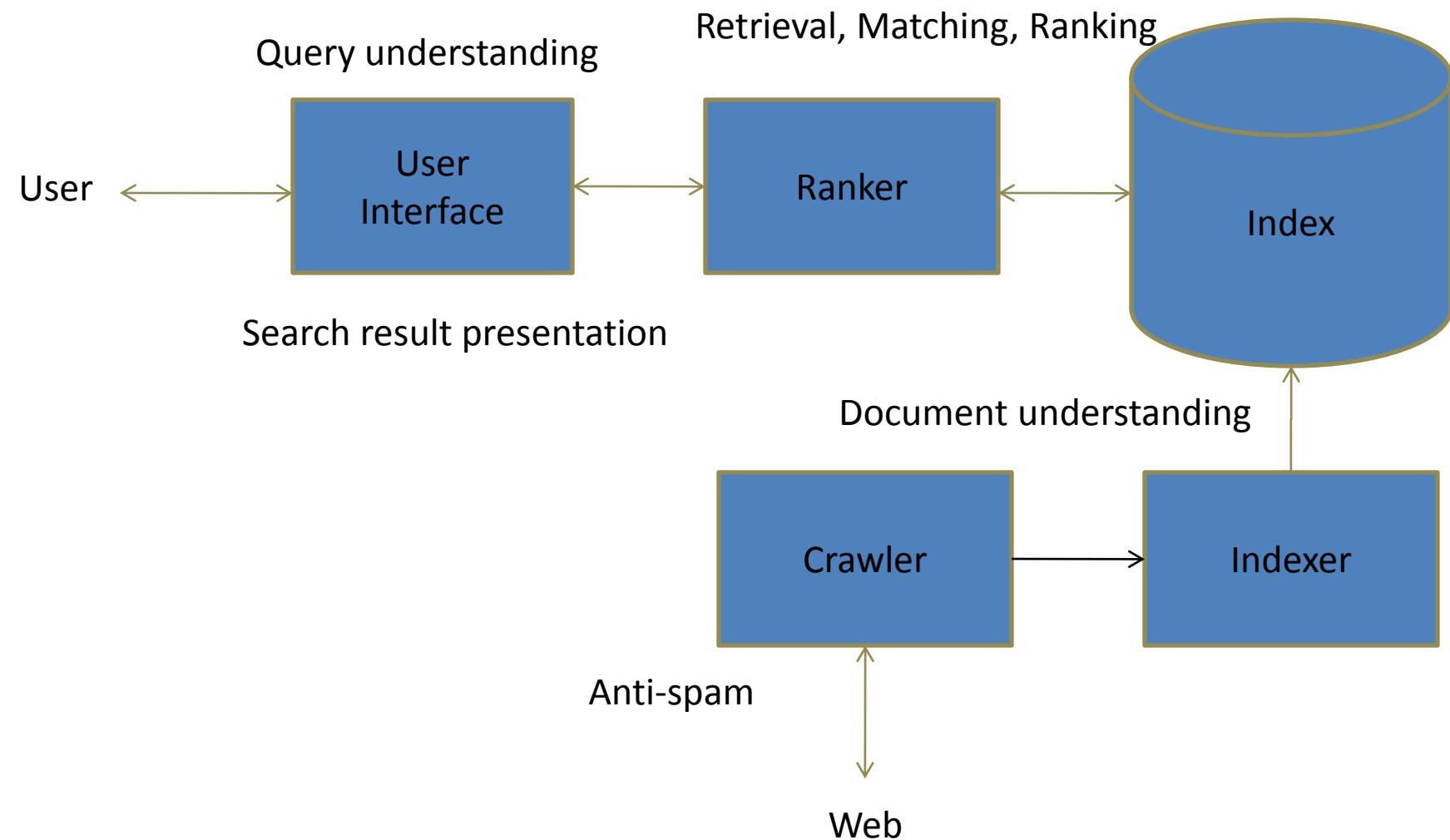
Overview of Web Search Technologies

General Web Search, Entity Search, Facet Search,
Question Answering, Image Search

Ranking, Matching, Retrieval
Document Understanding, Query Understanding,
Crawling, Indexing, Result Presentation,
Anti-Spam

Classification, Clustering, Ranking,
Graph Learning, Tagging, Distributed Computing

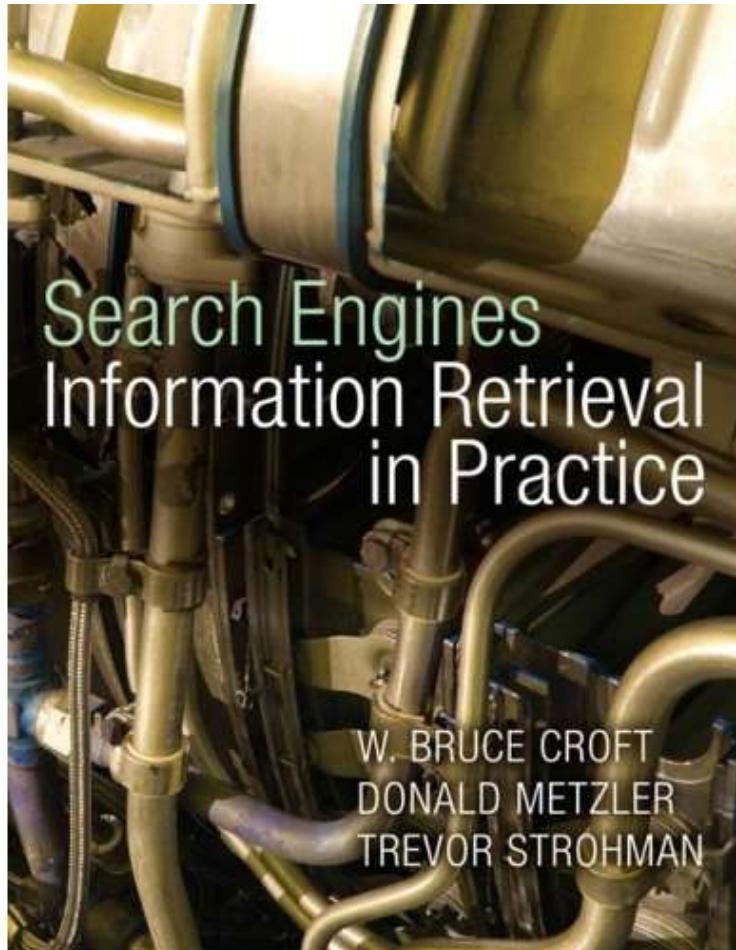
Example of Web Search Architecture



Component Technologies for Web IR

- Relevance Ranking
- Importance Ranking
- Document Understanding
- Query Understanding
- User Understanding
- Crawling
- Indexing
- Search Result Presentation
- Anti-Spam
- Search Log Data Mining

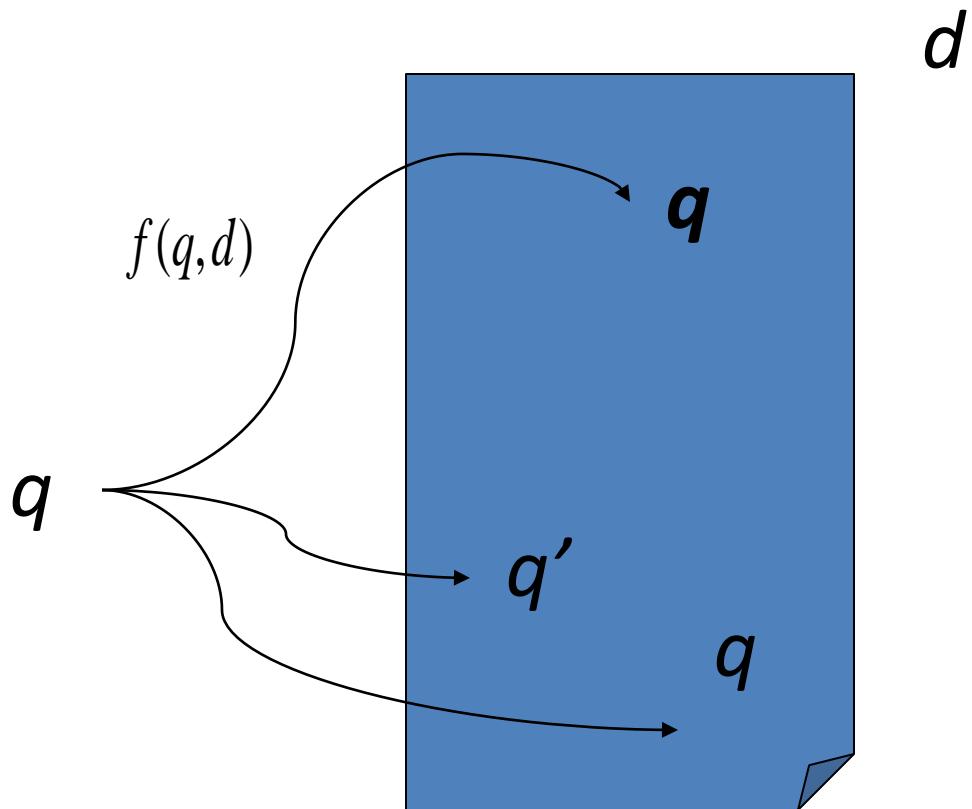
Book by Croft et al.



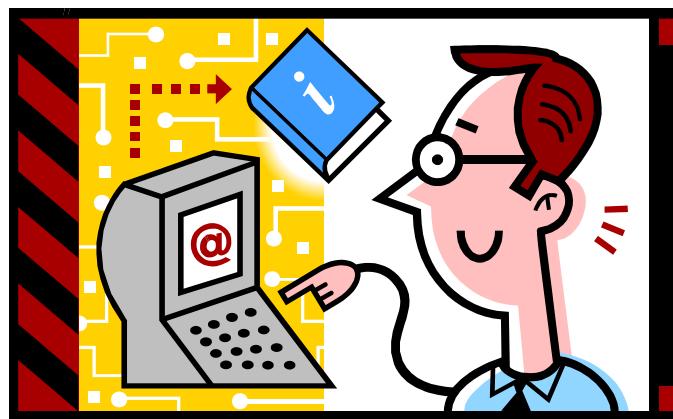
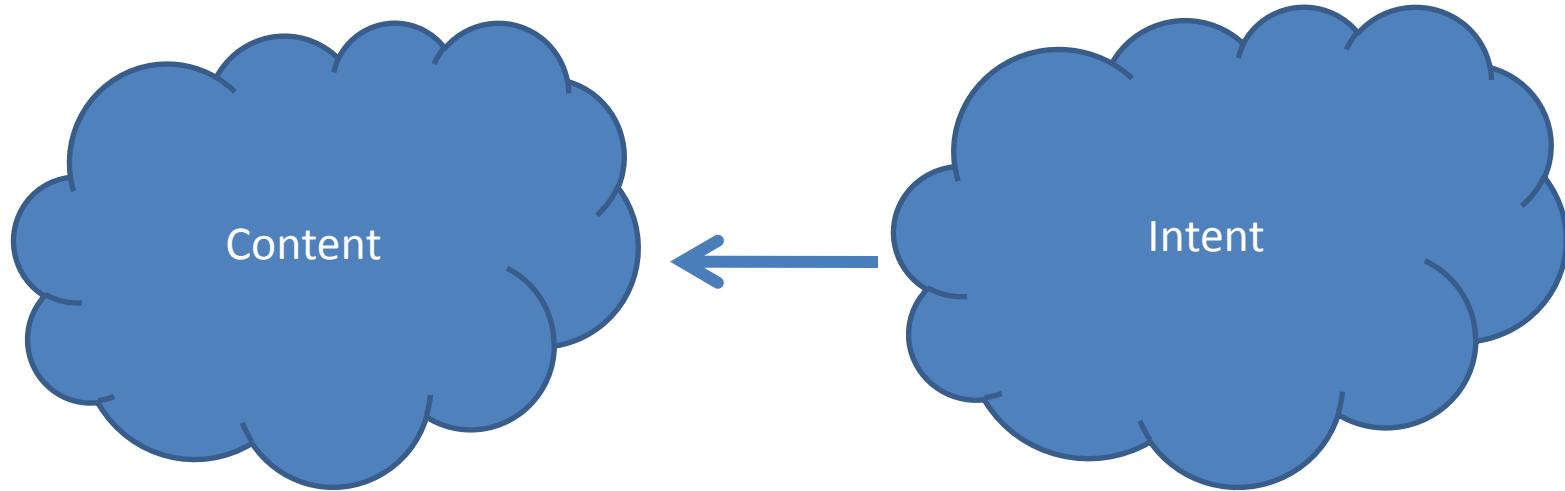
<http://www.search-engines-book.com/>

Relevance Model (Matching Model)

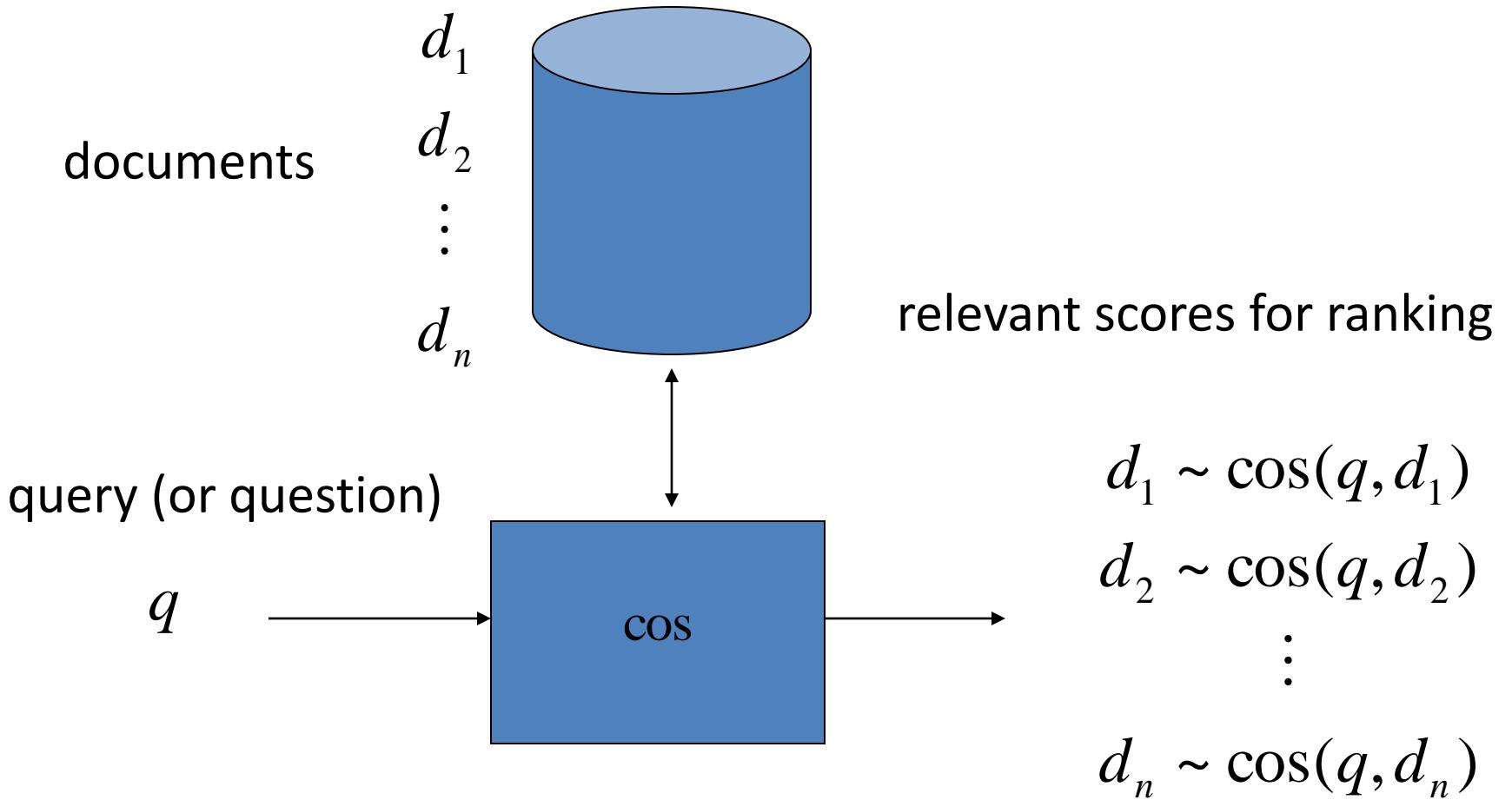
Matching between Query and Document



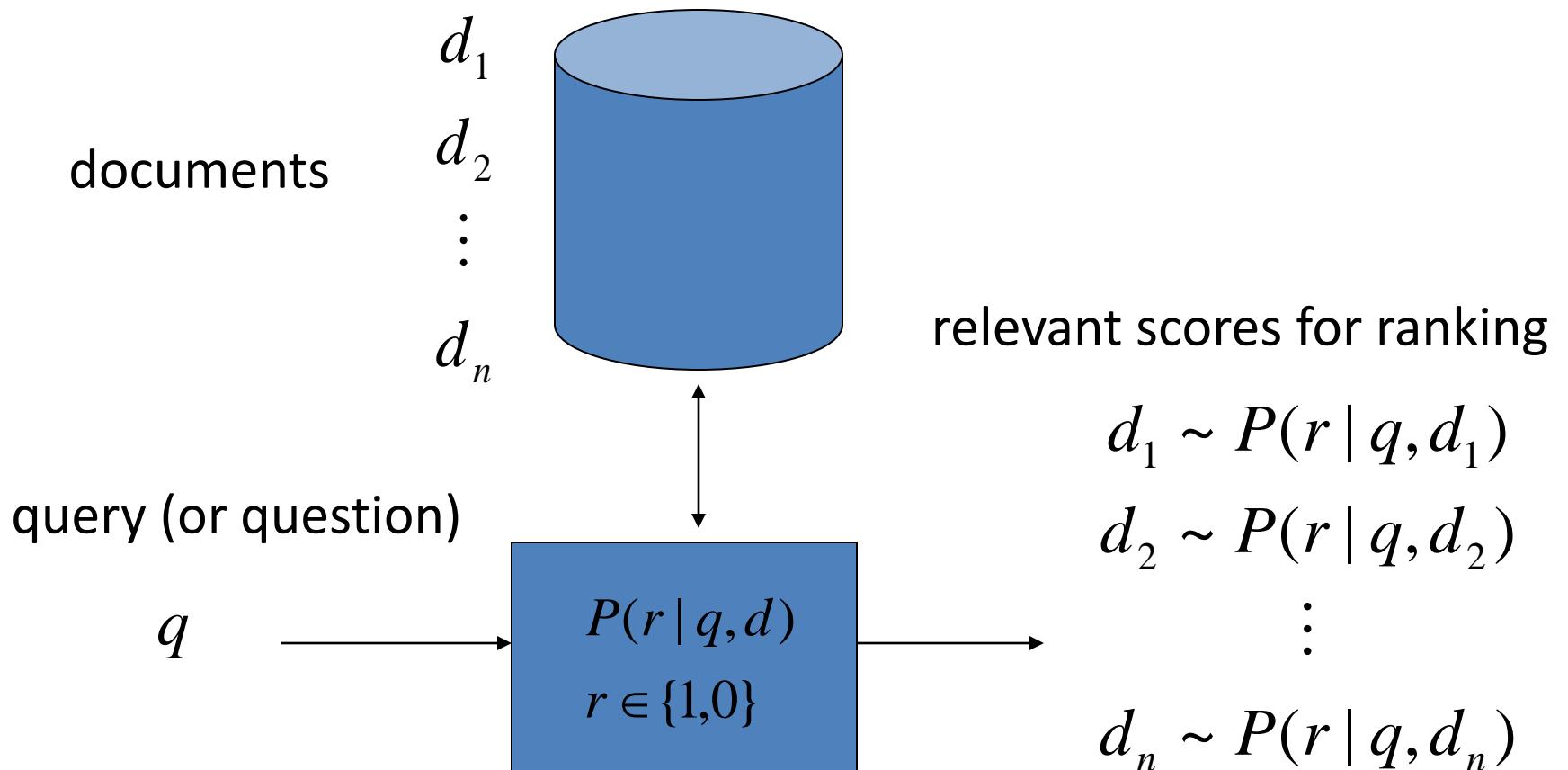
Matching between Two Worlds



Vector Space Model (Salton 1975)

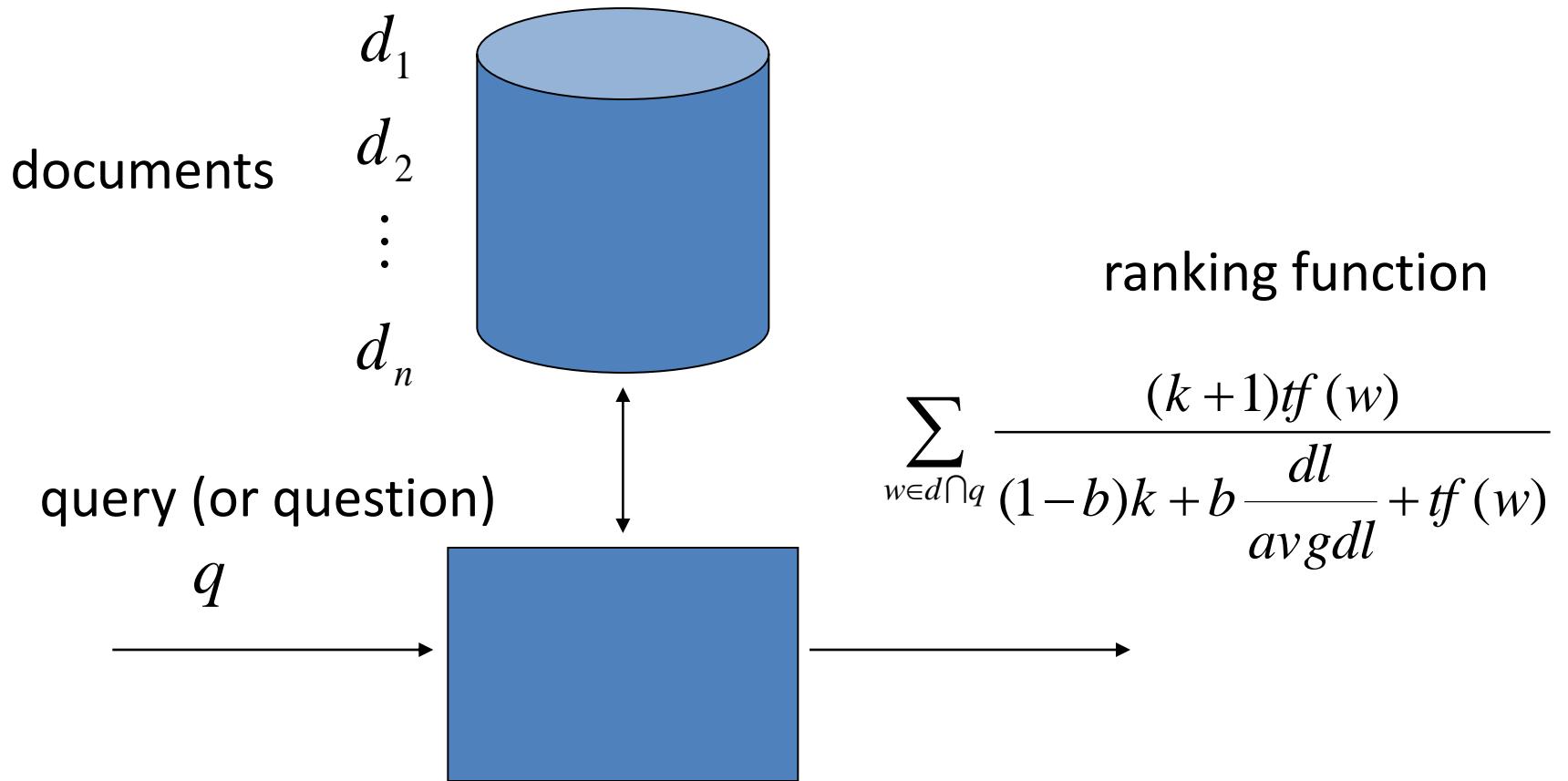


Probabilistic Model



Okapi or BM25

(Robertson and Walker 1994)



Language Mode

(Ponte and Croft 1998)

document = bag of words

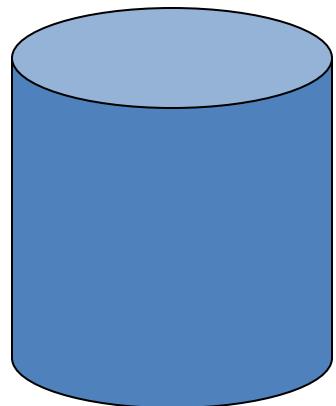
$$d_1 = w_{11} w_{12} \cdots w_{1l_1}$$

$$d_2 = w_{21} w_{22} \cdots w_{2l_2}$$

⋮

$$d_n = w_{n1} w_{n2} \cdots w_{nl_n}$$

$$q = w_{q1} w_{q2} \cdots w_{ql_q}$$



relevance scores for ranking

$$d_1 \sim P(q | d_1)$$

$$d_2 \sim P(q | d_2)$$

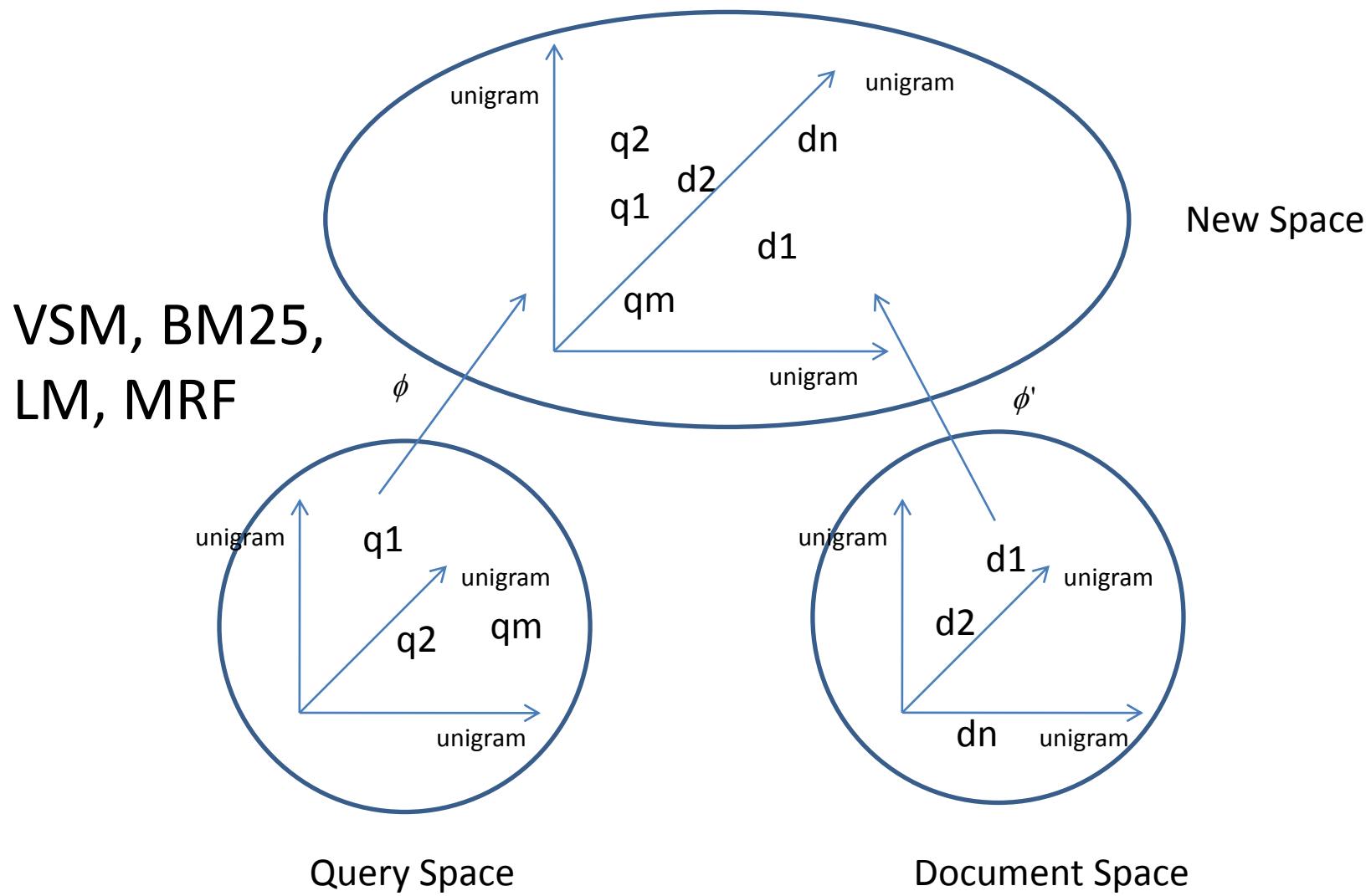
⋮

$$d_n \sim P(q | d_n)$$

Relevance Model as Similarity Function

Jun Xu, Hang Li, Chaoling Zhong
AIRS 2010

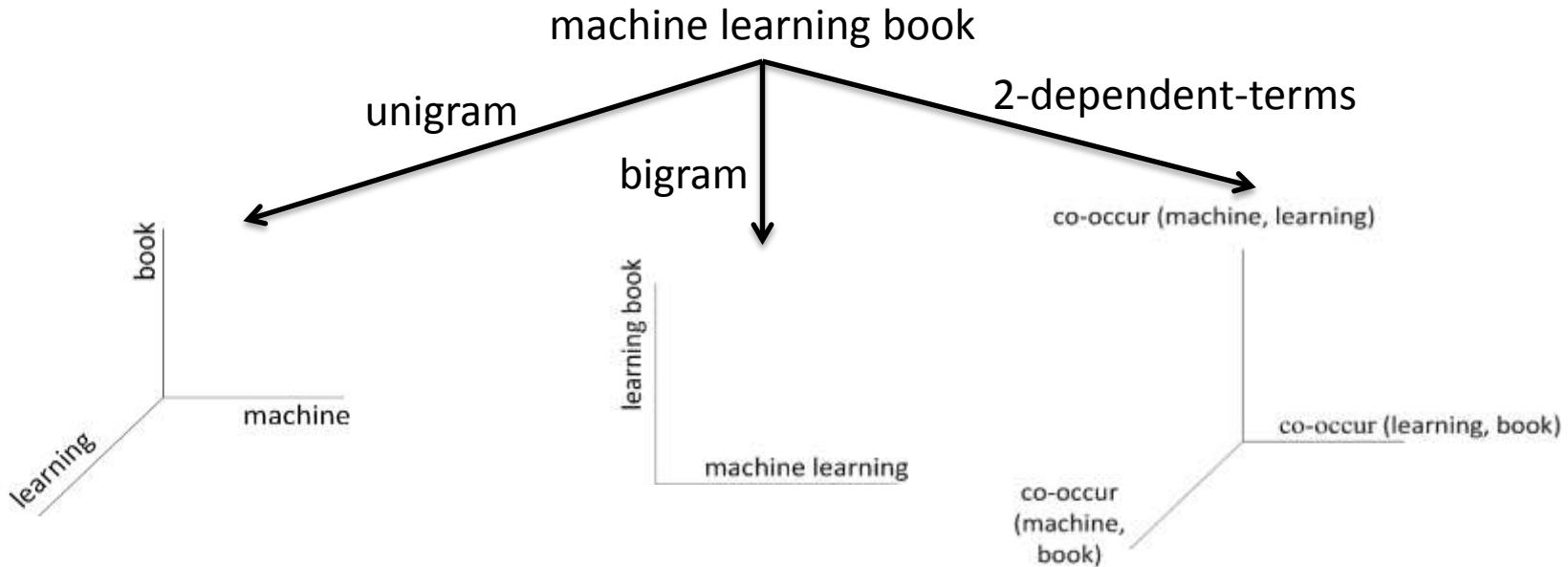
IR Models as Similarity Functions (Similarity Functions)



IR Models Are Similarity Functions

- VSM
 - $\text{BM25}(q, d) = \langle \phi_Q^{VSM}(q), \phi_D^{VSM}(d) \rangle$, for all $w \in V$
 $\phi_Q^{VSM}(q)_w = tfidf(w, q)$ and $\phi_D^{VSM}(d)_w = tfidf(w, d)$
- BM25
 - $\text{BM25}(q, d) = \langle \phi_Q^{BM25}(q), \phi_D^{BM25}(d) \rangle$, for all $w \in V$
 $\phi_Q^{BM25}(q)_w = \frac{(k_3+1) \times tf(w, q)}{k_3 + tf(w, q)}$
 $\phi_D^{BM25}(d)_w = \text{IDF}(w) \cdot \frac{(k_1+1) \times tf(w, d)}{k_1 \left(1 - b + b \cdot \frac{\text{len}(d)}{\text{avgDocLen}} \right) + tf(w, d)}$
- LMIR
 - $\text{LMIR}(q, d) = \langle \phi_Q^{LMIR}(q), \phi_D^{LMIR}(d) \rangle + \text{len}(q) \cdot \log \frac{\mu}{\text{len}(d) + \mu}$, for all $w \in V$
 $\phi_Q^{LMIR}(q)_w = tf(w, q)$
 $\phi_D^{LMIR}(d)_w = \log \left(1 + \frac{tf(w, d)}{\mu \cdot P(w)} \right)$, where $P(w)$ plays similar role as IDF in BM25

Relevance beyond Unigram



Extension of IR models

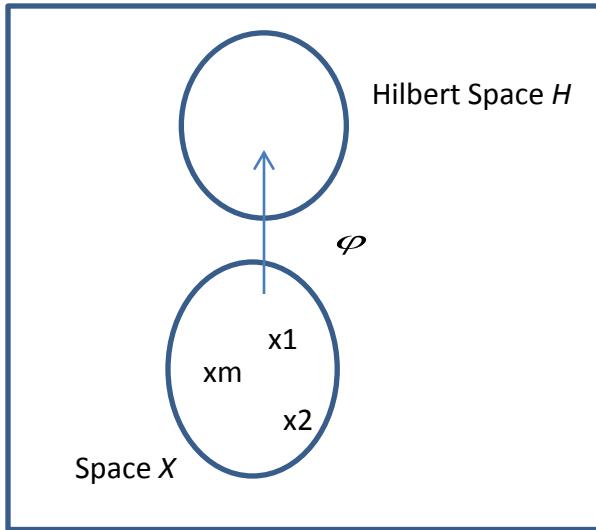
- BM25
 - $\text{BM25}(q, d) = \langle \phi_Q^{BM25}(q), \phi_D^{BM25}(d) \rangle$, and for all $w \in V$
$$\phi_Q^{BM25}(q)_w = \frac{(k_3+1) \times tf(w, q)}{k_3 + tf(w, q)}$$
$$\phi_D^{BM25}(d)_w = \text{IDF}(t) \cdot \frac{(k_1+1) \times tf(w, d)}{k_1 \left(1 - b + b \cdot \frac{\text{len}(d)}{\text{avgDocLen}}\right) + tf(w, d)}$$
- BM25_Kernel
 - $\text{BM25_Kernel}(q, d) = \sum_t \text{BM25_Kernel}_t(q, d)$ where t is dependence type
 - $\text{BM25_Kernel}_t(q, d) = \langle \phi_{Q,t}^{BM25}(q), \phi_{D,t}^{BM25}(d) \rangle$, and for all $x \in V_t$
$$\phi_{Q,t}^{BM25}(q)_x = \frac{(k_3+1) \times f_t(x, q)}{k_3 + f_t(x, q)}$$
$$\phi_{D,t}^{BM25}(d)_x = \text{IDF}_t(x) \cdot \frac{(k_1+1) \times f_t(x, d)}{k_1 \left(1 - b + b \cdot \frac{f_t(d)}{\text{avgDocLen}_t}\right) + f_t(x, d)}$$

Similarity Function

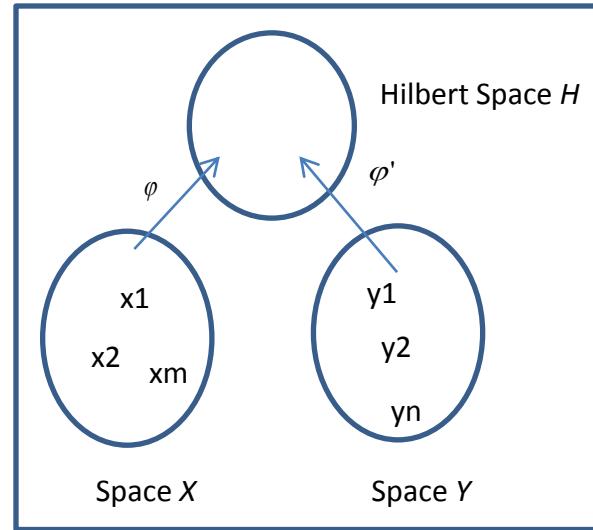
- Kernel $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$
 - Definition: $k(x, x') = \langle \phi(x), \phi(x') \rangle$, where $\phi: \mathcal{X} \rightarrow \mathcal{H}$
 - Given k_1 and k_2 are kernels, create new kernels:
 αk , where $\alpha \geq 0$; $k_1 + k_2$; $k_1 \cdot k_2$
- Similarity function: $k: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$
 - Definition: $k(x, y) = \langle \phi(x), \phi'(y) \rangle$, where $\phi: \mathcal{X} \rightarrow \mathcal{H}$ and $\phi': \mathcal{Y} \rightarrow \mathcal{H}$
 - Given k_1 and k_2 are similarity functions, create new similarity functions:
 αk , where $\alpha \in \mathbb{R}$; $k_1 + k_2$; $k_1 \cdot k_2$

Kernel vs Similarity Function

$$k(x, x') = \langle \phi(x), \phi(x') \rangle$$



$$k(x, y) = \langle \phi(x), \phi'(y) \rangle$$



Query Document Mismatch

Same Search Intent Different Query Representations

Example = “Distance between Sun and Earth”

- "how far" earth sun
- "how far" sun
- "how far" sun earth
- average distance earth sun
- average distance from earth to sun
- average distance from the earth to the sun
- distance between earth & sun
- distance between earth and sun
- distance between earth and the sun
- distance from earth to the sun
- distance from sun to earth
- distance from sun to the earth
- distance from the earth to the sun
- distance from the sun to earth
- distance from the sun to the earth
- distance of earth from sun
- distance between earth sun
- how far away is the sun from earth
- how far away is the sun from the earth
- how far earth from sun
- how far earth is from the sun
- how far from earth is the sun
- how far from earth to sun
- how far from the earth to the sun
- distance between sun and earth

Same Search Intent, Different Query Representations

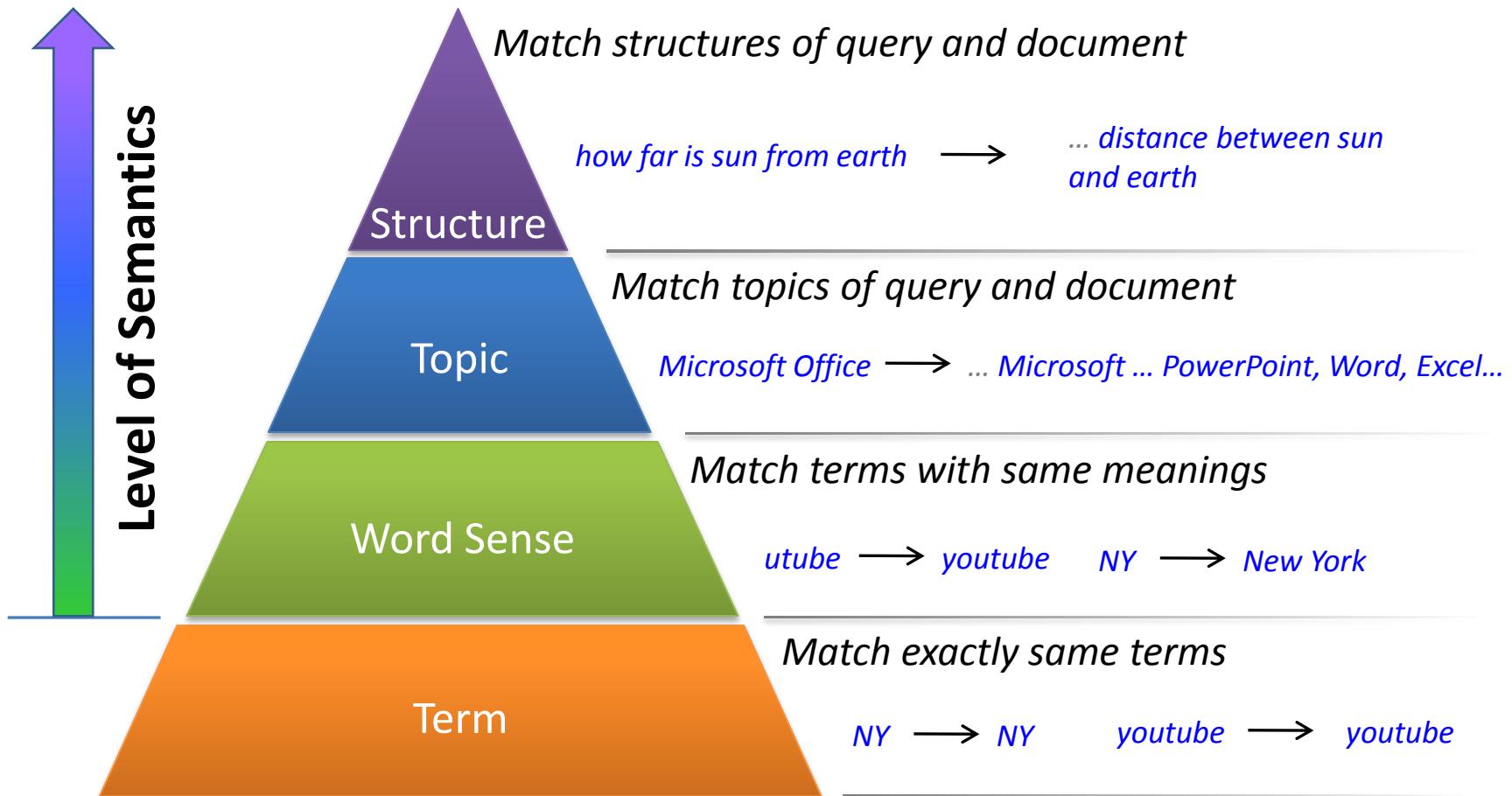
Example = “Youtube”

- | | | |
|-------------------|-----------------------|----------------------|
| • youtube | yuotube | yuo tube |
| • ytube | youtubr | yu tube |
| • youtubo | youtuber | youtubecom |
| • youtube om | youtube music videos | youtube videos |
| • youtube | youtube com | youtube co |
| • youtub com | you tube music videos | yout tube |
| • youtub | you tube com yourtube | your tube |
| • you tube | you tub | you tube video clips |
| • you tube videos | www you tube com | wwwwww youtube com |
| • www youtube | www youtube com | www youtube co |
| • yotube | www you tube | www utube com |
| • ww youtube com | www utube | www u tube |
| • utube videos | utube com | utube |
| • u tube com | utub | u tube videos |
| • u tube | my tube | toutube |
| • outube | our tube | touttube |

Examples of Term Mismatch

- **Query → Document**
- swimming pool schedule = pool schedule
- seattle best hotel = seattle best hotels
- natural logarithm transformation = logarithm transformation
- china kong ≠ china hong kong
- why are windows so expensive ≠ why are macs so expensive

Different Levels of Semantic Matching



Query Understanding (Online)

Keyphrase Identification
in Query

[michael I. jordan: **Keyphrase**]
[berkeley: **Attribute**]: *academic*

[michael jordan: **Keyphrase**]
[berkeley: **Attribute**]: *academic* Structure

Query Topic
Identification

michael I. jordan berkeley: *academic*
michael jordan berkeley: *academic*

Topic

Similar Query Finding

michael I. jordan berkeley
michael jordan berkeley

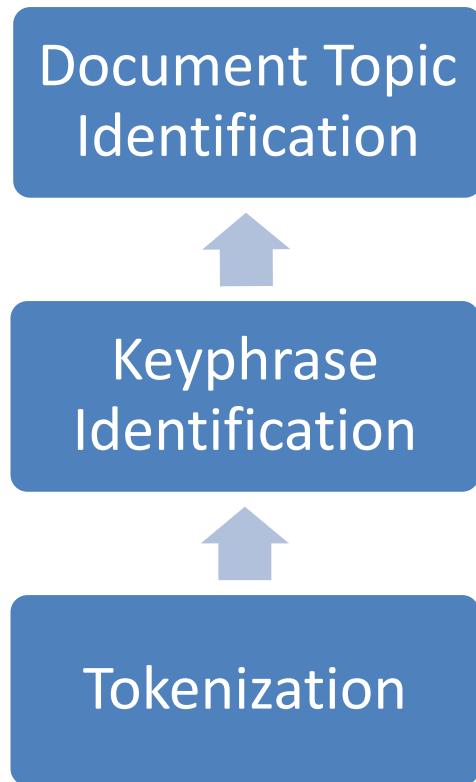
Spelling Error Correction

michael jordan berkeley

Sense

michael jordan berkele

Document Understanding (Offline)



*Michael Jordan is Professor in the
Department of Electrical Engineering*

**[Michael Jordan/M. Jordan] is [Professor]
in the [Department/Dept.] of
[Electrical Engineering/EE]: academic**

Topic

**[Michael Jordan/M. Jordan: Keyphrase] is
[Professor] in the [Department/Dept.] of
[Electrical Engineering/EE]**

Structure

**[Michael Jordan] is [Professor] in the
[Department] of [Electrical Engineering]**

Term

Online Semantic Matching

[Michael I. Jordan's Home Page](#)

Models of visuomotor and other learning (Univ. of California, Berkeley, USA)
www.cs.berkeley.edu/~jordan · Cached page · Mark as spam

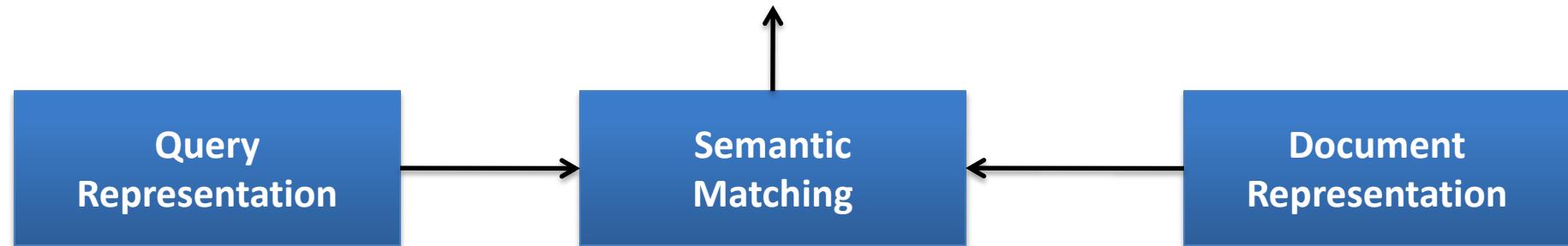
[Michael Jordan | EECS at UC Berkeley](#)

Michael Jordan Professor Research Areas Artificial Intelligence (AI) Biosystems & Computational Biology (BIO) Control, Intelligent Systems, and Robotics (CIR)
www.eecs.berkeley.edu/Faculty/Homepages/jordan.html · Cached page · Mark as spam

[Publications](#)

Jordan, In M.-H. Chen, D. Dey, P. Mueller, D. Sun, and K. Ye (Eds.), Frontiers of Technical Report 661, Department of Statistics, University of California, Berkeley, 2004.
www.cs.berkeley.edu/~jordan/publications.html · Cached page · Mark as spam

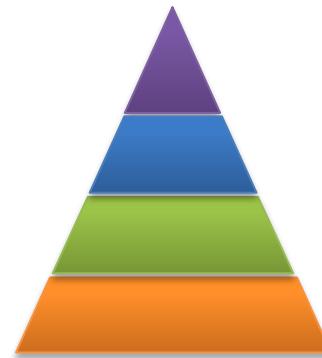
Ranking Results



[michael i. jordan: **Keyphrase**]
[berkeley: **Attribute**]: *academic*

[michael jordan: **Keyphrase**]
[berkeley: **Attribute**]: *academic*

[Michael Jordan/M. Jordan: Keyphrase]
is *[Professor]* in the *[Department/Dept.]* of
[Electrical Engineering/EE]: academic

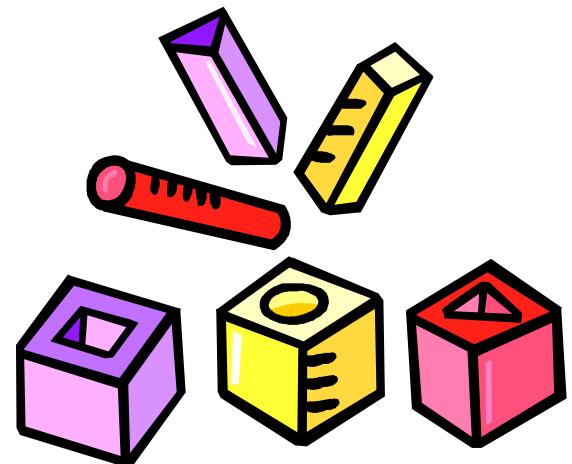


Matching can be conducted at different levels

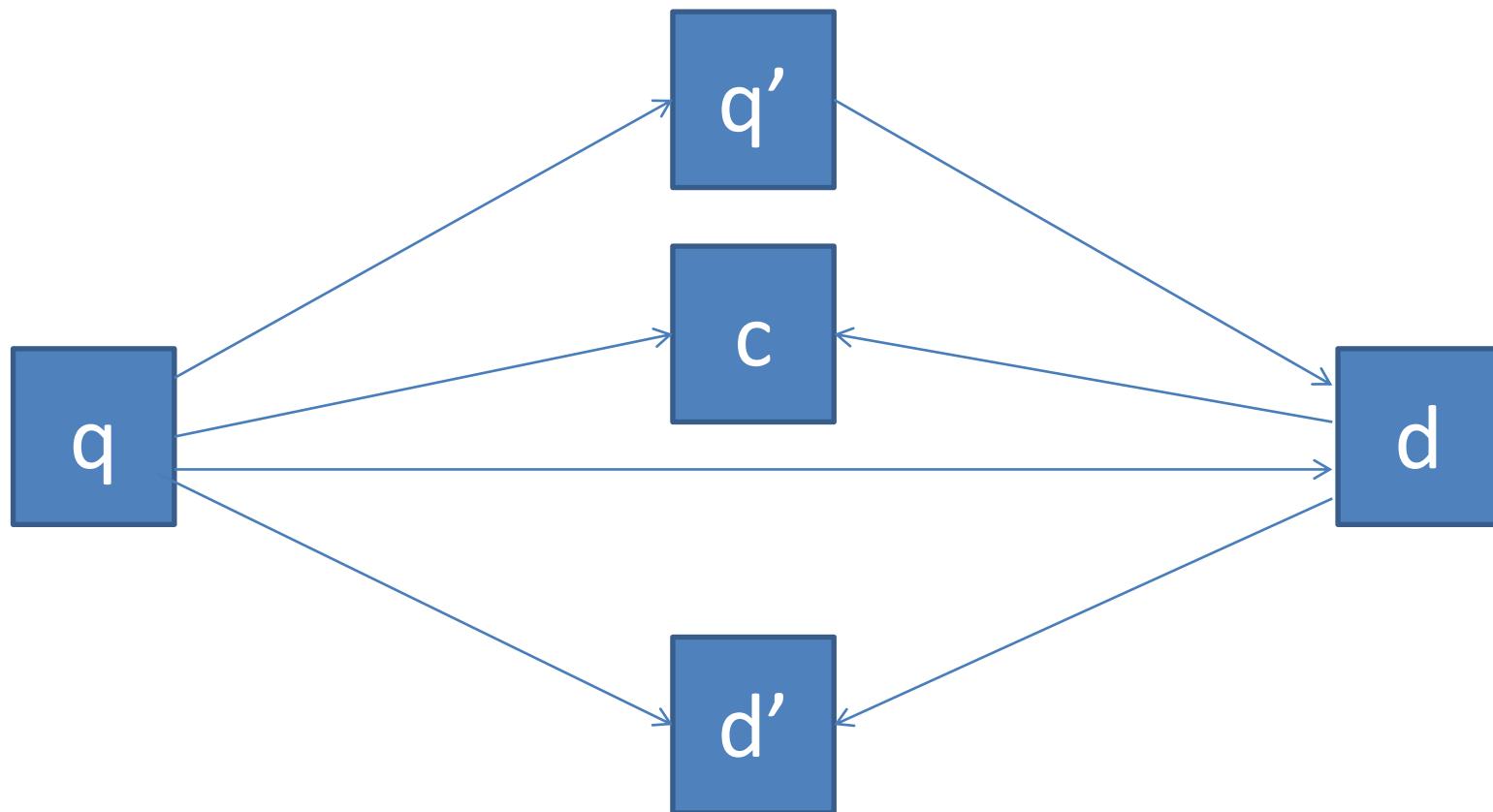
Related Work

- Studied in long history of IR
- Query expansion, pseudo relevance feedback
- Latent Semantic Indexing, Probabilistic Latent Semantic Indexing, Latent Dirichlet Allocation
-

Learning to Match



Four Ways to Match



Learning to Match

- Learning matching function

$$f_M(q, d)$$

- Using training data $(q_1, d_1), \dots, (q_N, d_N)$
- *Using prior knowledge or other data M*

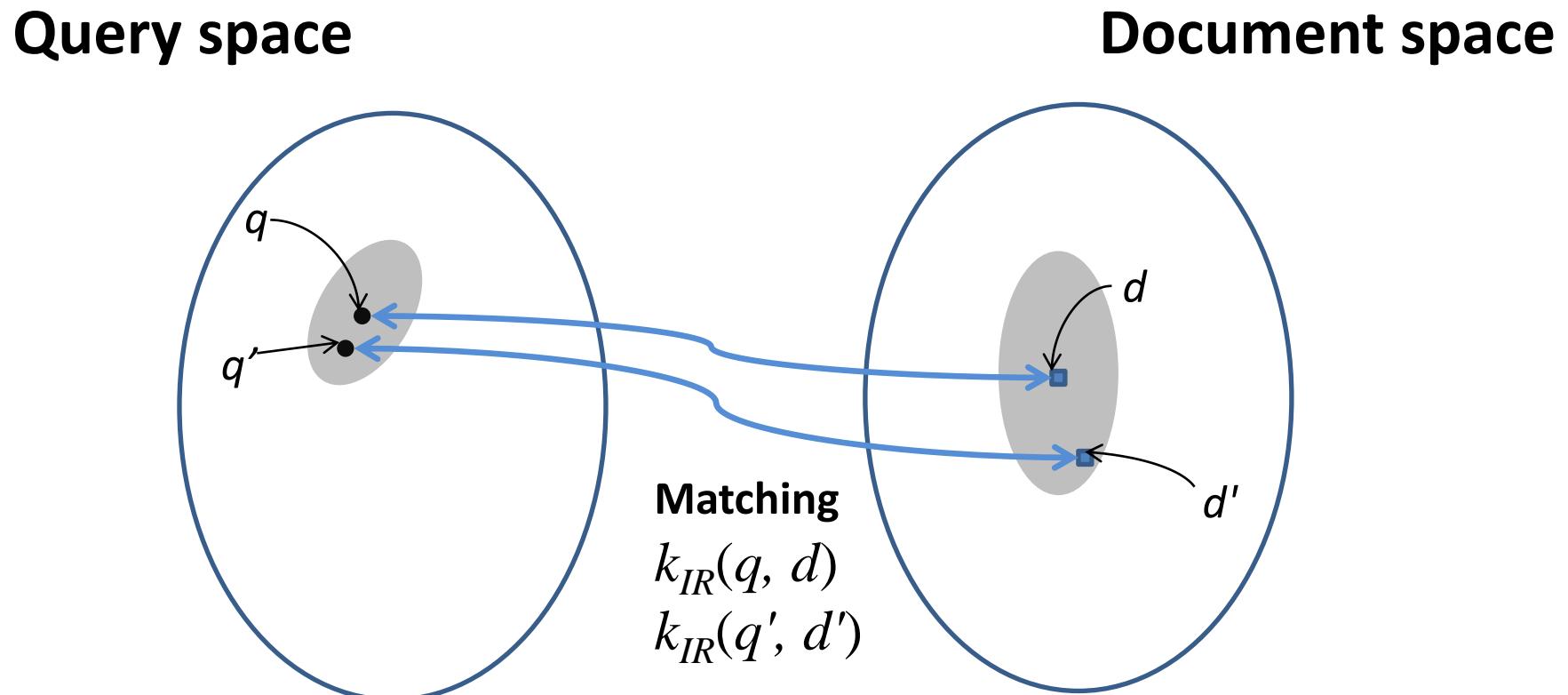
Challenges in Matching

- How to incorporate prior knowledge or other data into model
- Scale is very large

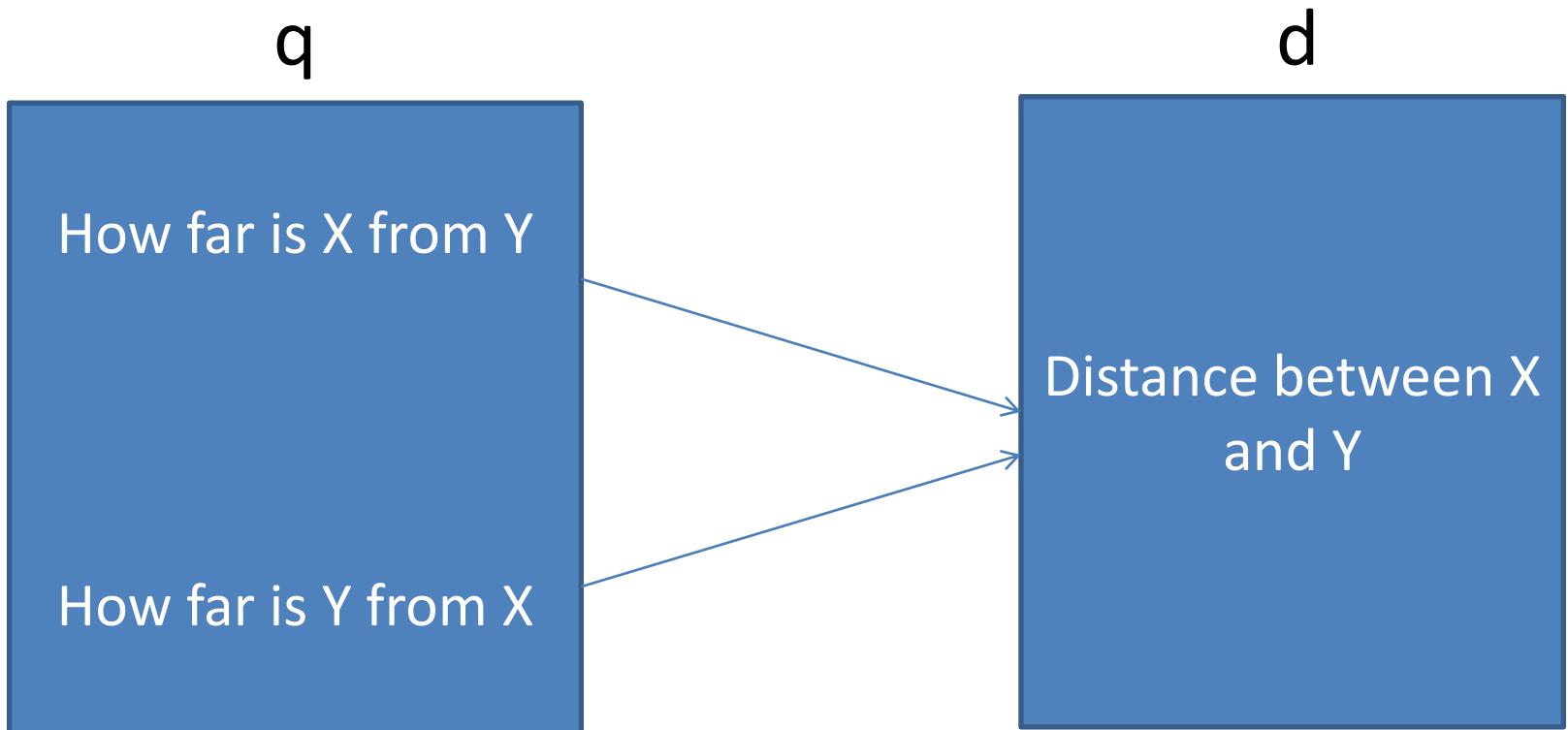
Matching Problem: Matrix Data View

	d1	d2	d3		dn
q1			1		
q1					1
q1				4	
qm		1			5

Matching Problem: Space View



Matching Problem: String Data View



Examples of Matching Models

- **Similarity Learning**

$$f_M(q, d) = f(q, d) + \sum_i k_Q(q, q_i) k_D(d, d_i) f(q_i, d_i)$$

- **Topic Modeling**

$$f_M(q, d) = \sum_k u(q, k) v(k, d)$$

- **String Transformation**

$$f_M(q, d) = f(q, d) + \sum_i k_T(q, q_i) k_T(d, d_i) f(q_i, d_i)$$

Matching vs Ranking

	Matching	Ranking
Prediction	Matching score between query and document	List of documents
Model	$f(q, d)$	$f(q,d_1), f(q,d_2), \dots f(q,d_n)$
Loss Function	Single query document pair	List of documents with respect to query
Challenge	Mismatch	Correct ranking on top

Matching between Heterogeneous Data is Everywhere

- Matching between user and product (collaborative filtering)
- Matching between text and image (image annotation)
- Matching between people (dating)
- Matching between languages (machine translation)

Our Methods of Learning to Match

Three Approaches of Learning to Match

- Similarity Learning → Word sense level
- Topic Modeling → Topic level
- String Transformation → Structure level

Summary of Our Technologies

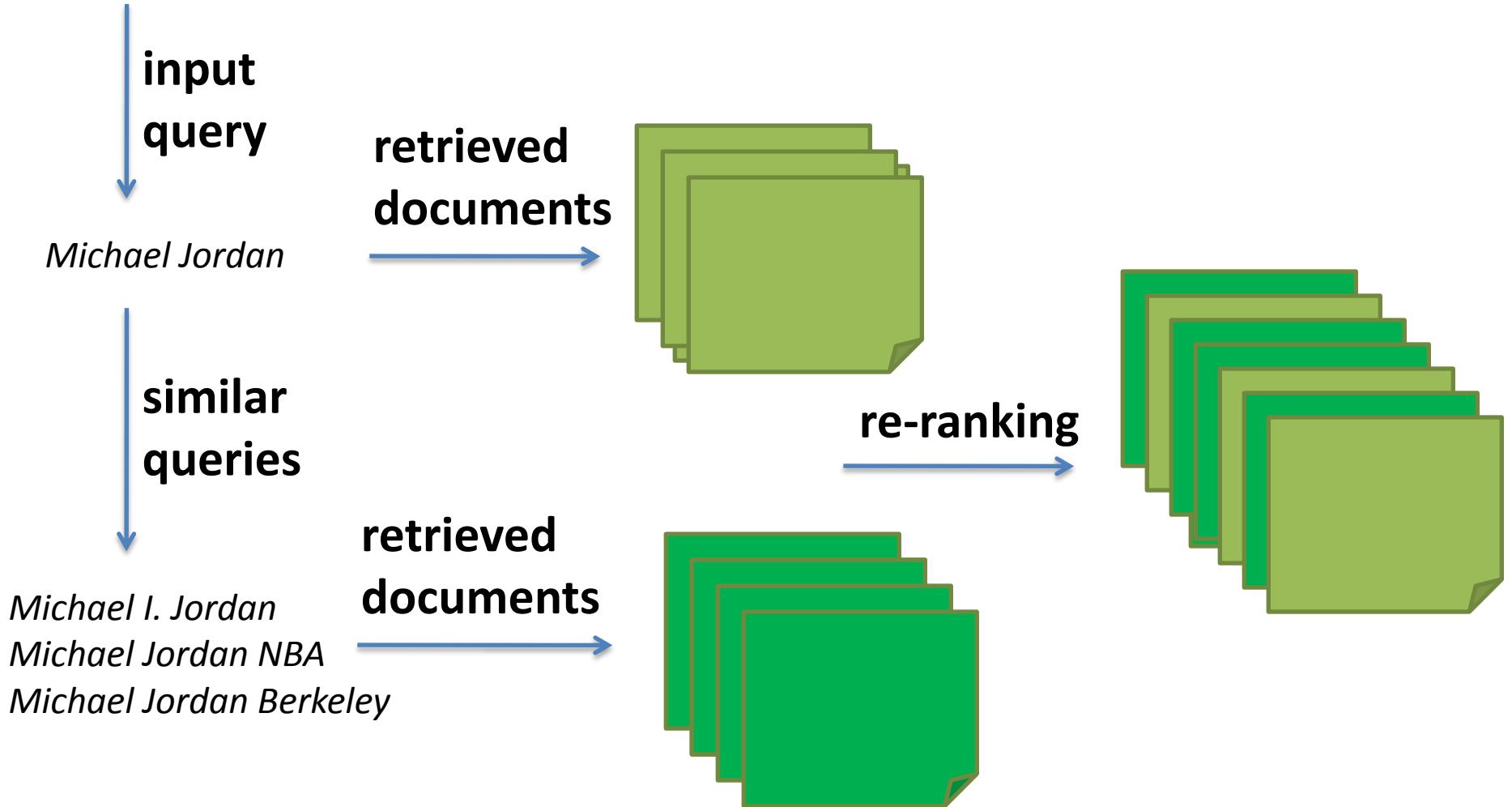
	Technologies	Current Status
Term matching	Key n-gram learning	ongoing
Term matching	Relevance model as similarity function	AIRS'10 best paper
Sense matching	Robust relevance model	JMLR, WWW'11 poster
Sense matching	Query similarity learning	WSDM'11
Sense matching	CRF model for candidate selection	SIGIR'08
Sense matching	Log linear model for candidate generation	ACL'11
Sense matching	Projection to latent structure	ongoing
Topic matching	Scalable and efficient topic modeling	SIGIR'11

Robust Similarity Function Learning Using Kernel Methods

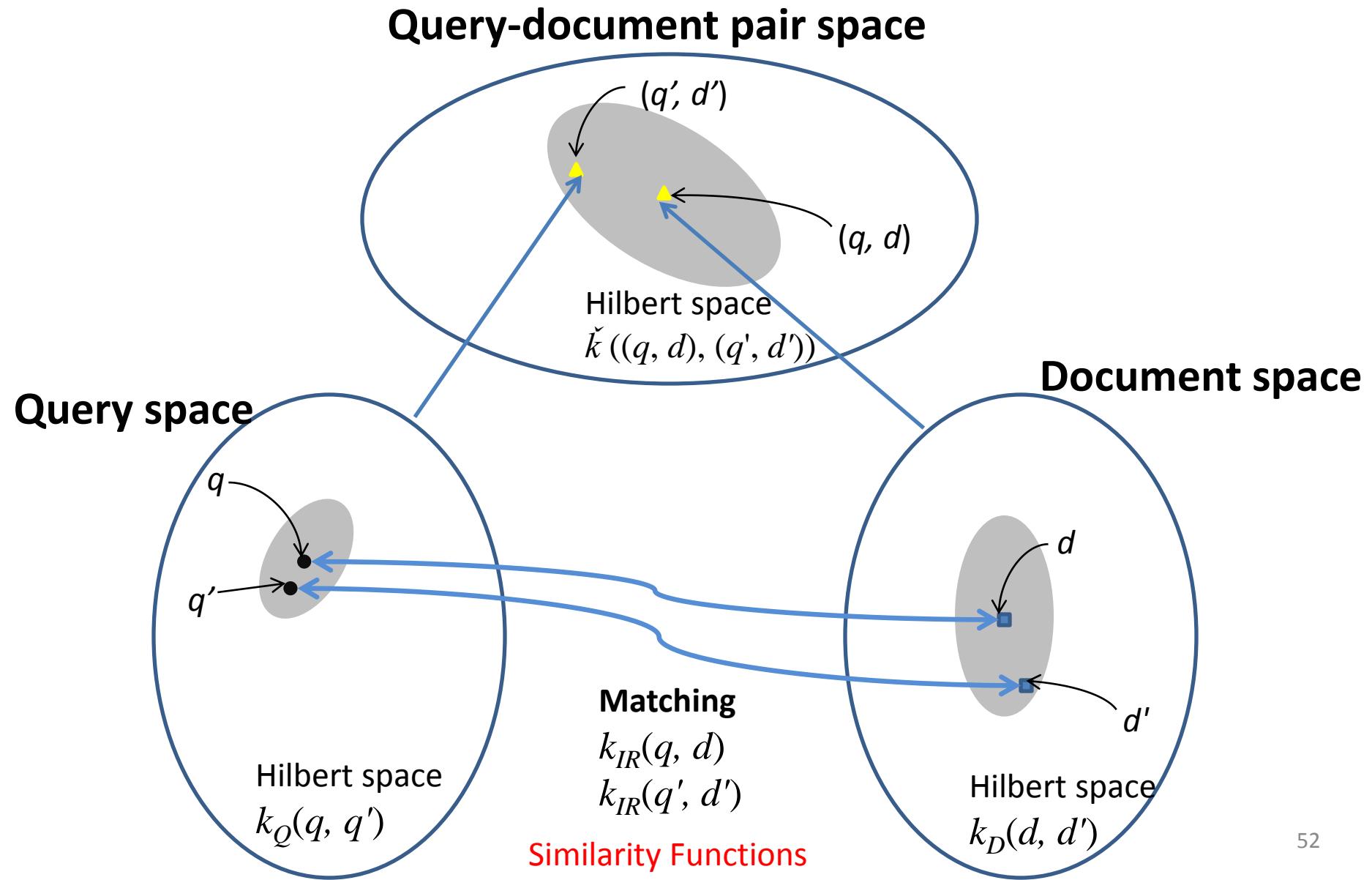
Wei Wu, Hang Li, Jun Xu, Satoshi
Oyama, JMLR 2011

Dealing with Mismatch with Re-Ranking

- Our Approach = Online Learning of Kernel Methods



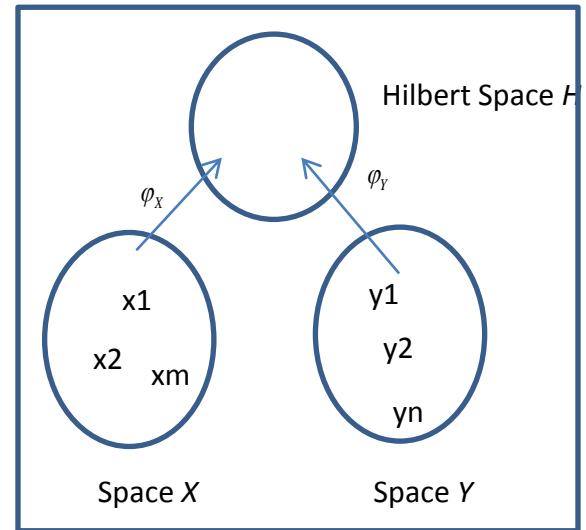
Mapping to Space of Query Document Pairs - Using Kernel Methods



Similarity Learning

- Similarity Function : $k(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}$
- Input
 - Training data $S = \{(x_i, y_i), t_i\}_{1 \leq i \leq N}$
- Output
 - Similarity Function
- Optimization

$$\min_{k \in \mathcal{K} \subseteq \mathcal{A}} \frac{1}{N} \sum_{i=1}^N l(k(x_i, y_i), t_i) + \Omega(k)$$



Similarity Learning Using Kernel Methods

- Assumption
 - Space of similarity functions is RKHS generated by positive-definite kernel $\bar{k}: (\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y})$
 -
- Optimization

$$\min_{k \in \mathcal{K}} \frac{1}{N} \sum_{i=1}^N l(k(x_i, y_i), t_i) + \frac{\lambda}{2} \|k\|_{\mathcal{K}}^2$$

- Solution
 - By representer theorem $k^*(x, y) = \sum_{i=1}^N \alpha_i \bar{k}((x_i, y_i), (x, y))$
 -
- $\bar{k}((x, y), (x', y')) = g(x, y)k_x(x, x')k_y(y, y')g(x', y')$

Learning Robust BM25

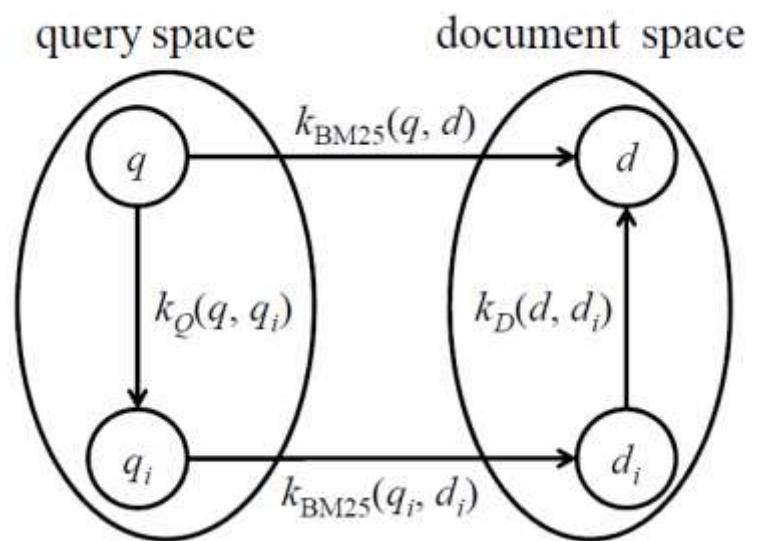
- BM25 :
- Kernel

$$\bar{k}((q, d), (q', d')) = k_{BM25}(q, d)k_Q(q, q')k_D(d, d')k_{BM25}(q', d')$$

- Solution (called Robust BM25)

$$k_{RBM25}(q, d) = k_{BM25}(q, d) \cdot \sum_{i=1}^N \alpha_i k_Q(q, q_i) k_D(d, d_i) k_{BM25}(q_i, d_i)$$

- Deal with term mismatch

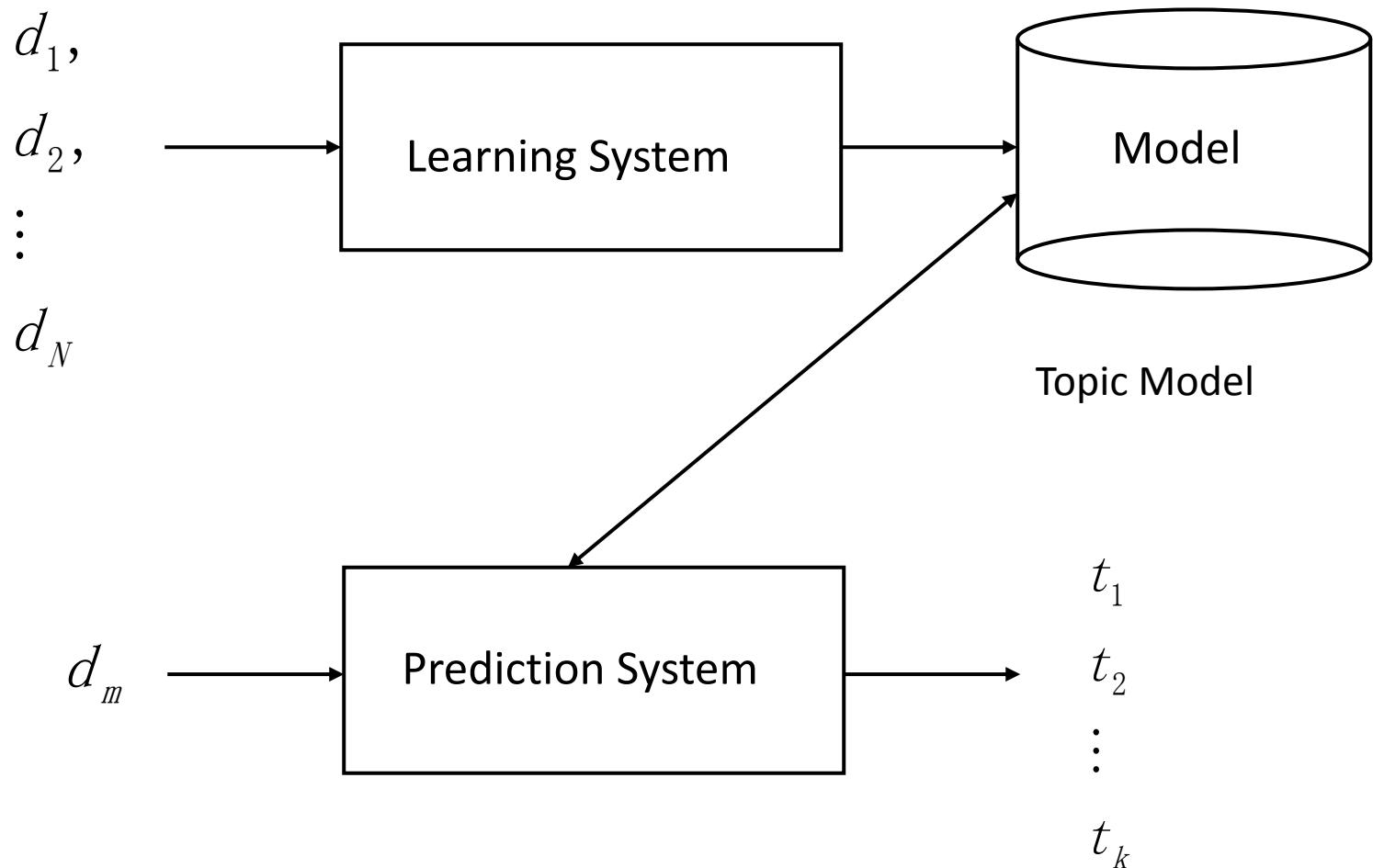


Regularized Latent Semantic Indexing

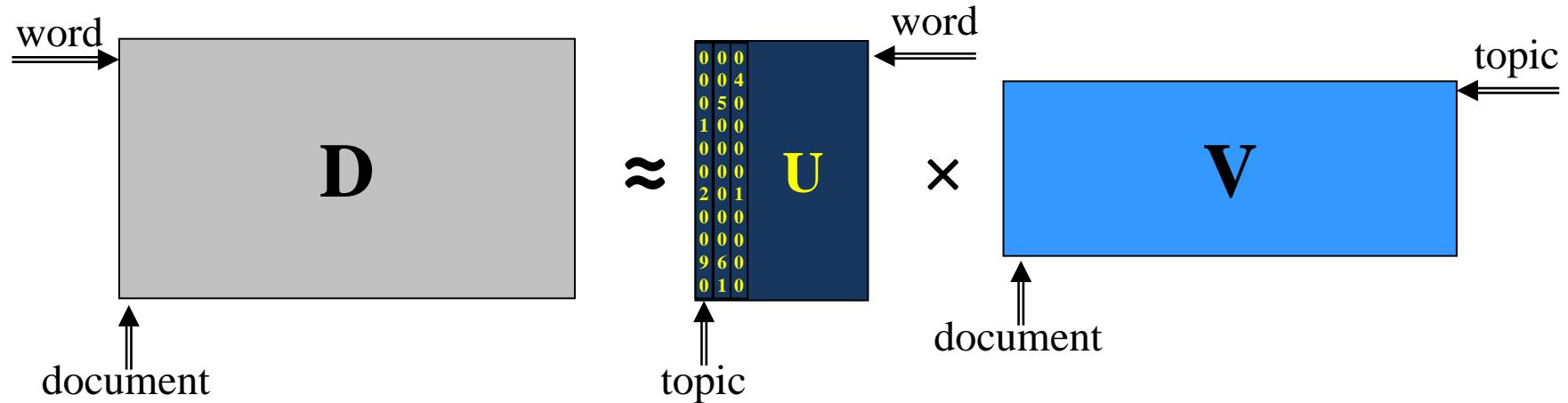
Quan Wang, Jun Xu, Hang Li,
Nick Craswell, SIGIR 2011

Topic Modeling

Our Approach = Regularized Latent Semantic Indexing



Regularized Latent Semantic Indexing



$$\min_{\mathbf{U}, \{\mathbf{v}_n\}} \sum_{n=1}^N \|\mathbf{d}_n - \mathbf{U}\mathbf{v}_n\|_2^2 + \lambda_1 \sum_{k=1}^K \|\mathbf{u}_k\|_1 + \lambda_2 \sum_{n=1}^N \|\mathbf{v}_n\|_2^2$$

topics are sparse

word representation of doc n

topic matrix

topic representation of doc n

Regularized Latent Semantic Indexing

- L1 on topics and L2 on documents
- L1 leads to sparse topics, a topic only contains a small number of words
- L2 leads to accurate modeling
- Formulation is simple
- Easy to scale up

Scalability Comparison

algorithm	max dataset applied (#docs; #words)	# topics	# processors used
PLDA and PLDA+ (by Google)	Wiki-200T (2,112,618; 200,000)	1000	2,048
AD-LDA (by UCI)	NY Times (300,000; 102,660)	200	16
RLSI	B01 (1,562,807; 7,014,881) Wikipedia (3,239,884; 1,689,193) Bing News (1,028,070; 940,702)	500 ~ 1000	16 single machine!

Regularized Topics

AP dataset, topic compactness: 0.0075

OPEC	Africa	contra	school	Noriega	firefight	plane	Saturday	Iran	senate
oil	South	Sandinista	student	Panama	ACR	crash	coastal	Iranian	Reagan
cent	African	rebel	teacher	Panamanian	forest	flight	estimate	Iraq	billion
barrel	Angola	Nicaragua	education	Delval	park	air	western	hostage	budget
price	apartheid	Nicaraguan	college	canal	blaze	airline	Minsch	Iraqi	trade
drug	soviet	aid	court	Jackson	percent	student	nuclear	Bush	Israel
cocaine	Afghanistan	virus	senate	Dukaki	billion	Korea	soviet	Dukaki	Palestinian
traffick	Afghan	infect	Reagan	democrat	rate	protest	treaty	campaign	Israeli
test	Gorbachev	test	house	delegate	0	Korean	missile	Quayle	Arab
enforce	Pakistan	patient	state	percent	trade	Chun	weapon	Bentsen	PLO

Optimization

Algorithm 2 Update \mathbf{U}

Require: $\mathbf{D} \in \mathbb{R}^{M \times N}$, $\mathbf{V} \in \mathbb{R}^{K \times N}$

```

1:  $\mathbf{S} \leftarrow \mathbf{V}\mathbf{V}^T$ 
2:  $\mathbf{R} \leftarrow \mathbf{D}\mathbf{V}^T$ 
3: for  $m = 1 : M$  do
4:    $\bar{\mathbf{u}}_m \leftarrow \mathbf{0}$ 
5:   repeat
6:     for  $k = 1 : K$  do
7:        $w_{mk} \leftarrow r_{mk} - \sum_{l \neq k} s_{kl} u_{ml}$ 
8:        $u_{mk} \leftarrow \frac{(|w_{mk}| - \frac{1}{2} \lambda N)_+ \text{sign}(w_{mk})}{s_{kk}}$ 
9:     end for
10:    until convergence
11:   end for
12:   return  $\mathbf{U}$ 

```

words
processed
in parallel

docs
processed
in parallel

Algorithm 3 Update \mathbf{V}

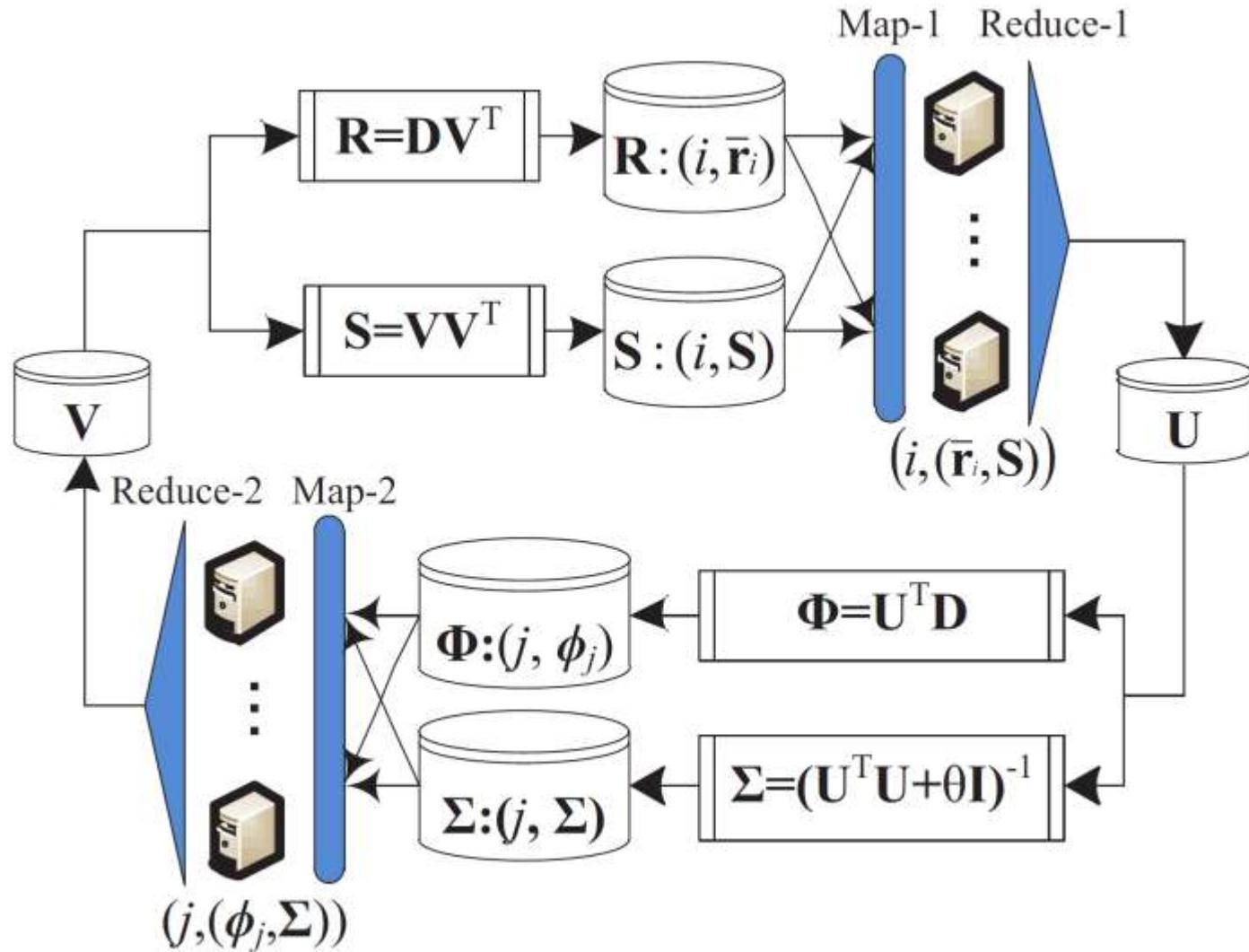
Require: $\mathbf{D} \in \mathbb{R}^{M \times N}$, $\mathbf{U} \in \mathbb{R}^{M \times K}$

```

1:  $\Sigma \leftarrow (\mathbf{U}^T \mathbf{U} + \theta \mathbf{I})^{-1}$ 
2:  $\Phi \leftarrow \mathbf{U}^T \mathbf{D}$ 
3: for  $n = 1 : N$  do
4:    $v_n \leftarrow \Sigma \phi_n$ , where  $\phi_n$  is the  $n^{th}$  column
5: end for
6: return  $\mathbf{V}$ 

```

Scaling up on MapReduce

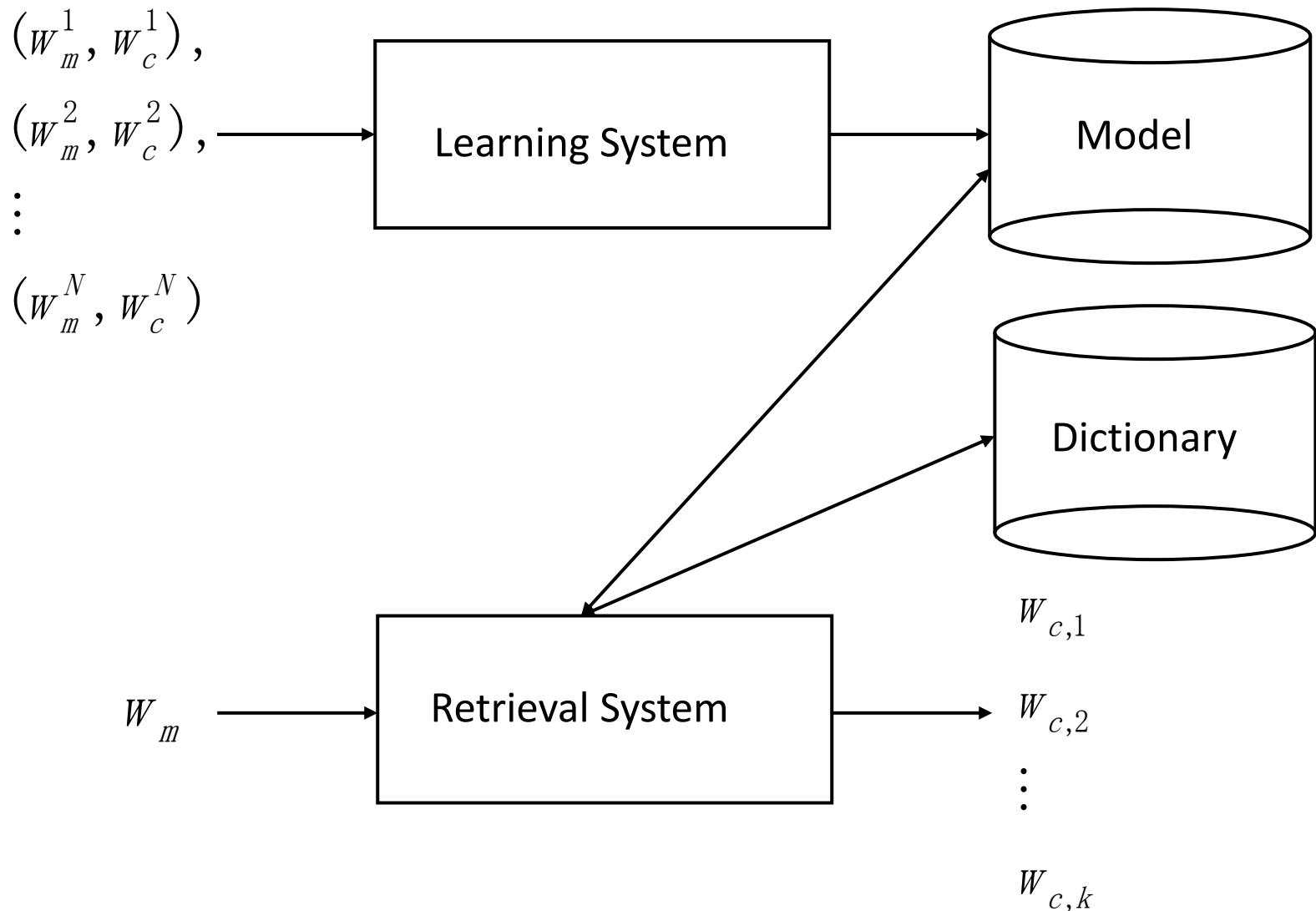


Query Generation Using Log Linear Model

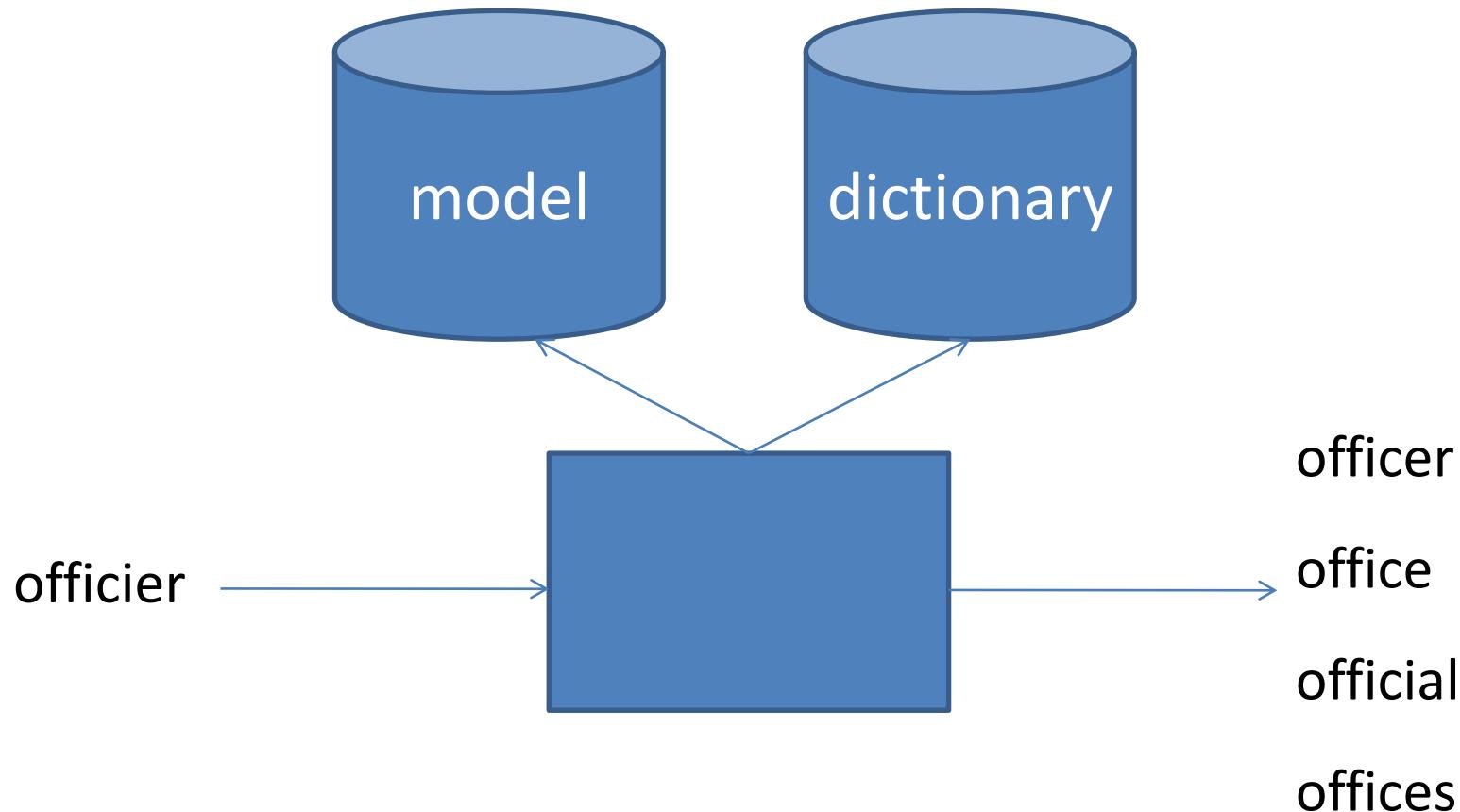
Ziqi Wang, Gu Xu, Hang Li, Ming Zhang
ACL 2011

Candidate Generation in Spelling Correction

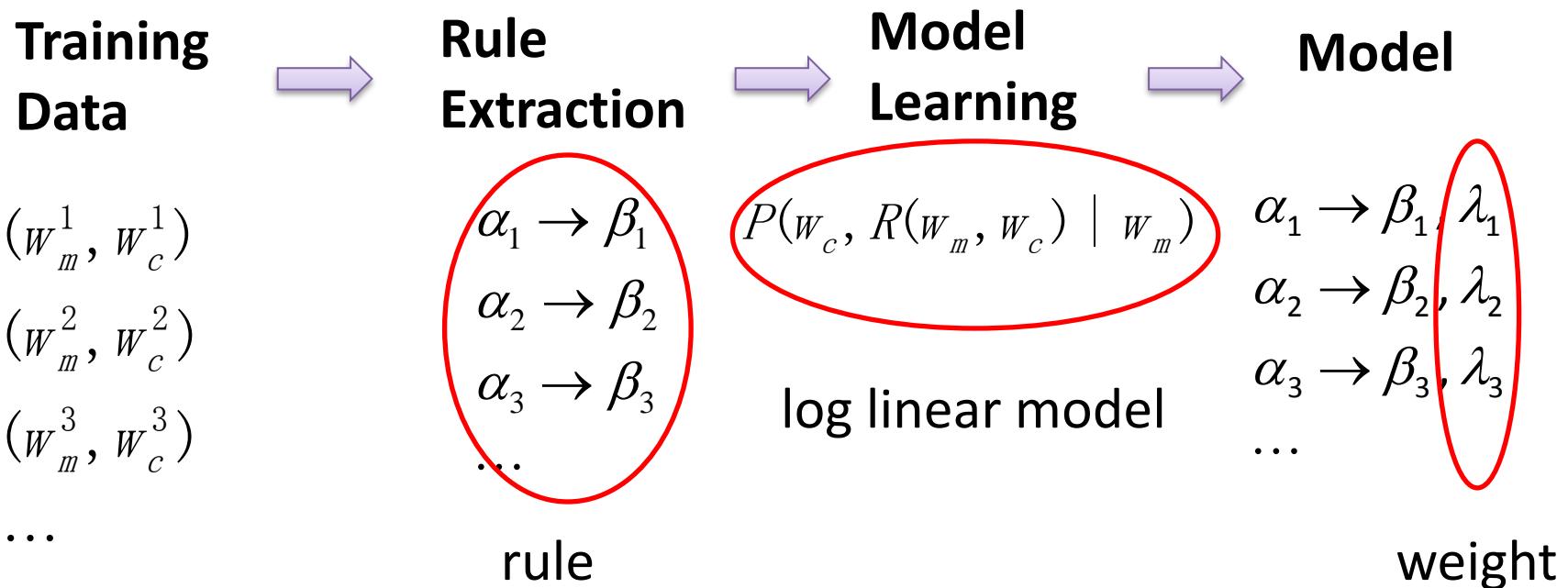
Our Approach = Log Linear Model



Candidate Generation in Spelling Error Correction



Learning



Rule Extraction

- Edit-distance based alignment:

Misspelled: ^ n i c o s o o f t \$
 ↓ ↓ ↓ ↓ ↘ ↘ ↓ ↓ ↓ ↓
Correct: ^ m i c r o s o f t \$

- Basic substitution rules:

$$n \rightarrow m, \phi \rightarrow r$$

- Contextual substitution rules

$$\wedge n \rightarrow \wedge m, ni \rightarrow mi, \wedge ni \rightarrow \wedge mi, c \rightarrow cr, \dots$$

Log Linear Model

- Model

$$P(w_c, R(w_m, w_c) | w_m) = \frac{\exp\left(\sum_{r \in R(w_m, w_c)} \lambda_r\right)}{\sum_{(w'_c, R(w_m, w'_c)) \in Z(w_m)} \exp\left(\sum_{o \in R(w_m, w'_c)} \lambda_o\right)}$$

Set of rules
rewrite w_m to w_c

All pairs of word w'_c and rule set $R(w_m, w'_c)$

Weight of rule

$$\forall \lambda_r \leq 0$$

Non-positive constraint, to improve efficiency in retrieval,
Natural assumption

- Candidate Generation

$$rank(w_c | w_m) = \max_{R(w_m, w_c)} \left(\sum_{r \in R(w_m, w_c)} \lambda_r \right)$$

Model Learning

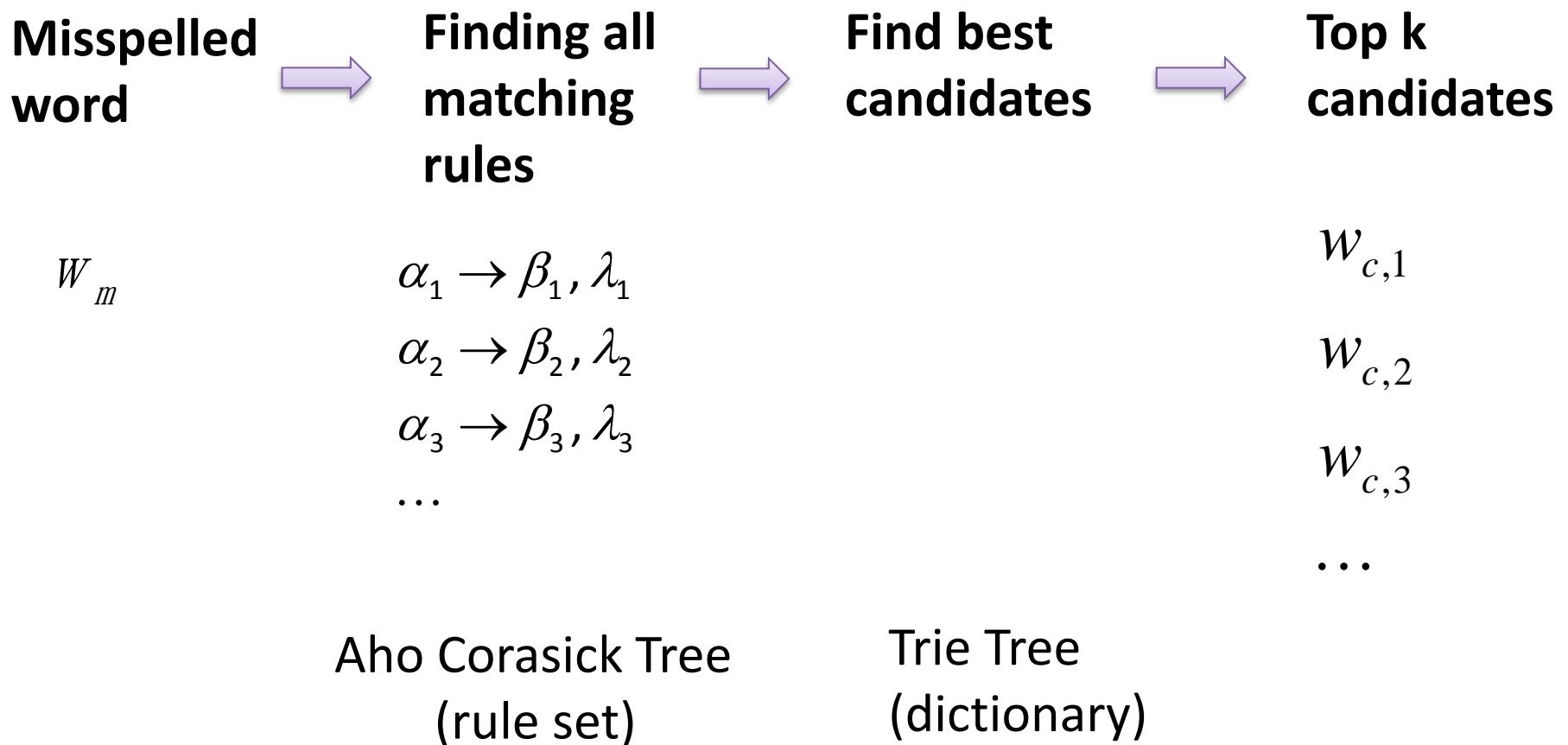
- Objective function

$$\lambda^* = \arg \max_{\lambda} \sum_i \max_{R(w_m^i, w_c^i)} \log P(w_c^i, R(w_m^i, w_c^i) \mid w_m^i)$$

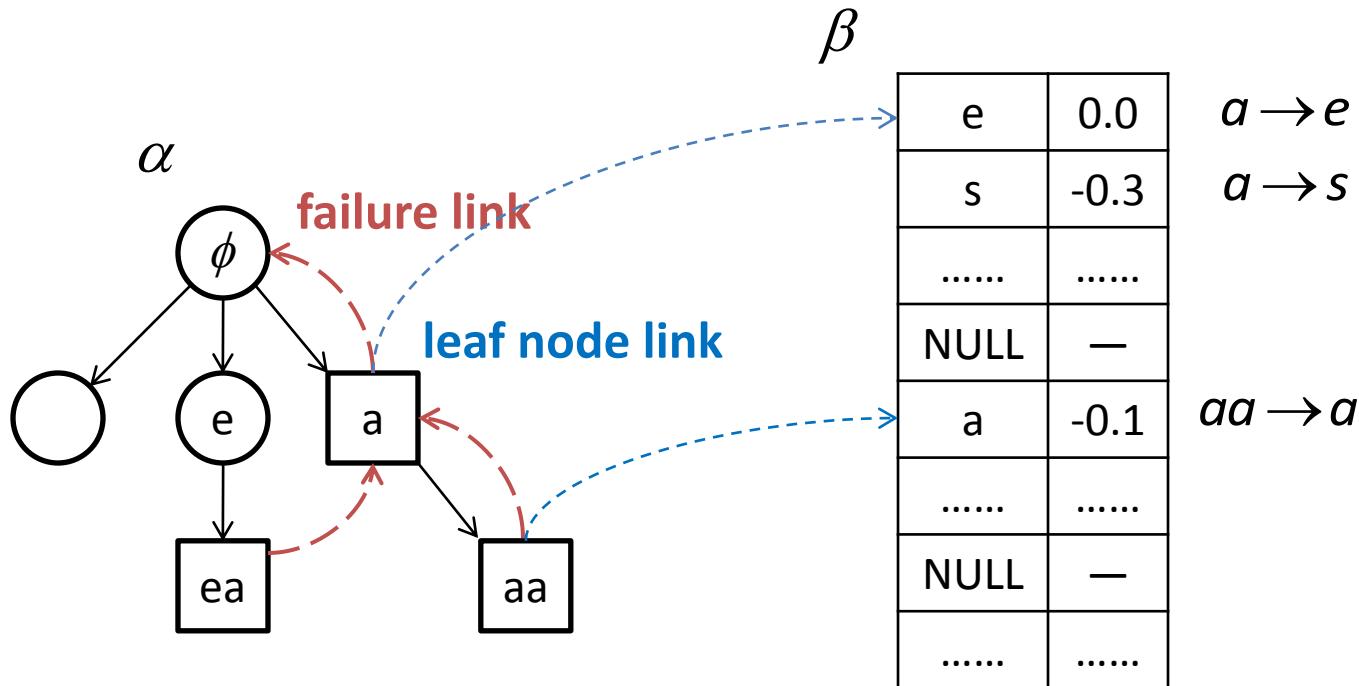
Take max over
transformations

- Algorithm
 - Constrained Quasi Newton Method (BFGS)

Retrieval



Aho Corasick Tree

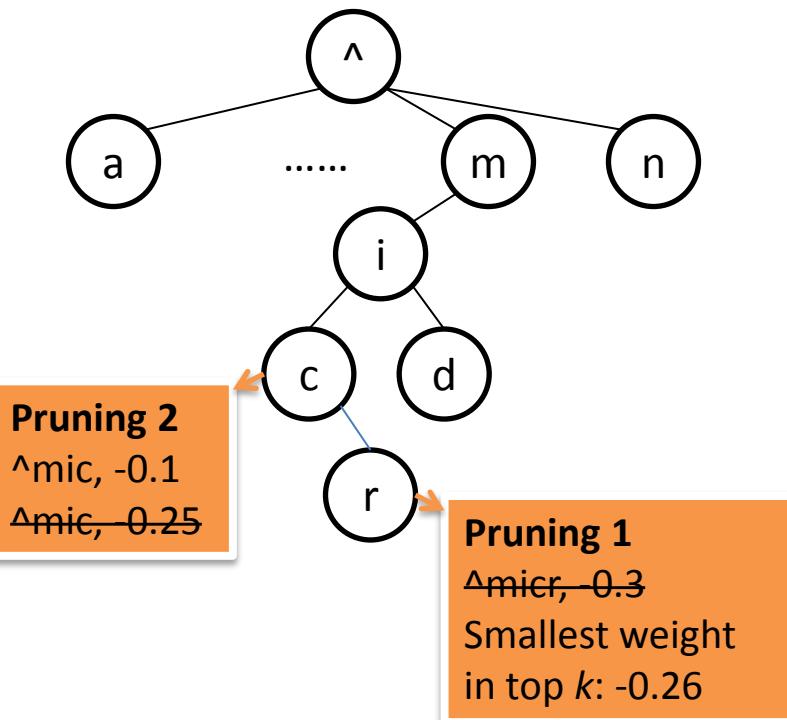


Index all the α 's in the rules on the AC tree

β are stored in an associated list

Retrieval with Dynamic Programming

- Traverse trie tree
 - Match the next position of w_m
 - Apply a rule at the current position of w_m
- Two pruning strategies
 - If the sum of weights is smaller than the smallest weight in the top k list, prune the branch
 - two search branches merge, prune the smaller branch

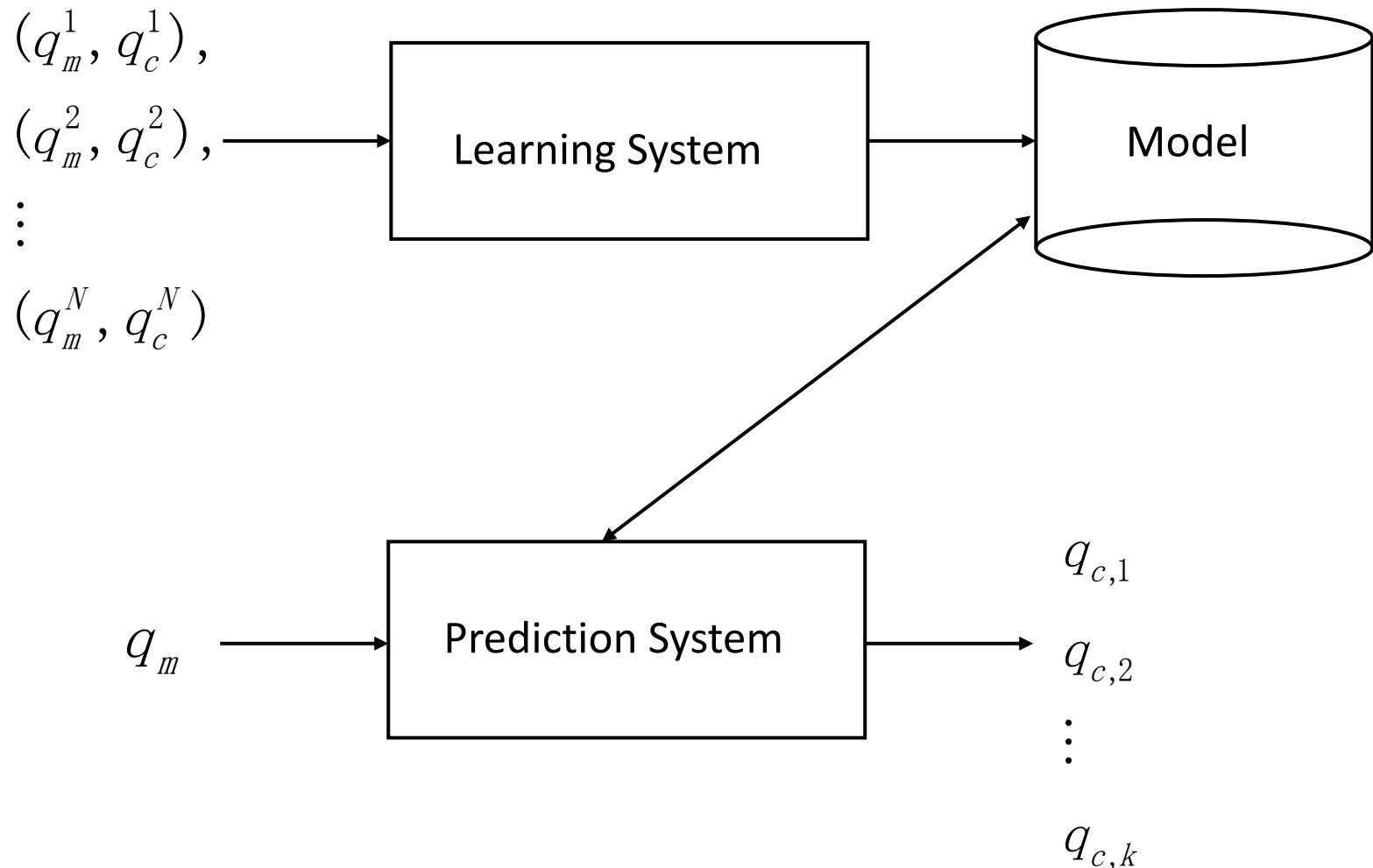


Query Rewriting Using Conditional Random Fields

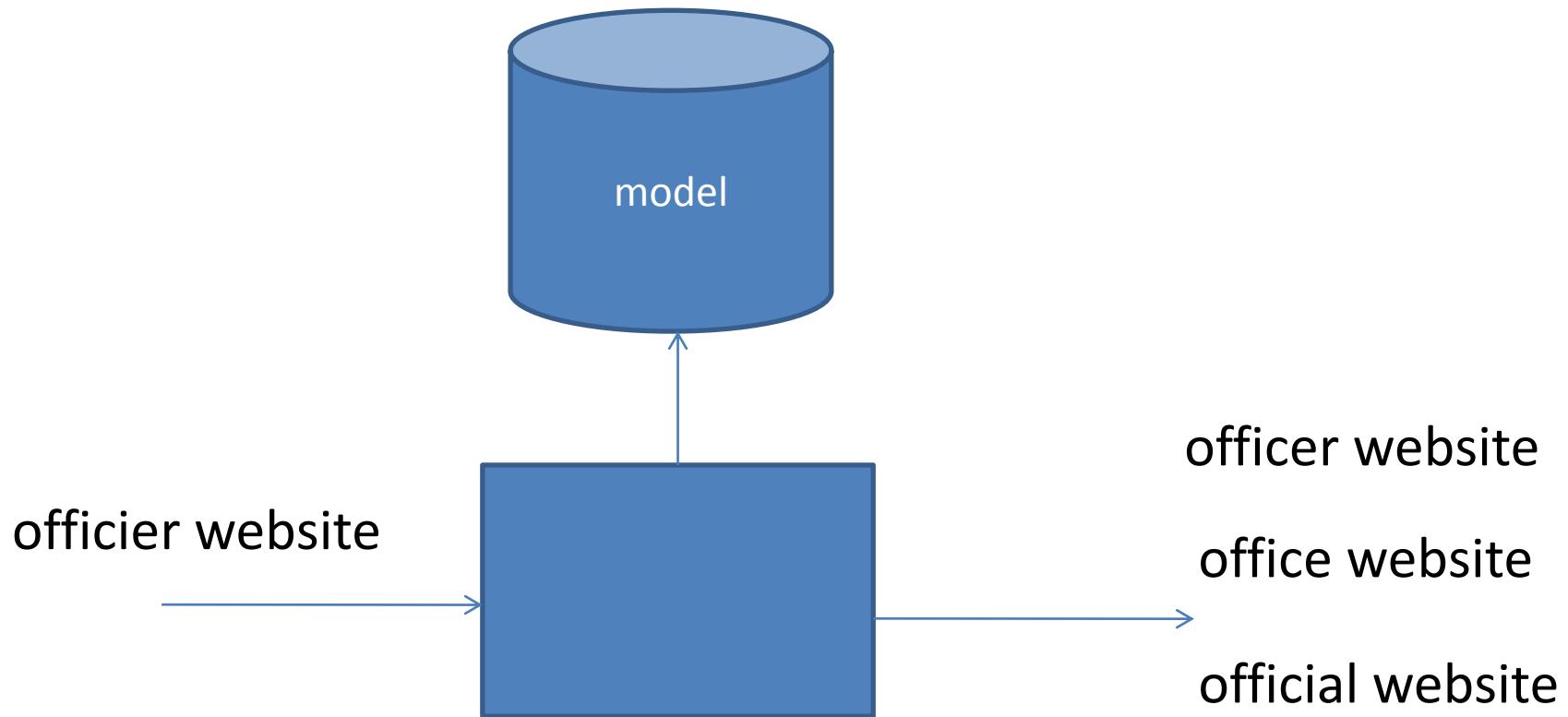
Jiafeng Guo, Gu Xu, Hang Li, Xueqi Cheng
SIGIR 2008

Candidate Selection in Spelling Error Correction

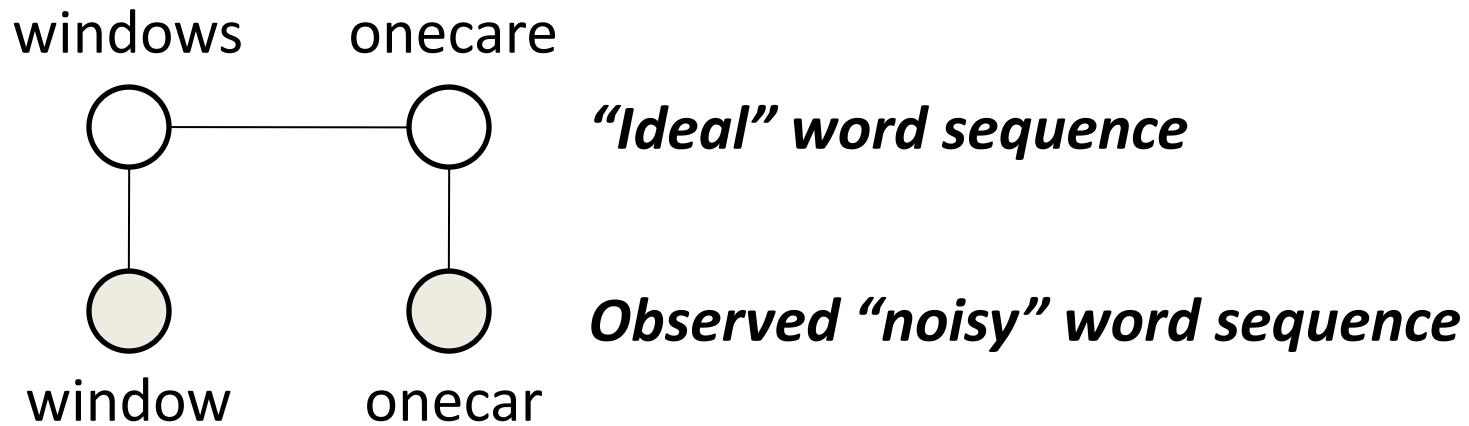
Our Approach = Conditional Random Fields



Candidate Selection in Spelling Error Correction



Candidate Selection Problem



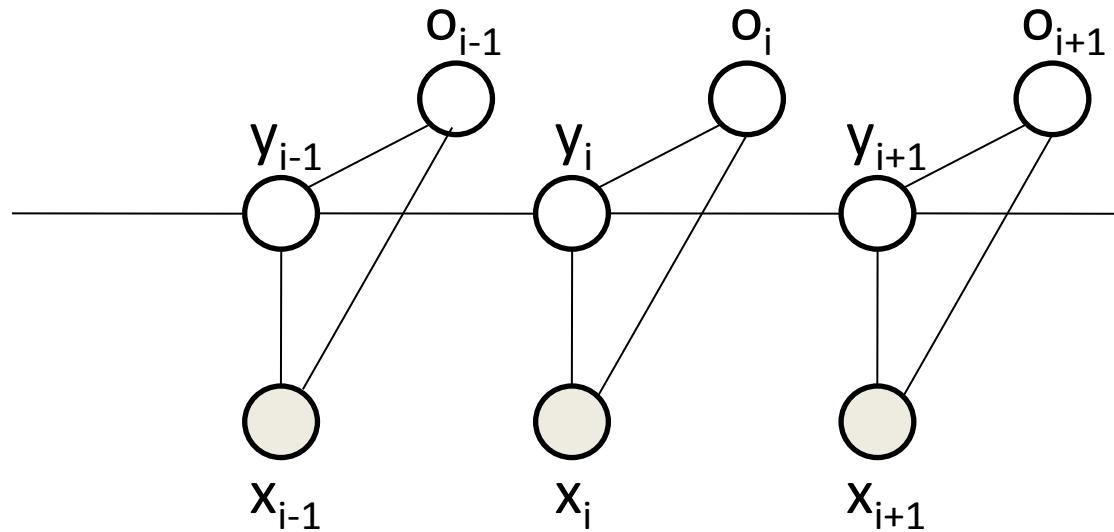
$$y^* = \arg \max_y \Pr(y|x)$$

*"ideal" query
word sequence*

*original query
word sequence*

Conditional Random Fields for Candidate Selection

Introducing Refinement Operations



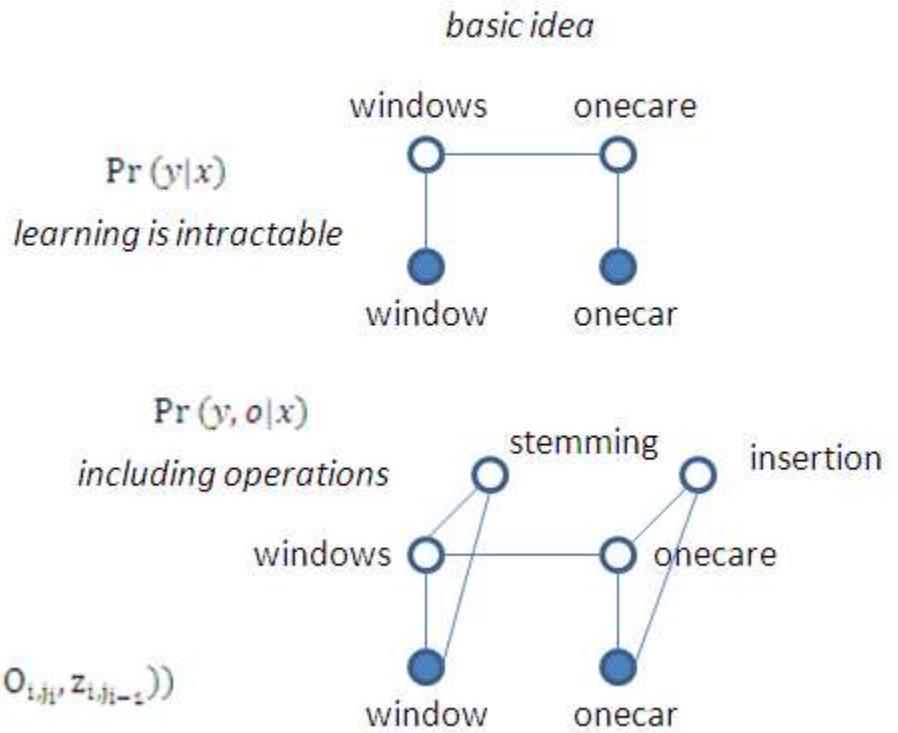
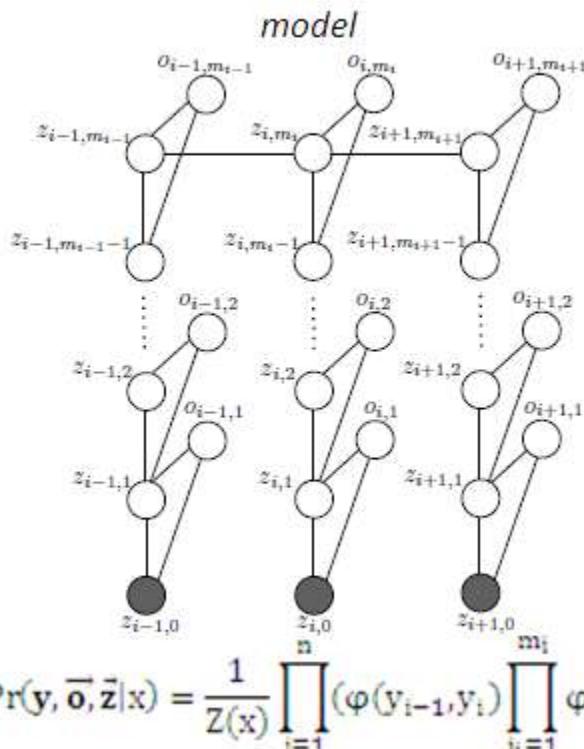
$$\Pr(y, o|x) = \frac{1}{Z(x)} \prod_{i=1}^n \phi(y_{i-1}, y_i) \phi(y_i, o_i, x)$$

Operations

Spelling: insertion, deletion, substitution, transposition, ...

Word Stemming: +s/-s, +es/-es, +ed/-ed, +ing/-ing, ...

Query Refinement Using Conditional Random Fields



IR Matching (Relevance) Models

	Probabilistic Approach	Non Probabilistic Approach
Term Matching (unigram)	BM25[Robertson], LM4IR [Zhai][Ponte & Croft]	Vector Space Model [Salton]
Term Matching (n-gram)	MRF[Metzler & Croft]	<i>Similarity Function [Xu & Li]</i>
Topic Matching	PLSI[Hoffman], LDA[Blei et al]	LSI[Deerwester et al], <i>RLSI[Xu et al]</i>

IR Matching (Relevance) Models

	Probabilistic Approach	Non Probabilistic Approach
Sense Matching (synonym)		Rocchio [Rocchio], <i>Kernel Method</i> [Wu et al]
Sense Matching (spelling)	Generative model [Brill & Moore], <i>Log linear model</i> [Wang et al], CRF [Guo et al]	
Structure Matching	Translation Model [Berger & Lafferty]	

Summary

Summary

- Introduction to Web Search
- Relevance Model (Matching Model)
- Query Term Mismatch
- Learning to Match
- Our Methods
 - Robust Similarity Function Learning Using Kernel Methods
 - Regularized Latent Semantic Indexing
 - Query Generation Using Log Linear Model
 - Query Rewriting Using Conditional Random Fields

Thank You!

hangli@microsoft.com

CCF ADL 2011
Beijing
Aug. 27, 2011

Learning to Rank

Hang Li

Microsoft Research Asia

Outline of Tutorial

1. Learning to Rank
2. Learning for Ranking Creation
3. Learning for Ranking Aggregation
4. Methods of Learning to Rank
5. Applications of Learning to Rank
6. Theory of Learning to Rank
7. Ongoing and Future Work

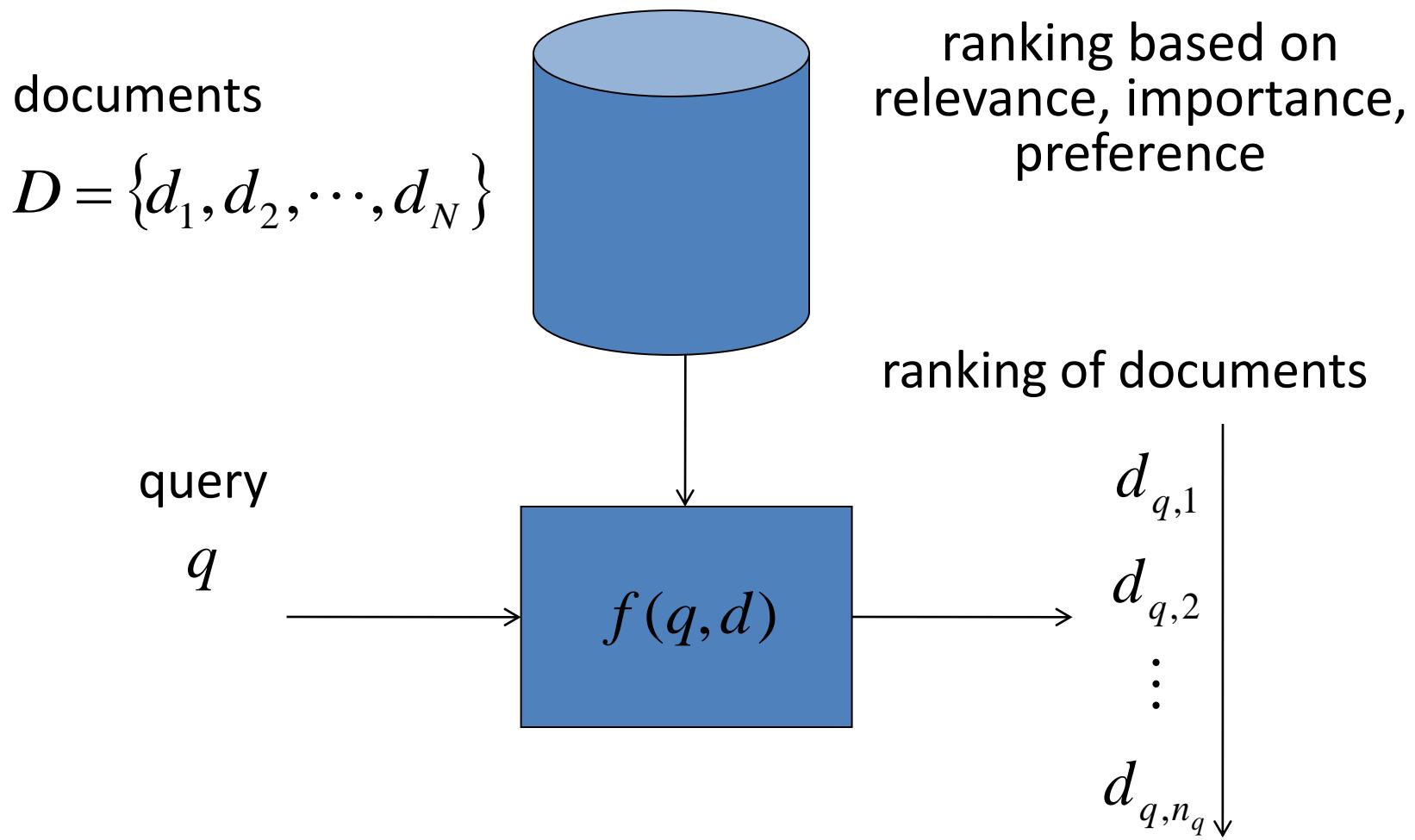
1. Learning to Rank

Ranking Plays Key Role in Many Applications



Ranking Problem:

Example = Document Search



Ranking Problem

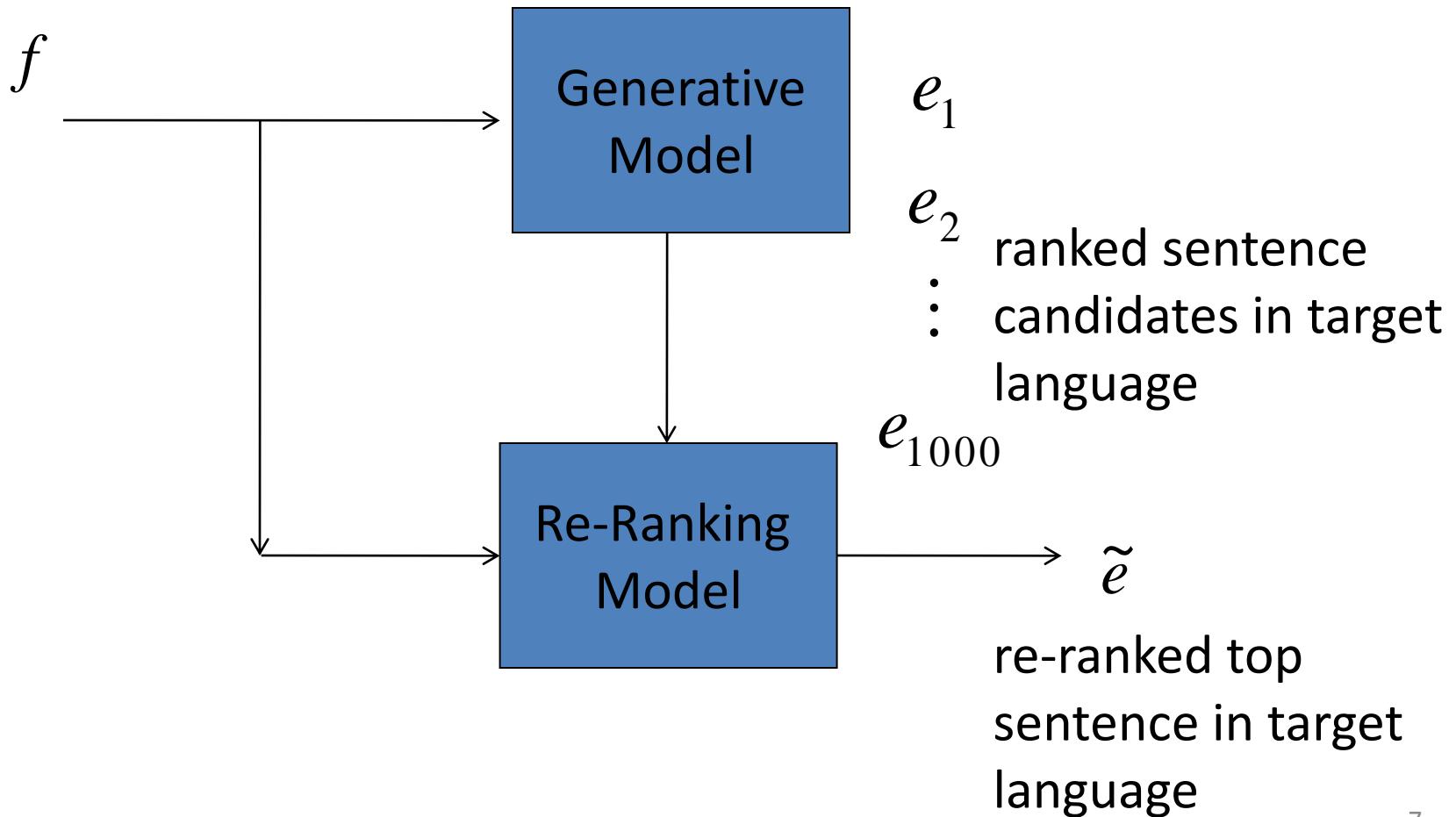
Example = Recommender System

	Item1	Item2	Item3	...	
User1	5	4			
User2	1		2		2
...		?	?	?	
UserM	4	3			

Ranking Problem

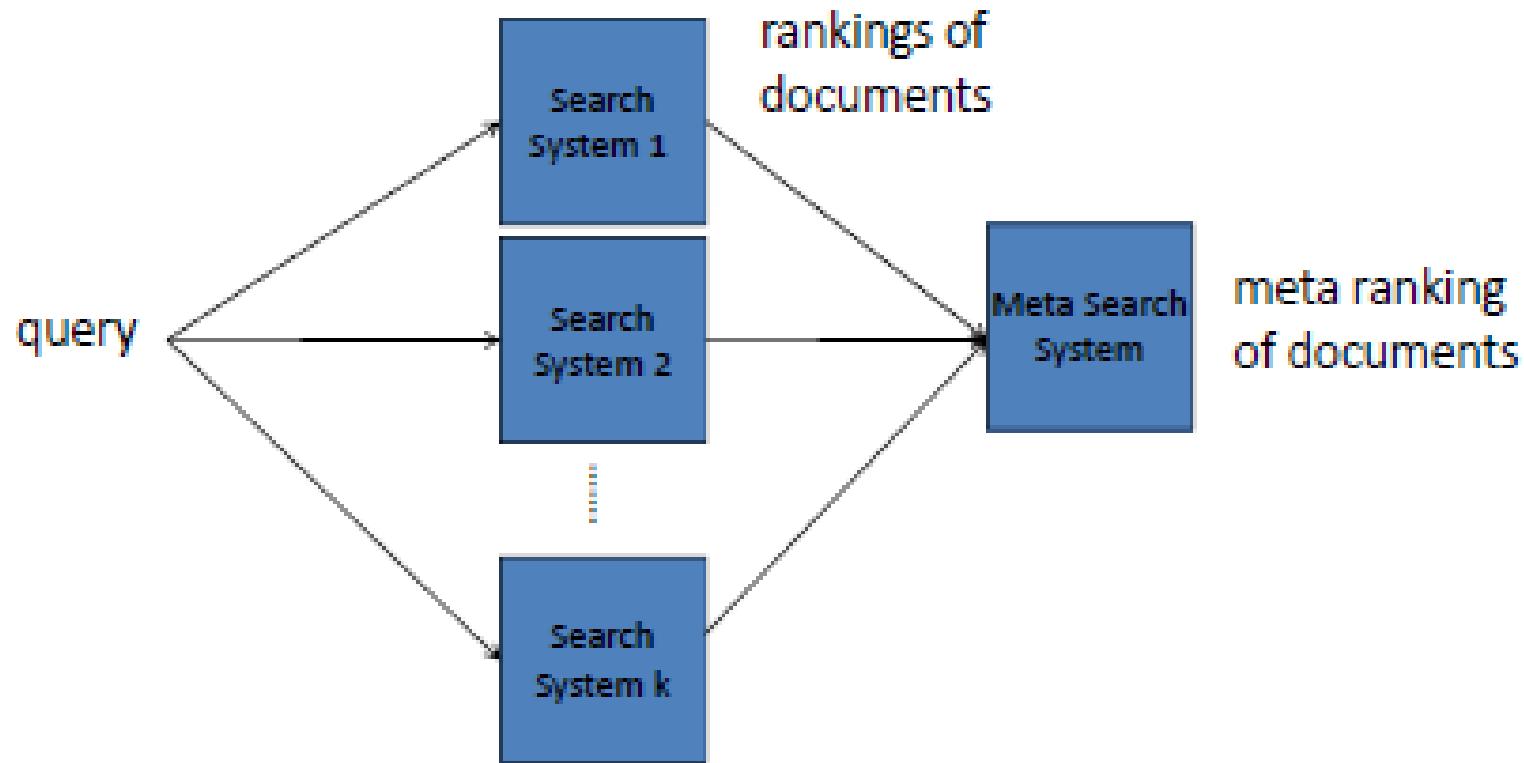
Example = Machine Translation

sentence source language



Ranking Problem

Example = Meta Search



Learning to Rank

- Definition 1 (in broad sense)

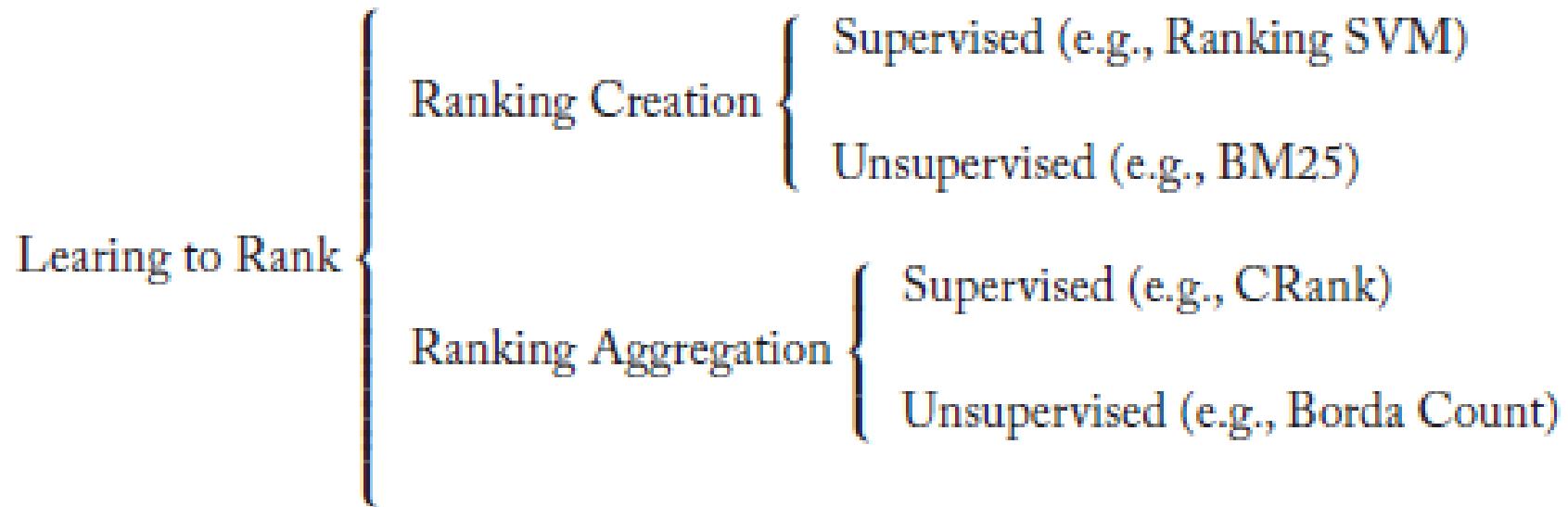
Learning to rank = any machine learning technology for ranking problem

- Definition 2 (in narrow sense)

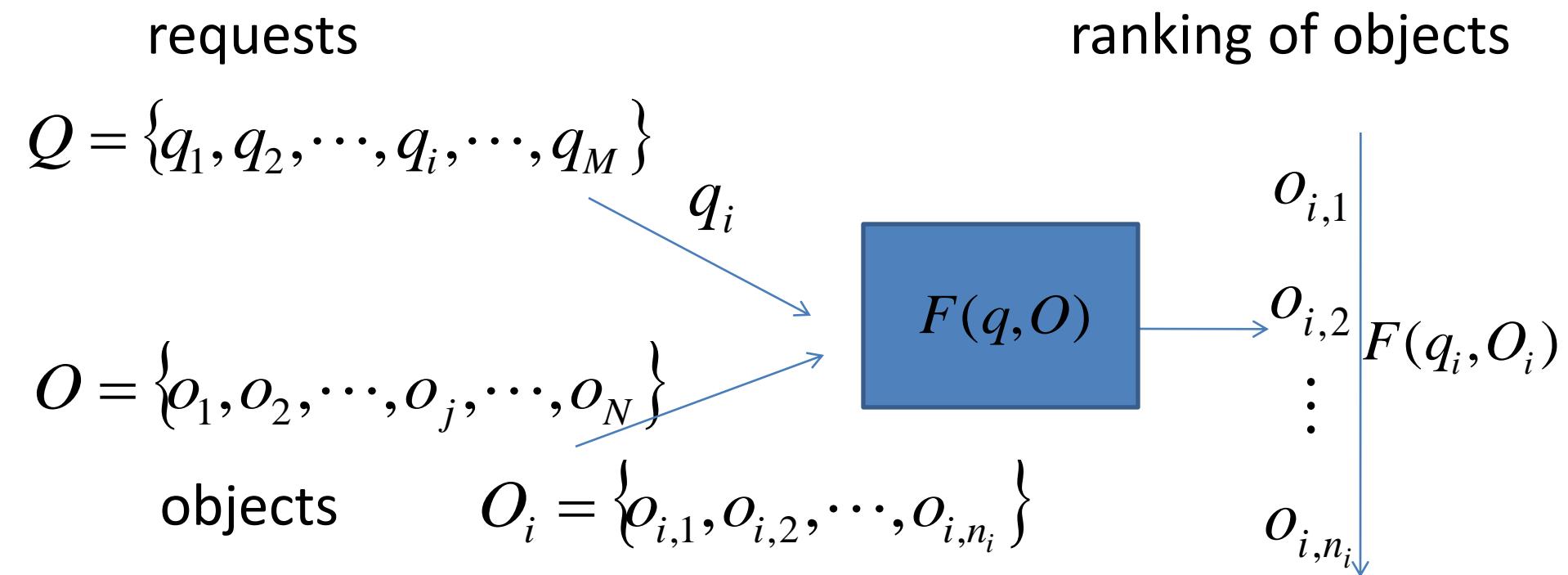
Learning to rank = machine learning technology for ranking creation and ranking aggregation

- This tutorial takes Definition 2

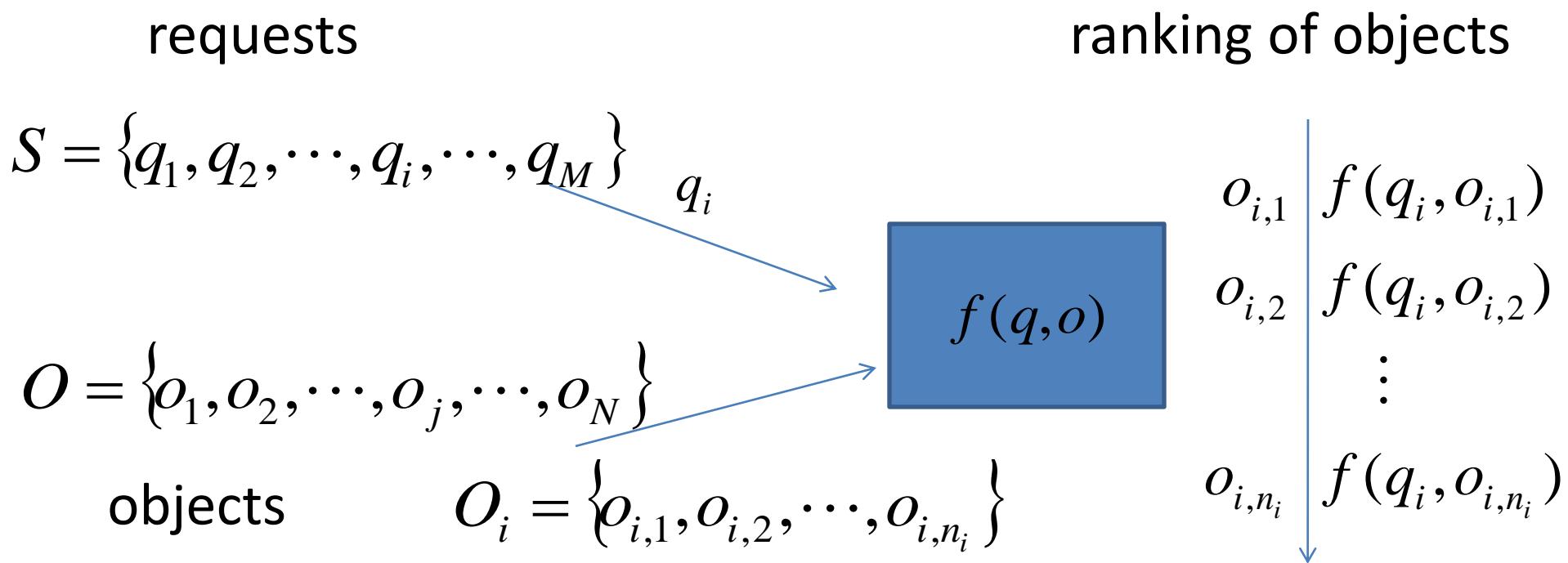
Taxonomy of Problems in Learning to Rank



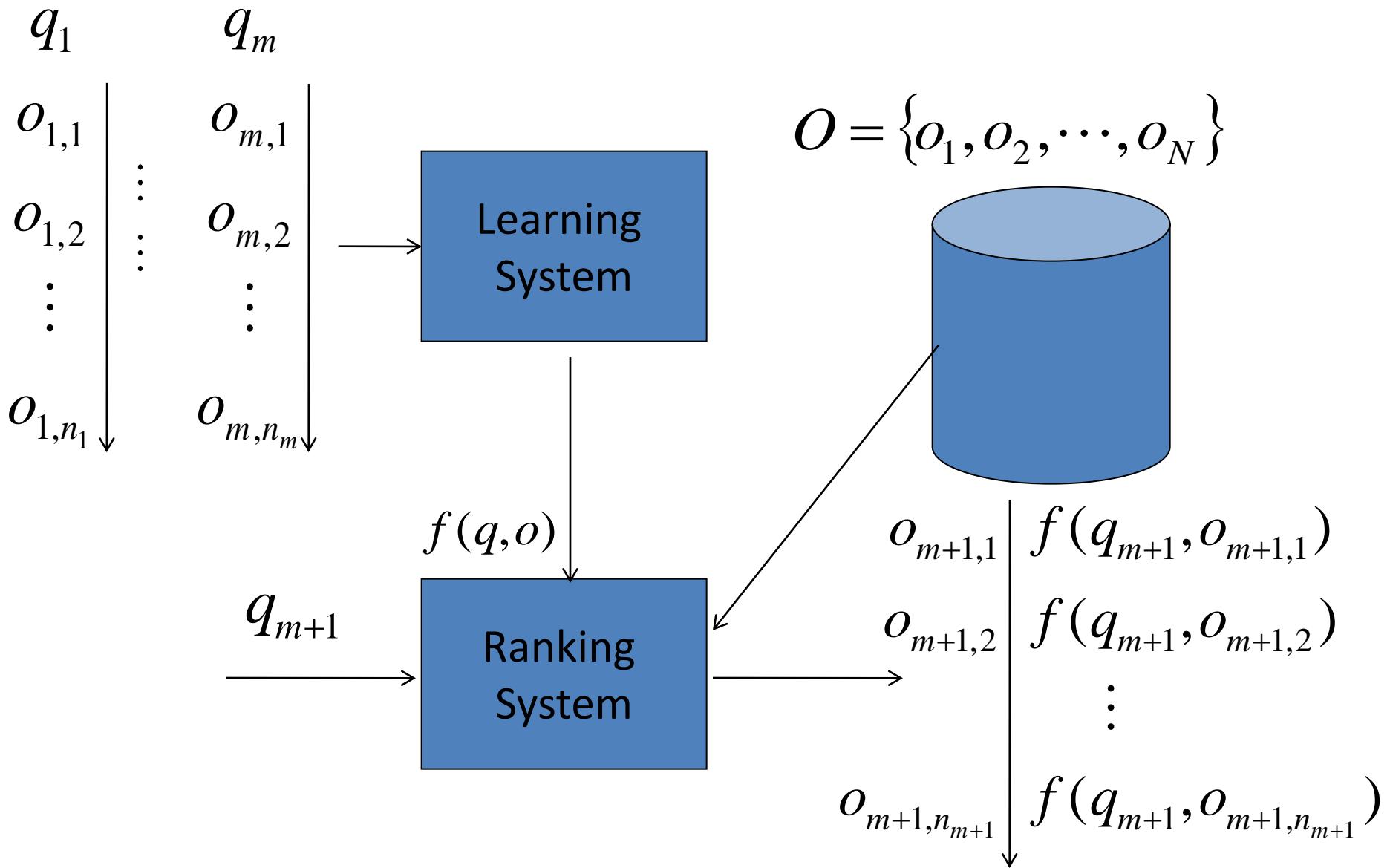
Ranking Creation (with Global Ranking Model)



Ranking Creation (with Local Ranking Model)



Learning for Ranking Creation



Model in Ranking Creation

- Global ranking model

$$s_O = F(q, O)$$

$$\pi = \text{sort}_{s_O}(O),$$

- Local ranking model

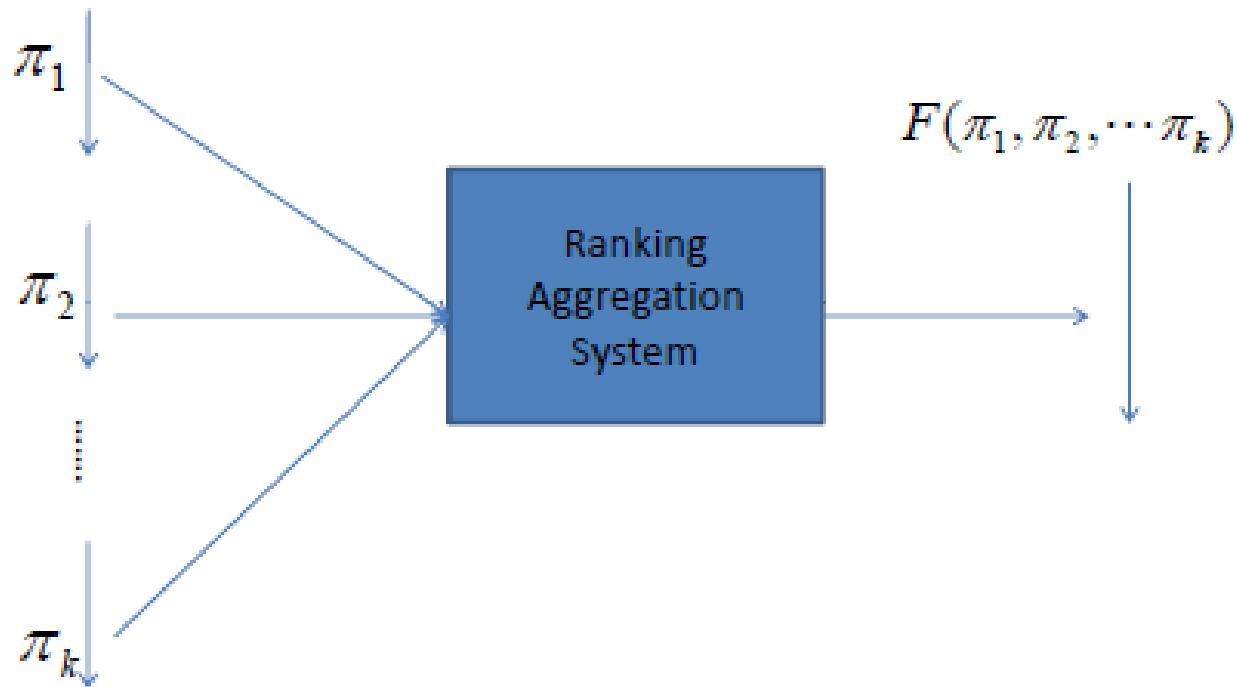
$$s_o = f(q, o)$$

$$\pi = \text{sort}_{s_o, o \in O}(O).$$

Learning for Ranking Creation

- Creating a ranking list of offerings based on request and offerings
- Feature-based
- Usually local ranking model
- Usually supervised learning

Ranking Aggregation



Model in Ranking Aggregation

- Global ranking

$$S_O = F(q, \Sigma)$$

$$\pi = sort_{S_O}(O)$$

Learning for Ranking Aggregation

- Aggregating a ranking list from multiple ranking lists of offerings
- Ranking based
- Usually global ranking model
- Both supervised and unsupervised learning

Technologies on Learning to Rank

- Methods
 - Pointwise Methods
 - Pairwise Methods
 - Listwise Methods
- Theory
 - Generalization
 - Consistency
- Applications
 - Search
 - Collaborative Filtering
 - Key Phrase Extraction

Recent Trends on Learning to Rank

- Successfully applied to web search
- Over 100 publications at SIGIR, ICML, NIPS, etc
- One book on Learning to rank for information retrieval
- 2 sessions at SIGIR every year
- 3 SIGIR workshops
- Special issue at Information Retrieval Journal
- Yahoo Learning to rank challenge
- LETOR benchmark dataset

[http://research.microsoft.com/en-
us/um/beijing/projects/letor/index.html](http://research.microsoft.com/en-us/um/beijing/projects/letor/index.html)



MORGAN & CLAYPOOL PUBLISHERS

Learning to Rank for Information Retrieval and Natural Language Processing

Hang Li

*SYNTHESIS LECTURES ON
HUMAN LANGUAGE TECHNOLOGIES*

Graeme Hirst, *Series Editor*

Scope of This Tutorial

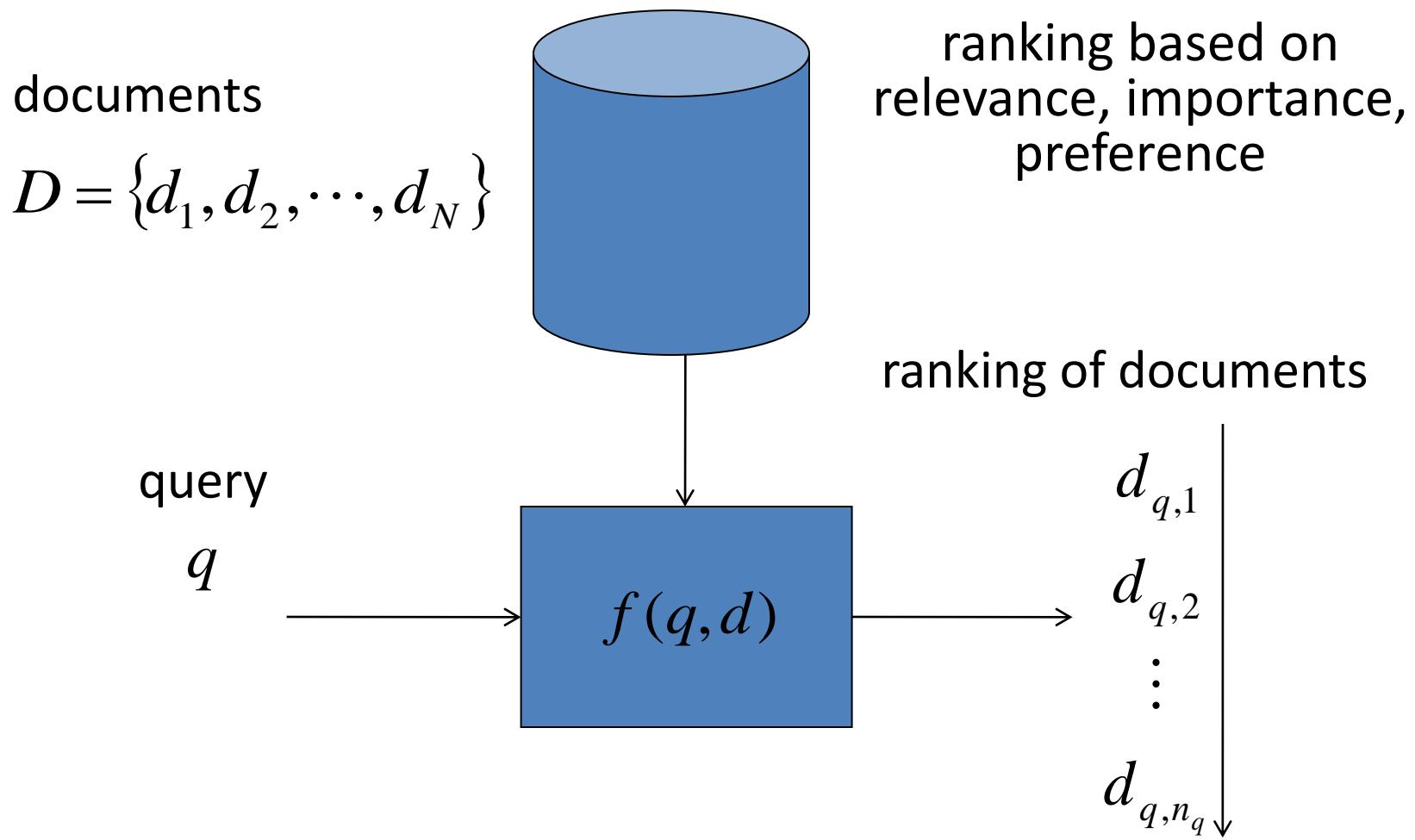
- Overview of Learning to rank technologies
- Focusing on Learning to rank methods
- Touching theoretical issues
- Showing future directions
- Knowledge necessary for this tutorial:
Machine Learning

2. Learning for Ranking Creation

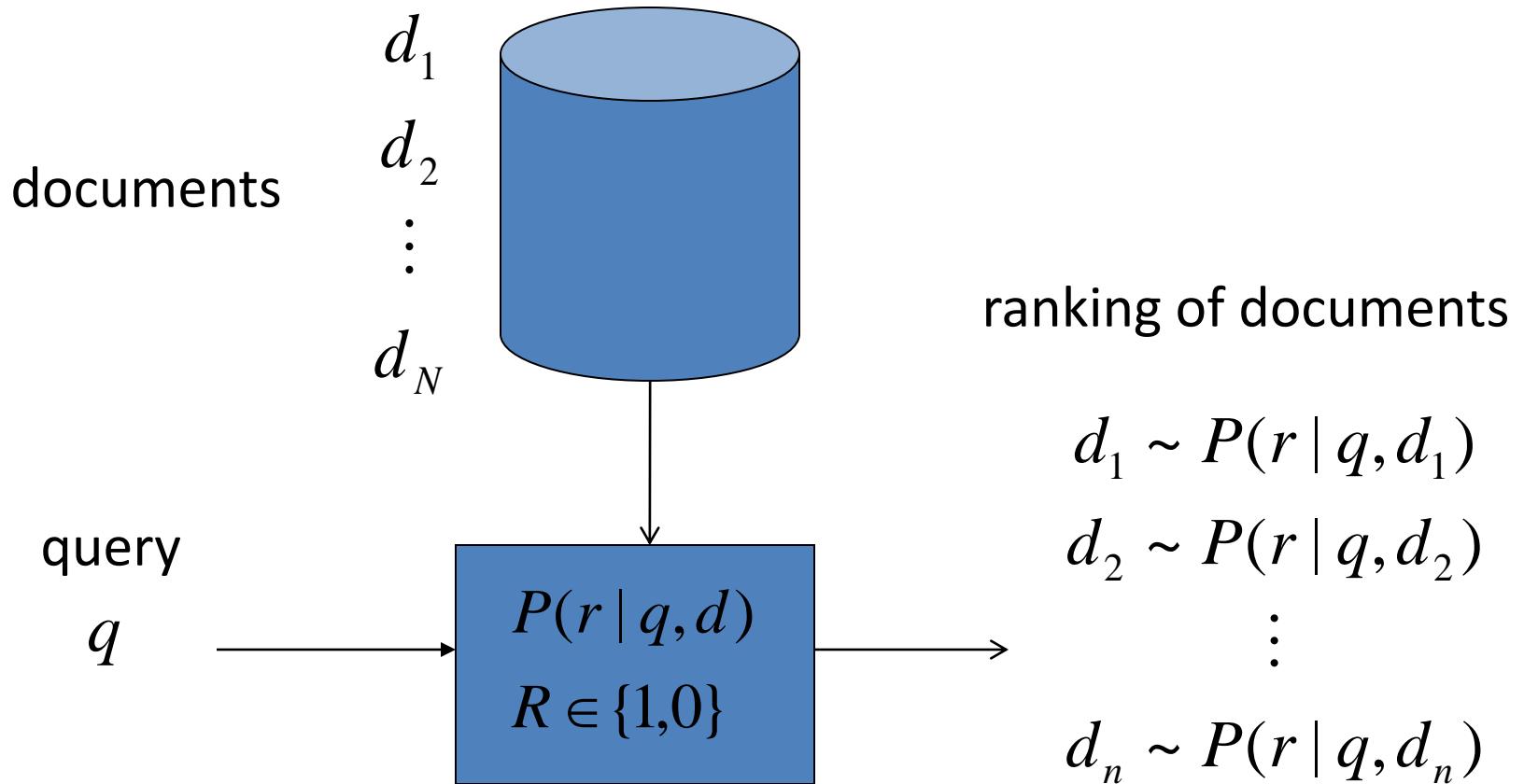
2.1 Document Retrieval as Example

Ranking Problem:

Example = Document Search

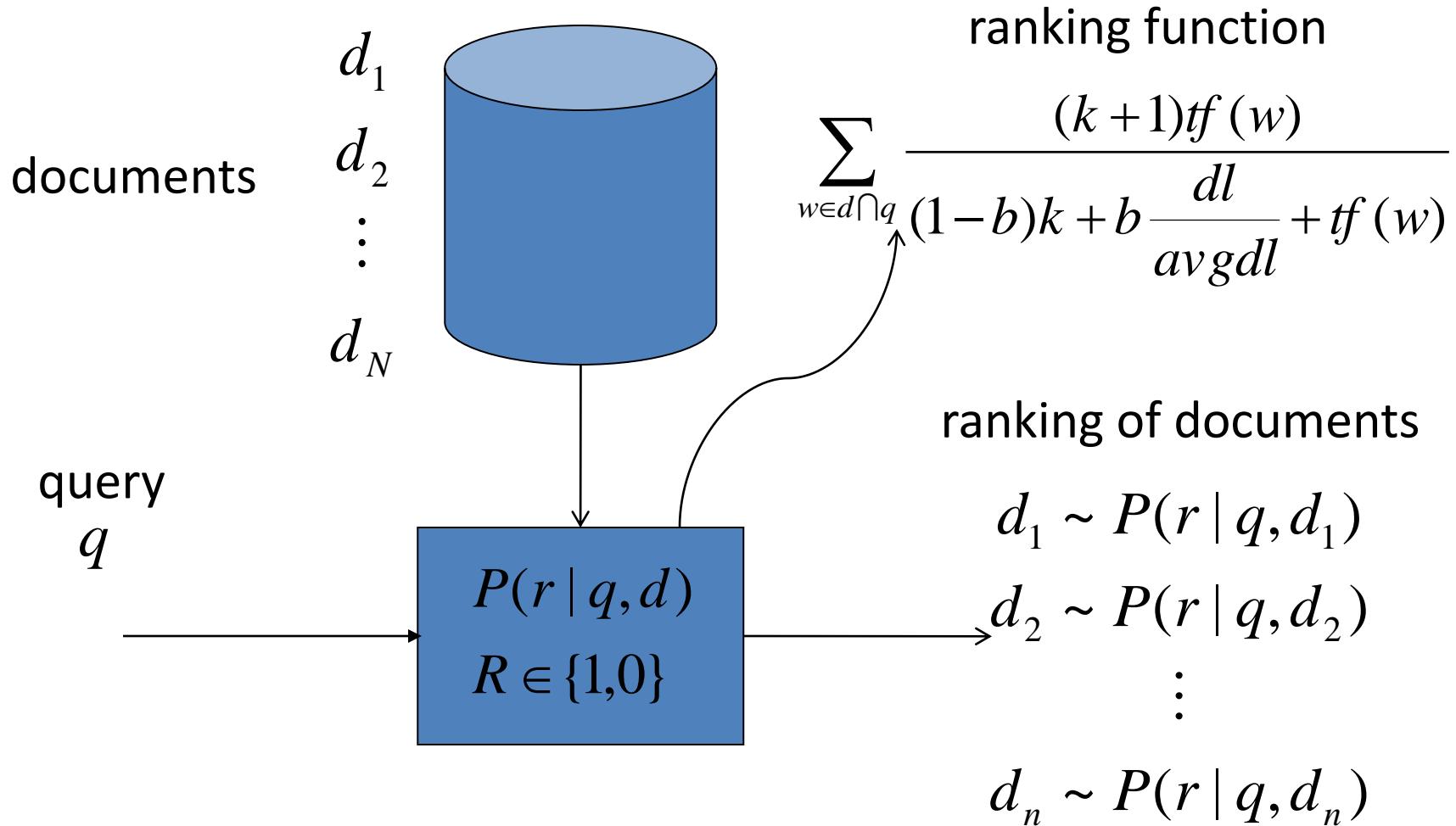


Traditional Approach = Probabilistic Model



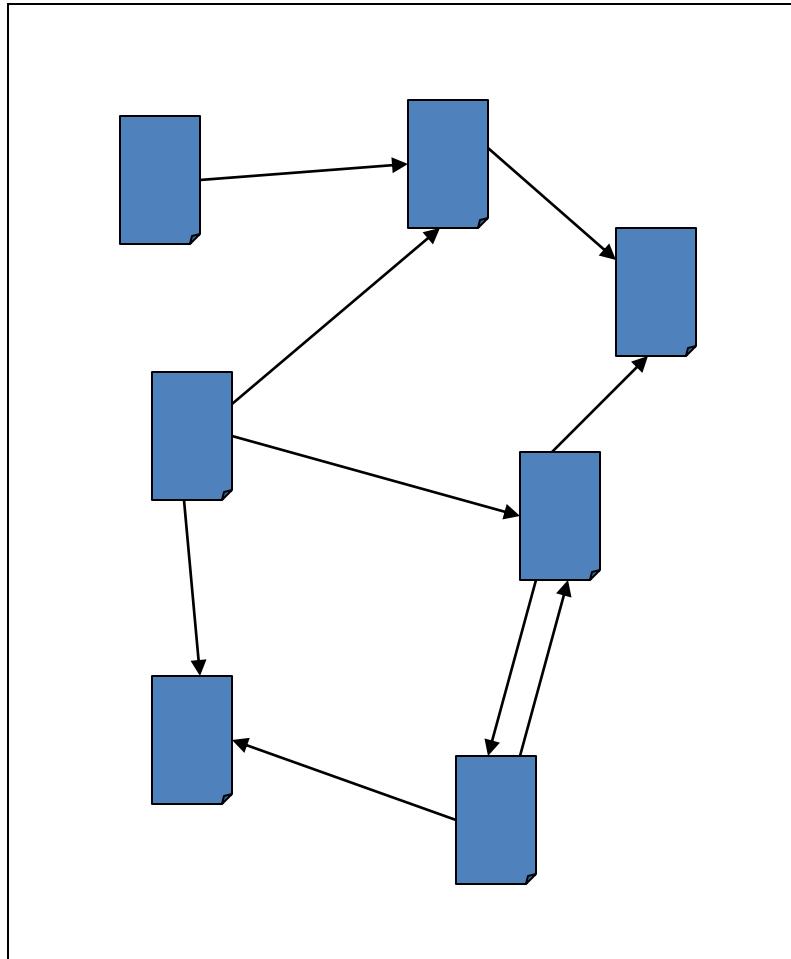
BM25

[Robertson & Walker 94]



PageRank

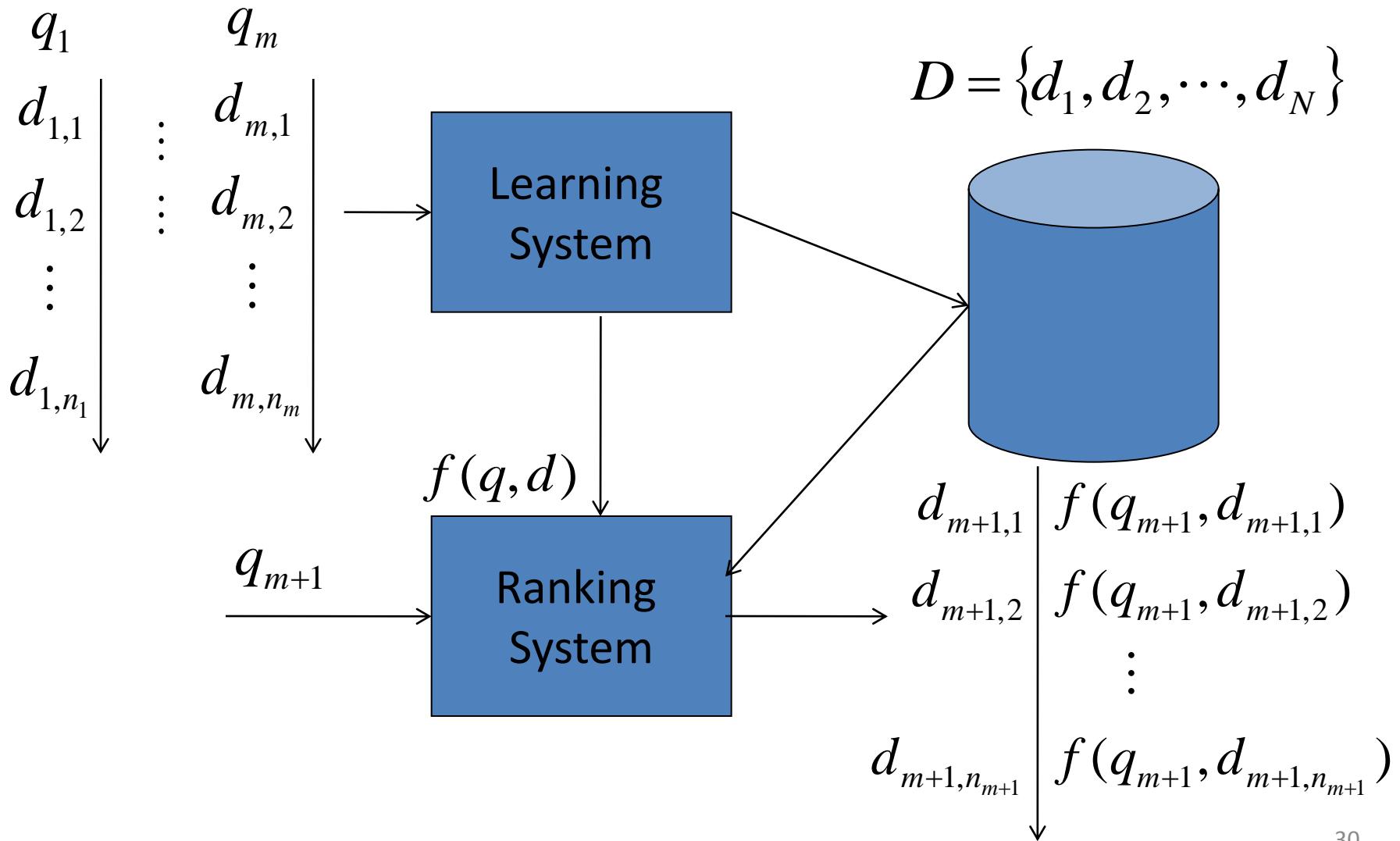
[Page et al, 1999]



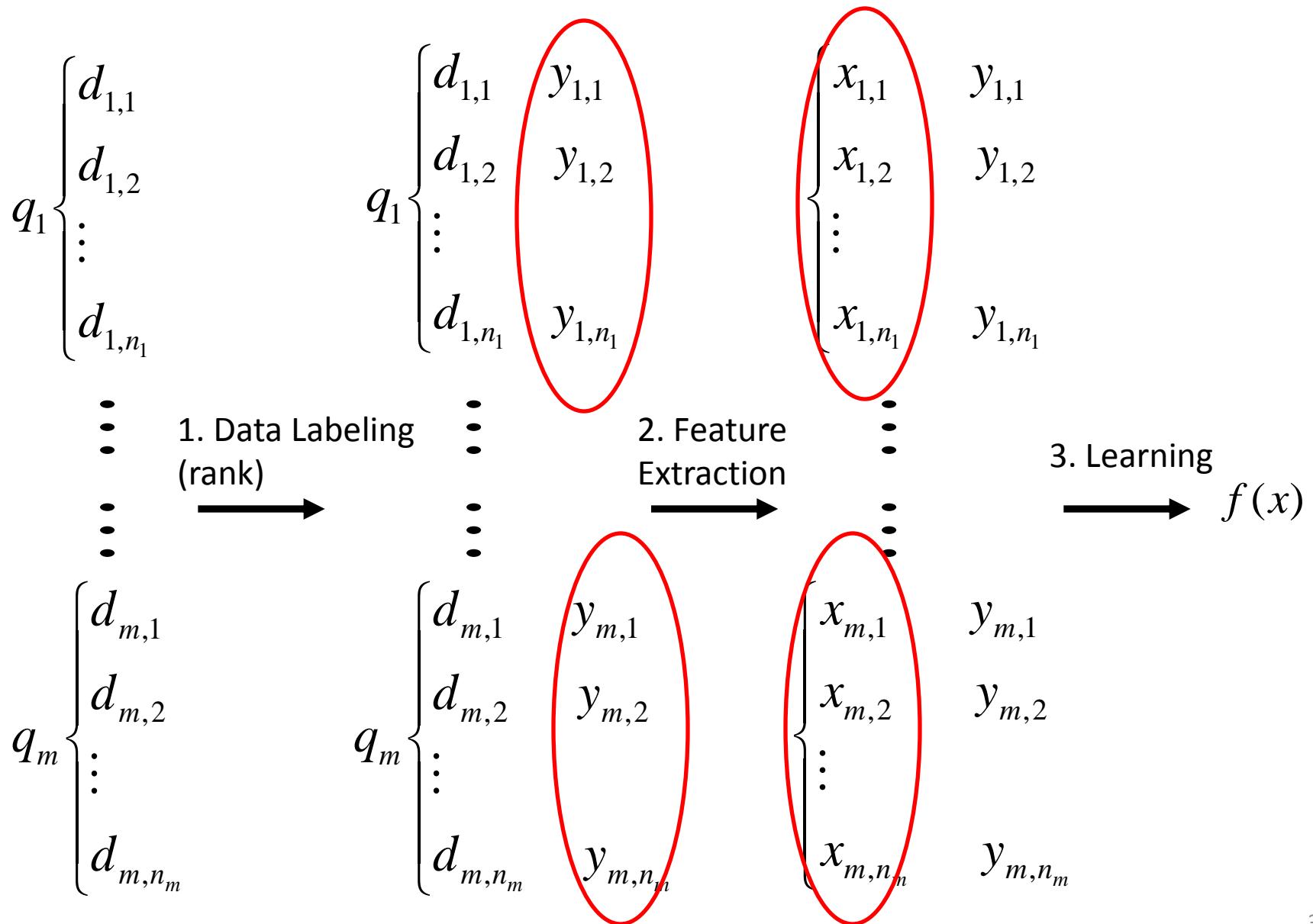
$$P(d_i) = \alpha \sum_{d_j \in M(d_i)} \frac{P(d_j)}{L(d_j)} + (1 - \alpha) \frac{1}{n}$$

2.1 Learning Task

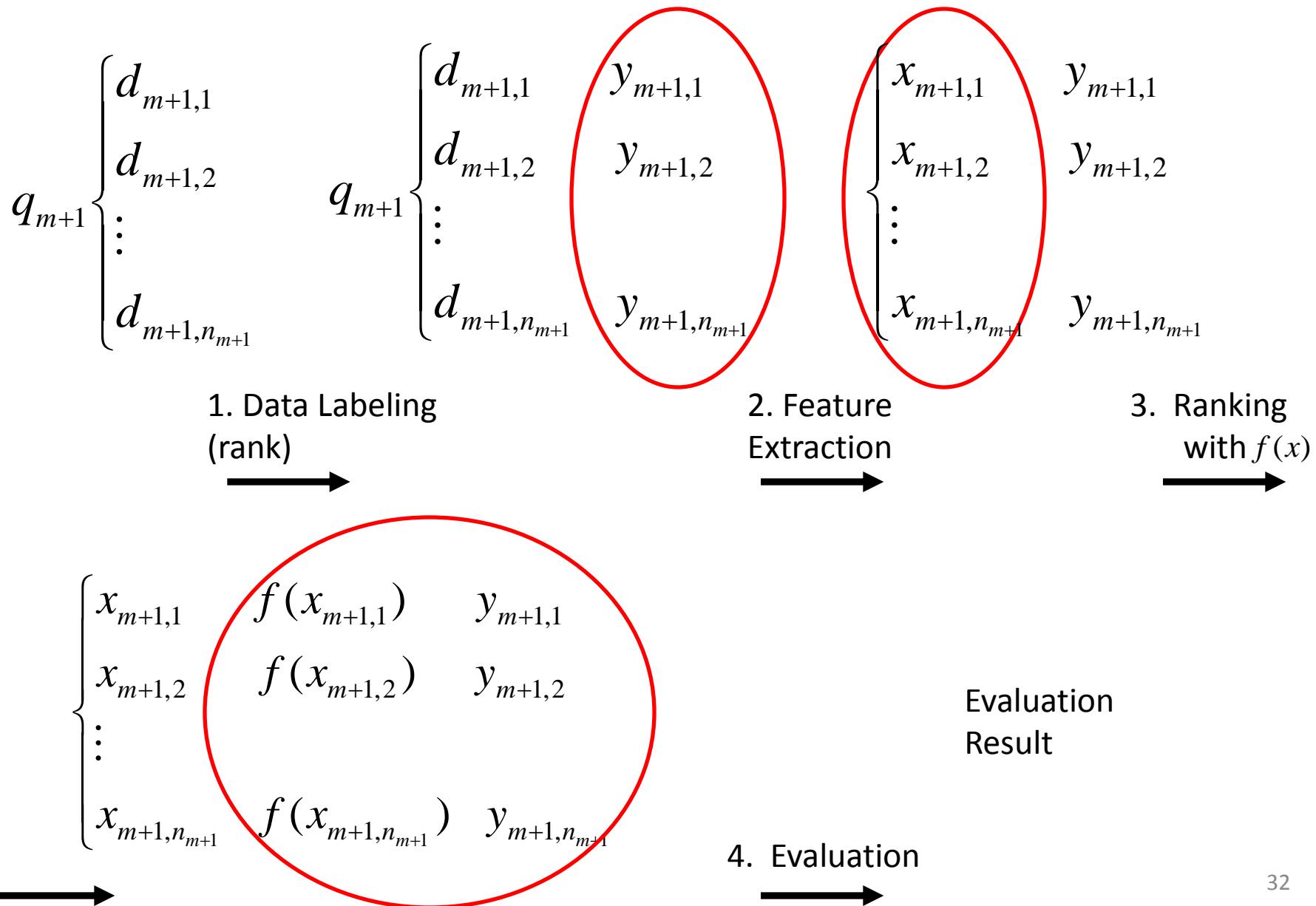
New Approach = Learning to Rank



Training Process



Testing Process



Notes

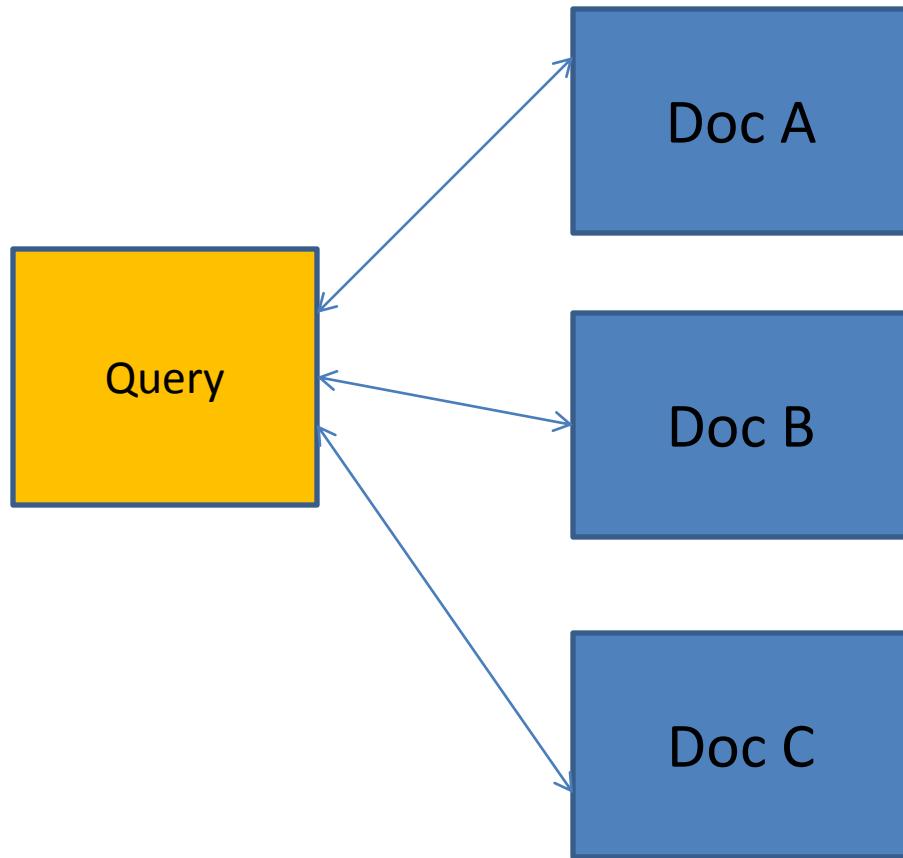
- Features are functions of query and document
- Query and associated documents form a group
- Groups are i.i.d. data
- Feature vectors within group are not i.i.d. data
- Ranking model is function of features
- Several data labeling methods (here labeling of grade)

Issues in Learning to Rank

- Data Labeling
- Feature Extraction
- Evaluation Measure
- Learning Method (Model, Loss Function, Algorithm)

Data Labeling Problem

- E.g., relevance of documents w.r.t. query

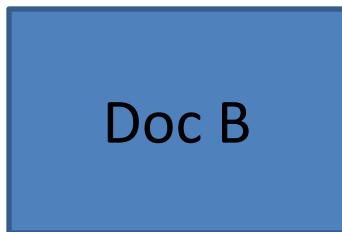
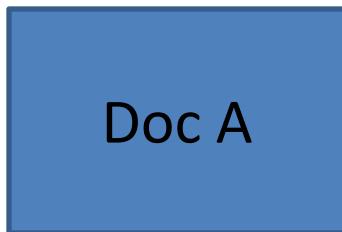


Data Labeling Methods

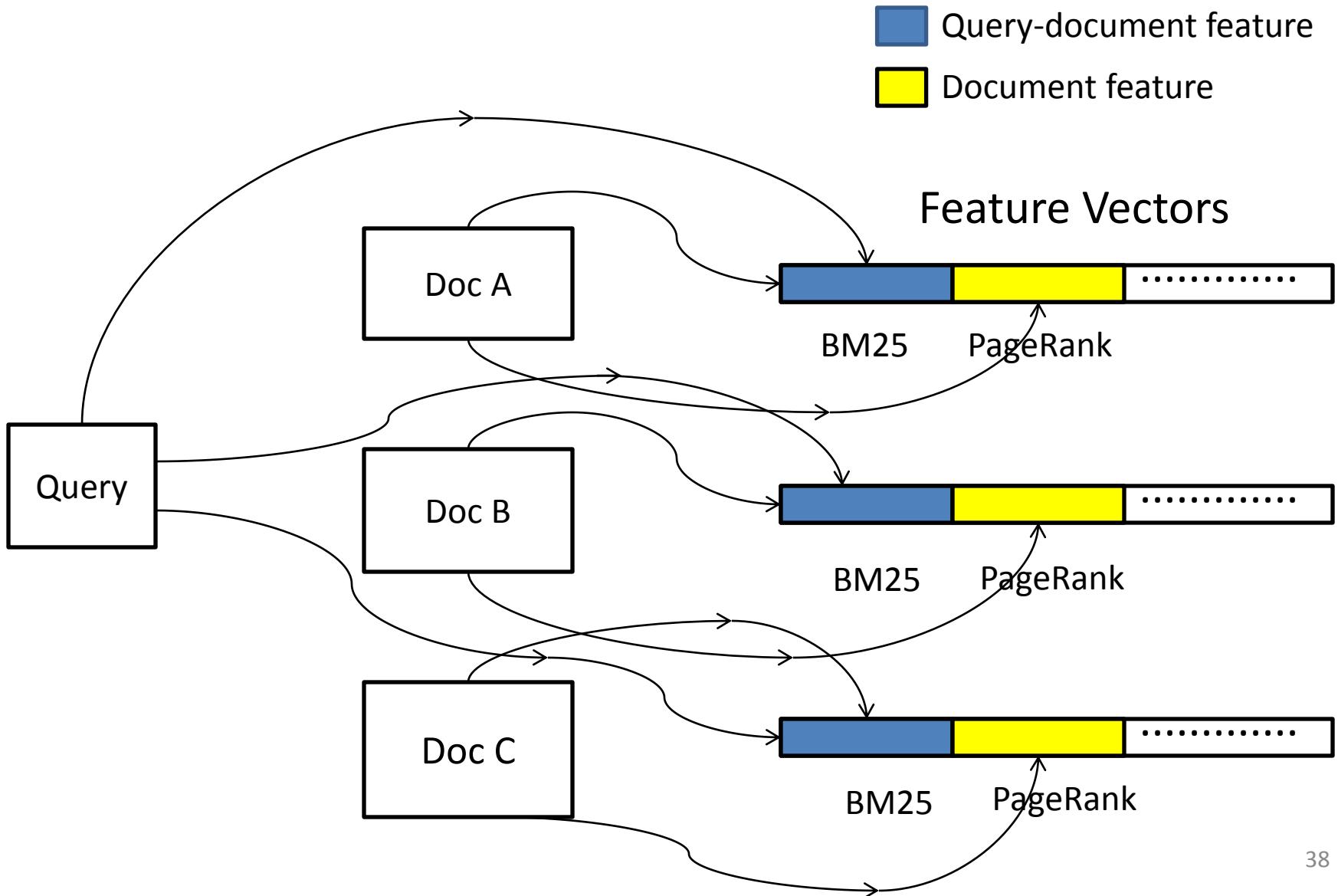
- Labeling of Grades
 - Multiple levels (e.g., relevant, partially relevant, irrelevant)
 - Widely used in IR
- Labeling of Ordered Pairs
 - Ordered pairs between documents (e.g. A>B, B>C)
 - Implicit relevance judgment: derived from click-through data
- Creation of List
 - List (or permutation) of documents is given
 - Ideal but difficult to implement

Implicit Relevance Judgment

ranking of documents at search system



Feature Extraction



Example Features

Table 2.3: Example Features of Learning to Rank for Web Search

Feature	Type	Explanation	Reference
Number of occurrences	Matching	number of times query exactly occurs in title, anchor, URL, extracted title, associated query, and body	
BM25	Matching	BM25 scores on title, anchor, URL, extracted title, associated query, and body	[90]
N-gram BM25	Matching	BM25 scores of n-grams on title, anchor, URL, extracted title, associated query, and body	[109]
Edit Distance	Matching	edit distance scores between query and title, anchor, URL, extracted title, associated query, and span in body (minimum length of text segment including all query words [94])	Our unpublished work
Number of in-links	Document	number of in-links to the page	
PageRank	Document	importance score of page calculated on web link graph	[78]
Number of clicks	Document	number of clicks on the page in search log	
BrowseRank	Document	importance score of page calculated on user browsing graph	[72]
Spam score	Document	likelihood of spam page	[45]
Page quality score	Document	likelihood of low quality page	[10]

Evaluation Measures

- Important to rank top results correctly
- Measures
 - NDCG (Normalized Discounted Cumulative Gain)
 - MAP (Mean Average Precision)
 - MRR (Mean Reciprocal Rank)
 - WTA (Winners Take All)
 - Kendall's Tau

NDCG

- Evaluating ranking using labeled grades
- NDCG at position j

$$\frac{1}{n_j} \sum_{i=1}^j (2^{r(i)} - 1) / \log(1+i)$$

NDCG (cont')

- Example: perfect ranking
 - $(3, 3, 2, 2, 1, 1, 1)$ grade $r=3,2,1$
 - $(7, 7, 3, 3, 1, 1, 1)$ gain $2^{r(j)} - 1$
 - $(1, 0.63, 0.5, 0.43, 0.39, 0.36, 0.33)$ position discount
 - $(7, 11.41, 12.91, \dots)$ DCG $\sum_{i=1}^j (2^{r(i)} - 1) / \log(1 + i)$
 - $(1/7, 1/11.41, 1/12.91, \dots)$ normalizing factor n_j
 - $(1, 1, 1, 1, 1, 1, 1)$ NDCG for perfect ranking

NDCG (cont')

- Example: imperfect ranking
 - $(2, 3, 2, 3, 1, 1, 1)$
 - $(3, 7, 3, 7, 1, 1, 1)$ Gain
 - $(1, 0.63, 0.5, 0.43, 0.39, 0.36, 0.33)$ Position discount
 - $(3, 7.41, 8.91, \dots)$ DCG
 - $(1/7, 1/11.41, 1/12.91, \dots)$ normalizing factor
 - $(0.43, 0.65, 0.69, \dots)$ NDCG
- Imperfect ranking decreases NDCG

MAP

- Evaluating ranking using two grades
- AP

$$AP = \frac{\sum_{j=1}^{n_i} P(j) \cdot y_{i,j}}{\sum_{j=1}^{n_i} y_{i,j}},$$

$$P(j) = \frac{\sum_{k:\pi_i(k) \leq \pi_i(j)} y_{i,k}}{\pi_i(j)},$$

MAP (cont')

- Example: perfect ranking
 - (1,0,1,1,0,0,0) grade $r=0,1$
 - (1, -, 0.67, 0.75, -, -, -) $P(j)$ precision at position j
 - 0.81 AP average precision

Relations with Other Learning Tasks

- No need to predict category
vs Classification
- No need to predict value of $f(q, d)$
vs Regression
- Relative ranking order is more important
vs Ordinal regression
- *Learning to rank can be approximated by classification, regression, ordinal regression*

Ordinal Regression (Ordinal Classification)

- Categories are ordered
 - 5, 4, 3, 2, 1
 - e.g., rating restaurants
- Prediction
 - Map to ordered categories

2.3 Learning Approaches

Three Major Approaches

- Pointwise approach
 - Pairwise approach
 - Listwise approach
-
- SVM based
 - Boosting based
 - Neural Network based
 - Others

Categorization of Learning to rank Methods

Table 2.6: Categorization of Learning to Rank Methods

	SVM	Boosting	Neural Net	Others
Pointwise	OC SVM [92]	McRank [67]		Prank [30] Subset Ranking [29]
Pairwise	Ranking SVM [48] IR SVM [13]	RankBoost [37] GBRank [115] LambdaMART [102]	RankNet [11]	
Listwise	SVM [111] PermuRank [110]	MAP AdaRank [108]	ListNet [14] ListMLE [104]	SoftRank [95] AppRank [81]

Pointwise Approach

- Transforming ranking to regression, classification, or ordinal classification
- Query-document group structure is ignored

Pointwise Approach

Table 2.7: Characteristics of Pointwise Approach

Pointwise Approach (Classification)		
	Learning	Ranking
Input	feature vector x	feature vectors $\mathbf{x} = \{x_i\}_{i=1}^n$
Output	category $y = \text{classifier}(f(x))$	ranking list $\text{sort}(\{f(x_i)\}_{i=1}^n)$
Model	$\text{classifier}(f(x))$	ranking model $f(x)$
Loss	classification loss	ranking loss
Pointwise Approach (Regression)		
	Learning	Ranking
Input	feature vector x	feature vectors $\mathbf{x} = \{x_i\}_{i=1}^n$
Output	real number $y = f(x)$	ranking list $\text{sort}(\{f(x_i)\}_{i=1}^n)$
Model	regression model $f(x)$	ranking model $f(x)$
Loss	regression loss	ranking loss

Pointwise Approach

Pointwise Approach (Ordinal Classification)		
	Learning	Ranking
Input	feature vector x	feature vectors $\mathbf{x} = \{x_i\}_{i=1}^n$
Output	ordered category $y = \text{threshold}(f(x))$	ranking list $\text{sort}(\{f(x_i)\}_{i=1}^n)$
Model	$\text{threshold}(f(x))$	ranking model $f(x)$
Loss	ordinal classification loss	ranking loss

Pairwise Approach

- Transforming ranking to pairwise classification
- Query-document group structure is ignored

Pairwise Approach

Table 2.8: Characteristics of Pairwise Approach

Pairwise Approach (Classification)		
	Learning	Ranking
Input	feature vectors $x^{(1)}, x^{(2)}$	feature vectors $\mathbf{x} = \{x_i\}_{i=1}^n$
Output	pairwise classification $\text{classifier}(f(x^{(1)}) - f(x^{(2)}))$	ranking list $\text{sort}(\{f(x_i)\}_{i=1}^n)$
Model	$\text{classifier}(f(x))$	ranking model $f(x)$
Loss	pairwise classification loss	ranking loss
Pairwise Approach (Regression)		
	Learning	Ranking
Input	feature vectors $x^{(1)}, x^{(2)}$	feature vectors $\mathbf{x} = \{x_i\}_{i=1}^n$
Output	pairwise regression $f(x^{(1)}) - f(x^{(2)})$	ranking list $\text{sort}(\{f(x_i)\}_{i=1}^n)$
Model	regression model $f(x)$	ranking model $f(x)$
Loss	pairwise regression loss	ranking loss

Listwise Approach

- List as instance
- Query-document group structure is used
- Straightforwardly represents learning to rank problem

Listwise Approach

Table 2.9: Characteristics of Listwise Approach

Listwise Approach		
	Learning	Ranking
Input	feature vectors $\mathbf{x} = \{x_i\}_{i=1}^n$	feature vectors $\mathbf{x} = \{x_i\}_{i=1}^n$
Output	ranking list $\text{sort}(\{f(x_i)\}_{i=1}^n)$	ranking list $\text{sort}(\{f(x_i)\}_{i=1}^n)$
Model	ranking model $f(x)$	ranking model $f(x)$
Loss	listwise loss function	ranking loss

Learning to rank Methods

- Pointwise Approach
 - Subset Ranking [Cossack and Zhang, 2006]: Regression
 - McRank [Li et al 2007]: Multi-Class Classification Using Boosting Tree
 - PRank [Crammer and Singer 2002]: Ordinal Classification Using Perceptron
 - OC SVM [Shashua & Levin 2002]: Ordinal Classification Using SVM

Learning to rank Methods

- Pairwise Approach
 - Ranking SVM: Pairwise Classification Using SVM
 - RankBoost [Freund et al 2003]: Pairwise Classification Using Boosting
 - RankNet [Burges et al 2005]: Pairwise Classification Using Neural Net
 - Frank [Tsai et al 2007]: Pairwise Classification Using Fidelity Loss and Neural Net
 - GBRank [Zheng et al 2007]: Pairwise Regression Using Boosting Tree
 - IR SVM [Cao et al 2006]: Cost-sensitive Pairwise Classification Using SVM
 - LambdaRank [Burges et al 2007]: Using Implicit Loss Function
 - LambdaMART [Wu et al 2010]: Using Implicit Loss Function

Learning to rank Methods

- Listwise Approach
 - ListNet [Cao et al 2007]: Probabilistic Ranking Model
 - ListMLE [Xia et al 2008]: Probabilistic Ranking Model
 - AdaRank [Xu and Li 2007]: Direct Optimization of Evaluation Measure
 - SVM Map [Yue et al 2007]: Direct Optimization of Evaluation Measure (Using Structure SVM)
 - PermuRank [Xu et al 2008]: Direct Optimization of Evaluation Measure
 - Soft Rank [Taylor et al 2008]: Approximation of Evaluation Measure
 - AppRank [Qin et al 2010]: Approximation of Evaluation Measure

LETOR Data Set

- Available at
 - <http://research.microsoft.com/~letor/>
- Data Corpora: TREC, OHSUMED
- Training/Validation/Test split
- Standard IR Features

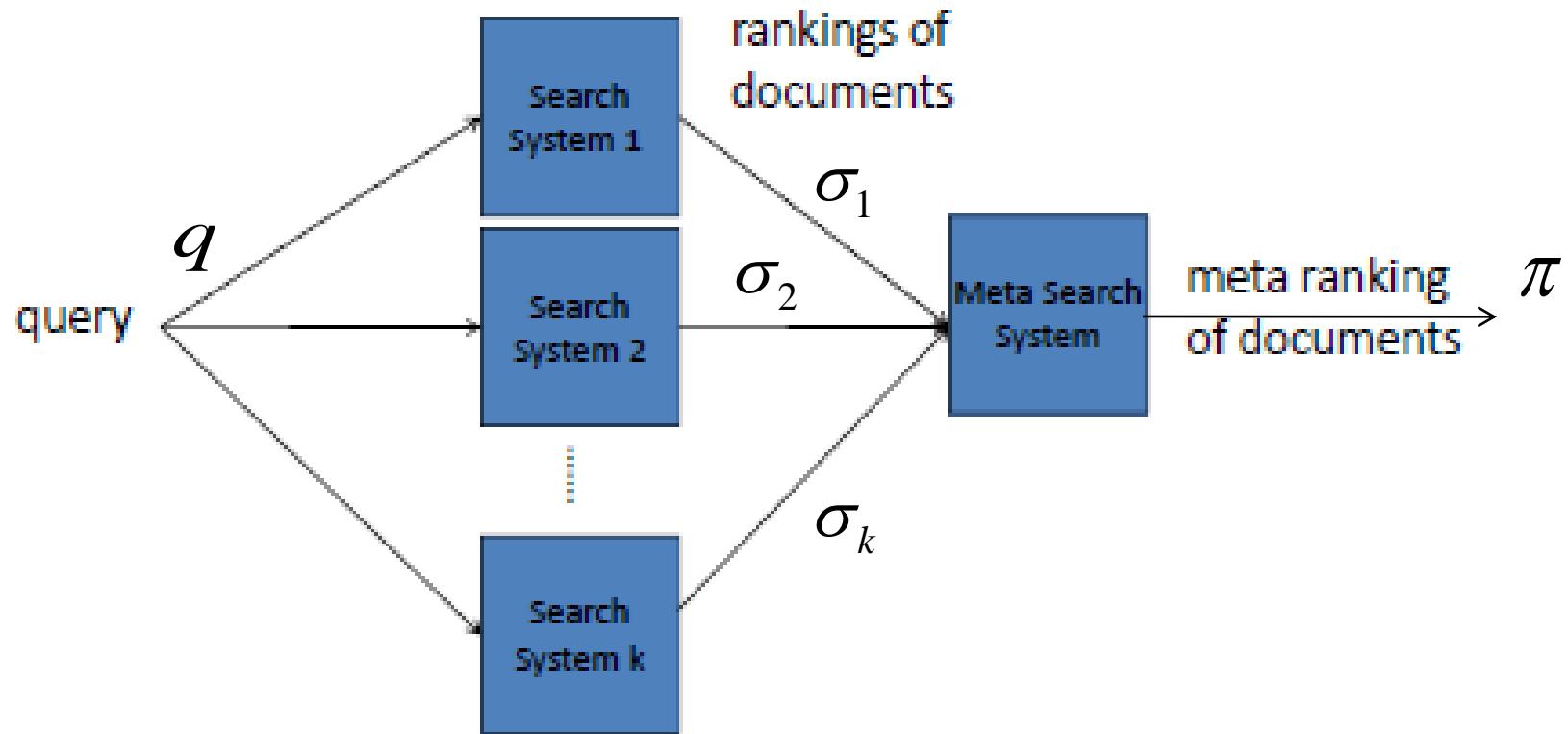
Evaluation Results

- Pairwise approach and listwise approach perform better than pointwise approach
- LabmdaMART performs best in Yahoo Learning to rank Challenge
- No significant difference among pairwise and listwise methods

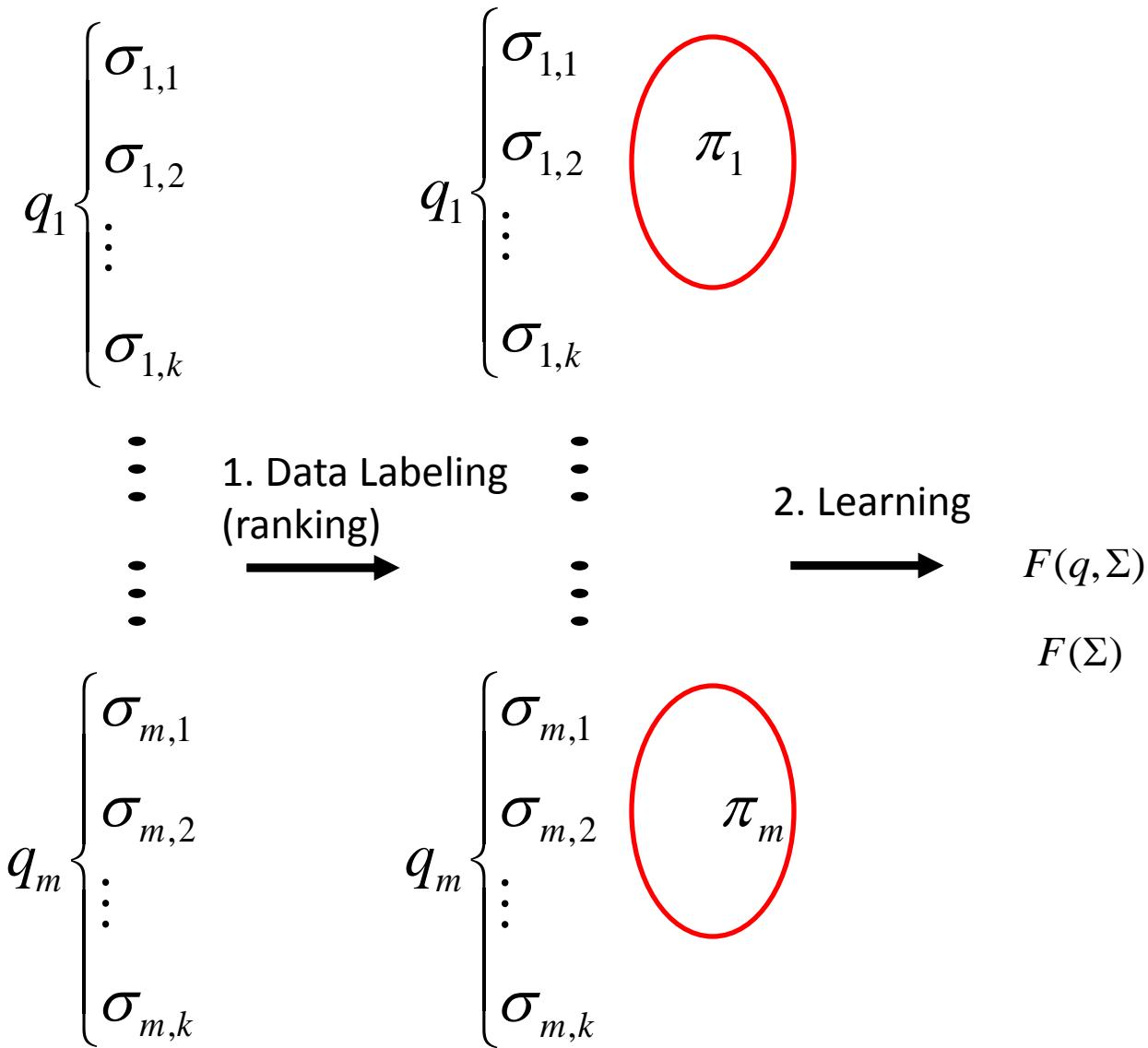
3. Learning for Ranking Aggregation

Ranking Problem

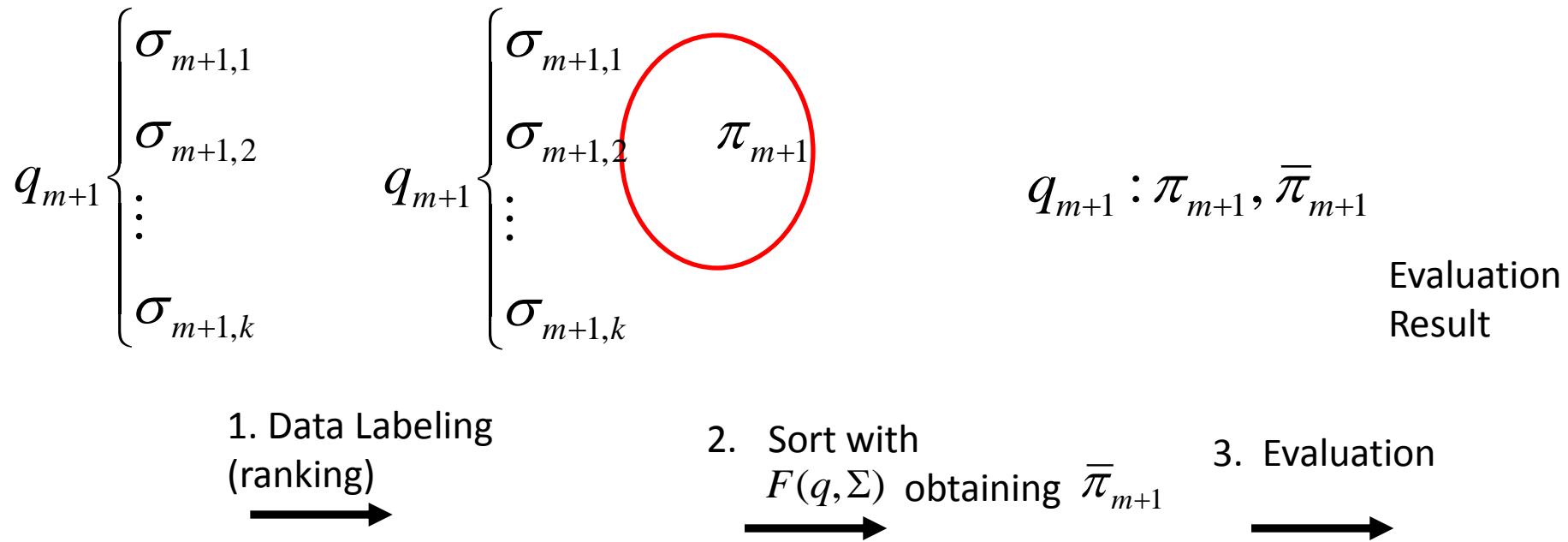
Example = Meta Search



Training Process



Testing Process



Learning Methods

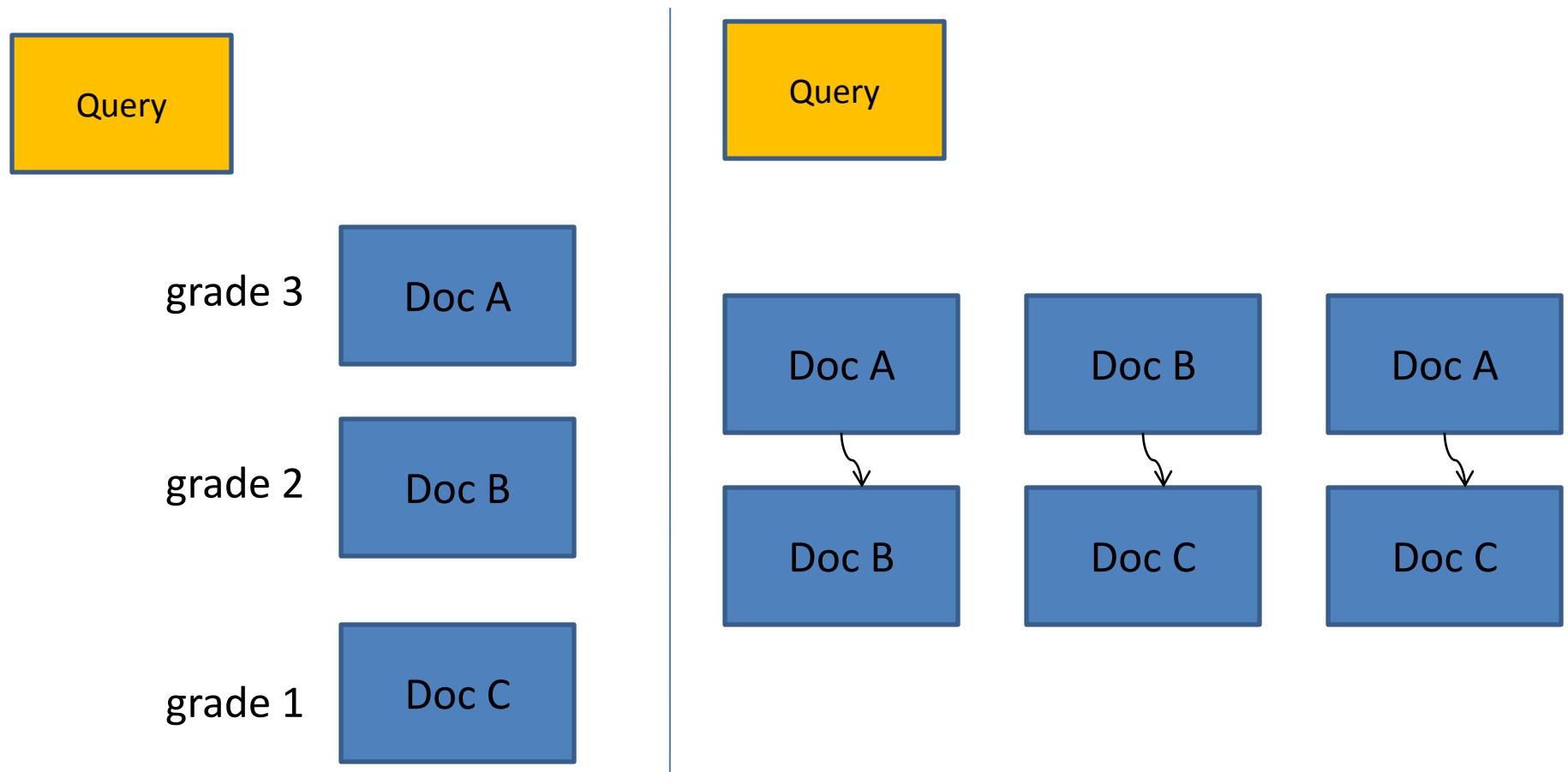
- Unsupervised Learning
 - Borda Count [Aslam & Montague 2001]
 - Markov Chain [Dwork et al 2001]
- Supervised learning
 - CRanking [Lebanon & Lafferty 2002]

4. Learning to rank Methods

Ranking SVM

Pairwise Classification

- Converting document list to document pairs

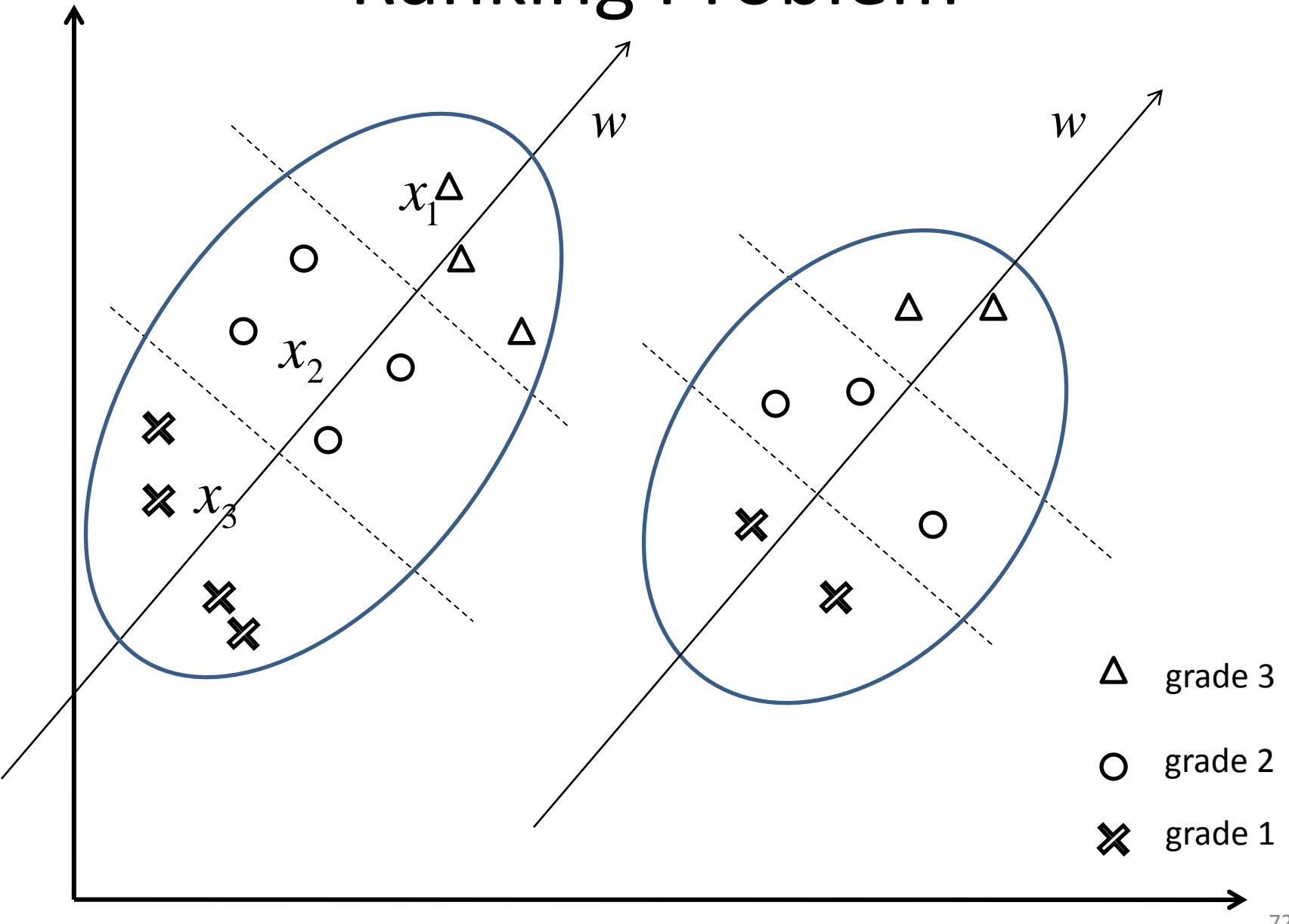


Transforming Ranking to Pairwise Classification

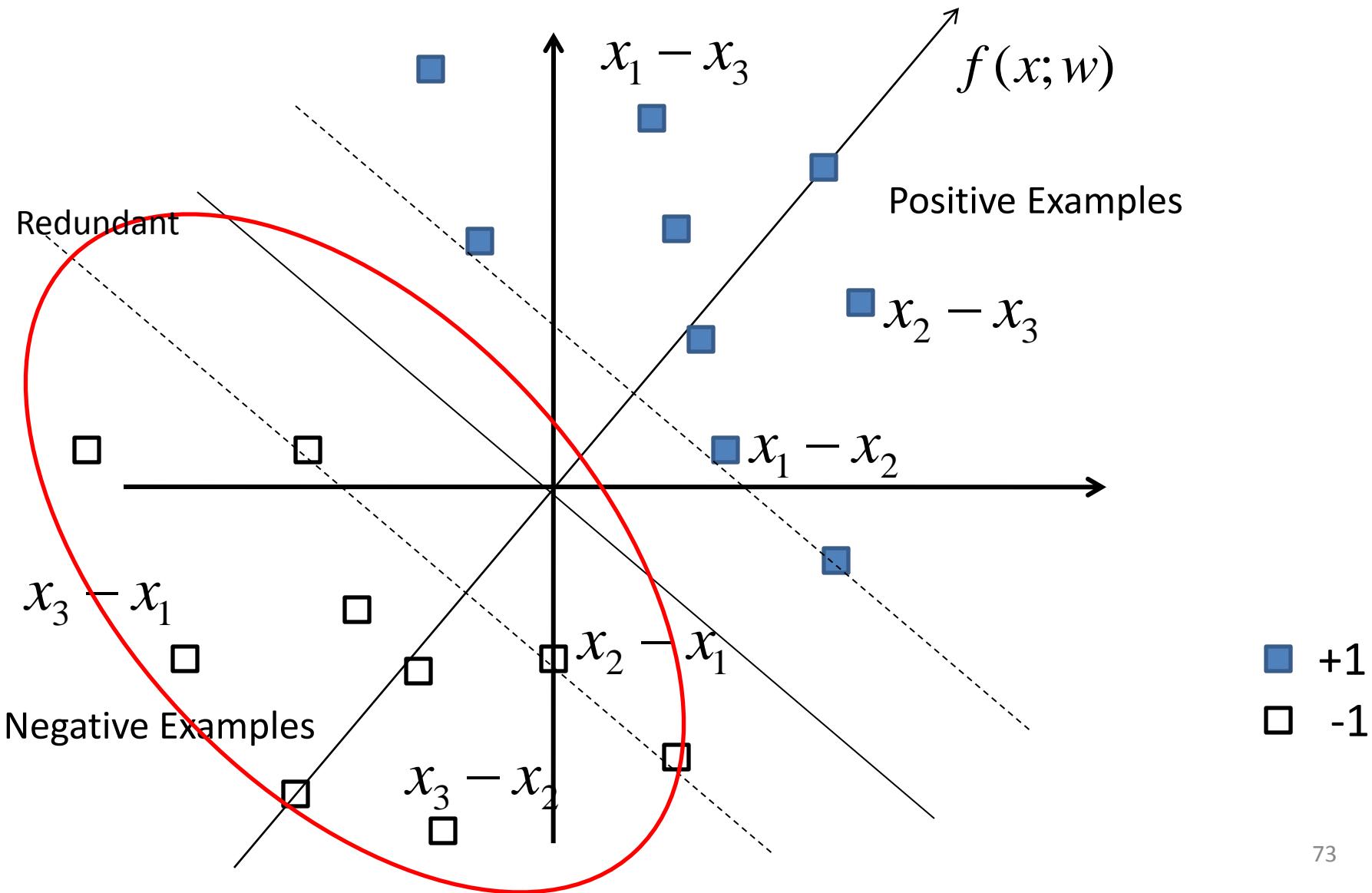
- Input space: X
- Ranking function $f : X \rightarrow R$
- Ranking: $x_i \succ x_j \Leftrightarrow f(x_i; w) > f(x_j; w)$
- Linear ranking function: $f(x; w) = \langle w, x \rangle$
 $\langle w, x_i - x_j \rangle > 0 \Leftrightarrow f(x_i; w) > f(x_j; w)$
- Transforming to pairwise classification:

$$(x_i - x_j, z), \quad y = \begin{cases} +1 & x_i \succ x_j \\ -1 & x_j \succ x_i \end{cases}$$

Ranking Problem



Transformed Pairwise Classification Problem



Ranking SVM

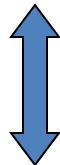
- Pairwise classification on differences of feature vectors
- Corresponding positive and negative examples
- Negative examples are redundant and can be discarded
- Hyper plane passes the origin
- Soft margin and kernel can be used
- *Ranking SVM* = pairwise classification SVM

Learning of Ranking SVM

$$\min_{w, \xi} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^N \xi_i$$

$$y_i \langle w, x_i^{(1)} - x_i^{(2)} \rangle \geq 1 - \xi_i \quad i = 1, \dots, N$$

$$\xi_i \geq 0$$



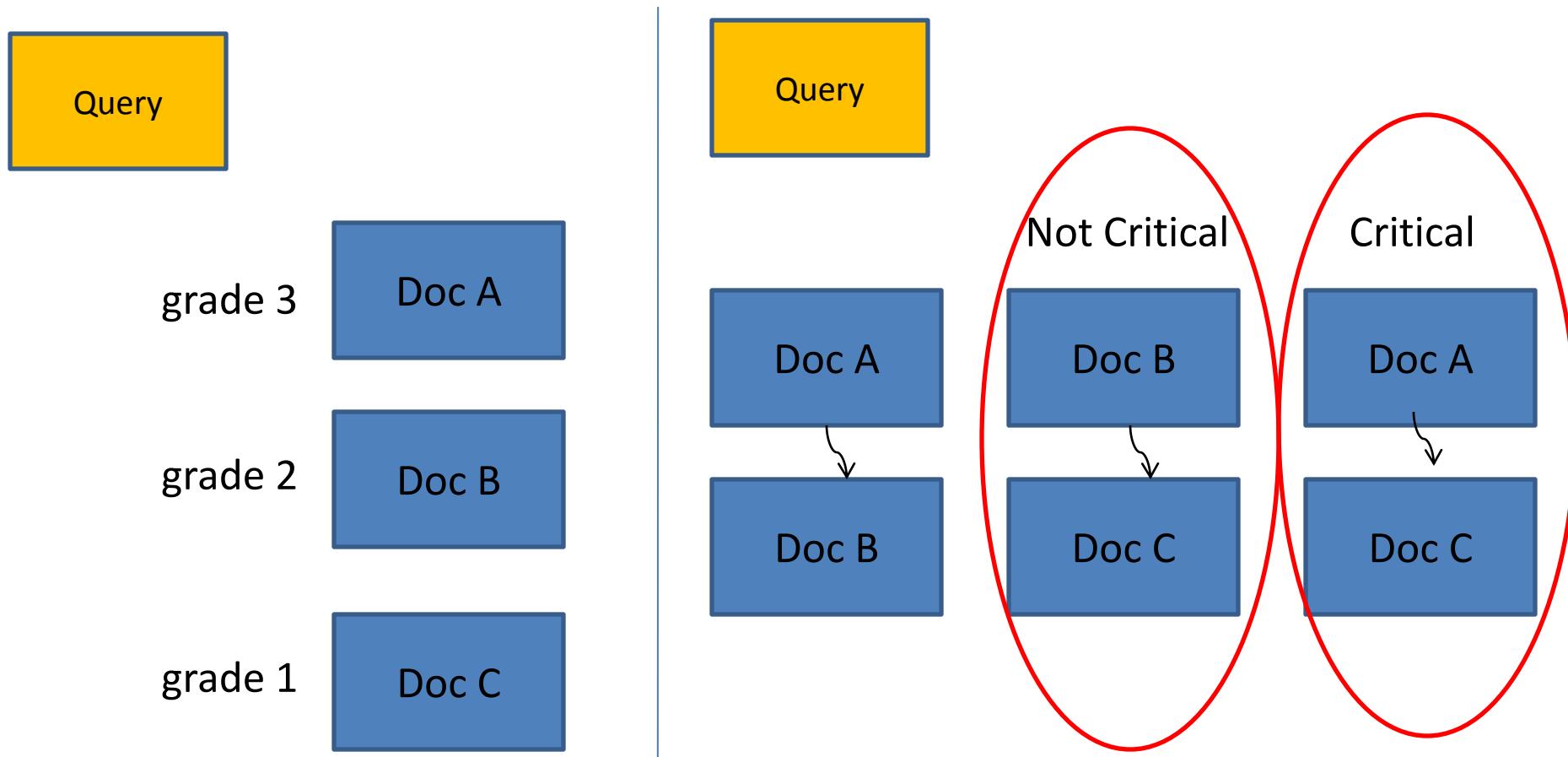
$$\min_w \sum_{i=1}^l \left[1 - y_i \langle w, x_i^{(1)} - x_i^{(2)} \rangle \right]_+ + \lambda \| w \|^2$$

$$[s]_+ = \max(0, s) \quad \lambda = \frac{1}{2C}$$

IR SVM

Cost-sensitive Pairwise Classification

- Converting to document pairs



Problems with Ranking SVM

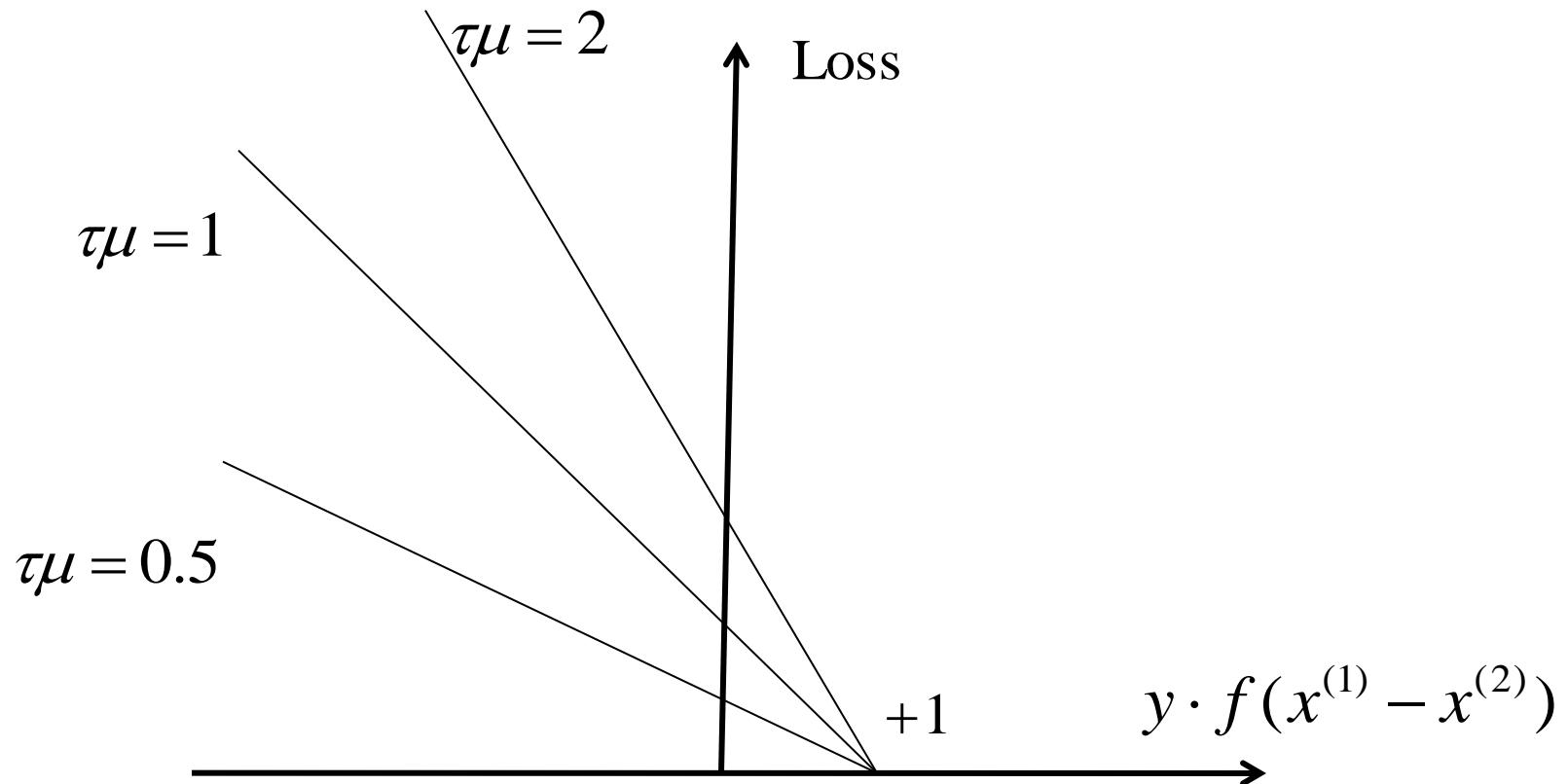
- Not sufficient emphasis on correct ranking on top grades: 3, 2, 1
 - ranking 1: 2 3 2 1 1 1 1
 - ranking 2: 3 2 1 2 1 1 1
 - ranking 2 should be better than ranking 1
 - Ranking SVM views them as the same
- Numbers of pairs vary according to queries
 - q1: 3 2 2 1 1 1 1
 - q2: 3 3 2 2 2 1 1 1 1 1
 - number of pairs for q1 : $2*(2-2) + 4*(3-1) + 8*(2-1) = 14$
 - number of pairs for q2: $6*(3-2) + 10*(3-1) + 15*(2-1) = 31$
 - Ranking SVM is biased toward q2

IR SVM

- Solving the two problems of Ranking SVM
- Higher weight on important grade pairs $\tau_{k(i)}$
- Normalization weight on pairs in query $\mu_{q(i)}$
- IR SVM = Ranking SVM using modified hinge loss

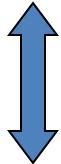
Modified Hinge Loss function

$$\min_w \sum_{i=1}^l \tau_{k(i)} \mu_{q(i)} \left[1 - y_i \langle w, x_i^{(1)} - x_i^{(2)} \rangle \right]_+ + \lambda \|w\|^2$$



Learning of IR SVM

$$\min_w \sum_{i=1}^l \tau_{k(i)} \mu_{q(i)} \left[1 - y_i \langle w, x_i^{(1)} - x_i^{(2)} \rangle \right]_+ + \lambda \|w\|^2$$



$$\min_{w,\xi} \frac{1}{2} \|w\|^2 + \sum_{i=1}^l C_i \xi_i$$

$$y_i \langle w, x_i^{(1)} - x_i^{(2)} \rangle \geq 1 - \xi_i \quad i = 1, \dots, l$$

$$\xi_i \geq 0$$

$$C_i = \frac{\tau_{k(i)} \mu_{q(i)}}{2\lambda}$$

ListNet

Plackett-Luce Model (Permutation Probability)

- Probability of permutation π is defined as

$$P(\pi) = \prod_{i=1}^n \frac{s_{\pi(i)}}{\sum_{j=i}^n s_{\pi(j)}}$$

- Example:

$$P(ABC) = \frac{s_A}{s_A + s_B + s_C} \cdot \frac{s_B}{s_B + s_C} \cdot \frac{s_C}{s_C}$$

$P(A \text{ ranked No.1})$

$P(B \text{ ranked No.2} | A \text{ ranked No.1})$

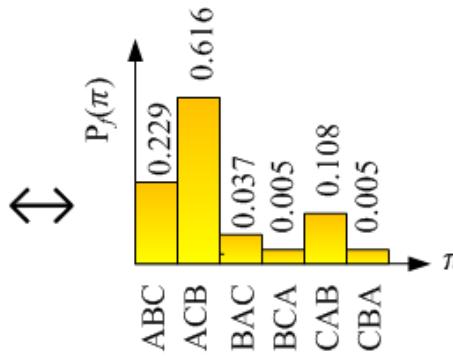
$P(C \text{ ranked No.3} | A \text{ ranked No.1}, B \text{ ranked No.2})$

Properties of Plackett-Luce Model

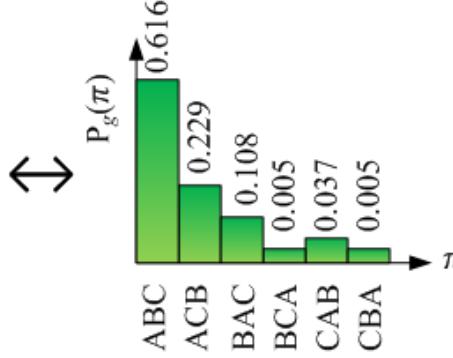
- Objects: ABC
- Scores: $s_A = 5, s_B = 3, s_C = 1$
- Property 1: $P(ABC)$ is largest, $P(CBA)$ is smallest
- Property 2: swap B and C in ABC, $P(ABC) > P(ACB)$

KL Divergence between Permutation Probability Distributions

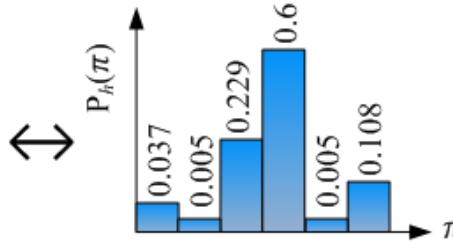
$f: f(A) = 3, f(B)=0, f(C)=1;$
Ranking by $f: \text{ABC}$



$g: g(A) = 6, g(B)=4, g(C)=3;$
Ranking by $g: \text{ABC}$



$h: h(A) = 4, h(B)=6, h(C)=3;$
Ranking by $h: \text{ACB}$



Plackett-Luce Model (Top-k Probability)

- Computation of permutation probabilities is intractable
- Top- k probability
 - Defining Top- k subgroup $G(o_1 \dots o_k)$ containing all permutations whose top- k objects are o_1, \dots, o_k
 - $P(G(o_1 \dots o_k)) = \prod_{i=1}^k \frac{s_{o_i}}{\sum_{j=i}^n s_{o_j}}$
 - Time complexity of computation : from $n!$ to $n!/(n-k)!$
- Example: $P(G(A)) = \frac{s_A}{s_A + s_B + s_C}$

ListNet

- Parameterized Plackett-Luce Model

$$s = \exp(f(x; w))$$

$$P(G(x_1 \cdots x_k)) = \prod_{i=1}^k \frac{s_{x_i}}{\sum_{j=i}^n s_{x_j}}$$

- Ranking Model: $f(x; w)$ = Neural Net

ListNet (cont')

- Loss function = KL-divergence between two Top- k probability distributions from ground truth and ranking model

$$L(w) = - \sum_{\pi \in \Omega^k} \left(\prod_{i=1}^k \frac{\exp(y_i))}{\sum_{j=i}^n \exp(y_j)} \right) \log \left(\prod_{i=1}^k \frac{\exp(f(x_i; w))}{\sum_{j=i}^n \exp(f(x_j; w))} \right)$$

- Algorithm = Gradient Descent

AdaRank

Listwise Loss

q_1	$x_{1,1}$	$\pi_{1,1}$	$y_{1,1}$
	$x_{1,2}$	$\pi_{1,2}$	$y_{1,2}$
	\vdots		
	x_{1,n_1}	π_{1,n_1}	y_{1,n_1}
	\vdots		
	\vdots		
q_m	$x_{m,1}$	$\pi_{m,1}$	$y_{m,1}$
	$x_{m,2}$	$\pi_{m,2}$	$y_{m,2}$
	\vdots		
	x_{m,n_m}	π_{m,n_m}	y_{m,n_m}

$$\max_{f \in \mathcal{F}} \sum_{i=1}^m E(\pi(q_i, \mathbf{d}_i, f), \mathbf{y}_i)$$

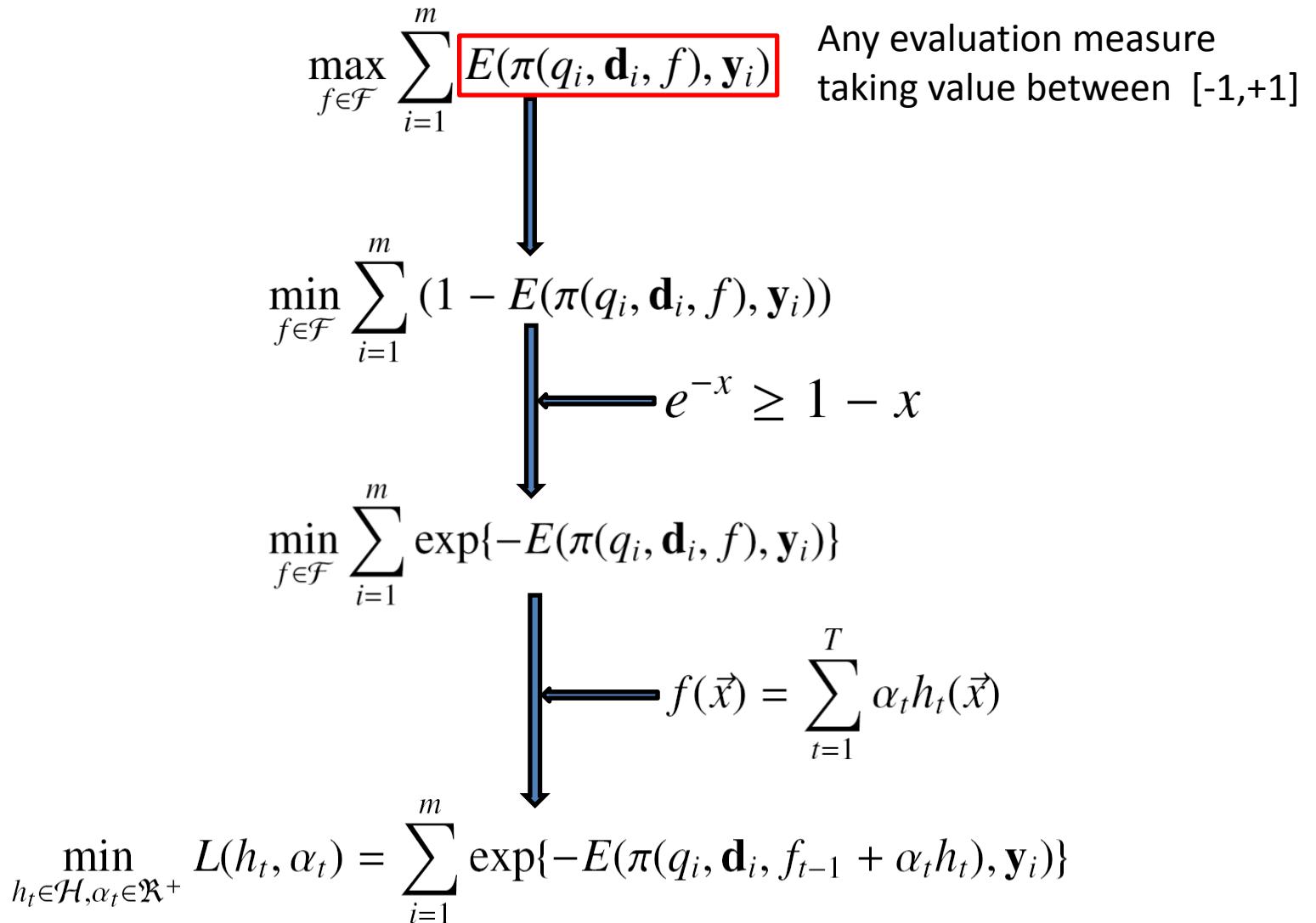


$$\min_{f \in \mathcal{F}} \sum_{i=1}^m (1 - E(\pi(q_i, \mathbf{d}_i, f), \mathbf{y}_i))$$

AdaRank

- Optimizing exponential loss function
- Algorithm: AdaBoost-like algorithm for ranking

Loss Function of AdaRank



AdaRank Algorithm

Input: $S = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^m$

Parameter: T (number of iterations)

Evaluation measure: E

Initialize $P_1(i) = 1/m$

For $t = 1, \dots, T$

- Create weak ranker h_t with weighted distribution P_t on training data S

- Choose α_t

$$\alpha_t = \frac{1}{2} \cdot \ln \frac{\sum_{i=1}^m P_t(i)(1 + E(\pi_i, \mathbf{y}_i))}{\sum_{i=1}^m P_t(i)(1 - E(\pi_i, \mathbf{y}_i))}$$

- where $\pi_i = \text{sort}_{h_t}(\mathbf{x}_i)$

- Create f_t

$$f_t(x) = \sum_{k=1}^t \alpha_k h_k(x)$$

- Update P_{t+1}

$$P_{t+1}(i) = \frac{\exp(-E(\pi_i, \mathbf{y}_i))}{\sum_{j=1}^m \exp(-E(\pi_j, \mathbf{y}_j))}$$

- where $\pi_i = \text{sort}_{f_t}(\mathbf{x}_i)$

End For

Output: the ranking model $f(x) = f_T(x)$

SVM MAP

Scoring Function

$$S(\mathbf{x}_i, \pi_i) = \langle w, \sigma(\mathbf{x}_i, \pi_i) \rangle,$$

$$\sigma(\mathbf{x}_i, \pi_i) = \frac{2}{n_i(n_i - 1)} \sum_{k,l:k < l} z_{kl}(x_{ik} - x_{il}),$$

$z_{kl} = +1$ if $\pi_i(k) < \pi_i(l)$ (x_{ik} is ranked ahead of x_{il} in π_i), and -1 , otherwise.

Scoring Function (cont')

- Ranking model is linear model
- The scoring function gives
 - highest score to the perfect ranking
 - lower scores to imperfect rankings

Example of Scoring Function

Objects: A, B, C

$$f_A = \langle w, x_A \rangle, f_B = \langle w, x_B \rangle, f_C = \langle w, x_C \rangle$$

Suppose $f_A > f_B > f_C$

For example:

Permutation1: ABC

Permutation2: ACB

$$S_{ABC} = \frac{1}{6} (w, ((x_A - x_B) + (x_B - x_C) + (x_A - x_C)))$$

$$S_{ACB} = \frac{1}{6} (w, ((x_A - x_C) + (x_C - x_B) + (x_A - x_B)))$$

$$S_{ABC} > S_{ACB}$$

Loss Function

$$\sum_{i=1}^m \left[\max_{\pi_i^* \in \Pi_i^*, \pi_i \in \Pi_i \setminus \Pi_i^*} \left((E(\pi_i^*, y_i) - E(\pi_i, y_i)) - (S(x_i, \pi_i^*) - S(x_i, \pi_i)) \right)_+ \right],$$

Difference between
Evaluation Measures

Difference between
Scoring Functions

π_i^* Perfect ranking

π_i Imperfect ranking

SVM MAP

$$\begin{aligned} & \min_{w; \xi \geq 0} \frac{1}{2} \|w\|^2 + \frac{C}{m} \sum_{i=1}^m \xi_i \\ \text{s.t. } & \forall i, \forall \pi_i^* \in \Pi_i^*, \forall \pi_i \in \Pi_i \setminus \Pi_i^* : \\ & S(\mathbf{x}_i, \pi_i^*) - S(\mathbf{x}_i, \pi_i) \geq E(\pi_i^*, y_i) - E(\pi_i, y_i) - \xi_i, \end{aligned}$$

$$\sum_{i=1}^m \left[\max_{\pi_i^* \in \Pi_i^*; \pi_i \in \Pi_i \setminus \Pi_i^*} (E(\pi_i^*, y_i) - E(\pi_i, y_i)) - (S(\mathbf{x}_i, \pi_i^*) - S(\mathbf{x}_i, \pi_i)) \right]_+ + \lambda \|w\|^2.$$

Borda Count

Ranking Function

- Sum of number of objects ranked behind

$$S_D = F(\Sigma) = \sum_{i=1}^k S_i$$

$$S_i \equiv \begin{pmatrix} s_{i,1} \\ \vdots \\ s_{i,j} \\ \vdots \\ s_{i,n} \end{pmatrix}$$

$$s_{i,j} = n - \sigma_i(j),$$

Example

- Three basic rankings

$$\begin{pmatrix} \sigma_1 \\ A \\ B \\ C \end{pmatrix}$$

$$\begin{pmatrix} \sigma_2 \\ A \\ C \\ B \end{pmatrix}$$

$$\begin{pmatrix} \sigma_3 \\ B \\ A \\ C \end{pmatrix}$$

- Ranking scores

$$s_D = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix} = \begin{pmatrix} 5 \\ 3 \\ 1 \end{pmatrix}$$

- Ranking

$$\begin{pmatrix} \pi \\ A \\ B \\ C \end{pmatrix}$$

5. Learning to rank Applications

Learning to rank Applications

- Web Search
- Recommender System
- Key Phrase Extraction
- Query Dependent Summarization
- Machine Translation

Recommender System (Collaborative Filtering)

- Problem formulation
 - Input: users' ratings on some items
 - Output: users' ratings on other items
 - Assumption: users sharing same ratings on input items tend to agree on new items
- Solutions
 - Classification
 - Ordinal Regression
 - Learning to Rank

Recommender System

	Item1	Item2	Item3	...	
User1	5	4			
User2	1		2		2
...		?	?	?	
UserM	4	3			

Recommender System Using RankBoost

- Ranking items according to users
- Justification: users tend to rate on different scales
- Method: RankBoost
- Result: RankBoost > Nearest Neighbor

Key Phrase Extraction

- Problem formulation
 - Input: document
 - Output: keyphrases of document
 - Two steps: phrase extraction and keyphrase identification
- Traditional approach
 - Classification: keyphrase vs non-keyphrase

Key Phrase Extraction Using Ranking SVM

- Ranking of phrases as keyphrases
- Justification: keyphrase or non-keyphrase is relative
- Method: Ranking SVM
- Result: Ranking SVM > SVM

6. Theory of Learning to Rank

Statistical Learning Formulation

- Input space: \mathcal{X} : lists of feature vectors
- Output space \mathcal{Y} : lists of grades
- Input \mathbf{x} : list of feature vectors
- Output \mathbf{y} : list of grades
- Training data: $(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{x}_m, \mathbf{y}_m)$.
- Global ranking model: $F(\mathbf{x}) = [f(x_1), f(x_2), \dots, f(x_n)]$
- Loss function: $L(F(\mathbf{x}), \mathbf{y})$.

Statistical Learning Formulation

- Risk function: $R(F) = \int_{\mathcal{X} \times \mathcal{Y}} L(F(\mathbf{x}), y) dP(\mathbf{x}, y).$
- Empirical risk: $\hat{R}(F) = \frac{1}{m} \sum_{i=1}^m L(F(\mathbf{x}_i), y_i).$
- Surrogate loss function: $L'(F(\mathbf{x}), y).$

Loss Functions

- True loss function

$$L(F(\mathbf{x}), \mathbf{y}) = 1 - NDCCG$$

$$L(F(\mathbf{x}), \mathbf{y}) = 1 - MAP.$$

- Difficult to optimize
 - Use of sorting
 - Non continuous
- Using surrogate loss functions

Loss Functions

- Pointwise loss (squared): $L'(F(\mathbf{x}), \mathbf{y}) = \sum_{i=1}^n (f(x_i) - y_i)^2.$
- Pairwise loss (hinge, exponential, logistic)

$$L'(F(\mathbf{x}), \mathbf{y}) = \sum_{i=1}^{n-1} \sum_{j=i+1}^n [1 - \text{sign}(y_i - y_j)(f(x_i) - f(x_j))]_+, \text{ when } y_i \neq y_j,$$

$$L'(F(\mathbf{x}), \mathbf{y}) = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \exp(-\text{sign}(y_i - y_j)(f(x_i) - f(x_j))), \text{ when } y_i \neq y_j.$$

$$L'(F(\mathbf{x}), \mathbf{y}) = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \log(1 + \exp(-\text{sign}(y_i - y_j)(f(x_i) - f(x_j)))), \text{ when } y_i \neq y_j.$$

Relations between Surrogate Loss and True Loss

- Pointwise loss

$$1 - NDCG \leq \frac{1}{G_{max}} \left(2 \sum_{i=1}^n D(\pi(i))^2 \right)^{1/2} L'(F(\mathbf{x}), \mathbf{y})^{1/2},$$

- Pairwise loss

$$1 - NDCG \leq \frac{\max_i(G(i)D(\pi(i)))}{G_{max}} L'(F(\mathbf{x}), \mathbf{y}),$$

- Listwise loss

$$1 - NDCG \leq \frac{\max_i(G(i)D(\pi(i)))}{\ln 2 \cdot G_{max}} L'(F(\mathbf{x}), \mathbf{y}),$$

Theoretical Analysis

- Generalization ability
- Consistency

7. Ongoing and Future Work

Future and Ongoing Work

- Training data creation
- Semi-supervised learning and active learning
- Feature learning
- Scalable and efficient training
- Domain adaptation
- Ranking by ensemble learning
- Global ranking
- Ranking of objects in graph

Summary

Outline of Tutorial

1. Learning to Rank
2. Learning for Ranking Creation
3. Learning for Ranking Aggregation
4. Methods of Learning to Rank
5. Applications of Learning to Rank
6. Theory of Learning to Rank
7. Ongoing and Future Work

Contact: hangli@microsoft.com

Large-Scale, Open Domain Semantic Mining: Basic Techniques

Shuming Shi
Microsoft Research Asia
August 2011

Outline

- Overview
- Semantic class mining
- Semantic hierarchy construction
- Mining attribute names and values
- General relation extraction
- Demo
- Summary

Semantic Mining: Introduction

- (Semi-)Automatically obtaining semantic knowledge
 - Semantic knowledge: Entities, concepts, relations
 - Similarity(significantly, substantially, 0.9)
 - Synonym(China, People's Republic of China)
 - IsA(pear, fruit)
 - Peer(Beijing, Shanghai, Guangzhou...)
 - InClass(Beijing, C1)
 - Attribute(Capital, China, Beijing)
 - BornIn(Barack Obama, 1961)
 - DefeatedIn(Dallas Mavericks, Miami Heat, 2011 NBA Finals)
 - Data sources:
 - Web documents, query logs, web search results
 - Existing dictionaries & knowledge-bases

Semantic Mining: Introduction (cont.)

- Motivation
 - Build “smarter” computer systems with the semantic knowledge-base
 - Better fulfill the information needs of end users
 - Better web search
 - Better QA
 - Better machine translation
 - ...

Outline

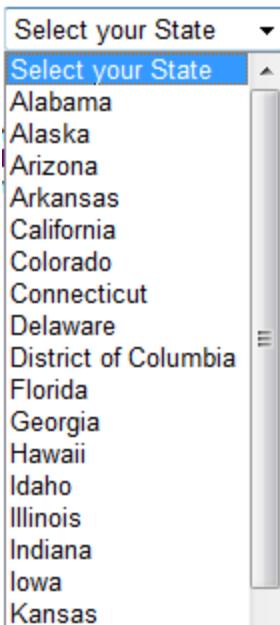
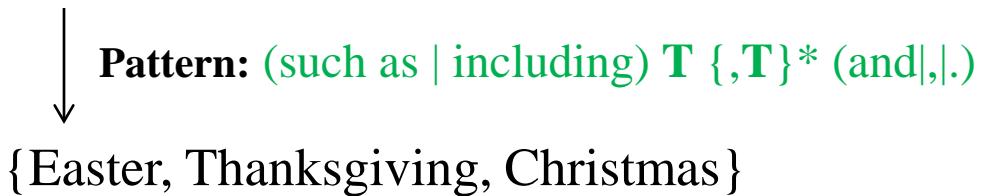
- Overview
- **Semantic class mining**
 - Semantic hierarchy construction
 - Mining attribute names and values
 - General relation extraction
 - Demo
 - Summary

Semantic Class Mining

- Goal
 - Discover peer terms (or coordinate terms)
 - Sample: {C++, C#, Java, PHP, Perl, ...}
- Main techniques
 - First-order co-occurrences
 - Standard co-occurrences
 - Patterns: Special first-order co-occurrences
 - Second-order co-occurrences
 - Distributional similarity

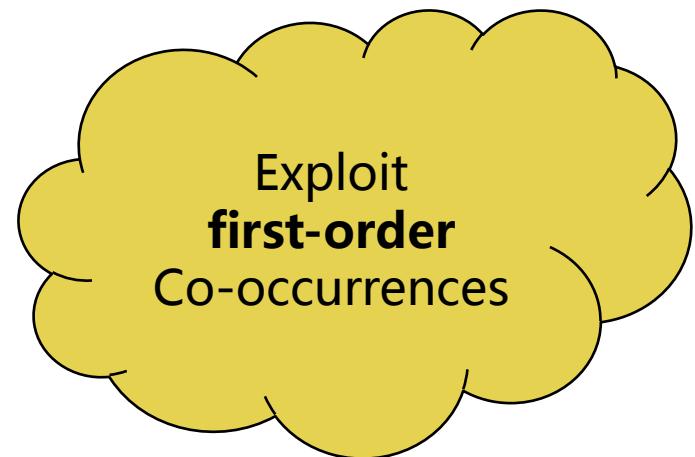
Pattern-Based (PB)

Hours may vary on holidays, such as Easter, Thanksgiving and Christmas.



Pattern:

```
<select>  
<option> T ... <option> T  
</select>
```



→ { Alabama, Alaska, Arizona... }

PB Implementation

- RASC mining
 - Employ predefined patterns to extract Raw Semantic Classes (RASCs)

Type	Pattern
Lexical	T {, T}*{,} (and or) {other} T
	(such as including) T {,T}* (and ,.)
	T, T, T {,T}*{,}
Tag	 T ... T
	 T ... T
	<select> <option> T ... <option> T </select>
	<table> <tr> <td> T </td> ... <td> T </td> </tr> ... </table>
	Other Html-tag repeat patterns

PB Implementation

- Compute Term Similarity
 - Based on the RASCs containing both terms

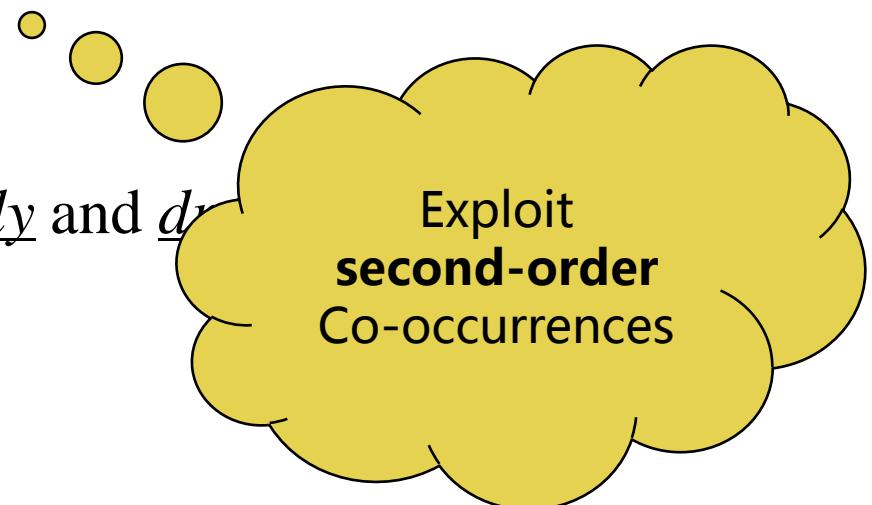
$$Sim(a, b) = \sum_{i=1}^m \log\left(1 + \sum_{j=1}^{k_i} w(P(C_{i,j}))\right) \quad (\text{Zhang et al., ACL'09})$$

$$Sim^*(a, b) = Sim(a, b) \cdot \sqrt{IDF(a) \cdot IDF(b)}$$

$$IDF(a) = \log\left(1 + \frac{N}{N(a)}\right)$$

Distributional Similarity (DS)

- Distributional hypothesis (Harris, 1985): Terms occurring in analogous (lexical or syntactic) contexts tend to be similar
- Contexts shared by Easter and Christmas
 - the date _ is celebrated
 - | _ is a religious festival
 - history of the _ festival
 - ...
- Contexts shared by significantly and dr
 - is _ improved by
 - unlikely to _ alter the
 - can _ increase health risks
 - ...



DS Implementation

- Define context
 - Syntactic context, lexical context...
- Represent each term by a feature vector
 - Feature: A context in which the term appears
 - Feature value: “Weight” of the context w.r.t. the term
- Compute term similarity
 - Term similarity = similarity between corresponding feature vectors

DS Implementation

Contexts	Text window (window size: 2, 4)
Feature value	Syntactic
Similarity measure	PMI
	Cosine, Jaccard

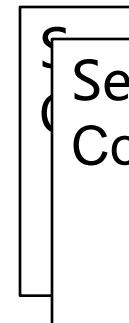
DS approach implemented in the study

Pointwise mutual information:

$$f_{w,c} = \text{PMI}_{w,c} = \log \frac{F(w, c) \cdot F(*, *)}{F(w, *) \cdot F(*, c)}$$

$$\text{Cosine}(\vec{x}, \vec{y}) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2} \cdot \sqrt{\sum_i y_i^2}}$$

$$\text{Jaccard}(\vec{x}, \vec{y}) = \frac{\sum_i \min(x_i, y_i)}{\sum_i x_i + \sum_i y_i - \sum_i \min(x_i, y_i)}$$



Compare DS and PB with Set Expansion

(Shi et al., COLING'2010)

- Set Expansion: Problem statement
 - Given a list of seed terms in a semantic class
 - $Q = \{s_1, s_2, \dots, s_k\}$ (e.g. $Q = \{Lent, Epiphany\}$)
 - To find other members of the class $R(t, s_i)$: the rank of term t among the neighbors of s_i
 - E.g., $\{Advent, Easter, Christmas\}$

- Set Expansion with a similarity graph G
 - Select the terms most similar to the seeds

$$f(t, Q) = \sum_{i=1}^k w_i \cdot \text{Sim}(t, s_i)$$

$$\text{Sim}(t, s_i) = \frac{1}{\log(\lambda + r(t, s_i))}$$

Compare and Combine PB & DS (cont.)

- Corpus: ClueWeb (500 million English pages)
- Five term categories: **proper nouns, common nouns, verbs, adjectives, adverbs**
- Key observations: PB performs better for proper nouns; DS has better performance for other term categories

Samples (Query: significantly)

1	significantly	187.419
2	and	54.8759
3	slightly	23.4412
4	but	21.8044
5	moderately	21.7083
6	english	20.4911
7	seriously	20.4479
8	yiddish	20.2321
9	hebrew	19.7871
10	too	19.6313
11	kigezi	18.4164
12	bunyoro	17.8679
13	specifically	17.8268
14	also	17.4519
15	mbale	17.3605
16	especially	17.2895
17	rich americans	17.1207
18	surely	16.7604
19	sharply	16.5638
20	it	15.6475

PB results

1	significantly	0.121576
2	substantially	0.0162357
3	considerably	0.0154982
4	greatly	0.0138213
5	dramatically	0.013429
6	slightly	0.0100923
7	drastically	0.0089119
8	somewhat	0.00800886
9	vastly	0.0074269
10	steadily	0.00731532
11	severely	0.00688791
12	importantly	0.00640118
13	remarkably	0.0061907
14	inherently	0.00606039
15	comparatively	0.00604854
16	strongly	0.0060448
17	consistently	0.00603508
18	sufficiently	0.00602135
19	rapidly	0.00601235
20	gradually	0.00590148

DS results

Samples (Query: Apple)

```
D:\Projects\NeedleSeek\bin\TermGraphClien... X
1      apple    1741.13
2      microsoft   639.909
3      ibm      617.503
4      sony     613.111
5      dell      601.909
6      hp       597.473
7      toshiba   546.464
8      orange     537.578
9      samsung   528.885
10     compaq     490.275
11     canon      476.098
12     cherry     472.247
13     pear       470.911
14     panasonic   467.727
15     peach      460.441
16     pineapple   444.158
17     intel      434.583
18     acer       433.825
19     lemon      424.788
20     strawberry  423.942
```

PB results

```
D:\Projects\NeedleSeek\bin\TermGraphClien... X
1      apple    0.0808821
2      microsoft   0.00336825
3      the government  0.00237455
4      the company   0.00223547
5      google     0.00212872
6      sony       0.00193015
7      ibm        0.00185744
8      obama      0.00163117
9      dell        0.00161188
10     nintendo    0.00135578
11     bush        0.00129623
12     hp         0.00127199
13     banana      0.00126387
14     intel       0.00124417
15     someone     0.00123563
16     mccain      0.00119849
17     congress     0.00114992
18     israel      0.00111276
19     the team     0.00110751
20     adobe       0.00108731
```

DS results

Explain by Frequency

- Normalized frequency (F_{norm}) of term t
$$\frac{\text{Frequency in the RASCs}}{\text{Frequency in the sentences of the original documents}}$$
- Mean normalized frequency (MNF) of a query set S

$$MNF(S) = \frac{\sum_{t \in S} F_{norm}(t)}{|S|}$$

Seed Categories	Terms	MNF
Proper nouns	40	0.2333
Common nouns	40	0.0716
Verbs	40	0.0099
Adjectives	40	0.0126
Adverbs	40	0.0053

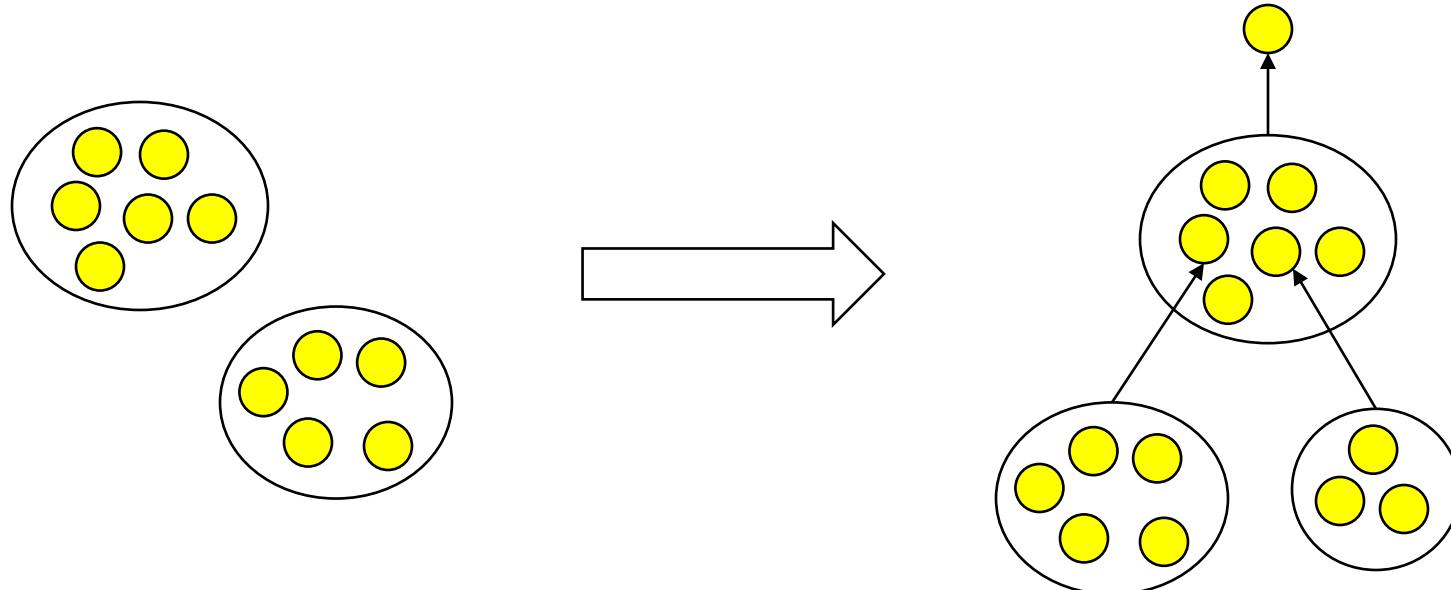
Related Papers

- Harris, 1985 (in The Philosophy of Linguistics)
Distributional Structure
- Pantel & Lin, SIGKDD'2002
Discovering Word Senses from Text
- Etzioni et al., WWW'2004
Web-Scale Information Extraction in KnowItAll
- Wang & Cohen, ICDM'2008
Iterative Set Expansion of Named Entities Using the Web
- Pantel, EMNLP'2009
Web-Scale Distributional Similarity and Entity Set Expansion
- Agirre et al., NAACL'2009
A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches
- Shi et al., COLING'2010
Corpus-based Semantic Class Mining: Distributional vs. Pattern-Based Approaches

Outline

- Overview
- Semantic class mining
- **Semantic hierarchy construction**
- Mining attribute names and values
- General relation extraction
- Demo
- Summary

Semantic Hierarchy Construction

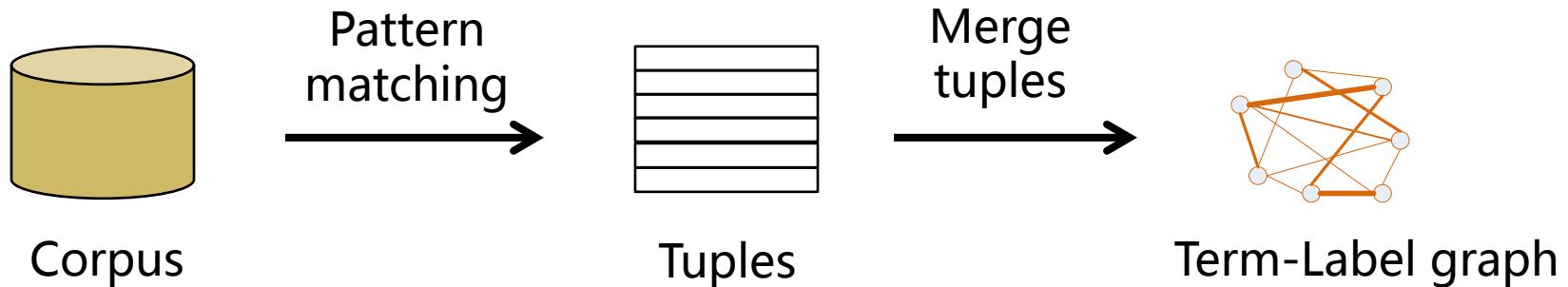


Semantic hierarchy construction

- Major subtasks
 - Assign category labels (hyponyms) to terms
 - Beijing → city, capital...
 - Apple → company, fruit...
 - Red → color...
 - Canon EOS 400D → digital camera, product...
 - Assign category labels to semantic classes
 - {Beijing, Shanghai, Dalian...} → cities, Chinese cities...
 - {Microsoft, IBM, Apple...} → companies, manufacturers...
 - Build the hierarchy

Subtask: Term→Label

- Approach: Pattern matching + counting



Tuple:
<term, label, pattern, source, weight>

<pear, fruit, P1, S1, 1.0>
<pear, shape, P2, S2, 1.0>
<pear, fruit, P3, S3, 1.0>
<New York, city, P1, S4, 1.0>
<New York, office, P2, S6, 1.0>
<New York, state, P4, S7, 1.0>

... ...

Subtask: Term→Label (cont.)

- Pattern matching
 - Manually designed or automatically generated patterns
 - Text patterns or HTML tables

Label	Label	Label
Term	Term	Term
Term	Term	Term
...

Type	Pattern
Hearst-I	$NP_L \{,\} \text{ (such as) } \{NP,\}^* \{\text{and} \text{or}\} NP$
Hearst-II	$NP_L \{,\} \text{ (include(s) including) } \{NP,\}^* \{\text{and} \text{or}\} NP$
Hearst-III	$NP_L \{,\} \text{ (e.g. e.g) } \{NP,\}^* \{\text{and} \text{or}\} NP$
IsA-I	$NP \text{ (is are was were being) } (a an) NP_L$
IsA-II	$NP \text{ (is are was were being) } \{\text{the, those}\} NP_L$
IsA-III	$NP \text{ (is are was were being) } \{\text{another, any}\} NP_L$

- Output: <term, label, pattern, source, weight> tuples
- Challenges
 - Boundary detection: term boundary, label boundary
 - Label selection

Subtask: Term \rightarrow Label (cont.)

- Merge tuples
 - For each term T and label L , compute $w(T, L)$
- Methods
 - Simple counting
 - Count the number of $\langle T, L, P, S, W \rangle$ tuples for each (T, L) pair
 - Or TF-IDF
 - Nonlinear evidence fusion (Shi et al., ACL'2011)

$$Score(T, L) = \left(\sum_{i=1}^K \sqrt[p]{m_i} \right) \cdot IDF(L)$$

m_i : #tuples for pattern i

$x_{i,j}$: Gain value given the j 'th tuple for pattern i

Subtask: Class→Label

- Input
 - Class C : {orange, apple, pear, banana...}
- Output
 - Label list for C : fruit, tree, flavor...
- Method: Voting
 - orange: color, flavor, client, network, fruit, county, tree...
 - apple: company, brand, fruit, manufacturer, client, tree...
 - pear: fruit, tree, shape, flavor, juice, cut, wood...
 - banana: fruit, crop, flavor, tree, food, plant, vegetable...



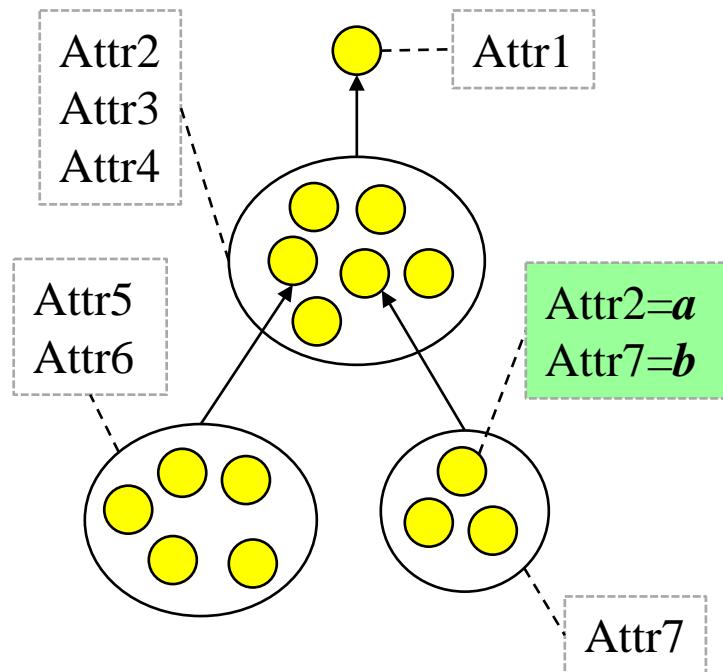
Related Papers

- Hearst, COLING'1992
Automatic Acquisition of Hyponyms from Large Text Corpora
- Pantel & Ravichandran, HLT-NAACL'2004
Automatically Labeling Semantic Classes
- Snow et al., COLING-ACL'2006
Semantic Taxonomy Induction from Heterogenous Evidence
- Banko et al., IJCAI'2007
Open Information Extraction from the Web
- Cafarella et al., VLDB'2008
WebTables: Exploring the Power of Tables on the Web
- Durme & Pasca, AAAI'2008
Finding cars, Goddesses and Enzymes: Parametrizable Acquisition of Labeled Instances for Open-Domain Information Extraction
- Zhang et al., ACL'2011
Nonlinear Evidence Fusion and Propagation for Hyponymy Relation Mining

Outline

- Overview
- Semantic class mining
- Semantic hierarchy construction
- **Mining attribute names and values**
- General relation extraction
- Demo
- Summary

Semantic Attributes



(city, population)

(country, flag)

(country, capital)

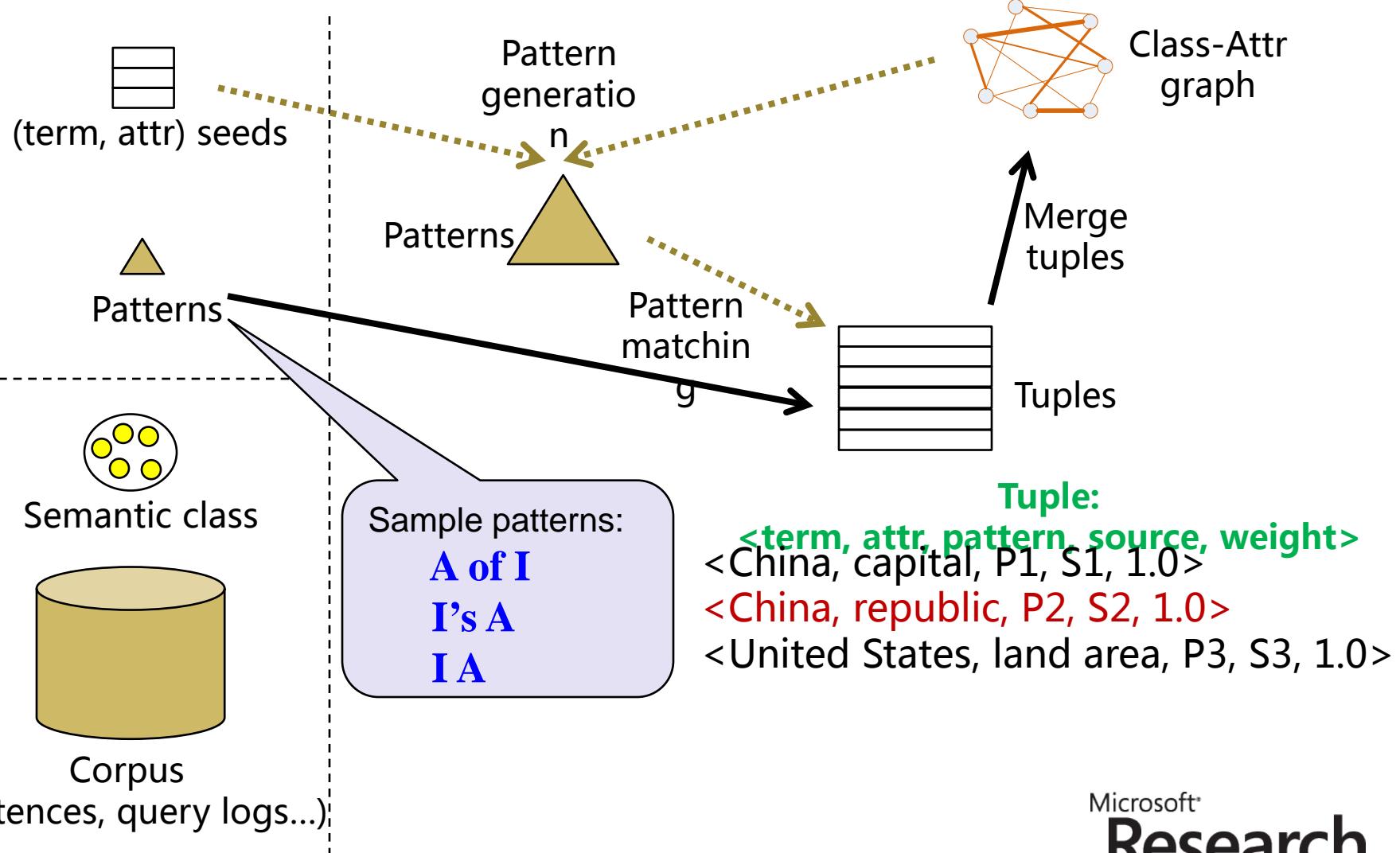
(company, CEO)

(China, capital, Beijing)

(Microsoft, CEO, Steve Ballmer)

(Barack Obama, Birth year, 1961)

Attribute Name Extraction from Unstructured Text



Attribute Name Extraction from Unstructured Text

- Major papers:
 - Pasca, WWW'2007
Organizing and Searching the World Wide Web of Facts Step Two: Harnessing the Wisdom of the Crowds
 - Durme et al., COLING'2008
Class-Driven Attribute Extraction
 - Pasca et al., CIKM'2007
The Role of Documents vs. Queries in Extracting Class Attributes from Text
 - Bellare et al., NIPS'2007
Lightly-Supervised Attribute Extraction
 - Reisinger & Pasca, 2009
Low-Cost Supervision for Multiple-Source Attribute Extraction
 - Tokunaga et al., IJCNLP'2005 (Japanese data)
Automatic Discovery of Attribute Words from Web Documents
 - ...

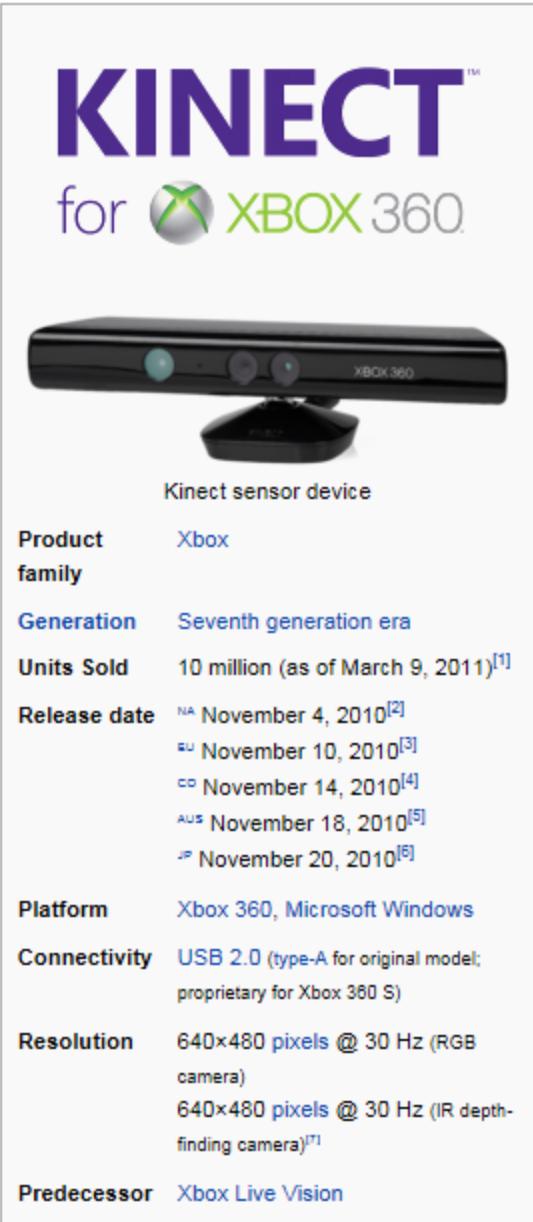
Attribute Name & Value Extraction

- From Unstructured Text

- Similar with extracting attribute names from unstructured text

	Mountain Peak	Continent	Height
C	Mount Everest	Asia	8,850 m
C	Aconcagua	South America	6,959 m
Fron	Mount McKinley (Denali)	North America	6,194 m
Fron	Kilimanjaro	Africa	5,895 m
Fron	Mount Elbrus	Europe	5,642 m
Fron	Vinson Massif	Antarctica	4,897 m
	Carstensz Pyramid	Australia - Oceania	4,884 m
	Mount Kosciuszko (The highest point on the Australian landmass)		2,228 m

Kinect for Xbox 360



Outline

- Overview
- Semantic class mining
- Semantic hierarchy construction
- Mining attribute names and values

➤ **General relation extraction**

- Demo
- Summary

General Relations

- Relations: Facts involving entities
 - [PER Susan Dumais] works for [ORG Microsoft Research], which is headquartered in [LOC Redmond, WA]
 - DefeatedIn(Dallas Mavericks, Miami Heat, 2011 NBA Finals)
- Relations vs. Events
 - Vague boundary
- History
 - Introduced in MUC-7 (1997) , extended by ACE, continued by KBP
 - Gain popularity in molecular biology, recent works including extracting protein-protein interaction

Type	Subtype
ART (artifact)	User-Owner-Inventor-Manufacturer
GEN-AFF (Gen-affiliation)	Citizen-Resident-Religion-Ethnicity, Org-Location
METONYMY*	<i>none</i>
ORG-AFF (Org-affiliation)	Employment, Founder, Ownership, Student-Alum, Sports-Affiliation, Investor-Shareholder, Membership
PART-WHOLE (part-whole)	Artifact, Geographical, Subsidiary
PER-SOC* (person-social)	Business, Family, Lasting-Personal
PHYS* (physical)	Located, Near

ACE' 05 relation types
Microsoft®
Research

Supervised Learning

- Treat relation mining as a classification problem
 - Use relational and non-relational mentions as positive and negative data, respectively
- Solve it with supervised Machine learning algorithms
 - Popular choices include SVM, MaxEnt, KNN
- Key: data representation
 - Feature based methods
 - Kernel based methods
- Evaluate metrics: Precision, Recall, F1 on relation mention level

Features

- List of common features (Kambhatla 2004)
 - **Words:** Words of both the entity mentions and all the words in between.
 - **Entity Type:** Entity type of both the mentions.
 - **Mention Level:** Mention level of both the mentions.
 - **Overlap:** Number of words separating the two mentions, number of other mentions in between, flags indicating whether the two mentions are in the same noun phrase, verb phrase or prepositional phrase.
 - **Dependency:** Words and PoS and chunk labels of the words on which the mentions are dependent in the dependency tree
 - **Parse Tree:** Path of non-terminals (removing duplicates) connecting the two mentions in the parse tree, and the path annotated with head words.
- Other features (Zhou et al. 2005)
 - **Based phrase chunking** chunk labels and chunk heads in between
 - **Semantic resources** (country list, etc)

Kernel based Methods

- Kernel (X, Y) defines similarity between X and Y
- X and Y can be
 - Vectors of features (as in previous slides)
 - Objects (string sequence, Parse trees)
- Kernel-based methods
 - Don't require extensive feature engineering
 - Maybe computational expensive
- Multiple Kernels can also be used in combination with a composite kernel (Zhao and Grishman, 2005)

Subsequence Kernel (Bunescu and Mooney, 2005)

- Implicit features are sequences of words anchored at the two entity names
 - s = a word sequence

<e₁> ... **bought** ... **<e₂>** ... **billion** ... **deal.**

- x = an example sentence, containing s as a subsequence

Google has **bought** **video-sharing website** **YouTube** in a controversial \$1.6 **billion deal.**

$\underbrace{\hspace{1cm}}_{g_1=1}$ $\underbrace{\hspace{3cm}}_{g_2=3}$ $\underbrace{\hspace{3cm}}_{g_3=4}$ $\underbrace{\hspace{1cm}}_{g_4=0}$

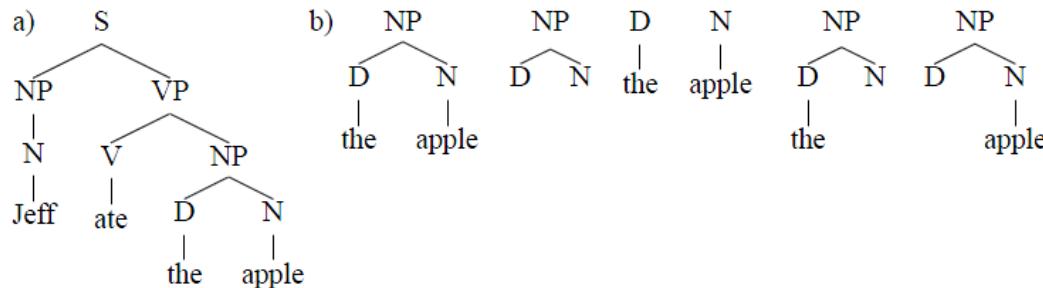
- $\varphi_s(x)$ = the value of feature s in example x

$$\varphi_s(x) = \lambda^{\sum g_i} = \lambda^{gap(s,x)} = \lambda^{1+3+4+0}$$

- $K(x_1, x_2) = \varphi(x_1)\varphi(x_2)$ = the number of common “anchored” subsequences between x_1 and x_2 , weighted by their total gap

Tree Kernel for RDC

- Convolution kernels for NLP (Collins and Duffy. 2001)
 - $K(T_1, T_2)$ defined over trees T_1 and T_2
 - Measured as number of overlapping fragments.

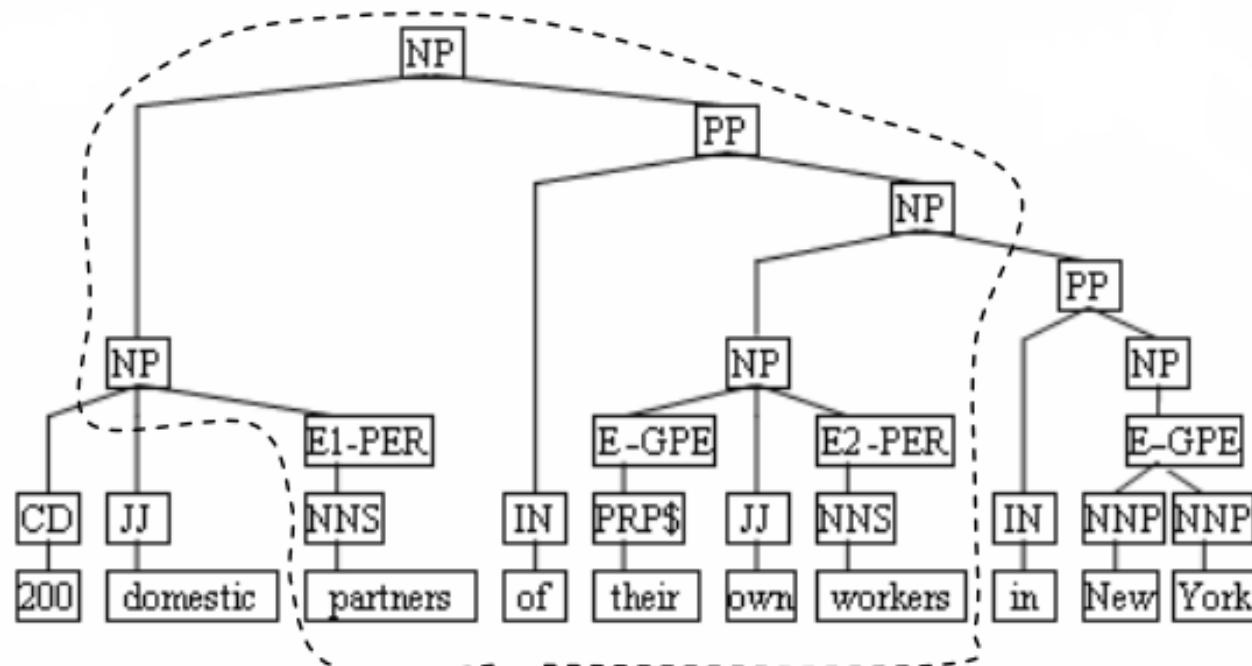


An example parse tree(a) and its sub-trees(b)

- Parse tree needs to be augmented before used for RDC
- Tree kernel for RDC differs in ways to augment/prune trees

Tree kernels for RDC

- An example of pruned parse tree augmented with entity types (Zhang et al. 2006)



Semi-Supervised Learning

- Supervised learning requires sufficient amount of annotated data
 - Expensive to obtain
 - Annotation error still occurs even dual annotated and adjudicated (ACE 2005)
- Semi-supervised learning (SSL) use a handful of seed tuples or patterns
- Bootstrapping alternates between finding pairs of arguments and contexts(pattern) of them

Bootstrapping

Initial Seed Tuples:

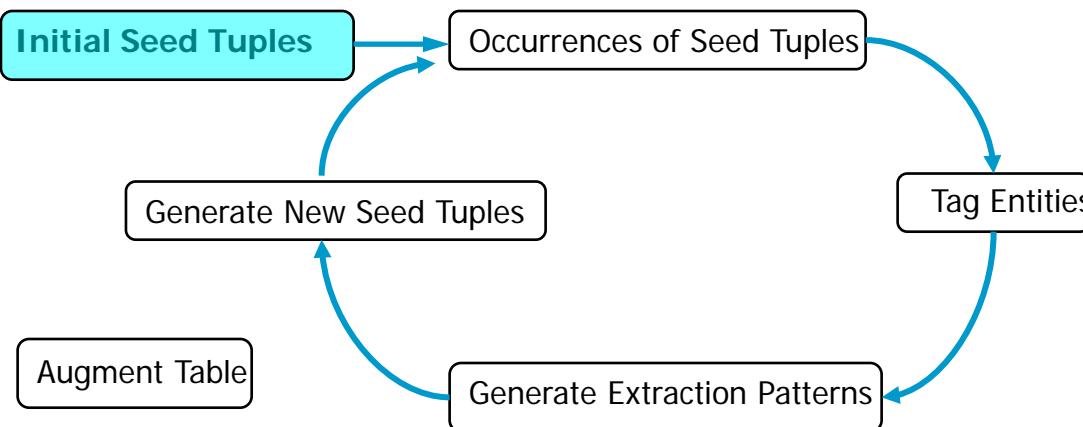
ORGANIZATION	LOCATION
MICROSOFT	REDMOND
IBM	ARMONK
BOEING	SEATTLE
INTEL	SANTA CLARA

DIRPRE (Brin 1998) patterns:

<*STRING1*>'s headquarters in <*STRING2*>

Snowball patterns:

<*left, NE tag1, middle, NE tag2, right*>,
left, middle, right are weighted terms



Evaluating Patterns and tuples
(Snowball)

$$Conf(Pat) = \frac{Positive}{Positive + Negative}$$

$$Conf(Tuple) = 1 - \prod(1 - Conf(P_i))$$

Weakly Supervision

- Handful of seeds for supervision

$+/-$	Arg a_1	Arg a_2
+	Google	YouTube
+	Adobe Systems	Macromedia
+	Viacom	DreamWorks
+	Novartis	Eon Labs
-	Yahoo	Microsoft
-	Pfizer	Teva

Table 1: Corporate Acquisition Pairs.

Bunescu and Mooney, 2007

$+/S_1$: Search engine giant **Google** has bought video-sharing website **YouTube** in a controversial \$1.6 billion deal.

$-/S_2$: The companies will merge **Google**'s search expertise with **YouTube**'s video expertise, pushing what executives believe is a hot emerging market of video offered over the Internet.

$+/S_3$: **Google** has acquired social media company, **YouTube** for \$1.65 billion in a stock-for-stock transaction as announced by Google Inc. on October 9, 2006.

$+/S_4$: Drug giant **Pfizer Inc.** has reached an agreement to buy the private biotechnology firm **Rinat Neuroscience Corp.**, the companies announced Thursday.

$-/S_5$: He has also received consulting fees from Alpharma, Eli Lilly and Company, **Pfizer**, Wyeth Pharmaceuticals, **Rinat Neuroscience**, Elan Pharmaceuticals, and Forest Laboratories.

Figure 1: Sentence examples.

Research

Weakly Supervision (cont.)

- A SVM solution to tolerate noisy positive instances

minimize:

$$J(w, b, \xi) = \frac{1}{2} \|w\|^2 + \frac{C}{L} \left(c_p \frac{L_n}{L} \Xi_p + c_n \frac{L_p}{L} \Xi_n \right)$$

$$\Xi_p = \sum_{X \in \mathcal{X}_p} \sum_{x \in X} \xi_x$$

$$\Xi_n = \sum_{X \in \mathcal{X}_n} \sum_{x \in X} \xi_x$$

Use a lower penalize factor for positive errors to tolerate noises from positive instances

subject to:

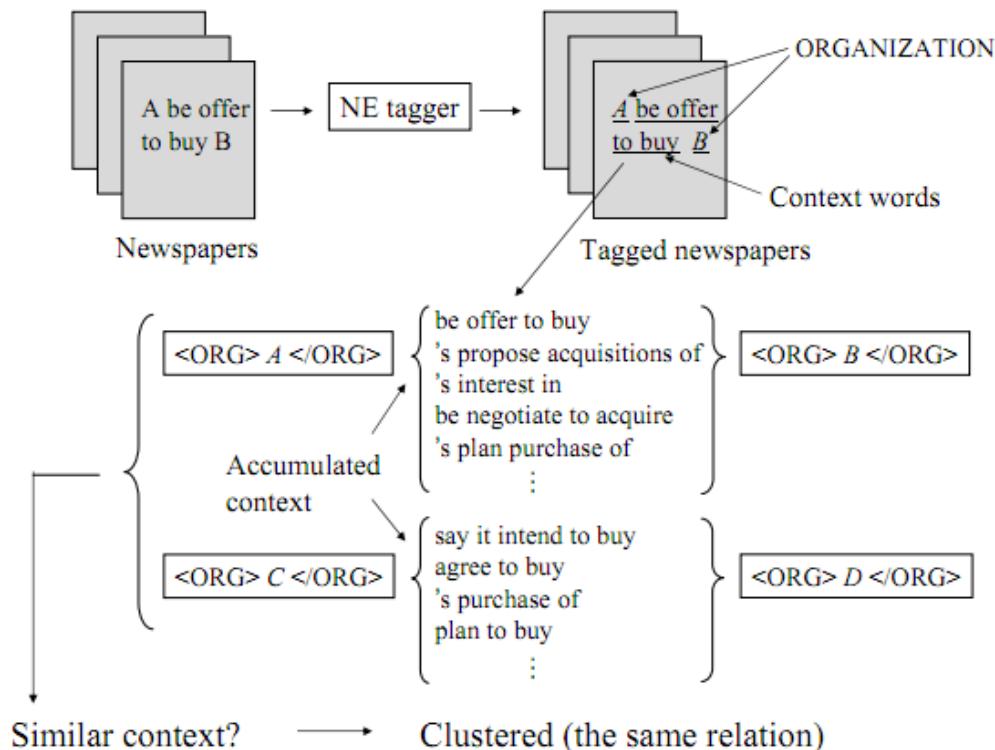
$$w \phi(x) + b \geq +1 - \xi_x, \quad \forall x \in X \in \mathcal{X}_p$$

$$w \phi(x) + b \leq -1 + \xi_x, \quad \forall x \in X \in \mathcal{X}_n$$

$$\xi_x \geq 0$$

Unsupervised Learning

- Automatically find major relations and respective arguments
- builds on the same duality of name pairs and contexts as relation bootstrapping methods



Hasegawa et al. 2004

- Uses Sekine's Extended NE tagger
- A domain is defined as a pair of name classes
- Bag-of-words features to model relational context
- hierarchical clustering

References for General Relation Mining

- MUC-7, http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_toc.html
- ACE, <http://www.itl.nist.gov/iad/mig/tests/ace/>
- KBP, <http://nlp.cs.qc.cuny.edu/kbp/2011/>
- Nanda Kambhatla. Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Information Extraction. ACL 2004
- GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. Exploring Various Knowledge in Relation Extraction. ACL 2005
- Shubin Zhao and Ralph Grishman. Extracting Relations with Integrated Information Using Kernel Methods. ACL 2005
- Razvan Bunescu and Raymond J. Mooney . Subsequence Kernels for Relation Extraction . In Proceedings of the 19th Conference on Neural Information Processing Systems (NIPS), Vancouver, BC, December 2005

References for General Relation Mining (cont.)

- Michael Collins and Nigel Duffy. Convolution Kernels for Natural Language. NIPS 2001.
- Min ZHANG, Jie ZHANG, Jian SU, Exploring Syntactic Features for Relation Extraction using a Convolution Tree Kernel, In ACL 2006.
- Sergei Brin. Extracting Patterns and Relations from the World Wide Web. In Proc. World Wide Web and Databases International Workshop, pages 172-183. Number 1590 in LNCS, Springer, March 1998.
- Eugene Agichtein and Luis Gravano, Snowball: Extracting Relations from Large Plain-Text Collections, In Proc. 5th ACM International Conference on Digital Libraries (ACM DL), 2000
- Razvan Bunescu and Raymond J. Mooney. Learning to Extract Relations from the Web using Minimal Supervision. ACL 2007
- Takaaki Hasegawa, Satoshi Sekine, Ralph Grishman Discovering Relations among Named Entities from Large Corpora. ACL 2004.

Outline

- Overview
 - Semantic class mining
 - Semantic hierarchy construction
 - Mining attribute names and values
 - General relation extraction
- **Demo**
- **Summary**

Demo: NeedleSeek



- A sub-project of the Sempute (*Semantic Computing*) project in WSM group, MSRA
- URL: <http://needleseek.msra.cn>

Semantic Mining: **Summary**

- Semantic class mining
 - Sample: {C++, C#, Java, PHP, Perl, ...}
 - Methods: Pattern matching (1st-order co-occurrences); distributional similarity (2nd-order co-occurrences)
- Semantic hierarchy construction
 - Key task: Hyponymy extraction (Beijing→city; pear→fruit; pear→shape)
 - Pattern matching; tuple aggregation; Label voting
- Mining attribute names and values
 - Samples: (company, CEO); (China, capital, Beijing)
 - Pattern learning; pattern matching; Table extraction; Wikipedia Infobox
- General relation extraction
 - Sample: WorkFor(Susan Dumais, Microsoft Research)
 - Supervised, semi-supervised, & unsupervised learning
 - Process contexts (especially middle contexts)



Paraphrases and Applications



Haifeng Wang, Shiqi Zhao

August 31st, 2011, CCFADL

Outline

- Part I
 - NLP for Web Applications
- Part II
 - Introduction
 - Paraphrase Identification
 - Paraphrase Extraction
- Part III
 - Paraphrase Generation
 - Applications of Paraphrases
 - Evaluation of Paraphrases
 - Conclusions and Future work

Outline

- Part I
 - NLP for Web Applications
- Part II
 - Introduction
 - Paraphrase Identification
 - Paraphrase Extraction
- Part III
 - Paraphrase Generation
 - Applications of Paraphrases
 - Evaluation of Paraphrases
 - Conclusions and Future work

CL vs. NLP

Computational Linguistics, CL

ACL: Association for Computational Linguistics

COLING: International Conference on Computational Linguistics

ICCL: International Committee on Computational Linguistics

CNCCL: Chinese National Conference on Computational Linguistics

ICL: Institute of Computational Linguistics

Natural Language Processing, NLP

EMNLP: Empirical Methods in Natural Language Processing

IJCNLP: International Joint Conference on Natural Language Processing

AFNLP: Asian Federation of Natural Language Processing

YSSNLP: Young Scholar Symposium on Natural Language Processing

****NLPLAB:** **Natural Language Processing LAB

Impact?

History
Theory
Methodology



NLP Areas

Area	#Submission	#Accepted	Rate
Machine Translation	82	23	28.0%
Semantics	67	14	20.9%
Syntax and Parsing	49	14	28.6%
Information Extraction	49	10	20.4%
Discourse, Dialogue and Pragmatics	43	9	20.9%
Summarization and Generation	44	8	18.2%
Phonology, Morphology, Segmentation, POS, Chunking	31	8	25.8%
Sentiment Analysis, Opinion Mining, Classification	45	7	15.6%
Statistical and Machine Learning Methods	40	6	15.0%
Spoken Language Processing	19	6	31.6%
Information Retrieval	28	4	14.3%
Language Resource	26	4	15.4%
Text Mining and NLP Applications	21	4	19.0%
Question Answering	25	3	12.0%
Total	569	120	21.1%

NLP Areas

Application

Area	#Submission	#Accepted	Rate
Machine Translation	82	23	28.0%
Semantics	67	14	20.9%
Syntax and Parsing	49	14	28.6%
Information Extraction	49	10	20.4%
Discourse, Dialogue and Pragmatics	43	9	20.9%
Summarization and Generation	44	8	18.2%
Phonology, Morphology, Segmentation, POS, Chunking	31	8	25.8%
Sentiment Analysis, Opinion Mining, Classification	45	7	15.6%
Statistical and Machine Learning Methods	40	6	15.0%
Spoken Language Processing	19	6	31.6%
Information Retrieval	28	4	14.3%
Language Resource	26	4	15.4%
Text Mining and NLP Applications	21	4	19.0%
Question Answering	25	3	12.0%
Total	569	120	21.1%

NLP Taxonomy

- Sub-task
 - Analysis & understanding, generation
- Level
 - Morphology, syntax, semantics, pragmatics
- Grammar
 - PS, DS, LFG, HPSG, CCG ...
- Unit
 - Character, word, phrase, sentence, paragraph ...
- Style
 - Spoken language, written language
- Application
 - Translation, information retrieval and extraction, sentiment, QA, summarization, grammar check ...
- Approach
 - Rationalist and empiricist approaches
- Data
 - Lexicon, rules, corpus (labeled and unlabeled)

Difficulties

- Complex structure
 - Mapping between string and structure
- Ambiguities
 - Disambiguation
- Examples
 - 打: 打酱油、打毛衣、打人、打针
 - pretty little girls' school
 - Does the school look little?
 - Do the girls look little?
 - Do the girls look pretty?
 - Does the school look pretty?

Approaches

- Rationalist approaches
 - Linguistic theory
 - Grammar system
 - Rules
 - Usually manually compiled
 - Popular in NLP application (e.g. RBMT)

Noam Chomsky

It must be recognized that the notion "probability of a sentence" is an entirely useless one, under any known interpretation of this term.

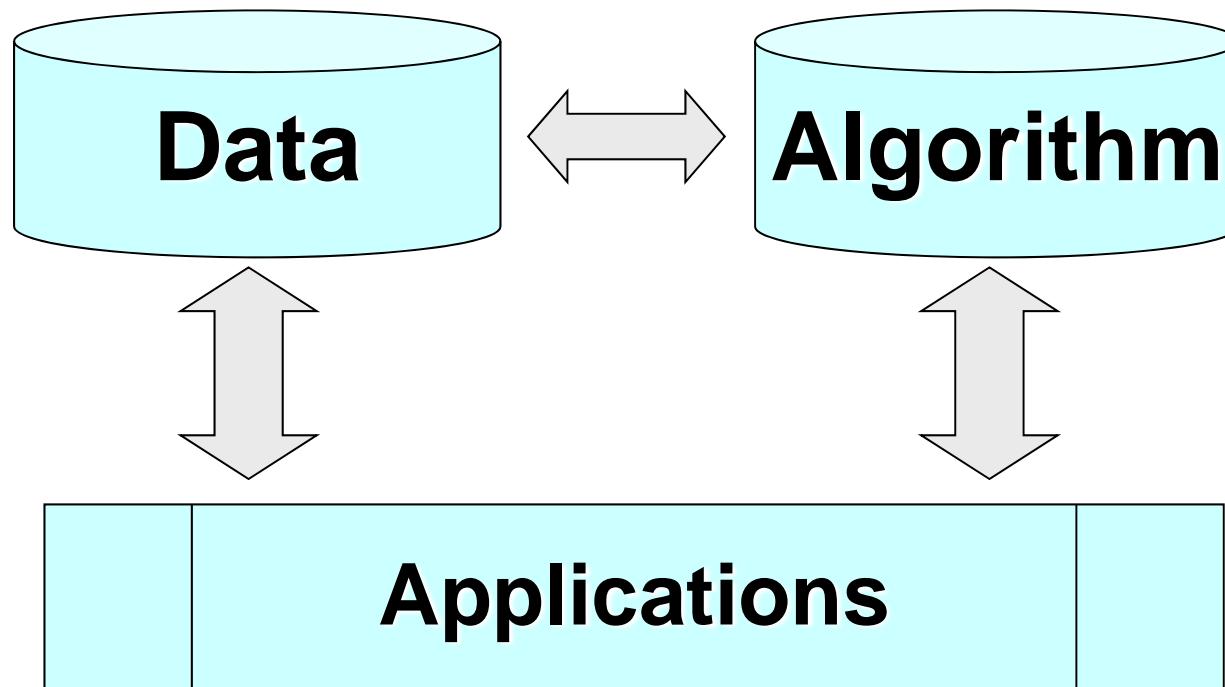
- Empiricist approaches
 - Corpus
 - Labeled, unlabeled
 - Monolingual, multilingual
 - Statistical and Machine Learning Approaches
 - Dominant approach in NLP research

Frederick Jelinek

Whenever I fire a linguist our system performance improves.



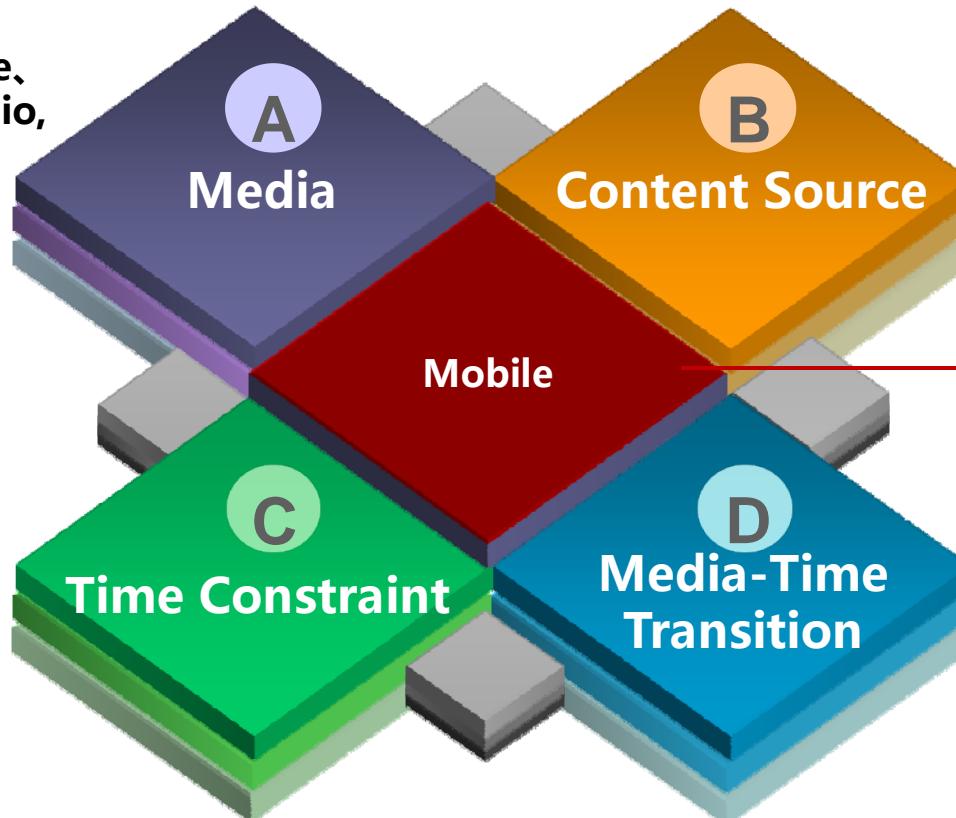
Data vs. Algorithm



Trends of the Internet

- The Internet and users have changed a lot in the past decade

- Text→Image, Video, Audio, etc.



- Users want real-time information when searching

- Professional Editors→ User Generated ; Static→Dynamically Generated

3.18 mobile users

- Simply Surfing→Online gaming, video-watching, Social interaction – All kinds of Application

Queries to Baidu Search Engine

听起来欢乐的歌曲

joyous song

百度一下

令人心情愉快的图片

Pleasant pictures

百度一下

现在几点了

What time is it

百度一下

电脑中毒了怎么办

How to deal with computer virus

百度一下

哪能买到漂亮衣服

Where could I buy some beautiful clothes

百度一下

北京哪能找到女朋友

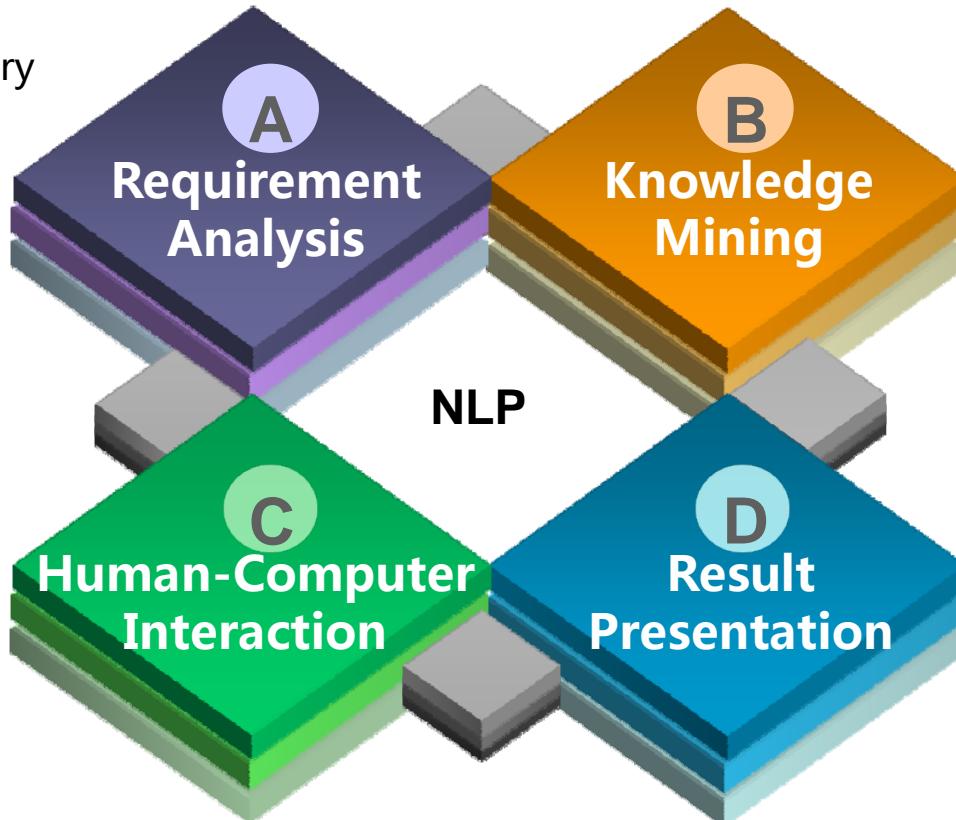
Where could I get a girlfriend in Beijing

百度一下



Challenge to NLP

- Complex Query
- Diversiform Requirement



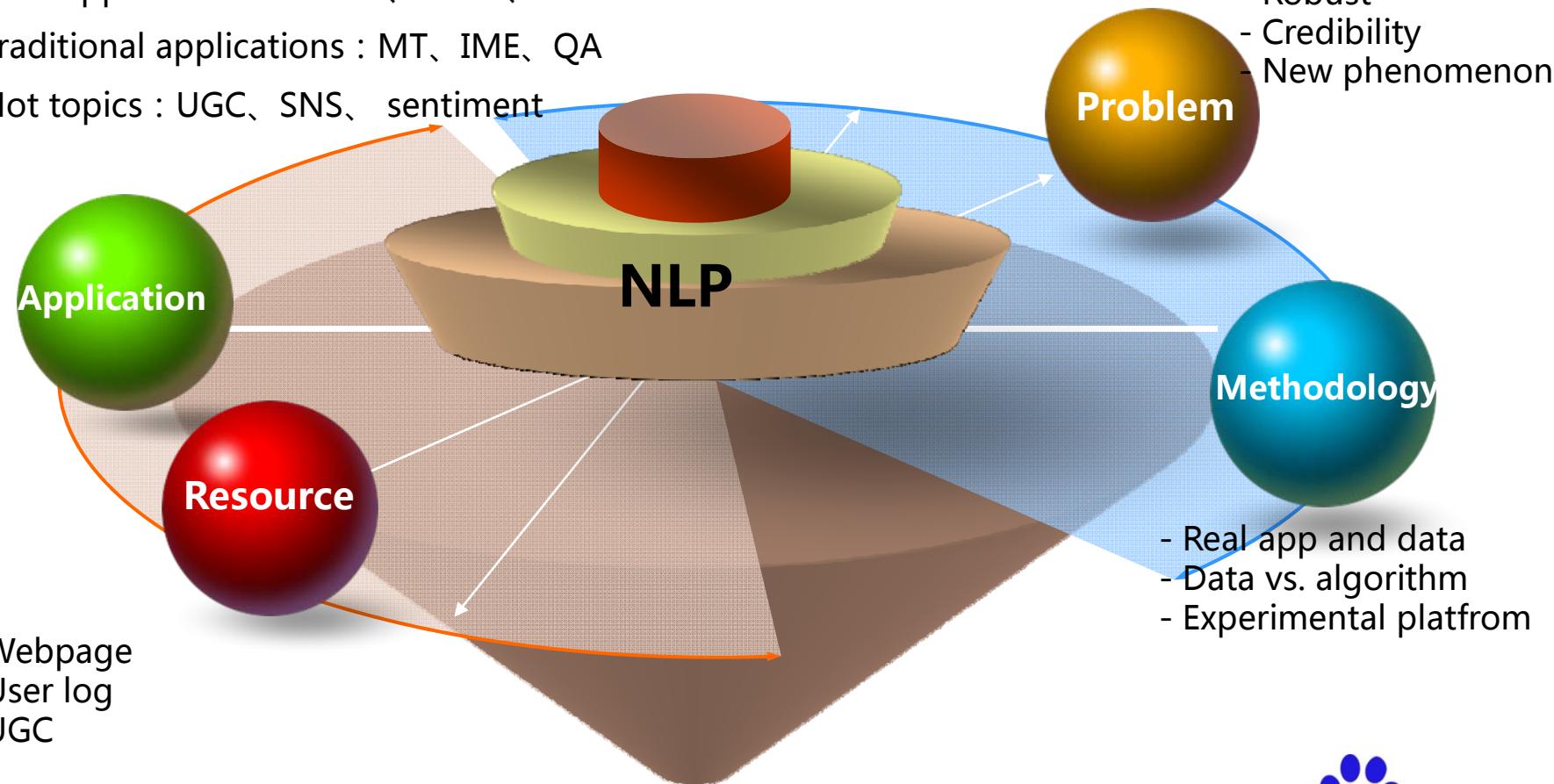
- Suggestion
- Extension
- Interaction

- Hidden web, hiding knowledge
 - Structured, semi-structured, unstructured
 - Various levels
-
- Direct answer
 - Clustering
 - Summarization
 - Relation Graph
 - intelligent push
 - Rich media

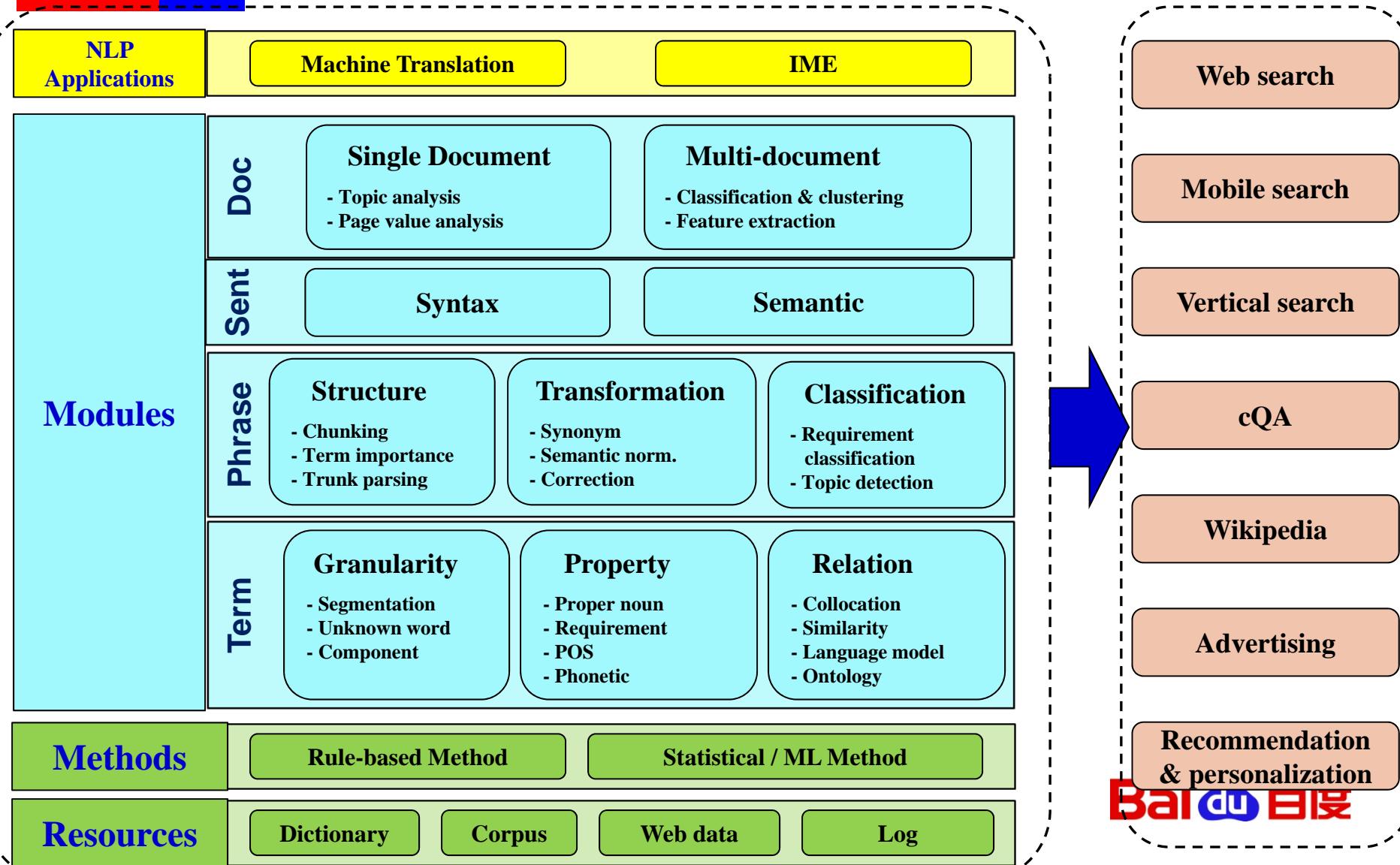
NLP for Web Applications

- Web applications : search, ecom, SNS
- Traditional applications : MT, IME, QA
- Hot topics : UGC, SNS, sentiment

- Platform
- Robust
- Credibility
- New phenomenon



NLP for Web Applications



Paraphrasing for Web Applications

Machine Translation

- rewrite input sentence
- alleviate data sparseness
- expand training data
- automatic evaluation

Summarization

- sentence clustering
- rewrite summaries
- automatic evaluation

Question Answering

- question rewriting
- answer extraction template
paraphrasing

Natural Language Generation

- rewriting of the automatically generated texts

Information Extraction

- template expansion

Information Retrieval

- query rewriting

Other applications

- identify plagiarism
- text simplification
- writing style transformation
- error correction
-

Examples

- 天龙八步 → 天龙八部
- 怎样能有归一证 → 怎样能有皈依证
- 宝马X6价钱 → 宝马X6报价
- 成都的哥罢工 → 成都出租车罢工
- 赞颂母爱的现代诗 → 母爱的现代诗
- 康柏笔记本vista系统一键恢复 → 康柏vista
一键恢复

Outline

- Part I
 - NLP for Web Applications
- Part II
 - Introduction
 - Paraphrase Identification
 - Paraphrase Extraction
- Part III
 - Paraphrase Generation
 - Applications of Paraphrases
 - Evaluation of Paraphrases
 - Conclusions and Future work

Definition

- Paraphrase
 - Noun
 - Alternative expressions of the same meaning
 - Verb
 - Generate paraphrases for the input expression
- “same meaning”?
 - Quite subjective
 - Different degrees of strictness
 - Depend on applications

Paraphrase (noun): Alternative expressions of the same meaning

8月29日，男子110米栏决赛，刘翔憾获银牌。原因则在于，遭罗伯斯犯规阻挠，虽然古巴人终受严惩，翔飞人最终铜牌变银牌。



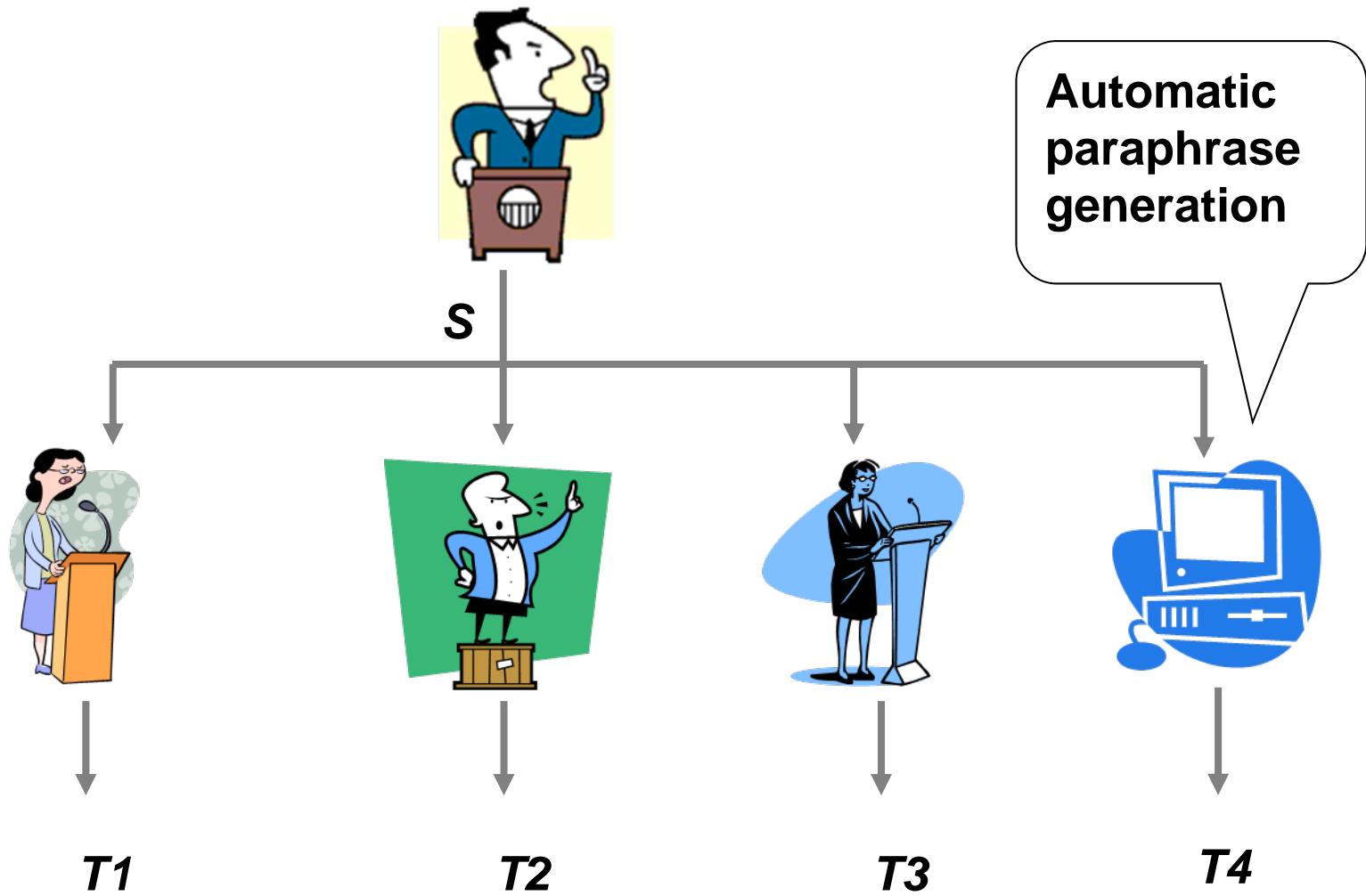
由于在比赛中对中国选手刘翔进行干扰，古巴名将罗伯斯尽管第一个冲过终点线，但随后被取消了比赛资格，无缘金牌，刘翔原本的铜牌也换成了一枚银牌。

北京时间8月29日，大邱田径世锦赛男子110米栏决赛的剧情有些跌宕起伏，古巴名将罗伯斯虽然率先到达终点，但最终因阻挡刘翔，因而被取消成绩，原本获得铜牌的刘翔名次递进一位，获得银牌。

北京时间8月29日，男子110米栏决赛结束，刘翔两次与罗伯斯碰撞，最后遗憾获得银牌。

日本的几名记者都向笔者道歉，说没太关注刘翔和罗伯斯的表现，尽管知道罗伯斯被取消了成绩，刘翔递补了银牌。

Paraphrase (verb): Generate paraphrases for an input S.



Classification of Paraphrases

- According to granularity
 - Surface paraphrases
 - Lexical level
 - Phrase level
 - Sentence level
 - Discourse level
 - Structural paraphrases
 - Pattern level
 - Collocation level

Example

- Lexical paraphrases (generally synonyms)
 - 笔记本 vs. 本本
- Paraphrase phrases
 - 列车/出轨 vs. 火车/脱轨
- Paraphrase sentences
 - 減肥/中/水果/可以/吃/什么
 - 吃/什么/水果/可以/瘦身
- Paraphrase patterns
 - [x]/文件/怎么/打开
 - 如何/打开/[x]/文件
- Paraphrase collocations
 - (捧走 OBJ 奖杯) vs. (获得 OBJ 奖杯)

Classification of Paraphrases

- According to paraphrase style
 - Trivial change
 - Phrase replacement
 - Phrase reordering
 - Sentence split & merge
 - Complex paraphrases

Example

- Trivial change
 - 考研/失败/怎么办 vs. 考研/失败/怎么办/**呢**
- Phrase replacement
 - 咖啡斑/的/治疗/**多少钱**
 - 咖啡斑/的/治疗/**费用/是多少**
- Phrase reordering
 - 红烧肉/菜谱 vs. 菜谱/红烧肉
- Sentence split & merge
 - 给/女朋友/买/什么/生日礼物
 - 女朋友/过生日/, /买/什么/礼物
- Complex paraphrases
 - 菜谱/红烧肉
 - 红烧肉/怎么烧/好吃

Research on Paraphrasing

- Paraphrase identification
 - Identify (sentential) paraphrases
- Paraphrase extraction
 - Extract paraphrase instances (different granularities)
- Paraphrase generation
 - Generate (sentential) paraphrases
- Paraphrase applications
 - Apply paraphrases in other areas

Textual Entailment – A Similar Direction

- Textual entailment:
 - A directional relation between two text fragments
 - T : the entailing text
 - H : the entailed hypothesis
 - T entails H if, typically, people reading T would infer that H is most likely true.
 - Compare entailment with paraphrase
 - Paraphrase is bidirectional entailment

Text Entailment – A Similar Direction

- Recognizing Textual Entailment Track (RTE)
 - RTE-1 (2004) to RTE-5 (2009)
 - RTE-6 (2010) is in progress
- Example:
 - *T*: A shootout at the Guadalajara airport in May, 1993, killed Cardinal Juan Jesus Posadas Ocampo.
 - *H*: Juan Jesus Posadas Ocampo died in 1993.

Outline

- Part I
 - NLP for Web Applications
- Part II
 - Introduction
 - Paraphrase Identification
 - Paraphrase Extraction
- Part III
 - Paraphrase Generation
 - Applications of Paraphrases
 - Evaluation of Paraphrases
 - Conclusions and Future work

Paraphrase Identification



- Specially refers to sentential paraphrase identification
 - Given any pair of sentences, automatically identifies whether these two sentences are paraphrases
- Paraphrase identification is not trivial

Susan often goes to see movies with her boyfriend.
Susan never goes to see movies with her boyfriend.



He said there will be major cuts in the salaries of high-level civil servants.
He claimed to implement huge salary cut to senior civil servants.



Overview

- Classification based methods
 - Reviewed as a binary classification problem, i.e., input s_1 and s_2 to a classifier and output 0/1
 - Compute the similarities between s_1 and s_2 at different levels, which are then used as classification features
- Alignment based methods
 - Align s_1 and s_2 first, and score the sentence pair based on the alignment results
 - Alignment based on ITG
 - Alignment based on quasi-synchronous dependency grammars

Classification based Methods

- Brockett and Dolan, 2005

- Features:

- String similarity features
 - Sentence length, word overlap, edit distance, ...

- Morphological variants
 - Word pairs with the same stem

- WordNet lexical mappings
 - Synonym pairs / word-hypernym pairs from WordNet

- Word association pairs
 - Automatically learned synonym pairs

- Classifier

- SVM classifier



orbit | orbital



operation | procedure



vendors | suppliers

Classification based Methods (cont')

- Finch et al., 2005
 - Using MT evaluation techniques to compute sentence similarities, which are then used as classification features
 - WER, PER, BLEU, NIST
 - Feature vector $\text{vec}(\mathbf{s}_1, \mathbf{s}_2)$
 - $\text{vec1}(\mathbf{s}_1, \mathbf{s}_2)$: \mathbf{s}_1 as reference, \mathbf{s}_2 as MT system output;
 - $\text{vec2}(\mathbf{s}_1, \mathbf{s}_2)$: \mathbf{s}_2 as reference, \mathbf{s}_1 as MT system output;
 - $\text{vec}(\mathbf{s}_1, \mathbf{s}_2)$: average of $\text{vec1}(\mathbf{s}_1, \mathbf{s}_2)$ and $\text{vec2}(\mathbf{s}_1, \mathbf{s}_2)$:
 - Classifier
 - SVM classifier

Classification based Methods (cont')

- Malakasiotis, 2009
 - Combining multiple classification features
 - String similarity (various levels)
 - Tokens, stems, POS tags, nouns only, verbs only, ...
 - Different measures
 - Edit distance, Jaro-Winkler distance, Manhattan distance...
 - Synonym similarity
 - Treat synonyms in two sentences as identical words
 - Syntax similarity
 - Dependency parsing of two sentences and compute the overlap of dependencies
 - Classifier
 - Maximum Entropy classifier



Alignment based Methods

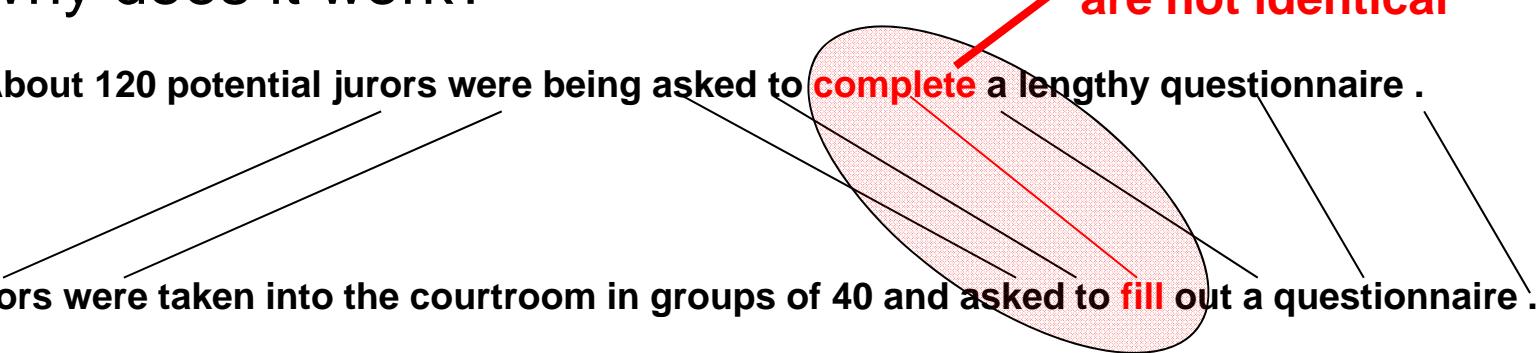
- Wu, 2005
 - Conduct alignment based on Inversion Transduction Grammars (ITG)
 - Sensitive to the differences in sentence structures
 - Without using any thesaurus to deal with lexical variation
 - Performance is comparable to the classification based methods
 - Also performs well in recognizing textual entailment

Alignment based Methods (cont')

- Das and Smith, 2009
 - Conduct alignment based on Quasi-Synchronous Dependency Grammar (QG)
 - Alignment between two dependency trees
 - Assumption: the dependency trees of two paraphrase sentences should be aligned closely
 - Why does it work?

About 120 potential jurors were being asked to **complete** a lengthy questionnaire .

The jurors were taken into the courtroom in groups of 40 and asked to **fill** out a questionnaire .



- Performs competitively with classification based methods

A Summary

- Classification based method is still the mainstream method, since:
 - Binary classification problem is well defined;
 - Classification algorithms and tools are readily available;
 - It can combine various features in a simple way;
 - It achieves state-of-the-art performance.

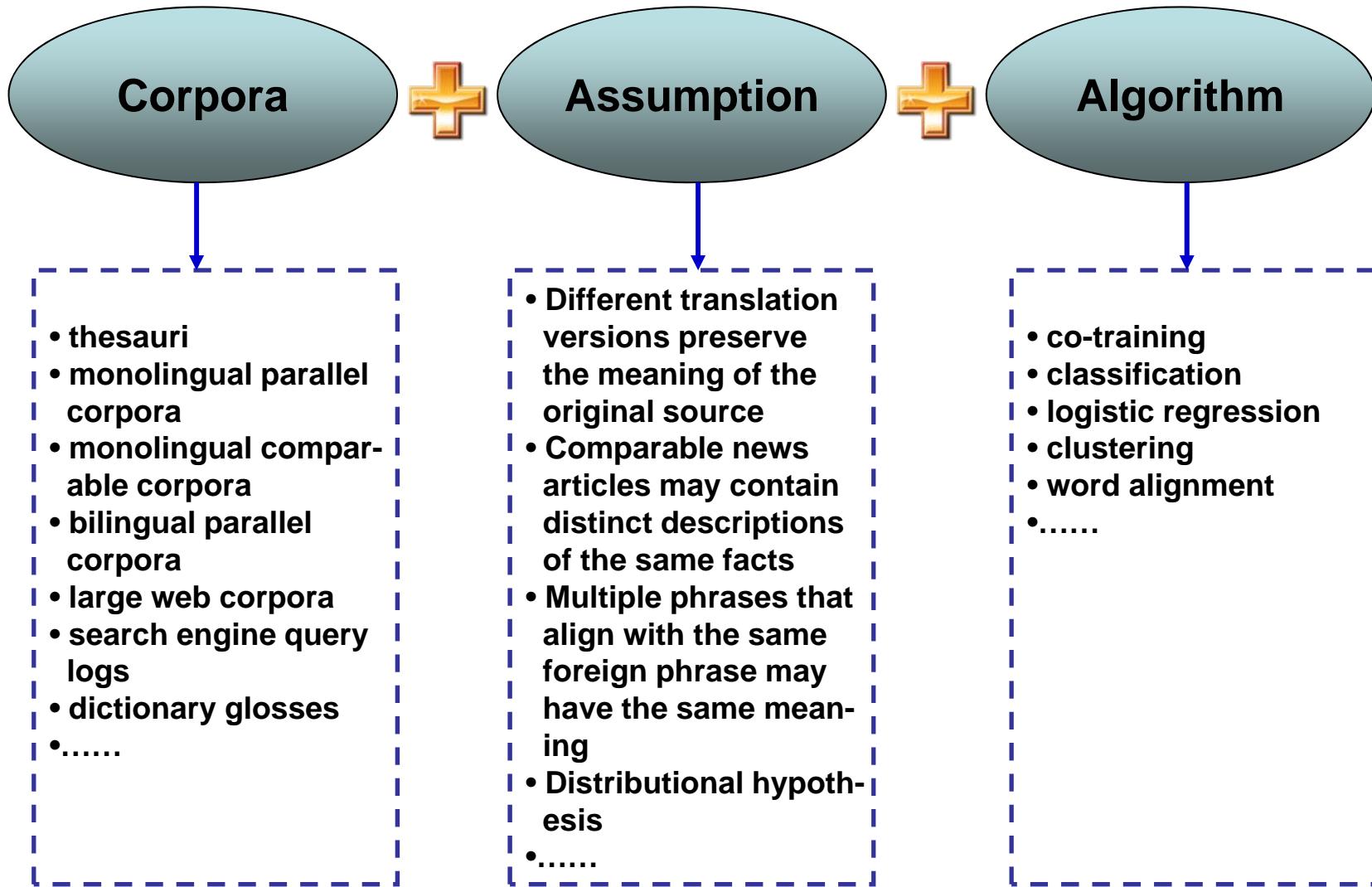
References

- Brockett and Dolan. 2005. Support Vector Machines for Paraphrase Identification and Corpus Construction.
- Finch et al. 2005. Using Machine Translation Evaluation Techniques to Determine Sentence-level Semantic Equivalence.
- Wu. 2005. Recognizing Paraphrases and Textual Entailment using Inversion Transduction Grammars.
- Malakasiotis. 2009. Paraphrase Recognition Using Machine Learning to Combine Similarity Measures.
- Das and Smith. 2009. Paraphrase Identification as Probabilistic Quasi-Synchronous Recognition.

Outline

- Part I
 - NLP for Web Applications
- Part II
 - Introduction
 - Paraphrase Identification
 - Paraphrase Extraction
- Part III
 - Paraphrase Generation
 - Applications of Paraphrases
 - Evaluation of Paraphrases
 - Conclusions and Future work

Three Elements for Paraphrase Extraction



Outline

- **Part II**
 - Introduction
 - Paraphrase Identification
 - **Paraphrase Extraction**
 - **From Thesauri**
 - From Monolingual Parallel Corpora
 - From Monolingual Comparable Corpora
 - From Bilingual Parallel Corpora
 - From Large Web Corpora
 - From Other Resources

Method Overview

- Extract words with specific semantic relations as paraphrases
 - Most common: synonyms
 - Other relations: hypernyms, hyponyms...
- Widely used thesauri
 - In English
 - WordNet
 - In other languages
 - E.g., HowNet, Tongyici Cilin in Chinese

Pros and Cons

- Pros
 - Existing resources
 - High quality
 - Thesauri are hand crafted
- Cons
 - Language limitation
 - Thesauri are not available in many languages
 - Difficult to update
 - Disambiguation

Outline

- **Part II**
 - Introduction
 - Paraphrase Identification
 - **Paraphrase Extraction**
 - From Thesauri
 - **From Monolingual Parallel Corpora**
 - From Monolingual Comparable Corpora
 - From Bilingual Parallel Corpora
 - From Large Web Corpora
 - From Other Resources

Method Overview

- Corpus
 - Multiple translations of the same foreign literary work
- Assumption
 - Different translation versions preserve the meaning of the original source, but may use different expressions

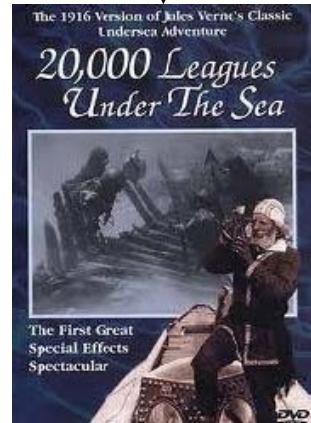
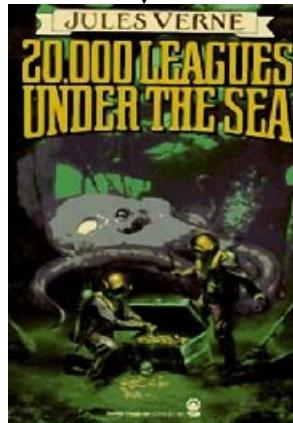
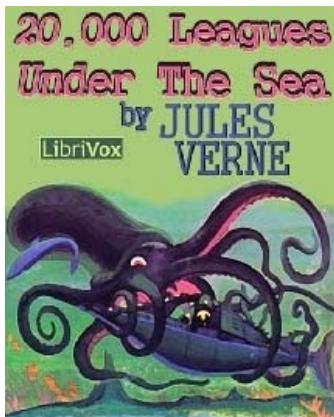
Example



Vingt mille lieues sous les mers
(in French)

20000 Leagues Under the Sea

(different English translation versions)



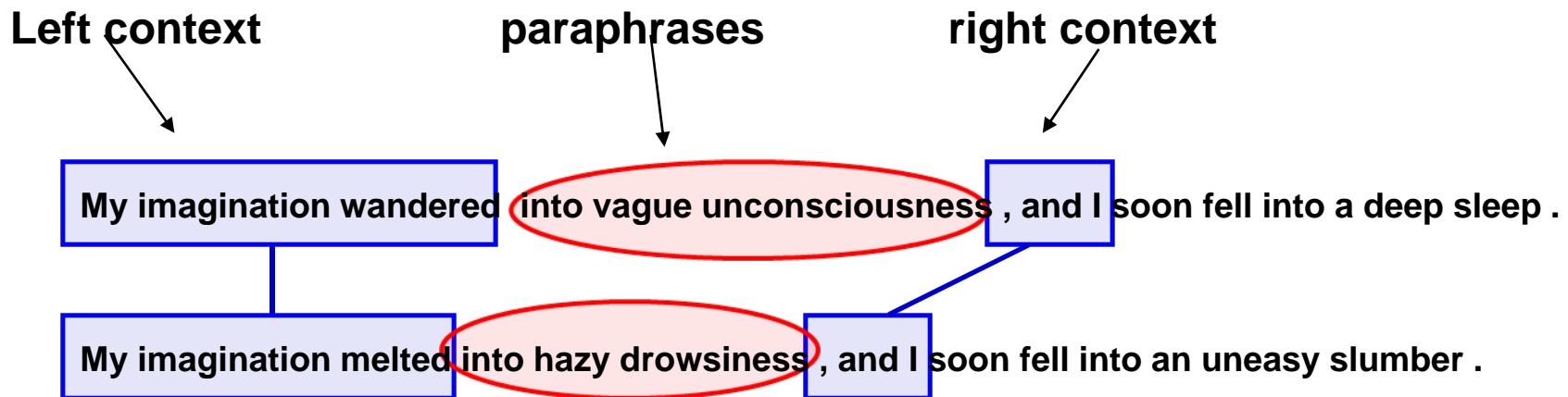
.....

Sentence Alignment and Preprocessing

- Barzilay and McKeown, 2001
 - Collected 11 English translations for 5 foreign novels
 - E.g., *Madame Bovary*, *Fairy Tale*, *Twenty Thousand Leagues under the sea...*
 - Sentence alignment
 - A dynamic programming algorithm
 - Produced 44,562 pairs of parallel sentences
 - Precision is 94.5%
 - Other preprocessing
 - POS tagging and chunking
 - Phrases are the atomic units in paraphrase extraction

Paraphrase Phrase Extraction

- Barzilay and McKeown, 2001 (cont')
 - Extracting paraphrase phrases
 - Assumption: phrases in aligned sentences which appear in similar contexts are paraphrases
 - Method: co-training
 - Iteratively learn contexts and paraphrases



Pros and Cons

- Pros
 - Easy to align monolingual parallel sentences
- Cons
 - Domain limitation
 - Limited in literary works
 - Scale limitation
 - The size of the corpus is relatively small
 - Context dependence
 - E.g., “*John said*” and “*he said*”

Other Monolingual Parallel Corpora

- Paraphrasing with definition sentences
 - Hashimoto et al., ACL-2011
 - Basic assumption
 - Sentences defining the same concept may mean the same thing
 - Two main steps:
 - Definition sentence collection
 - Paraphrase phrase recognition

Other Monolingual Parallel Corpora (cont.)

- Paraphrasing with definitions (cont.)
 - Definition sentence collection
 - Resource: Web corpora & wikipedia
 - Method: Simple template & SVM classifier
 - Paraphrase phrase recognition
 - Candidate phrase pair extraction
 - Dependency parsing on the sentences
 - Pair any two phrases across parallel sentences
 - SVM classifier
 - Features: surface similarity & context similarity

Pros and Cons

- Pros
 - Resources (Web corpora) are available
 - Precision is high (94%)
- Cons
 - Volume of the extracted paraphrases
 - 300,000 paraphrases from 600 Million web docs

Alignment on Monolingual Parallel Data

- Alignment by edit rate computation
 - Bouamor et al., ACL-2011
 - TER-plus (Translation Edit Rate Plus)
 - Originally designed for MT evaluation
 - Can also be used in paraphrase scoring
 - Computes an optimal set of word edits that can transform a candidate paraphrase into a reference paraphrase
 - TER-plus can exploit a paraphrase table and deal with paraphrase substitutes between sentences

Alignment on Monolingual Parallel Data (cont.)



- Alignment by edit rate computation (cont.)
 - Other techniques
 - Statistical word alignment
 - Symbolic expressions of linguistic variation
 - Syntactic similarity
 - The above techniques can be combined with TER-plus
 - Paraphrases yielded by these techniques can be used as a paraphrase table in TER-plus

Outline

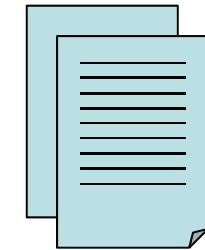
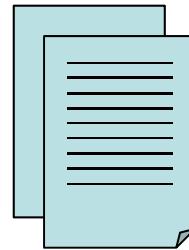
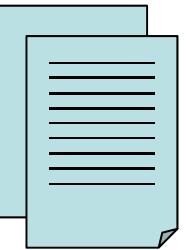
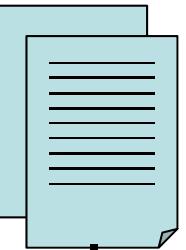
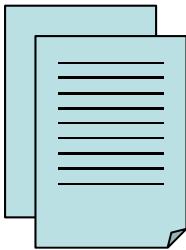
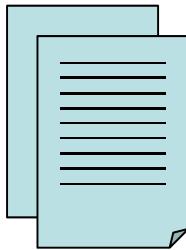
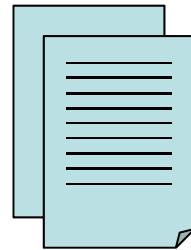


- **Part II**
 - Introduction
 - Paraphrase Identification
 - **Paraphrase Extraction**
 - From Thesauri
 - From Monolingual Parallel Corpora
 - **From Monolingual Comparable Corpora**
 - From Bilingual Parallel Corpora
 - From Large Web Corpora
 - From Other Resources

Method Overview

- Corpus
 - News articles that report the same event within a brief period of time
 - Produced by different news agencies
- Assumption
 - Comparable news articles may contain distinct descriptions of the same facts

Example



Comparable documents

Home > SPORT > TENNIS > FRENCH OPEN

French Open 2010: Justine Henin defeats Maria Sharapova

Four-time champion Justine Henin beat Maria Sharapova of Russia 6-2, 3-6, 6-3 in delayed French Open third-round match.

Published: 12:34PM BST 30 May 2010

« Previous 1 of 2 Images Next »



Share | Digg submit | Email | Text Size | French Open | Sport | Tennis | Maria Sharapova | Ads by Google | French Open Tennis Telegraph Tennis Match V

Respect: French Open 2010: Justine Henin (right) acknowledges Maria Sharapova after defeating the Russian at the French Open. Photo: GETTY IMAGES

With the match between two former world number ones held over at a set-all

Home / Topics / Story

PARIS — One winner-take-all set seemed like a final, and Justine Henin emerged the winner. Back on center court Sunday following an overnight suspension of play, Henin outslugged Maria Sharapova in a third-round showdown at the French Open, 6-2, 3-6, 6-3.

FULL ARTICLE AT ESPN

2010 French Open: Justine Henin tops Maria Sharapova in 3 sets

PARIS -- One winner-take-all set seemed like a final, and Justine Henin emerged the winner. Back on center court Sunday following an overnight suspension of play, Henin outslugged Maria Sharapova in a third-round showdown at the French Open, 6-2, 3-6, 6-3.

FULL ARTICLE AT BELLINGTON TELEGRAPH

Related Articles

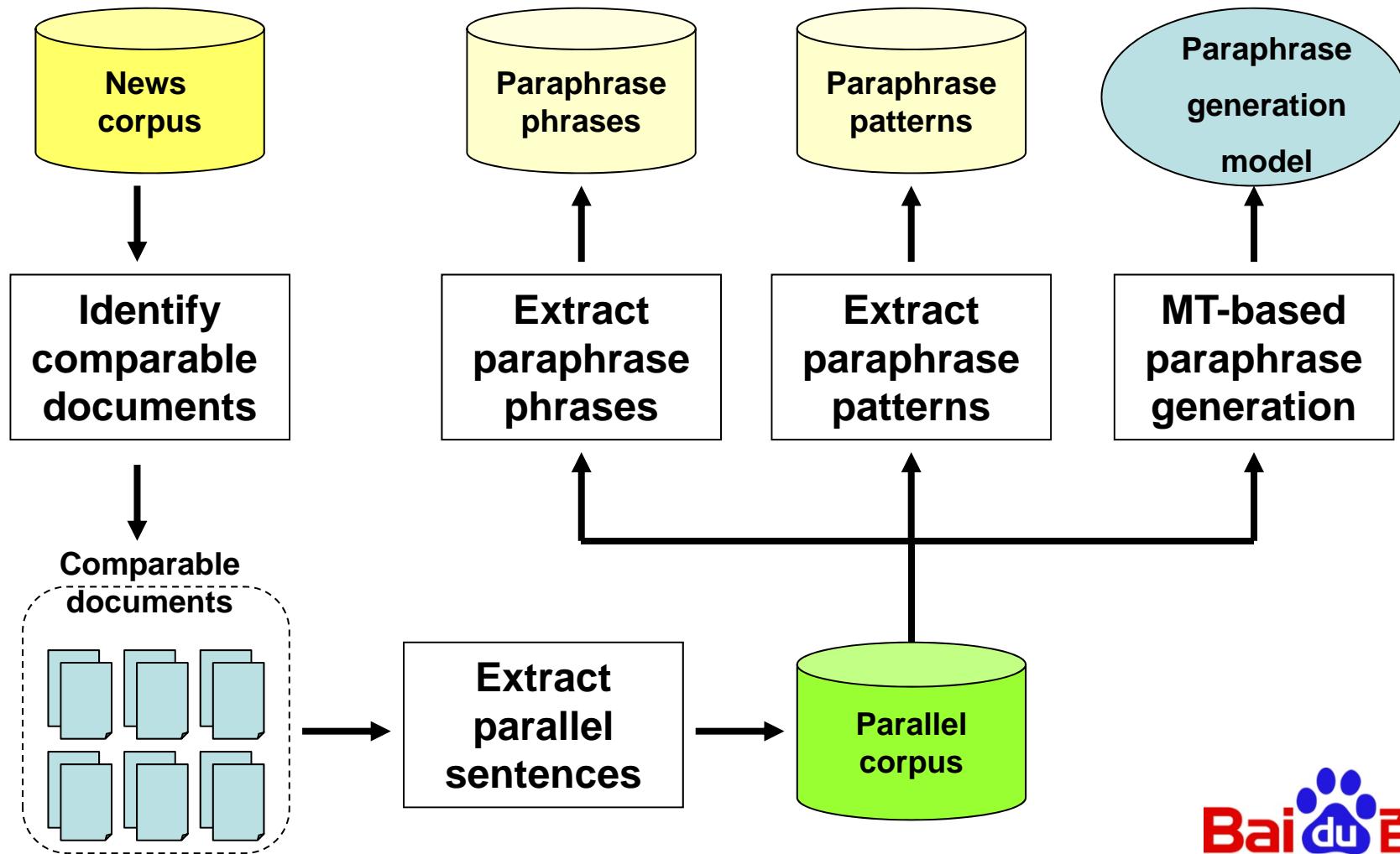
- Super Stosur proves step too far for shattered Henin 3 HOURS AGO
Justine Henin revealed she was emotionally exhausted after slipping to her first French Open defeat for six years yesterday. Four-time champion Henin surrendered a one-set lead to lose 2-6 6-1 6-4 in round four to Samantha Stosur on a stunned Suzanne...

- Match too far for weary Henin as Stosur earns shock win 3 HOURS AGO

d2

Baidu 百度

Procedure



Identify Comparable Documents

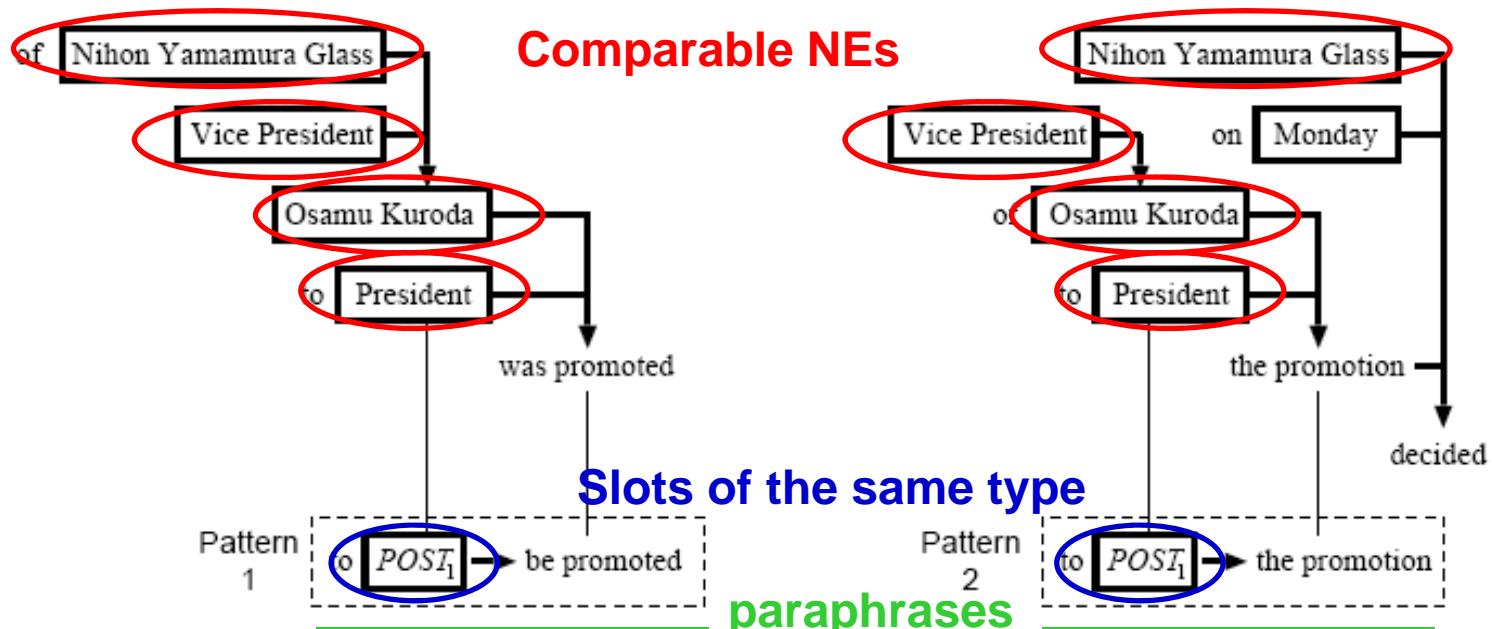
- Input
 - News articles from different news agencies
 - E.g., CNN, New York Times, Washington Post...
- Processing
 - Method-1: Retrieve documents on a given topic or event
 - Needs predefined topics or events
 - Method-2: Cluster documents
 - Content similarity; time interval
- Output
 - Corpus of comparable documents

Extract Parallel (Paraphrase) Sentences

- Input
 - Corpus of comparable documents
- Processing
 - Sentence clustering
 - Method-1: based on an assumption: first sentences of a news article usually summarize its content
 - Method-2: based on computing the content similarity
- Output
 - Corpus of parallel (paraphrase) sentences

Extract Paraphrase Patterns

- Using NEs as anchors
 - Shinyama et al., 2002
 - Basic idea: paraphrase sentences should contain comparable NEs

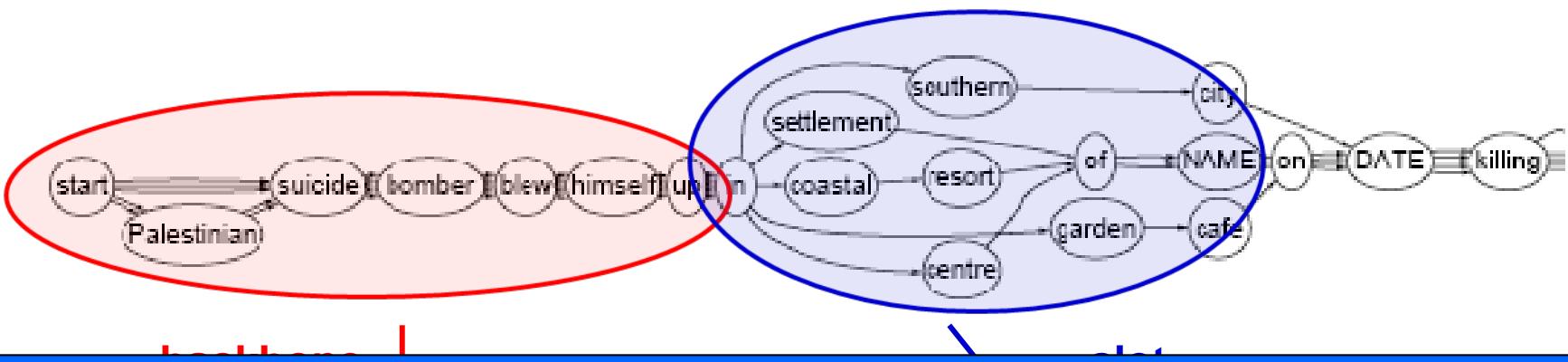


A: Vice President Osamu Kuroda of Nihon Yamamura Glass Corp. was promoted to President.

B: Nihon Yamamura Glass Corp., decided the promotion of Vice President Osamu Kuroda to President on Monday.

Extract Paraphrase Patterns

- Multiple-sequence alignment
 - Barzilay and Lee, 2003



Extracted paraphrase patterns

X (injured/wounded) Y people, Z of them seriously

Y were (wounded/hurt) by X, among them Z were in serious condition

Pros and Cons

- Pros
 - Language-independent
 - Comparable news can be found in many languages
- Cons
 - Domain-dependent
 - Paraphrases are extracted from specific domains or topics
 - Sentence clustering
 - Either too strict or too loose

Coffee Break!



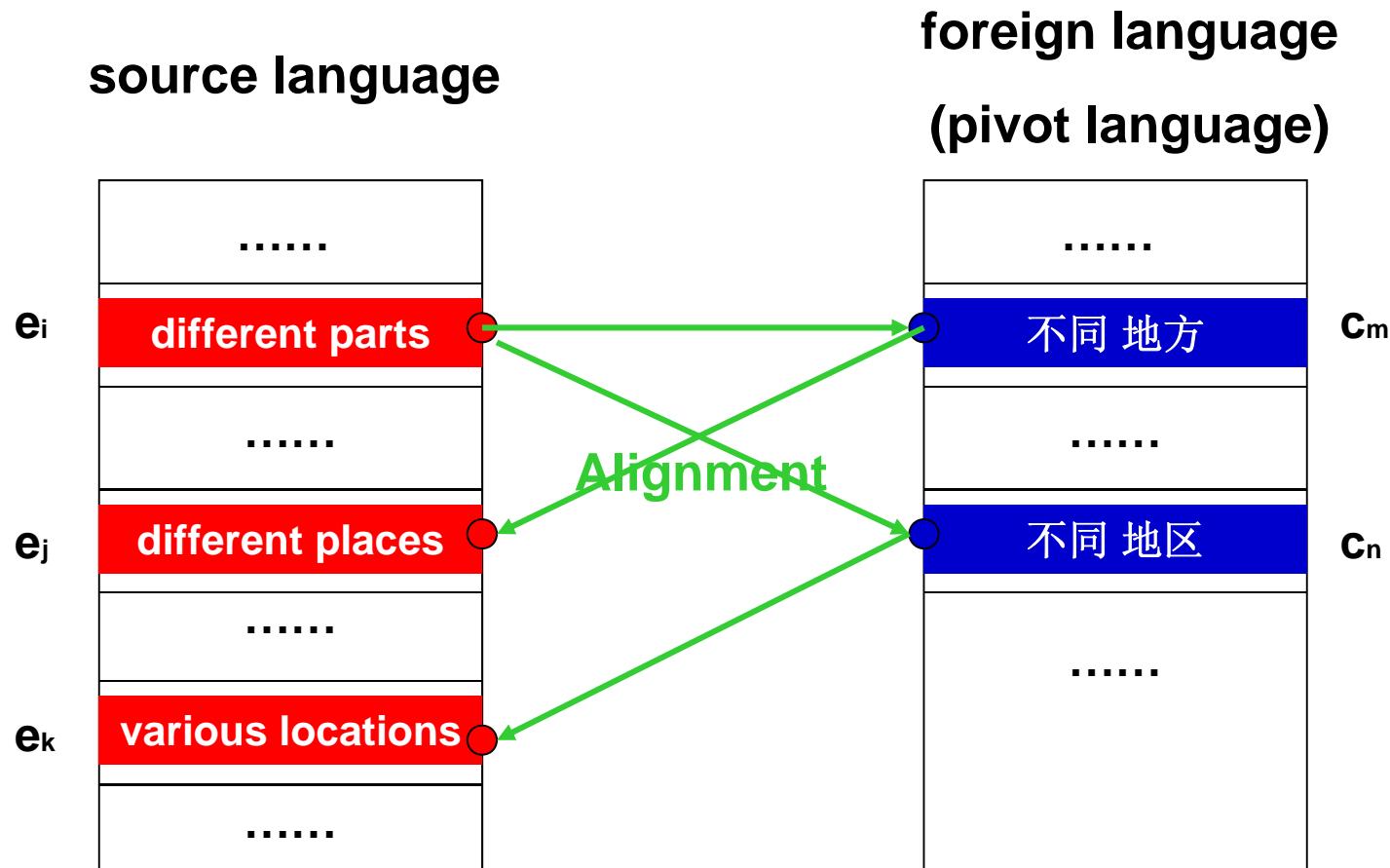
Outline

- **Part II**
 - Introduction
 - Paraphrase Identification
 - **Paraphrase Extraction**
 - From Thesauri
 - From Monolingual Parallel Corpora
 - From Monolingual Comparable Corpora
 - **From Bilingual Parallel Corpora**
 - From Large Web Corpora
 - From Other Resources

Method Overview

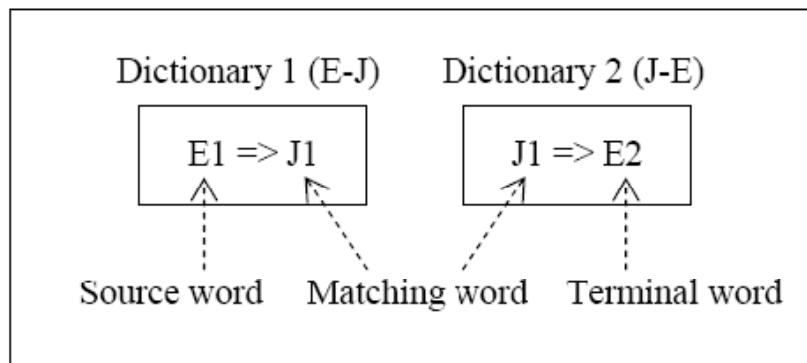
- Corpus
 - A parallel corpus of the source language and a foreign language
- Assumption
 - Multiple phrases that align with the same foreign phrase may have the same meaning
- The method is also termed as “*pivot approach*”

Example



A Simple Version

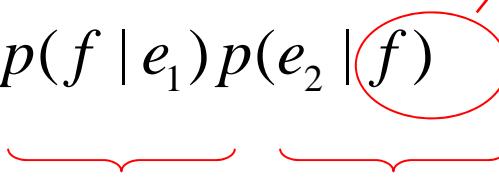
- Takao et al., 2002
 - Basic idea:
 - Generating lexical paraphrases using 2-way dictionaries
 - English word e_1 can be translated to a Japanese word j with an E-J dic. D_1 , and then j can be translated back to an English word e_2 with a J-E dictionary D_2 . e_1 and e_2 are extracted as paraphrases



Extracting Paraphrase Phrases

- Bannard and Callison-Burch, 2005
 - Word alignment and phrase extraction
 - Basic assumption:
 - If two English phrases e_1 and e_2 can be aligned with the same foreign phrase f , e_1 and e_2 are likely to be paraphrases.
 - Paraphrase probability:

$$\begin{aligned}\hat{e}_2 &= \arg \max_{e_2 \neq e_1} p(e_2 | e_1) \\ &= \arg \max_{e_2 \neq e_1} \sum_f p(f | e_1) p(e_2 | f)\end{aligned}$$



Translation probability

Pivot in a foreign language

Bannard & Callison-Burch (2005) 's results:

...should take the matter into consideration...

...应当考虑这种情况...



take the matter into consideration
take the matter into account

...must take the matter into account...

...必须考虑这种情况...



take the matter into consideration
the consideration of this matter

The consideration of this matter will...

考虑这种情况会...



take the matter into account
the consideration of this matter

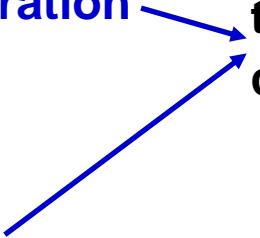
He'll take the matter into consideration

他将考虑这一问题

take the matter into consideration
consider this matter

We need to consider this matter

大家需要考虑这一问题



Add Syntactic Constraints

- Callison-Burch, 2008

- Basic idea:

- Two paraphrase phrases should have the same syntactic type.

- Paraphrase probability:

$$\begin{aligned}\hat{e}_2 &= \arg \max_{e_2: e_2 \neq e_1 \wedge s(e_2) = s(e_1)} p(e_2 | e_1, s(e_1)) \\ &= \arg \max_{e_2: e_2 \neq e_1 \wedge s(e_2) = s(e_1)} \sum_f p(f | e_1, s(e_1)) p(e_2 | f, s(e_1))\end{aligned}$$

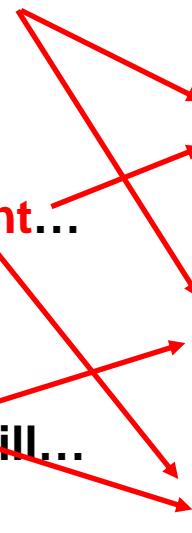
given the syntactic type

- Syntactic constraints are also used when substituting paraphrases in sentences

Callison-Burch (2008) 's results:

...should **take the matter into consideration**...

...应当**考虑这种情况**...



take the matter into consideration
take the matter into account

...must **take the matter into account**...

...必须**考虑这种情况**...



~~**take the matter into consideration**~~
~~**the consideration of this matter**~~

The consideration of this matter will...

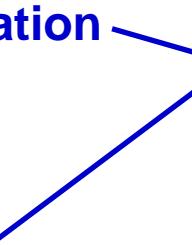
考虑这种情况会...



~~**take the matter into account**~~
~~**the consideration of this matter**~~

He'll take the matter into consideration

他将**考虑这一问题**



take the matter into consideration
consider this matter

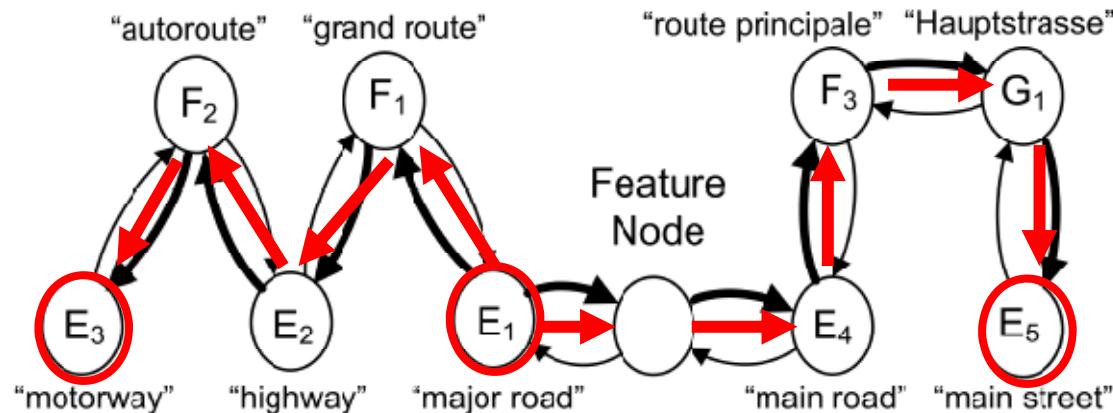
We need to consider this matter

大家需要**考虑这一问题**



Learning Paraphrases from Graphs

- Kok and Brockett, 2010
 - Basic idea:
 - Convert aligned phrases into a graph, extract paraphrases based on random walks and hitting times



Kok and Brockett (2010) 's results:

...should take the matter into consideration...

...应当考虑这种情况...



take the matter into consideration
take the matter into account

...must take the matter into account...

...必须考虑这种情况...



take the matter into account
consider this matter

The consideration of this matter will...

考虑这种情况会...

He'll take the matter into consideration

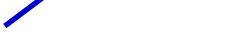
他将考虑这一问题



take the matter into consideration
consider this matter

We need to consider this matter

大家需要考虑这一问题



take the matter into account
consider this matter

Extracting Paraphrase Patterns

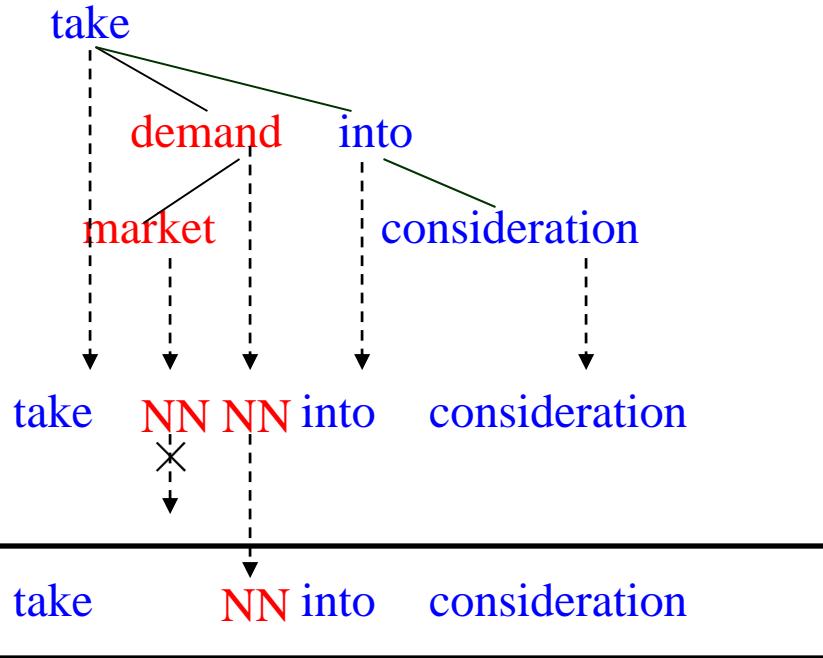
- Zhao et al., 2008
 - Basic idea:
 - Generate paraphrase patterns that include part-of-speech slots.
 - Paraphrase probability:

$$score(e_2 | e_1) = \sum_c \exp\left[\sum_{i=1}^N \lambda_i h_i(e_1, e_2, c)\right]$$

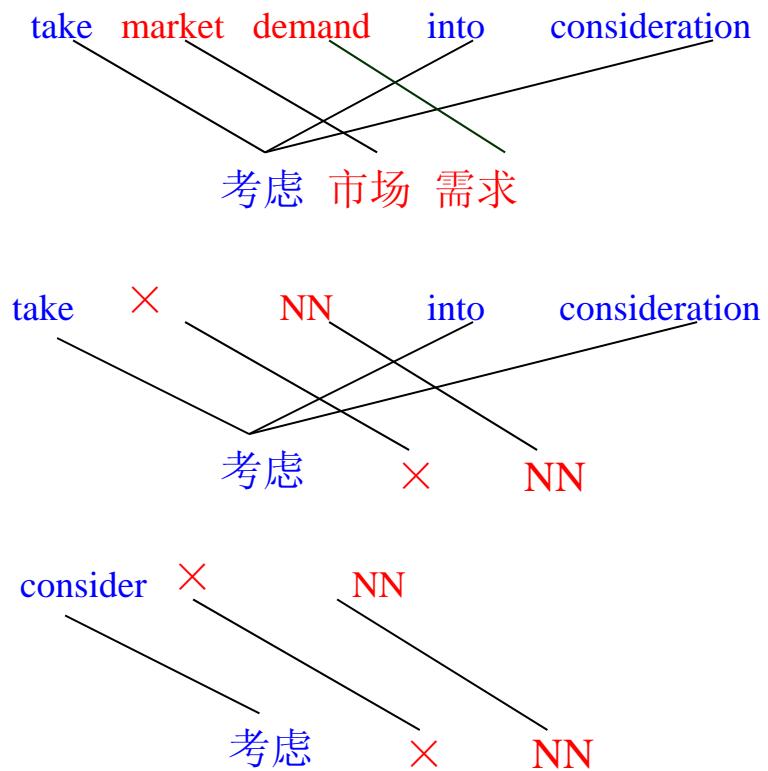
$$\begin{aligned} h_1(e_1, e_2, c) &= score_{MLE}(c | e_1) \\ h_2(e_1, e_2, c) &= score_{MLE}(e_2 | c) \\ h_3(e_1, e_2, c) &= score_{LW}(c | e_1) \\ h_4(e_1, e_2, c) &= score_{LW}(e_2 | c) \end{aligned} \quad \left. \begin{array}{l} \text{Based on maximum likelihood estimation} \\ \text{Based on lexical weighting} \end{array} \right\}$$

Example

Inducing English patterns



Inducing Chinese patterns



Extract paraphrase patterns

take NN into consideration & consider NN

Zhao et al (2008) 's results:

...should take the matter into consideration...

...应当考虑这种情况...



...must take the matter into account...

...必须考虑这种情况...



The consideration of this matter will...

考虑这种情况会...



He'll take the matter into consideration → take [NN] into consideration

他将考虑这一问题



We need to consider this matter

大家需要考虑这一问题



Pros and Cons

- Pros
 - The method proves effective, hence it's widely used
 - High precision
 - Large scale
- Cons
 - Language limitation
 - Cannot work where the large-scale bilingual parallel corpora are not available

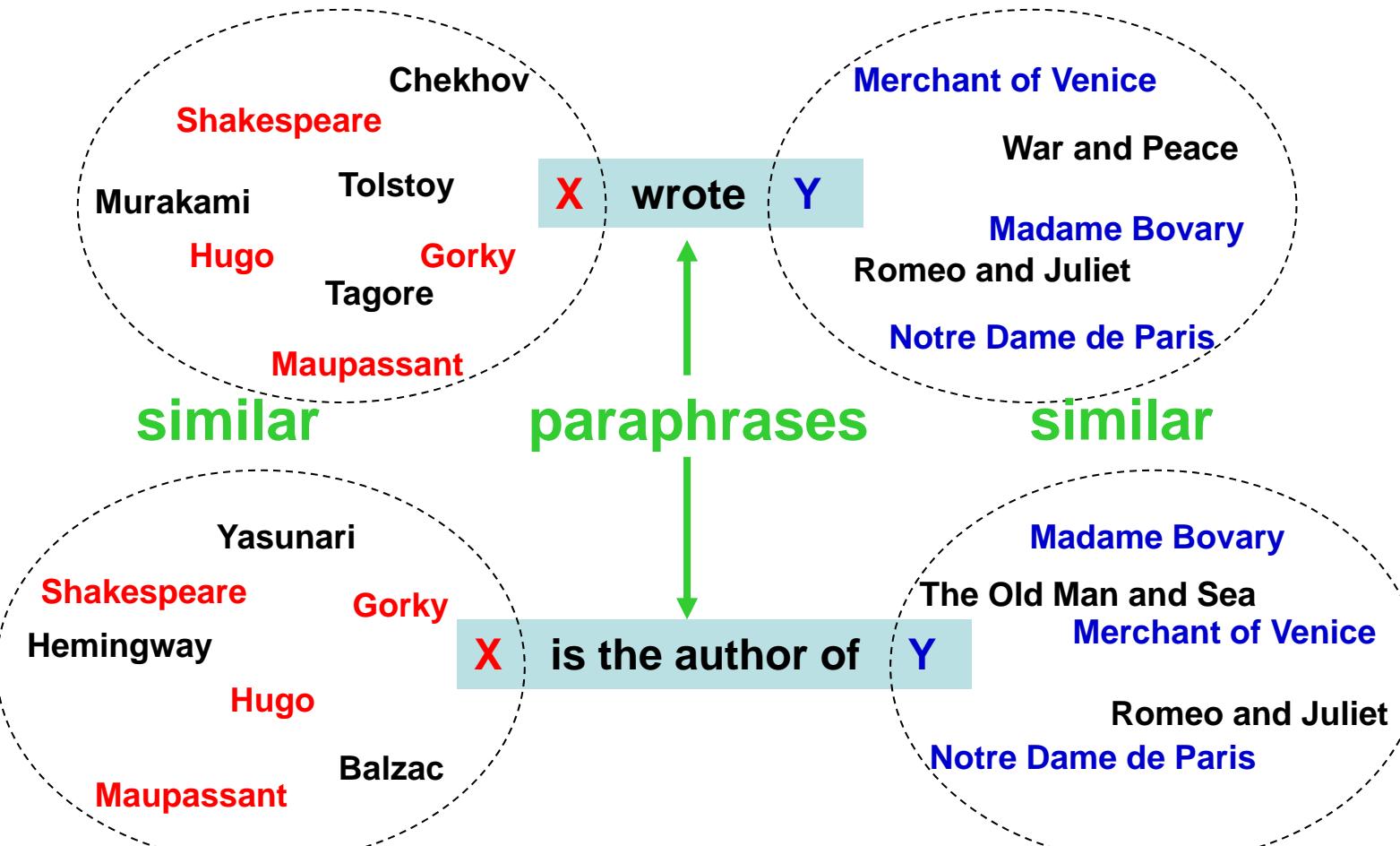
Outline

- **Part II**
 - Introduction
 - Paraphrase Identification
 - **Paraphrase Extraction**
 - From Thesauri
 - From Monolingual Parallel Corpora
 - From Monolingual Comparable Corpora
 - From Bilingual Parallel Corpora
 - **From Large Web Corpora**
 - From Other Resources

Method Overview

- Corpus
 - Large corpus of web documents
 - Or directly based on web mining
- Assumption
 - Distributional hypothesis
 - If two words / phrases / patterns often occur in similar contexts, their meanings tend to be similar

Example



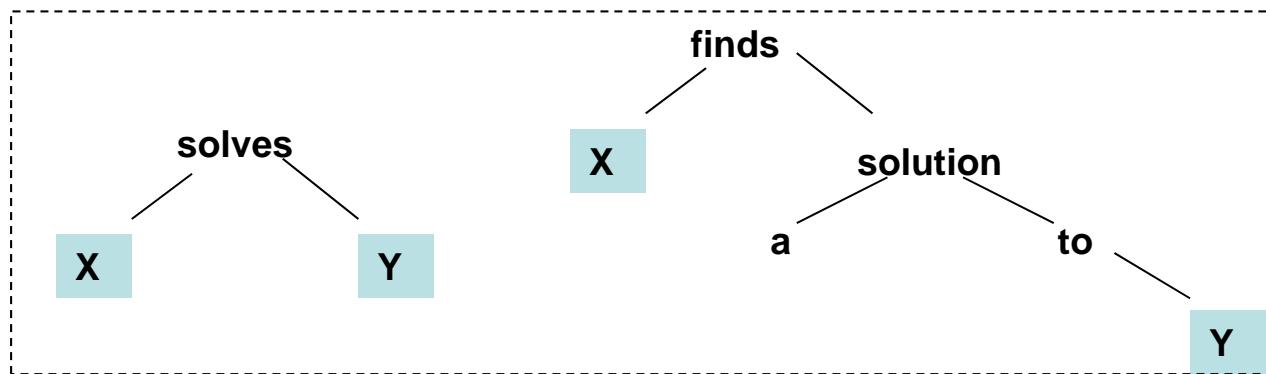
Extracting Lexical Paraphrases (Word Clustering)

- Lin, 1998
 - Basic idea
 - Measure words' similarity based on the distributional pattern of words
 - Corpus
 - A (dependency) parsed corpus
 - Word similarity

$$sim(w_1, w_2) = \frac{\sum_{(r,w) \in T_r(w_1) \cap T_r(w_2)} (I(w_1, r, w) + I(w_2, r, w))}{\sum_{(r,w) \in T_r(w_1)} I(w_1, r, w) + \sum_{(r,w) \in T_r(w_2)} I(w_2, r, w)}$$

Extracting Syntactic Paraphrase Patterns

- Lin and Pantel, 2001
 - Basic idea: extended distributional hypothesis
 - Corpus: a large corpus of parsed monolingual sentences
 - pattern pairs



- Pattern similarity

$$sim(p_1, p_2) = \sqrt{sim(SlotX_1, SlotX_2) \times sim(SlotY_1, SlotY_2)}$$

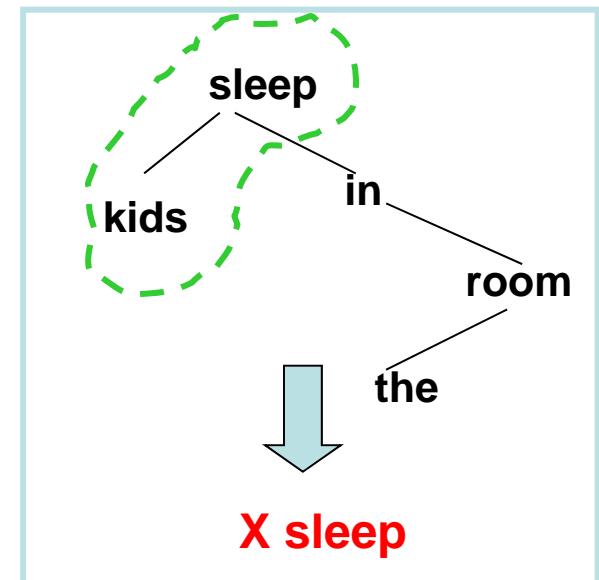
Similarity of the slot fillers

Extracting Surface Paraphrases

- Bhagat and Ravichandran, 2008
 - Basic idea is the same as the above work
 - Corpus:
 - a large corpus of monolingual sentences without parsing
 - 150GB, 25 billion words
 - Surface paraphrases
 - Pairs of n-grams
 - E.g., “*X acquired Y*” and “*X completed the acquisition of Y*”
 - Techniques
 - Apply locality sensitive hashing (LSH) to speed up the computation

Learning Unary Paraphrase Patterns

- Szpector and Dagan, 2008
 - Binary paraphrase patterns (most of the previous work)
 - Each pattern has two slots at both ends
 - E.g., “ X solves Y ” and “ X finds a solution to Y ”
 - Unary paraphrase patterns
 - Each pattern has a single slot
 - E.g., “ X take a nap” and “ X sleep”
 - Method
 - The same with the above works
 - Based on distributional hypothesis



Extracting Paraphrases based on Web Mining

- Ravichandran and Hovy, 2002
 - Basic idea
 - Learn paraphrase patterns with search engines
 - Corpus
 - The whole internet
 - Method
 - Extract paraphrase patterns for each type, e.g., “*BIRTHDAY*”
 - Provide hand-crafted seeds, e.g., “*Mozart, 1756*”
 - Retrieve sentences containing the seeds from the web with a search engine
 - Extract patterns, e.g.,
 - *born in <ANSWER>, <NAME>*
 - *<NAME> was born on <ANSWER>,*
 -

Pros and Cons

- Pros
 - Language independent
- Cons
 - For methods based on large web corpora
 - Computation complexity is high
 - Needs to process an extremely large corpus
 - Needs to compute pair-wise similarity for all candidates
 - For methods based on web mining
 - Extract paraphrase patterns type by type
 - Needs to prepare seeds beforehand

Outline

- **Part II**
 - Introduction
 - Paraphrase Identification
 - **Paraphrase Extraction**
 - From Thesauri
 - From Monolingual Parallel Corpora
 - From Monolingual Comparable Corpora
 - From Bilingual Parallel Corpora
 - From Large Web Corpora
 - **From Other Resources**

Paraphrasing with Search Engine Query Logs

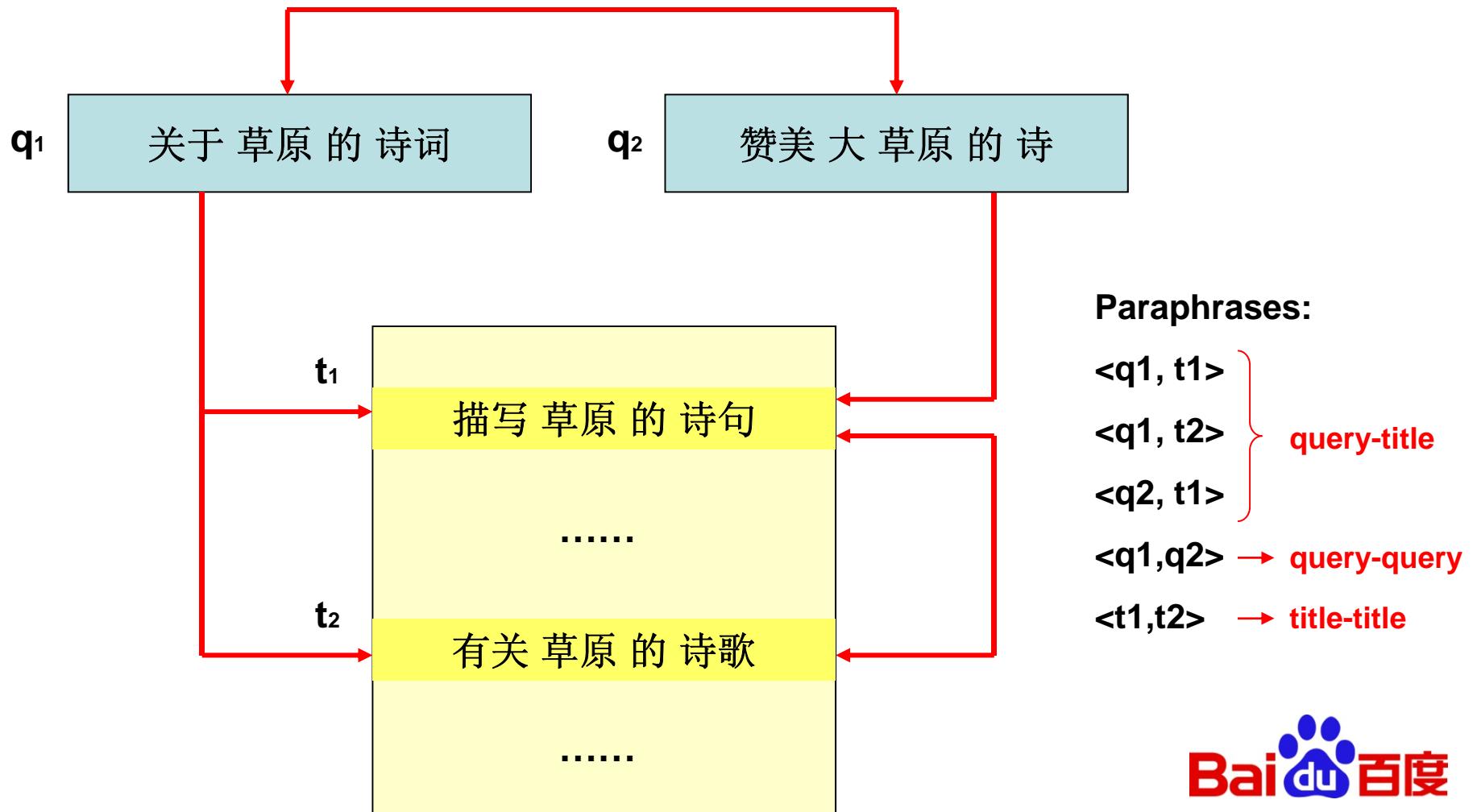
- Zhao et al., 2010
 - Corpus
 - Query logs (queries and titles) of a search engine
 - Assumption

H-1: If a query q hits a title t , then q and t are likely to be paraphrases

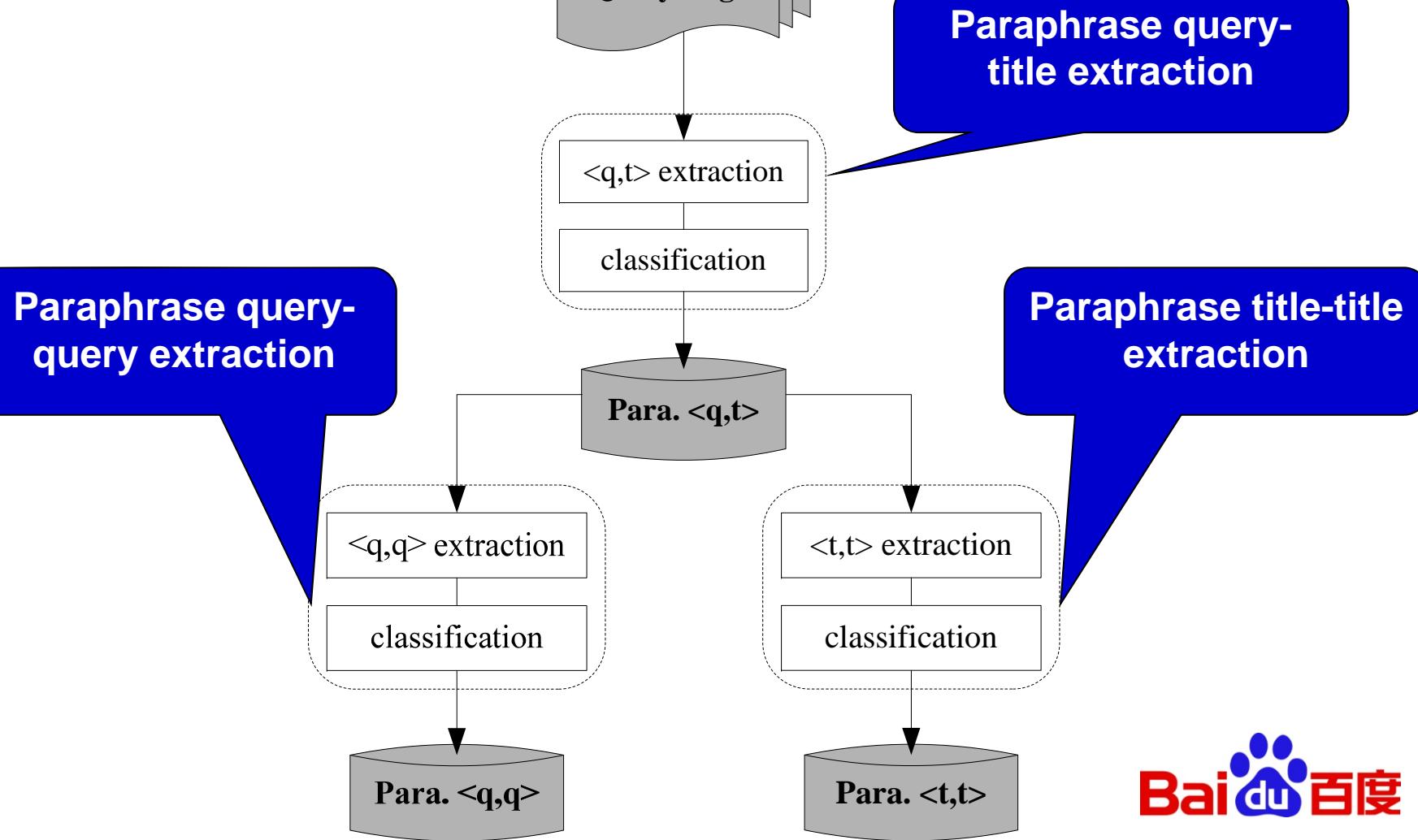
H-2: If queries q_1 and q_2 hit the same title t , then q_1 and q_2 are likely to be paraphrases

H-3: If a query q hits titles t_1 and t_2 , then t_1 and t_2 are likely to be paraphrases

Example



Method



Classification-based Paraphrase Validation

- Classification features

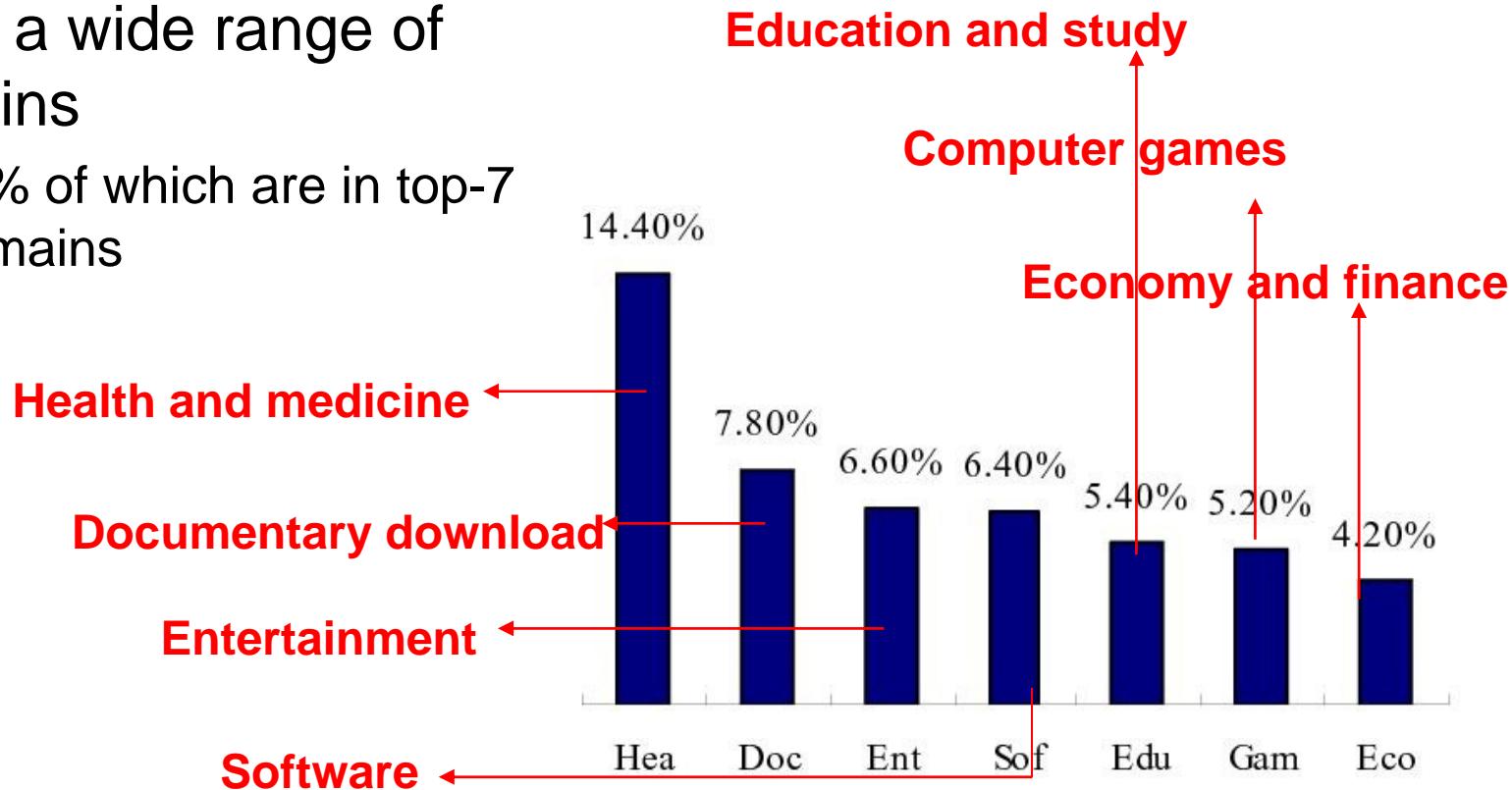
F_F	Frequency feature
F_{LR}	Length rate feature
F_{WOR}	Word overlap rate feature
F_{COR}	Character overlap rate feature
F_{Cs}	Cosine similarity feature
F_{ED}	Edit distance feature
F_{NE}	Named entity similarity feature
F_{PF}	Pivot fertility feature

most
useful

- Classifier: support vector machines (SVM)

Domains of the Extracted Paraphrases

- Extracted paraphrases cover a wide range of domains
 - 50% of which are in top-7 domains



Pros and Cons

- Pros
 - No scale limitation
 - Query logs keep growing
 - A large volume of paraphrases can be extracted
 - Query logs reflect web users' real needs
- Cons
 - Query logs data are only available in IR companies
 - User queries are noisy
 - Spelling mistakes, grammatical errors...

Extracting Paraphrases from Dictionary Glosses

- Corpus
 - Glosses of dictionaries
- Assumption
 - A word and its definition (gloss) in the dictionary have the same meaning

Example (Encarta Dictionary)

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

hurling
hurly-burly
Huron (1)
Huron (2)
Huron, Lake
hurrah
hurray
Hurrian
▶ **hurricane**
hurricane deck
hurricane lamp
hurried
hurry
hurry-come-up
hurry-scurry
hurry sickness
hurst

hurricane



hur·ri·cane [húrri káyn] (*plural*
hur·ri·canes)

noun

Definition:

1. **severe storm:** a severe tropical storm with torrential rain and extremely strong winds. Hurricanes originate in areas of low pressure in equatorial regions of the Atlantic or Caribbean, and then strengthen, traveling northwest, north, or northeast.

2. **high wind:** a wind of above 119 km (74 mi) per hour, classified as force 12 or above on the Beaufort scale

3. **fast and forceful person or thing:** somebody or something resembling a violent storm in force, speed, or effect

hurricane

severe storm

high wind

fast and forceful person or thing

Also available:

World English Dictionary
Dictionnaire Français

Baidu 百度

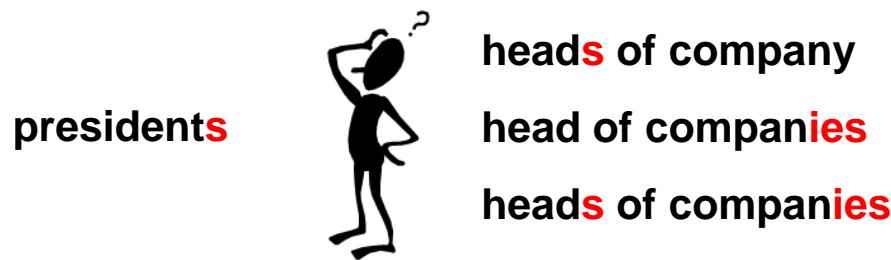
Method

- Prune and reformulate the definitions
 - For a verb v , extracts the head of the definition (h) and h 's adverb modifier m as v 's paraphrase
 - Kaji et al., 2002
 - E.g., *shout* -> *say something loudly*
 - Rule based method for extracting the appropriate part from the definition
 - Higashinaka and Nagao, 2002
 - E.g., w should not be in def ; ignore contents in parentheses in def ; avoid double negation...

Pros and Cons

- Pros
 - Explain unfamiliar words with simpler definitions
 - E.g., *amnesia* -> *memory loss*
- Cons
 - Transformation of *person*, *number*, *tense*

E.g., president → head of company



References

- From monolingual parallel corpora
 - Barzilay and McKeown. 2001. Extracting Paraphrases from a Parallel Corpus.
 - Hashimoto et al. 2011. Extracting Paraphrases from Definition Sentences on the Web.
 - Bouamor et al. 2011. Monolingual Alignment by Edit Rate Computation on Sentential Paraphrase Pairs.
- From monolingual comparable corpora
 - Yusuke Shinyama, Satoshi Sekine, Kiyoshi Sudo. 2002. Automatic Paraphrase Acquisition from News Articles.
 - Regina Barzilay and Lillian Lee. 2003. Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment.
 - Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources.

References (cont')

- From bilingual parallel corpora
 - Takao et al. 2002. Comparing and Extracting Paraphrasing Words with 2-Way Bilingual Dictionaries.
 - Bannard and Callison-Burch. 2005. Paraphrasing with Bilingual Parallel Corpora.
 - Callison-Burch. 2008. Syntactic Constraints on Paraphrases Extracted from Parallel Corpora.
 - Kok and Brockett. 2010. Hitting the Right Paraphrases in Good Time.
 - Zhao et al. 2008. Pivot Approach for Extracting Paraphrase Patterns from bilingual corpora.

References (cont')

- From large web corpora
 - Lin. 1998. Automatic Retrieval and Clustering of Similar Words.
 - Lin and Pantel. 2001. Discovery of Inference Rules for Question Answering.
 - Bhagat and Ravichandran. 2008. Large Scale Acquisition of Paraphrases for Learning Surface Patterns.
 - Szpector and Dagan. 2008. Learning Entailment Rules for Unary Templates.
 - Ravichandran and Hovy. 2002. Learning Surface Text Patterns for a Question Answering System.

References (cont')

- From other resources
 - Zhao et al. 2010. Paraphrasing with Search Engine Query Logs.
 - Kaji et al. 2002. Verb Paraphrase based on Case Frame Alignment.
 - Higashinaka and Nagao. 2002. Interactive Paraphrasing Based on Linguistic Annotation.

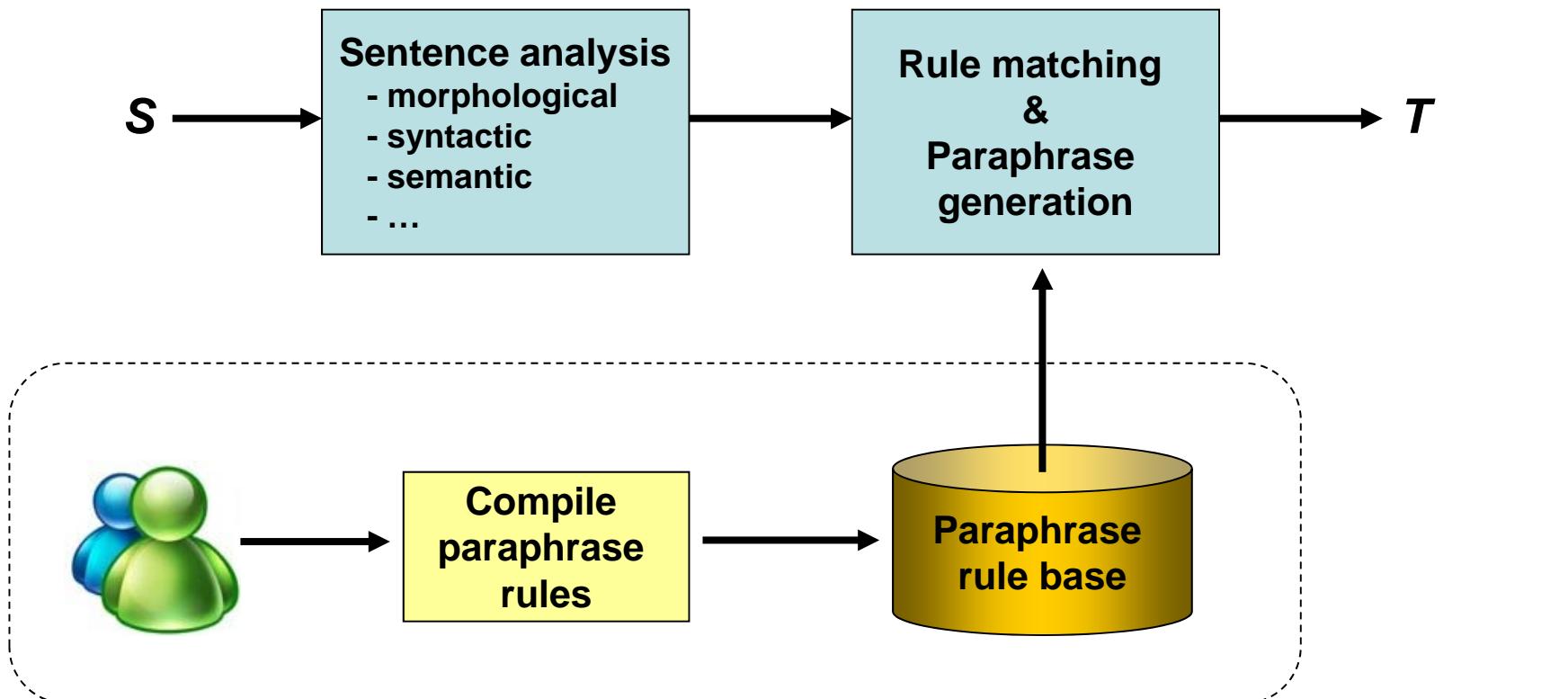
Outline

- **Part III**
 - **Paraphrase Generation**
 - Rule based Method
 - Thesaurus based Method
 - NLG based Method
 - MT based Method
 - Pivot based Method
 - Applications of Paraphrases
 - Evaluation of Paraphrases
 - Conclusions and Future work

Rule based Method

- Two types:
 - Based on hand-crafted rules
 - Widely used in early studies of paraphrase generation
 - McKeown, 1979; Zong et al., 2001; Tetsuro et al., 2001; Zhang and Yamamoto, 2002.....
 - Based on automatically extracted rules
 - Extract paraphrase patterns from corpora
 - Barzilay and Lee, 2003, Zhao et al., 2009a.....

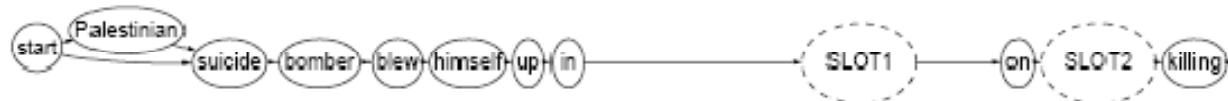
Based on Hand-crafted Rules



Based on Hand-crafted Rules

- Examples of paraphrase rules
 - Change the positions of adverbials
 - *He booked a single room in Beijing **yesterday**.* =>
 - **Yesterday**, he booked a single room in Beijing.
 - Split a compound sentence into a group of simple sentences
 - *He booked a single room in Beijing yesterday* =>
 - *He booked a single room in Beijing.*
 - *He booked a single room yesterday.*
 - *He booked a room.*
 - Rewrite a sentence using hand-crafted patterns
 - **Can I have a cup of tea?** =>
 - **May** I have a cup of tea?
 - **I would like** a cup of tea, please.
 - **Give me** a cup of tea.

Based on Automatically Extracted Rules

- Studies on paraphrase patterns extraction has been introduced above
- Some of them have tried to apply the extracted paraphrase patterns in paraphrase generation
 - Complex paraphrase patterns
 - Barzilay and Lee, 2003
 - E.g.,

 - Short and simple paraphrase patterns
 - Zhao et al., 2009a
 - E.g., *consider [NN]* and *take [NN] into consideration*

Based on Automatically Extracted Rules (cont.)

- A generate and rank approach for sentence paraphrasing
 - Malakasiotis and Androutsopoulos, EMNLP-2011
 - A two-stage approach
 - Generate
 - Generate candidate paraphrases with paraphrase patterns extracted with a pivot-approach (Zhao et al., 2009b)
 - Rank
 - Rank candidates with an SVR ranker
 - Features include: language model, patterns' paraphrasing probabilities, kinds of similarity measurements

Pros and Cons

- Methods based on hand-crafted rules
 - Pros
 - Can design paraphrase rules for specific applications and requirements
 - Cons
 - It is time-consuming to construct paraphrase rules
 - Problem of rules conflict
 - Coverage of paraphrase rules is limited
- Methods based on automatically extracted rules
 - Pros
 - Can generate paraphrases with structural changes
 - Cons
 - Coverage of paraphrase rules is limited

References

- McKeown. 1979. Paraphrasing Using Given and New Information in a Question-Answer System.
- Zong et al. 2001. Approach to Spoken Chinese Paraphrasing Based on Feature Extraction.
- Tetsuro et al.. 2001. KURA: A Transfer-Based Lexico-Structural Paraphrasing Engine.
- Zhang and Yamamoto. 2002. Paraphrasing of Chinese Utterances.
- Barzilay and Lee. 2003. Learning to Paraphrase - An Unsupervised Approach Using Multiple-Sequence Alignment.
- Zhao et al. 2009a. Application-driven Statistic Paraphrase Generation.
- Malakasiotis and Androutsopoulos. 2011. A Generate and Rank Approach to Sentence Paraphrasing.
- Zhao et al. 2009b. Extracting Paraphrase Patterns from Bilingual Parallel Corpora.

Outline

- Part III
 - Paraphrase Generation
 - Rule based Method
 - **Thesaurus based Method**
 - NLG based Method
 - MT based Method
 - Pivot based Method
 - Applications of Paraphrases
 - Evaluation of Paraphrases
 - Conclusions and Future work

Thesaurus based Method

- Also known as lexical substitution
 - Substitute words in a sentence with their synonyms that fit in the given context
 - SemEval-2007: English lexical substitution task
 - SemEval-2010: Cross-lingual lexical substitution
 - Example:
 - *There will be major cuts in the salaries of high-level civil servants.*
 - *There will be major cuts in the wages of high-level civil servants.*

Thesaurus based Method

- Include two stages
 - **Stage-1:** extract candidate substitutes from predefined inventories.
 - E.g., WordNet
 - **Stage-2:** find substitutes that fit in the given context
 - Using language model or web data (e.g., Google 5-gram) for evaluating the fitness in the context
 - Disambiguation may also be useful

Stage-1: Candidate Extraction

- Various thesauri have been tried
 - WordNet:
 - the most commonly used
 - Others:
 - Encarta, Roget, Oxford American Writer's Thesaurus...
- Extracting different information as candidates
 - Synsets (all synsets vs. best synset)
 - Hypernyms, similar-to, also-see...
 - Words in glosses

Example:

WordNet

different
synsets

WordNet 2.1 Browser

File History Options Help

Search Word: bright Redisplay Overview

Searches for bright: Adjective Adverb Senses:

11 senses of bright

Sense 1

bright (vs. dull) -- (emitting or reflecting light readily or in large amounts; "the sun was bright and hot"; "bright sunlit room")

=> agleam, gleaming, nitid -- (bright with a steady but subdued shining; "from the plane we saw the city below agleam with lights"; "the gleaming brass on the altar"; "Nereids beneath the nitid moon")

=> aglow(predicate), lambent, lucent, luminous -- (softly bright or radiant; "a house aglow with lights"; "glowing embers"; "lambent tongues of flame"; "the lucent moon"; "a sky luminous with stars")

=> aglitter(predicate), coruscant, fulgid, glinting, glistening, glittering, glittery, scintillant, scintillating, sparkly -- (having brief brilliant points or flashes of light; "bugle beads all aglitter"; "glinting eyes"; "glinting water"; "his glittering eyes were cold and malevolent"; "shop window full of glittering Christmas trees"; "glittery costume jewelry"; "scintillant mica"; "the scintillating stars"; "a dress with sparkly sequins"; "'glistening' is an archaic term")

=> beady, beadlike, buttony, buttonlike -- (small and round and shiny like a shiny bead or button; "bright beady eyes"; "black buttony eyes")

=> beaming, beam, effulgent, radiant, refulgent -- (radiating or as if radiating light; "the beaming sun"; "the effulgent daffodils"; "a radiant sunrise"; "a refulgent sunset")

=> blazing, blinding, dazzling, fulgent, glaring, glary -- (shining intensely; "the blazing sun"; "blinding headlights"; "dazzling snow"; "fulgent patterns of sunlight"; "the glaring sun")

=> bright as a new penny(predicate) -- ((metaphor) shining brightly)

"Synonyms/Related Nouns" search for adjective "bright"

Example:

Encarta

Dictionary Thesaurus Translations

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

bright (adj)

Synonyms: brilliant, vivid, intense,
dazzling, light, clear

Antonym: dark

Synonyms: intelligent, clever, smart,
brainy, quick, sharp-witted

Antonym: unintelligent

Synonyms: cheerful, happy, lively,
optimistic, positive, upbeat, sunny,
perky

Antonym: gloomy

Up Arrow
Down Arrow

definition of the synset

synset

Stage-2: Substitute Selection

- Rank the candidates and select the one fits best in the given context
- Context constraints
 - Semantic constraints
 - Select substitutes with the correct meaning wrt the given context
 - Syntactic constraints
 - The sentence generated after substitution should keep grammatical

SubFinder: A Lexical Substitution System

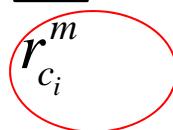
- SubFinder
 - University of North Texas
 - Performs well in SemEval-2007 English lexical substitution task
- Candidate extraction
 - WordNet
 - Encarta
 - Others
 - Prove to be useless

SubFinder: A Lexical Substitution System

- Substitute selection (**5 ranking methods R1~R5**)
 - Language model (**R1**)
 - Google 1T 5-gram
 - Information Retrieval (**R2**)
 - Search on the web using a web search engine
 - Latent semantic analysis (LSA) (**R3**)
 - Rank a candidate by its relatedness to the context sentence
 - Word sense disambiguation (WSD) (**R4**)
 - Disambiguate the target word and select the synset of the right sense
 - Pivot approach (**R5**)
 - Check whether a candidate substitute can be generated via a 2-way translation

SubFinder: A Lexical Substitution System

- Combine R1~R5:
 - Voting mechanism

$$score(c_i) = \sum_{m \in rankings} \lambda_m \frac{1}{r_{c_i}^m}$$


Ranks according to R1-R5

- Contribution of each ranking method is not analyzed ☹

Pros and Cons

- Pros
 - Based on existing inventories
- Cons
 - Cannot generate structural paraphrases
 - Language limitation
- Question
 - *How to merge different thesauri?*
 - Thesauri have different forms of synset clustering

References

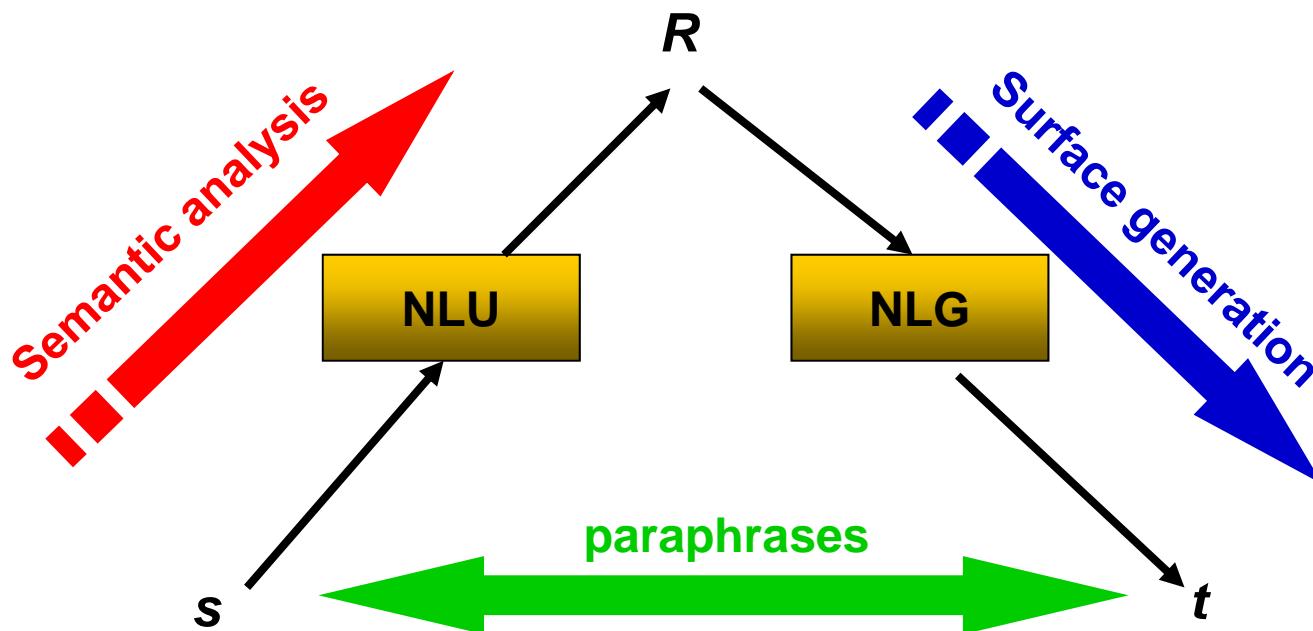
- McCarthy and Navigli. 2007. SemEval-2007 Task 10: English Lexical Substitution Task.
- Hassan et al. 2007. UNT: SubFinder: Combining Knowledge Sources for Automatic Lexical Substitution.
- Yuret. 2007. KU: Word Sense Disambiguation by Substitution.
- Giuliano et al. 2007. FBK-irst: Lexical Substitution Task Exploiting Domain and Syntagmatic Coherence.
- Martinez et al. 2007. MELB-MKB: Lexical Substitution System based on Relatives in Context.
- Kauchak and Barzilay. 2006. Paraphrasing for Automatic Evaluation.

Outline

- Part III
 - Paraphrase Generation
 - Rule based Method
 - Thesaurus based Method
 - **NLG based Method**
 - MT based Method
 - Pivot based Method
 - Applications of Paraphrases
 - Evaluation of Paraphrases
 - Conclusions and Future work

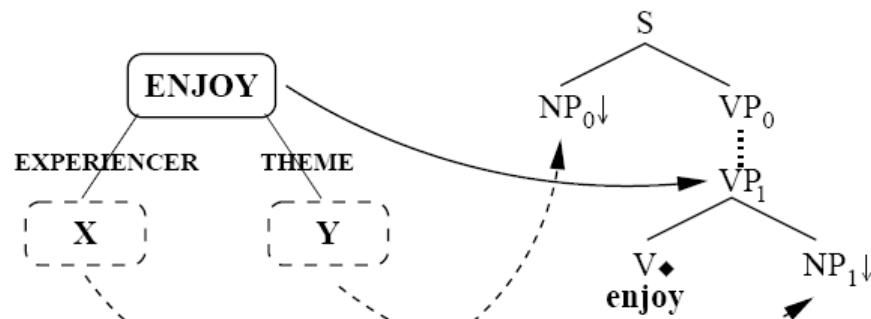
Overview

- Two steps
 - (1) analysis and (2) generation



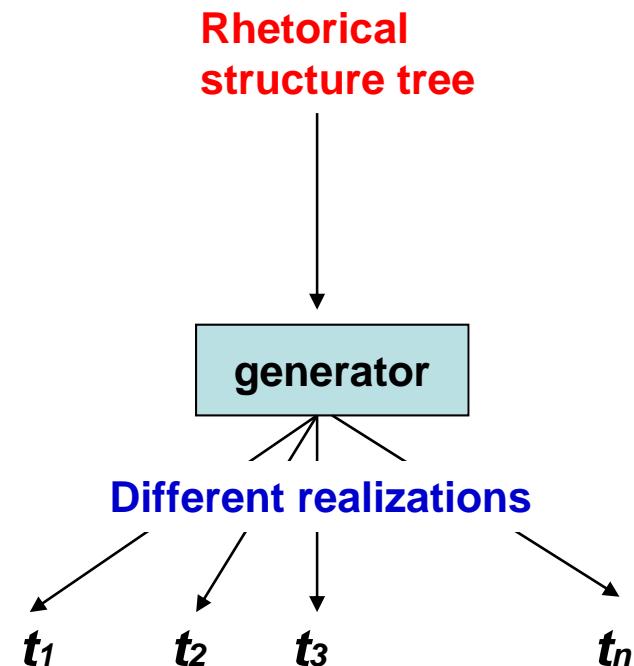
NLG based Methods

- Kozlowski et al., 2003
 - Generate single-sentence paraphrases
 - Input: predicate/argument structure
 - Not natural language sentences ☹
 - Based on lexico-grammatical resource
 - Map elementary semantic structures with syntactic realization



NLG based Methods (cont')

- Power and Scott, 2005
 - Concerning *larger-scale* paraphrases
 - Paraphrases of multiple sentences or even the whole text
 - Paraphrases vary not only at lexical and syntactic levels, but also in document structure and layout
 - Problem:
 - The input is not natural language texts ☹



NLG based Methods (cont')

- Power and Scott, 2005 (cont')
 - Example:

Rhetorical
structure tree

reason
NUCLEUS: recommend(doctors, elixir)
SATELLITE: conjunction
1: quick-results(elixir)
2: few-side-effects(elixir)

solution1

Doctors recommend Elixir since it gives quick results and it has few side effects.

solution2

(1) Elixir gives quick results.
(2) Elixir has few side effects.
(3) Therefore, it is recommended by doctors.

NLG based Methods (cont')

- Fujita et al., 2005
 - Paraphrase *light-verb constructions* (LVC) in sentences
 - LVC: a light-verb that syntactically governs a noun
 - E.g., “give + impression”
 - Semantic representation
 - LCS: Lexical Conceptual Structure
 - Procedure
 - Semantic analysis
 - Semantic transformation
 - Surface generation

Pros and Cons

- Pros
 - It simulates human being's behavior when generating paraphrases:
 - Step-1: understand the meaning of a sentence
 - Step-2: generate a new sentence expressing the meaning
- Cons
 - Both deep analysis of sentences and NLG are difficult to realize

References

- Kozlowski et al. 2003. Generation of single-sentence paraphrases from predicate/argument structure using lexico-grammatical resources.
- Power and Scott. 2005. Automatic generation of large-scale paraphrases.
- Fujita et al. 2005. Exploiting Lexical Conceptual Structure for Paraphrase Generation.

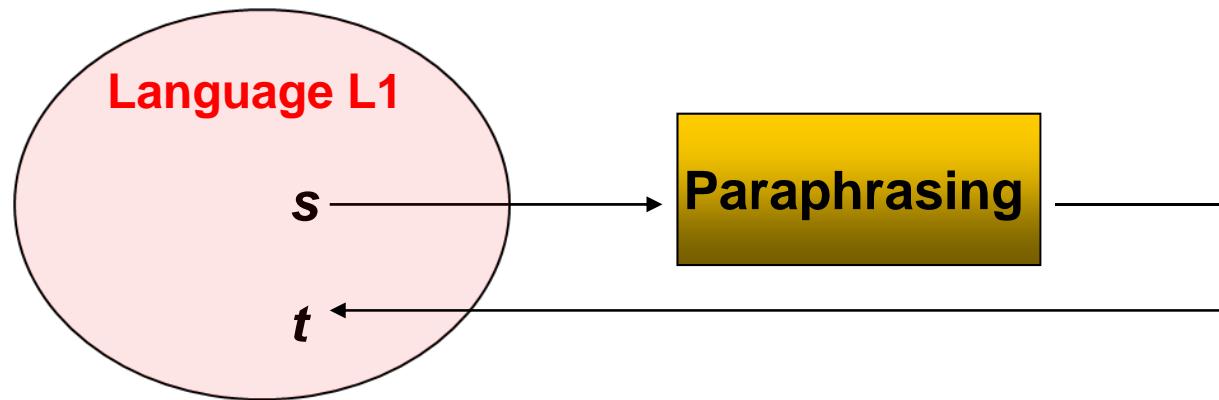
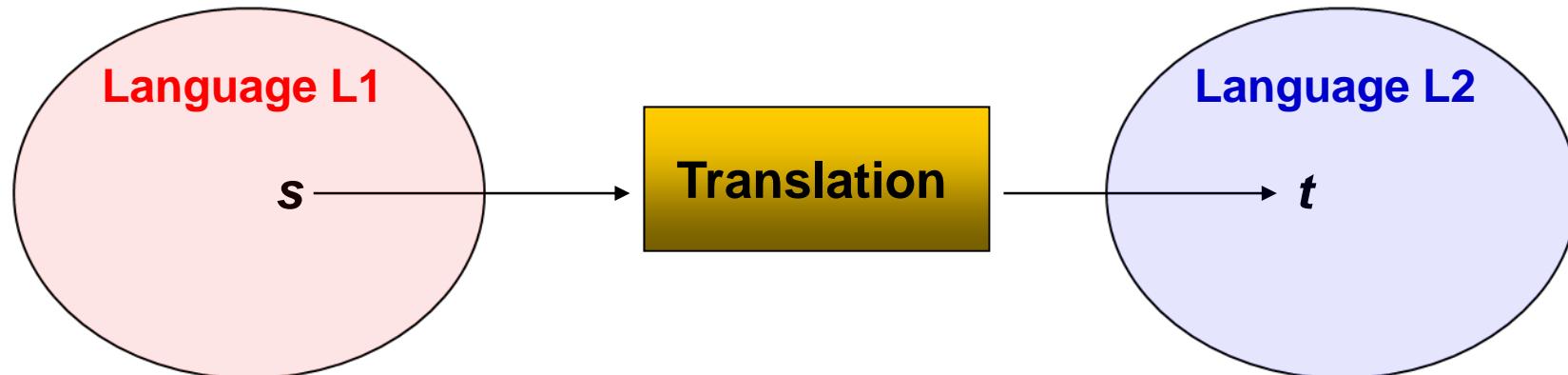
Lunch Time!



Outline

- Part III
 - Paraphrase Generation
 - Rule based Method
 - Thesaurus based Method
 - NLG based Method
 - **MT based Method**
 - Pivot based Method
 - Applications of Paraphrases
 - Evaluation of Paraphrases
 - Conclusions and Future work

Machine Translation vs. Paraphrase Generation

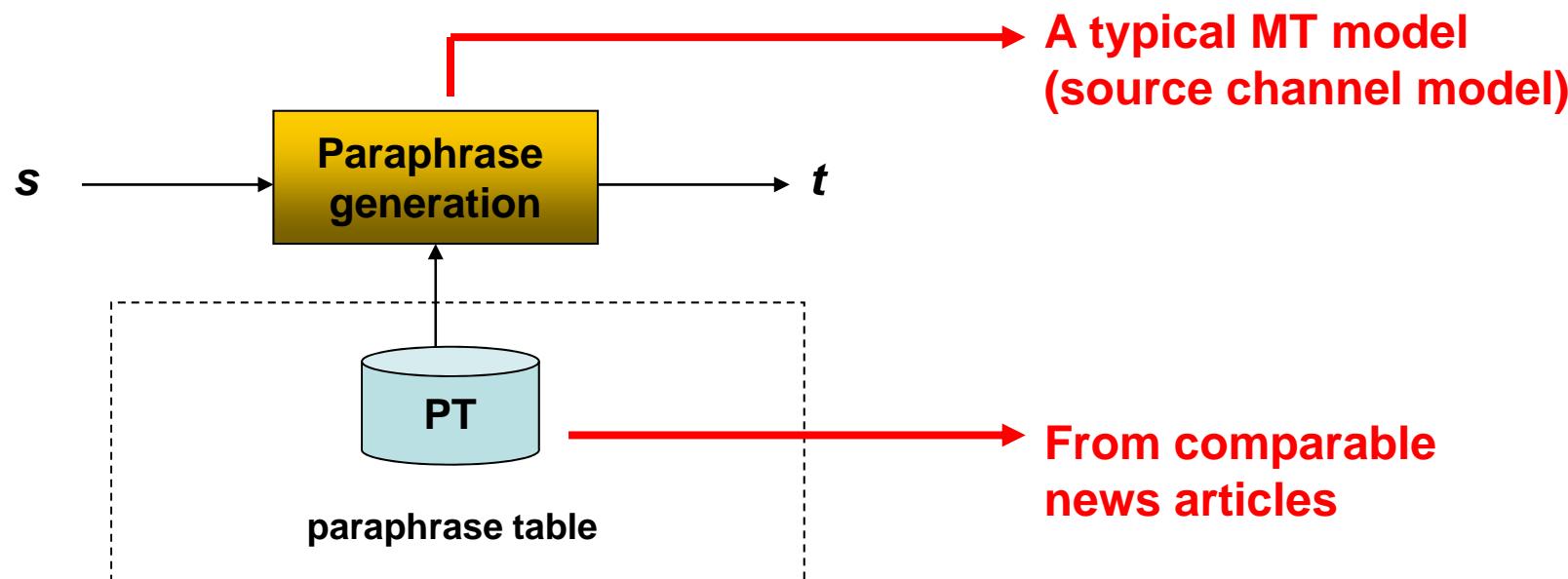


For both machine translation and paraphrase generation:

- (1) t should preserve the meaning of s
- (2) t should be a fluent sentence

Paraphrase Generation as Machine Translation

- Quirk et al., 2004
 - First recast paraphrase generation as a monolingual machine translation task



Paraphrase Generation as Machine Translation (cont')

- Model
 - Source channel model

$$t^* = \arg \max_t p(t | s)$$

$$= \arg \max_t p(s | t) p(t)$$

→ Language model



“Translation” model

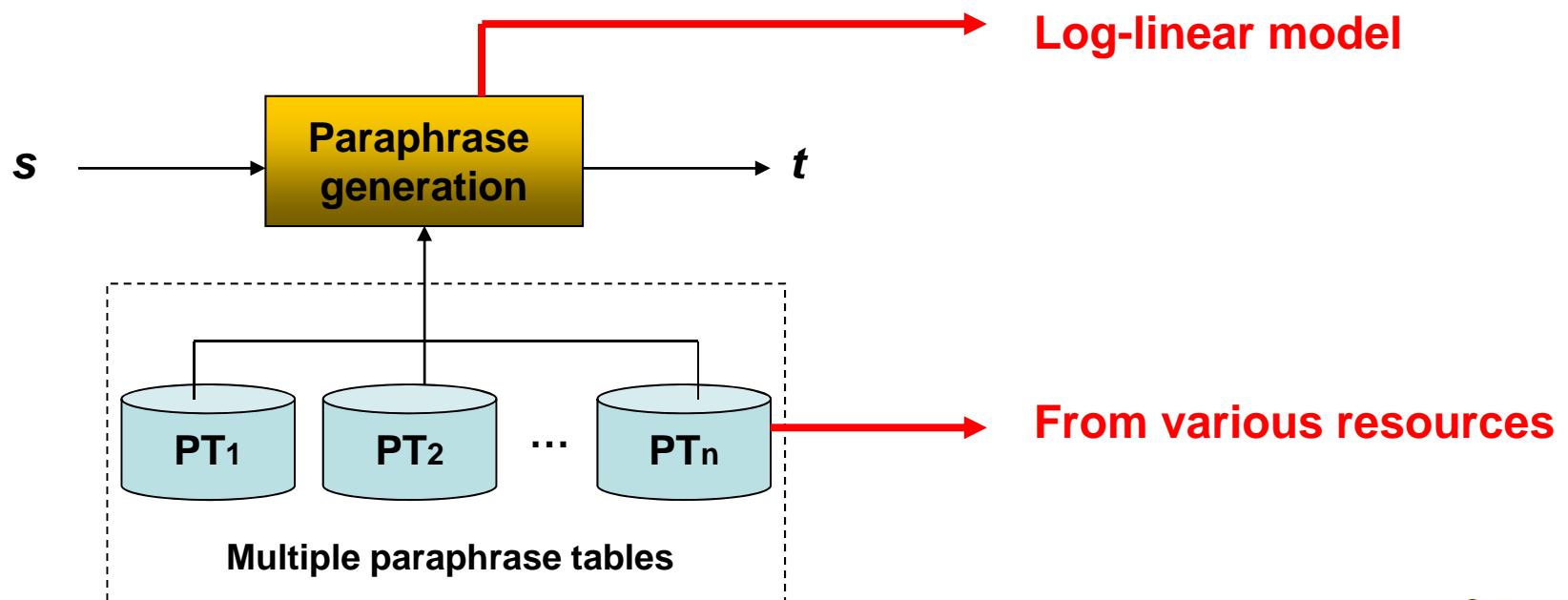
(based on a phrasal
paraphrase table)

Paraphrase Generation as Machine Translation (cont')

- Paraphrase table
 - Monolingual parallel sentences
 - Extracted from comparable news articles
 - 139K pairs
 - Word alignment & phrase pair extraction
 - With Giza++
- Limitation
 - Lack of monolingual parallel corpora to train the paraphrase table!!!

Paraphrase Generation as Machine Translation (cont')

- Zhao et al., 2008
 - Combine multiple resources to improve paraphrase generation



Paraphrase Generation as Machine Translation (cont')

- Model
 - Log-linear model

$$t^* = \arg \max_t \left\{ \sum_{i=1}^N \lambda_{TM_i} h_{TM_i}(t, s) + \lambda_{LM} h_{LM}(t, s) \right\}$$

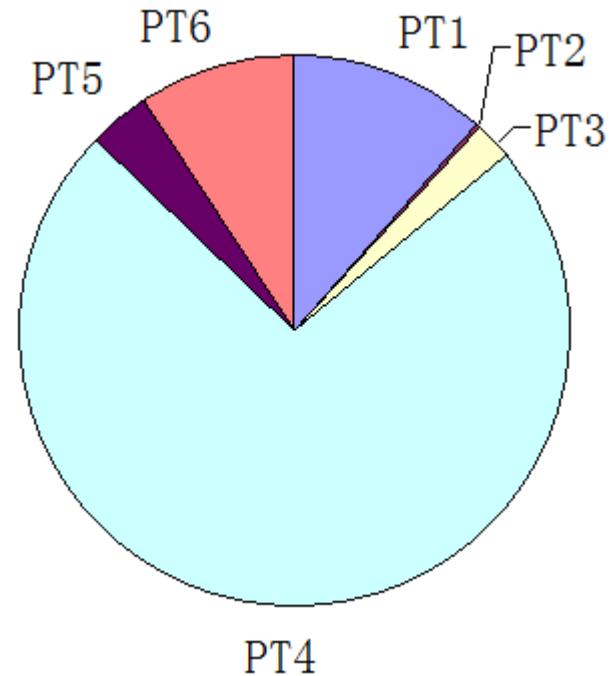
N paraphrase tables, each feature corresponds to a paraphrase table

Language model

$$h_{TM_i}(t, s) = \log \prod_{k=1}^{K_i} score_i(t_k, s_k)$$

Paraphrase Generation as Machine Translation (cont')

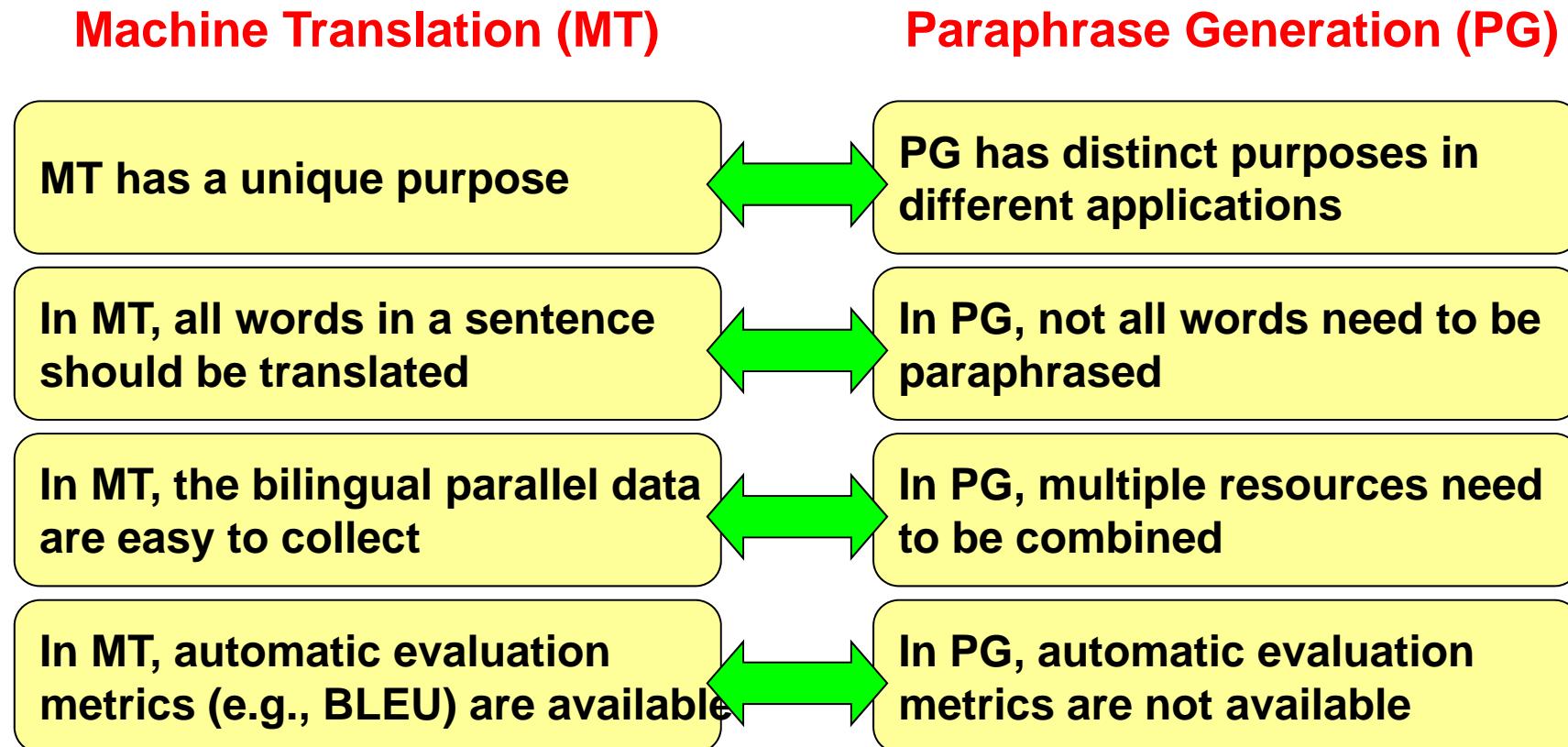
- Paraphrase tables
 - PT1: from word clusters (Lin, 1998)
 - PT2: from monolingual parallel corpora
 - PT3: from monolingual comparable corpora
 - PT4: from bilingual parallel corpora
 - PT5: from Encarta dictionary glosses
 - PT6: from clusters of similar user queries
- Volumes of the PTs:



Proves most useful!

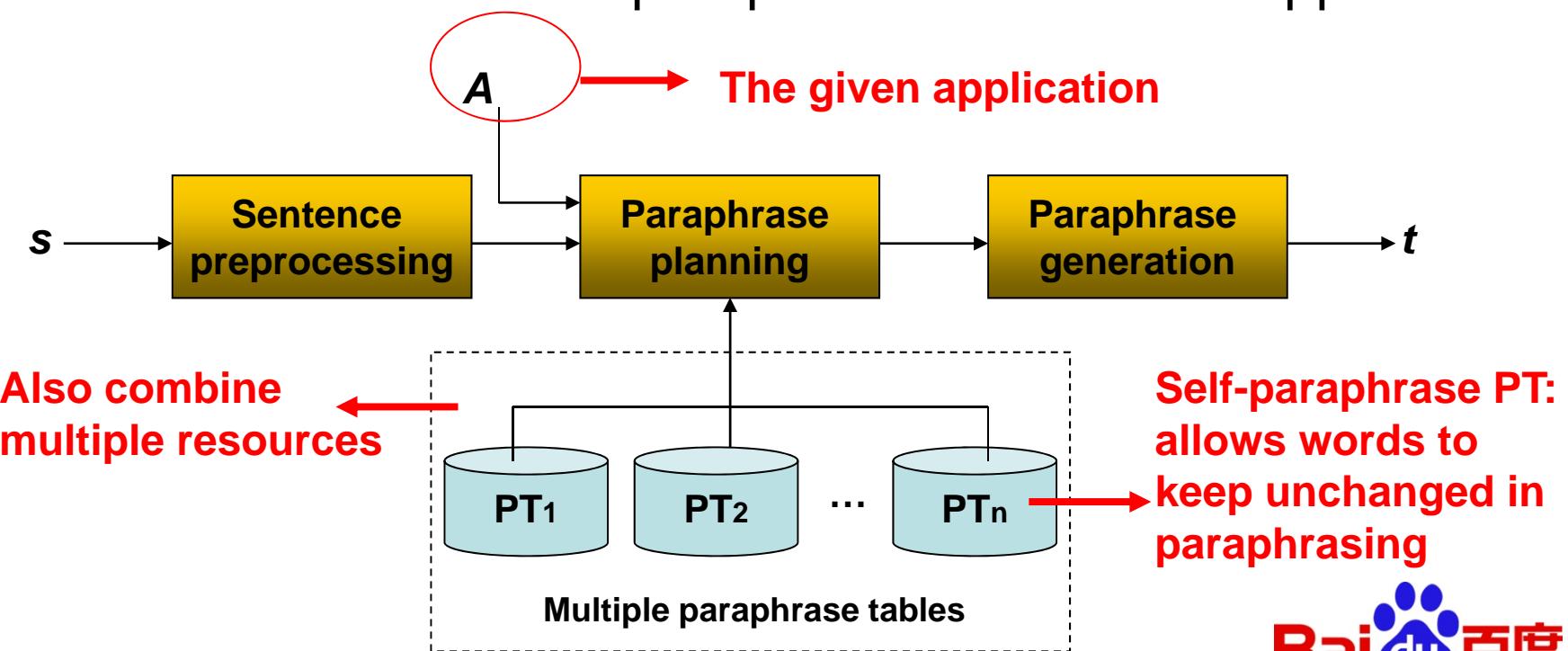
Paraphrase Generation vs. Machine Translation

- Differences between machine translation and paraphrase generation (Zhao et al., 2009):



Application-driven Statistical Paraphrase Generation

- Zhao et al., 2009
 - Propose a statistical model for paraphrase generation
 - Generate different paraphrases in different applications



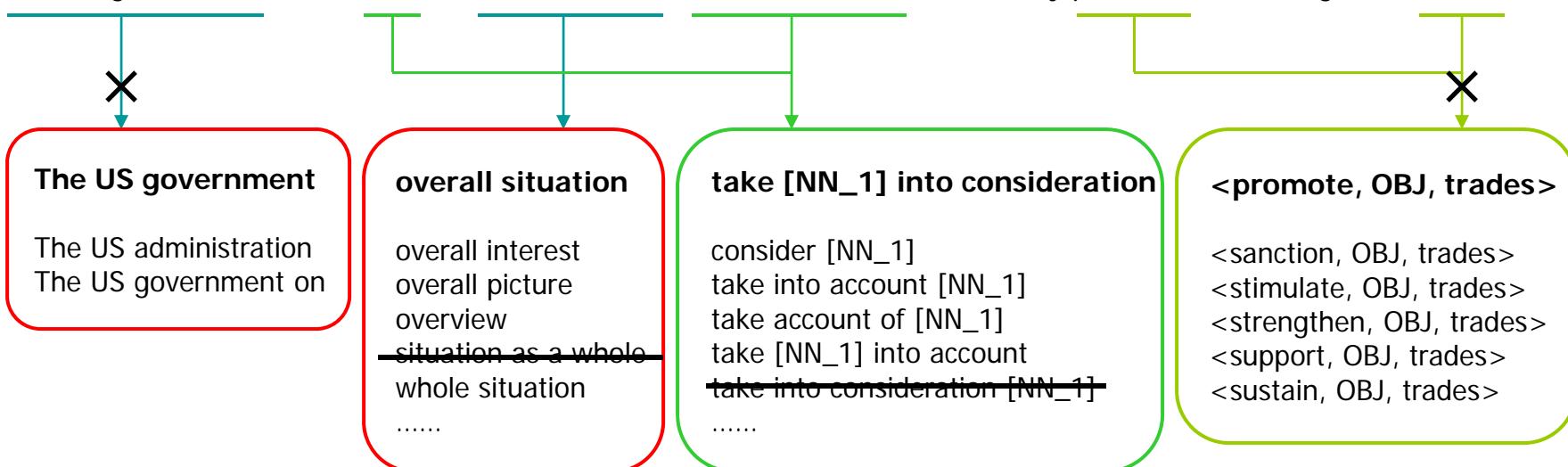
Application-driven Statistical Paraphrase Generation (cont')

- Paraphrase planning
 - When an application *A* is given, only the paraphrase pairs that can achieve *A* are kept

Example:

Paraphrase application: sentence compression

The US government should take the overall situation into consideration and actively promote bilateral high-tech trades.



Application-driven Statistical Paraphrase Generation (cont')

- Model:
 - Log-linear model

$$p(\mathbf{t} \mid \mathbf{s}) = \sum_{k=1}^K (\lambda_k \sum_{k_i} \log \phi_k(\bar{s}_{k_i}, \bar{t}_{k_i})) \rightarrow \text{Paraphrase model}$$
$$+ \lambda_{lm} \sum_{j=1}^J \log p(t_j \mid t_{j-2} t_{j-1}) \rightarrow \text{Language model}$$
$$+ \lambda_{um} \sum_{i=1}^I \mu(\bar{s}_i, \bar{t}_i) \rightarrow \text{Usability model
(defined for each application)}$$

Application-driven Statistical Paraphrase Generation (cont')

- Ganitkevitch et al., EMNLP-2011
 - Similar to the above work of (Zhao et al., 2009)
 - Extract paraphrase patterns from bilingual corpora based on a pivot approach

Paraphrase Rule	Foreign Pivot Phrase
Lexical paraphrase: $JJ \rightarrow \text{offensive} \mid \text{insulting}$	$JJ \rightarrow \text{beleidigend} \mid \text{offensive}$ $JJ \rightarrow \text{beleidigend} \mid \text{insulting}$
Reduced relative clause: $NP \rightarrow NP \text{ that } VP \mid NP VP$	$NP \rightarrow NP \text{ die } VP \mid NP VP$ $NP \rightarrow NP \text{ die } VP \mid NP \text{ that } VP$
Pred. adjective copula deletion: $VP \rightarrow \text{are } JJ \text{ to } NP \mid JJ \text{ NP}$	$VP \rightarrow \text{sind } JJ \text{ f\"ur } NP \mid \text{are } JJ \text{ to } NP$ $VP \rightarrow \text{sind } JJ \text{ f\"ur } NP \mid JJ \text{ NP}$
Partitive construction: $NP \rightarrow CD \text{ of the NNS} \mid CD \text{ NNS}$	$NP \rightarrow CD \text{ der NNS} \mid CD \text{ of the NNS}$ $NP \rightarrow CD \text{ der NNS} \mid CD \text{ NNS}$

Application-driven Statistical Paraphrase Generation (cont')

- Ganitkevitch et al., EMNLP-2011 (cont.)
 - Paraphrase generation
 - Be regarded as monolingual translation task
 - Consider certain application
 - Add features for the given application
 - E.g., sentence compression
 - » Features: source / target length (word number); length difference
 - Change object function during parameter tuning

References

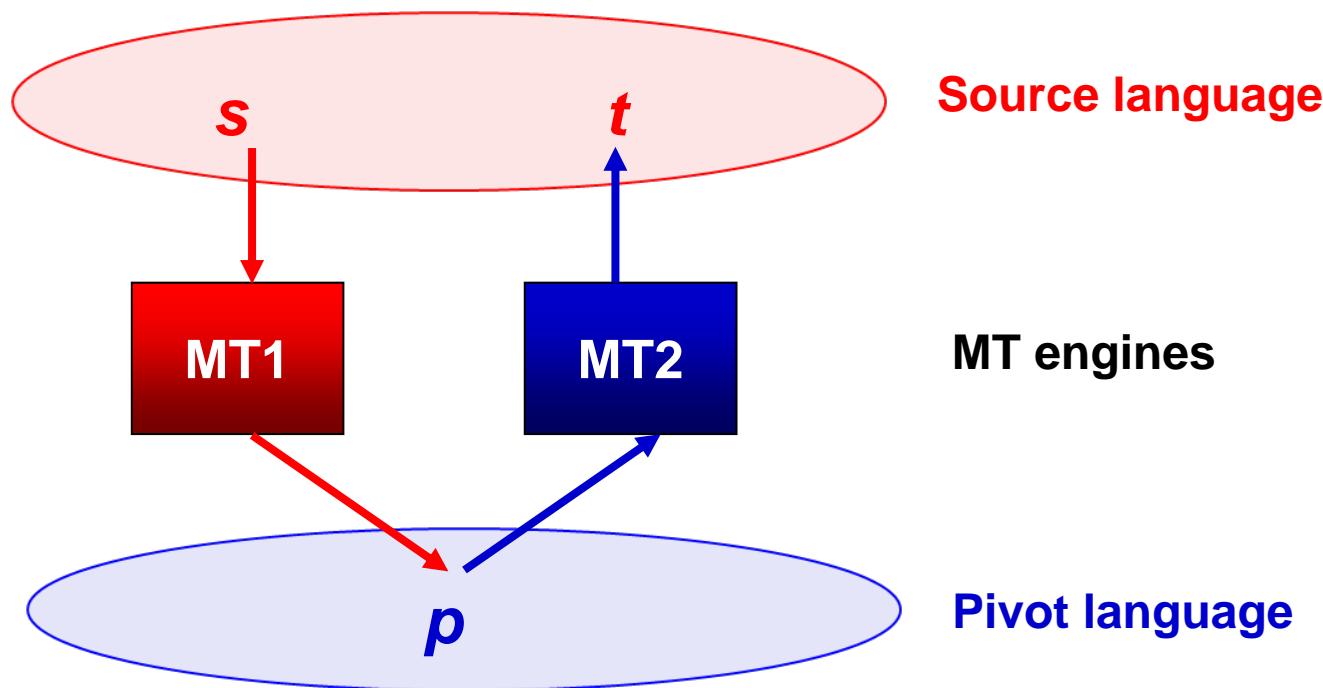
- Lin. 1998. Automatic Retrieval and Clustering of Similar Words.
- Quirk et al. 2004. Monolingual Machine Translation for Paraphrase Generation.
- Finch et al. 2004. Paraphrasing as Machine Translation.
- Zhao et al. 2008. Combining Multiple Resources to Improve SMT-based Paraphrasing Model.
- Zhao et al. 2009. Application-driven Statistical Paraphrase Generation.
- Ganitkevitch et al. 2011. Learning Sentential Paraphrases from Bilingual Parallel Corpora for Text-to-Text Generation.

Outline

- **Part III**
 - **Paraphrase Generation**
 - Rule based Method
 - Thesaurus based Method
 - NLG based Method
 - MT based Method
 - **Pivot based Method**
 - Applications of Paraphrases
 - Evaluation of Paraphrases
 - Conclusions and Future work

Overview

- Basic idea
 - We can generate a paraphrase t for a sentence s by translating s into a foreign language, and then translating it back into the source language.



Overview (cont')

- Example:

English **What toxins are most hazardous to expectant mothers?**



Italian **Che tossine sono più pericolose alle donne incinte?**



English **What toxins are more dangerous to pregnant women?**

- Single-pivot

- Using a single pivot language

- Multi-pivot

- Using multiple pivot languages

Pivot based Methods

- Duboue and Chu-Carroll, 2006
 - Applied in QA systems
 - Paraphrase the input questions so as to improve the coverage in answer extraction
 - Pivot languages
 - 11
 - MT engines
 - 2: Babelfish (**B**) and Google MT (**G**)
 - 4 combinations: B+B, B+G, G+G, G+B

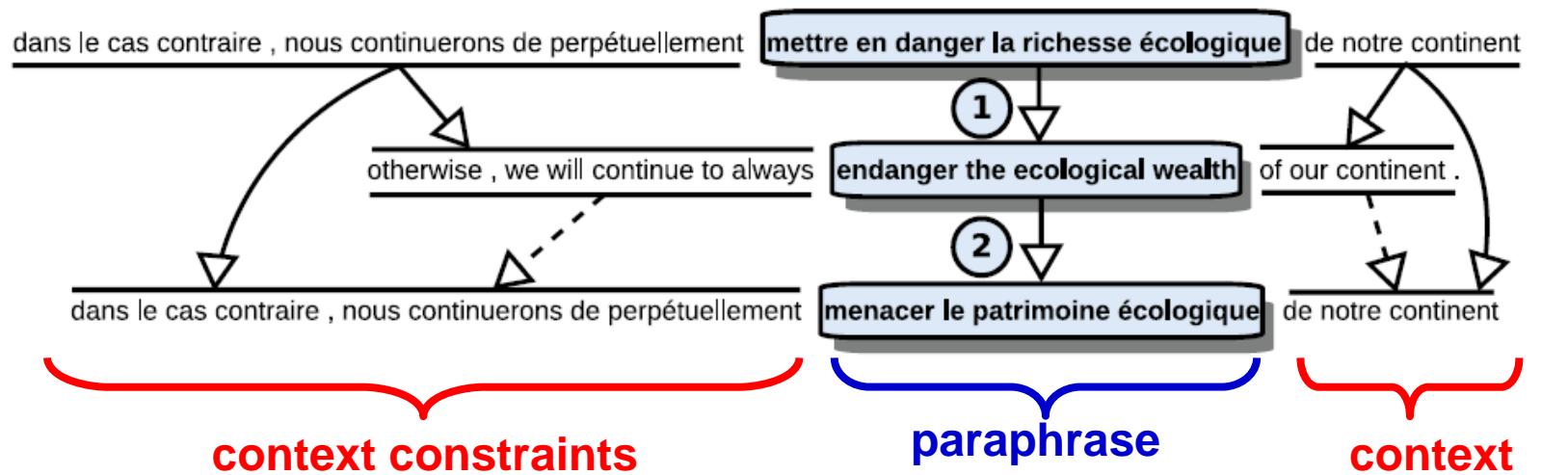
Pivot based Methods (cont')

- Duboue and Chu-Carroll, 2006 (cont')
 - Given a list of automatically generated paraphrases, we need to select a **best** one.
 - For QA, we need to select the paraphrase that can find the answer more easily than the original question.

Features for paraphrase selection (in a classification framework)	
SUM IDF	The sum of the IDF scores for all terms in the original question and the paraphrase. (prefer paraphrases with more informative terms)
Lengths	Number of query terms for each of the paraphrase and the original question. (prefer shorter paraphrases)
Cosine Distance	The distance between the vectors of both questions, IDF-weighted. (filter paraphrases that diverge too much from the original)
Answer Types	Whether answer types, as predicted by the question analyzer, are the same or overlap. (the answer type should be the same)

Pivot based Methods (cont')

- Max, 2009
 - Paraphrasing sub-sentential fragments
 - Allows the exploitation of context during both source-pivot translation and pivot-source back-translation

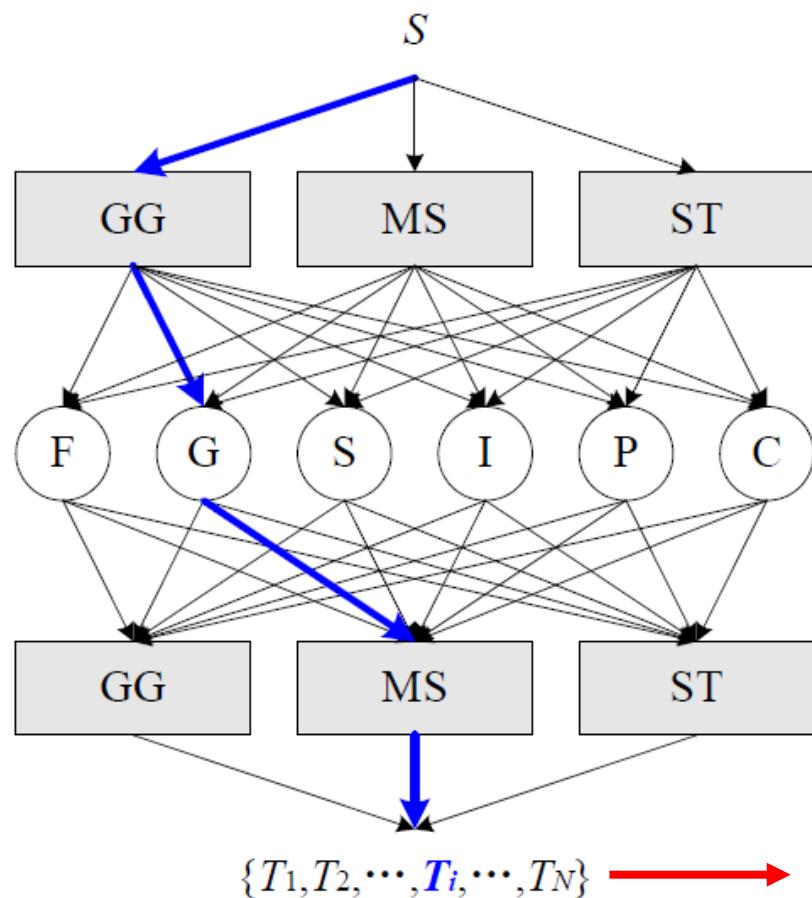


Pivot based Methods (cont')

- Max, 2009 (cont')
 - Application
 - Text revision
 - Pivot language
 - English
 - Paraphrases are acquired for French sub-sentences
 - MT engine
 - Source context aware SMT (Stroppa et al., 2007)

Pivot based Methods (cont')

- Zhao et al., 2010



3 MT engines: (1) Google translator (GG), (2) Microsoft translator (MS), (3) Systran translator (ST)

6 pivot languages: (1) French (F), (2) German (G), (3) Spanish (S), (4) Italian (I), (5) Portuguese (P), (6) Chinese (C)

Pivot based Methods (cont')

- Zhao et al., 2010 (cont')
 - Produce a high-quality paraphrase using the list of candidates

Source	he said there will be major cuts in the salaries of high-level civil servants
(GG, G, MS)	he said there are significant cuts in the salaries of high-level officials
(GG, F, GG)	he said there will be significant cuts in the salaries of top civil level
(GG, P, GG)	he said there will be big cuts in salaries of high-level civil
(MS, C, MS)	he said that there will be a major senior civil service pay cut
(MS, S, GG)	he said there will be significant cuts in the salaries of senior officials
(MS, F, ST)	he said there will be great cuts in the wages of the high level civils servant
(ST, G, GG)	he said that there are major cuts in the salaries of senior government officials
.....



Good paraphrases



Bad paraphrases

Pivot based Methods (cont')

- Zhao et al., 2010 (cont')
 - Two techniques for producing high-quality paraphrases using the candidates
 - Selection-based technique
 - Select a best paraphrase from the 54 candidates based on Minimum Bayes Risk (MBR)
 - Decoding-based technique
 - Train a MT model using the 54 candidates, and generate a new paraphrase with it

References

- Duboue and Chu-Carroll. 2006. Answering the Question You Wish They Had Asked: The Impact of Paraphrasing for Question Answering.
- Stroppa et al. 2007. Exploiting Source Similarity for SMT using Context-informed Features.
- Max. 2009. Sub-sentential Paraphrasing by Contextual Pivot Translation.
- Zhao et al. 2010. Leveraging Multiple MT Engines for Paraphrase Generation.

Outline

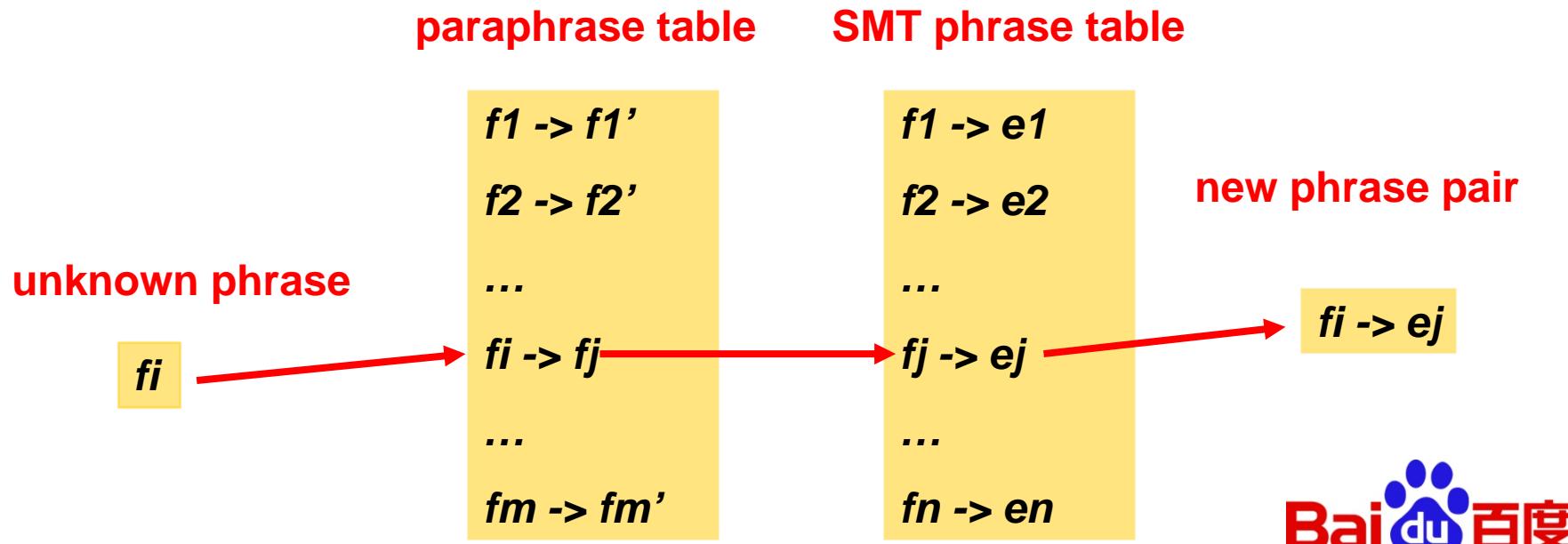
- **Part III**
 - Paraphrase Generation
 - **Applications of Paraphrases**
 - **Paraphrasing for MT**
 - Other Applications
 - Evaluation of Paraphrases
 - Conclusions and Future work

Paraphrasing for MT

- Applications:
 - Translate unknown terms (phrases)
 - Expand training data
 - Rewrite input sentences
 - Improve automatic evaluation
 - Tune parameters

Translate Unknown Terms (Phrases)

- Basic idea:
 - In SMT, when encountering an unknown source term (phrase), we can substitute a paraphrase for it and then proceed using the translation of that paraphrase



Translate Unknown Terms (Phrases) (cont')

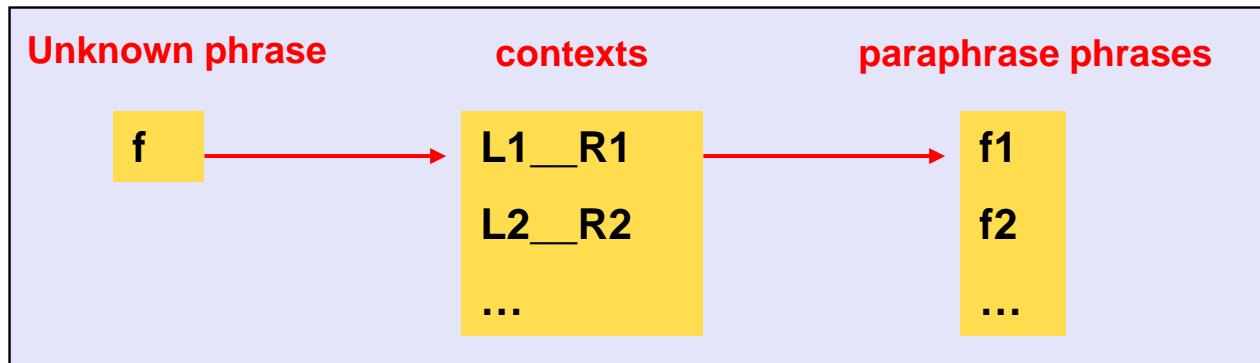
- Callison-Burch et al., 2006
 - Paraphrases are extracted from bilingual parallel corpora using the pivot approach
 - New phrase pairs generated through paraphrasing are incorporated into the phrase table
 - The paraphrase probability is added as a new feature function:

paraphrase
probability

$$h(e, f_1) = \begin{cases} p(f_2 | f_1) & \text{If phrase table entry } (e, f_1) \\ & \text{is generated from } (e, f_2) \\ 1 & \text{Otherwise} \end{cases}$$

Translate Unknown Terms (Phrases) (cont')

- Marton et al., 2009
 - Paraphrases are extracted from monolingual corpora, based on distributional hypothesis



- Combine the new phrase pairs in the phrase table

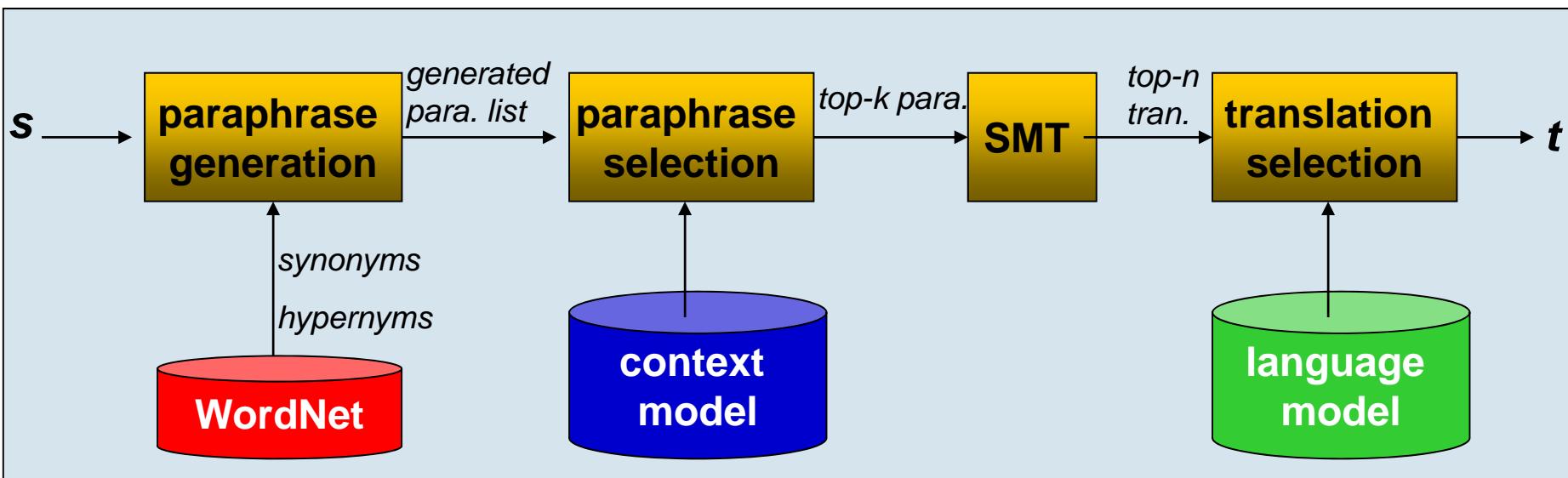
Context
similarity

$$h(e, f_1) = \begin{cases} psim(DP_{f_1}, DP_{f_2}) & \text{If phrase table entry } (e, f_1) \\ & \text{is generated from } (e, f_2) \\ 1 & \text{Otherwise} \end{cases}$$

If phrase table entry (e, f_1)
is generated from (e, f_2)
Otherwise

Translate Unknown Terms (Phrases) (cont')

- Mirkin et al., 2009
 - Use not only paraphrases but also entailment rules
 - From WordNet
 - Paraphrases: *synonyms* in WordNet
 - Entailment rules: *hypercnyms* in WordNet



Translate Unknown Terms (Phrases) (cont')

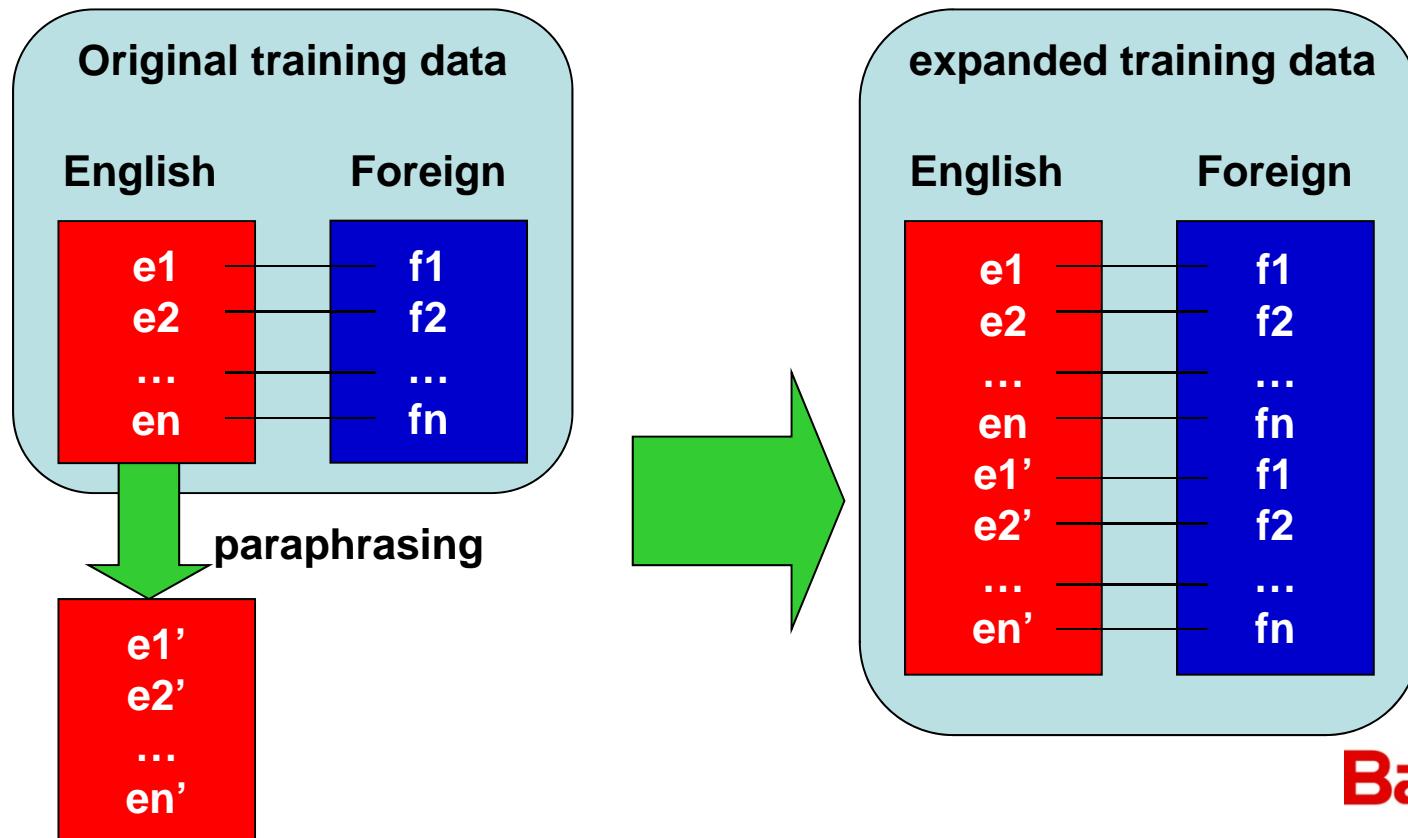
- Onishi et al., 2010
 - Using paraphrase lattices for SMT
 - **Step-1:** Paraphrase the input sentence, and generate a paraphrase lattice
 - Paraphrases are extracted from bilingual parallel corpora based on the pivot approach
 - **Step-2:** Give the paraphrase lattice as the input to the lattice decoder

Translate Unknown Terms (Phrases) (cont')

- Effectiveness
 - When the training data of SMT is small
 - Effective 😊
 - Problem of unknown terms is more serious when the training data is small
 - When the training data of SMT is large
 - Ineffective 😞
 - Unknown terms can be covered by adding more training data

Expand Training Data

- Enlarge training data via paraphrasing the source-side sentences in the parallel corpus



Rewrite Input Sentences

- Paraphrase the sentence to be translated, so as to make it more translatable
 - Yamamoto, 2002; Zhang and Yamamoto, 2002
 - Rule-based Paraphraser for simplifying the source sentences
 - Shimohata et al., 2004
 - Shorten long sentences and sentences with redundant information in a speech translation system
 - Nakov and Ng., 2011
 - Paraphrase morphological complex sentences to simpler ones in order to improve the translation quality (e.g., Malay-English translation)

Improve Automatic Evaluation

- Automatic evaluation of MT
 - Based on counting the overlaps between the references and machine outputs
 - E.g., BLEU, NIST...
 - Only computing the surface similarity is limited
 - A meaning may be expressed in a way that is not included in the references
 - Human references are expensive to produce
 - Solution: paraphrase the references so as to include as many correct expressions as possible!

Improve Automatic Evaluation (cont')

- Kauchak and Barzilay, 2006
 - Find a paraphrase of the reference that is closer in wording to the system output
 - Extract candidates from WordNet synonyms

System output

It is **hard** to believe that such tremendous changes have taken **place** for those people and lands that I have never stopped missing while living abroad.

Correct

Wrong

Reference

For someone born here but has been sentimentally attached to a foreign country far from **home**, it is **difficult** to believe this kind of changes.

- Filter the invalid substitution given the context
 - Binary classification
 - Features: context n-grams and local collocations

Improve Automatic Evaluation (cont')

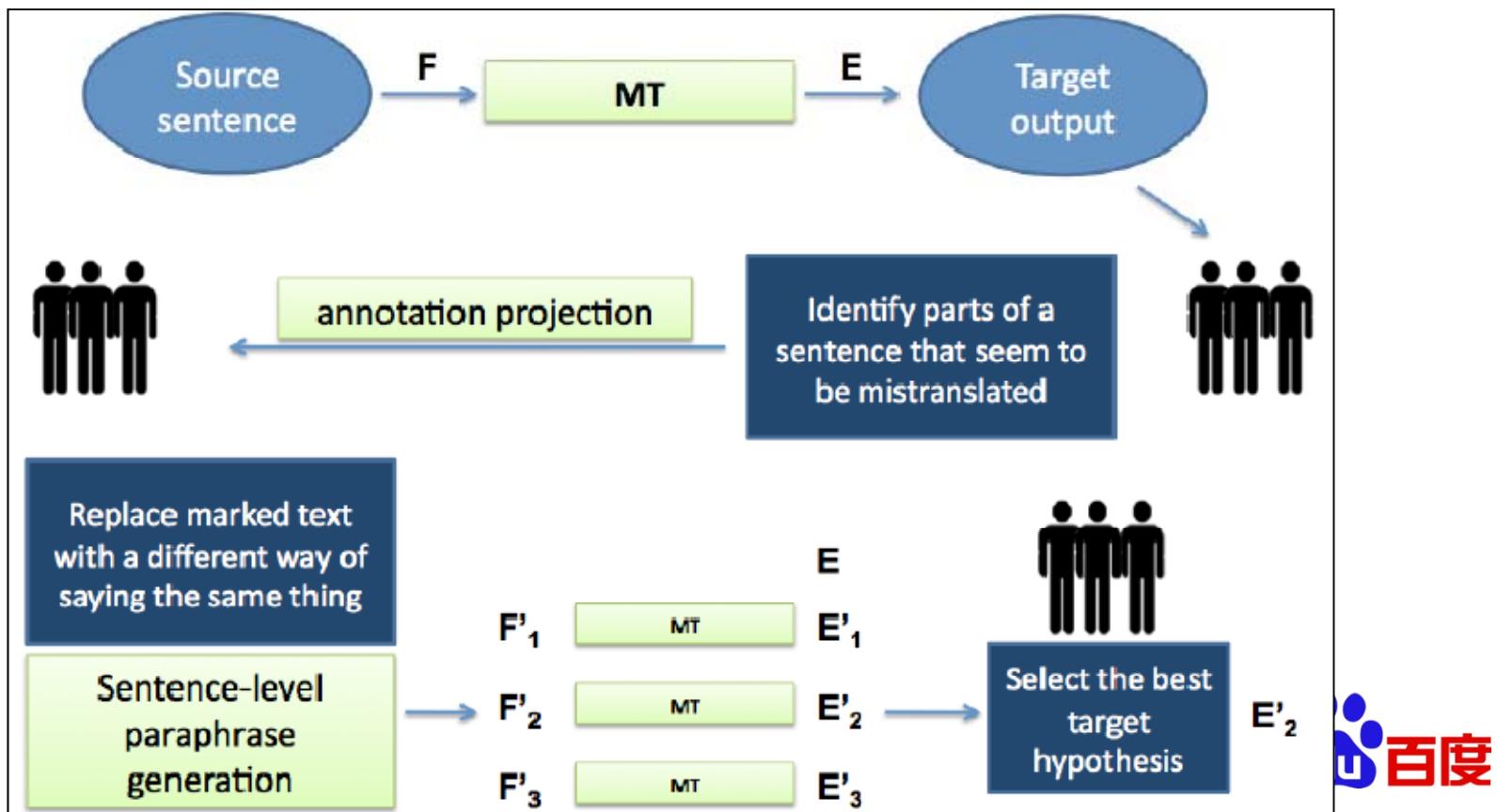
- Zhou et al., 2006
 - ParaEval: Compute the similarity of reference and system output using paraphrases
 - Paraphrases are learned from bilingual parallel corpora with a pivot approach
 - Two-tier matching strategy for SMT evaluation
 - First tier: paraphrase match
 - Second tier: unigram match for words not matched by paraphrases

Tune Parameters

- Madnani et al. 2007
 - Similar to the studies using paraphrases to improve automatic evaluation of MT
 - Parameter tuning in SMT also needs references
 - Parameter estimation of SMT:
 - optimize BLEU on a development set
 - Expand the references automatically via paraphrasing
 - Paraphrase generation
 - Paraphrase resources are acquired based on a pivot approach
 - Recast paraphrase generation as a monolingual MT problem and decode with a typical SMT decoder

Targeted Paraphrasing for MT

- Using targeted paraphrasing and monolingual crowdsourcing to improve translation (Resnil et al., 2010)



References

- Translate unknown terms (phrases)
 - Callison-Burch et al. 2006. Improved Statistical Machine Translation Using Paraphrases.
 - Marton et al. 2009. Improved Statistical Machine Translation Using Monolingually-Derived Paraphrases.
 - Mirkin et al. 2009. Source-Language Entailment Modeling for Translating Unknown Terms.
 - Onishi et al. 2010. Paraphrase Lattice for Statistical Machine Translation.
- Expand training data
 - Nakov. 2008. Improved Statistical Machine Translation Using Monolingual Paraphrases.
 - Bond et al. 2008. Improving Statistical Machine Translation by Paraphrasing the Training Data.
- Targeted paraphrasing
 - Resnik et al., 2010. Improving Translation via Targeted Paraphrasing.

References (cont')

- Rewrite input sentences
 - Yamamoto. 2002. Machine Translation by Interaction between Paraphraser and Transfer.
 - Zhang and Yamamoto. 2002. Paraphrasing of Chinese Utterances.
 - Shimohata et al. 2004. Building a Paraphrase Corpus for Speech Translation.
 - Nakov and Ng. 2011. Translating from Morphologically Complex Languages: A Paraphrase-Based Approach
- Improve automatic evaluation
 - Kauchak and Barzilay. 2006. Paraphrasing for Automatic Evaluation.
 - Zhou et al. 2006. Re-evaluating Machine Translation Results with Paraphrase Support.
- Tune parameters
 - Madnani et al. 2007. Using Paraphrases for Parameter Tuning in Statistical Machine Translation.

Outline

- **Part III**
 - Paraphrase Generation
 - **Applications of Paraphrases**
 - Paraphrasing for MT
 - **Other Applications**
 - Evaluation of Paraphrases
 - Conclusions and Future work

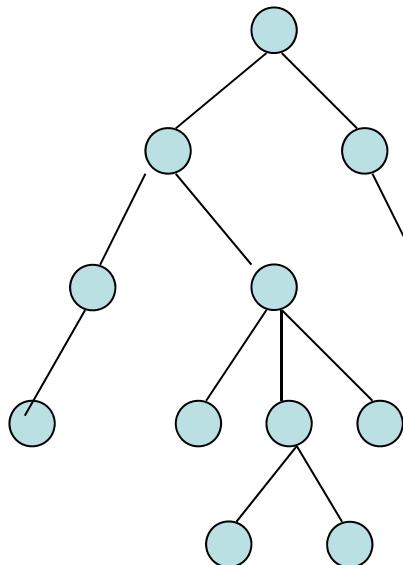
Paraphrasing for QA

- Goal:
 - Alleviate the problem of *word mismatch* between questions and answers
- Two directions:
 - Paraphrase questions
 - Rewrite a question into a group of paraphrases, so as to improve the coverage in answer extraction
 - Paraphrase answer extraction patterns
 - Generate answer extraction patterns as many as possible

Paraphrasing for QA

- Ravichandran and Hovy, 2002.
 - Mining paraphrase patterns from the web
 - Using hand-crafted seeds (e.g., (*Mozart, 1756*) for **BIRTHDAY**)
 - Mining patterns containing the seeds

Question taxonomy



	BIRTHDAY	Paraphrase patterns
1.00	<NAME> (<ANSWER> -)	
0.85	<NAME> was born on <ANSWER>	
0.60	<NAME> was born in <ANSWER>	
0.59	<NAME> was born <ANSWER>	
0.53	<ANSWER> <NAME> was born	
0.50	- <NAME> (<ANSWER>	
0.36	<NAME> (<ANSWER> -	

Given seed (*Mozart, 1756*)

Paraphrasing for Summarization

- Improve automatic evaluation of summaries
 - Zhou et al., 2006
 - Similar to the automatic evaluation of MT
 - Measure the similarity between references and system outputs using paraphrase match as well as exact match
- Improve sentence clustering
 - Barzilay et al., 1999
 - Considering paraphrase match when Computing sentence similarity

Paraphrasing for Error Correction

- Correcting semantic collocation errors with paraphrases
 - Dahlmeier and Ng, EMNLP-2011
 - Learn paraphrase collocations from bilingual corpora based on a pivot approach
 - Generate correction for erroneous collocations with a log-linear model

Other Applications

- Paraphrasing for NLG
 - Text revision and transformation
 - Dras, 1997
 - Text transformation in order to meet external constraints, such as length and readability
- Paraphrasing for IR
 - Query rewriting
 - Zukerman and Raskutti. 2002.
 - Paraphrase user queries with WordNet synonyms

Other Applications (cont')

- Writing style transformation
 - Kaji et al., 2004
 - Paraphrasing predicates from written language to spoken language
- Text simplification
 - Carroll et al. 1999
 - Simplifying texts for language-impaired readers or non-native speakers
- Identify plagiarism
 - Uzuner et al. 2005
 - Using paraphrases to better identify plagiarism

References

- Paraphrasing for QA
 - Ravichandran and Hovy. 2002. Learning Surface Text Patterns for a Question Answering System.
 - Duboue and Chu-Carroll. 2006. Answering the Question You Wish They Had Asked: The Impact of Paraphrasing for Question Answering.
- Paraphrasing for summarization
 - Barzilay et al. 1999. Information Fusion in the Context of Multi-Document Summarization.
 - Zhou et al. 2006. ParaEval: Using Paraphrases to Evaluate Summaries Automatically.
- Paraphrasing for error correction
 - Dahlmeier and Ng. 2011. Correcting Semantic Collocation Errors with L1-induced Paraphrases.

References (cont')

- Paraphrasing for NLG
 - Dras. 1997. Reluctant Paraphrase: Textual Restructuring under an Optimisation Model.
- Paraphrasing for IR
 - Zukerman and Raskutti. 2002. Lexical Query Paraphrasing for Document Retrieval.
- Writing style transformation
 - Kaji et al. 2004. Paraphrasing Predicates from Written Language to Spoken Language Using the Web.
- Text simplification
 - Carroll et al. 1999. Simplifying Text for Language-Impaired Readers.
- Identify plagiarism
 - Uzuner et al. 2005. Using Syntactic Information to Identify Plagiarism.

Outline

- **Part III**
 - Paraphrase Generation
 - Applications of Paraphrases
 - Paraphrasing for MT
 - Other Applications
 - **Evaluation of Paraphrases**
 - Conclusions and Future work

Evaluation of Paraphrases

- No widely accepted evaluation criteria 😞
 - **Problem-1:** Researchers define various evaluation methods in their studies
 - Difficult to make a direct comparison among different works
 - **Problem-2:** Human evaluation is commonly used
 - Human evaluation is rather subjective
 - Difficult to replicate

Evaluation of Paraphrase Identification

- Human evaluation
- Automatic evaluation
 - Brockett and Dolan, 2005
 - **A**lignment **E**rror **R**ate (AER)
 - AER is indicative of how far the corpus is from providing a solution under a standard SMT tool

$$AER = \frac{|A \cap P| + |A \cap S|}{|A + S|}$$

Automatic alignment

POSSIBLE + SURE
alignment in the gold
standard

SURE alignment in
the gold standard

Evaluation of Lexical Substitution

- Automatic evaluation
 - McCarthy and Navigli, 2007
 - Construction of gold standard data
 - Five annotators, who are native speakers
 - For each test word, each annotator provides up to three substitutes
 - Evaluation:
 - Precision and Recall

Evaluation of Paraphrase Phrases



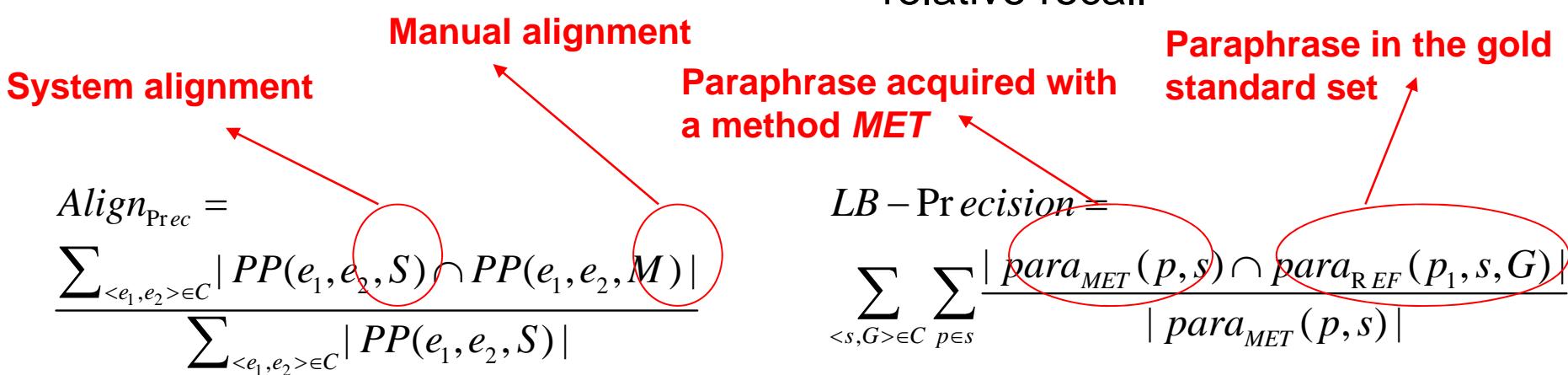
- Human evaluation
 - Ask judges:
 - Whether paraphrases were approximately conceptual equivalent
 - Whether the paraphrases were roughly interchangeable given the genre
 - Whether the substitutions preserved the meaning and remained grammatical
 -
 - The criteria above are vaguely defined and not easy to reproduce

Evaluation of Paraphrase Phrases (cont')

- Automatic evaluation
 - Callison-Burch et al., 2008
 - Data:
 - Parallel sentences, in which paraphrases are annotated through manual alignment (gold standard)
 - Two fashions of evaluation
 - Calculate how well an automatic paraphrasing technique can **align** the paraphrases in a sentence pair
 - Calculate the **lower-bound precision** and **relative recall** of a paraphrasing technique (which extracts paraphrases from other resources)

Evaluation of Paraphrase Phrases (cont')

- Alignment precision and recall



- Alignment precision and recall

- Lower-bound precision and relative recall

Evaluation of Paraphrase Patterns

- Human evaluation
 - Paraphrase patterns cannot be evaluated without context information
 - E.g., *X acquire Y*, *X buy Y*
 - Correct or not? It depends on what fill in slots X and Y
 - Common view:
 - A pair of paraphrase patterns is considered correct if the judge could think of contexts under which it holds
 - Problem:
 - Different judges may think of totally distinct contexts, thus the agreement among the judges could be low

Evaluation of Paraphrase Patterns (cont')

- Szpektor et al., 2007
 - Evaluate paraphrase patterns (and entailment rules) with instances rather than directly evaluate patterns
 - Judges are presented not only with a pair of patterns, but also a sample of sentences that match its left-hand side
 - Judges assess whether two patterns are paraphrases under each specific example
 - A pair of paraphrase patterns is considered as correct only when the percentage of correct examples is high enough

Evaluation of Paraphrase Sentences

- Human evaluation
 - Similar to human evaluation of SMT
 - Criteria (Zhao et al., 2009, 2010)
 - **Adequacy**: If the meaning of the source sentence is preserved in the paraphrase?
 - **Fluency**: if the generated paraphrase is well-formed?
 - **Usability** (Zhao et al., 2009): If the paraphrase meets the requirement of the given application?
 - **Paraphrase rate** (Zhao et al., 2009): How different the paraphrase is from the source sentence?

Evaluation of Paraphrase Sentences (cont')

- Three scales for adequacy, fluency, and usability (Zhao et al., 2009)

Adequacy	1	The meaning is evidently changed.
	2	The meaning is generally preserved.
	3	The meaning is completely preserved.
Fluency	1	The paraphrase t is incomprehensible.
	2	t is comprehensible.
	3	t is a flawless sentence.
Usability	1	t is opposite to the application purpose.
	2	t does not achieve the application.
	3	t achieves the application.

- Five scales for adequacy and fluency (Zhao et al., 2010)

Evaluation of Paraphrase Sentences (cont')

- Paraphrase rate (Zhao et al., 2010):
 - PR-1: based on word overlap rate

$$PR1(T) = 1 - \frac{OL(S, T)}{L(S)}$$

—————→ Word overlap rate
—————→ Number of words
in the source sen.

- PR-2: based on edit distance

$$PR2(T) = \frac{ED(S, T)}{L(S)}$$

—————→ Edit distance

Evaluation of Paraphrase Sentences (cont')

- Two questions:
 - **Q1:** Why not adopt automatic MT methods here, e.g., BLEU, NIST, TER...?
 - **Reason-1:** It is much more difficult to construct human references in paraphrase generation than MT
 - **Reason-2:** Paraphrases that change less will get larger scores in criteria like BLEU
 - **Q2:** How to combine the evaluation of *paraphrase correctness* and *paraphrase rate*?
 - They seem to be incompatible

Evaluation of Paraphrase Sentences (cont')

- Crowdsourcing-based method
 - Chen and Dolan, ACL-2011
 - Crowdsource large set of parallel sentences
 - A worker watches a very short video clip, and writes a description of the content
 - Descriptions from different workers form a parallel sentence collection
 - Amazon's Mechanical Turk is used

Evaluation of Paraphrase Sentences (cont')

- Crowdsourcing-based method (cont.)
 - Paraphrase evaluation metrics:
 - BLEU:
 - Measure the similarity between a candidate paraphrase and reference paraphrases (the larger the better)
 - PINC:
 - Measure the similarity between a candidate paraphrase and the source sentence (the smaller the better)
 - Use BLEU and PINC together but treat them separately

Evaluation within Applications

- Evaluate the role of a paraphrasing module within a certain application system
 - E.g., in MT, examine whether a paraphrasing module helps to alleviate the unknown term problem
 - E.g., in QA, whether paraphrasing the answer patterns can improve the coverage of answer extraction
- Problems:
 - Whether the result can hold for a different application?
 - How to evaluate the role of the paraphrase module independently (not influenced by other modules)?

References

- Brockett and Dolan. 2005. Support Vector Machines for Paraphrase Identification.
- Szpektor et al. 2007. Instance-based Evaluation of Entailment Rule Acquisition.
- McCarthy and Navigli. 2007. SemEval-2007 Task 10: English Lexical Substitution Task.
- Callison-Burch et al. 2008. ParaMetric: An Automatic Evaluation Metric for Paraphrasing.
- Zhao et al. 2009. Application-driven Statistical Paraphrase Generation.
- Zhao et al. 2010. Leveraging Multiple MT Engines for Paraphrase Generation.
- Chen and Dolan. 2011. Collecting Highly Parallel Data for Paraphrase Evaluation.

Outline

● Part III

- Paraphrase Generation
- Applications of Paraphrases
 - Paraphrasing for MT
 - Other Applications
- Evaluation of Paraphrases
- Conclusions and Future work

Conclusions and Future Work

- Conclusions
 - Paraphrasing is important in various research areas
 - Many different kinds of corpora and data resources have been investigated for paraphrase extraction
 - Paraphrase generation is a task similar to MT, but not the same
 - Paraphrase evaluation is problematic. Automatic evaluation methods are in need

Conclusions and Future Work (cont')

- Future work
 - Paraphrase extraction
 - Improve the quality of the extracted paraphrases
 - Paraphrase generation
 - Application-driven paraphrase generation
 - Paraphrase application
 - Apply paraphrasing techniques in commercial NLP systems, rather than merely in labs
 - Paraphrase evaluation
 - Come up with evaluation methods that can be widely accepted

Thanks!

QA



百度大厦/外观

SNS Text Mining and Search

Xiaohua Liu and Furu Wei

Natural Language Computing Group

Microsoft Research Asia

Agenda

- QuickView: A Research Platform of SNS Text Mining and Search
- SNS Text Mining
 - Semantic Role Labeling
 - Sentiment Analysis
- Future Work

Agenda

- QuickView: A Research Platform of SNS Text Mining and Search
- SNS Text Mining
 - Semantic Role Labeling
 - Sentiment Analysis
- Future Work



A Research Platform of SNS Text Mining and Search

System Overview

- Input
 - Tweets
 - Facebook updates
 - Chinese Weibo, e.g., Sina microblog
 - ...
- Process
 - NLP pipeline for social content: POS, shallow parsing, SRL, etc.
 - Single instance level mining: entities, events, opinions, etc.
 - Collectively mining: hot topics, hot entities, hot opinions, etc.
- Output
 - A multi-level index of social contents
 - Recent statistical information: e.g., hot topics

Data processed per day

- Tweets
 - 1,500M English tweets
- Facebook
 - 15M updates
 - 100M likes
 - 8M fan pages
- SINA weibo
 - 1.5M micro-blogs

SNS Research Platform

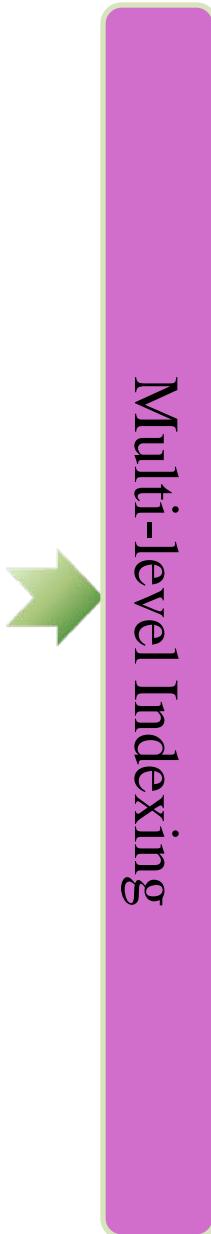
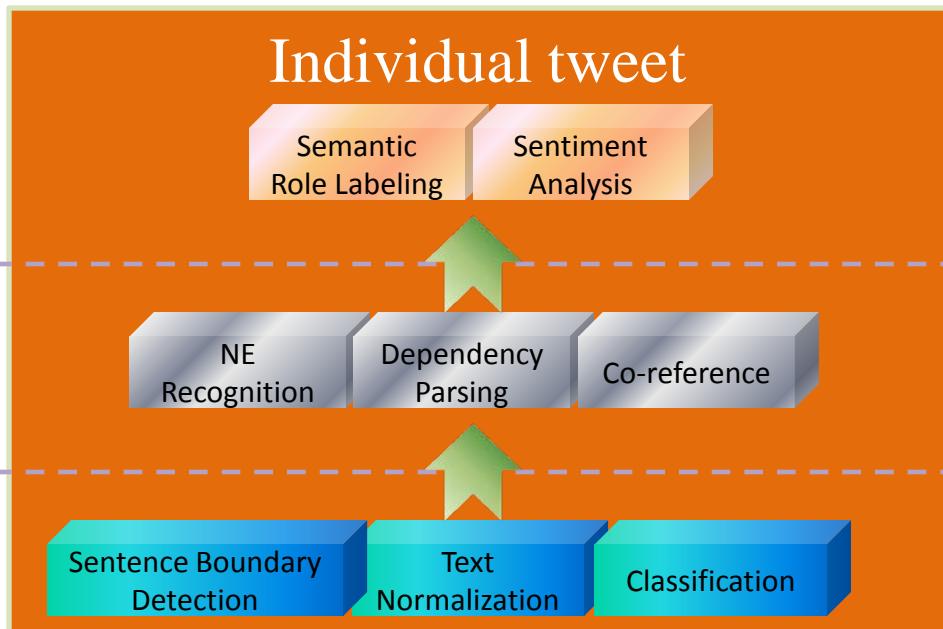
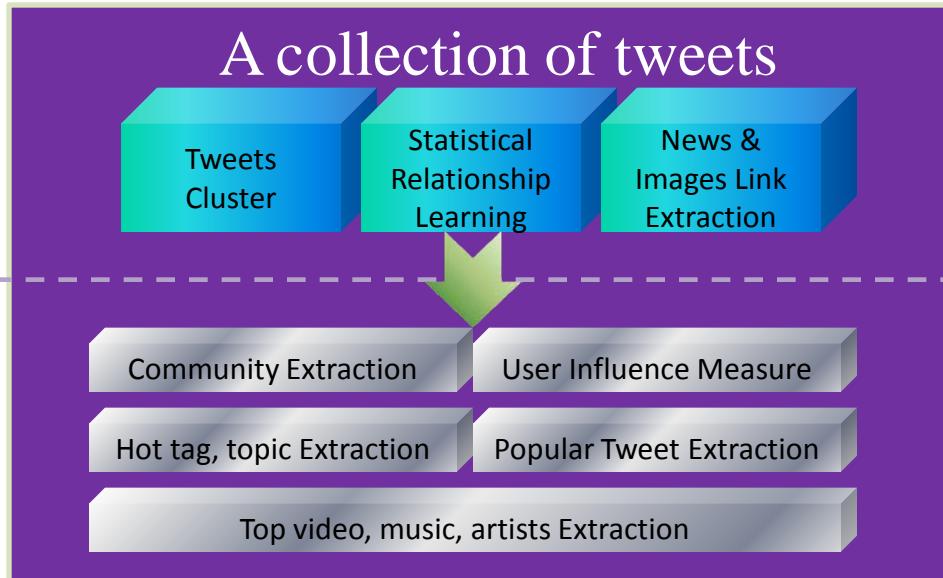
- Twitter, FaceBook, Chinese micro-blogs
- 3000+ News sources
- Multiple languages
- Text mining
- Personalized search and recommendation



Noise
Filtering



Raw Data



Semantic Search

The QuickView Demo

Our Current Research

- Sentiment analysis
 - Target-dependent Twitter Sentiment Classification (ACL 2011)
 - User-level sentiment analysis incorporating social networks (KDD 2011)
 - A Graph-based Hashtag Sentiment Classification Approach. (CIKM 2011)
- NER
 - Recognizing Named Entities in Tweets (ACL 2011)
 - Mining Entity Translations from Comparable Corpora: A Holistic Graph Mapping Approach. (CIKM 2011)
- SRL
 - Collective Semantic Role Labeling (IJCAI 2011)
 - Improving Semantic Role Labeling using Semi-supervised Learning (AAAI 2011)
 - Semantic Role Labeling for News Tweets (Coling 2010)
- SNS Search
 - An Empirical Study on Learning to Rank of Tweets. (Coling 2010)
 - QuickView: Semantic Search For Tweets (SIGIR 2011 demo)

Agenda

- QuickView: A Research Platform of SNS Text Mining and Search
- SNS Text Mining
 - Semantic Role Labeling
 - Sentiment Analysis
- Future Work

Semantic Role Labeling for Tweets

Outline

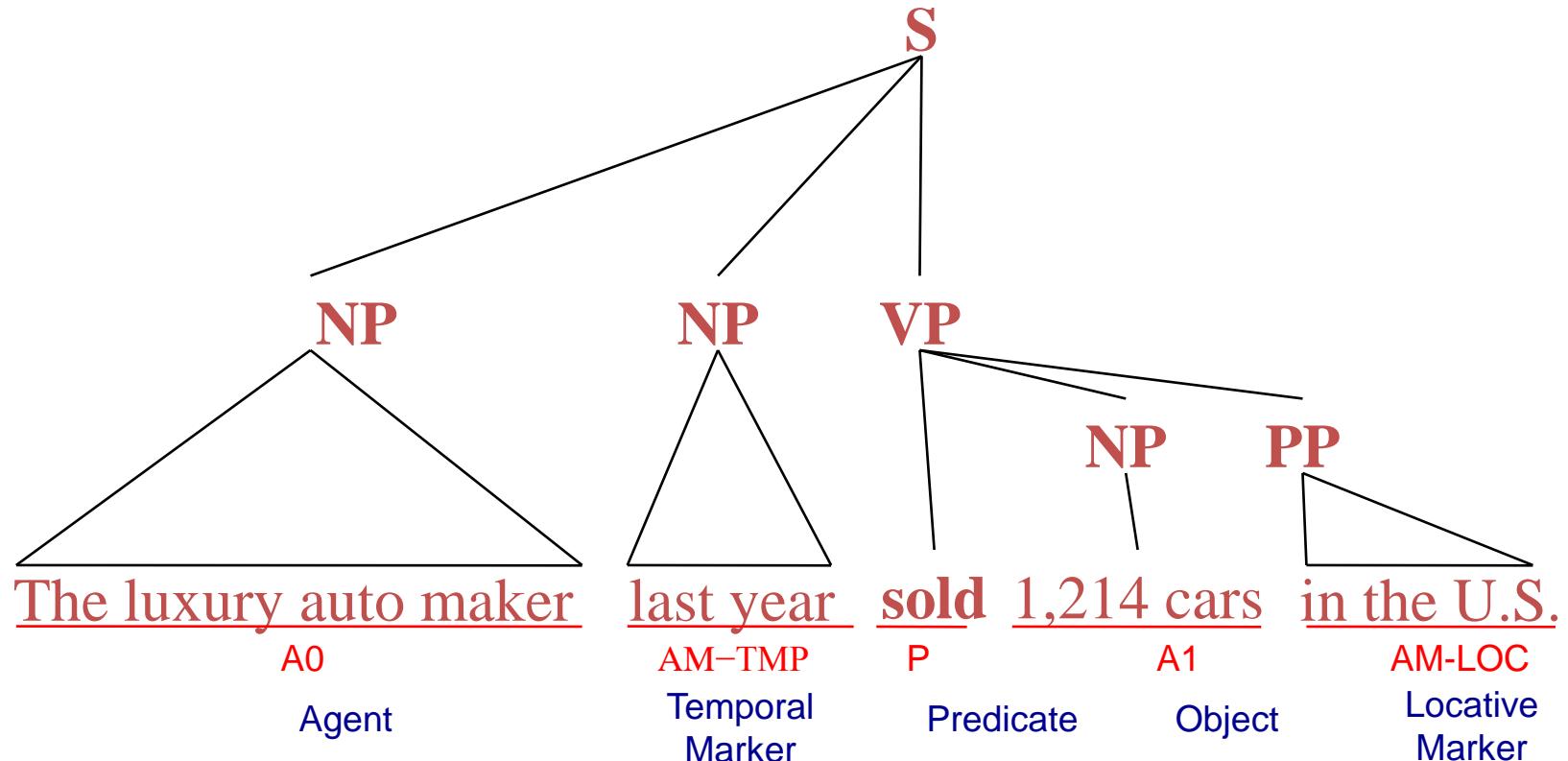
- Introduction
 - SRL task definition
 - Application to twitter search
- General approaches to SRL
 - Resources
 - Typical systems
- SRL on tweets
 - Challenges
 - Method

Outline

- Introduction
 - SRL task definition
 - Application to twitter search
- General approaches to SRL
 - Resources
 - Typical systems
- SRL on tweets
 - Challenges
 - Method

Semantic Role Labeling

- Detect basic event structures such as *who* did *what* to *whom*, *when* and *where*



Predicate

- Verbal predicate (PropBank)
 - *Chile [earthquake]_{A0} shorten the [day]_{A1}*
- Other types of predicate (NomBank)
 - *[Her]_{A0} gift of [a book]_{A1} [to John]_{A2}*

Predicate Arguments

- Core arguments
 - A0, A1: agent and patient
- 13 adjunctive arguments
 - Temporal, manner, location, etc.
- Phrase level vs. word level argument
 - Word level: *Chile [earthquake]*_{A0} *shorten the [day]*_{A1}
 - Phrase level: *[Chile earthquake]*_{A0} *shorten [the day]*_{A1}

Evaluation of SRL

- ▶ Evaluation metrics
 - ▶ Precision
 - ▶ How many output labels are correct
 - ▶ Recall
 - ▶ How many labels in the gold standard dataset are correctly labeled
 - ▶ F1 score
 - ▶ The harmonic mean of Precision and Recall or
 - ▶
$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Evaluation of SRL

- Test datasets (from PropBank)
 - WSJ (Wall Street Journal) : mainly news
 - Brown: more balanced corpus, including news, reports and others
- The state-of-the-art results
 - CoNLL-2005 : 81.52% F1 on WSJ
 - CoNLL-2008: 87.69% F1 on WSJ, 69.06% F1 on Brown
 - CoNLL-2009: 80.47 F1 on WSJ
- Best systems are pipelined or based on MLN

Outline

- Introduction
 - SRL Task definition
 - Application to twitter search
- General approaches to SRL
 - Resources
 - Typical systems
- SRL on tweets
 - Challenges
 - Method

SRL Helps Twitter Search

- Twitter search is now keyword search, unable to answer questions, like how many people were killed in Algeria earthquake?



[OrganicUniverse](#) 2 killed, 43 injured in Algeria **earthquake**|Two people were **killed** and 43 others injured in an **earthquake** ..
<http://oohja.com/xdij2>
about 6 hours ago via API

SRL Helps Twitter Search (2)

- SRL extracts who acted what
 - *oh yea and Chile [earthquake]_{A0} the earth off it's axis according to NASA and shorten the [day]_{A1} by a wee second :-(→ [earthquake]_{A0} shorten the [day]_{A1}*
 - Beyond keyword search, e.g., *what shorten the day?*

SRL Helps Twitter Search

- SRL abstracts away syntax variances
 - *Chile Earthquake Shortened Earth Day*
 - *The Chile earthquake shortened the length of an Earth day*
 - ...
 - → [earthquake]_{A0} **shorten** [day]_{A1}

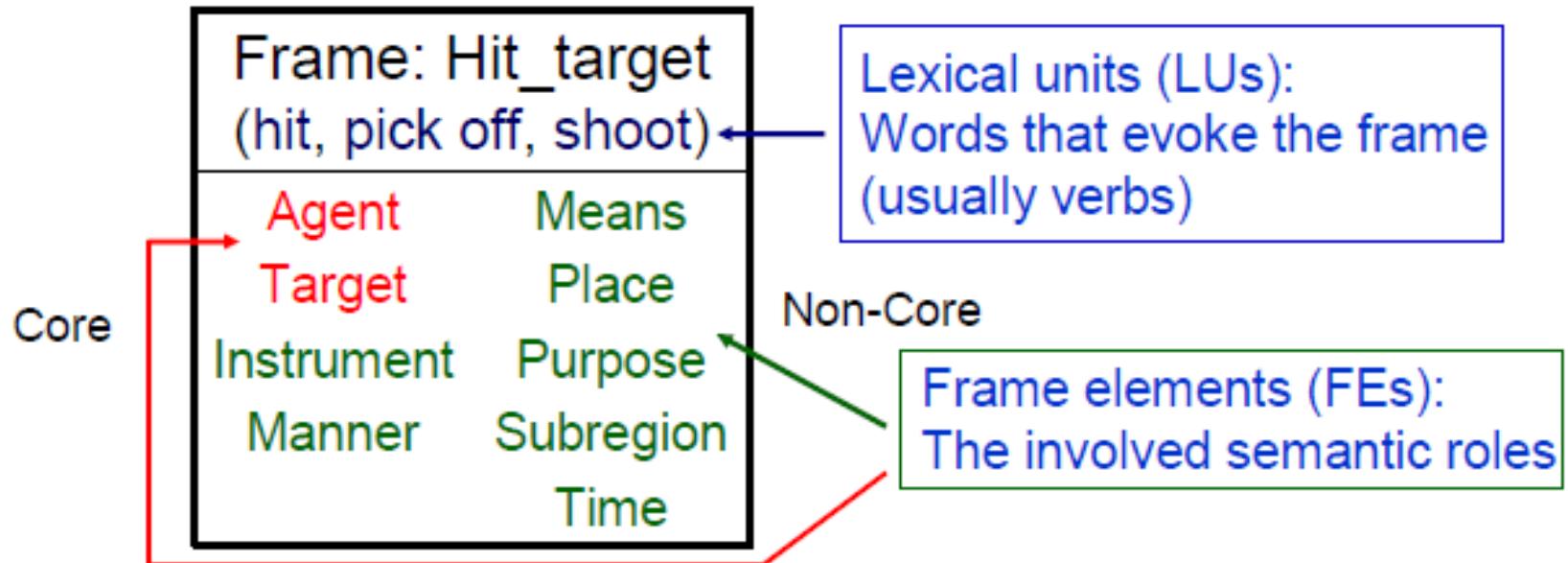
Outline

- Introduction
 - SRL Task definition
 - Application to twitter search
- General approaches to SRL
 - Resources
 - Typical systems
- SRL on tweets
 - Challenges
 - Method

FrameNet(Fillmore et al., 2004)

- Computational frame lexicon + corpus of examples annotated with semantic roles (mostly BNC)
 - ~800 semantic frames
 - >9,000 lexical units
 - ~150,000 annotated sentences

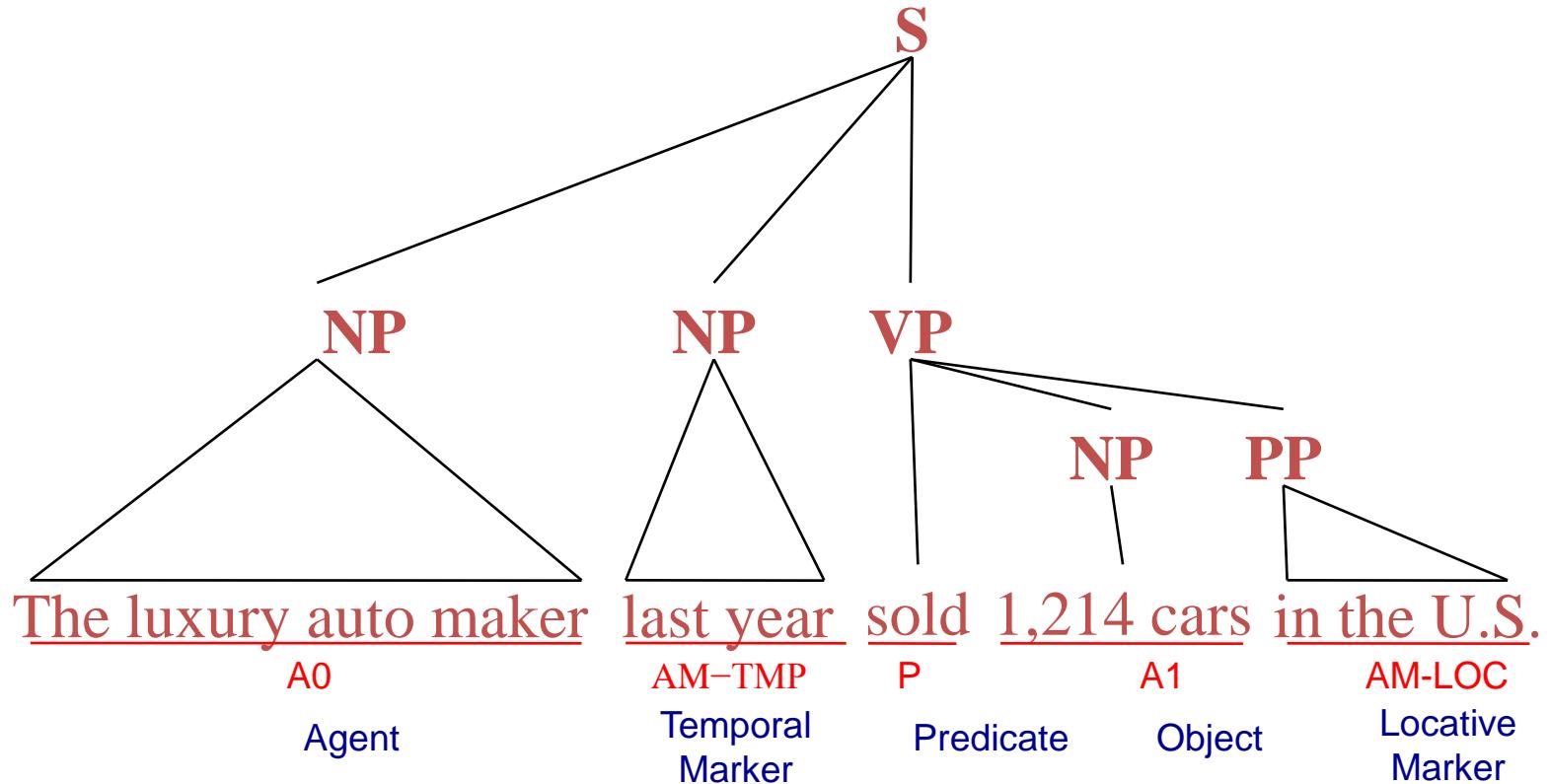
A Frame Example



[Agent *Kristina*] **hit** [Target *Scott*] [Instrument *with a baseball*] [Time *yesterday*].

PropBank (Palmer et al., 2005)

- The **primary resource** for research in SRL
- Annotation of all verbal predicates in Penn Treebank



An Example: Argument Structure Depends on Verb and Its Meaning

sell.01: commerce: seller

A0=“seller” (*agent*); A1=“thing sold” (*theme*); A2=“buyer” (*recipient*); A3=“price paid”; A4=“benefactive”
[Al Brownstein]_{A0} **sold** [it]_{A1} [for \$60 a bottle]_{A3}

sell.02: give up

A0=“entity selling out”
[John]_{A0} **sold out**

sell.03: sell until none is/are left

A0=“seller”; A1=“thing sold”; ...

[The new Harry Potter]_{A1} **sold out** [within 20 minutes]_{AM-TMP}

NomBank (Meyers et al., 2004)

- Annotation of the nominal predicates in Penn TreeBank
 - [IBM]_{A0}'s **appointment** of [John]_{A1}
 - The **appointment** of [John]_{A1} by [IBM]_{A0}
 - [John]_{A1} is the current [IBM]_{A0} **appointee**

Outline

- Introduction
 - SRL Task definition
 - Application to twitter search
- General approaches to SRL
 - Resources
 - Typical systems
- SRL on tweets
 - Challenges
 - Method

Typical systems

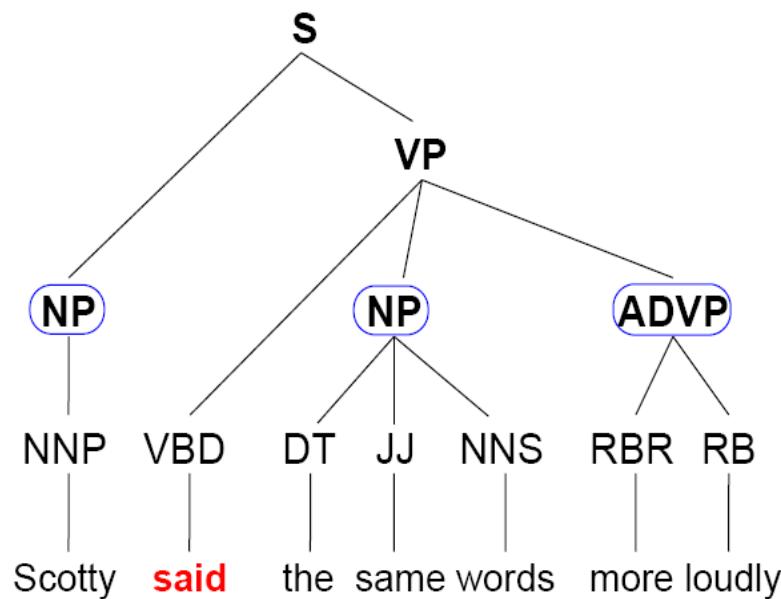
- Pipelined system
- System based on sequential labeling
- System using Markov Logic Networks
- Collective SRL (jointly conduct SRL on multi sentences)

Typical systems

- Pipelined system
- System based on sequential labeling
- System using Markov Logic Networks
- Collective SRL (jointly conduct SRL on multi sentences)

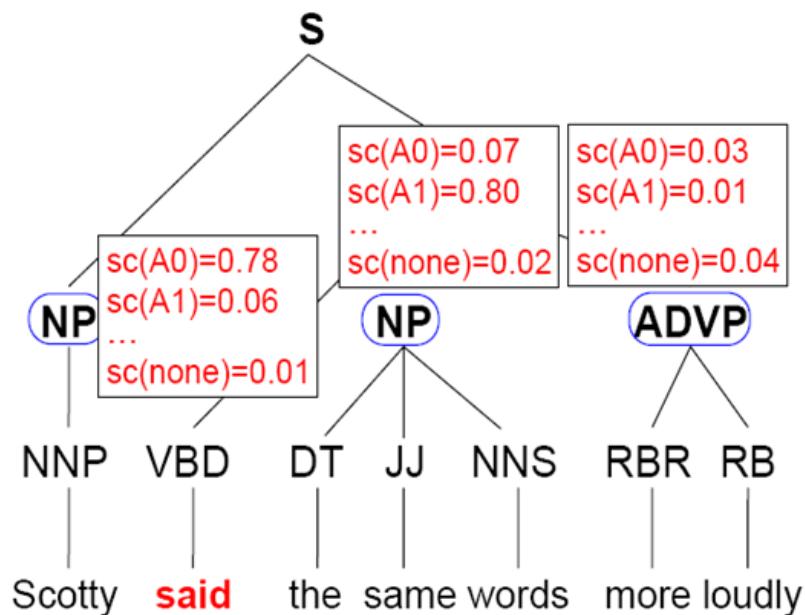
Pipelined SRL

- Argument candidates generation



Pipelined SRL

- Argument candidates generation
- Argument classification



Pipelined SRL

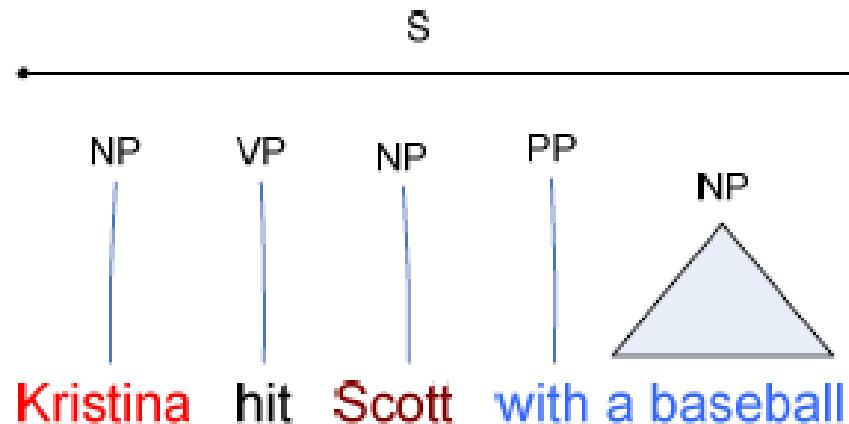
- Argument candidates generation
- Argument classification
- Global inference
 - Find the best solution from all possible solutions
 - E.g., Re-ranking of N best solutions(Haghghi et al., 2005; Toutanova et al., 2008)

Typical Systems

- Pipelined system
- System based on sequential labeling
- System using Markov Logic Networks
- Collective SRL (jointly conduct SRL on multi sentences)

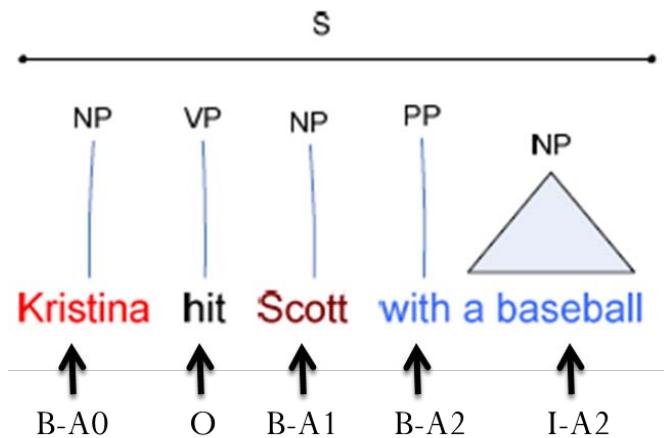
System Based on Sequential Labeling (Marques et al., 2005)

- Break into base chunks
 - Chunker: Yamcha (Kudo & Matsumoto, 2001)



System Based on Sequential Labeling

- Break into base chunks
- Labeling each chunk
 - B/I marks the beginning/ continuation of an argument span; and O non-arguments



Tool: CRF++ <http://crfpp.sourceforge.net/>

Typical Systems

- Pipelined system
- System based on sequential labeling
- **System using Markov Logic Networks**
- Collective SRL (jointly conduct SRL on multi sentences)

System using Markov Logic Networks

(Sebastian Riedel and Ivan Meza-Ruiz, 2008)

- Define formulae

$$\text{lemma}(p, +l_1) \wedge \text{lemma}(a, +l_2) \Rightarrow \text{hasRole}(p, a)$$
$$(\text{role}(p, a, r_1) \wedge r_1 \neq r_2 \Rightarrow \neg \text{role}(p, a, r_2))$$

System using Markov Logic Networks

- Define formulae
- Learning formula weights
 - To allocate high probability to correctly identified predicate argument structures

I swim → {lemma(1,I) , lemma(2, swim), isPredicate(2)} > {lemma(1,I) , lemma(2, swim), isPredicate(1)}

System using Markov Logic Networks

- Define formulae
- Learning formula weights
- Inference
 - Jointly determine predicate argument structures that best fit the formulae

Toolkit: thebeast <http://code.google.com/p/thebeast/>

Typical systems

- Pipelined system
- System based on sequential labeling
- System using Markov Logic Networks
- Collective SRL (jointly conduct SRL on multi sentences)

Task Definition of Collective SRL

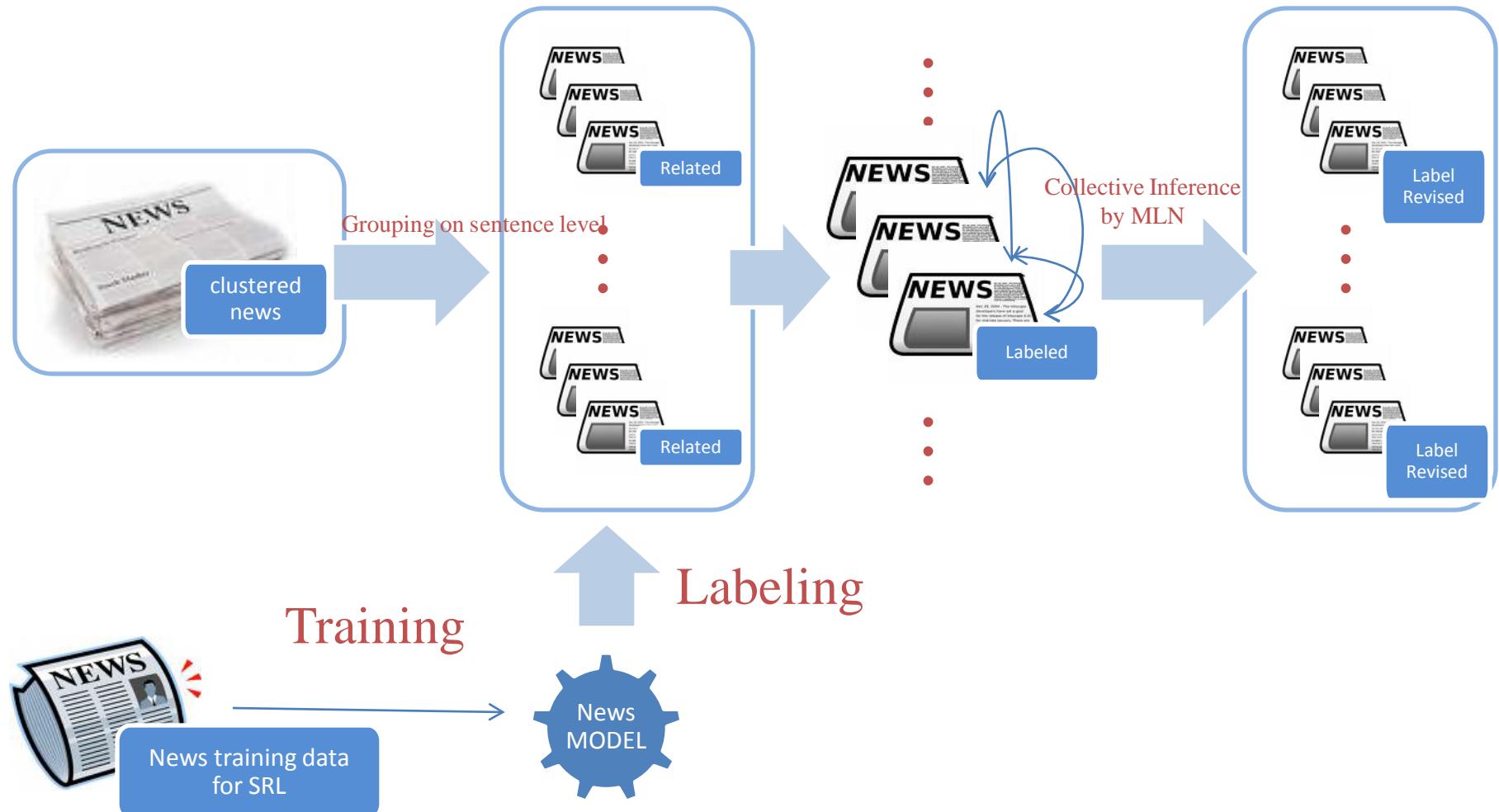
- Input: a set of sentences from news articles
 - 1. *Hurricane Ida, the first Atlantic hurricane to **target** the U.S. this **year**, **plodded** yesterday **toward** the Gulf Coast...*
 - 2. *Hurricane Ida **trudged toward** the Gulf Coast...*
 - ...
- Output: predicate-argument-role structures
 - 1. (plodded, Ida, A0), (plodded, toward, AM-DIR), (target, Ida, A0), (target, U.S., A1), (target, year, AM-TMP)
 - 2. (trudged, Ida, A0), (trudged, toward, AM-DIR)
- Role sets (following PropBank)

Collective SRL

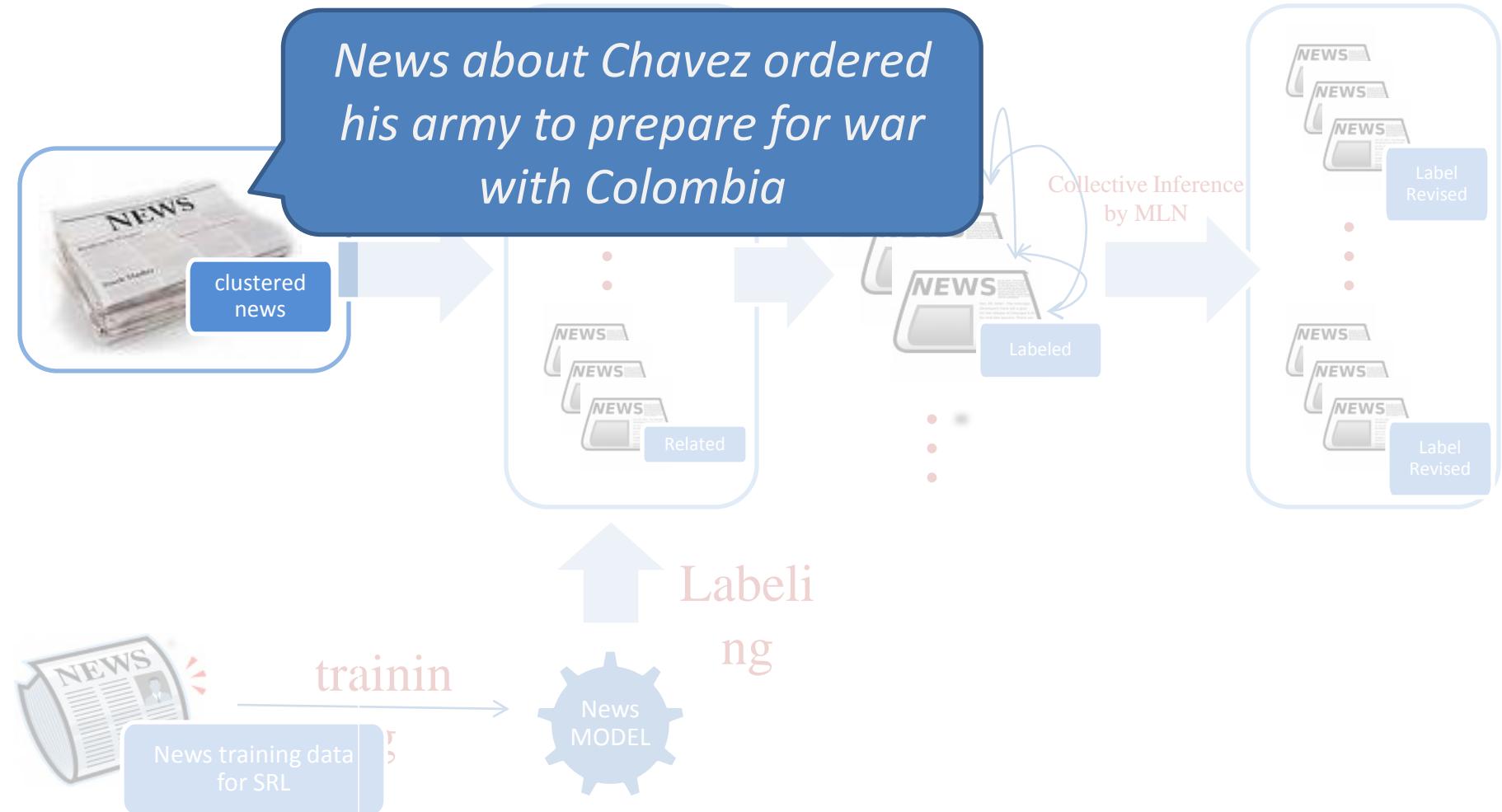
- Motivated by the fact SRL on one sentence can help that on other differently phrased sentences with similar meaning
 - *A suicide **bomber** blew himself up Sunday in market in Pakistan's **northwest** crowded with shoppers ahead of a Muslim holiday, killing 12 people, including a mayor who*
 - *Police in northwestern Pakistan say that a suicide **bomber** has killed at least 13 people and wounded dozens of others.*

Implementation of Collective SRL

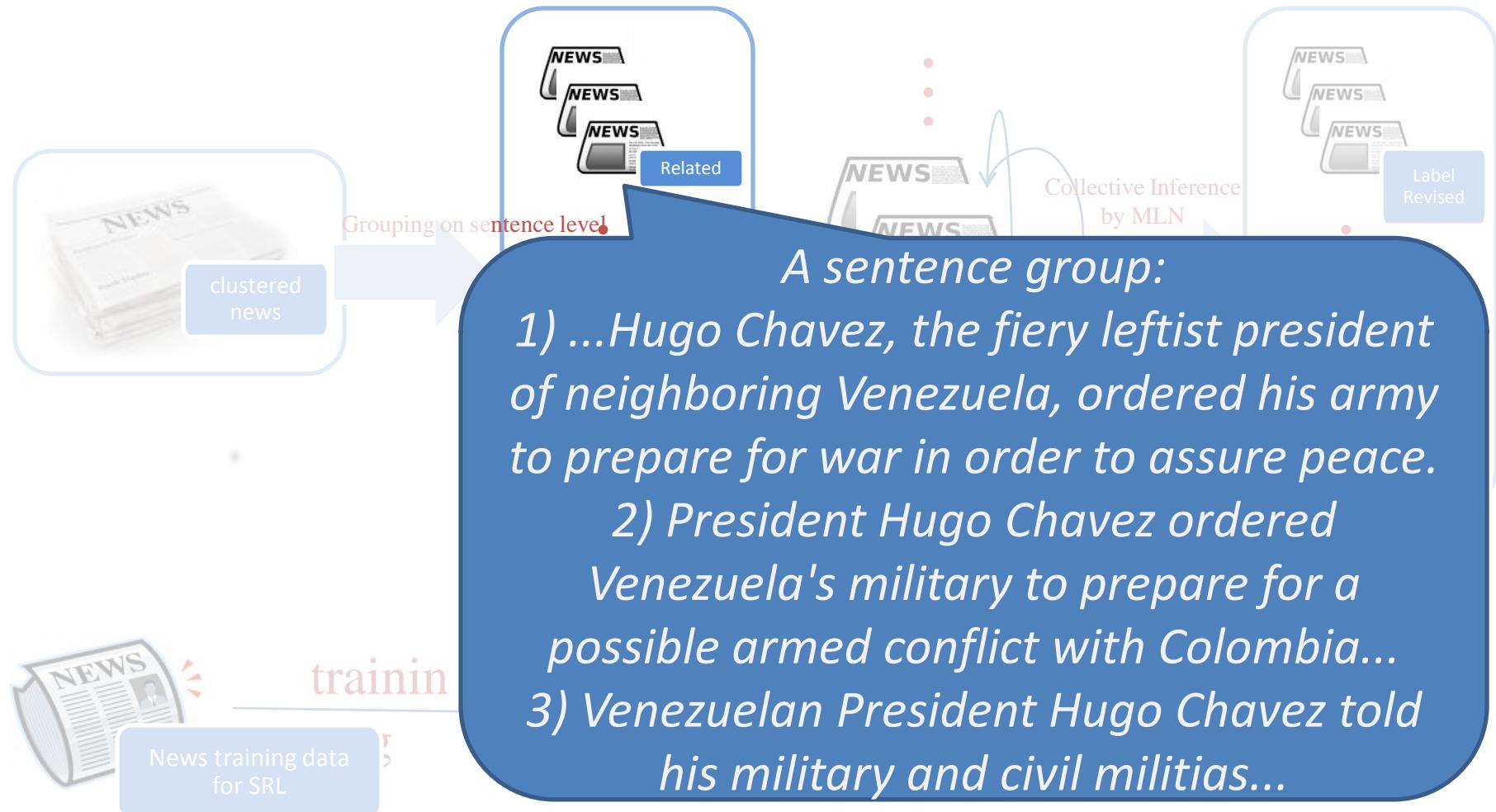
(Xiaohua Liu et al., 2010)



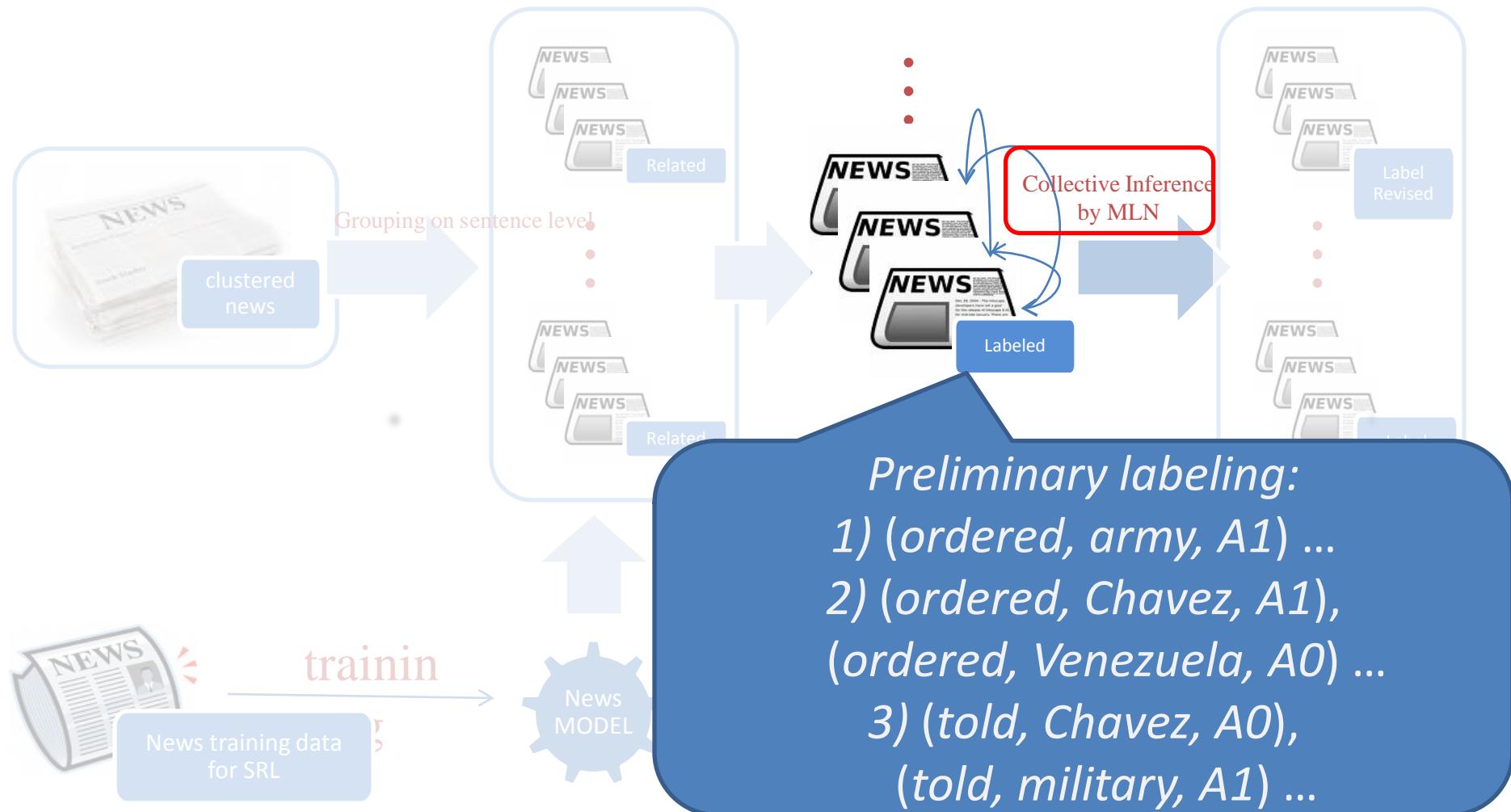
Implementation of Collective SRL



Implementation of Collective SRL



Implementation of Collective SRL



Implementation of collective SRL

Collective inference with MLN:

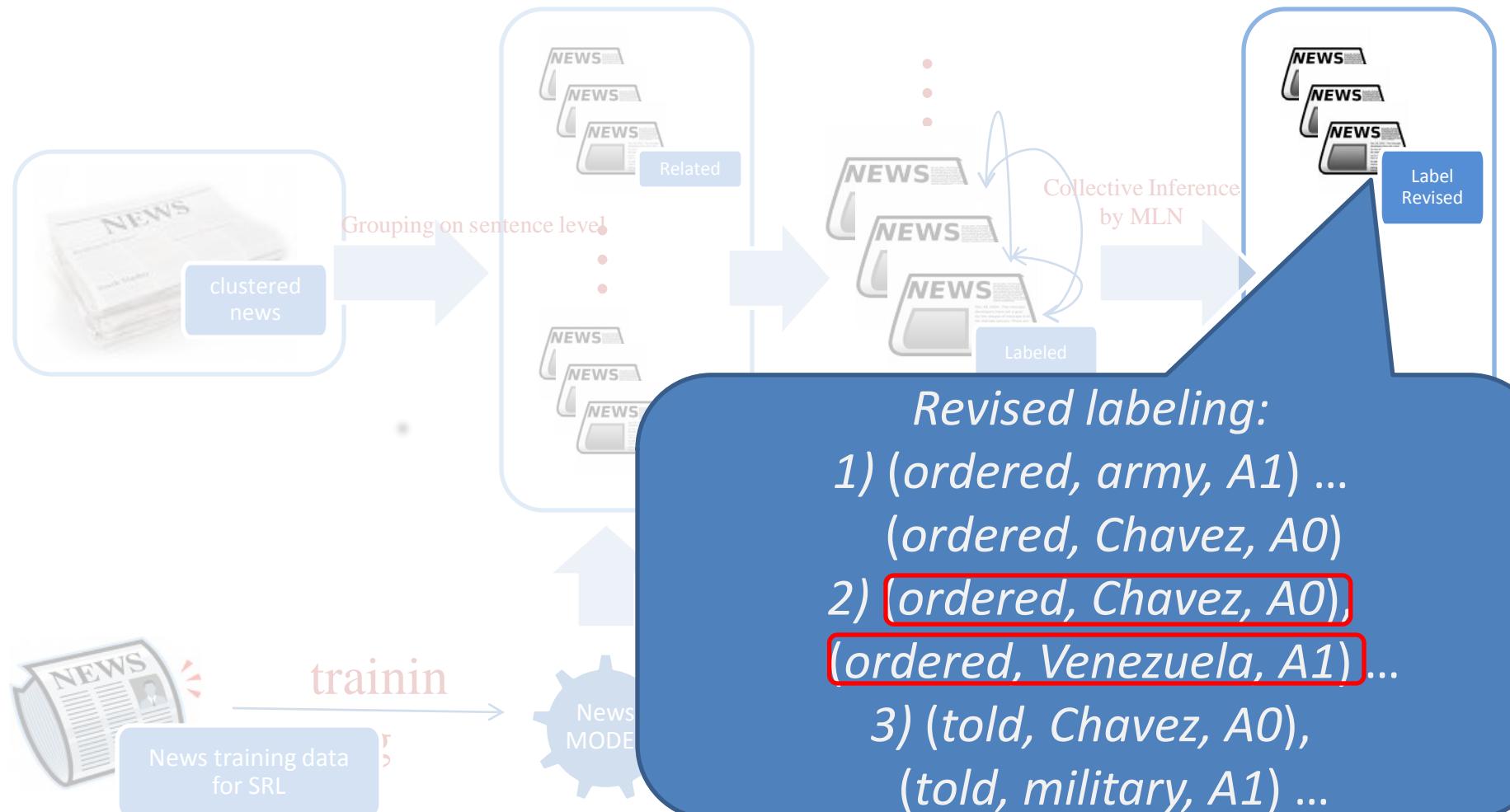
introduce two formulas (the second is for collective inference)

$$\text{role}(s, p, a, +r) \Rightarrow \text{final_role}(s, p, a, +r) \quad (1)$$

$$\begin{aligned} s_1 \neq s_2 \wedge & \text{lemma}(s_1, p_1, p_lemma) \wedge \text{lemma}(s_2, p_2, p_lemma) \\ & \wedge \text{lemma}(s_1, a_1, a_lemma) \wedge \text{lemma}(s_2, a_2, a_lemma) \\ & \wedge \text{role}(s_2, p_2, a_2, +r) \Rightarrow \text{final_role}(s_1, p_1, a_1, +r) \quad (2) \end{aligned}$$



Implementation of Collective SRL

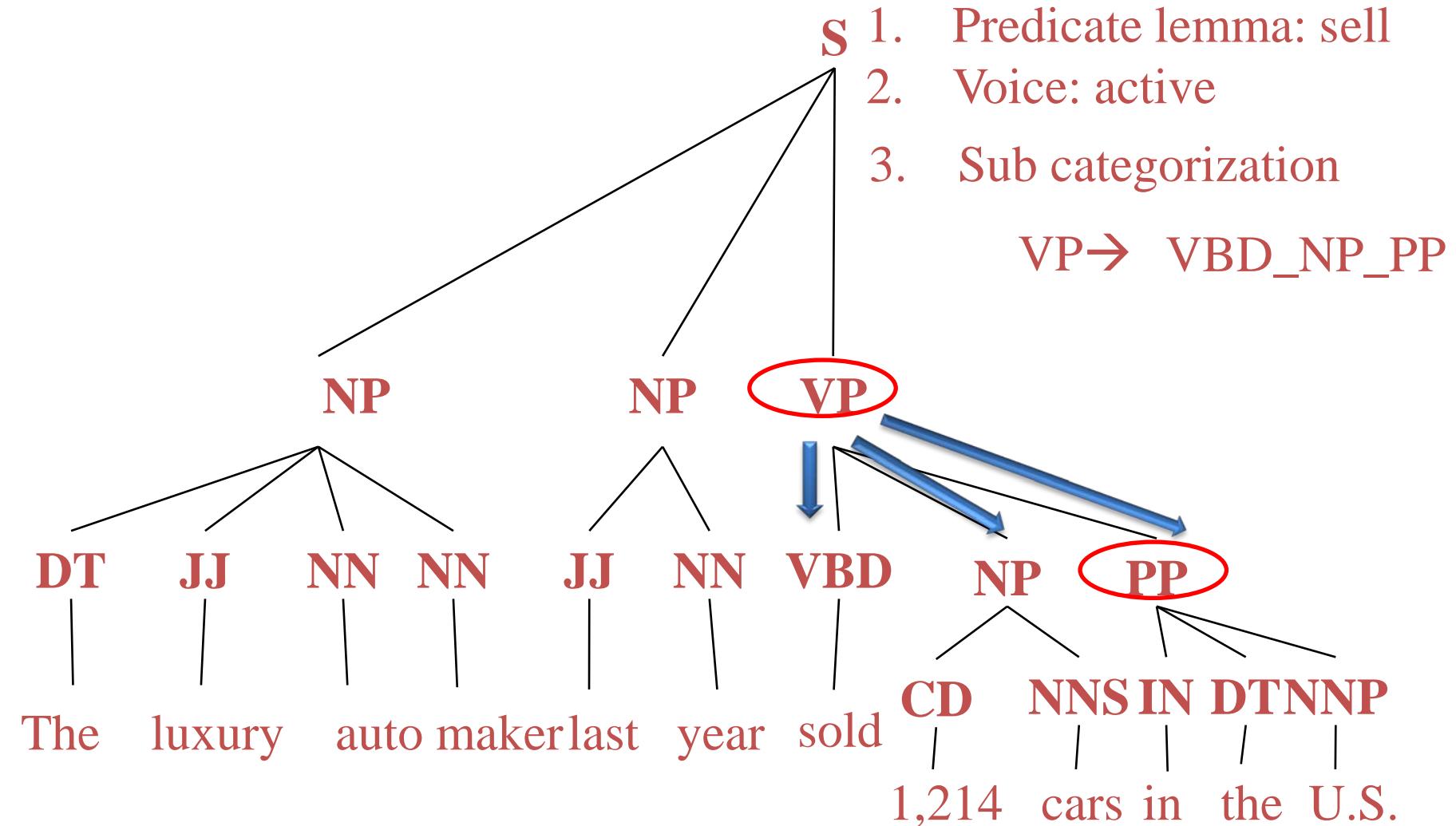


Experimental Results of Collective SRL

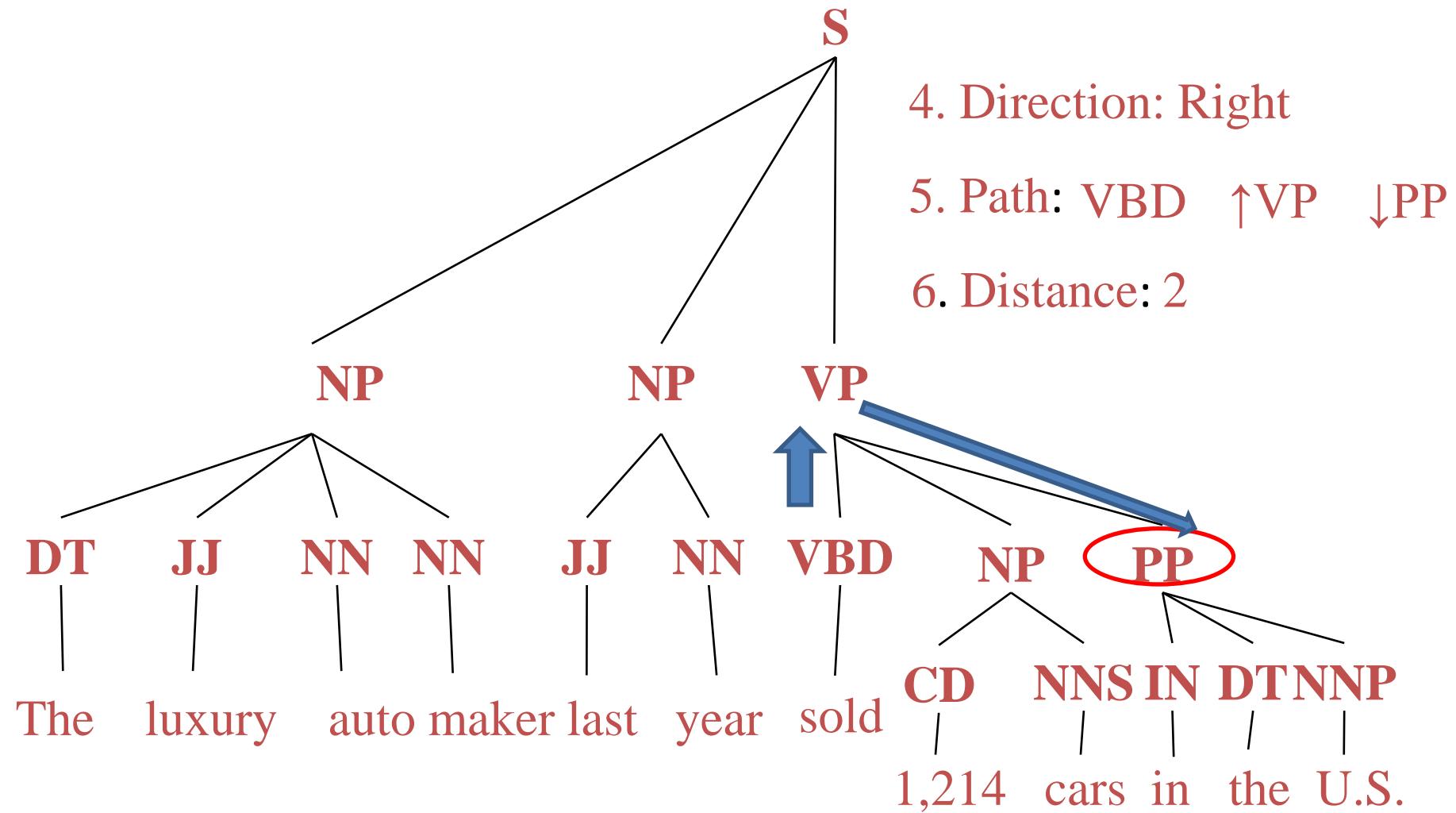
- Data
 - 1000 sentences from news clusters, grouped into 200 clusters
- Results (10-fold cross validation)

Systems	Precision	Recall	F-Score
Baseline	69.87%	59.26%	64.13%
Our method	67.01%	68.33%	67.66%

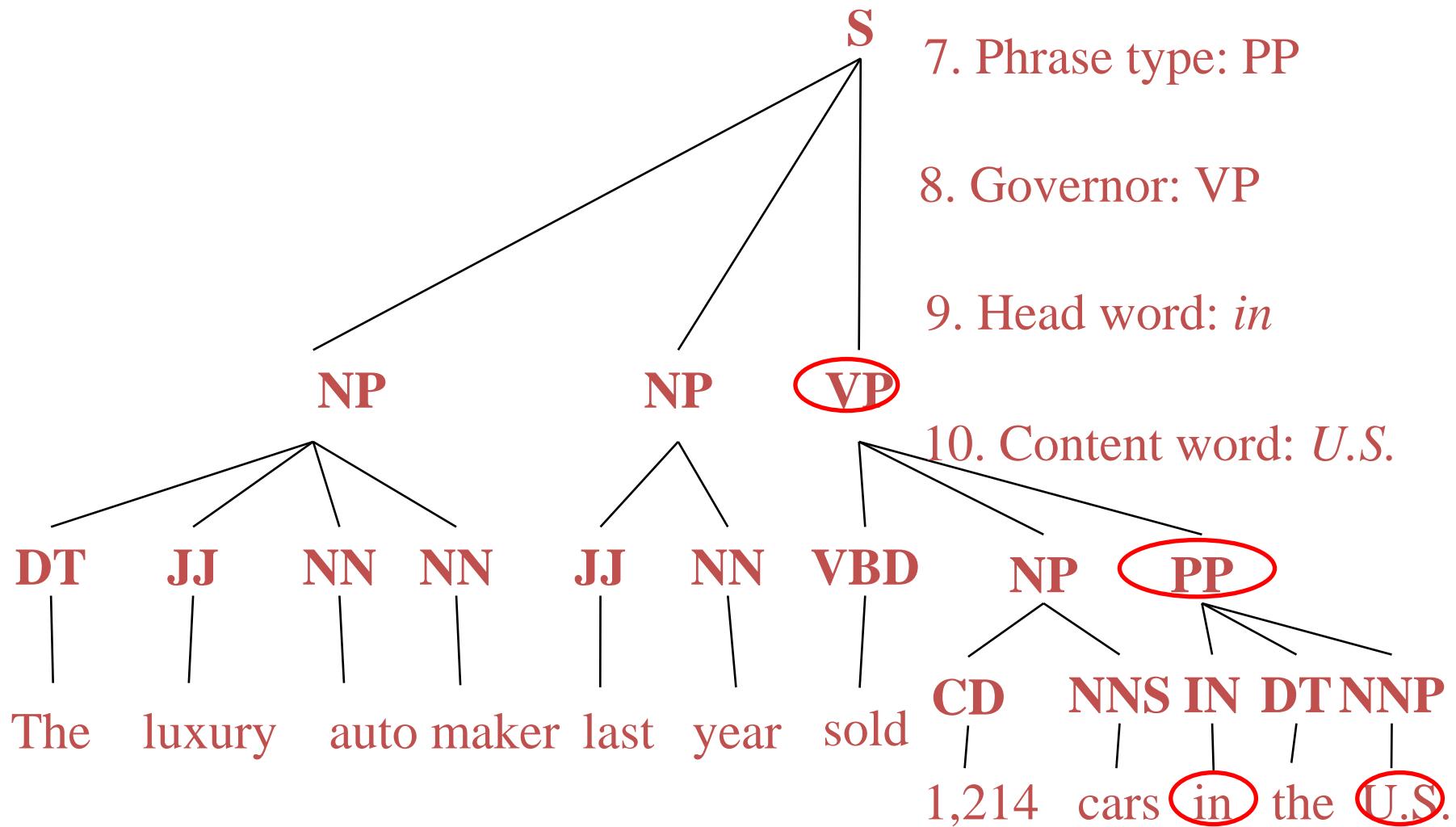
Feature Engineering in SRL



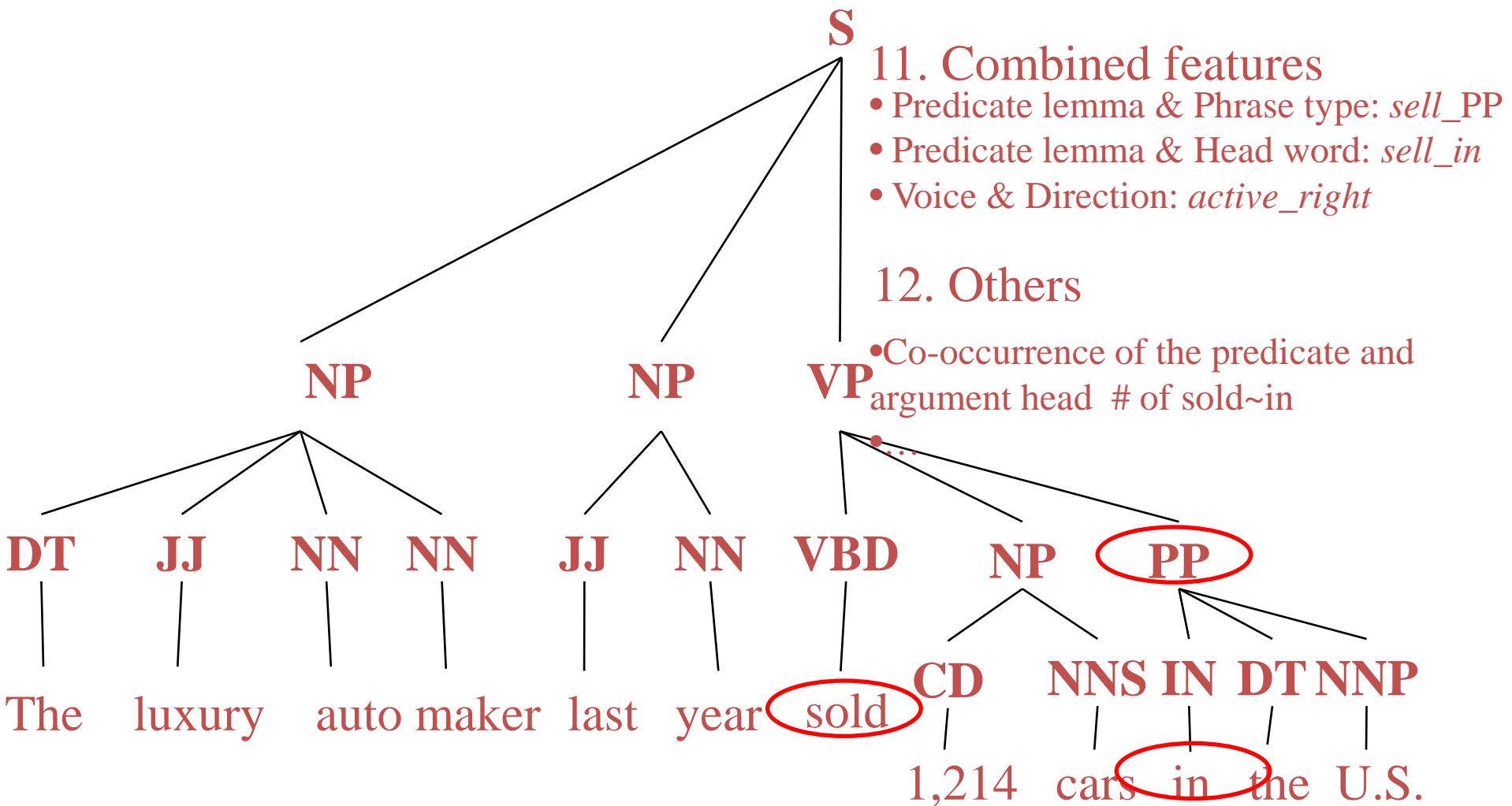
Feature Engineering in SRL



Feature Engineering in SRL



Feature Engineering in SRL



Outline

- Introduction
 - SRL Task definition
 - Application to twitter search
- General approaches to SRL
 - Resources
 - Typical systems
- SRL on tweets
 - Challenges
 - Method

Task Definition of Tweet Level SRL

- Input: a tweet
 - *oh yea and Chile **earthquake** the earth off it's axis according to NASA and **shorten** the **day** by a wee second :-(*
- Output: predicate-argument structures
 - (shorten, earthquake, A0), (shorten, day, A1)

Research Challenges

- SRL system for news does not work: F1 90.0%
→ 43.3%
 - Reason: tweets are greatly different from news in written styles
 - Formal vs. informal; and Human edited vs. freely written; long vs. short;
 - Question: how to leverage existing SRL resources?

Research Challenges

- Building a SRL for tweets requires a huge number of training data
 - Manually labeling is prohibitively affordable
 - Question: can we train a system without much human labeling?

Outline

- Introduction
 - SRL Task definition
 - Application to twitter search
- General approaches to SRL
 - Resources
 - Typical systems
- SRL on tweets
 - Challenges
 - Method

SRL for Tweets

- SRL for news tweets
 - Focus on news tweets, tweets that report news
 - Relatively formal and less noisy
 - Easy to leverage related news
 - Using the redundancy between news and news tweets
(Xiaohua Liu et al., 2010)
- SRL for general tweets
 - Collective SRL using clustering **(Xiaohua Liu et al., 2011)**
 - Enhancing SRL for Tweets using Self-training **(Xiaohua Liu et al., 2011)**

SRL for News Tweets

Key Observations(1)

- There are strong content connection between news and tweets
 - Tweets directly excerpted from news articles or Links in tweets point to news articles

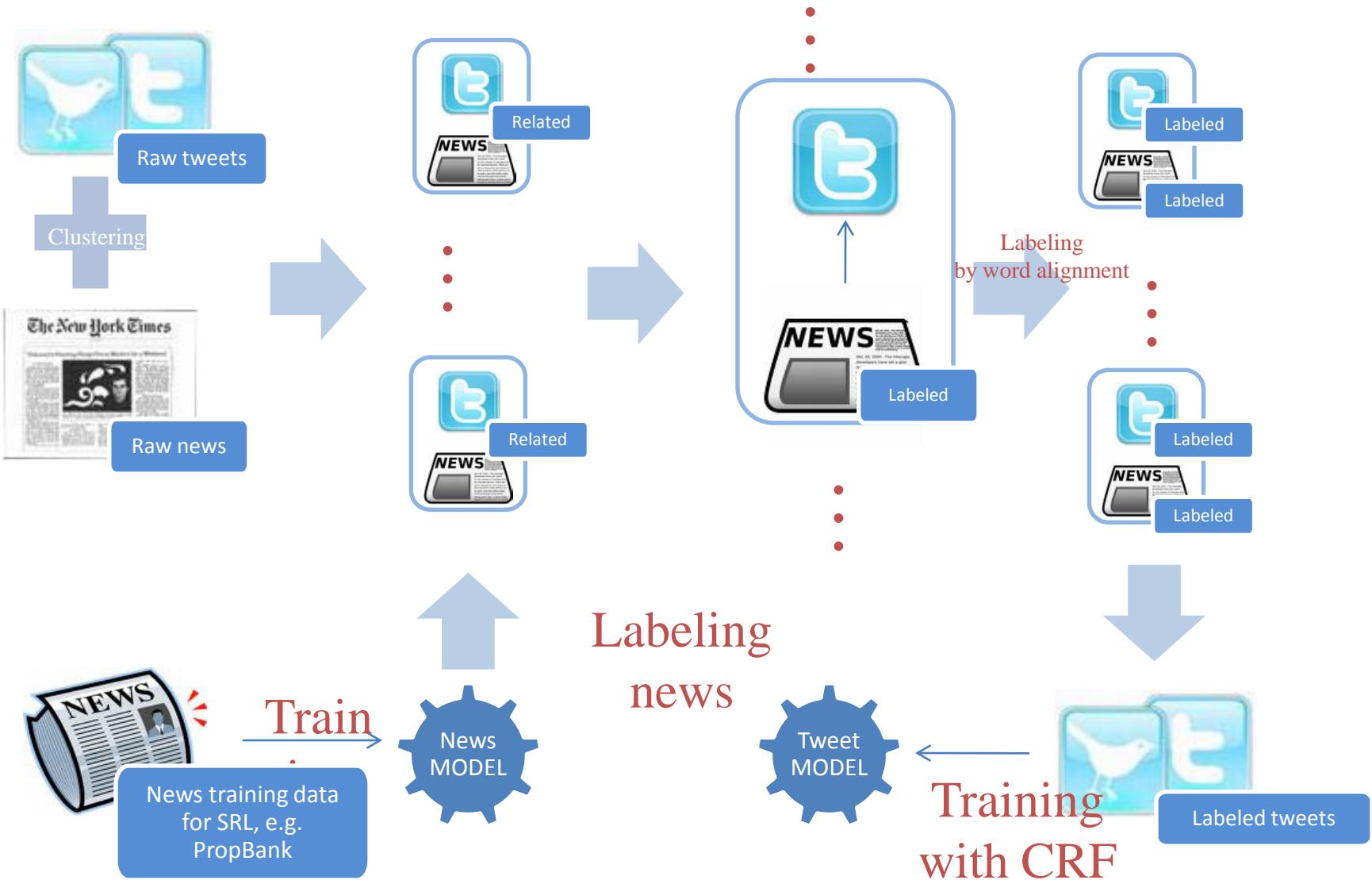
A screenshot showing a tweet and its source news article. The tweet is from a user named 'pctquimelp' (@pctquimelp) and reads: "Obama Limits When U.S. Would Use Nuclear Arms - http://nyti.ms/9mo4Dj". Below the tweet, it says "16 minutes ago from TimesPeople · Reply · View Tweet". A red arrow points from the link in the tweet to the URL in the news article below. The news article title is "Obama Limits When U.S. Would Use Nuclear Arms" (underlined). It was published on April 6, 2010. The article text begins: "In the year since Mr. Obama gave a speech in Prague declaring that he would shift the policy of the United States toward the elimination of nuclear weapons, his staff has been meeting — and arguing — over how to turn that commitment into a workable policy, without undermining the credibility of the country's nuclear deterrent." To the right of the article, there is a sidebar with options: COMMENTS (527), SIGN IN TO E-MAIL, PRINT, SINGLE PAGE, and REPRINTS.

- Official news that follow hot tweets
 - E.g., For *Chile earthquake* on March 2nd, 2010, 261 news and 722 news tweets published on the same day that described this event

Key Observations(2)

- News and tweets that describe similar content often have similar predicate argument structures
 - Chile *Earthquake Shortened Earth Day*
 - Chile *Earthquake Shortened Day*
 - *oh yea and Chile earthquake the earth off it's axis according to NASA and shorten the day by a wee second :-(*

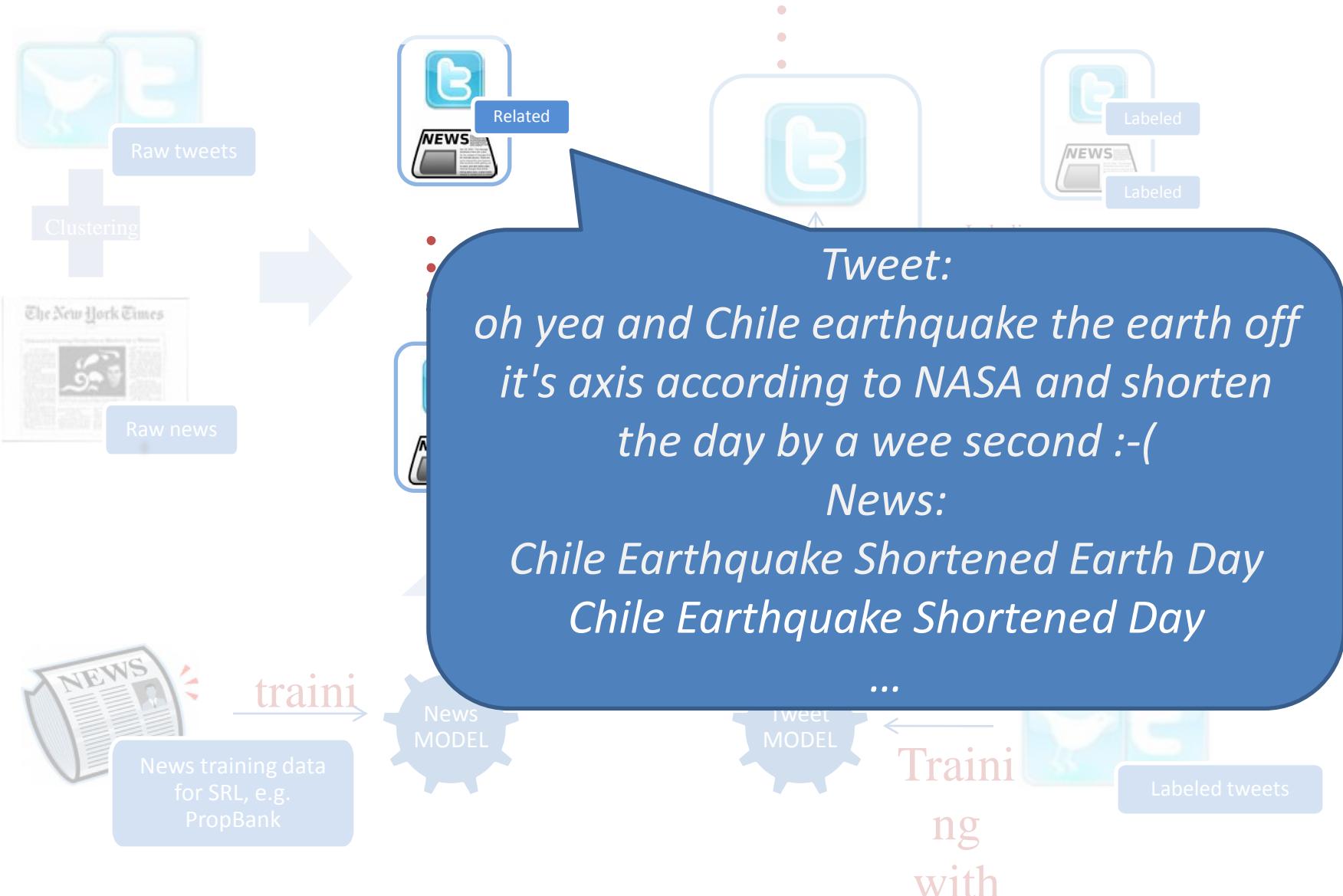
Label News Tweets with News



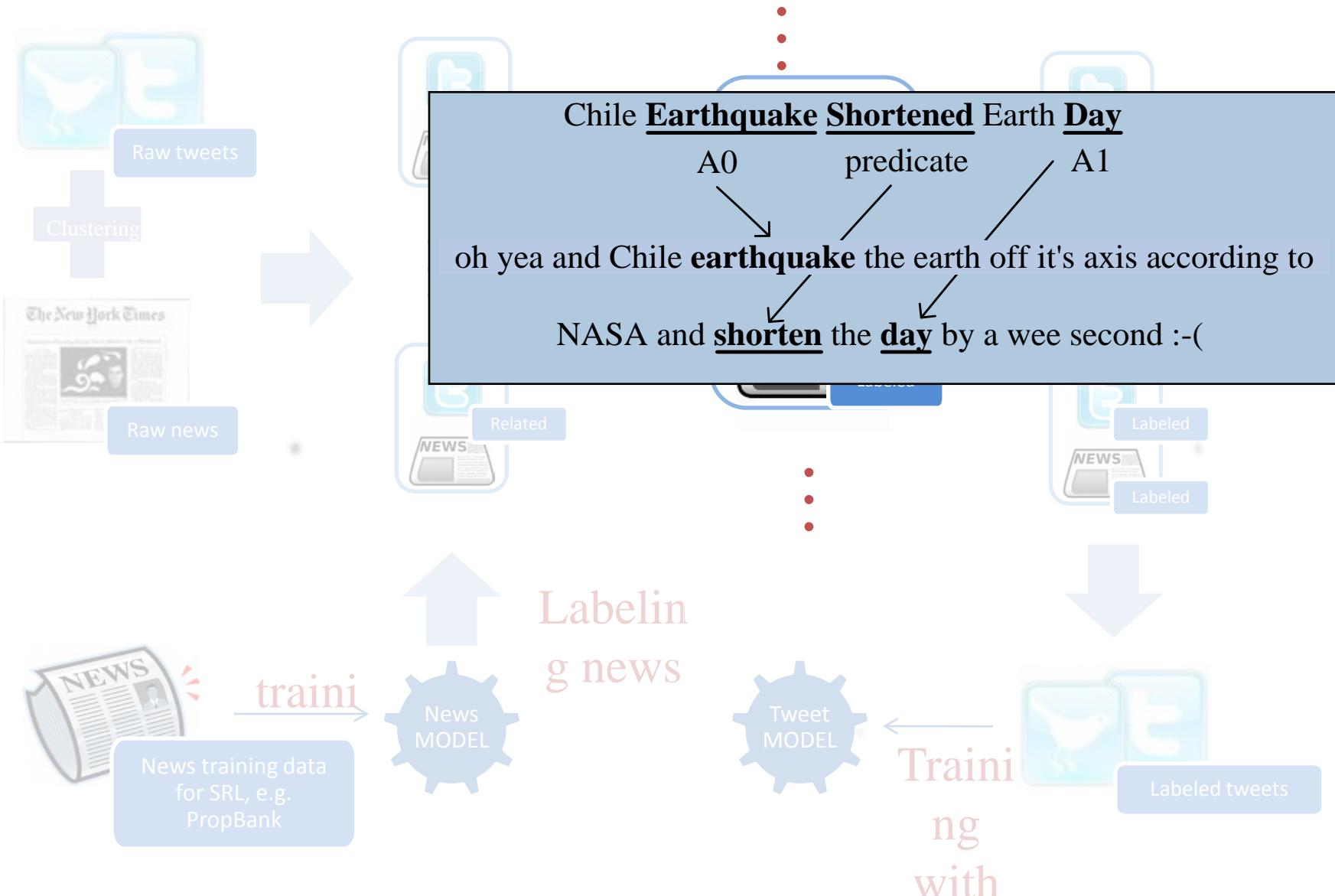
Label News Tweets with News



Label News Tweets with News



Label News Tweets with News



Label News Tweets with News

:

Conflict resolution

Conflicts are cases that violate any of the two structure constraints(Meza-Ruiz and Riedel, 2009)

1. one (predicate, argument) pair has only one role label in one sentence;

E.g., *(shorten, earthquake, A0) vs.*
(shorten, earthquake, A1)

2. one predicate can have each of the proper arguments (A0~A5) once at most in one sentence.

E.g., *(shorten, earthquake, A0) vs. (shorten, axis, A0),*

News training data
for SRL, e.g.
PropBank



Training
with

Labeled tweets

Experiment Setting

- Evaluation metric: precision, recall and F1
- Baseline: SRL system trained on news (Meza-Ruiz and Riedel, 2009)
- Data preparation
 - Training dataset: 10,000 mechanically labeled tweets
 - Testing dataset: 1,110 human labeled tweets

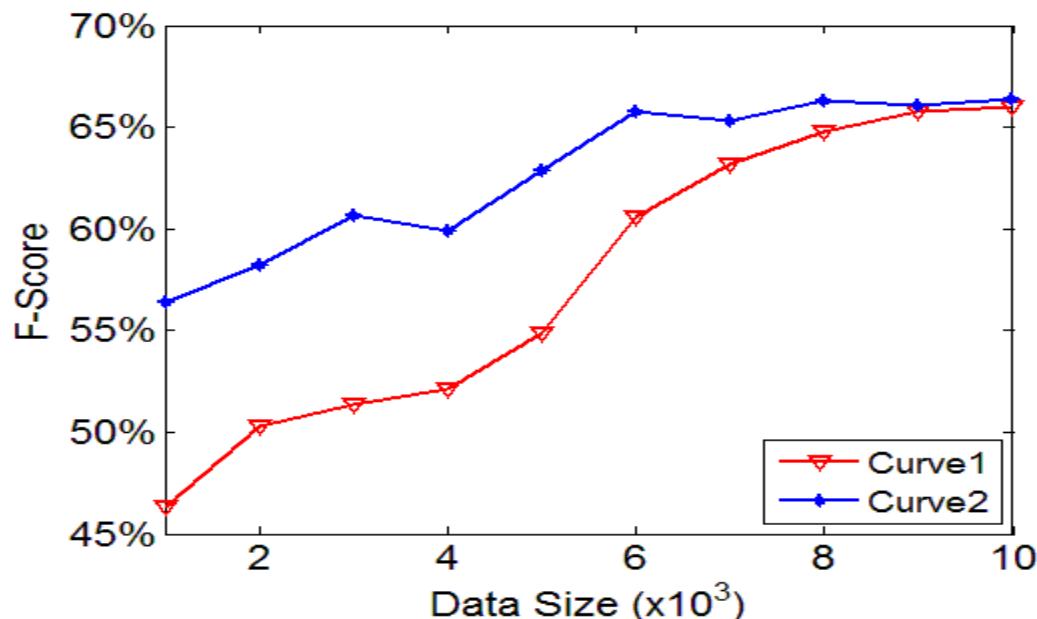
Experimental Results (1)

- Basic results
 - SRL-TS: our system; SRL-BS: baseline

	Precision	Recall	F1
SRL-BS	36.0 %	54.5%	43.3%
SRL-TS	78.0%	57.1%	66.0%

Experimental Results (2)

- Influence of training data size
 - Curve1: no test data is used for training
 - Curve2: half of the test data is used as training data



Collective SRL for Tweets with Clustering

Observation

- Great redundancy in tweets, i.e., the same predicate-argument structures occur in multiple tweets.
- Example:
 - oh yea and Chile earthquake the earth off it's axis according to NASA and shorten the day by a second :-
 - Chile Earthquake Shortened Earth Day
 - item Chile Earthquake Shortened Day

Collective SRL with Clustering

- Two round labeling to leverage the redundancy in tweets
 - #1 round: labeling on single tweet level using a conventional CRF based labeler
 - #2 round: labeling on cluster level using a CRF based labeler enhanced by cluster-level features
- Cluster-level features
 - The top 3 most frequent roles played by the word in the cluster

Algorithm 1 Collective SRL for tweets with Clustering.

Require: Tweet stream i ;sequential labelers l_1, l_2 ;output stream o .

```
1: Initialize clusters  $cl: cl = \emptyset$ .
2: while Pop a tweet  $t$  from  $i$  and  $t \neq null$  do
3:   Put  $t$  to a cluster  $c: (c, cl) = cluster(cl, t)$ .
4:   if  $|c| > N$  then
5:     Initialize cache of labeled results  $cs: cs = \emptyset$ .
6:     for  $\forall t' \in c$  do
7:       Label  $t'$  with  $l_1:(t', \{(p, s, cf)\}) = label(l_1, t')$ .
8:       for  $\forall cf \in \{(p, s, cf)\} > \alpha$  do
9:         Cache labeled results: $cs = cs \cup \{(t', p, s, cf)\}$ .
10:      end for
11:    end for
12:    for  $\forall t' \in c$  do
13:      Label  $t'$  with  $l_2:(t', \{(p, s, cf)\}) = label(cs, l_2, t')$ .
14:      Output labeled results  $(t', \{(p, s, cf)\})$ .to  $o$ .
15:    end for
16:    Remove  $c$  from  $cl: cl = cl - \{c\}$ .
17:  end if
18: end while
19: for  $\forall c \in cl, \forall t' \in c$  do
20:   Label  $t'$  with  $l_1:(t', \{(p, s, cf)\}) = label(l_1, t')$ .
21:   Output labeled results  $(t', \{(p, s, cf)\})$  to  $o$ .
22: end for
23: return  $o$ .
```

Implementation Details

- The CRF labelers
 - Training : CRF++
 - Decoder: Viterbi algorithm
 - Feature extraction: OpenNLP and the Stanford parser
- Clustering
 - Bottom-up online clustering based on merge
 - Features: bag-of-words
 - Similarity computing: cosine similarity, i.e.,

$$sim(\vec{t}_1, \vec{t}_2) = \frac{\vec{t}_1 \cdot \vec{t}_2}{|\vec{t}_1| |\vec{t}_2|}$$

Algorithm 2 Clustering a tweet.

Require: Clusters cl ; tweet for clustering t .

- 1: Get the reference of the most near cluster c^* , for $t:c^* = \operatorname{argmax}_{c' \in cl} sim(t, c')$.
 - 2: Get the similarity s between t and c^* : $s = sim(t, c^*)$.
 - 3: **if** $s < \beta$ **then**
 - 4: Create a new cluster for t : $c^* = \{t\}$.
 - 5: Add c^* to cl : $cl = cl \cup \{c^*\}$.
 - 6: **if** $|cl| > M$ **then**
 - 7: Merge clusters: $cl = merge(cl)$.
 - 8: **end if**
 - 9: **end if**
 - 10: **return** c^* and cl .
-

Experiment Setting

- Dataset: 6,670 manually annotated dataset, randomly divided into three parts
 - 1,000 for development
 - 2,394 for training
 - The remaining for testing
- Baseline: a conventional CRF based labeler
 - BIO labeling schema: ...<B-A0>earthquake<O> shorten<B-A1>day...
 - Features: lemma/POS tag of the current/previous/next token, the lemma of the predicate and its combination with the lemma/POS tag of the current token, the voice of the predicate (active/passive), the distance between the current token and the predicate, and the relative position of the current token to the predicate
- Evaluation metrics: Precision, Recall and F1

Experiment Results

- An absolute gain of 3.1% in terms of average F1 measure

System	Pre. (%)	Rec. (%)	F1 (%)
SRL_{CL}	61.9	56.7	59.2
SRL_{BA}	62.7	50.8	56.1

- Main error sources:
 - Irregular words in tweets: "...thank **youuuu sweedie pops..."**
 - Unknown words in tweets: "**Bacteria** in the gut shown to lower obesity..."

Enhancing SRL for Tweets with Self-training

Motivation

- Use abundant unlabeled data to overcome the lack of annotated tweets
 - Repeatedly re-train the model using the tweets labeled by itself
 - Consider two factors while selecting: correctness of labeling and informativeness

Algorithm 1 Self-training based SRL for tweets.

Require: Tweet stream i ; training tweets ts ; output stream o .

- 1: Initialize two CRF based labelers l and l' : $(l, l') = train(cl)$.
 - 2: Initialize the number of new accumulated tweets for training n : $n = 0$.
 - 3: **while** Pop a tweet t from i and $t \neq null$ **do**
 - 4: Label t with l : $(t, \{(p, s, cf)\}) = label(l, c, t)$.
 - 5: Label t with l' : $(t, \{(p, s, cf)\}') = label(l', c, t)$.
 - 6: Output labeled results $(t, \{(p, s, cf)\})$ to o .
 - 7: **if** $select(t, \{(p, s, cf)\}, \{(p, s, cf)\}')$ **then**
 - 8: Add t to training set ts : $ts = ts \cup \{t, \{(p, s, cf)\}\}$; $n = n + 1$.
 - 9: **end if**
 - 10: **if** $n > N$ **then**
 - 11: Retrain labelers: $(l, l') = train(cl)$; $n = 0$.
 - 12: **end if**
 - 13: **if** $|ts| > M$ **then**
 - 14: shrink the training set: $ts = shrink(ts)$.
 - 15: **end if**
 - 16: **end while**
-

Features

- Lemma/POS tag of the current/previous/next token
- Lemma of the predicate and its combination with the lemma/POS tag of the current token
- The voice of the predicate (active/passive)
- The distance between the current token and the predicate
- The relative position of the current token to the predicate
- Dependencies parsing related features

Selection Criteria

- Select such tweets that the performance can be most improved if they are selected for training
 - Correctness : These tweets are correctly labeled
 - Train two independent models and consider a tweet is labeled correctly if the two models give the same results confidently
 - Informativeness: These tweets provide new information to the existing training set
 - If a labeled tweet is not much similar to any tweet in the training set

Algorithm 2 Selection of a training tweet.

Require: Training tweets ts ; tweet t ; labeled results by l $\{(p, s, cf)\}$; labeled results by l' $\{(p, s, cf)\}'$.

- 1: **if** $\{(p, s, cf)\} \neq \{(p, s, cf)\}'$ **then**
- 2: **return** FALSE.
- 3: **end if**
- 4: **if** $\exists cf \in \{(p, s, cf)\} \cup \{(p, s, cf)\}' < \alpha$ **then**
- 5: **return** FALSE.
- 6: **end if**
- 7: **if** $\exists t' \in ts \text{ sim}(t, t') > \beta$ **then**
- 8: **return** FALSE.
- 9: **end if**
- 10: **return** TRUE.

Experiment Setting

- Dataset: 7,171 manually annotated dataset, randomly divided into three parts
 - 583 as seeds
 - 5,421 for self-training development
 - The remaining for blind testing
- Baseline: a conventional CRF based labeler
- Evaluation metrics: Precision, Recall and F1

Experiment Results

- A gain of 3.4% F1
- Effects of correctness
- Effects of informativeness

Table 2: Basic experimental results.

System	Pre.(%)	Rec.(%)	F1(%)
SRL_{SE}	59.2	45.9	51.7
SRL_{BA}	46.7	50.0	48.3

System	Pre.(%)	Rec.(%)	F1(%)
SRL_{SE}	59.2	45.9	51.7
SRL_{SE-C}	48.4	36.5	41.6

System	#T	F1(%)
SRL_{SE}	2,557	51.7
SRL_{RD}	2,557	47.7
SRL_{SP}	4,000	42.5
SRL_{CF}	4,277	44.9

References

- Màrquez, Lluís. 2009. *Semantic Role Labeling Past, Present and Future*, Tutorial of ACL-IJCNLP 2009.
- Meza-Ruiz, Ivan and Sebastian Riedel. 2009. Jointly Identifying Predicates, Arguments and Senses using Markov Logic. *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, pages: 155-163.
- Xiaohua Liu, Kuan Li, Bo Han, Ming Zhou, Long Jiang, Daniel Tse and Zhongyang Xiong. 2010. Collective Semantic Role Labeling on Open News Corpus by Leveraging Redundancy. COLING 2010
- Xiaohua Liu, Kuan Li, Bo Han, Ming Zhou, Long Jiang, Zhongyang Xiong and Changning Huang. 2010. Semantic Role Labeling for News Tweets. COLING 2010
- Xiaohua Liu, Kuan Li, Ming Zhou and Zhongyang Xiong. 2011. Collective Semantic Role Labeling for Tweets with Clustering. IJCAI 2011
- Xiaohua Liu, Kuan Li, Ming Zhou and Zhongyang Xiong. 2011. Enhancing Semantic Role Labeling for Tweets with Self-training. AAAI 2011

Agenda

- QuickView: A Research Platform of SNS Text Mining and Search
- SNS Text Mining
 - Semantic Role Labeling
 - Sentiment Analysis
- Future Work

Outline

- Sentiment analysis
 - Introduction
 - Definition, application, components
 - Approaches for SA subtasks
 - Holder detection
 - Target detection
 - Polarity classification
- Twitter Sentiment Analysis
 - Goals and challenges
 - Existing systems
 - Target-dependent twitter sentiment analysis

Outline

- **Sentiment analysis**
 - Introduction
 - Definition, application, components
 - Approaches for SA subtasks
 - Holder detection
 - Target detection
 - Polarity classification
- Twitter Sentiment Analysis
 - Goals and challenges
 - Existing systems
 - Target-dependent twitter sentiment analysis

Sentiment Analysis (SA)

- Also known as **opinion mining**: to understand the attitude of a speaker or a writer with respect to some topic
 - The attitude may be their judgment or evaluation, their affective state or the intended emotional communication
 - Most popular classification of sentiment: positive or negative
- For example
 - The pictures are very clear.
 - *In his recent State of the Union address, US President Bush quite unexpectedly labeled Iran, Iraq, and the DPRK as an “axis of evil”.*

Applications of SA

- Business intelligence system
- Purchase planning
- Public opinion management
- Web advertising

Sentiment Components

- **Holder**
 - who expresses the sentiment
- Target
 - what the sentiment is expressed to
- Polarity
 - the nature of the sentiment (e.g., **positive/negative**)
- *In his recent State of the Union address, US President Bush quite unexpectedly labeled Iran, Iraq, and the DPRK as an “axis of evil”.*

Outline

- Sentiment analysis
 - Introduction
 - Definition, application, components
 - Approaches for SA subtasks
 - Holder detection
 - Target detection
 - Polarity classification
- Twitter Sentiment Analysis
 - Goals and challenges
 - Existing systems
 - Target-dependent twitter sentiment analysis

Outline

- **Sentiment analysis**
 - Introduction
 - Definition, application, components
 - Approaches for SA subtasks
 - Holder detection
 - Target detection
 - Polarity classification
- Twitter Sentiment Analysis
 - Goals and challenges
 - Existing systems
 - Target-dependent twitter sentiment analysis

Holder Detection

- Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns
 - (Choi et al., HLT/EMNLP-05)

International officers believe that the EU will prevail.
International officers said **US officials** want the EU to prevail.

- View *source identification* as an information extraction task and tackle the problem using sequence tagging and pattern matching techniques simultaneously
 - *Linear-chain CRF model* to identify opinion sources
 - Patterns incorporated as features

CRF for Holder Detection

- Given a sentence X, to seek for a label sequence Y that maximizes

$$P(y|x) = \frac{1}{Z_x} \exp\left(\sum_{i,k} \lambda_k f_k(y_{i-1}, y_i, x) + \sum_{i,k} \lambda'_k f'_k(y_i, x)\right)$$

- y_i belongs to {‘S’, ‘T’, ‘-’}
- λ_k and λ'_k are parameters, f_k and f'_k are feature functions
- Z_x is the normalization factor

International	officers	believe	that	the	EU	will	prevail
S	T	-	-	-	-	-	-

Basic Features

- **Capitalization features:** all-capital, initial-capital
- **Part-of-speech features $([-2,+2])$:** noun, verb, adverb, wh-word, determiner, punctuation, etc
- **Opinion lexicon features:** $[-1,+1]$ whether or not the word is in the *opinion lexicon*
- **Dependency tree features**
 - the grammatical role of its chunk
 - the grammatical role of x_{i-1} 's chunk
 - whether the parent chunk includes an opinion word
 - whether x_i 's chunk is in an argument position with respect to the parent chunk
 - whether x_i represents a constituent boundary
- **Semantic class features:** the semantic class of each word: authority, government, human, media, organization or company, proper name, and other

Extraction Pattern Learning

- Looking at the context surrounding each answer and proposes a lexico-syntactic pattern
 - *[They]_h complained about the deficiencies of the benefits given to them.*
 - *<subj> complained*
- Compute the probability that the pattern will extract an opinion source

$$P(\text{source} \mid \text{pattern}_i) = \frac{\text{correct sources}}{\text{correct sources} + \text{incorrect sources}}$$

Extraction Pattern Features

- Four IE pattern-based features for each token x_i
 - SourcePatt-Freq, SourcePatt-Prob,
 - SourceExtr-Freq, SourceExtr-Prob
 - Where
 - SourcePatt indicates whether a word activates any source extraction pattern. E.g., “complained” activates the pattern “ $\langle \text{subj} \rangle$ complained”
 - SourceExtr indicates whether a word is extracted by any source pattern. E.g., “They” would be extracted by the “ $\langle \text{subj} \rangle$ complained”

Experimental Results

- MPQA data
 - In total, 535 documents where targets are annotated by human
 - 135 as development set and feature engineering, and the remaining 400 for evaluation, performing 10-fold cross validation
- 3 measures: overlap match (OL), head match (HM), and exact match (EM)

		Recall	Prec	F1
Extraction Patterns	OL	48.5	81.3	60.8
	HM	46.9	78.5	58.7
	EM	41.9	70.2	52.5
CRF: basic features	OL	56.1	81.0	66.3
	HM	55.1	79.2	65.0
	EM	50.0	72.4	59.2
CRF: basic + IE pattern features	OL	59.1	82.4	68.9
	HM	58.1	80.5	67.5
	EM	52.5	73.3	61.2

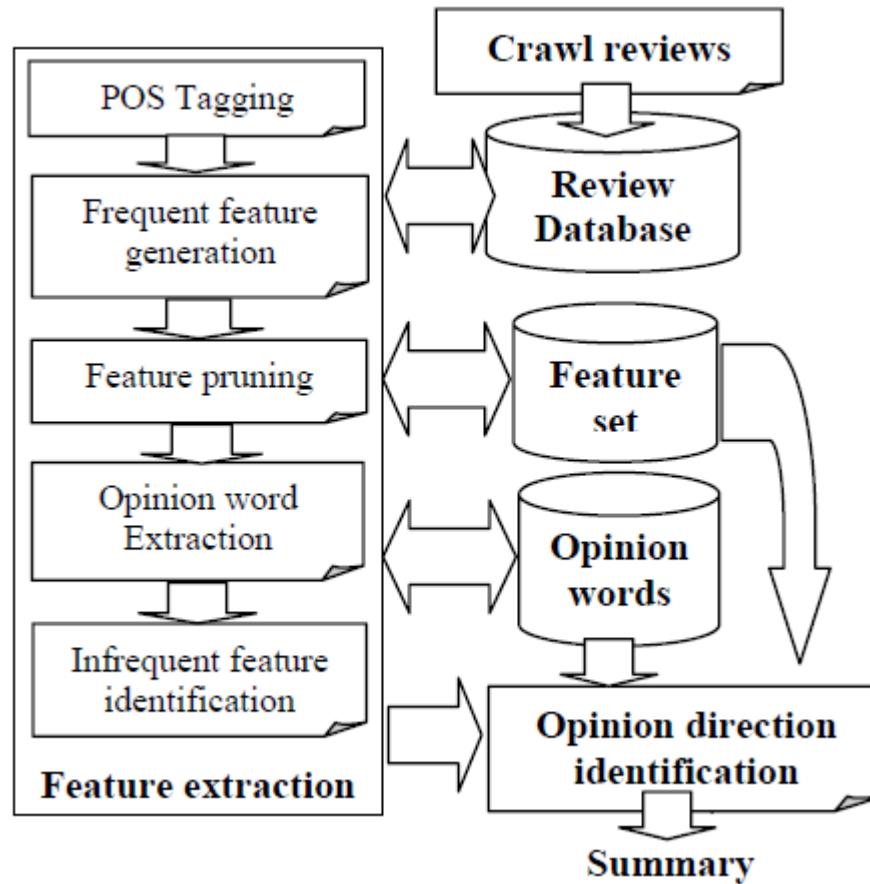
Outline

- **Sentiment analysis**
 - Introduction
 - Definition, application, components
 - Approaches for SA subtasks
 - Holder detection
 - Target detection
 - Polarity classification
- Twitter Sentiment Analysis
 - Goals and challenges
 - Existing systems
 - Target-dependent twitter sentiment analysis

Target Detection

- Mining Opinion Features in Customer Reviews
 - (Minqing Hu and Bing Liu, AAAI 2004)
 - *Explicit feature*
 - *The pictures are very clear.*
 - *Implicit feature*
 - *While light, it will not easily fit in pockets. (size)*
- Task definition
 - Given a product name and all the reviews of the product, to find the features of the product that appear explicitly as nouns or noun phrases in the reviews

Approach Overview



Frequent Features Detection

- Association rule mining
 - Find frequent features with three words or fewer
 - Appears in more than 1% of the review sentences (minimum support)
- Feature Pruning
 - Compactness: compact in at least 2 *sentences*
 - *p-support (pure support)*: a p-support lower than the minimum p-support (3)

Infrequent Feature Detection

- People use the same adjective words to describe different subjects
 - “*Red eye* is very *easy* to correct.”
 - “The camera comes with an excellent *easy* to install *software*”
 - “The *pictures* are absolutely *amazing*”
 - “The *software* that comes with it is *amazing*”

Infrequent Feature Detection

- Opinion word identification
 - For each sentence in the review database, if it contains any frequent feature, extract the nearby *adjective* as opinion word
- Infrequent feature detection
 - For each sentence in the review database, if it contains no frequent feature but one or more opinion words, find the *nearest noun/noun phrase of the opinion word* as an infrequent feature

Experimental Results

- Data: customer reviews of five electronics products from Amazon.com and C|net.com

Product name	No. of manual Features	Frequent features (association mining)		Compactness pruning		P-support pruning		Infrequent feature identification	
		Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
Digital camera1	79	0.671	0.552	0.658	0.634	0.658	0.825	0.822	0.747
Digital camera2	96	0.594	0.594	0.594	0.679	0.594	0.781	0.792	0.710
Cellular phone	67	0.731	0.563	0.716	0.676	0.716	0.828	0.761	0.718
Mp3 player	57	0.652	0.573	0.652	0.683	0.652	0.754	0.818	0.692
DVD player	49	0.754	0.531	0.754	0.634	0.754	0.765	0.797	0.743
Average	69	0.68	0.56	0.67	0.66	0.67	0.79	0.80	0.72

Outline

- **Sentiment analysis**
 - Introduction
 - Definition, application, components
 - Approaches for SA subtasks
 - Holder detection
 - Target detection
 - Polarity classification
- Twitter Sentiment Analysis
 - Goals and challenges
 - Existing systems
 - Target-dependent twitter sentiment analysis

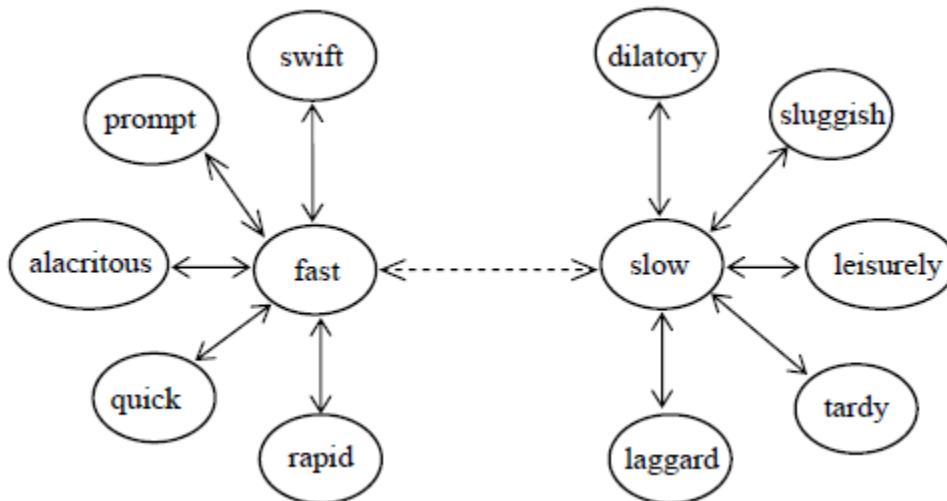
Lexicon Based Polarity Classification

- Mining and Summarizing Customer Reviews
 - (Hu and Liu, KDD-2004)
- Basic idea
 - Use the dominant orientation of the opinion words in the sentence to determine the orientation of the sentence.
 - That is, if positive/negative opinion prevails, the opinion sentence is regarded as a positive/negative one.

Lexicon Building

(Hu and Liu, KDD-2004)

- Utilize the adjective synonym set and antonym set in WordNet to predict the semantic orientations of adjectives
 - Adjectives share the same orientation as their synonyms and opposite orientations as their antonyms.
- Start with several seeds, iteratively expand to cover most opinion words



Hatzivassiloglou and McKeown (1997)

- Predicting the Semantic Orientation of Adjectives
 - (Hatzivassiloglou and McKeown, ACL-97)
- Assumption: adjectives connected by “and”/“but” tend to have same/opposite polarities

The tax proposal was

1. simple and well-received
2. simplistic but well-received
3. *simplistic and well-received

by the public.

ML-based Approaches for Polarity Classification

- Thumbs up? Sentiment Classification using Machine Learning Techniques
 - (Pang et al., 2002)
- Basic idea
 - Treat sentiment classification simply as a special case of topic-based categorization
 - With the two “topics” being positive sentiment and negative sentiment
 - Use three standard algorithms: Naive Bayes classification, maximum entropy classification, and support vector machines

Approach Details

- Document representation
 - Each document d is represented by a feature vector $\tilde{d} := (n_1(d), n_2(d), \dots, n_m(d))$
 - $n_i(d)$ could indicate presence, term frequency
- Classification algorithms
 - Naive Bayes, Maximum Entropy, SVM

Data

- Movie reviews
 - From Internet Movie Database (IMDb)
 - <http://www.cs.cornell.edu/people/pabo/movie-review-data/>
 - <http://reviews.imdb.com/Reviews/>
 - 700 positive / 700negative
- Experiment setting for ML classifiers
 - 3-fold cross validation
 - Treating punctuation as separate lexical items
 - No stemming or stoplists were used

Experimental Results

- Baseline: use a few words written by human to classify

	Proposed word lists	Accuracy
Human 1	positive: <i>dazzling, brilliant, phenomenal, excellent, fantastic</i> negative: <i>suck, terrible, awful, unwatchable, hideous</i>	58%
Human 2	positive: <i>gripping, mesmerizing, riveting, spectacular, cool, awesome, thrilling, badass, excellent, moving, exciting</i> negative: <i>bad, cliched, sucks, boring, stupid, slow</i>	64%

- ML-based methods

	Features	# of features	frequency or presence?	NB	ME	SVM
(1)	unigrams	16165	freq.	78.7	N/A	72.8
(2)	unigrams	"	pres.	81.0	80.4	82.9
(3)	unigrams+bigrams	32330	pres.	80.6	80.8	82.7
(4)	bigrams	16165	pres.	77.3	77.4	77.1
(5)	unigrams+POS	16695	pres.	81.5	80.4	81.9
(6)	adjectives	2633	pres.	77.0	77.7	75.1
(7)	top 2633 unigrams	2633	pres.	80.3	81.0	81.4
(8)	unigrams+position	22430	pres.	81.0	80.1	81.6

Other Related Approaches

- Topic sentiment mixture
 - Mei et al., 2007
- Semi-supervised approach
 - Li et al., 2010
- Domain Adaptation
 - Blitzer et al., 2007

Summary

1. Sentiment analysis refers to a set of subtasks
 - Holder, target, polarity
2. Sentiment analysis is a challenging task and more difficult than traditional topic-based classification
 - Understanding of the semantics is often needed
 - How could anyone sit through this movie?
 - Same word/phrase may have different polarities in different domains
 - An unpredictable movie (positive)
 - An unpredictable politician (negative)

Outline

- Sentiment analysis
 - Introduction
 - Definition, application, components
 - Approaches for SA subtasks
 - Holder detection
 - Target detection
 - Polarity classification
- Twitter Sentiment Analysis
 - Goals and challenges
 - Existing systems
 - Target-dependent twitter sentiment analysis

Outline

- Sentiment analysis
 - Introduction
 - Definition, application, components
 - Approaches for SA subtasks
 - Holder detection
 - Target detection
 - Polarity classification
- Twitter Sentiment Analysis
 - Goals and challenges
 - Existing systems
 - Target-dependent twitter sentiment analysis

Twitter Sentiment Analysis

- Aiming to find positive and negative tweets about a given topic
 - Focusing on polarity classification
- Target-dependent sentiment classification
 - Given a target, classifying a tweet as positive, negative or neutral (no sentiment) towards the target
 - Input: a tweet “*Windows 7 is much better than Vista!*” and a target “*Windows 7*”
 - Output: positive

Advantages of Twitter SA

- Large amount
- Wide coverage of domain
- Fresh
- From grass roots

Special Challenges

- Short and ambiguous
- Informal and unedited texts
 - “another part of me by Micheal Jackson is soo nicee! *Loooveeeee ittttttttt!*”

Outline

- Sentiment analysis
 - Introduction
 - Definition, application, components
 - Approaches for SA subtasks
 - Holder detection
 - Target detection
 - Polarity classification
- Twitter Sentiment Analysis
 - Goals and challenges
 - Existing systems
 - Target-dependent twitter sentiment analysis

Existing Twitter SA Systems

- Lexicon-based method
 - Twittratr
- Rule-based
 - Tweetfeel
- Machine learning based
 - Twitter sentiment
- Unknown
 - Twendz
 - Tweetsentiments

Twitrrratr

Microsoft | twitrratr - Windows Internet Explorer

http://twitrratr.com/search/Microsoft

Bing

Favorites Microsoft | twitrratr

Page Safety Tools ?

 twitrratr
TRACKING OPINIONS ON TWITTER

SEARCH

SEARCHED TERM	POSITIVE TWEETS	NEUTRAL TWEETS	NEGATIVE TWEETS	TOTAL TWEETS
Microsoft	619	9470	317	10406

5.95% POSITIVE

-  at a columbia engineering job fair. my state dot is here. the city of portland. nyc dot. real networks. microsoft. great for stuff ([view](#))
-  i love how unhelpful microsoft help articles are ([view](#))
-  new microsoft blog: the windows 7 blog for developers <http://is.gd/4isg> i think this blog could also interesting for admins. ([view](#))
-  microsoft gemini is what you always wanted excel to be. its an excellent step forward to bi for the masses. ([view](#))
-  microsoft gemini is what you always wanted excel to be. its an excellent step forward to bi for the masses. ([view](#))
-  @zik actually, i know people @ msnbc.com and already knew that... still, i don't think microsoft is happy with nbc making them look dumb. ([view](#))
-  @zik actually, i know people @ msnbc.com and already knew that... still, i don't think microsoft is happy with nbc making them look dumb. ([view](#))

91.01% NEUTRAL

-  Reading Harvard Business columnist Stew Friedman that says Microsoft could learn from Neil Young. Hmmm.. OK. ([view](#))

-  Just ordered Microsoft Office 2007 Professional for \$66 and change after shipping. Much better than the \$390 that Amazon wants. ([view](#))
-  @bryansimpson I have been hearing that a lot over the past month. Maybe them and Microsoft need to get together. ([view](#))
-  Microsoft Ad Business Strong. But Display Ads Threatened: Among Microsoft's diverse revenue streams, it's d.. <http://bit.ly/2owGq0> ([view](#))
-  Got myself a Microsoft Natural 4000 keyboard and wireless optical mouse. Need a USB hub methinks ([view](#))
-  [PCWorld] Microsoft Taps Telefonica to Deliver Live Messenger VOIP <http://tinyurl.com/5g57pj> ([view](#))
-  Found this: meta name="GENERATOR" content="Microsoft FrontPage 4.0". I think I just threw up in my mouth a little. ([view](#))

3.05% NEGATIVE

-  why microsoft develop ie ? it sucks grrrrrr ([view](#))
-  why microsoft develop ie ? it sucks grrrrrr ([view](#))
-  why microsoft develop ie ? it sucks grrrrrr ([view](#))
-  why microsoft develop ie ? it sucks grrrrrr ([view](#))
-  why microsoft develop ie ? it sucks grrrrrr ([view](#))
-  f*ck you internet explorer!! microsoft, you should be utterly ashamed of yourselves. ([view](#))
-  f*ck you internet explorer!! microsoft, you should be utterly ashamed of yourselves. ([view](#))
-  f*ck you internet explorer!! microsoft, you should be utterly ashamed of yourselves. ([view](#))
-  f*ck you internet explorer!! microsoft, you should be utterly ashamed of yourselves. ([view](#))
-  f*ck you internet explorer!! microsoft, you should be utterly ashamed of yourselves. ([view](#))

Internet | Protected Mode: On 100%

Twitrratr

- <http://twitrratr.com>
- Feature
 - 3 classes (positive, negative, neutral)
 - Highlight the sentiment expressions
- Method
 - Lexicon-based
 - Words, phrases, emoticons (☺, :D, :-(...))
 - Manually-made lexicon
 - Still contains errors (e.g., *fail* in the positive list)
 - Simple string (not word) match (“*unhelpful*”)

Tweetfeel

What do tweeples think about microsoft? Check out Tweetfeel twitter sentiment here - Windows Internet Explorer
http://www.tweetfeel.com/#microsoft

Favorites What do tweeples think about microsoft? Check o... □

Tweetfeel Biz | FAQ | Contact Us | Biz Login

 A New Level of Tweetfeel

|| microsoft Search

Try some Twitter trends: [Bondan Prakoso](#) [Scott Pilgrim](#) [Ramadan](#) [Inception](#) [Meteor Shower](#) [iOS](#) [Poulsen](#)

 60  89 = 60%

LOL RT @gmilh: RT @BillGayt's I love **microsoft** prodott. I li prend and me li ficc in my ass.

@leighpenny1 Is it OK that I love **microsoft** Excel and would marry it if I could figure out have to have sex with it?

I love **microsoft** Office 2010....Yeah i kno im a nerd at heart!

RearType? Expliquenme el prototipo de Table que desarrolla **microsoft** #fail TCTAL

RT @BillGayt: I love **microsoft** prodott. I li prend and me li ficc in my ass.

@dancerlindsey nope PS3...**microsoft** sucks

Read our FAQ Subscribe to our API Legal Stuff 100% Guarantee Share Email us Follow us Brought to you by **conversion** Powered by **twitter**

Done, but with errors on page. Internet | Protected Mode: On 100%

Tweetfeel

- <http://www.tweetfeel.com>
- Feature
 - 2 classes: positive and negative
- Method
 - Probably rule based
 - Positive patterns
 - pos_verb [Query], [Query] pos_verb, [Query] is pos_adj
 - Negative patterns
 - neg_verb [Query], [Query] neg_verb, [Query] is neg_adj
 - High precision, low recall

Twitter Sentiment

Twitter Sentiment - Windows Internet Explorer
http://twittersentiment.appspot.com/search?query=Microsoft

Favorites Twitter Sentiment

Twitter Sentiment

Type in a word and we'll highlight the good and the bad

microsoft Save this search

Sentiment analysis for microsoft

Sentiment by Percent

Negative (47%)
Positive (53%)

Sentiment by Count

Positive (18038) Negative (15789)

Zoom: 1d 5d 1m 3m 6m 1y Max

Positive 281 Negative 336 | August 11, 2010

Jul 11 Jul 18 Jul 25 Aug 1 Aug 8

1 K 500

Show Hourly Change Percent

Tweets about: microsoft

Choose a date range: to Update Note: We can only remember as far back as November 26, 2009

Reclassify the sentiment as: [Negative Positive]

abbykutiwa: @iamtomwah the creator of apple or **microsoft** didn't go to college, can't remember who...
Posted 20 minutes ago

SivadYar: @Catawampus25 the joys of working at **Microsoft** eh? :P
Posted 23 minutes ago

tomarbuthnot: @sevanjaniyan lol, that's your answer to all **Microsoft** related problems! :-)
Posted 28 minutes ago

mamotokyo: Internet Explorer 7 of **Microsoft** is SUCK.
Posted 39 minutes ago

BrianaThaBombb: @ClickClackAG I have a xbox 360 asshole! But my uncle works for **Microsoft** thas why lol
Posted 40 minutes ago

comparelaptops: Fury as users locked out of Hotmail: **Microsoft** says users should switch to ChromeUsers have hit back at **Microsoft** ... <http://bit.ly/b5d73C>

t by

Sentiment timeline

Detailed tweets

Twitter Sentiment

- <http://twittersentiment.appspot.com>
 - Created by some graduate students at Stanford University
- Features
 - 2 classes: positive and negative
 - Timeline: how the number of pos/neg sentiments change over time
 - Allows users to correct wrongly classified tweets
- Method
 - Machine learning-based (maximum entropy classifier)
 - Unsupervised training data construction by making use of emoticons (☺ for positive, ☹ for negative)

Summary

- Twitter SA has its own characteristics
 - Short, informal text
 - Pictograms (<3) and emoticons (☺, ☹, :D,...)
- However, not intensively studied yet
 - Traditional SA methods are employed
 - No paper published in top conferences yet
 - Lacking of large amount of publicly available annotated data for system evaluation and comparison
- Potential directions
 - Tweet normalization
 - Context aware sentiment analysis

References

- J. Blitzer, R. McDonald, and F. Pereira. Domain adaption with structural correspondence learning. In EMNLP, 2006.
- Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. 2005. Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. Proceedings of HLT/EMNLP-05.
- Alec Go, Richa Bhayani, Lei Huang. Twitter Sentiment Classification using Distant Supervision. <http://www.stanford.edu/~alecmgo/papers/TwitterDistantSupervision09.pdf>
- Vasileios Hatzivassiloglou and Kathleen McKeown. 1997. Predicting the semantic orientation of adjectives. In Proc. of the 35th ACL/8th EACL, pages 174–181.
- Hu, M., and Liu, B. 2004. Mining Opinion Features in Customer Reviews. To appear in AAAI'04, 2004.
- Hu, M., and Liu, B. 2004a. Mining and summarizing customer reviews. In Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining, pages 168–177. ACM Press New York, NY, USA.
- S Li, CR Huang, G Zhou, SYM Lee. 2010. Employing Personal/Impersonal Views in Supervised and Semi-supervised Sentiment Classification. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 414–423, Uppsala, Sweden, 11-16 July 2010.
- Q.Mei, X. Ling,M.Wondra, H. Su, and C.X. Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In Proceedings of the 16th International Conference onWorldWideWeb, pages 171–180.
- Bo Pang, Lillian Lee and Shivakumar Vaithyanathan: Thumbs up? Sentiment Classification using Machine Learning Techniques. In Proc. Conf. on EMNLP 2002.

Outline

- Sentiment analysis
 - Introduction
 - Definition, application, components
 - Approaches for SA subtasks
 - Holder detection
 - Target detection
 - Polarity classification
- Twitter Sentiment Analysis
 - Goals and challenges
 - Existing systems
 - Target-dependent twitter sentiment analysis

Target-dependent Twitter Sentiment Analysis

- Twitter: an important source for mining people's sentiments
 - Many existing systems for Twitter sentiment analysis
 - Tweetfeel, Twendz, and Twitter Sentiment
 - Typical scenario: the user inputs a sentiment target as a query, and searches for tweets containing positive or negative sentiments towards the target
- Target-dependent Sentiment Classification of Tweets
 - Given a query, classify the sentiments of the tweets as positive, negative or neutral according to whether they contain positive, negative or neutral sentiments about that query
 - Here the query serves as the target of the sentiments

Related Work

- Machine learning based (target-independent) twitter sentiment classification
 - Barbosa and Feng, 2010: Two-step approach to classify the sentiments of tweets using SVM classifiers
 - Davidiv et al., 2010 : Classify tweets into multiple sentiment types using hashtags and smileys as labels
 - Go et al., 2009: SVM classifier + collect training data using emoticons

Motivating Examples

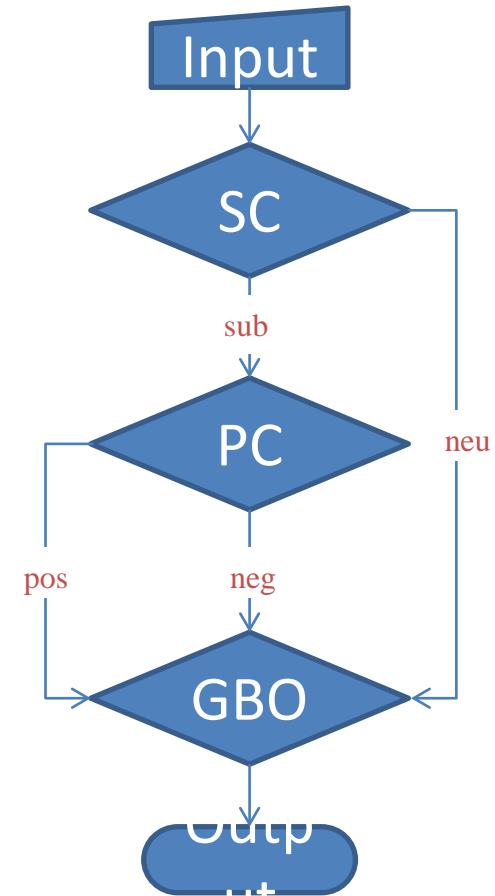
- Most current systems do not consider the target when classifying the sentiment
 - “*Bringing iPhone and iPad apps into cars? http://www.speakwithme.com/ will be out soon and alpha is awesome in my car.*”
 - (positive by Twitter Sentiment)
 - “*Here's a great article about Monte Veronese cheese. It's in Italian so just put the url into Google translate and enjoy http://ow.ly/3oQ77*”
 - (positive by Twitter Sentiment)
 - “*No debate needed, heat can't beat lakers or celtics*”
 - (negative by Twitter Sentiment)

Motivations

- Consider the relation between the target and the sentiment word
 - *Windows 7 is much **better** than Vista!*
 - “Windows 7” is connected with “better” by a copula while “Vista” is connected by a preposition “than”
 - *People everywhere **love** Windows & vista. Bill Gates*
 - “love” is not connected to “Bill Gates”
- Consider the context of the tweet
 - *First game: Lakers!*
 - *Too short even for human to decide the polarity*

Overview of Our Approach

- Task definition
 - Input
 - a collection of tweets containing the target (or query)
 - Output
 - labels assigned to each of the tweets
- Three steps
 - Subjectivity classification (SC)
 - Polarity classification (PC)
 - Graph-based optimization (GBO)



Preprocessing

- Tweet normalization
 - A simple rule-based model
 - “gooood” to “good”, “luve” to “love”
- POS tagging
 - OpenNLP POS tagger
- Word stemming
 - A word stem mapping table (about 20,000 entries)
- Syntactic parsing
 - A Maximum Spanning Tree dependency parser
(McDonald et al., 2005)

Subjectivity and Polarity Classification

- Binary SVM classifiers with linear kernel
 - Target-independent features
 - Content features
 - Words, punctuations, emoticons, and hashtags
 - Sentiment lexicon features
 - The number of positive or negative words in the tweet according to a sentiment lexicon (General Inquirer)
 - Target-dependent features

Target-dependent Features

- *Rules for generating target-dependent features*
 - *Subject/object* of a transitive verb w_i
 - wi_arg2 , e.g., “I love iPhone”, => “love_arg2”
 - wi_arg1 , e.g., “Obama reaffirms ..” => reaffirm_arg1
 - *Subject of* a intransitive verb
 - Wi_it_arg1
 - *Head of* an adjective or noun
 - Wi_arg1
 - Connected by a copula with an adjective or noun
 - Wi_cp_arg1
 - w_i is an adjective or intransitive verb appearing alone as a sentence and *the target* appears in the previous sentence
 - Wi_arg
 - E.g., “**John** did that. Great!” => great_arg
 - w_i is an adverb, and the verb it modifies has the target as its subject
 - $Arg1_v_wi$
 - E.g., “**iPhone** works better with the CellBand” => arg1_v_well
- Handle negations by adding “*neg-*”
 - “*iPhone does not work better with the CellBand*” => neg-work_it_arg1, arg1_v_neg-well

Target Expansion

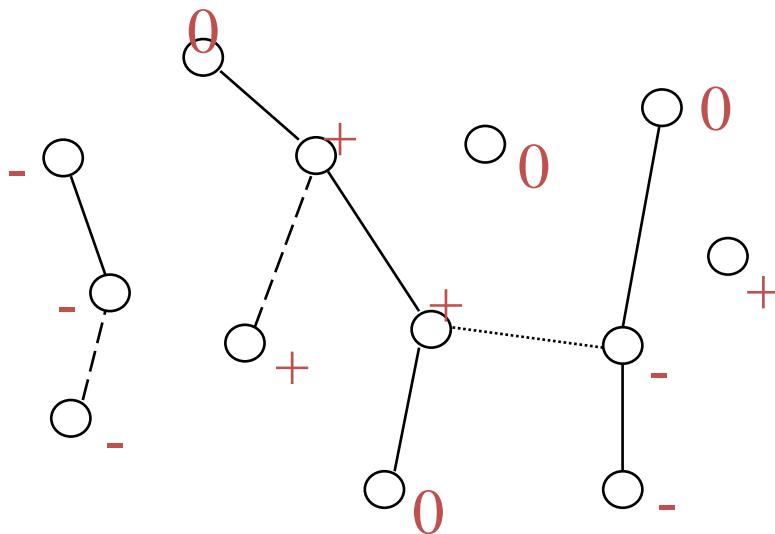
- Many sentiments are not expressed exactly towards the target
 - *I am passionate about Microsoft technologies especially Silverlight.*
 - *Microsoft vs. Microsoft technologies*
- Extended target identification
 - All noun phrases including the target
 - Mentions co-referring to the target
 - The top K nouns and noun phrases which have strong association with the target
 - Head nouns of all extended targets, which have strong association with the target

Examples for Target Expansion

- I am passionate about Microsoft technologies especially Silverlight.
- Oh, Jon Stewart. How I love you so ...
- Top K nouns and noun phrases with “Lady Gaga”: ladygaga, dressing, songs ...
- Microsoft Technologies

Graph-based Sentiment Optimization

- Relation types among the input tweets
 - Retweeting
 - Being published by the same person
 - Replying



Graph-based Sentiment Optimization

$$p(c | \tau) \longrightarrow$$

Relaxation Labeling
(Angelova and
Weikum, 2006)

$$\longrightarrow p(c | \tau, G)$$

The output score by SC and
PC to calculate $p(c | \tau)$

Iteratively optimize $p(c | \tau, G)$



$$p(c | \tau, G) = p(c | \tau) \sum_{N(d)} p(c | N(d)) p(N(d))$$

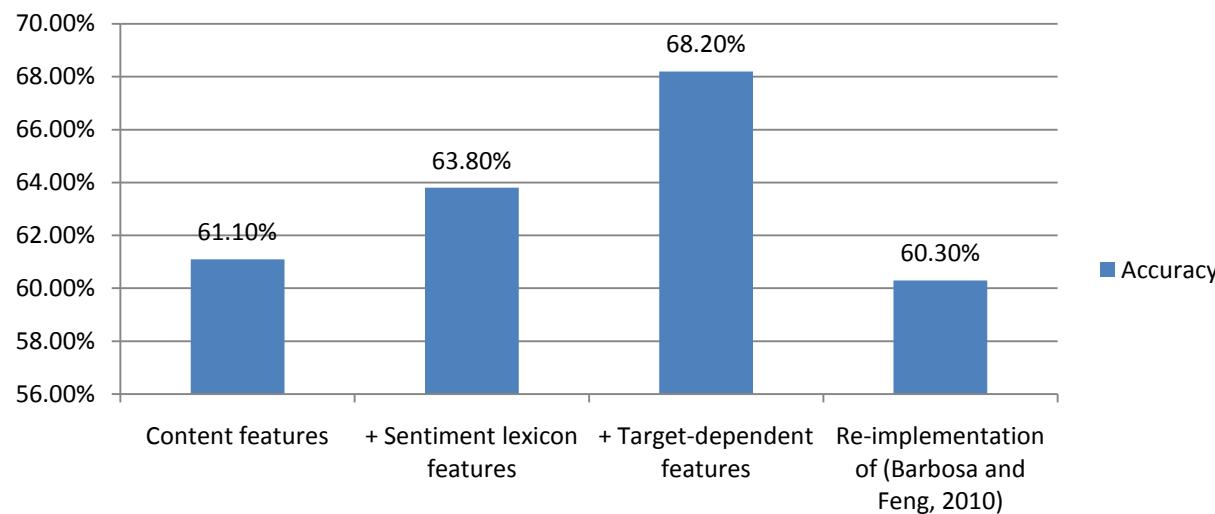
- c is the sentiment label of a tweet which belongs to {positive, negative, neutral}
- G is the tweet graph
- $N(d)$ is a specific assignment of sentiment labels to all immediate neighbors of the tweet
- τ is the content of the tweet

Experimental Setting

- Raw data
 - 5 queries: *Obama, Google, iPad, Lakers, Lady Gaga*
 - 400 English tweets downloaded for each
- Annotation
 - 2 human annotators
 - 3 labels: positive, negative or neutral
 - 459 positive, 268 negative and 1,212 neutral tweets
- Inter-annotator study
 - For 86% of tweets, two annotators give identical labels
 - For 13%, neutral-subjective disagreement
 - For 1%, positive-negative disagreement

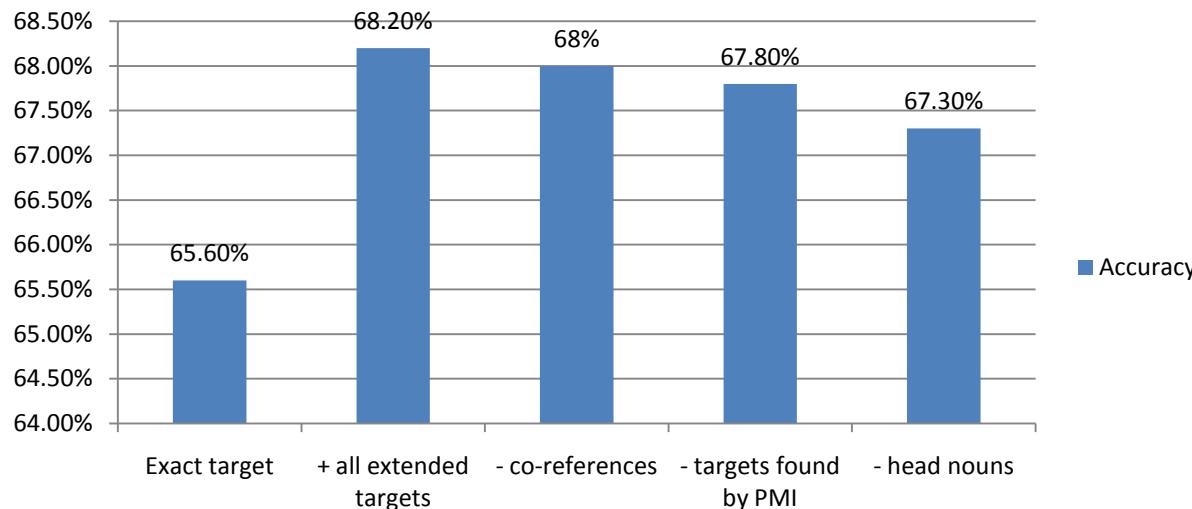
Subjectivity Classification Evaluation

- Data
 - 727 subjective (positive + negative) tweets and 1212 neutral tweets
 - 5 fold cross validation



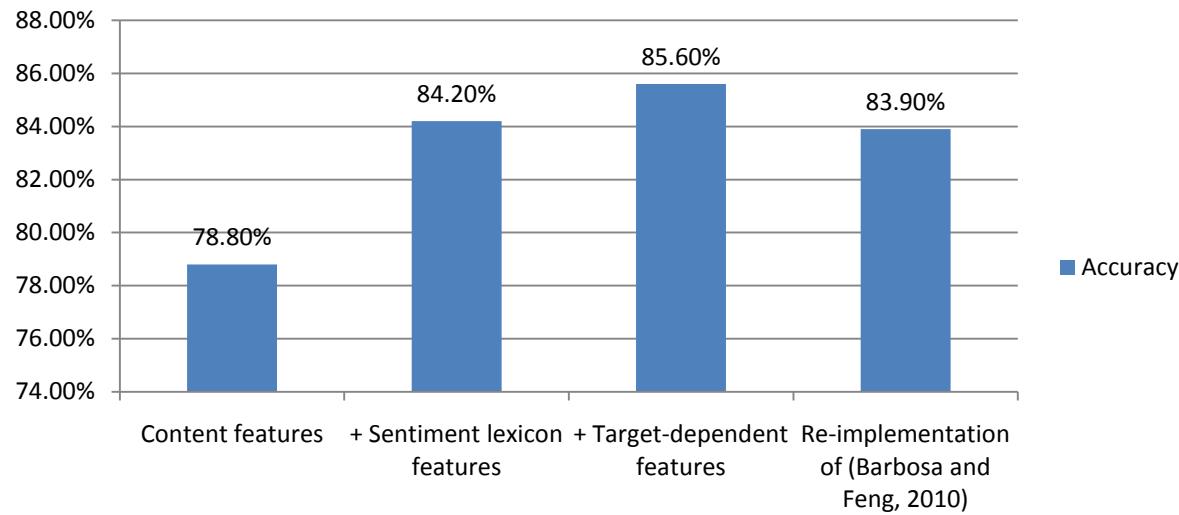
Evaluation of Extended Targets for Subjectivity Classification

- Data
 - 727 subjective (positive + negative) tweets and 1212 neutral tweets
 - 5 fold cross validation



Polarity Classification Evaluation

- Data
 - 268 negative and 459 positive tweets
 - 5 fold cross validation



Evaluation of Graph-based Optimization

- Data
 - 459 positive, 268 negative and 1,212 neutral tweets

System	Accuracy(%)	F1-score (%)		
		pos	neu	neg
Target-dependent sentiment classifier	66.0	57.5	70.1	66.1
+Graph-based optimization	68.3	63.5	71.0	68.5

Summary

- Target-dependent Twitter sentiment classification
 - Target-dependent features can improve the performance, especially for subjectivity classification
 - Incorporating related tweets can further improve the performance
- Future work
 - More types of extended targets
 - Exploring relations between Twitter accounts for classifying the sentiments of the tweets

Agenda

- QuickView: A Research Platform of SNS Text Mining and Search
- SNS Text Mining
 - Semantic Role Labeling
 - Sentiment Analysis
- Future Work

Research Task

- Fundamental NLP tasks
 - Tokenization, POS tagging, parsing, etc.
 - Text normalization
- Named entity extraction
 - Beyond PLO: movie, TV show, music, etc.
 - Normalization
- Hashtag understanding
 - Summarization
 - Relation recognition

Model and Approach

- Graph and graphical models
- Semi-supervised learning
- Online learning and data stream mining

Thanks

计算的未来

周明 博士
微软亚洲研究院主任研究员

中国计算机学会面向互联网的自然语言处理技术暑期学校， 2011

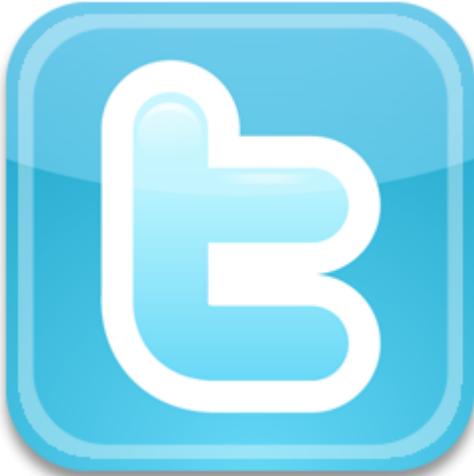
讲座内容

- 互联网的发展趋势和自然语言处理的新战略
- 微软公司、微软研究院和研究院的人才观
- 暑期学校微软段的课程简介

讲座内容

- 互联网的发展趋势和自然语言处理的新战略
- 微软公司、微软研究院和研究院的人才观
- 暑期学校微软段的课程简介

Understanding SNS



SNS Impacts

- Personalized search
- Personalized recommendation
- Enrich the search results with fresh, fine granularity, location-sensitive contents from SNS
- What is the best way to do SNS search and ads?
- The business model of SNS

Understanding Apps



中国移动通信
CHINA MOBILE



Apps Impact

- Change user behavior
 - From search documents to search apps
 - From search information to take action
- Build a new eco system
 - E.g. Baidu box computing, QQ+
 - Share users and markets to third parties
- Incubate new opportunities of app search and app ads

Mobile Internet

- 10+ times larger than desktop internet
- Location based service
- SNS empowered and local search oriented
- More search apps than docs
- From search engine to task engine
- Connection people for search, decision and task completion

The Ideal Search Engine

- Multi-modal query input and understanding
- Accurate answer rather than 10 blue links
- Super fresh results
- Aggregation of information from different sources, different media, different languages, if they are important
- Help make decision
- Personalized search results
- Automatic recommendation
- Help to complete tasks

NLP plays crucial roles for all of these tasks

Impact and Challenge to NLP Research

- Impact
 - Biggest database ever – connects data
 - Biggest social network – connects people
 - Contextual information processing: User, user's social network, location, time
 - Real-time information processing: Collection, index, operation without delay
- Challenge
 - How to leverage data, people, contextual information to reach real-time information processing?

Problems of Traditional NLP Approaches (NLP 1.0)

- Deep in individual component technologies but reach ceilings
- Less consider scenarios, user's need, market need
- Serious data sparseness with human annotation
- Evaluation bottleneck
- Slow deployment
- Lack effective framework to involve users' feedback

Traditional Research Approach Less Support to Internet Innovation

- Methodology oriented vs. solution oriented
- New but complicated method vs. simple but scalable method
- Local optimal vs. global optimal
- Call for new method for Internet Innovation

New Strategy of NLP (NLP2.0)

- Data collection from the web
- Contextual NLP (time, place, people)
- Maximize on the system level, not on the individual component
- Earlier deployment on Internet
- Make best use of social factors

Chinese Couplets (<http://duilian.msra.cn>)



<http://video.sina.com.cn/v/b/10937201-1452530713.html>

未来的技术是怎样的？

网络版本

本地版本

讲座内容

- 互联网的发展趋势和自然语言处理的新战略
- 微软公司、微软研究院和研究院的人才观
- 暑期学校微软段的课程简介

微软公司起源

- 成立于 1975
 - 致力于为个人计算机编写软件
 - 第一个产品: **BASIC 编译器**
 - **11 个人, 5万美金的营业额**



微软的新愿景

云 + 端

(Cloud + Client)

电脑



手机



电视



三屏融合

(自然用户界面)
NATURAL USER INTERFACE

创新，从未止步

PCs



手机



电视及娱乐



服务器



沟通
与生产力



创造力和
社交网络

搜索与在线



企业架构



2010财年，研发投入高达 95
亿美元

精彩，纷呈不断



Windows® 7

bing™



Microsoft® Online

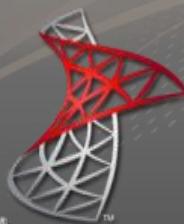


XBOX 360



Windows® Azure™

Microsoft®
Silverlight™



Microsoft®
SharePoint® Server

Microsoft®
SQL Server® 2008 R2



Microsoft®
Exchange



Windows®
Internet Explorer® 8



Windows Server® 2008 R2



Windows® phone



Microsoft®
Office 2010

XBOX体感游戏设备 Kinect

[网络版本](#)

[本地版本](#)

微软亚洲研究院概览

- 自1998年11月5日建院以来，十年创新路！
- 19个研究及工程小组
- 累计发表3000多篇高水平论文
- 260多项技术转化到微软产品中
- 过去4年20多项专利技术授权给国内外企业



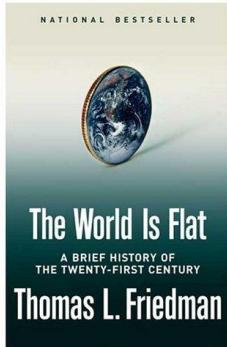
研究领域



创新的企业文化

- “每天清晨当你醒

都会为技术进步及其为人类生活
不已。” -Bill Gates



-
- 500多个专利
- 产品转化

水平
自微
自微
来自
自微



- The New York Times

中国，潜力无限

全球第一大手机市场 超过7亿手机用户(MIIT)



**全球最大互联网市场
超过3.8亿互联网用户(CNNIC)
2.3亿手机互联网用户**



**全球第二大电脑市场，
2011年预计成为全球最大电脑市场
(IDC)**



全球服务器第二大市场



微软亚洲研究院

创 新

- 尖端的技术
- 技术向产品的转化

人 才

- 招募一流人才
- 实习生项目

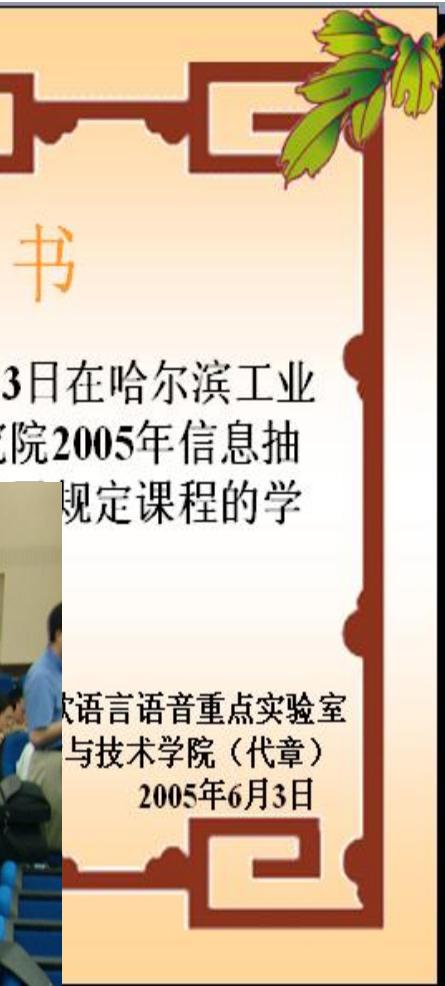
合 作

- 与高校建立联合实验室
- 与教育部合作人才培养计划

合格人才的 5C

- Competence 技术和技能
 - 坚实的专业基础
 - 学习能力
- Confidence 自信心
 - 积极主动，独立思考，相信自己
 - 对自己对世界有清晰的认识
- Creativity 创造力，自主创新
 - 初生牛犊不畏虎，大胆提出问题
 - 持之以恒的努力
- Communicating 沟通能力
 - 谋略，表达能力，倾听能力，说服技巧
 - 获取信息
- Cooperation 合作
 - 统一战线，海纳百川，共享成功
 - 合作能力

丰富多彩的暑期学校



Academic Collaboration in Asia

On an annual basis:

~ 20,000 attendees to events

~ 180 universities & institutes

~ 30-40 visiting researchers

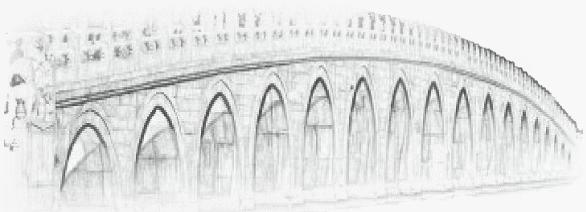
~ 100 funded projects

~ 200 interns at MS Research Asia

~ 8,000 attendees to campus activities



Microsoft Research Connections is a bridge to build long-term and mutually beneficial partnerships between Microsoft Research and academia in Asia Pacific.



人才培养



“明日之星”实习生项目（1998年）

培养高素质的计算机基础研究人才



微软学者奖学金

微软学者奖学金（1999年）

鼓励和支持亚太地区最优秀的计算机及相关专业的博士生



微软学生技术俱乐部（1999年）

为全国30所一流高校提供最新的技术资源和多元的交流平台

联合培养博士生项目

联合培养计划（2005年）

微软亚洲研究院研究员与高校教授一起联合培养博士生



微软小学者奖学金（2006年）

激励具有发展潜力优秀本科生的科技创新之路



微软铸星计划（2008年）

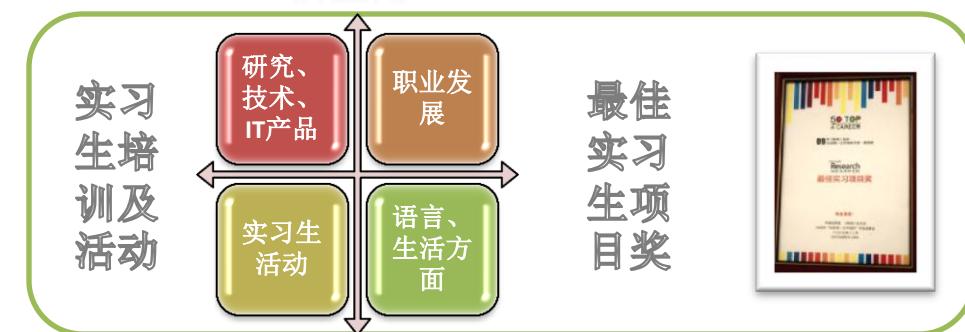
培养学术界未来的领导者

明日之星实习生项目

- 开始于1998年
- 全年性项目，每年向学生们提供400多个实习机会



陶李天
● 北京航空航天大学博士生，视觉计算组实习生
● 2009年8月的SIGGRAPH会议上作为第一作者成功发表论文 "SkyFinder: Attribute-Based Sky Image Search".



在这里你将得到



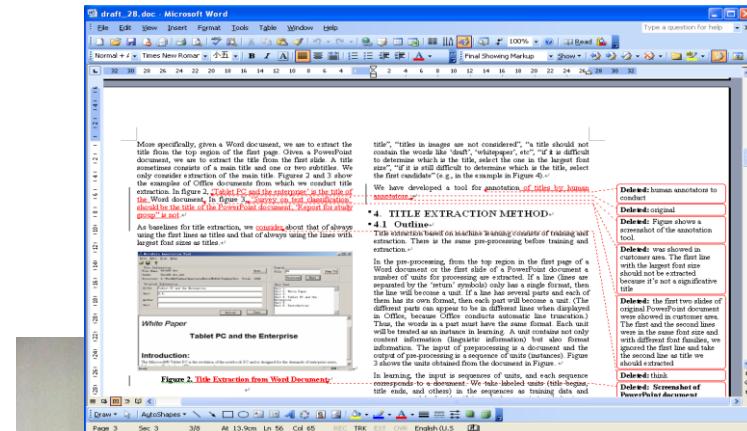
“一对一”的导师制

• Mentor

- Mentor一词来自希腊神话，意思是良师益友，贤明的顾问
- 1:1的指导学业、解决问题
- 传授做人做事的道理



如师如父，亦师亦友



帮实习生修改论文

亲密的合作伙伴

“有的放矢”的培训

- 前沿性学术讲座和沙龙：开拓学生视野
- 基础研究讲座系列：传授研究心得
 - How to do research
 - How to write the good paper
 - Research as a career
- 素质与技能培训：增强学生的综合素质，塑造职业品质
 - 专业技能(包括代码规范，设计模式，单元测试等)
 - ⑩ 软件项目项目实战(包括开发流程，开发团队，项目文档等)
 - ⑩ 职业素质(团队精神，沟通技巧，时间管理等)



研究院的经历能够帮助你

- ➔ English as the working language
- ➔ Competitive stipend and benefits
- ➔ Various of Trainings and Events



实习申请

Open Throughout the year

msraih@microsoft.com

With 400+ opportunities



Phone interview

CRITERIA

- Current Bachelor's, Master's, or Doctoral Candidate
- Major in CS, EE, math, physics, related fields
- Strong programming & research skills
- Strong written & verbal English



Resume



Acceptance

Application Form

On-Boarding

讲座内容

- 互联网的发展趋势和自然语言处理的新战略
- 微软公司、微软研究院和研究院的人才观
- 暑期学校微软段的课程简介

暑期学校微软研究院段(Day 1)

第三讲 互联网自然语言处理技术

组织者：MSRA主任研究员 周 明

- 第一课 09:00-10:00 计算的未来
- 第二课 10:10-11:30 机器翻译基础
- 11:30-13:00 午餐、休息
 - 地点：员工餐厅，微软1号楼3层
 - 携带餐券
- 13:00-14:00 参观微软体验中心(optional)
 - 分成三组，20个人左右一组
 - 每隔10分钟出发一组
- 第三课 14:00-14:50 微博语义分析和搜索(1)
- 第四课 15:00-15:50 微博语义分析和搜索(2)
- 第五课 16:00-17:00 深入英库和微软对联

暑期学校微软研究院段(Day 2)

第四讲 互联网语义挖掘与智能计算

组织者：MSRA主任研究员 林钦佑

- 第一课 09:00-10:00 大规模开放域语义挖掘技术概览
- 第二课 10:20-11:30 从物理世界建构智能 - 移动计算及普适计算的前沿
- 12:00-14:00 午餐、休息
 - 地点：员工餐厅，微软1号楼3层
 - 携带餐券
 - 自由沟通，多结识朋友
- 第三课 14:00-15:00 在线广告
- 第四课 15:10-16:10 微软学术搜索
- 第五课 16:20-17:00 总结暨问答互动

PAPER *Special Section on Information-Based Induction Sciences and Machine Learning*

A Short Introduction to Learning to Rank

Hang LI[†], Nonmember

SUMMARY Learning to rank refers to machine learning techniques for training the model in a ranking task. Learning to rank is useful for many applications in Information Retrieval, Natural Language Processing, and Data Mining. Intensive studies have been conducted on the problem and significant progress has been made [1], [2]. This short paper gives an introduction to learning to rank, and it specifically explains the fundamental problems, existing approaches, and future work of learning to rank. Several learning to rank methods using SVM techniques are described in details.

key words: *Learning to rank, information retrieval, natural language processing, SVM*

1. Ranking Problem

Learning to rank can be employed in a wide variety of applications in Information Retrieval (IR), Natural Language Processing (NLP), and Data Mining (DM). Typical applications are document retrieval, expert search, definition search, collaborative filtering, question answering, keyphrase extraction, document summarization, and machine translation [2]. Without loss of generality, we take document retrieval as example in this article.

Document retrieval is a task as follows (Fig. 1). The system maintains a collection of documents. Given a query, the system retrieves documents containing the query words from the collection, ranks the documents, and returns the top ranked documents. The ranking task is performed by using a ranking model $f(q, d)$ to sort the documents, where q denotes a query and d denotes a document.

Traditionally, the ranking model $f(q, d)$ is created without training. In the BM25 model, for example, it is assumed that $f(q, d)$ is represented by a conditional probability distribution $P(r|q, d)$ where r takes on 1 or 0 as value and denotes being relevant or irreverent, and q and d denote a query and a document respectively. In Language Model for IR (LMIR), $f(q, d)$ is represented as a conditional probability distribution $P(q|d)$. The probability models can be calculated with the words appearing in the query and document, and thus no training is needed (only tuning of a small number of parameters is necessary) [3].

Manuscript received December 31, 2010.

Manuscript revised June 1, 2011.

[†]The author is with Microsoft Research Asia
DOI: 10.1587/transinf.E94.D.1

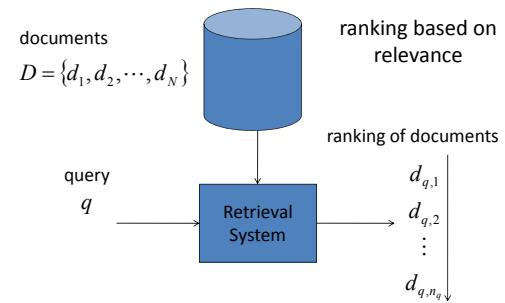


Fig. 1 Document Retrieval

A new trend has recently arisen in document retrieval, particularly in web search, that is, to employ machine learning techniques to automatically construct the ranking model $f(q, d)$. This is motivated by a number of facts. At web search, there are many signals which can represent relevance, for example, the anchor texts and PageRank score of a web page. Incorporating such information into the ranking model and automatically constructing the ranking model using machine learning techniques becomes a natural choice. In web search engines, a large amount of search log data, such as click through data, is accumulated. This makes it possible to derive training data from search log data and automatically create the ranking model. In fact, learning to rank has become one of the key technologies for modern web search.

We describe a number of issues in learning for ranking, including training and testing, data labeling, feature construction, evaluation, and relations with ordinal classification.

1.1 Training and Testing

Learning to rank is a supervised learning task and thus has training and testing phases (see Fig. 2).

The training data consists of queries and documents. Each query is associated with a number of documents. The relevance of the documents with respect to the query is also given. The relevance information can be represented in several ways. Here, we take the most widely used approach and assume that the relevance of a document with respect to a query is represented by

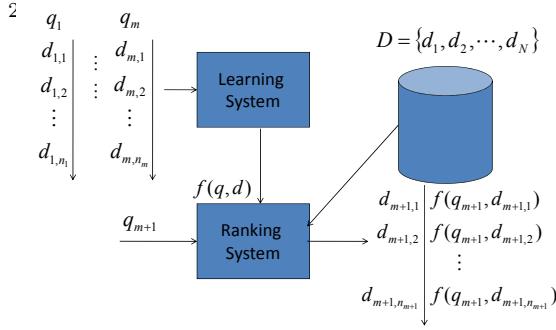


Fig. 2 Learning to Rank for Document Retrieval

a label, while the labels denote several grades (levels). The higher grade a document has, the more relevant the document is.

Suppose that \mathcal{Q} is the query set and \mathcal{D} is the document set. Suppose that $\mathcal{Y} = \{1, 2, \dots, l\}$ is the label set, where labels represent grades. There exists a total order between the grades $l \succ l-1 \succ \dots \succ 1$, where \succ denotes the order relation. Further suppose that $\{q_1, q_2, \dots, q_m\}$ is the set of queries for training and q_i is the i -th query. $D_i = \{d_{i,1}, d_{i,2}, \dots, d_{i,n_i}\}$ is the set of documents associated with query q_i and $\mathbf{y}_i = \{y_{i,1}, y_{i,2}, \dots, y_{i,n_i}\}$ is the set of labels associated with query q_i , where n_i denotes the sizes of D_i and \mathbf{y}_i ; $d_{i,j}$ denotes the j -th document in D_i ; and $y_{i,j} \in \mathcal{Y}$ denotes the j -th grade label in \mathbf{y}_i , representing the relevance degree of $d_{i,j}$ with respect to q_i . The original training set is denoted as $S = \{(q_i, D_i), \mathbf{y}_i\}_{i=1}^m$.

A feature vector $x_{i,j} = \phi(q_i, d_{i,j})$ is created from each query-document pair $(q_i, d_{i,j})$, $i = 1, 2, \dots, m$; $j = 1, 2, \dots, n_i$, where ϕ denotes the feature functions. That is to say, features are defined as functions of a query document pair. For example, BM25 and PageRank are typical features [2]. Letting $\mathbf{x}_i = \{x_{i,1}, x_{i,2}, \dots, x_{i,n_i}\}$, we represent the training data set as $S' = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^m$. Here $x \in \mathcal{X}$ and $\mathcal{X} \subseteq \mathbb{R}^d$.

We aim to train a (local) ranking model $f(q, d) = f(x)$ that can assign a score to a given query document pair q and d , or equivalently to a given feature vector x . More generally, we can also consider training a global ranking model $F(q, D) = F(\mathbf{x})$. The local ranking model outputs a single score, while the global ranking model outputs a list of scores.

Let the documents in D_i be identified by the integers $\{1, 2, \dots, n_i\}$. We define a permutation (ranking list) π_i on D_i as a bijection from $\{1, 2, \dots, n_i\}$ to itself. We use Π_i to denote the set of all possible permutations on D_i , use $\pi_i(j)$ to denote the rank (or position) of the j -th document (i.e., $d_{i,j}$) in permutation π_i . Ranking is nothing but to select a permutation $\pi_i \in \Pi_i$ for the given query q_i and the associated documents D_i using the scores given by the ranking model $f(q_i, d_i)$.

The test data consists of a new query q_{m+1} and associated documents D_{m+1} . $T = \{(q_{m+1}, D_{m+1})\}$.

We create feature vector \mathbf{x}_{m+1} , use the trained ranking model to assign scores to the documents D_{m+1} , sort them based on the scores, and give the ranking list of documents as output π_{m+1} .

The training and testing data is similar to, but different from the data in conventional supervised learning such as classification and regression. Query and its associated documents form a group. The groups are i.i.d. data, while the instances within a group are not i.i.d. data. A local ranking model is a function of a query and a document, or equivalently, a function of a feature vector derived from a query and a document.

1.2 Data Labeling

Currently there are two ways to create training data. The first one is by human judgments and the second one is by derivation from search log data. We explain the first approach here. Explanations on the second approach can be found in [2]. In the first approach, a set of queries is randomly selected from the query log of a search system. Suppose that there are multiple search systems. Then the queries are submitted to the search systems and all the top ranked documents are collected. As a result, each query is associated with multiple documents. Human judges are then asked to make relevance judgments on all the query document pairs. Relevance judgments are usually conducted at five levels, for example, perfect, excellent, good, fair, and bad. Human judges make relevance judgments from the viewpoint of average users. For example, if the query is ‘Microsoft’, and the web page is microsoft.com, then the label is ‘perfect’. Furthermore, the Wikipedia page about Microsoft is ‘excellent’, and so on. Labels representing relevance are then assigned to the query document pairs. Relevance judgment on a query document pair can be performed by multiple judges and then majority voting can be conducted. Benchmark data sets on learning to rank have also been released [4].

1.3 Evaluation

The evaluation on the performance of a ranking model is carried out by comparison between the ranking lists output by the model and the ranking lists given as the ground truth. Several evaluation measures are widely used in IR and other fields. These include NDCG (Normalized Discounted Cumulative Gain), DCG (Discounted Cumulative Gain), MAP (Mean Average Precision), and Kendall’s Tau.

Given query q_i and associated documents D_i , suppose that π_i is the ranking list (permutation) on D_i and \mathbf{y}_i is the set of labels (grades) of D_i . DCG [5] measures the goodness of the ranking list with the labels. Specifically, DCG at position k is defined as:

$$DCG(k) = \sum_{j:\pi_i(j) \leq k} G(j)D(\pi_i(j)),$$

where $G_i(\cdot)$ is a gain function and $D_i(\cdot)$ is a position discount function, and $\pi_i(j)$ is the position of $d_{i,j}$ in π_i . The summation is taken over the top k positions in the ranking list π_i . DCG represents the cumulative gain of accessing the information from position one to position k with discounts on the positions. NDCG is normalized DCG and NDCG at position k is defined as:

$$NDCG(k) = G_{max,i}^{-1}(k) \sum_{j:\pi_i(j) \leq k} G(j)D(\pi_i(j)),$$

where $G_{max,i}(k)$ is the normalizing factor and is chosen such that a perfect ranking π_i^* 's NDCG score at position k is 1. In a perfect ranking, the documents with higher grades are always ranked higher. Note that there can be multiple perfect rankings for a query and associated documents.

The gain function is normally defined as an exponential function of grade. That is to say, the satisfaction of accessing information exponentially increases when the grade of relevance of information increases.

$$G(j) = 2^{y_{i,j}} - 1, \quad (1)$$

where $y_{i,j}$ is the label (grade) of document $d_{i,j}$ in ranking list π_i . The discount function is normally defined as a logarithmic function of position. That is to say, the satisfaction of accessing information logarithmically decreases when the position of access increases.

$$D(\pi_i(j)) = \frac{1}{\log_2(1 + \pi_i(j))}, \quad (2)$$

where $\pi_i(j)$ is the position of document $d_{i,j}$ in ranking list π_i .

Hence, DCG and NDCG at position k become

$$DCG(k) = \sum_{j:\pi_i(j) \leq k} \frac{2^{y_{i,j}} - 1}{\log_2(1 + \pi_i(j))}, \quad (3)$$

$$NDCG(k) = G_{max,i}^{-1}(k) \sum_{j:\pi_i(j) \leq k} \frac{2^{y_{i,j}} - 1}{\log_2(1 + \pi_i(j))}. \quad (4)$$

In evaluation, DCG and NDCG values are further averaged over queries.

Table 1 gives examples of calculating NDCG values of two ranking lists. NDCG (DCG) has the effect of giving high scores to the ranking lists in which relevant documents are ranked high. For perfect rankings, the NDCG value at each position is always one, while for imperfect rankings, the NDCG values are usually less than one.

MAP is another measure widely used in IR. In MAP, it is assumed that the grades of relevance are at two levels: 1 and 0. Given query q_i , associated documents D_i , ranking list π_i on D_i , and labels \mathbf{y}_i of D_i , Average Precision for q_i is defined as:

$$AP = \frac{\sum_{j=1}^{n_i} P(j) \cdot y_{i,j}}{\sum_{j=1}^{n_i} y_{i,j}},$$

Table 1 Examples of NDCG Calculation.

Perfect ranking	Formula	Explanation
(3, 3, 2, 2, 1, 1, 1)		grades:3,2,1
(7, 7, 3, 3, 1, 1, 1)	Eq.(1)	gains
(1, 0.63, 0.5, ...)	Eq.(2)	discounts
(7, 11.41, 12.91, ...)	Eq.(3)	DCG
(1/7, 1/11.41, 1/12.91, ...)		normalizers
(1,1,1,...)	Eq.(4)	NDCG
Imperfect ranking	Formula	Explanation
(2, 3, 2, 3, 1, 1, 1)		grades:3,2,1
(3, 7, 3, 7, 1, 1, 1)	Eq.(1)	gains
(1, 0.63, 0.5, ...)	Eq.(2)	discounts
(3, 7.41, 8.91, ...)	Eq.(3)	DCG
(1/7, 1/11.41, 1/12.91, ...)		normalizers
(0.43, 0.65, 0.69, ...)	Eq.(4)	NDCG

where $y_{i,j}$ is the label (grade) of $d_{i,j}$ and takes on 1 or 0 as a value, representing being relevant or irrelevant. $P(j)$ for query q_i is defined as:

$$P(j) = \frac{\sum_{k:\pi_i(k) \leq \pi_i(j)} y_{i,k}}{\pi_i(j)},$$

where $\pi_i(j)$ is the position of $d_{i,j}$ in π_i . $P(j)$ represents the precision until the position of $d_{i,j}$ for q_i . Note that labels are either 1 or 0, and thus ‘precision’ can be defined. Average Precision represents averaged precision over all the positions of documents with label 1 for query q_i .

Average Precision values are further averaged over queries to become Mean Average Precision (MAP).

1.4 Relation with Ordinal Classification

Ordinal classification (also known as ordinal regression) is similar to ranking, but is also different. The input of ordinal classification is a feature vector x and the output is a label y representing a grade, where the grades are classes in a total order. The goal of learning is to construct a model which can assign a grade label y to a given feature vector x . The model mainly consists of a scoring function $f(x)$. The model first assigns a real number to x using $f(x)$ and then determines the grade y of x using a number of thresholds. Specifically, it partitions the real number axis into intervals and aligns each interval to a grade. It takes the grade of the interval that $f(x)$ falls into as the grade of x .

In ranking, one cares more about accurate ordering of objects, while in ordinal classification, one cares more about accurate ordered-categorization of objects. A typical example of ordinal classification is product rating. For example, given the features of a movie, we are to assign a number of stars (ratings) to the movie. In that case, correct assignment of the number of stars is critical. In contrast, in ranking such as document retrieval, given a query, the objective is to correctly sort related documents, although sometimes training data and testing data are labeled at multiple grades as in ordinal classification. The number of documents to be

ranked can vary from query to query. There are queries for which more relevant documents are available in the collection, and there are also queries for which only weakly relevant documents are available.

2. Formulation

We formalize learning to rank as a supervised learning task. Suppose that \mathcal{X} is the input space (feature space) consisting of lists of feature vectors, and \mathcal{Y} is the output space consisting of lists of grades. Further suppose that \mathbf{x} is an element of \mathcal{X} representing a list of feature vectors and \mathbf{y} is an element of \mathcal{Y} representing a list of grades. Let $P(X, Y)$ be an unknown joint probability distribution where random variable X takes \mathbf{x} as its value and random variable Y takes \mathbf{y} as its value.

Assume that $F(\cdot)$ is a function mapping from a list of feature vectors \mathbf{x} to a list of scores. The goal of the learning task is to automatically learn a function $\hat{F}(\mathbf{x})$ given training data $(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{x}_m, \mathbf{y}_m)$. Each training instance is comprised of feature vectors \mathbf{x}_i and the corresponding grades \mathbf{y}_i ($i = 1, \dots, m$). Here m denotes the number of training instances.

$F(\mathbf{x})$ and \mathbf{y} can be further written as $F(\mathbf{x}) = (f(x_1), f(x_2), \dots, f(x_n))$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)$. The feature vectors represent objects to be ranked. Here $f(x)$ denotes the local ranking function and n denotes the number of feature vectors and grades.

A loss function $L(\cdot, \cdot)$ is utilized to evaluate the prediction result of $F(\cdot)$. First, feature vectors \mathbf{x} are ranked according to $F(\mathbf{x})$, then the top n results of the ranking are evaluated using their corresponding grades \mathbf{y} . If the feature vectors with higher grades are ranked higher, then the loss will be small. Otherwise, the loss will be large. The loss function is specifically represented as $L(F(\mathbf{x}), \mathbf{y})$. Note that the loss function for ranking is slightly different from the loss functions in other statistical learning tasks, in the sense that it makes use of sorting.

We further define the risk function $R(\cdot)$ as the expected loss function with respect to the joint distribution $P(X, Y)$,

$$R(F) = \int_{\mathcal{X} \times \mathcal{Y}} L(F(\mathbf{x}), \mathbf{y}) dP(\mathbf{x}, \mathbf{y}).$$

Given training data, we calculate the empirical risk function as follows,

$$\hat{R}(F) = \frac{1}{m} \sum_{i=1}^m L(F(\mathbf{x}_i), \mathbf{y}_i).$$

The learning task then becomes the minimization of the empirical risk function, as in other learning tasks. The minimization of the empirical risk function could be difficult due to the nature of the loss function (it is not continuous and it uses sorting). We can consider using a surrogate loss function $L'(F(\mathbf{x}), \mathbf{y})$.

The corresponding empirical risk function is defined as follows.

$$\hat{R}'(F) = \frac{1}{m} \sum_{i=1}^m L'(F(\mathbf{x}_i), \mathbf{y}_i).$$

We can also introduce a regularizer to conduct minimization of the regularized empirical risk. In such cases, the learning problem becomes minimization of the (regularized) empirical risk function based on the surrogate loss.

Note that we adopt a machine learning formulation here. In IR, the feature vectors \mathbf{x} are derived from a query and its associated documents. The grades \mathbf{y} represent the relevance degrees of the documents with respect to the query. We make use of a global ranking function $F(\cdot)$. In practice, it can be a local ranking function $f(\cdot)$. The possible number of feature vectors in \mathbf{x} can be very large, even infinite. The evaluation (loss function) is, however, only concerned with n results.

In IR, the true loss functions can be those defined based on NDCG (Normalized Discounted Cumulative Gain) and MAP (Mean Average Precision). For example, we can have

$$L(F(\mathbf{x}), \mathbf{y}) = 1.0 - NDCG.$$

Note that the true loss functions (NDCG loss and MAP loss) makes use of sorting based on $F(\mathbf{x})$.

For the surrogate loss function, there are also different ways to define it, which lead to different approaches to learning to rank. For example, one can define pointwise loss, pairwise loss, and listwise loss functions.

The squared loss used in Subset Regression is a pointwise surrogate loss [6]. We call it pointwise loss, because it is defined on *single objects*.

$$L'(F(\mathbf{x}), \mathbf{y}) = \sum_{i=1}^n (f(x_i) - y_i)^2.$$

It is actually an upper bound of $1.0 - NDCG$.

Pairwise losses can be the hinge loss, exponential loss, and logistic loss on *pairs of objects*, which are used in Ranking SVM [7], RankBoost [8], and RankNet [9], respectively. They are also upper bounds of $1.0 - NDCG$ [10].

$$L'(F(\mathbf{x}), \mathbf{y}) = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \phi(\text{sign}(y_i - y_j), f(x_i) - f(x_j)),$$

where it is assumed that $L' = 0$ when $y_i = y_j$ and ϕ is the hinge loss, exponential loss, or logistic loss function.

Listwise loss functions are defined on *lists of objects*, just like the true loss functions, and thus are more directly related to the true loss functions. Different listwise loss functions are exploited in the listwise methods. For example, the loss function in AdaRank is a listwise

loss.

$$L'(F(\mathbf{x}), \mathbf{y}) = \exp(-NDCG),$$

where $NDCG$ is calculated on the basis of $F(\mathbf{x})$ and \mathbf{y} . Obviously, it is also an upper bound of $1.0 - NDCG$.

3. Pointwise Approach

In the pointwise approach, the ranking problem (ranking creation) is transformed to classification, regression, or ordinal classification, and existing methods for classification, regression, or ordinal classification are applied. Therefore, the group structure of ranking is ignored in this approach.

The pointwise approach includes Subset Ranking [6], McRank [11], Prank [12], and OC SVM [13]. We take the last one as an example and describe it in detail.

3.1 SVM for Ordinal Classification

The method proposed by Shashua & Levin [13] utilizes a number of parallel hyperplanes as a ranking model. Their method, referred to as OC SVM in this article, learns the parallel hyperplanes by the large margin principle. In one implementation, the method tries to maximize a fixed margin for all the adjacent classes (grades).[†]

Suppose that $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = \{1, 2, \dots, l\}$ where there exists a total order on \mathcal{Y} . $x \in \mathcal{X}$ is an object (feature vector) and $y \in \mathcal{Y}$ is a label representing a grade. Given object x , we aim to predict its label (grade) y . That is to say, this is an ordinal classification problem. We employ a number of linear models (parallel hyperplanes) $\langle w, x \rangle - b_r, (r = 1, \dots, l - 1)$ to make the prediction, where $w \in \mathbb{R}^d$ is a weight vector and $b_r \in \mathbb{R}, (r = 1, \dots, l)$ are biases satisfying $b_1 \leq \dots \leq b_{l-1} \leq b_l = +\infty$. The models correspond to parallel hyperplanes $\langle w, x \rangle - b_r = 0$ separating grades r and $r + 1$, $(r = 1, \dots, l - 1)$. Figure 3 illustrates the model. If x satisfies $\langle w, x \rangle - b_{r-1} \geq 0$ and $\langle w, x \rangle - b_r < 0$, then $y = r$, $(r = 1, \dots, l)$. We can write it as $\min_{r \in \{1, \dots, l\}} \{r | \langle w, x \rangle - b_r < 0\}$.

Suppose that the training data is given as follows. For each grade $r = 1, \dots, l$, there are m_r instances: $x_{r,i}, i = 1, \dots, m_r$. The learning task is formalized as the following Quadratic Programming (QP) problem.

$$\begin{aligned} \min_{w, b, \xi} & \frac{1}{2} \|w\|^2 + C \sum_{r=1}^{l-1} \sum_{i=1}^{m_r} (\xi_{r,i} + \xi_{r+1,i}^*) \\ \text{s. t. } & \langle w, x_{r,i} \rangle + b_r \geq 1 - \xi_{r,i} \\ & \langle w, x_{r+1,i} \rangle + b_r \leq 1 - \xi_{r+1,i}^* \\ & \xi_{r,i} \geq 0, \quad \xi_{r+1,i}^* \geq 0 \\ & i = 1, \dots, m_r, \quad r = 1, \dots, l - 1 \\ & m = m_1 + \dots + m_l, \end{aligned}$$

where $x_{r,i}$ denotes the i -th instance in the r -th grade,

[†]The other method maximizes the sum of all margins.

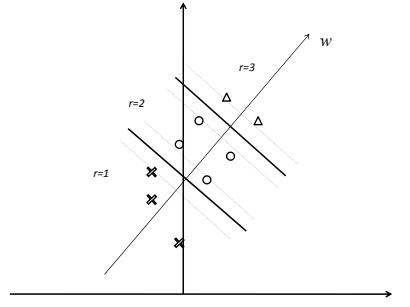


Fig. 3 SVM for Ordinal Classification

$\xi_{r+1,i}$ and $\xi_{r+1,i}^*$ denote the corresponding slack variables, $\|\cdot\|$ denotes L_2 norm, m denotes the number of training instances, and $C > 0$ is a coefficient. The method tries to separate the instances in the neighboring grades with the same margin.

4. Pairwise Approach

In the pairwise approach, ranking is transformed into pairwise classification or pairwise regression. In the former case, a classifier for classifying the ranking orders of document pairs is created and is employed in the ranking of documents. In the pairwise approach, the group structure of ranking is also ignored.

The pairwise approach includes Ranking SVM [7], RankBoost [8], RankNet [9], GBRank [14], IR SVM [15], Lambda Rank [16], and LambdaMART [17]. We introduce Ranking SVM and IR SVM in this article.

4.1 Ranking SVM

We can learn a classifier, such as SVM, for *classifying the order of pairs of objects* and utilize the classifier in the ranking task. This is the idea behind the Ranking SVM method proposed by Herbrich et al. [7].

Figure 4 shows an example of the ranking problem. Suppose that there are two groups of objects (documents associated with two queries) in the feature space. Further suppose that there are three grades (levels). For example, objects x_1, x_2 , and x_3 in the first group are at three different grades. The weight vector w corresponds to the linear function $f(x) = \langle w, x \rangle$ which can score and rank the objects. Ranking objects with the function is equivalent to projecting the objects into the vector and sorting the objects according to the projections on the vector. If the ranking function is ‘good’, then there should be an effect that objects at grade 3 are ranked ahead of objects at grade 2, etc. Note that objects belonging to different groups are incomparable.

Figure 5 shows that the ranking problem in Figure 4 can be transformed to Linear SVM classification. The differences between two feature vectors at different

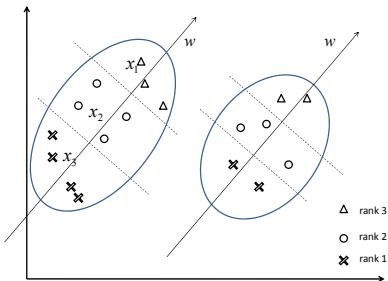


Fig. 4 Example of Ranking Problem

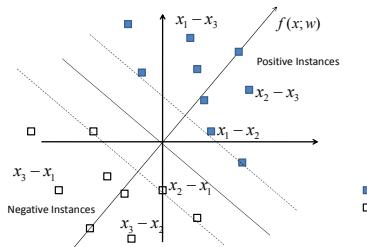


Fig. 5 Transformation to Pairwise Classification

grades in the same group are treated as new feature vectors, e.g., $x_1 - x_2$, $x_1 - x_3$, and $x_2 - x_3$. Furthermore, labels are also assigned to the new feature vectors. For example, $x_1 - x_2$, $x_1 - x_3$, and $x_2 - x_3$ are positive. Note that feature vectors at the same grade or feature vectors from different groups are not utilized to create new feature vectors. One can train a Linear SVM classifier which separates the new feature vectors as shown in Figure 5. Geometrically, the margin in the SVM model represents the closest distance between the projections of object pairs in two grades. Note that the hyperplane of the SVM classifier passes the original and the positive and negative instances form corresponding pairs. For example, $x_1 - x_2$ and $x_2 - x_1$ are positive and negative instances respectively. The weight vector w of the SVM classifier corresponds to the ranking function. In fact, we can discard the negative instances in learning, because they are redundant.

Training data is given as $\{((x_i^{(1)}, x_i^{(2)}), y_i)\}, i = 1, \dots, m$ where each instance consists of two feature vectors $(x_i^{(1)}, x_i^{(2)})$ and a label $y_i \in \{+1, -1\}$ denoting which feature vector should be ranked ahead.

The learning of Ranking SVM is formalized as the following QP problem.

Grade: 3, 2, 1

Documents are represented by their grades

Example 1:

ranking-1: 2 3 2 1 1 1 1

ranking-2: 3 2 1 2 1 1 1

Example 2:

ranking for query-1: 3 2 2 1 1 1 1

ranking for query-2: 3 3 2 2 2 1 1 1 1 1

Fig. 6 Example Ranking Lists

$$\begin{aligned} & \min_{w, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i \\ \text{s. t. } & y_i \langle w, x_i^{(1)} - x_i^{(2)} \rangle \geq 1 - \xi_i \\ & \xi_i \geq 0 \\ & i = 1, \dots, m, \end{aligned}$$

where $x_i^{(1)}$ and $x_i^{(2)}$ denote the first and second feature vectors in a pair of feature vectors, $\|\cdot\|$ denotes L_2 norm, m denotes the number of training instances, and $C > 0$ is a coefficient.

It is equivalent to the following non-constrained optimization problem, i.e., the minimization of the regularized hinge loss function.

$$\min_w \sum_{i=1}^m [1 - y_i \langle w, x_i^{(1)} - x_i^{(2)} \rangle]_+ + \lambda \|w\|^2, \quad (5)$$

where $[x]_+$ denotes function $\max(x, 0)$ and $\lambda = \frac{1}{2C}$.

4.2 IR SVM

IR SVM proposed by Cao et al. [15] is an extension of Ranking SVM for Information Retrieval (IR), whose idea can be applied to other applications as well.

Ranking SVM transforms ranking into pairwise classification, and thus it actually makes use of the 0-1 loss in the learning process. There exists a gap between the loss function and the IR evaluation measures. IR SVM attempts to bridge the gap by modifying 0-1 loss, that is, conducting cost sensitive learning of Ranking SVM.

We first look at the problems caused by straightforward application of Ranking SVM to document retrieval, using examples in Figure 6.

One problem with the direct application of Ranking SVM is that Ranking SVM equally treats document pairs across different grades. Example 1 indicates the problem. There are two rankings for the same query. The documents at positions 1 and 2 are swapped in ranking-1 from the perfect ranking, while the documents at positions 3 and 4 are swapped in ranking-2 from the perfect ranking. There is only one error for each ranking in terms of the 0-1 loss, or difference in order of pairs. They have the same effect on the training of Ranking SVM, which is not desirable. Ranking-2 should be better than ranking-1, from the viewpoint of IR, because the result on its top is better. Note that to have high accuracy on top-ranked documents is crucial for an IR system, which is reflected in the IR evaluation

measures.

Another issue with Ranking SVM is that it equally treats document pairs from different queries. In example 2, there are two queries and the numbers of documents associated with them are different. For query-1 there are 2 document pairs between grades 3-2, 4 document pairs between grades 3-1, 8 document pairs between grades 2-1, and in total 14 document pairs. For query-2, there are 31 document pairs. Ranking SVM takes 14 instances (document pairs) from query-1 and 31 instances (document pairs) from query-2 for training. Thus, the impact on the ranking model from query-2 will be larger than the impact from query-1. In other words, the model learned will be biased toward query-2. This is in contrast to the fact that in IR evaluation queries are evenly important. Note that the numbers of documents usually vary from query to query.

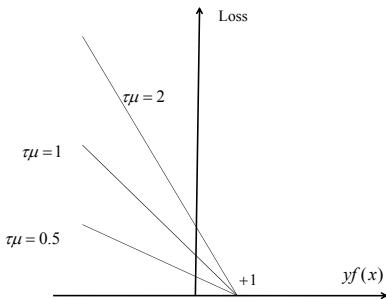


Fig. 7 Modified Hinge Loss Functions

IR SVM addresses the above two problems by changing the 0-1 pairwise classification into a cost sensitive pairwise classification. It does so by modifying the hinge loss function of Ranking SVM. Specifically, it sets different losses for document pairs across different grades and from different queries. To emphasize the importance of correct ranking on the top, the loss function heavily penalizes errors related to the top. To increase the influence of queries with less documents, the loss function heavily penalizes errors from the queries.

Figure 7 plots the shapes of different hinge loss functions with different penalty parameters. The x-axis represents $yf(x)$ and the y-axis represents loss. When $yf(x_i^{(1)} - x_i^{(2)}) \geq 1$, the losses are zero. When $yf(x_i^{(1)} - x_i^{(2)}) < 1$, the losses are represented by linearly decreasing functions with different slopes. If the slope equals -1 , then the function is the normal hinge loss function. IR SVM modifies the hinge loss function, specifically modifies the slopes for different grade pairs and different queries. It assigns higher weights to document pairs across important grade pairs and assigns normalization weights to document pairs according to

queries.

The learning of IR SVM is equivalent to the following optimization problem. Specifically, the minimization of the modified regularized hinge loss function,

$$\min_w \sum_{i=1}^m \tau_{k(i)} \mu_{q(i)} [1 - y_i \langle w, x_i^{(1)} - x_i^{(2)} \rangle]_+ + \lambda \|w\|^2,$$

where $[x]_+$ denotes $\max(x, 0)$, $\lambda = \frac{1}{2C}$, and $\tau_{k(i)}$ and $\mu_{q(i)}$ are weights. See the loss function of Ranking SVM (5).

Here $\tau_{k(i)}$ represents the weight of instance (document pair) i whose label pair belongs to the k -th type. Xu et al. propose a heuristic method to determine the value of τ_k . The method takes the average drop in NDCG@1 when randomly changing the positions of documents belonging to the grade pair as the value of a grade pair τ_k . Moreover, $\mu_{q(i)}$ represents the weight of instance (document pair) i which is from query q . The value of $\mu_{q(i)}$ is simply determined by $\frac{1}{|n_q|}$, where n_q is the number of document pairs for query q .

The equivalent QP problem is as below.

$$\begin{aligned} & \min_{w, \xi} \frac{1}{2} \|w\|^2 + C_i \sum_{i=1}^m \xi_i \\ & \text{s. t. } y_i \langle w, x_i^{(1)} - x_i^{(2)} \rangle \geq 1 - \xi_i, \\ & C_i = \frac{\tau_{k(i)} \mu_{q(i)}}{2\lambda} \\ & \xi_i \geq 0, \\ & i = 1, \dots, m. \end{aligned}$$

5. Listwise Approach

The listwise approach addresses the ranking problem in a more straightforward way. Specifically, it takes ranking lists as instances in both learning and prediction. The group structure of ranking is maintained and ranking evaluation measures can be more directly incorporated into the loss functions in learning.

The listwise approach includes ListNet [18], ListMLE [19], AdaRank [20], SVM MAP [21], and Soft Rank [22]. SVM MAP and related methods are explained in this article.

5.1 SVM MAP

The algorithm SVM MAP developed by Yue et al. [21] is designed to directly optimize MAP [2], but it can be easily extended to optimize NDCG. Xu et al. [23] further generalize it to a group of algorithms.

In ranking, for query q_i the ranking model $f(x_{ij})$ assigns a score to each associated document d_{ij} or feature vector x_{ij} where x_{ij} is the feature vector derived from q_i and d_{ij} . The documents \mathbf{d}_i (feature vectors \mathbf{x}_i) are then sorted based on their scores and a permutation denoted as π_i is obtained. For simplicity, suppose that the ranking model $f(x_{ij})$ is a linear model:

$$f(x_{ij}) = \langle w, x_{ij} \rangle, \quad (6)$$

where w denotes a weight vector.

Suppose that labels for the feature vectors \mathbf{x}_i are also given as \mathbf{y}_i . We consider using a scoring function $S(\mathbf{x}_i, \pi_i)$ to measure the goodness of ranking π_i . $S(\mathbf{x}_i, \pi_i)$ is defined as

$$S(\mathbf{x}_i, \pi_i) = \langle w, \sigma(\mathbf{x}_i, \pi_i) \rangle,$$

where w is still the weight vector and vector $\sigma(\mathbf{x}_i, \pi_i)$ is defined as

$$\sigma(\mathbf{x}_i, \pi_i) = \frac{2}{n_i(n_i - 1)} \sum_{k,l:k < l} [z_{kl}(x_{ik} - x_{il})],$$

where $z_{kl} = +1$, if $\pi_i(k) < \pi_i(l)$ (x_{ik} is ranked ahead of x_{il} in π_i), and $z_{kl} = -1$, otherwise. Recall that n_i is the number of documents associated with query q_i .

For query q_i , we can calculate $S(\mathbf{x}_i, \pi_i)$ for each permutation π_i and select the permutation $\tilde{\pi}_i$ with the largest score:

$$\tilde{\pi}_i = \arg \max_{\pi_i \in \Pi_i} S(\mathbf{x}_i, \pi_i), \quad (7)$$

where Π_i denotes the set of all possible permutations for \mathbf{x}_i .

It can be easily shown that the ranking $\tilde{\pi}_i$ selected by Eq.(7) is equivalent to the ranking created by the ranking model $f(x_{ij})$ (when both of them are linear functions). Figure 8 gives an example. It is easy to verify that both $f(x)$ and $S(\mathbf{x}_i, \pi)$ will output ABC as the most preferable ranking (permutation).

Objects: A, B, C
 $f_A = \langle w, x_A \rangle, f_B = \langle w, x_B \rangle, f_C = \langle w, x_C \rangle$
Suppose $f_A > f_B > f_C$
For example:
Permutation1: ABC
Permutation2: ACB
 $S_{ABC} = \frac{1}{6} \langle w, ((x_A - x_B) + (x_B - x_C) + (x_A - x_C)) \rangle$
 $S_{ACB} = \frac{1}{6} \langle w, ((x_A - x_C) + (x_C - x_B) + (x_A - x_B)) \rangle$
 $S_{ABC} > S_{ACB}$

Fig. 8 Example of Scoring Function

In learning, we would ideally create a ranking model that can maximize the accuracy in terms of a listwise evaluation measure on training data, or equivalently, minimizes the loss function defined below,

$$L(f) = \sum_{i=1}^m (E(\pi_i^*, \mathbf{y}_i) - E(\pi_i, \mathbf{y}_i)), \quad (8)$$

where π_i is the permutation on feature vector \mathbf{x}_i by ranking model f and \mathbf{y}_i is the corresponding list of grades. $E(\pi_i, \mathbf{y}_i)$ denotes the evaluation result of π_i in terms of an evaluation measure (e.g., NDCG). Usually $E(\pi_i^*, \mathbf{y}_i) = 1$.

We view the problem of learning a ranking model as the following optimization problem in which the following loss function is minimized.

$$\sum_{i=1}^m \max_{\pi_i^* \in \Pi_i^*; \pi_i \in \Pi_i \setminus \Pi_i^*} (E(\pi_i^*, \mathbf{y}_i) - E(\pi_i, \mathbf{y}_i)) \cdot [[S(\mathbf{x}_i, \pi_i^*) \leq S(\mathbf{x}_i, \pi_i)]], \quad (9)$$

where $[[c]]$ is one if condition c is satisfied, otherwise it is zero. $\pi_i^* \in \Pi_i^* \subseteq \Pi_i$ denotes any of the perfect permutations for q_i .

The loss function measures the loss when the most preferred ranking list by the ranking model is not the perfect ranking list. One can prove that the true loss function such as that in (8) is upper-bounded by the new loss function in (9).

The loss function (9) is still not continuous and differentiable. We can consider using continuous, differentiable, and even convex upper bounds of the loss function (9).

1) The 0-1 function in (9) can be replaced with its upper bounds, for example, hinge functions, yielding

$$\begin{aligned} & \sum_{i=1}^m \max_{\pi_i^* \in \Pi_i^*; \pi_i \in \Pi_i \setminus \Pi_i^*} (E(\pi_i^*, \mathbf{y}_i) - E(\pi_i, \mathbf{y}_i)) \cdot \\ & [1 - (S(\mathbf{x}_i, \pi_i^*) - S(\mathbf{x}_i, \pi_i))]_+ \\ & \sum_{i=1}^m \left[\max_{\pi_i^* \in \Pi_i^*; \pi_i \in \Pi_i \setminus \Pi_i^*} ((E(\pi_i^*, \mathbf{y}_i) - E(\pi_i, \mathbf{y}_i)) \right. \\ & \left. - (S(\mathbf{x}_i, \pi_i^*) - S(\mathbf{x}_i, \pi_i))) \right]_+, \end{aligned}$$

2) The max function can also be replaced with its upper bound, the sum function. This is because $\sum_i x_i \geq \max_i x_i$ if $x_i \geq 0$ holds for all i .

3) Relaxations 1 and 2 can be applied simultaneously.

For example, using the hinge function and taking the true loss as $1.0 - MAP$, we obtain SVM MAP. More precisely, SVM MAP solves the following QP problem:

$$\begin{aligned} & \min_{w; \xi \geq 0} \frac{1}{2} \|w\|^2 + \frac{C}{m} \sum_{i=1}^m \xi_i \\ & s.t. \quad \forall i, \forall \pi_i^* \in \Pi_i^*, \forall \pi_i \in \Pi_i \setminus \Pi_i^* : \\ & S(\mathbf{x}_i, \pi_i^*) - S(\mathbf{x}_i, \pi_i) \geq E(\pi_i^*, \mathbf{y}_i) - E(\pi_i, \mathbf{y}_i) - \xi_i, \end{aligned} \quad (10)$$

where C is a coefficient and ξ_i is the maximum loss among all the losses for permutations of query q_i .

Equivalently, SVM MAP minimizes the following regularized hinge loss function

$$\sum_{i=1}^m \left[\max_{\pi_i^* \in \Pi_i^*; \pi_i \in \Pi_i \setminus \Pi_i^*} (E(\pi_i^*, \mathbf{y}_i) - E(\pi_i, \mathbf{y}_i)) \right. \\ \left. - (S(\mathbf{x}_i, \pi_i^*) - S(\mathbf{x}_i, \pi_i)) \right]_+ + \lambda \|w\|^2. \quad (11)$$

Intuitively, the first term calculates the total maximum loss when selecting the best permutation for each of the queries. Specifically, if the difference between the permutations $S(\mathbf{x}_i, \pi_i^*) - S(\mathbf{x}_i, \pi_i)$ is less than the difference between the corresponding evaluation measures $E(\pi_i^*, \mathbf{y}_i) - E(\pi_i, \mathbf{y}_i)$, then there will be a loss, otherwise not. Next, the maximum loss is selected for each query and they are summed up over all the queries.

Since $c \cdot [[x \leq 0]] < [c - x]_+$ holds for all $c \in \mathbb{R}^+$ and $x \in \mathbb{R}$, it is easy to see that the loss in (11) also bounds the true loss function in (8).

6. Ongoing and Future Work

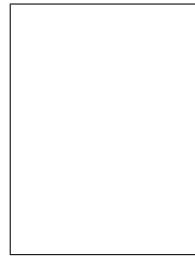
It is still necessary to develop more advanced technologies for learning to rank. There are also many open

questions with regard to theory and applications of learning to rank [2], [24]. Current and future research directions include

- training data creation
- semi-supervised learning and active learning
- feature learning
- scalable and efficient training
- domain adaptation and multi-task learning
- ranking by ensemble learning
- global ranking
- ranking of nodes in a graph.

References

- [1] T.Y. Liu, "Learning to rank for information retrieval," *Foundations and Trends in Information Retrieval*, vol.3, no.3, pp.225–331, 2009.
- [2] H. Li, "Learning to rank for information retrieval and natural language processing," *Synthesis Lectures on Human Language Technologies*, 2011, Morgan & Claypool Publishers.
- [3] W.B. Croft, D. Metzler, and T. Strohman, *Search Engines - Information Retrieval in Practice*, Pearson Education, 2009.
- [4] T.Y. Liu, J. Xu, T. Qin, W. Xiong, and H. Li, "LETOR: Benchmark dataset for research on learning to rank for information retrieval," *Proceedings of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval*, 2007.
- [5] K. Järvelin and J. Kekäläinen, "IR evaluation methods for retrieving highly relevant documents," *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, New York, NY, USA, pp.41–48, ACM, 2000.
- [6] D. Cossack and T. Zhang, "Subset ranking using regression..," *COLT '06: Proceedings of the 19th Annual Conference on Learning Theory*, pp.605–619, 2006.
- [7] R. Herbrich, T. Graepel, and K. Obermayer, *Large Margin rank boundaries for ordinal regression*, MIT Press, Cambridge, MA, 2000.
- [8] Y. Freund, R.D. Iyer, R.E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences..," *Journal of Machine Learning Research*, vol.4, pp.933–969, 2003.
- [9] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pp.89–96, 2005.
- [10] W. Chen, T.Y. Liu, Y. Lan, Z.M. Ma, and H. Li, "Ranking measures and loss functions in learning to rank," *NIPS '09*, 2009.
- [11] P. Li, C. Burges, and Q. Wu, "McRank: Learning to rank using multiple classification and gradient boosting," in *Advances in Neural Information Processing Systems 20*, ed. J. Platt, D. Koller, Y. Singer, and S. Roweis, pp.897–904, MIT Press, Cambridge, MA, 2008.
- [12] K. Crammer and Y. Singer, "Pranking with ranking..," *NIPS*, pp.641–647, 2001.
- [13] A. Shashua and A. Levin, "Ranking with large margin principle: Two approaches," in *Advances in Neural Information Processing Systems 15*, ed. S.T. S. Becker and K. Obermayer, MIT Press.
- [14] Z. Zheng, H. Zha, T. Zhang, O. Chapelle, K. Chen, and G. Sun, "A general boosting method and its application to learning ranking functions for web search," in *Advances in Neural Information Processing Systems 20*, ed. J. Platt, D. Koller, Y. Singer, and S. Roweis, pp.1697–1704, MIT Press, Cambridge, MA, 2008.
- [15] Y. Cao, J. Xu, T.Y. Liu, H. Li, Y. Huang, and H.W. Hon, "Adapting ranking SVM to document retrieval," *SIGIR' 06*, pp.186–193, 2006.
- [16] C. Burges, R. Ragno, and Q. Le, "Learning to rank with nonsmooth cost functions," in *Advances in Neural Information Processing Systems 18*, pp.395–402, MIT Press, Cambridge, MA, 2006.
- [17] Q. Wu, C.J.C. Burges, K.M. Svore, and J. Gao, "Adapting boosting for information retrieval measures," *Inf. Retr.*, vol.13, no.3, pp.254–270, 2010.
- [18] Z. Cao, T. Qin, T.Y. Liu, M.F. Tsai, and H. Li, "Learning to rank: from pairwise approach to listwise approach," *ICML '07: Proceedings of the 24th international conference on Machine learning*, pp.129–136, 2007.
- [19] F. Xia, T.Y. Liu, J. Wang, W. Zhang, and H. Li, "Listwise approach to learning to rank: theory and algorithm," *ICML '08: Proceedings of the 25th international conference on Machine learning*, New York, NY, USA, pp.1192–1199, ACM, 2008.
- [20] J. Xu and H. Li, "AdaRank: a boosting algorithm for information retrieval," *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, pp.391–398, ACM, 2007.
- [21] Y. Yue, T. Finley, F. Radlinski, and T. Joachims, "A support vector method for optimizing average precision," *Proceedings of the 30th annual international ACM SIGIR conference*, pp.271–278, 2007.
- [22] M. Taylor, J. Guiver, S. Robertson, and T. Minka, "Soft-Rank: optimizing non-smooth rank metrics," *WSDM '08: Proceedings of the international conference on Web search and web data mining*, New York, NY, USA, pp.77–86, ACM, 2008.
- [23] J. Xu, T.Y. Liu, M. Lu, H. Li, and W.Y. Ma, "Directly optimizing evaluation measures in learning to rank," *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, pp.107–114, ACM, 2008.
- [24] O. Chapelle, Y. Chang, and T.Y. Liu, "Future directions in learning to rank," *Journal of Machine Learning Research - Proceedings Track*, vol.14, pp.91–100, 2011.



Hang Li Senior researcher and research manager in Web Search and Mining Group at Microsoft Research Asia. He joined Microsoft Research in June 2001. Prior to that, He worked at the Research Laboratories of NEC Corporation. He obtained a B.S. in Electrical Engineering from Kyoto University in 1988 and a M.S. in Computer Science from Kyoto University in 1990. He earned his Ph.D. in Computer Science from the University of Tokyo in 1998. He is interested in statistical learning, information retrieval, data mining, and natural language processing.

of Tokyo in 1998. He is interested in statistical learning, information retrieval, data mining, and natural language processing.

中国计算机学会《学科前沿讲习班》第 21 期
面向互联网的自然语言处理技术——理论、方法与应用问题研究

第二讲

自然语言处理和信息抽取

赵军 (jzhao@nlpr.ia.ac.cn)

共同讲者：刘康、韩先培、周光有、蔡黎

中国科学院自动化研究所
模式识别国家重点实验室

引言(1/3)

- 60-80年代：自然语言处理是研究如何使机器具有语言分析和理解智能的一门学科，是随着人工智能学科的发展而发展起来的
 - 最早的自然语言理解研究工作是机器翻译，20世纪60年代，国外对机器翻译曾有大规模的研究工作
 - 普遍采用基于规则的方法，或者基于知识库的方法，在限定领域取得成功
 - 但人们普遍低估了自然语言的复杂性，开放领域自然语言处理遇到很大困难
- 90年代开始：随着大规模词典和真实语料库的研制，给自然语言处理领域的研究带来了巨大变化
 - 输入：自然语言处理技术开始面向大规模真实文本的处理，所研制的系统开始面向实用
 - 输出：系统并不要求对自然语言文本进行深层理解，而是从中抽取一些有用信息，作为自然语言部分理解的一种形式——信息抽取
 - 方法：基于语料库的统计自然语言学习成为一种重要的方法

引言(2/3)

- **过去10年**: 随着互联网的普及, 为自然语言处理领域提供了强有力的应用牵引和海量语言资源
 - 自然语言处理技术和信息检索技术相结合, 自然语言处理技术的应用领域大大扩大——问答系统等
 - 统计自然语言学习方法受限于语料库的规模, 过拟合问题严重, 缺乏推广能力, 遇到瓶颈
- **目前**: 随着Web2.0的普及, 网络上积累了规模巨大的用户生成内容(User Generated Content), 为自然语言处理技术的发展提供了新的资源和技术创新的源泉
 - 例如Wikipedia、社区问答资源等, 为建立大规模知识库奠定基础
 - 基于知识的方法在开放域自然语言处理任务中的应用成为可能
 - 基于知识的方法和基于统计的方法的融合开始受到关注

引言(3/3)

- 本课程将面向互联网应用，选取依存句法分析、信息抽取、观点挖掘和倾向性分析、问答系统等四个自然语言处理领域的研究方向，系统介绍其中的基本概念、主要方法、最新研究进展、需要解决的问题和发展趋势
- 目标：能够对以上几个研究方向的基本轮廓和发展脉络有较为系统的了解

主要内容

- 第一课 09:00-10:00 信息抽取
- 第二课 10:20-11:30 观点挖掘和倾向性分析
- 第三课 14:00-15:00 问答系统
- 第四课 15:20-16:20 依存句法分析
- 第五课 16:30-17:00 互动课

主要内容之间的关系

- 信息抽取：事实性信息的抽取
- 观点挖掘和倾向性分析：主观性信息的抽取
- 问答系统：网络搜索、自然语言处理、信息抽取技术相结合的应用
- 依存句法分析：自然语言处理关键技术

第一课

信息抽取

中国科学院自动化研究所
模式识别国家重点实验室

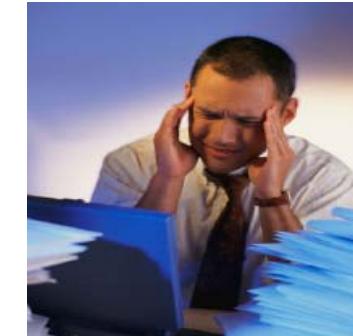
概述

- 引言
- 实体识别与抽取
- 实体消歧
- 关系抽取
- 问题与挑战

引言

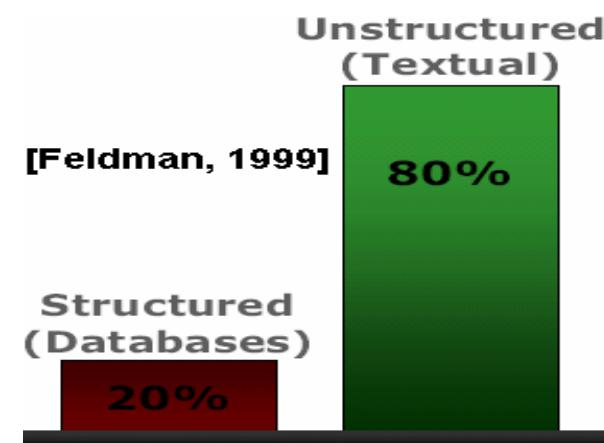
□ 互联网的迅速普及和发展

- 信息资源极大丰富
- 但“信息过载”问题日趋严重



□ 迫切需要快速、准确获取信息的技术手段

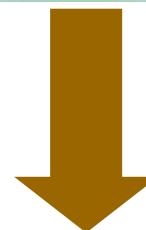
- 信息抽取技术应运而生
 - 文本信息抽取
 - 自然语言文本信息抽取



信息抽取与信息检索的区别 (1/2)



信息抽取与信息检索的区别 (2/2)



需要文本信息抽取技术做支撑

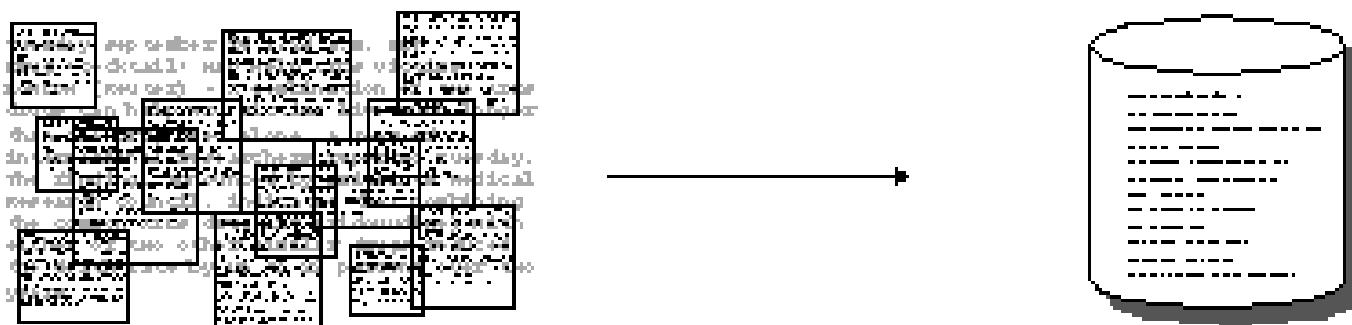
信息抽取定义

□ 信息抽取定义(Grishman, 1997)

- 从自然语言文本中抽取指定类型的实体、关系、事件等事实信息，并形成结构化数据输出的文本处理技术

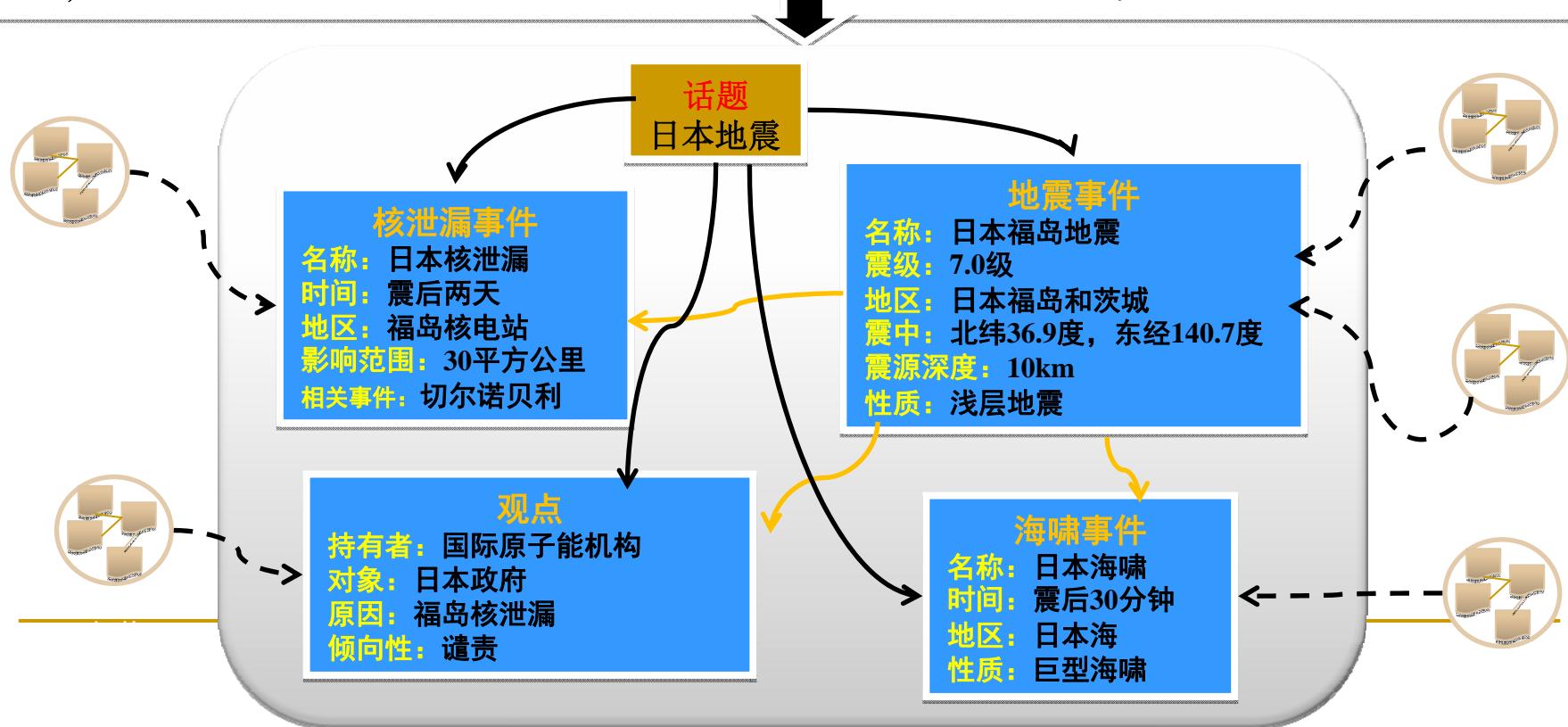
□ 目标：

- Making information more machine-readable (Wu, 2010)



信息抽取的目标示例

2011年4月11日17点16分，日本东北部的福岛和茨城地区发生里氏7.0级强烈地震（震中北纬36.9度、东经140.7度，即福岛西南30公里左右的地方，震源深度10公里，属于浅层地震）当局已经发布海啸预警震后约30分钟后在日本海地区发生巨型海啸，同时造成福岛核电站出现核泄漏震后第十天，国际原子能机构对于日本政府反应迟钝进行了谴责



信息抽取的历史(1/2)

- 信息抽取的研究最早于上世纪70年代末期
- 最早的信息抽取系统于上世纪80年代中期面世，由路透社研制的JASPER，用作向金融从业者提供结构化的金融新闻

信息抽取的历史 (2/2)

- MUC (Message Understanding Conferences, 1987-1997)
 - 由美国国防高级研究计划委员会DARPA资助
 - 主要是英文，后两届扩展到中文
 - 任务命名实体识别、共指消解、模板关系抽取等
- ACE (Automatic Content Extraction, 1999-2008)
 - 由美国国家标准与技术研究所NIST主办
 - 2009起，ACE变成了TAC (Text Analysis Conference) 的一项子任务
 - 英文、中文、阿拉伯文等
 - 任务命名实体识别、关系抽取、事件抽取等等
- TAC-KBP (Knowledge Base Population)子任务 (2009-今)
 - 实体链接、属性抽取

小结

时间	评测阶段	任务	语料
1987-1993	MUC1-MUC5	命名实体识别, 共指消解, 模板关系抽取等等	限定领域文本 (海军军事情报、恐怖袭击)
1995-1997	MUC6-MUC7	模板填充、命名实体识别、共指关系确定等	限定领域文本 (人事职位变动、飞机失事)
1999	ACE-Pilot---- ACE-1	命名实体识别	新闻语料
2002	ACE2	命名实体识别、关系识别与描述等	新闻语料
2003-2007	ACE2003- ACE2007	命名实体识别、关系识别与描述、 时间表达式识别、事件抽取等	新闻语料、对话语料
2009-2011	TAC1-TAC3	实体链接、属性抽取	新闻语料、Web页面

- 封闭语料→开放语料：限定领域的新闻语料→Web页面
- 限定类别→开放类别：有限类别的实体、关系、事件→维基百科条目
- 文本内信息抽取→与真实世界关联

概述

- 引言
- 实体识别与抽取
 - 实体识别
 - 开放域实体抽取
- 实体消歧
- 关系抽取
- 问题与挑战

命名实体识别的任务

- 识别出待处理文本中七类（人名、机构名、地名、时间、日期、货币和百分比）命名实体
- 两个子任务：实体边界识别和确定实体类别

2011年4月11日17点16分，日本东北部的福岛和茨城地区发生里氏7.0级强烈地震（震中北纬36.9度、东经140.7度，即福岛西南30公里左右的地方，震源深度10公里，属于浅层地震）当局已经发布海啸预警震后约30分钟后在日本海地区发生巨型海啸，同时造成福岛核电站出现核泄漏震后第十天，国际原子能机构对于日本政府反应迟钝进行了谴责



2011年4月11日17点16分，日本东北部的福岛和茨城地区发生里氏7.0级强烈地震（震中北纬36.9度、东经140.7度，即福岛西南30公里左右的地方，震源深度10公里，属于浅层地震）当局已经发布海啸预警震后约30分钟后在日本海地区发生巨型海啸，同时造成福岛核电站出现核泄漏震后第十天，国际原子能机构对于日本政府反应迟钝进行了谴责

命名实体识别的特点

- 时间、日期、货币和百分比的构成有比较明显的规律，识别起来相对容易
- 人名、地名、机构名的用字灵活，识别的难度很大
 - 内部结构复杂，形式多变，中文命名实体中这种情况更为严重
 - 人名：杜甫、杜子美、子美、杜工部、李杜
 - 机构名：北京百富勤投资咨询公司、北京大学附属小学、中国奥委会、北师大二附
 - 上下文密切相关
 - 不同语境下，可能具有不同的实体类型；或者在某些条件下是实体，在另外的条件下就不是实体
 - 彩霞、河南、新世界

命名实体识别的方法

- 命名实体的内部构成和外部语言环境具有一些特征
- 无论何种方法，都在试图充分发现和利用实体所在的上下文特征和实体的内部特征
- 考虑到每一类命名实体都具有不同的特征，不同类别的实体适合用不同的识别模型
 - 人名：用基于字的模型描述其内部构成
 - 地名和机构名：用基于词的模型描述
 - 不同类型的外国人名用字存在较大差别，如果按照人名的用字和构成特点，把人名分成多个类别并分别利用不同模型进行识别，对于提高人名识别的正确率是非常有益的（Wu Youzheng, EMNLP 2005）
- 利用序列标注工具计算特征权重
 - MEMM、HMM、CRF

命名实体识别的评测

- 国际会议：MUC、SigHAN、CoNLL、IEER 和 ACE
 - MUC-6和MUC-7设立的命名实体识别专项评测大大推动了英语命名实体识别技术的发展
 - MUC-6和MUC-7还设立了多语言实体识别评测任务MET，对日语、西班牙语、汉语等多种语言命名实体识别任务进行评测
 - SigHAN从2003年开始举办第一届中文分词评测BAKEOFF，2006年和2008年举行的BAKEOFF-3和BAKEOFF-4设立了命名实体识别专项评测
 - 2003年和2004年举办的863计划“中文信息处理与智能人机接口技术评测”中设立了中文命名实体识别评测任务

英文命名实体识别的技术水平

- 英文：Language Technology Group Summary开发的英语命名实体识别系统在MUC-7评测中取得第一名，其准确率和召回率分别达到95%和92%
(吴友政, 2006)
- 许多英语命名实体系统已经具备了相当程度的大规模文本处理能力

汉语命名实体识别的技术水平

- 参加MET-2评测的汉语命名实体识别系统对人名、地名、机构名识别的最优性能指标（准确率，召回率）只有(66%，92%)、(89%，91%)和(89%，88%)（吴友政，2006）

汉语命名实体识别的技术水平：BAKEOFF-3

数据来源	测试类别	测试性能						
		P	R	F	ORG-F	LOC-F	PER-F	GPE-F
MSRC (简)	Closed	0.8894	0.8420	0.8651	0.8310	0.8545	0.9009	~
	Open	0.9220	0.9018	0.9118	0.8590	0.9034	0.9604	~
LDC (简)	Closed	0.8026	0.7265	0.7627	0.6585	0.3046	0.7884	0.8204
	Open	0.7616	0.6621	0.7084	0.5209	0.2857	0.7422	0.7930
CITYU (繁)	Closed	0.9143	0.8676	0.8903	0.8046	0.9211	0.9087	~
	Open	0.8692	0.7498	0.8051	0.6801	0.8604	0.8098	~

数据来源	简繁体类别	训练集规模	测试集规模
MSRC	简体	1.3M/63K (词/词次)	100K/13K (词/词次)
LDC	简体	632K (词)	61K (词)
CITYU	繁体	1.6M/76K (词/词次)	220K/23K (词/词次)

(Levow, 2006)

汉语命名实体识别的技术水平：863

数据 来源	测试 类别	测试性能								
		ORG			LOC			PER		
		P	R	F	P	R	F	P	R	F
SXU (简)	开放	.646 4	.5741	.6081	.8702	.7843	.8251	.8138	.8847	.8478
CITYU (繁)	开放	.398 6	.2532	.3097	.6839	.7004	.6921	.3986	.2532	.3097

数据来源	简繁体类别	训练集规模 (词/词次)	测试集规模 (字)
SXU	简体	NONE	约400K
CITYU	繁体	NONE	约400K

从评测结果看汉语命名实体识别的技术水平

- 在BAKEOFF-3 MSRC语料和BAKEOFF-3 CITYU语料上的评测结果要好于BAKEOFF-3 LDC语料上的评测结果以及863语料上的评测结果
- 其中一个很重要原因是：BAKEOFF-3 MSRC和CITYU评测提供了相当规模的训练集，而BAKEOFF-3 LDC只提供了小规模的训练集，而863评测根本不提供训练集
- 因为训练集和测试集在题材和体裁方面比较类似，可能使得各个系统在BAKEOFF-3 MSRC语料和BAKEOFF-3 CITYU语料上的评测性能较高
- 在真实的应用环境中，命名实体识别的性能会大打折扣

小结

- 受限于训练语料规模，系统的自适应能力不强
 - 网页信息：不规范、存在很多噪音，有些根本就不构成自然语言句子，因此通常的命名实体识别模型所依赖的上下文特征发生了明显变化，使得识别性能剧烈下降
- 类别数限定，不满足实际的应用
 - 摩托罗拉V8088折叠手机、第6届苏迪曼杯羽毛球混合团体赛、胆结石腹腔镜手术等
- 需要开放域实体抽取
 - 实体类型更多、更细，而且有些实体类别是未知、或者是随时间演化的

概述

- 引言
- 实体识别与抽取
 - 实体识别
 - 开放域实体抽取
- 实体消歧
- 关系抽取
- 问题与挑战

开放域实体抽取

Output

- 不限定实体类别
- 不限定目标文本
- 任务：给定某一类别的实体实例，从网页中抽取同一类别其他实体实例
 - 给出<中国，美国，俄罗斯>（称为“种子”）
 - 找出其他国家或地区<德国，英国，法国……>

Input

中国
美国
俄罗斯



Predicted Items
中国
俄罗斯
美国
日本
德国
英国
法国
韩国
印度
香港
意大利
巴西
加拿大
新加坡

开放域实体抽取的主要方法

□ 基本思路

- 种子词与目标词在网页中具有相同或者类似的上下文
 - 网页结构
 - 上下文
- Step1：种子词→模板
- Step2：模板→更多同类实体

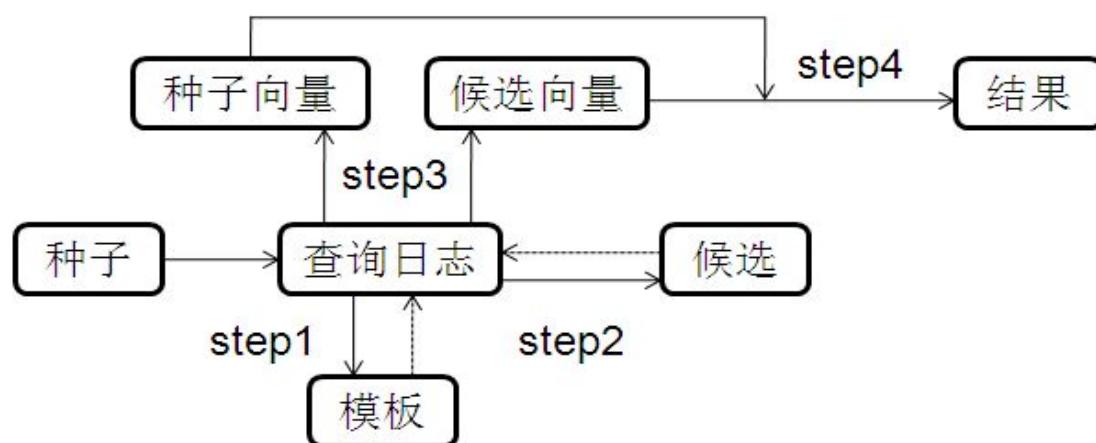
□ 处理实例扩展问题的主流框架



- 利用不同数据源（例如查询日志、网页文档、知识库文档等）的不同特点，设计方法

开放域实体抽取的主要方法: Query Log (Pasca CIKM 2007)

- 通过分析种子实例在查询日志中的上下文得模板，再利用模板找到同类别的实例
 - 联想 笔记本 如何
 - 苹果 笔记本 如何
 - 戴尔 笔记本 如何
- 构造候选与种子上下文向量，计算相似度



开放域实体抽取的主要方法: Web Page (Wang ICDM 2007) (1/2)

□ Motivation

□ 处理列表型网页

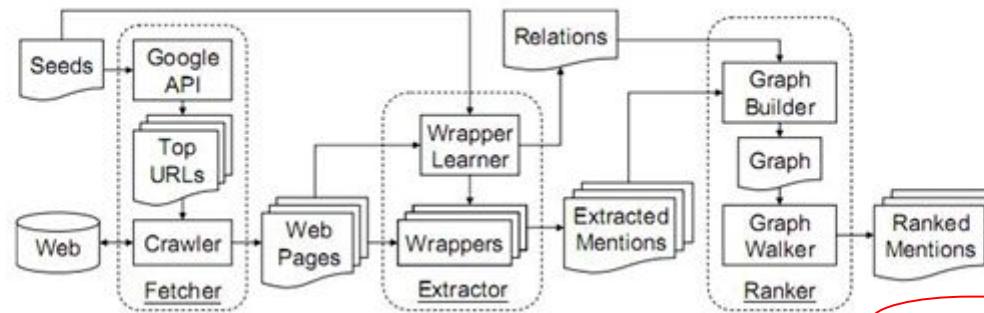
□ 在列表中，种子与目标实体具有相同的网页结构

The screenshot shows a search interface for car brands. At the top, there's a logo for '车168' and a navigation bar with categories like '微型车', '小型车', and '紧凑型车'. Below the navigation bar, there are links for '首页', '汽车报价', '汽车图片', '车型对比', and '二手车'. A search bar contains the placeholder '按价格查找 | 按...' and a dropdown menu for '按拼音查品牌' with options from A to J. A red box highlights a section titled '热门品牌' containing a grid of car brands: Audi, BMW, Porsche, Mercedes-Benz, Great Wall, Volkswagen, Toyota, Ford, Kia, Nissan, MG, and Mitsubishi. The entire grid is enclosed in a red box.

```
<li> · <a target="_self" href="#brand_A_2">奥迪</a></li>
<li> · <a target="_self" href="#brand_B_3">宝马</a></li>
<li> · <a target="_self" href="#brand_B_4">保时捷</a></li>
<li> · <a target="_self" href="#brand_B_5">奔驰</a></li>
<li> · <a target="_self" href="#brand_B_7">本田</a></li>
<li> · <a target="_self" href="#brand_B_8">比亚迪</a></li>
<li> · <a target="_self" href="#brand_B_9">标致</a></li>
<li> · <a target="_self" href="#brand_B_10">别克</a></li>
<li> · <a target="_self" href="#brand_C_15">长城</a></li>
<li> · <a target="_self" href="#brand_D_18">大众</a></li>
<li> · <a target="_self" href="#brand_F_24">丰田</a></li>
```

开放域实体抽取的主要方法: Web Page (Wang ICDM 2007) (2/2)

□ 系统框架



■ 爬取模块 (Fetcher)

把种子送到搜索引擎，把返回的前100个网页抓取下来作为语料

一个网页包含的高质量模板
越多，则该网页质量越高
(反之亦然)

■ 抽取模块 (Extractor)

针对单个网页学习模板，再使用模板抽取候选实例

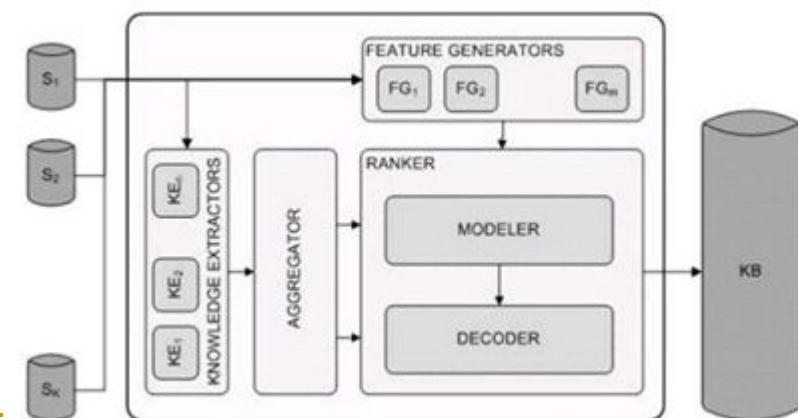
■ 排序模块 (Ranker)

利用种子、网页、模板、候选构造一个图，综合考虑网页和模板
的质量，使用Random Walk算法为候选打分并排序

一个模板抽出的正确实例越
多，则该模板的质量越高 (反
之亦然)

开放域实体抽取的主要方法: 融合多个数据源 (Pennacchiotti EMNLP 2009)

- 针对不同数据源，选取不同特征分别进行实例扩展，对结果进行融合
 - 6亿个网页
 - 1年的查询日志
 - Wikipedia
- 针对不同数据源选取不同的模板和特征
- 使用不同特征计算候选的置信度
 - 结果融合



评价指标与技术水平

- 针对实例扩展问题，目前缺少公认的评测，研究者在自己构建的数据上进行测试
- 评价方法：以平均准确度MAP为主，召回无法评价，覆盖度替代
 - MAP
 - 因为系统输出是一个ranked list，单纯考察准确率无法体现出rank的作用
 - 采用TREC中常用的MAP（平均正确率均值）

$$\text{AvgPrec}(L) = \frac{\sum_{r=1}^{|L|} \text{Prec}(r) \times \text{NewEntity}(r)}{\# \text{True Entities}}$$

- 每个类别做N次实验，每次都随机选取种子，对其求平均

评价指标与技术水平(1/2)

- Wang 2007在12个自制数据集的结果,

取前100个网页作为语料

Table 5. Experimental results

English	G.Sets	Max. 100 Results			Max. 200 E2+GW	Max. 300 E2+GW
		E1+EF	E2+EF	E2+GW		
classic-disney	37.62%	79.36%	74.45%	84.42%	88.20%	89.39%
cmu-buildings	0.00%	87.85%	87.75%	87.83%	87.83%	87.83%
common-diseases	1.12%	17.94%	52.84%	57.46%	75.79%	76.87%
constellations	10.45%	89.61%	99.97%	100.00%	100.00%	100.00%
countries	14.24%	95.95%	97.86%	98.17%	98.67%	98.53%
mlb-teams	70.06%	98.61%	99.50%	99.80%	99.84%	99.81%
nba-teams	90.73%	100.00%	100.00%	100.00%	100.00%	100.00%
nfl-teams	94.26%	99.22%	99.98%	100.00%	100.00%	100.00%
periodic-comets	0.22%	69.24%	79.04%	84.78%	84.77%	84.77%
popular-car-makers	73.61%	79.18%	88.23%	95.16%	96.23%	96.95%
us-presidents	56.77%	91.64%	97.07%	99.99%	100.00%	100.00%
us-states	76.00%	99.96%	93.55%	100.00%	100.00%	100.00%
Average	43.76%	84.05%	89.19%	92.30%	94.28%	94.51%

小结

- 针对不同数据源的特点设计方法，其针对性、灵活性很强
- 方法一般分为模板抽取和实例候选置信度计算两个模块，两部分迭代进行，相互依赖
- 以无监督的方法为主
- 缺少公认的数据集或相关评测

概述

- 引言
- 实体识别与抽取
- 实体消歧
 - 实体消歧任务定义
 - 基于聚类的实体消歧
 - 基于链接的实体消歧
- 关系抽取
- 问题与挑战

实体消歧定义

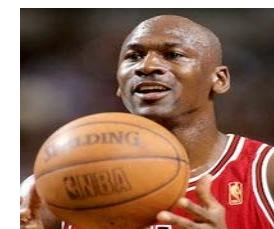
- 命名实体的歧义指的是一个实体指称项可对应到多个真实世界实体，例如，给定如下的四个实体指称项“*Michael Jordan*”

MJ1: Michael Jordan is a researcher in machine learning.

MJ2: Learning in Graphical Models: Michael Jordan

MJ3: M. Jordan wins NBA MVP.

MJ4 : Michael Jordan plays basketball in Chicago Bulls.



- 确定一个实体指称项所指向的真实世界实体，这就是命名实体消歧

实体消歧分类

□ 基于聚类的实体消歧

- 把所有实体指称项按其指向的目标实体进行聚类
- 每一个实体指称项对应到一个单独的类别

MJ1: Michael Jordan is a researcher in machine learning.

MJ2: Research in Graphical Models: Michael Jordan

MJ3: M. Jordan wins NBA MVP.

MJ4 : Michael Jordan plays basketball in Chicago Bulls

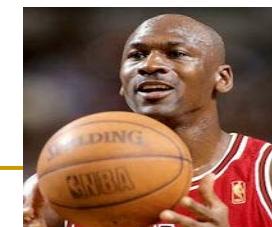
}

}

□ 基于实体链接的实体消歧

- 将实体指称项与目标实体列表中的对应实体进行链接实现消歧

MJ4 : Michael Jordan plays basketball in Chicago Bulls



概述

- 引言
- 实体识别与抽取
- 实体消歧
 - 实体消歧任务定义
 - 基于聚类的实体消歧
 - 基于链接的实体消歧
- 关系抽取
- 问题与挑战

基于聚类的实体消歧

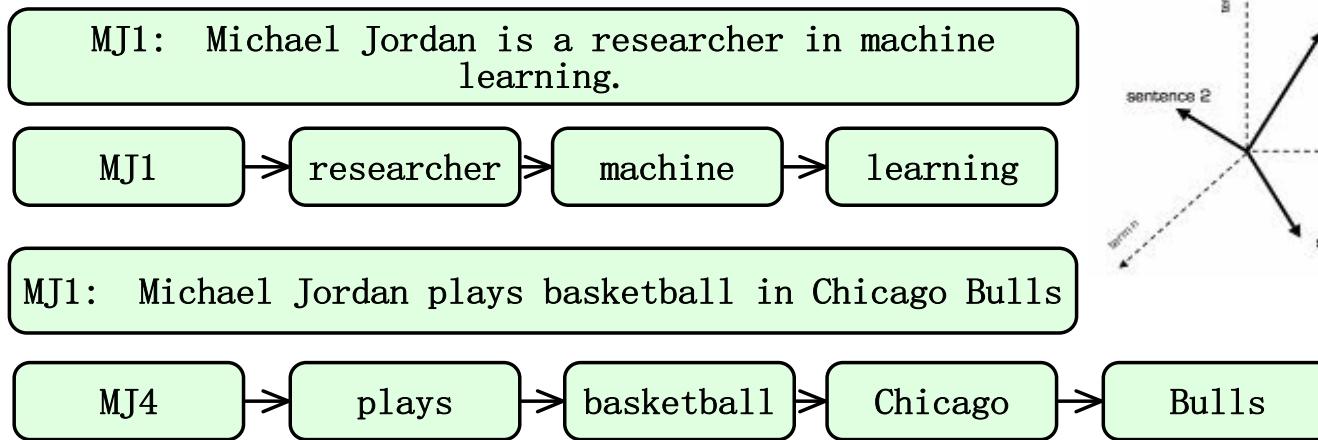
□ 基本思路

- 同一指称项具有近似的上下文
- 利用聚类算法进行消歧
- 核心问题：选取何种特征对于指称项进行表示
 - 词袋模型(Bagga et al., COLING, 1998)
 - 语义特征(Pederson et al., CLITP, 2005)
 - 社会化网络(Bekkerman et al., WWW, 2005)
 - 维基百科的知识(Han and Zhao, CIKM, 2009)
 - 多源异构语义知识融合(Han and Zhao, ACL, 2010)

基于聚类的实体消歧:词袋模型

(Bagga et al. COLING 1998)

- 利用待消歧实体周边的词来构造向量
- 利用向量空间模型来计算两个实体指称项的相似度，进行聚类



- 无法处理

*MJ1: Michael Jordan is a researcher in machine learning.
MJ2: Research in Graphical Models: Michael Jordan*

基于聚类的实体消歧: 语义特征

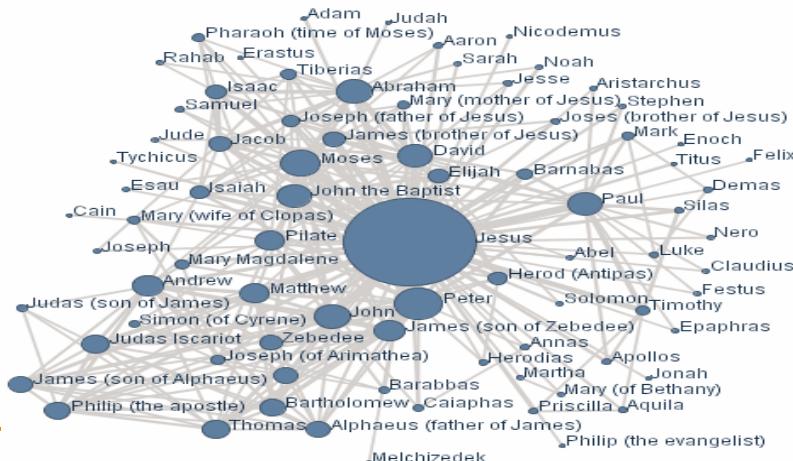
(Pederson et al. CLITP 2005)

- 词袋模型, 没有考虑词的语义信息
- 利用SVD分解挖掘词的语义信息
- 利用词袋和浅层语义特征, 共同来表示指称项, 利用余弦相似度来计算两个指称项的相似度

$$m \begin{matrix} n \\ A \end{matrix} = m \begin{matrix} r \\ U_1 \end{matrix} \begin{matrix} r \\ \Delta \end{matrix} \begin{matrix} n \\ V'_1 \end{matrix} r$$

基于聚类的实体消歧:社会化网络 (Bekkerman et al. WWW 2005)

- 不同的人具有不同的社会关系
 - MJ (BasketBall) : Pippen, Buckley, Ewing, Kobe...
 - MJ (Machine Learning) : Liang, Mackey, Wauthier...
 - 根据MJ, Pippen, Buckley, Ewing, Kobe等的社会化关联所表现出来的网页链接特征, 对网页进行聚类,从而实现网页内的人名聚类消歧



基于聚类的实体消歧: Wikipedia (Han CIKM 2009) (1/3)

□ Wikipedia中相关实体具有链接关系

Plant

From Wikipedia, the free encyclopedia

For other uses, see [Plant \(disambiguation\)](#).

Plants are [living organisms](#) belonging to the [kingdom Plantae](#). Precise definitions of the kingdom vary, but as the term is used here, plants include familiar organisms such as [trees](#), [flowers](#), [herbs](#), [bushes](#), [grasses](#), [vines](#), [ferns](#), [mosses](#), and [green algae](#). The group is also called [green plants](#) or [Viridiplantae](#) in [Latin](#). They obtain most of their energy from [sunlight](#) via [photosynthesis](#) using [chlorophyll](#) contained in [chloroplasts](#), which gives them their green color.

Precise numbers are difficult to determine, but as of 2010, there are thought to be 300–315 thousand [species](#) of plants, of which the great majority, some 260–290 thousand, are [seed plants](#) (see the [table below](#)).^[2]

The scientific study of plants is known as [botany](#).

Tree

From Wikipedia, the free encyclopedia

For other uses, see [Tree \(disambiguation\)](#).

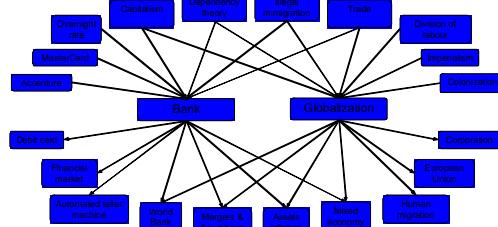
A [tree](#) is a perennial woody plant. It is most often defined as a woody plant that has many secondary branches supported clear of the ground on a single main stem or trunk with clear apical dominance.^[1] A minimum height specification at maturity is cited by some authors, varying from 3 m^[2] to 6 m.^[3] Some authors set a minimum of 10 cm trunk diameter (30 cm girth).^[4] Woody plants that do not meet these definitions by having multiple stems and/or small size are called [shrubs](#). Compared with most other plants, trees are long-lived, some reaching several thousand years old and growing to up to 115 m (379 ft) high.^[5] Trees are an important component of the natural landscape because of their prevention of [erosion](#) and the provision of a weather-sheltered [ecosystem](#) in and under their [foliage](#). They also play an important role in producing [oxygen](#) and reducing [carbon dioxide](#) in the atmosphere, as well as moderating ground temperatures. They are also elements in [landscaping](#) and [agriculture](#), both for their aesthetic appeal and their [orchard crops](#) (such as [apples](#)). [Wood](#) from trees is a [building material](#), as well as a primary energy source in many developing countries. Trees also play a role in many of the world's [mythologies](#) (see [trees in mythology](#)).^[6]

[Contribute](#) [Details](#)



Trees on a mountain in northern Utah

□ 这种链接关系反映条目之间的语义相关度

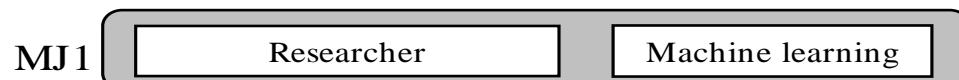


$$sr(a, b) = \frac{\log(\max(|A|, |B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|, |B|))}$$

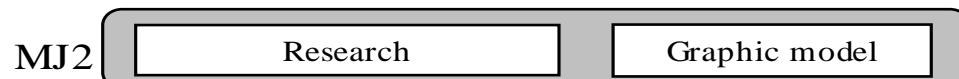
基于聚类的实体消歧: Wikipedia (Han CIKM 2009) (2/3)

□ 用实体上下文的维基条目对于实体进行向量表示

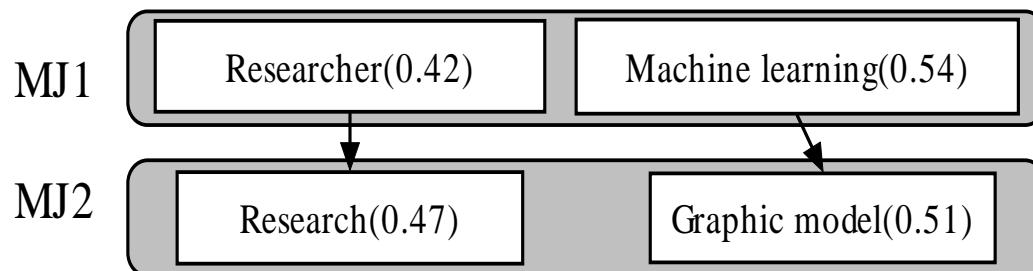
MJ1: Michael Jordan is a Researcher in machine learning.



MJ2: Research in Graphical Models: Michael Jordan



□ 利用维基条目之间的相关度计算指称项之间的相似度（解决数据稀疏问题）



实验比较 (Han CIKM 2009) (3/3)

- 使用WePS数据集测试
- 使用结构化关联语义核的实体相似度能够提升10.7%的消歧性能

Method	WePS1_training		
	Pur	Inv_Pur	F
<i>BOW</i>	0.71	0.88	0.78
<i>SocialNetwork</i>	0.66	0.98	0.76
<i>WikipediaConcept</i>	0.80	0.88	0.82
<i>WS-SameWeight</i>	0.84	0.89	0.85
<i>WS</i>	0.88	0.89	0.87

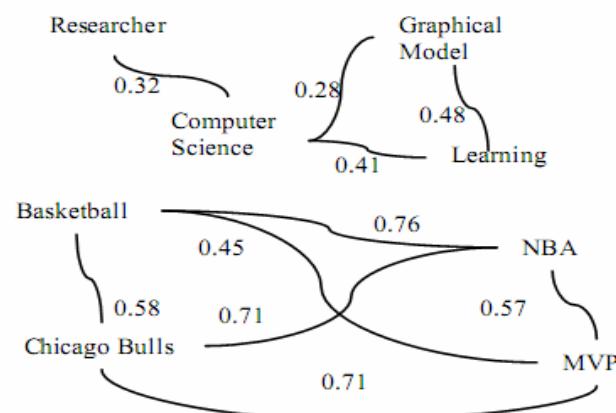
基于聚类的实体消歧: 多源异构知识 (Han ACL 2010) (1/3)

- 仅仅考虑Wikipedia一种知识源, 覆盖度有限
- 多源异构知识的挖掘与集成
 - 知识源中存在大量的多源异构知识
 - 挖掘和集成多源异构知识可以提高实体消歧的性能
 - Wikipedia
 - 用于捕捉概念之间的语义关联
 - WordNet
 - 用于捕捉普通词语之间的语言学关联
 - Web网页库
 - 用于捕捉命名实体之间的社会化关联



基于聚类的实体消歧:多源异构知识 (Han ACL 2010) (2/3)

- 多源异构知识的表示框架
- 如何处理知识源的多源异构性
- 提出了统一的语义知识表示模型——语义图
 - 结构化知识源中的显式语义知识都可以表示成概念之间的语义关联
 - 用语义图把WordNet语义关联、WikiPedia语义关联和Web



语义图

基于聚类的实体消歧:多源异构知识 (Han ACL 2010) (3/3)

□ 实验比较

- 使用WePS数据集测试
- 使用多源知识能够有效提高消歧的准确度

	WePS2_test		
	Pur	Inv_Pur	F
<i>BOW</i>	0.80	0.80	0.77
<i>SocialNetwork</i>	0.62	0.93	0.70
<i>SSR-NoKnowledge</i>	0.84	0.80	0.80
<i>SSR-NoStructure</i>	0.84	0.83	0.81
<i>SSR-NE</i>	0.78	0.88	0.80
<i>SSR-WordNet</i>	0.85	0.82	0.83
<i>SSR-Wikipedia</i>	0.84	0.81	0.82
SSR	0.90	0.86	0.88

基于聚类的实体消歧：评测 (1/2)

- WePS： Web People Search Evaluation
 - WePS1是SEMEVAL2007的子任务
 - WePS2是WWW的一个workshop
 - 任务：Web环境中的人名消歧，即给定一个包含某个歧义人名的网页集合，按照网页中人名指称项所指向的人物概念来对网页进行聚类，以及抽取一个网页中关于某个人的特定属性来辅助进行人名消歧
 - 评测方法

$$\text{Purity} = \sum_i \frac{|C_i|}{n} \max \text{Precision}(C_i, L_j)$$

$$F = \frac{1}{\alpha \frac{1}{\text{Purity}} + (1 - \alpha) \frac{1}{\text{Inverse Purity}}}$$

$$\text{Inverse Purity} = \sum_i \frac{|L_i|}{n} \max \text{Precision}(L_i, C_j)$$

基于聚类的实体消歧：评测 (2/2)

rank	team-id	Macro-averaged Scores			
		$\alpha = .5$	$\alpha = .2$	Pur	Inv_Pur
1	CU_COMSEM	,78	,83	,72	,88
2	IRST-BP	,75	,77	,75	,80
3	PSNUS	,75	,78	,73	,82
4	UVA	,67	,62	,81	,60
5	SHEF	,66	,73	,60	,82
6	FICO	,64	,76	,53	,90
7	UNN	,62	,67	,60	,73
8	ONE-IN-ONE	,61	,52	1,00	,47
9	AUG	,60	,73	,50	,88
10	SWAT-IV	,58	,64	,55	,71
11	UA-ZSA	,58	,60	,58	,64
12	TITPI	,57	,71	,45	,89
13	JHU1-13	,53	,65	,45	,82
14	DFKI2	,50	,63	,39	,83
15	WIT	,49	,66	,36	,93
16	UC3M_13	,48	,66	,35	,95
17	UBC-AS	,40	,55	,30	,91
18	ALL-IN-ONE	,40	,58	,29	1,00

WePS 1

rank	run	Macro-averaged Scores			
		$\alpha = .5$	$\alpha = .2$	Pur	Inv_Pur
1	<i>BEST-HAC-TOKENS</i>	,90	,89	,93	,88
2	<i>BEST-HAC-BIGRAMS</i>	,90	,87	,94	,86
3	PolyUHK	,88	,87	,91	,86
4	UVA_1	,87	,87	,89	,87
5	ITC-UT_1	,87	,83	,95	,81
6	<i>CHEAT-SYS</i>	,87	,94	,78	1,00
7	UMD_4	,81	,76	,95	,72
8	XMEDIA_3	,80	,76	,91	,73
9	UCI_2	,80	,84	,75	,89
10	LANZHOU_1	,80	,78	,85	,77
11	FICO_3	,80	,76	,90	,73
12	<i>HAC-BIGRAMS</i>	,78	,64	,96	,67
13	UGUELPH_1	,74	,84	,64	,95
14	CASIANED_4	,73	,77	,72	,83
15	<i>HAC-TOKENS</i>	,71	,64	,96	,60
16	AUG_4	,69	,68	,79	,68
17	UPM-SINT_4	,67	,70	,69	,74
18	<i>ALL_IN_ONE</i>	,67	,79	,56	1,00
19	UNN_2	,64	,59	,80	,57
20	ECNU_1	,53	,56	,60	,63
21	PRIYAVEN	,53	,49	,71	,48
22	UNED_3	,51	,48	,71	,48
23	BUAP_1	,37	,30	,89	,27
24	<i>ONE_IN_ONE</i>	,34	,27	1,00	,24

WePS 2

小结

- 主要研究集中在实体指称项的语义表示
- 已有工作大多是通过扩展特征，增加更多的知识来提高消歧精度
- 挑战
 - 消歧目标难以确定
 - 缺乏实体的显式表示

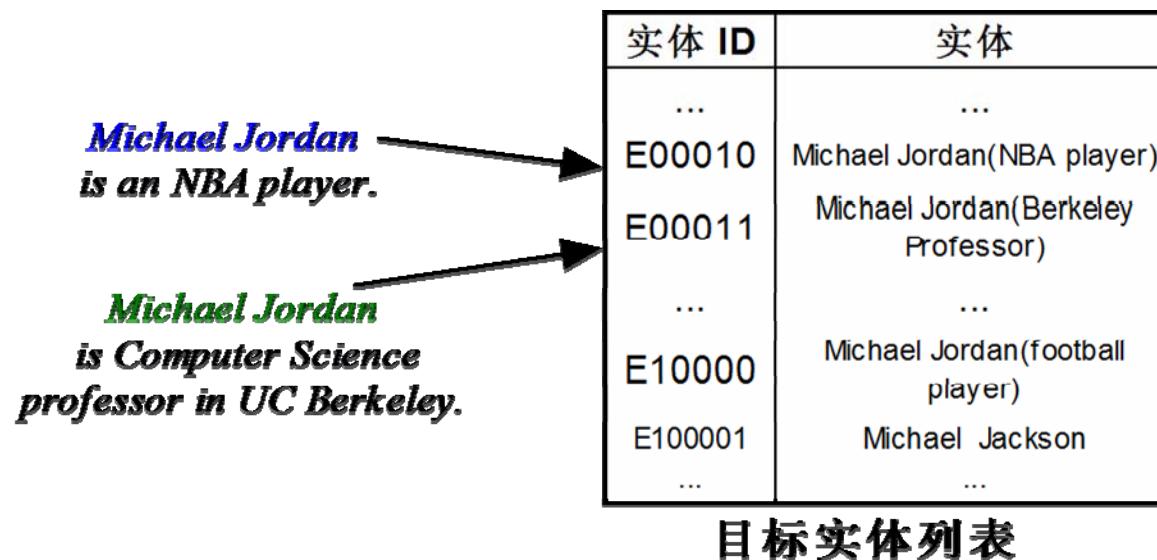
概述

- 引言
- 实体识别与抽取
- 实体消歧
 - 实体消歧任务定义
 - 基于聚类的实体消歧
 - 基于链接的实体消歧
- 关系抽取
- 问题与挑战

实体链接的任务

□ 任务

- 给定实体指称项和它所在的文本，将其链接到给定知识库中的相应实体上



实体链接主要步骤

□ 主要步骤

□ 候选实体的发现

- 给定实体指称项，链接系统根据知识、规则等信息找到实体指称项的候选实体

□ 候选实体的链接

- 系统根据指称项和候选实体之间的相似度等特征，选择实体指称项的目标实体

实体指称项文本：Michael Jordan is
a former NBA player, active
businessman and majority owner of
the Charlotte Bobcats.

候选实体：



- Michael Jordan (basketball player)
- Michael Jordan (mycologist)
- Michael Jordan (footballer)
- Michael B. Jordan
- Michael H. Jordan
- Michael-Hakim Jordan
- Michael Jordan (Irish politician)

...

候选实体发现

□ 如何根据实体指称项找出候选实体

- 利用Wikipedia的信息
- 利用上下文信息

实体指称项	候选实体
Michael Jordan	Michael Jordan (basketball) Michael Jordan (mycologist) Michael Jordan (football) Michael B. Jordan (American actor) ...
AI	Artificial intelligence Ai (singer) ...
...	...

利用Wikipedia信息获取候选实体

□ 利用Wikipedia中锚文本的超级链接关系

□ Michal Jordan is a former NBA player

□ 利用Wikipedia中的消歧页面

Michael Jordan is an American basketball player.

Michael Jordan may also refer to:

- Michael Jordan (mycologist), English mycologist
- Michael Jordan (footballer) (born 1986), English goalkeeper (A)
- Michael B. Jordan (born 1987), American actor
- Michael I. Jordan (born 1957), American researcher in machine learning
- Michael H. Jordan (d. 2010), American executive for CBS, PepsiCo
- Michael-Hakim Jordan (born 1977), American professional basketball player
- Michael Jordan (Irish politician), Irish Farmers' Party TD from V

□ 利用Wikipedia中的重定向页面

Michael jordan

From Wikipedia, the free encyclopedia
Redirect page

↳ Michael Jordan



利用上下文获取缩略语候选实体

(Zhang IJCAI 2011)

□ 问题

- 缩略语在实体指称项中十分常见，据统计，在KBP2009的测试数据，在3904个实体指称项中有827个为缩略语

□ 动机

- 缩略语指称项具有很强的歧义性，但它的全称往往是没有歧义的
- ABC和American Broadcasting Company AI和Artificial Intelligence等
- 在实体指称项文本中，缩略语的全称出现过

□ 解决方法

- 利用人工规则抽取实体候选

候选实体链接

□ 如何进行实体链接

- 基本方法：计算实体指称项和候选实体的相似度，选择相似度最大的候选实体

□ 单一实体链接

- BOW模型 (Honnibal TAC 2009, Bikel TAC 2009)
- 加入候选实体的类别特征 (Bunescu et al., EACL 2006)
- 加入候选实体的流行度等特征 (Han et al., ACL 2011)

□ 协同实体链接

- 利用实体之间类别的共现特征 (Cucerzan, EMNLP 2007)
- 利用实体之间连接关系 (Kulkarni et al., KDD 2009)
- 利用同一篇文档中不同实体之间存在着语义关联的特征 (Han et al., SIGIR 2011)

单一实体链接方法：词袋模型

(Honnibal TAC 2009, Bikel TAC 2009)

- 基于词袋子模型计算相似度
 - 将实体指称项上下文文本与候选实体上下文文本表示成词袋子向量形式，通过计算向量间的夹角确定指称项与候选实体相似度，系统选择相似度最大的候选实体进行链接

$$score(q, e_k) = \cos(q.T, e_k.T) = \frac{q.T}{\|q.T\|} \frac{e_k.T}{\|e_k.T\|}$$

$$\hat{e} = \arg \max_{e_k} score(q, e_k)$$

单一实体链接方法：类别特征

(Bunescu EACL 2006)

□ 动机

- 候选实体的文本内容可能太短，会导致相似度计算的不准确
- 加入指称项文本中的词与候选实体类别的共现特征
 - 例：除了计算待消歧文本和实体John Williams (composer)的 Wikipedia文本的相似度外，还考虑当前文本中的词语与Music, Art 等类别的共现信息

□ 方法

- 训练SVM分类器对候选实体进行选择
- 训练数据由 Wikipedia 中的超级链接获得
- 所采用的特征
 - 文本相似度
 - 指称项文本中词与候选实体类别的共现信息

John Williams (composer): Category={Music, Art...}
John Williams (wrestler): Category={Sport,...}
John Williams (VC): Category={Bank,...}

Williams has also composed
numerous classical concerti, and he
served as the principal conductor of
the Boston Pops Orchestra from
1980 to 1993

类别：music

单一实体链接方法：实体流行度等特征 (Han ACL 2011)

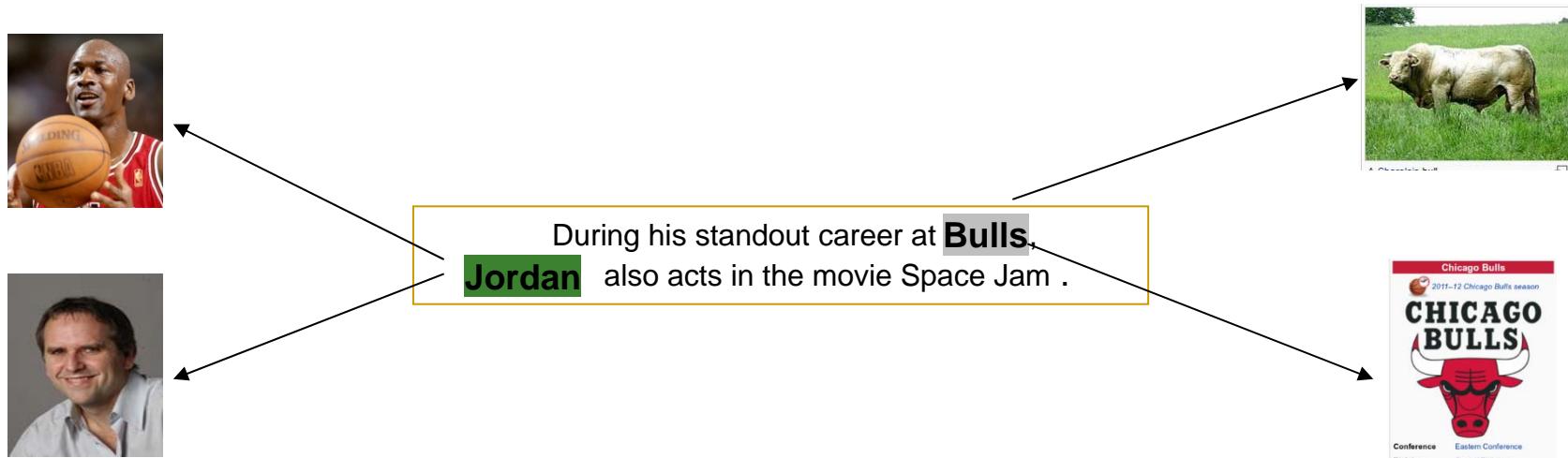
□ 动机

- 传统的方法仅仅是计算实体指称项与候选实体的相似度，忽略了候选实体的背景知识与先验信息，如实体本身的流行度、实体与指称项的关系等

□ 方法

- 考虑实体的背景知识，将实体的背景知识融入到实体链接的过程，实体的背景知识和先验信息主要有
 - 实体流行度：实体 e 在知识库中的概率 $P(e)$
 - 名称的知识：指称项 s 指向实体 e 的概率 $P(s|e)$
 - 上下文知识：实体 e 出现在特定上下文环境 c 的概率 $P(c|e)$

协同实体链接：基本思想



- 实体指称项与目标实体的语义相似度
- 目标实体之间的语义相似度

协同学习策略

□ 动机

- 同一篇文档中实体之间具有语义相关性
- 利用Pairwise优化策略

$$\frac{1}{\binom{|S_0|}{2}} \sum_{s \neq s' \in S_0} r(y_s, y_{s'}) + \frac{1}{|S_0|} \sum_{s \in S_0} w^\top f_s(y_s).$$

任意两个目标实体
之间的语义相关度

实体指称项到目标
实体的语义相似度

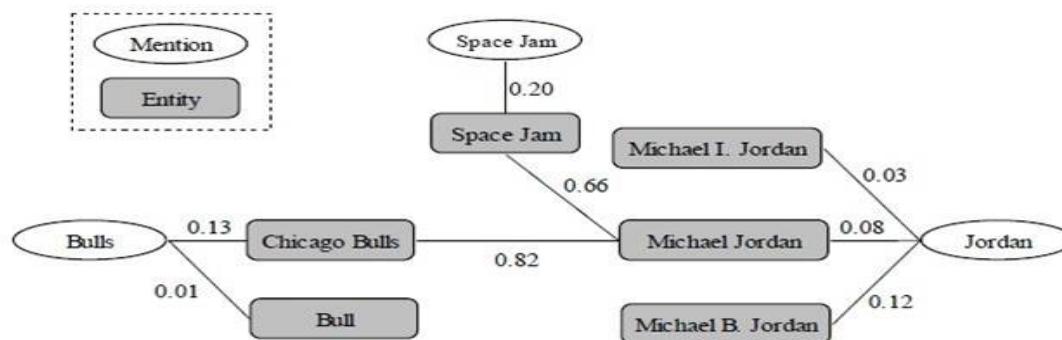
□ 语义相关度的计算方法：

- 利用实体类别重合度计算目标实体语义相似度 (Cucerzan, EMNLP 2007)
- 利用实体之间连接关系计算目标实体语义相似度 (Kulkarni , KDD 2009)

基于图的协同链接 (Han SIGIR 2011)

□ 动机

- Pairwise策略只考虑两两关系，结果不是全局最优的
- 采用**图方法**，**全局**考虑目标实体之间的语义关联
- **方法：** Referent Graph
 - **局部关系：**指称项与实体之间的关系，即该指称项文本与实体文本的相似度，由传统的VSM模型得到
 - **全局关系：**利用目标实体之间的链接关系计算实体之间的语义相关度



实体链接评测 (1/2)

- TAC-KBP (2009-Now): Entity Linking
 - 任务: 将文本中的目标实体链接到Wikipedia中的真实概念, 达到消歧的目的
 - 评测方法:

$$Accuracy_{micro} = \frac{NumCorrect}{NumQueries}$$

以指称项为单位计算的准确率

$$Accuracy_{macro} = \frac{\sum_i^{NumEntities} \frac{NumCorrect(E_i)}{NumQueries(E_i)}}{NumEntities}$$

以实体为单位计算的准确率

实体链接评测 (2/2)

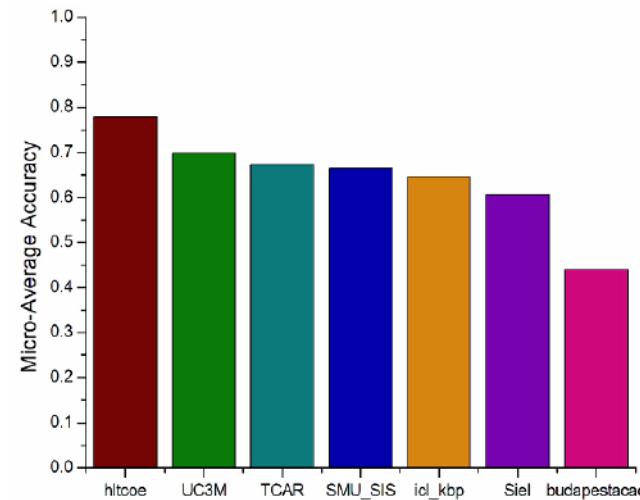
评测结果(Micro Accuracy)

Team	All	in KB	NIL
Siel_093	0.8217	0.7654	0.8641
QUANTA1	0.8033	0.7725	0.8264
hltcoe1	0.7984	0.7063	0.8677
Stanford_UBC2	0.7884	0.7588	0.8107
NLPR_KBP1	0.7672	0.6925	0.8232

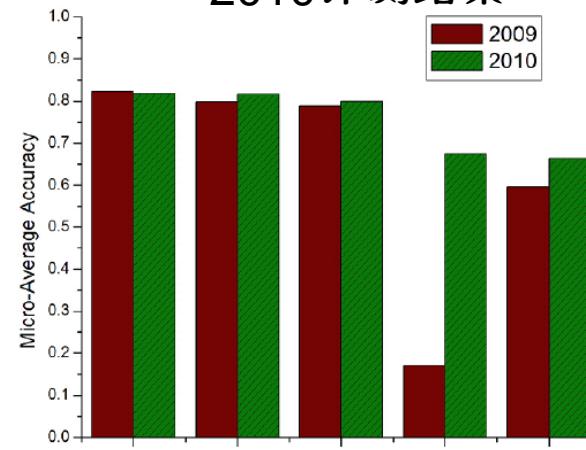
2009评测结果

	All	in-KB	NIL
All	0.8217 (3904)	0.7654 (1675)	0.8641 (2229)
PER	0.8309 (627)	0.8039 (255)	0.8495 (372)
ORG	0.8151 (2710)	0.7305 (1013)	0.8696 (1697)
GPE	0.8480 (567)	0.8280 (407)	0.8812 (160)

2009分实体类型结果



2010评测结果



2009 VS. 2010

小结

- 目前实体链接方法主要是如何更有效挖掘实体指称项信息，如何更准确地计算实体指称项和实体概念之间的相似度
- 由单一实体链接向协同实体链接发展
- 难点：未登录实体的处理

概述

- 引言
- 实体识别与抽取
- 实体消歧
- 关系抽取
 - **关系抽取任务定义**
 - 传统关系抽取
 - 开放域关系抽取
- 问题与挑战

关系抽取的定义

- Alexander Schutz等人认为关系抽取是自动识别由一对概念和联系这对概念的关系构成的相关三元组
 - Example1: 比尔盖茨是微软的CEO
 - CEO(比尔盖茨, 微软)
 - Example2: CMU坐落于匹兹堡
 - Located-in(CMU, 匹兹堡)
 - 高阶关系: Michael Jordan获得1997/98赛季的MVP
 - Award(Michael Jordan, 1997/98赛季, MVP)
- 关系抽取很重要
 - 属性抽取和事件抽取都可以归结为关系抽取

关系抽取分类

□ 传统关系抽取

□ 评测语料 (MUC, ACE, ...)

□ 专家标注语料

□ 开放域关系抽取

□ 数据源

□ 协作式知识库(Wikipedia, DBpedia, ...)

□ 半结构化，高质量

□ 网络语料

□ 海量，噪音大

□ 查询日志

□ 长度短，用户行为信息

概述

- 引言
- 实体识别与抽取
- 实体消歧
- 关系抽取
 - 关系抽取任务定义
 - 传统关系抽取
 - 开放域关系抽取
- 问题与挑战

传统关系抽取

□ 任务

- 给定实体关系类别，给定语料，抽取目标关系对
- 评测语料 (MUC, ACE, KBP)
 - 专家标注语料，语料质量高
 - 抽取的目标类别已经定义好
 - 有公认的评价方式

ACE抽取的目标关系列表

	PER	ORG	GPE	LOC	FAC	WEA	VEH
PER	Per_Social.Bus Per_Social.Family, Per_Social.Lasting, Gen_Aff.Ideology, Gen_Aff.CRRE	Org_Aff.Employment, Org_Aff.Ownership, Org_Aff.Student/Alum, Org_Aff.Sports_Affiliation, Org_Aff.Investor/Shareholder, Org_Aff.Membership, Org_Aff.Founder, Gen_Aff.CRRE	Physical.Located, Physical.Near, Org_Aff.Employment, Org_Aff.Investor/Shareholder, Org_Aff.Founder, Gen_Aff.CRRE	Physical.Located, Physical.Near, Gen_Aff.CRRE	Physical.Located Physical.Near, Agent/Artifact.UOIM	Agent/Artifact.UOIM	Agent/Artifact.UOIM
ORG		Part_Whole.Subsidiary, Org_Aff.Investor/Shareholder, Org_Aff.Membership	Part_Whole.Subsidiary, Org_Aff.Investor/Shareholder, Gen_Aff.Loc/Origin	Gen_Aff.Loc/Origin	Agent/Artifact.UOIM	Agent/Artifact.UOIM	Agent/Artifact.UOIM
GPE		Org_Aff.Investor/Shareholder,Org_Aff.Membership,	Physical.Near, Part_Whole.Geographical Org_Aff.Investor/Shareholder	Physical.Near, Part_Whole.Geo graphical	Agent/Artifact.UOIM	Agent/Artifact.UOIM	Agent/Artifact.UOIM
LOC			Physical.Near, Part_Whole.Geographical	Physical.Near, Part_Whole.Geo graphical	Physical.Near, Part_Whole.Geo graphical		
FAC			Physical.Near, Part_Whole.Geographical	Physical.Near, Part_Whole.Geo graphical	Physical.Near, Part_Whole.Geo graphical		
WEA						Part_Whole.Artifact	
VEH							Part_Whole.Artifact

KBP抽取的目标关系列表

Person	Organization	Geo-Political Entity
alternate names	alternate names	alternate names
age	political/religious affiliation	capital
birth: date, place	top members/employees	subsidiary orgs
death: date, place, cause	number of employees	top employees
national origin	members	political parties
residences	member of	established
spouse	subsidiaries	population
children	parents	currency
parents	founded by	
siblings	founded	
other family	dissolved	
schools attended	headquarters	
job title	shareholders	
employee-of	website	
member-of		
religion		
criminal charges		

(Paul McNamee, 2009)

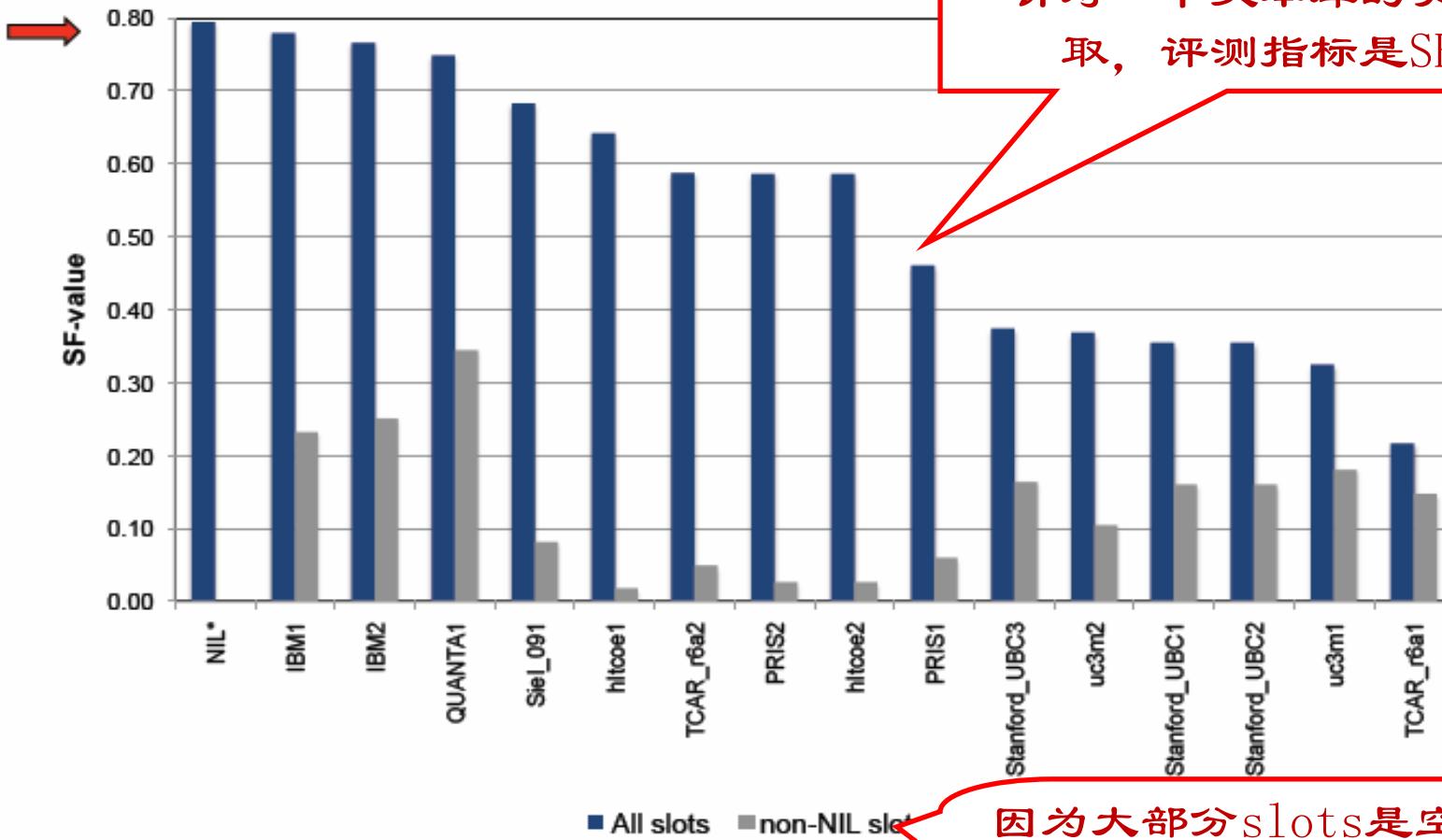
针对每个句子的关系识别，评测指标是F值

ACE(2005-2007)评测结果

语种	语料来源类型	ACE 2005				ACE 2007			
		EDR	RDR	TERN	VAL	VDR	EDR	RDR	TERN
Chinese	Broadcast News	70.5	24.4	81.8	48.7	11.2	49.7	16.8	
	Newswire	70.1	28.6	84.3	42.1	10.8	46.9	18.7	9.9
	Weblogs	67.1	26.6	86.2	71.2	2	35	15.7	24.8
	总分	69.2	26.8	83.7	49.7	10.2	45	17.6	14.8
English	Broadcast News	77.3	33.1	65.6	12.6	12.3	65.9	24.7	68.6
	Newswire	73.1	27.3	72.6	43.4	17.8	58.1	21.2	67.4
	Weblogs	71.6	20.5	62.8	38.5	15.4	52.7	18.2	54.8
	Broadcast Conversation	72.7	16	48	30	6.2	50.5	11	58.2
Arabic	Telephone	67.7	35.6	58.6	25	15.2	49.2	32.4	64.2
	Usenet	61.5	19.6	63.4	49.7	13.1	44	19.6	59
	总分	71.9	25.2	63.7	34.8	14.4	56.3	21.6	61.6
	Broadcast News						51.9		
Spanish	Newswire						49.4		
	Weblogs						42.1		
	总分						48.8		
	Newswire						51	46.5	
总分							46.5		

(http://www.nist.gov/speech/tests/ace/ace05/doc/acee05eval_official_results_20060110.htm
http://www.nist.gov/speech/tests/ace/ace05/doc/acee05eval_official_results_20070402.htm)

KBP 2009 Slot Filling Track评测结果



针对一个文本库的关系抽取，评测指标是SF值

(Paul McNamee, 2

因为大部分slots是空的，
所以non-NIL指标反映了系
统的性能

传统关系抽取方法 (1/4)

- 目前主要采用统计机器学习方法，将关系实例转换成高维空间中的特征向量或直接用离散结构来表示，在标注语料库上训练生成分类模型，然后再识别实体间关系
 - 基于特征向量方法：最大熵模型(Kambhatla 2004)和支持向量机(Zhao et al., 2005; Zhou et al., 2005; Jiang et al., 2007)等
 - 基于核函数的方法：浅层树核 (Zelenko et al., 2003)、依存树核 (Culotta et al., 2004)、最短依存树核 (Bunescu et al., 2005)、卷积树核 (Zhang et al., 2006; Zhou et al., 2007)

传统关系抽取方法 (2/4)

□ 基于特征向量方法：

- **主要问题**：如何获取各种有效的词法、句法、语义等特征，并把它们有效地集成起来，从而产生描述实体语义关系的各种局部特征和简单的全局特征
- **特征选取**：从自由文本及其句法结构中抽出取出各种表面特征以及结构化特征
 - 实体词汇及其上下文特征
 - 实体类型及其组合特征
 - 实体参照方式
 - 交叠特征
 - 基本短语块特征
 - 句法树特征

传统关系抽取方法 (3/4)

□ 基于核函数方法:

- **主要问题**: 如何有效挖掘反映语义关系的结构化信息及如何有效计算结构化信息之间的相似度
- **卷积树核**: 用两个句法树之间的公共子树的数目来衡量它们之间的相似度
 - 标准的卷积树核(CTK)
 - 在计算两棵子树的相似度时，只考虑子树本身，不考虑子树的上下文信息
 - 上下文相关卷积树核函数(CS-CTK)
 - 在计算子树相似度量时，同时考虑子树的祖先信息，如子树根结点的父结点、祖父结点信息，并对不同祖先的子树相似度加权平均

传统关系抽取方法 (4/4)

□ 基于特征向量方法 vs. 基于核函数方法

	基于特征向量方法	基于核函数方法
优点	简单实用 计算速度较快	能够有效挖掘结构化信息
缺点	难以进一步挖掘有效的平面特征 , 性能很难进一步提高	句法分析的错误引入了噪声, 同时由于树核的计算速度非常慢, 很难开发实用系统

小结

- 受限于训练语料规模
- 关系类别数限定，在实际应用中具有局限性
- 需要开放域关系抽取
 - 实体类型也更丰富、关系类型更多
 - 语料规模不受限

概述

- 引言
- 实体识别与抽取
- 实体消歧
- 关系抽取
 - 关系抽取任务定义
 - 传统关系抽取
 - **开放域关系抽取**
- 问题与挑战

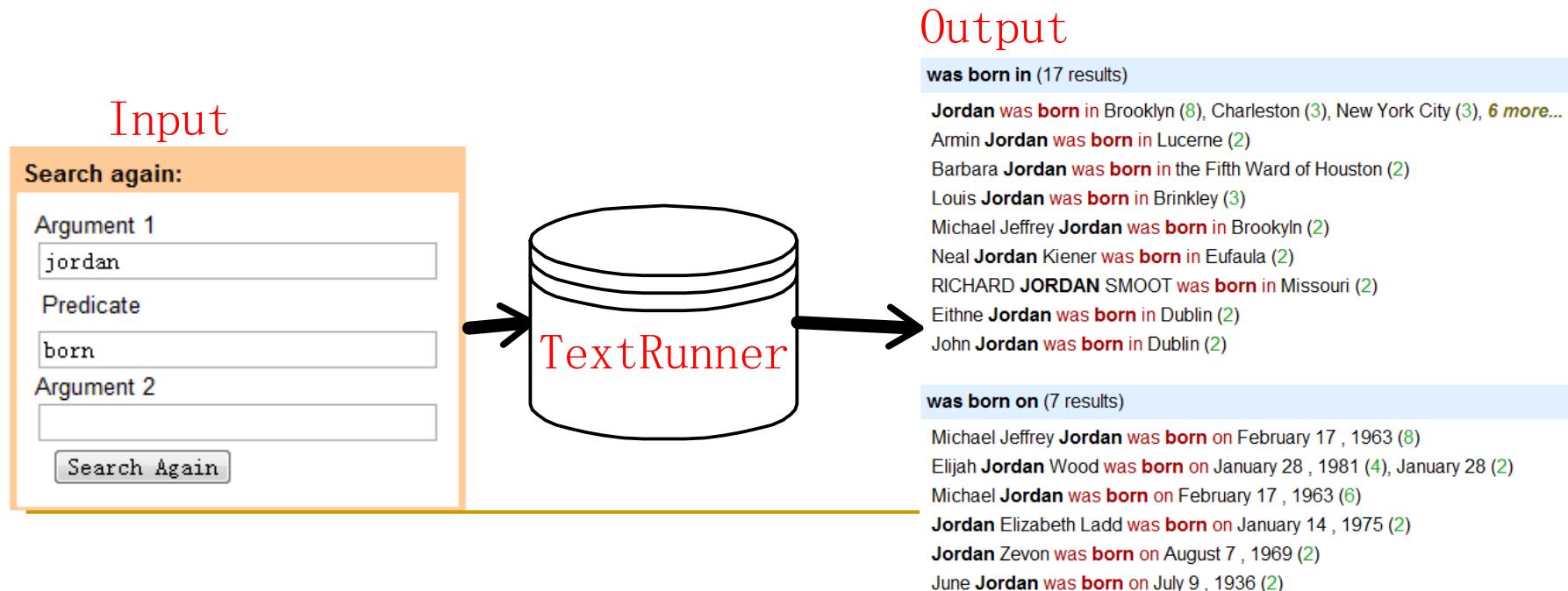
开放域关系抽取特点

- 不限定实体类别、不限定关系类别
- 不限定目标文本
 - Web Page
 - Wikipedia
 - Query Log
- 难点问题
 - 如何获取训练语料
 - 如何获取实体关系类别
 - 如何针对不同类型目标文本抽取关系

开放域关系抽取: Web Page (Banko IJCAI 2007) (1/2)

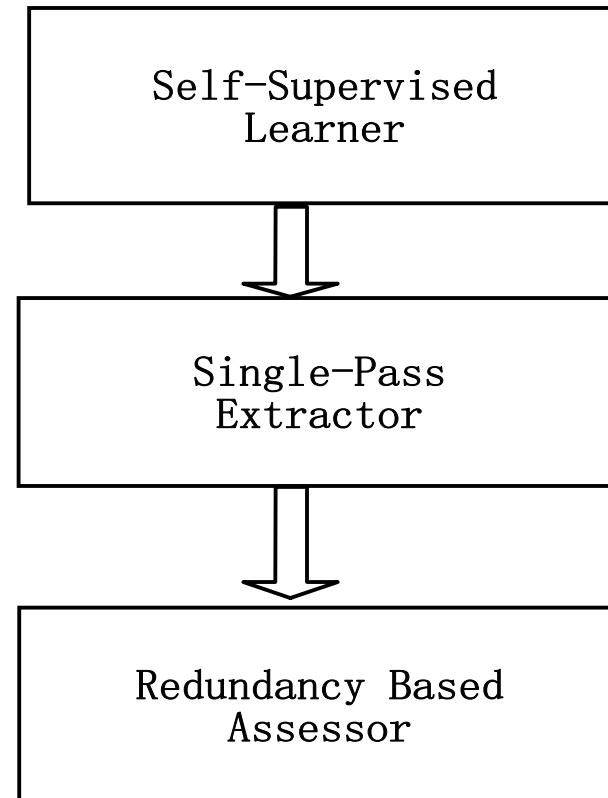
□ TextRunner: 利用宾州树库产生训练语料

- 利用宾州树库和shallow parser产生训练语料, 训练分类器用以对网络上的语料做浅层关系提取
- 利用搜索引擎返回用户输入论元相关的句子
- 用shallow parser和分类器抽取出句子中和输入相关的三元组



开放域关系抽取: Web Page (Banko IJCAI 2007) (2/2)

- 利用简单的启发式规则，在宾州树库上产生训练语料，提取一些浅层句法特征，训练分类器，用来判断一个元组是否构成关系
- 在网络语料上，找到候选句子，提取浅层句法特征，利用分类器，判断抽取的关系对是否“可信”
- 利用网络海量语料的冗余信息，对可信的关系对，进行评估



开放域关系抽取: Wikipedia (Wu CIKM 2007) (1/2)

□ 任务: 从Wikipedia文本中抽取关系(属性)信息

□ 难点

- 无法确定关系类别
- 无法获取训练语料

□ 方法

- **类别信息:** 在Infobox抽取关系类别信息
- **训练语料:** 在Wikipedia条目文本中进行回标, 产生训练语料



开放域关系抽取: Wikipedia (Wu CIKM 2007) (2/2)

Clearfield County, Pennsylvania	
Statistics	
Founded	March 26, 1804
Seat	Clearfield
Area	
- Total	2,988 km ² (1,154 mi ²)
- Land	sq mi (km ²)
- Water	17 km ² (6 mi ²), 0.56%
Population	
- (2000)	83,382
- Density	28/km ²

Clearfield County was created on 1804 from parts of Huntingdon and Lycoming Counties but was administered as part of Centre County until 1812.

Its county seat is Clearfield.

2,972 km² (1,147 mi²) of it is land and 17 km² (7 mi²) of it (0.56%) is water.

As of 2005, the population density was 28.2/km².

开放域关系抽取：从Query Log中抽取关系类别(Pasca WWW 2007)

- 利用Query Log中实体、属性和关系词的共现信息
 - 联想 笔记本 如何
 - 苹果 笔记本 如何
 - 戴尔 笔记本 售后 如何
 - 联想 笔记本 售后 服务点 上海 哪里有
 - 联想 CEO
 -
- Input：目标类别的代表性实体，例如{联想，苹果，戴尔}for 电脑厂商
- Output：排序后的目标类别的属性列表
 - {笔记本，售后，CEO，...}for 电脑厂商

小结

- 由识别到抽取
 - 规范文本→有噪音、有冗余的海量网络数据
- 限定类别→开放类别
- 难点：关系类别缺乏体系结构

概述

- 引言
- 实体识别与抽取
- 实体消歧
- 关系抽取
- 问题与挑战

问题与挑战 (1/2)

□ 封闭走向开放

- 训练语料限制传统信息抽取的发展
- 大规模网络信息不仅提供丰富的语料资源，同时提供了大量的抽取目标信息
- 抽取与识别相结合

□ 更鲁棒的自然语言理解技术

- 句法分析技术
- 实体识别技术

问题与挑战(2/2)

- 大规模信息抽取
 - 海量、分布式
- 深层次的挖掘信息背后的语义信息
 - 不仅仅是抽取
 - 更需要理解
 - 将信息与客观世界的真实目标相联系
- 结合知识库
 - 统计+约束

信息抽取评测

- MUC
 - ▣ http://www.itl.nist.gov/iaui/894.02/related_projects/muc/
- ACE
 - ▣ <http://www.itl.nist.gov/iad/894.01/tests/ace/>
- TAC
 - ▣ <http://www.nist.gov/tac/>
- CONLL
 - ▣ <http://www.cnts.ua.ac.be/conll/>
- Sighan
 - ▣ <http://www.sighan.org/>

信息抽取资源

- Unsupervised Information Extraction
 - ▣ SNOWBALL [Agichtein & Gravano ICDL00]
 - ▣ MULDER [Kwok et al. TOIS01]
 - ▣ AskMSR [Brill et al. EMNLP02]
- Ontology Driven Information Extraction
 - ▣ SemTag and Seeker [Dill WWW03]
 - ▣ PANKOW [Cimiano WWW05]
 - ▣ OntoSyphon [McDowell & Cafarella ISWC06]
- Other Wikipedia Systems
 - ▣ Yago [Suchanek et al. WWW07]
 - ▣ DBpedia [Auer & Lehmann ESWC07]
 - ▣ Wikipedia Reputation System [Adler & Alfaro WWW07]

参考文献

- Bagga, A. & Baldwin, B. 1998. Entity-based cross-document coreferencing using the vector space model. Proceedings of the 17th international conference on Computational linguistics-Volume 1, pp. 79-85.
- M. Banko, M. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the web. In IJCAI, 2007.
- Bekkerman, R. & McCallum, A. 2005. Disambiguating web appearances of people in a social network. Proceedings of the 14th international conference on World Wide Web, pp. 463-470.
- D.Bikel et al. Entity Linking and Slot Filling through Statistical Processing and Inference Rules. In Proceeding of TAC. 2009.
- R.Bunescu and M.Pasca. Using Encyclopedic Knowledge for Named Entity Disambiguation. In Proceeding of EACL. 2006.
- S.Cucerzan. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In Proceeding of EMNLP. 2007.
- GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. Exploring various knowledge in relation extraction. In Proceedings of ACL.

参考文献

- S.Kulkarni et al. Collective Annotation of Wikipedia Entities in Web Text. In Proceeding of KDD. 2009.
- Han, X. & Zhao, J. 2009. Named entity disambiguation by leveraging Wikipedia semantic knowledge. Proceeding of the 18th ACM conference on Information and knowledge management, pp. 215-224.
- Han, X. & Zhao, J. 2010. Structural Semantic Relatedness: A Knowledge-Based Method to Named Entity Disambiguation. Proceeding of ACL, pp. 50-59.
- XP.Han and L.Sun. A Generative Entity-Mention Model for Linking Entities with Knowledge Base. In Proceeding of ACL. 2011.
- XP.Han et al. Collective Entity Linking in Web Text: A Graph-Based Method. In Proceeding of SIGIR. 2011.
- M.Honnibal and R.Dale. DAMSEL: The DSTO/Macquarie System for Entity-Linking. In Proceeding of TAC. 2009.
- Gina-Anne Levow. The Third International Chinese Language Processing Backoff: Word Segmentation and Name Entity Recognition [C]. Proceedings of the Fifth SigHAN Workshop on Chinese Language Processing, Sydney: Association for Computational Linguistics, 2006: 108-117.

参考文献

- Medelyan, O., Witten, I.H. and Milne, D. (2008) Topic Indexing with Wikipedia. In Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence (WIKIAI 2008), Chicago, IL.
- Mihalcea, R. and Csomai, A. (2007) Wikify!: linking documents to encyclopedic knowledge. In Proceedings of the 16th ACM Conference on Information and Knowledge management (CIKM'07), Lisbon, Portugal, pp. 233-242.
- Milne, D. and Witten, I. (2008) Learning to link with Wikipedia. In Proceedings of the 16th ACM Conference on Information and Knowledge management (CIKM'08), Napa Valley, California, USA pp519-529.
- Pedersen, T., Purandare, A. & Kulkarni, A. 2005. Name discrimination by clustering similar contexts. Computational Linguistics and Intelligent Text Processing, pp. 226-237.
- Youzheng Wu, Jun Zhao, Xu Bo, Chinese Named Entity Recognition Model Based on Multiple Features. In: Proceedings of the Joint Conference of Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), Vancouver, 2005:427~434
- F. Wu and D. Weld. Autonomously semantifying Wikipedia. In CIKM, 2007.
- F. Wu and D. Weld. Open information extraction using Wikipedia. In ACL, 2010.

参考文献

- ZHAO Jun, LIU Feifan, Product Named Entity Recognition in Chinese Texts, International Journal of Language Resource and Evaluation (LRE), Vol.42 No.2 132-152, 2008 (SCI).
- W.Zhang et al. Entity Linking with Effective Acronym Expansion, Instance Selection and Topic Modeling. In Proceeding of IJCAI. 2011
- Min Zhang, Jie Zhang, and Jian Su. 2006a. Exploring syntactic features for relation extraction using a convolution tree kernel. In Proceedings of HLT/NAACL
- NIST 2005. Automatic Content Extraction Evaluation Official Results[2007-09-28].http://www.nist.gov/speech/tests/ace/ace05/doc/acee05eval_official_results_20060110.htm
- NIST 2007. Automatic Content Extraction Evaluation Official Results[2007-09-28].http://www.nist.gov/speech/tests/ace/ace05/doc/acee05eval_official_results_20070402.htm
- Paul McNamee, Overview of the TAC 2009 Knowledge Base Population Track, In Proceedings of TAC workshop, 2009.
- 863计划中文信息处理与智能人机接口技术评测组. 2004年度863计划中文信息处理与智能人机交互技术评测: 命名实体评测结果报告[R]. 北京: 863计划中文信息处理与智能人机接口技术评测组, 2004
- 吴友政. 问答系统关键技术研究. 中国科学院自动化研究所博士论文. 2006.

Q&A

Thanks

观点挖掘与倾向性分析

中国科学院自动化研究所
模式识别国家重点实验室

目录

□ 第一部分：

- 我们为什么需要观点挖掘与倾向性分析？
- 什么是观点挖掘与倾向性分析？

□ 第二部分：

- 如何进行观点挖掘与倾向性分析？
 - 任务、方法、资源、评测

□ 第三部分：

- 问题与挑战

为什么需要

- 文本信息主要包含两类
 - 客观性事实(Facts)
 - 主观性观点(Opinions)
- 随着Web2.0的飞速发展以及Web3.0的兴趣，互联网中出现大量的UGC数据，其中包含了大量观点信息
 - 博客、微博、商品评论、论坛....
- 44%新闻文本包含观点信息 (Wiebe ACL 2001)
- 已有文本分析方法主要侧重于客观性文本内容(factual information)的分析和挖掘



有什么用

□ 企业对观点挖掘和倾向性分析的需求

- Automatically find consumer sentiments and opinions (market intelligence)
- Capture public trends
- Capture commercial opportunity
- Online reputation management
- Precision Advertising



□ 普通用户对观点挖掘和倾向性分析的需求

- Helpful for purchasing a product
- Find opinions on political topics



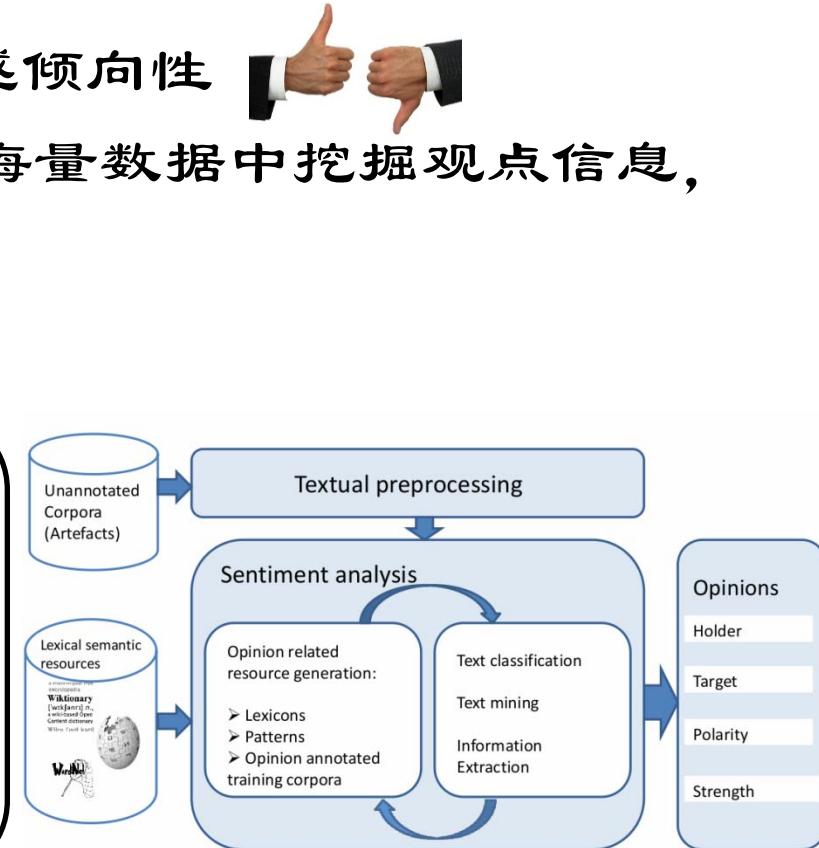
□ 政府对观点挖掘和倾向性分析的需求

- Control the public opinions
- Monitor the public event

定义

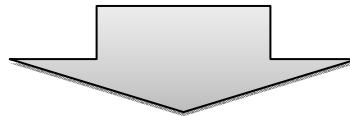
- **观点**: 人们对事物的看法，具有明显的主观性，不同人对同一事物的看法存在差异
- **倾向性**: 观点中所包含的情感倾向性 
- **观点挖掘与倾向性分析**: 从海量数据中挖掘观点信息，并分析观点信息的倾向性
- 非结构化→结构化

Sentiment analysis or opinion mining (in Wikipedia) refers to a broad area of natural language processing, computational linguistics and text mining. Generally speaking, it aims to determine the attitude of a speaker or a writer with respect to some topic.



例子

“我今年天让入手诺基亚5800，把玩不到24小时，**目前感觉**5800**屏幕很好，操作也很方便，通话质量也不错，但是外形有些偏女性化，不适合男生。这些都是小问题，最主要的问题是电池不耐用，只能坚持一天，反正我觉得对不起这个价格。”**



- 外形
- 电池



- 屏幕
- 操作
- 通话质量



观点挖掘与倾向性分析相关任务

□ 观点及倾向性识别

- Sentiment Identification

□ 观点信息抽取

- Opinion Attribute Extraction
- Opinion Summarization

□ 观点检索

- Opinion Retrieval
-

观点及倾向性识别 (1/2)

□ Opinion Identification (Subjective/Objective)

- 中美两方的代表就朝鲜核问题进行了磋商。(Objective)
- 中方发言人就美国近期对阿富汗的行动进行了强烈的谴责 (Subjective)

□ Polarity Classification (Positive/Negative/Neutral)

- 这家餐厅总体来说还可以。(Neutral)
- 但是价格偏贵，人均消费100块。(Negative)
- 抛开价格的因素还是很不错的。(Positive)

□ Strength Rating (Sentiment Strength Identification)

- iPhone 的价格太贵了。 (Strong against)
- iPhone 的价格有点贵。 (Something to be bad)



观点及倾向性识别 (2/2)

- Word Level
 - 识别一个词的倾向性
- Feature Level (Aspect Level)
 - 识别一个Aspect的倾向性
 - “这家餐厅**价格**偏贵，人均消费100块”→ **价格**
- Sentence Level
 - 识别一个句子的观点倾向性
- Document Level
 - 识别一篇文本（包含多个句子）整体的倾向性

观点信息抽取 (1/2)

□ Opinion Holder Extraction

- “中方发言人”就美国近期对阿富汗的行动进行了强烈的谴责”
- 在新闻语料中大量出现，通常为命名实体、名词性短语或者术语
- 在商品评论文本中很少出现

□ Opinion Target Extraction

- “中方发言人就美国近期对阿富汗的行动进行了强烈的谴责”
- “这款手机的屏幕太小，分辨率不足”
- 术语、事件、实体等

观点信息抽取 (2/2)

“I bought an iPhone a few days ago. It was such a nice phone. The touch screen was really cool. The voice quality was clear too. Although the battery life was not long, that is ok for me. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, and wanted me to return it to the shop. ...”

....

Opinion Summary:

Feature 1: Touch screen

Positive: 212

- *The touch screen was really cool.*
- *The touch screen was so easy to use and can do amazing things.*

...

Negative: 6

- The screen is easily scratched.
- I have a lot of difficulty in removing finger marks from the touch screen.

...

Feature 2: battery life

...

观点检索

- 根据用户的查询从文档中找出对于主题信息发表了观点的文档
 - 主题相关并且具有主观倾向性
 - Blog Search, Twitter, Forum.....
 - “奥巴马这人怎么样？”
 - “国际社会对于卡扎菲的态度如何？”
 - “iphone4如何？”

应用

新浪首页 转帖 - 开心网 谷歌iphone - Google 谷歌iphone - Google Apple iPhone - Google http://www.google.com/... 谷歌iphone - 必应 中国科学院自动化研究所 liuk.ia.ac@gmail.com | My Shopping List | Settings ▾ | Sign out

Web Images Videos Maps News Shopping Gmail more ▾

liuk.ia.ac@gmail.com | My Shopping List | Settings ▾ | Sign out

Google products

iphone

Search Products

Apple iPhone 8 GB (first generation) (AT&T - GSM)

from Apple in Mobile Phones

Overview · Compare prices · Reviews · Technical specifications · Similar items · Accessories

 \$500 new, \$180 used from 4 sellers

★★★★★ 1,784 reviews

Reviews

Summary - Based on 1,784 reviews

1 2 3 4 stars 5 stars

"Easy texting and buttons are good size to heat." "Awful speaker though."

"The photos have good clarity and is very handy." "With its touch screen it makes life so much easier."

"You can download music, videos, and movies (love)" "The worst thing about the iPhone is the price."

Apple iPhone - 8GB (AT&T) Review

★★★★★ By Kent German - Jun 30, 2007 - Editorial review - CNET

Pros: The Apple iPhone has a stunning display, a sleek design, and an innovative multitouch user interface. Its Safari browser makes for a superb Web surfing experience, and it offers easy-to-use apps. As an iPod, it shines.

Cons: The Apple iPhone has variable call quality and lacks some basic features found in many cell phones, including stereo Bluetooth support and 3G compatibility. Integrated memory is stingy for an iPod, and you have to sync the iPhone to manage music content.

Bottom Line: Despite some important missing features, a slow data network, and call quality that doesn't always deliver, the Apple iPhone sets a new benchmark for an integrated cell phone and MP3 player.

Editor's note (7/11/2008): This is the review of the original first-generation iPhone model, released June 2007. Coverage of the 3G iPhone model released July 11, 2008 is available [here](#).

Photo gallery: Apple iPhone

Editor's note: Apple eliminated the 4GB model on September 5, 2007, two months after the iPhone's initial

Show reviews that mention

Features (946)
Design (597)
Camera (486)
Screen (396)
Battery (366)
Music (347)
Video (288)

Show reviews by source

Editorial reviews (5)
User reviews (1,779)

Amazon.com (69)
CNET (2)
DigitalTrends.com (1)
Epinions (103)
PriceGrabber.com (15)
Viewpoints (1592)
Wired (2)

Sort reviews

Windows 7 taskbar: 开始, 收件箱, Apple iPhone, 新-搭隔, v_liukan..., Microsoft Word, Windows, 我的文档, 周报模板, 10:12

应用

ELECTRONICS > CELL PHONES & PLANS



Apple iPhone 3GS 16GB - smartphone - WCDMA (UMTS) / GSM

\$449 and up (6 stores)

★★★★★ User reviews (285)

★★★★★ Expert reviews (1)

SHARE [Facebook](#) [Twitter](#) [Messenger](#) [Email](#)

[See larger photo](#)

See also: [Product Summary](#) · [Where to Buy](#) · **User Reviews** · [Expert Reviews](#) · [Specifications](#)

USER REVIEWS



SCORECARD: EASE OF USE (See all)

87 positive reviews | no negative comments

Email Emailing is very basic, easy to use and works brilliantly.

★★★★★ KylieLandymore · 10/6/2010 · [www.ciao.co.uk](#) [see all](#)

Pros: good maps, memory, quick , easy to use , access to loads of different features

★★★★★ anklechris · 9/30/2010 · [www.ciao.co.uk](#) [see all](#)

Pros: Has everything, brilliant music player and camera, great for games, easy to use .

★★★★★ doodles12 · 5/24/2010 · [www.ciao.co.uk](#) [see all](#)

Safari: This is the Iphone Web browser, its very easy to use, very quick and you cant go wrong with it, it has a "Bookmarks" Section so can store all your favorites websites in specific folders, to can navigate forwards to backwards on the web...

★★★★★ alanh087 · 8/28/2010 · [www.ciao.co.uk](#) [see all](#)

The phone is very easy to use, everything is very intuitive.

★★★★★ Melski1979 · 9/20/2010 · [www.ciao.co.uk](#) [see all](#)



应用



twitrratr

SEARCH

SEARCHED TERM

iphone

POSITIVE TWEETS

2775

NEUTRAL TWEETS

19720

NEGATIVE TWEETS

846

TOTAL TWEETS

23341

11.89% POSITIVE

@schwa now there's a blast from the past. but it occurs to me that gliderpro would make a great iphone app. ([view](#))

alas fair iphone, you served me well and will be missed. ([view](#))

@mikediliberto @downtownrob @mitchwagner funny that i ended up following smoke signals as

84.49% NEUTRAL

view from the iPhone: <http://www.floodgap.com/iv/197> ([view](#))

That's "Memphis" Taproom. Goddamn iPhone. ([view](#))

@mothermusings This is the iPhone thingie, huh? Sooooo sorry! ([view](#))

3.62% NEGATIVE

@mikef1182 as bad as exchange on the iphone? ([view](#))

<http://twitpic.com/i0se> - iphone typing auto-correct changes 'just sayin' to 'just satin' - wrong msg indeed! ([view](#))

iphone applications don't whine about being left outside or going hungry or manual labor or using

应用

www.tweetfeel.com/?x=22&y=30#Kate_Middleton

Tweetfeel Biz | FAQ | Contact Us | Biz Login

tweetfeel



Kate Middleton

Searching and Analyzing

Try some Twitter trends: [Micky Christmas](#) [Eid Mubarak](#) [Kate Middleton](#) [Prince William](#) [Sharpe Playing God](#) [SARESP](#)

22 : 3 = 88%

Style watch: **Kate Middleton** rocks the royal scene: Speaking of giant rocks...are they or are they not? Rumors ar...
<http://bit.ly/citOrP>

Blech. I don't like **Kate Middleton**. #notjealousatall #coughcough

Kate Middleton Rocks.

@ambergunn "I love **Kate Middleton**'s hair. I want mine cut like that." @audreyrella "I bet no one said that about Diana's hair." #badhairday

Prince William & **Kate Middleton** ftw !!! ^_^

Read our FAQ | Subscribe to our API | Legal Stuff | 100% Guarantee |  Share | Follow us | Email us | Brought to you by **conversion** | Powered by **twitter**

目录

□ 第一部分：

- 我们为什么需要观点挖掘与倾向性分析？
- 什么是观点挖掘与倾向性分析？

□ 第二部分：

- 如何进行观点挖掘与倾向性分析？
 - 任务、方法、资源、评测

□ 第三部分：

- 存在的问题以及面临的挑战

内容

- Sentiment Identification
- Opinion Mining
- Opinion Retrieval
- Resources and Evaluations

Sentiment Identification

- Word Level
- Sentence Level
- Document Level
- Others

Word Level Sentiment Identification

□ 任务：

- 识别词语的情感倾向性，构建词典资源

□ 方法：

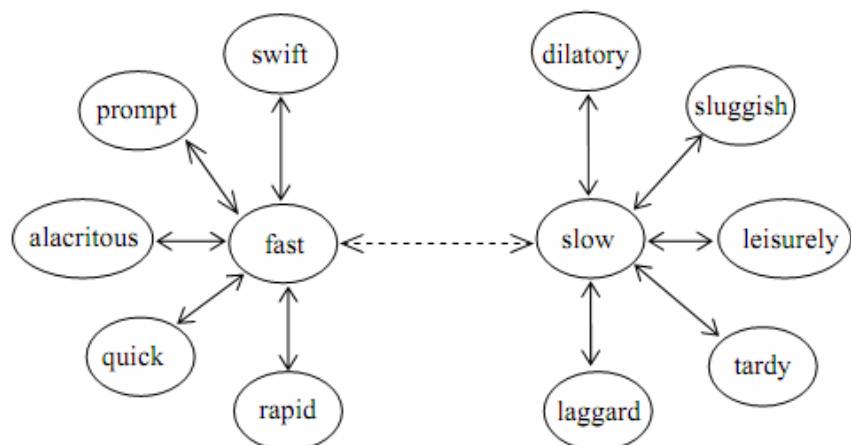
- 基本思路：利用词语之间的相似度进行词典扩展
- Dictionary-based approaches
- Corpus-based approaches

Dictionary-based Approaches (1/2)

□ Hu (KDD 2004)

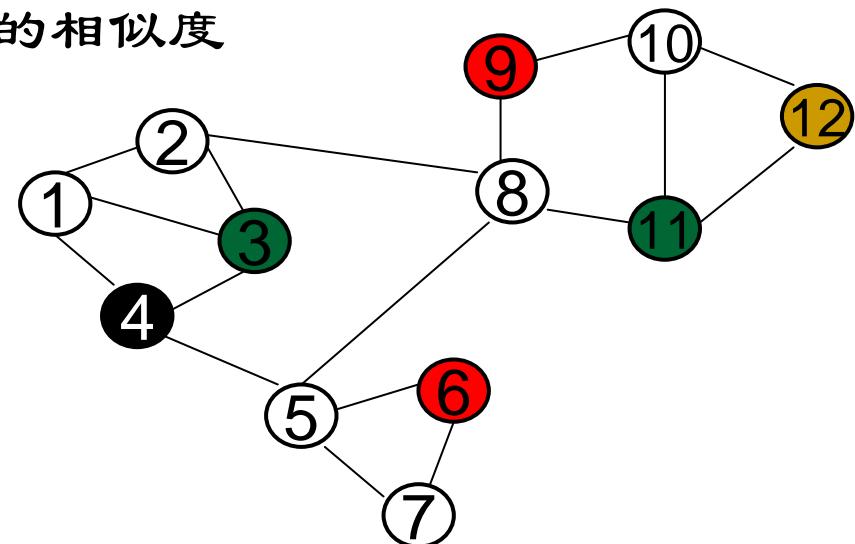
□ 利用词与词之间在WordNet中的同义、反义关系
对于情感词典进行扩展

```
1. Procedure OrientationSearch(adjective_list, seed_list)
2. begin
3.   for each adjective  $w_i$  in adjective_list
4.   begin
5.     if ( $w_i$  has synonym  $s$  in seed_list)
6.       {  $w_i$ 's orientation =  $s$ 's orientation;
7.         add  $w_i$  with orientation to seed_list; }
8.     else if ( $w_i$  has antonym  $a$  in seed_list)
9.       {  $w_i$ 's orientation = opposite orientation of  $a$ 's
          orientation;
10.      add  $w_i$  with orientation to seed_list; }
11.   endfor;
12. end
```



Dictionary-based Approaches (2/2)

- Hassan (ACL 2010), Kamps (LREC 2004)
 - 利用WordNet计算词之间的相似度，识别词的情感倾向性
 - 根据WordNet，计算词之间的相似度，建立词之间的语义图，边上的权重表示词之间的相似度
 - 利用图算法识别词的倾向性
 - Random Walk (Hassan)
 - Shortest Distance (Kamps)



Corpus-based Approaches (1/2)

□ Turney (ACL 2002)

- 利用网络资源计算两个词之间的相关度（互信息）
- 利用相关度识别词语的情感倾向性
- 使用‘Near’算子 (AltaVista)

$$\text{PMI}(word_1, word_2) = \log_2 \left(\frac{\frac{1}{N} \text{hits}(word_1 \text{ NEAR } word_2)}{\frac{1}{N} \text{hits}(word_1) \frac{1}{N} \text{hits}(word_2)} \right)$$

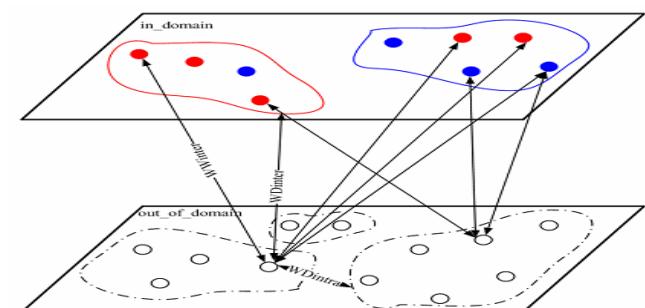
$$\begin{aligned} \text{SO}(phrase) &= \text{PMI}(phrase, "excellent") \\ &\quad - \text{PMI}(phrase, "poor") \end{aligned}$$

Corpus-based Approaches (2/2)

□ 建立领域情感词典 (Du WSDM 2010)

- 不同领域具有不同的领域情感词
- 缺乏目标领域训练语料，利用其他领域的标注语料，领域迁移的问题
- 不仅仅考虑词与词之间的关系
 - Word-Doc relation, Word-Doco relation
- 利用Information bottleneck method (co-clustering)
 - 对于文档、词同时进行聚类

$$\begin{aligned} & I(D_o; \hat{W}_o) - I(\hat{D}_o; \hat{W}_o) \\ & + \alpha \cdot \left[\left(I(D_i; \hat{W}_o) - I(\hat{D}_i; \hat{W}_o) \right) + \left(I(W_i; \hat{W}_o) - I(\hat{W}_i; \hat{W}_o) \right) \right] \end{aligned}$$



小结

- 基本思路：利用词之间的相似度对于情感词典进行扩展
(Dictionary-based, Corpus-based)
- Pros:
 - 模型直观，易于计算
- Cons:
 - 利用词典或者大规模语料方法计算词之间相似性易产生噪音
 - 部分词语的倾向性与上下文相关，与主题相关
 - 屏幕大
 - 体积太大
 - 大部分方法只计算了形容词的倾向性，忽略了动词、形容词、名词以及网络用语的情感倾向性
 - 小瘪三！
 - 做人不能CCTV

Sentiment Identification

- Word Level
- Sentence Level
- Document Level
- Others

Sentence Level Sentiment Identification

- 任务：识别句子的情感倾向性
 - “7.23动车追尾事故给铁道部一记响亮的耳光。”
- 关键问题
 - 如何进行特征表示
- 分类：
 - Corpus-based approaches (监督学习)
 - Lexicon-based approaches (非监督学习)
 - Combined approaches



与传统文本分类的区别

- Topic-based text categorization
 - 侧重于主题词特征
 - “这款手机的屏幕太大了”（科技、手机）
- Sentiment classification
 - 表示倾向性的词语更加重要。
 - “这款手机的屏幕好大了”（主观、褒义）

Corpus-based Approaches: 特征选择 (1/2)

□ 利用传统文本分类方法处理情感分类任务 (Pang EMNLP 2002)

□ 比较多种特征的效果

□ Unigram、bigram、POS、Adj.、Position

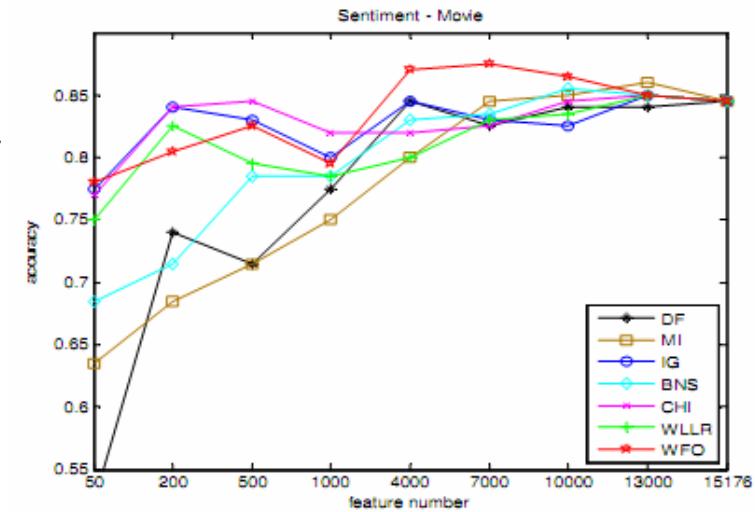
□ 比较多个分类器性能

□ SVM、Naïve Bayes、Maximum Entropy

	Features	# of features	frequency or presence?	NB	ME	SVM
(1)	unigrams	16165	freq.	78.7	N/A	72.8
(2)	unigrams	"	pres.	81.0	80.4	82.9
(3)	unigrams+bigrams	32330	pres.	80.6	80.8	82.7
(4)	bigrams	16165	pres.	77.3	77.4	77.1
(5)	unigrams+POS	16695	pres.	81.5	80.4	81.9
(6)	adjectives	2633	pres.	77.0	77.7	75.1
(7)	top 2633 unigrams	2633	pres.	80.3	81.0	81.4
(8)	unigrams+position	22430	pres.	81.0	80.1	81.6

Corpus-based Approaches : 特征选择 (2/2)

- 比较各种特征选择方法在情感分类中的作用 (Li ACL 2009)
- DF、MI、IG、CHI、BNS、WLLR、WFO
- 学习触发观点的模板(Riloff ACL 2003)



seems to be <dob>
underlined <dob>
pretext of <np>
atmosphere of <np>
<subj> reflect
to satisfy <dob>
way with <np>
bring about <np>
expense of <np>
voiced <dob>
turn into <np>

I am pleased that there now **seems to be** broad political consensus ...
Jiang's subdued tone . . . **underlined** his desire to avoid disputes ...
On the **pretext of** the US opposition . . .
Terrorism thrives in an **atmosphere of** hate . . .
These are fine words, but they do not **reflect** the reality . . .
The pictures resemble an attempt **to satisfy** a primitive thirst for revenge . . .
... to ever let China use force to have its **way with** . . .
"Everything must be done by everyone to **bring about** de-escalation," Mr Chirac added.
at the **expense of** the world's security and stability
Khatami . . . **voiced** Iran's displeasure.
... the surging epidemic could **turn into** "a national security threat," he said.

Corpus-based Approaches：上下文影响 (1/2)

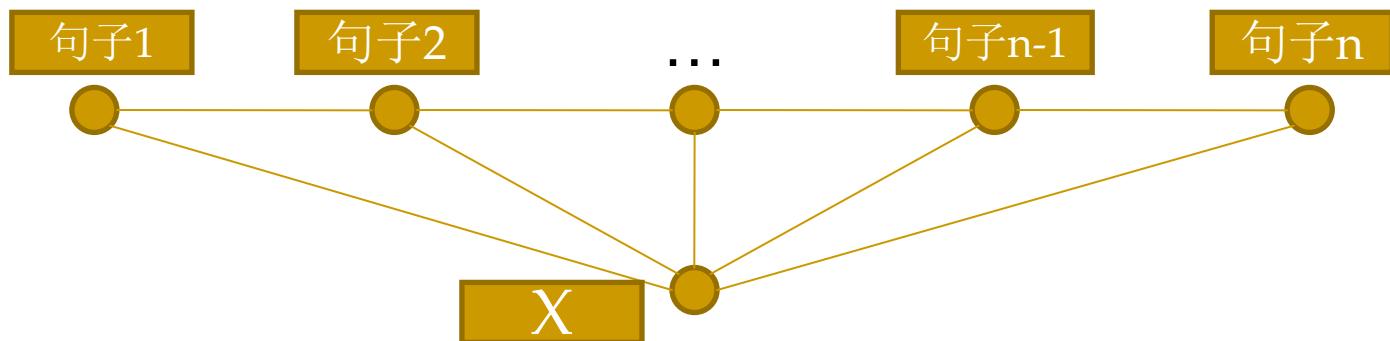
□ 上下文的影响

“1) 这是一个挺不错的电影院。2) 因为优惠很多，来的人还是比较多的，于是带起了时代广场地下一层的餐饮。3) 虽然硬件条件虽说赶不上星美，但也服务是不错的了。4) 同时看电影院周围有商场，电影开演之前可以逛逛商场。5) 总之，这里已经成为我和老公的定点影院了。”

- 句子的倾向性与句子所在上下文密切相关
- 分类任务->序列标注任务

Corpus-based Approaches : 上下文影响 (2/2)

- Yi (ICML 2006) and Zhao (EMNLP 2008)
 - 将篇章中每个句子看作是一个序列上的点
 - 利用CRFs进行学习和标注



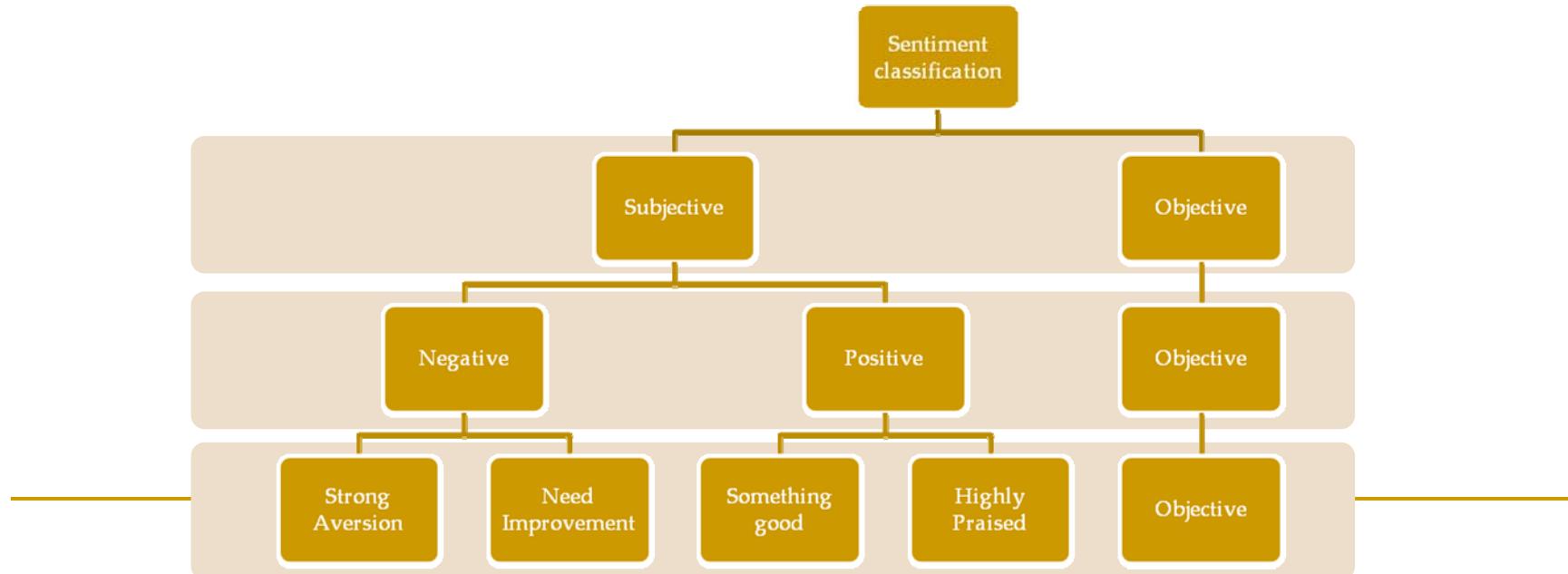
Corpus-based Approaches: 上下文+标记间冗余关系

□ Zhao (EMNLP 2008)

□ 情感倾向性标记之间具有冗余关系

□ 多任务联合处理

□ 主客观分类、褒贬分类、强度分类



Label	NB	SVM	MaxEnt	Standard CRF	Cascaded CRF	Our Method
PP	0.1745	0.2219	0.2055	0.2027	0.2575	0.2167
P	0.2049	0.2877	0.2353	0.2536	0.2881	0.3784
Neu	0.8083	0.8685	0.8161	0.8273	0.8554	0.8269
N	0.2636	0.3014	0.2558	0.2981	0.3092	0.4204
NN	0.0976	0.1162	0.1148	0.1379	0.1510	0.2967
Total	0.6442	0.6786	0.6652	0.6856	0.7153	0.7521

Strength Rating

Label	NB	SVM	MaxEnt	Standard CRF	Cascaded-CRF	Our Method
Pos	0.4218	0.4743	0.4599	0.4405	0.5122	0.6008
Neu	0.8147	0.8375	0.8424	0.8260	0.8545	0.8269
Neg	0.3217	0.3632	0.2739	0.3991	0.4067	0.5481
Total	0.7054	0.7322	0.7318	0.7327	0.7694	0.7855

Polarity Classification

Label	NB	SVM	MaxEnt	Standard CRF	Our Method
Subjective	0.4743	0.5847	0.4872	0.5594	0.6764
Objective	0.8170	0.8248	0.8212	0.8312	0.8269
Total	0.7238	0.7536	0.7518	0.7561	0.8018

Subjective Identification

Corpus-based Approaches: Polarity Shift (1/2)

- Polarity Shift
 - 多样语言现象造成的句子内部词的倾向性转移
 - “整个店面的装修 **不是很**漂亮”
 - 在这种情况下，如何减少学习错误？
 - 方法
 - 在句子中检测出Polarity Shift
 - 判别句子倾向性时对于Polarity Shift专门处理

Corpus-based Approaches: Polarity Shift (2/2)

- Polarity Shift的检测

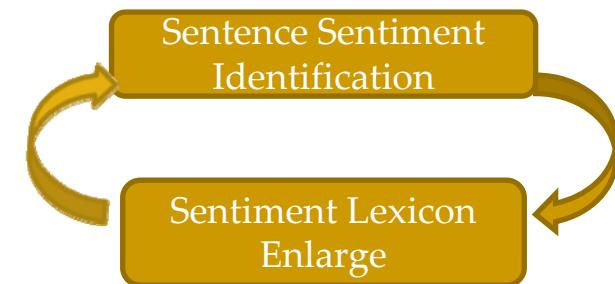
- 利用上下文信息



- 词典信息 (Ikeka IJCNLP 2008)
 - 特征选择 (Li COLING 2010)

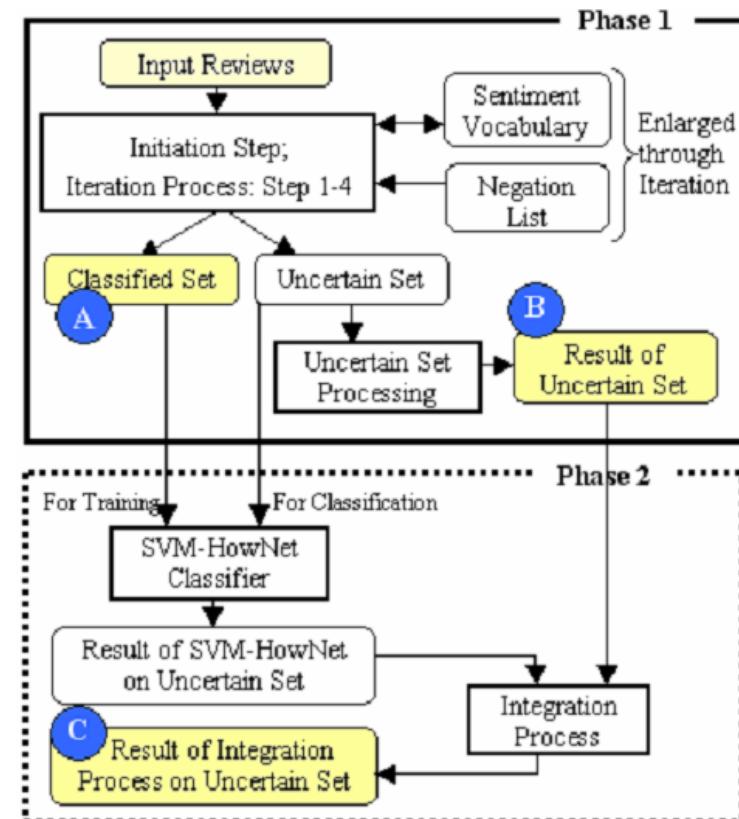
Lexicon-based Approach (1/2)

- 利用句子中词的倾向性来确定句子的倾向性
 - 关键问题：词的倾向性识别
- Turney (ACL 2002)
 - Step1: POS and select sentiment phase by patterns
 - Step2: Use PMI to compute the phase sentiments
 - Step3: Compute average sentiment of all phases in a sentence
 - Car: 84%, Banks 80%, Movies 65.83%, Travel 70.53%
- Taras (COLING 2008)
 - 句子、词的情感倾向性联合识别



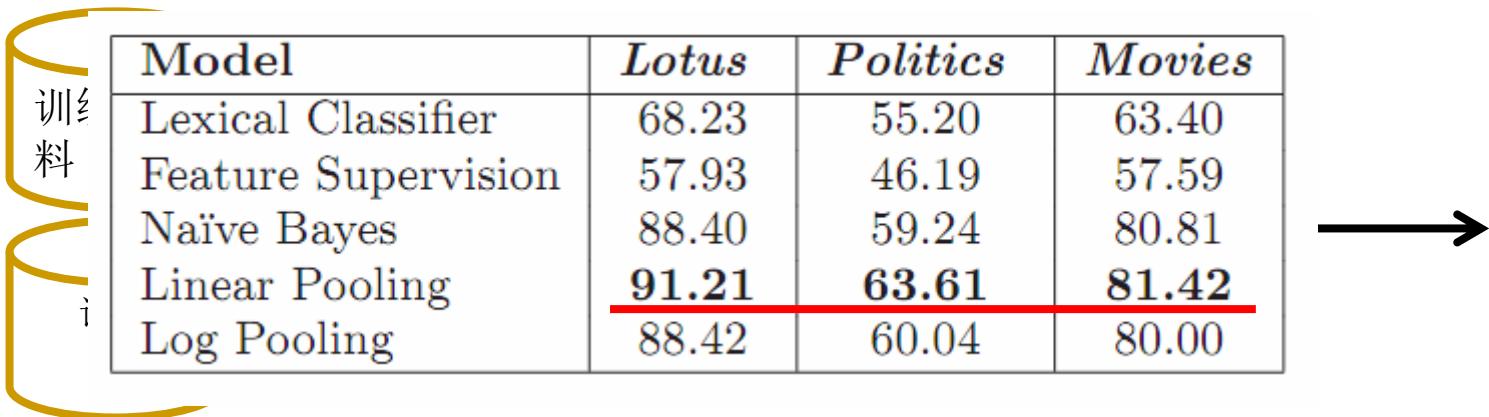
Lexicon-based Approach (2/2)

- 自学习方法 (Qiu CIKM 2009)
 - 利用词典信息产生初始标注
 - 利用置信度高的样本作为训练集，训练分类器
 - 利用启发式规则对于多个分类器进行集成



Combined Approaches (1/2)

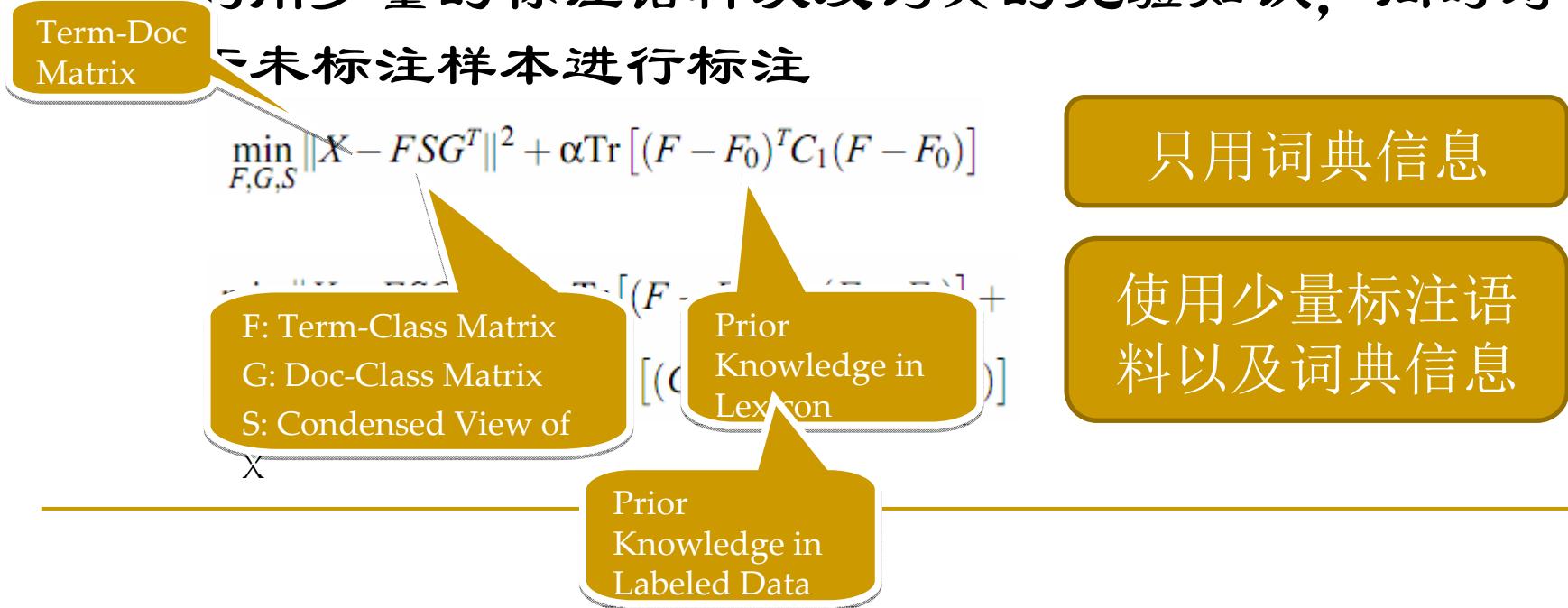
- 利用词典信息对于分类结果进行修正，主要解决训练语料不足的问题
- 分类器集成 (Melville KDD 2009)
 - 分别用语料和词典训练两个NB分类器
 - 对于分类器进行集成



Model	<i>Lotus</i>	<i>Politics</i>	<i>Movies</i>
Lexical Classifier	68.23	55.20	63.40
Feature Supervision	57.93	46.19	57.59
Naïve Bayes	88.40	59.24	80.81
Linear Pooling	91.21	63.61	81.42
Log Pooling	88.42	60.04	80.00

Combined Approaches (2/2)

- Semi-supervised Clustering (Li ACL 2009)
 - 建立文档与词的共现矩阵
 - 训练Matrix Factorization Model (cluster-based learning approach)
 - 利用少量的标注语料以及词典的先验知识，同时对未标注样本进行标注



小结

- Corpus-based VS. Lexicon-based
 - 基于训练语料的监督学习方法受到领域限制，需要对于每个领域都进行人工训练语料的标注
 - 基于词典的无监督方法具有领域独立性，但是缺乏领域词典，因此效果不如监督学习的方法
 - 结合两方面的优势
- 结合句子现象，还有很多问题需要处理
 - 比较句
 - 诺基亚5800比5230更超值
 - 否定词
 - ...

Sentiment Identification

- Word Level
- Sentence Level
- Document Level
- Others

Document Level Sentiment Identification

□ 任务：识别篇章整体观点倾向性

诺基亚5800屏幕很好，操作也很方便，通话质量也不错，但是外形偏女性化，而且电池不耐用，只能坚持一天，价格也偏贵，反正我觉得不值。

□ 绝大多数方法与句子级别方法类似

□ 特征+分类器

□ 关键问题

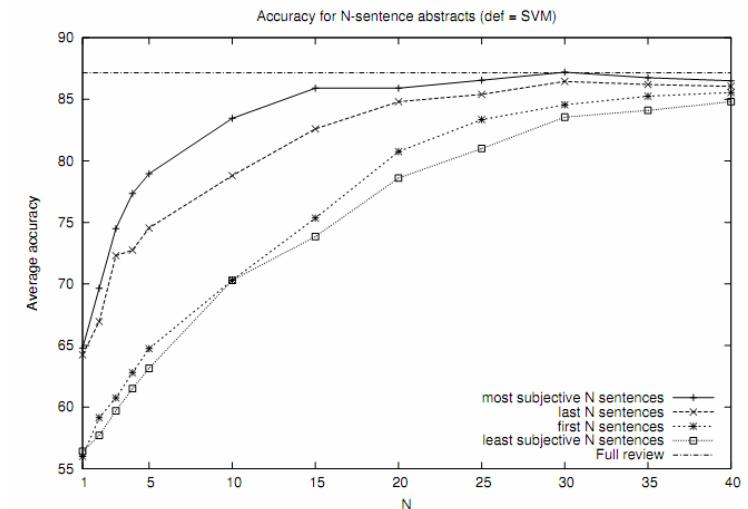
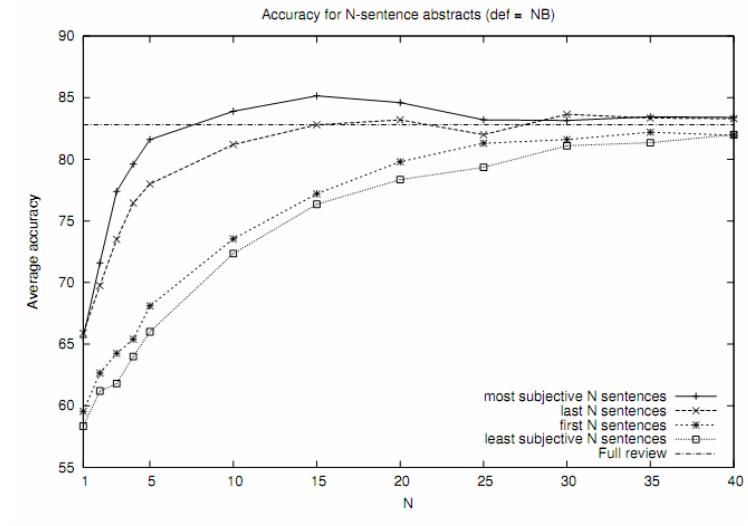
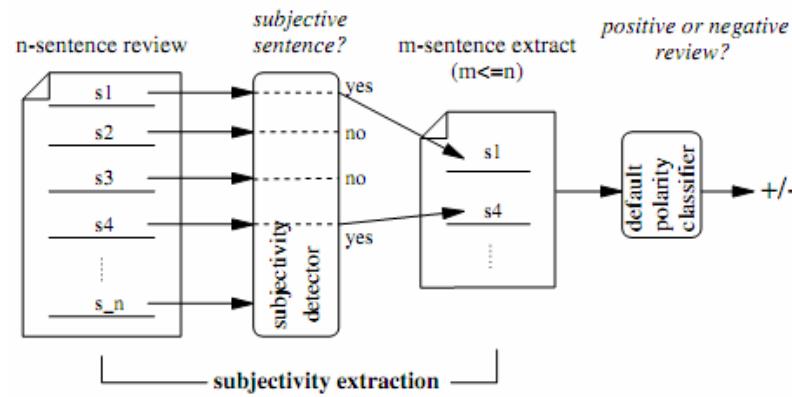
□ 多观点倾向性：一篇商品评论中可能包含对于商品多方面的观点，每个观点的倾向性也可能不同，如何识别篇章整体的观点倾向性

□ 按照句子划分

□ 按照主题划分

基于句子划分 (1/2)

- 篇章中的客观句子对于篇章整体的观点倾向性没有意义
(Pang ACL 2004)
- 利用图算法从篇章中识别出观点句，剔除客观句
- 只利用观点句来识别篇章整体的观点倾向性



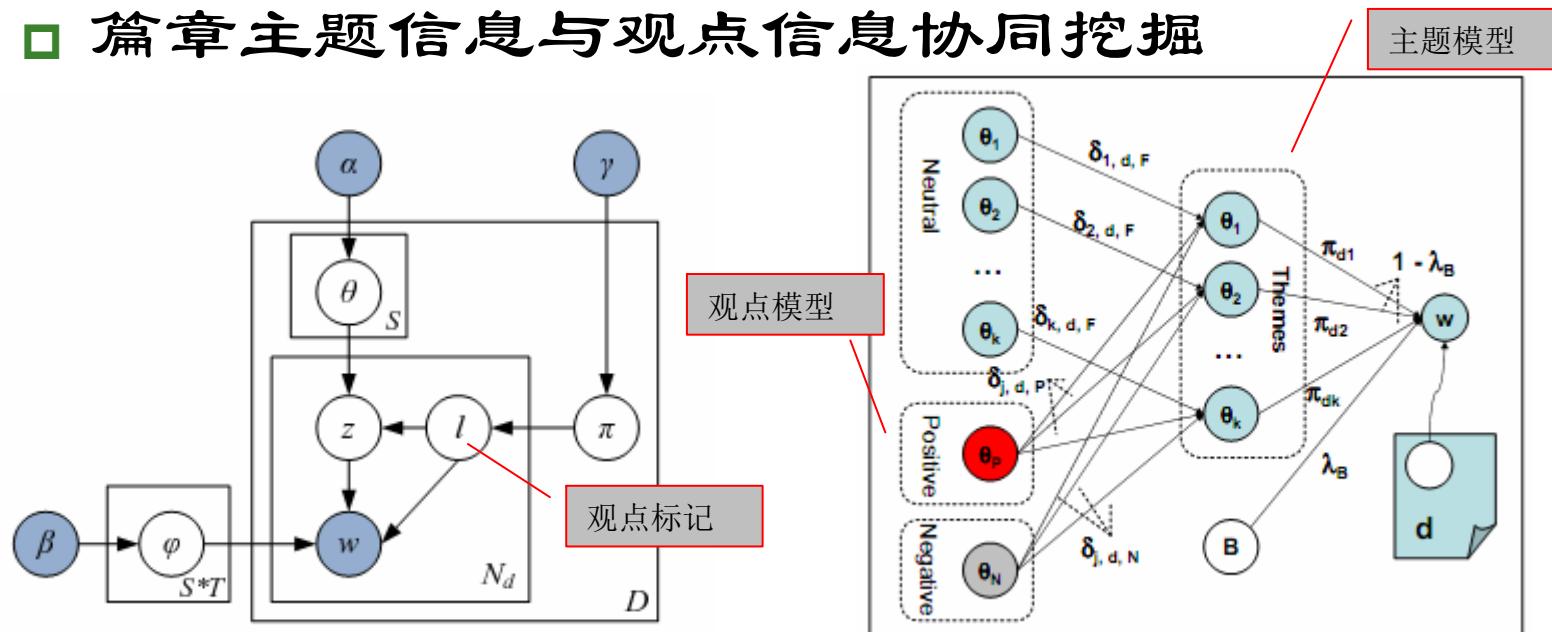
基于句子划分(2/2)

- 考慮篇章中每一个句子对于篇章整体倾向性的贡献 (McDonald ACL 2007)
 - 句子级倾向性识别与篇章级倾向性识别一体化
 - 结构化CRFs模型
 - 考慮句子的上下文，句子与篇章的关系

	Sentence Accuracy				Document Accuracy			
	Car	Fit	Mp3	Total	Car	Fit	Mp3	Total
Document-Classifier	-	-	-	-	72.8	80.1	87.2	80.3
Sentence-Classifier	54.8	56.8	49.4	53.1	-	-	-	-
Sentence-Structured	60.5	61.4	55.7	58.8	-	-	-	-
Joint-Structured	63.5*	65.2**	60.1**	62.6**	81.5*	81.9	85.0	82.8
Cascaded Sentence → Document	60.5	61.4	55.7	58.8	75.9	80.7	86.1	81.1
Cascaded Document → Sentence	59.7	61.0	58.3	59.5	72.8	80.1	87.2	80.3

基于主题的划分 (1/2)

- Lin (CIKM 2009), Mei (WWW 2007)
 - 篇章整体的观点倾向性是篇章中针对每个子主题的观点倾向性的集成
 - 篇章主题信息与观点信息协同挖掘



小结

- 篇章级观点倾向性识别仍然可以看做是一个text categorization 任务
 - 如果仅仅是用词袋子模型，那么 document level 与 sentence level 在处理方法上没有区别
- 主要问题在多观点混合问题
 - 篇章中局部观点与整体观点具不一致

Sentiment Identification

- Word Level
- Sentence Level
- Document Level
- Others
 - 跨语言观点识别与分析
 - 领域适应性

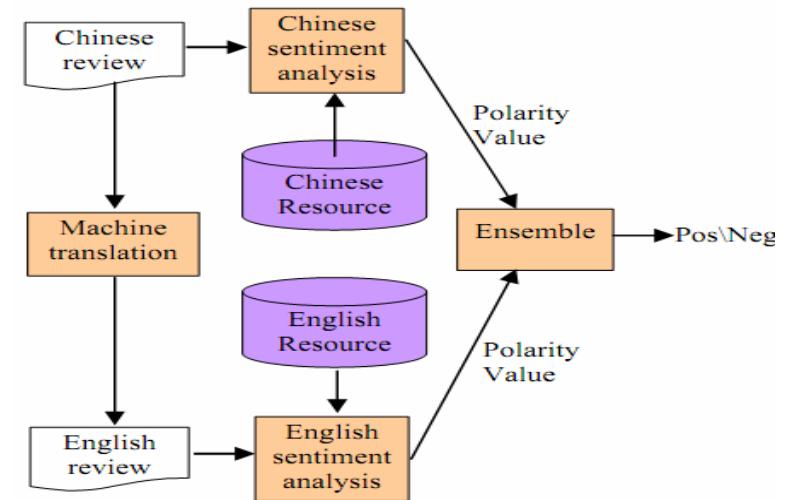
Cross-lingual Sentiment Classification

□ 任务

- 缺乏训练数据
- 利用其他语言资源
- 主要借鉴跨语言分类方法

□ 方法

- 借助于翻译系统
- 比较不同翻译系统的作用(Wan EMNLP 2008)
- 采用多视角学习策略 (Wan ACL 2009)



Sentiment Transfer (1/2)

□ 问题

- 不同领域的情感倾向性具有差异性
- 同样的词在不同的领域的情感倾向性不同
 - Screen is big (positive) Phone's size is big (negative)
- 不同领域的用词不相同
 - Car domain: faster, power,.....
 - Phone domain: colorful,
- 训练语料规模有限
 - 需要其他领域的标注数据
- 传统统计机器学习假设：训练数据与预测数据具有相同的分布

INVALID



Sentiment Transfer (2/2)

□ 方法 (两类)

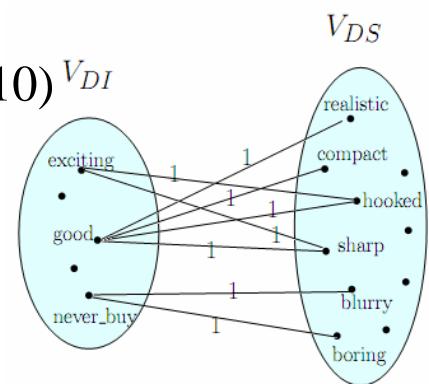
□ Instance view

- 假设：不同领域的数据的特征表示一致，数据分布不同
- 方法：调节样本的权重
- Jiang (ACL 2007), Dai (AAAI 2007)
 - The weight of the similar out-domain instances
 - The weight of the unlike out-domain instances



□ Feature representation view

- 假设：不同领域的数据的特征表示不一致
- 方法：统一特征表示
- Blitzer (ACL 2007), Liu (CIKM 2009), Pan (WWW 2010)
 - Select pivot features in two domains
 - Using pivot features to represent other features
 - Different data are represented in a unified feature space
 - Different features can correspond



内容

- Sentiment Identification
- Opinion Mining
 - Opinion Target Extraction
 - Opinion Holder Extraction
- Opinion Retrieval
- Resources and Evaluations

Opinion Target Extraction (1/4)

□ 任务：抽取观点评价的对象

- 中方发言人就美国近期对阿富汗的行动进行了强烈的谴责。 (新闻) 
- iphone4的屏幕简直太酷了！ (商品评论)
 - Product Feature: 商品、商品属性、商品的部件、商品部件的属性 (Popescu EMNLP 2005)

Explicit Features	Examples	% Total
Properties	ScannerSize	7%
Parts	ScannerCover	52%
Features of Parts	BatteryLife	24%
Related Concepts	ScannerImage	9%
Related Concepts' Features	ScannerImageSize	8%

□ 不是所有的商品属性都是评价的对象

- 诺基亚C1的屏幕尺寸有1.8寸。 
- iphone的价格太贵了 

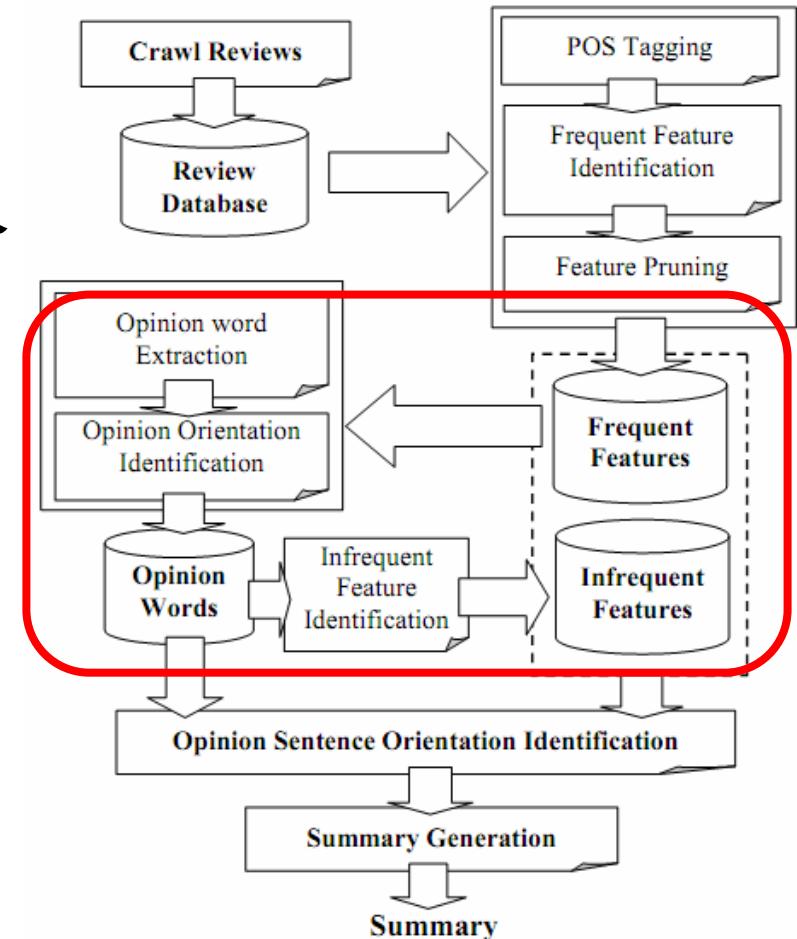
Opinion Target Extraction (2/4)

- 迭代抽取 (Liu KDD 2004, Liu WWW 2005)
- 商品属性词与评价词在评论文本中共同出现



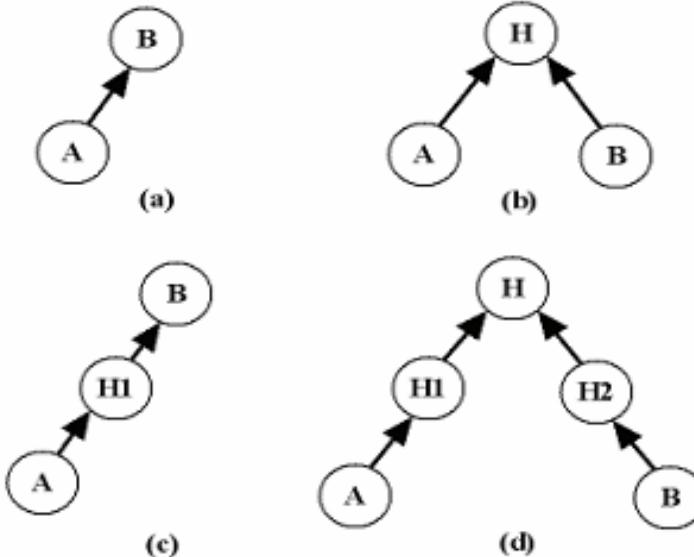
- 商品属性词分为两类
 - Frequent与Infrequent

Step 1: Frequent features extraction
Step 2: Opinion word extraction
Step 3: Infrequent features extraction
Step 4: Summarization



Opinion Target Extraction : 句法结构 (3/4)

□ 利用属性词与评价词之间的依存句法关系 (Popescu
EMNLP 2005, Qiu IJCAI 2009)

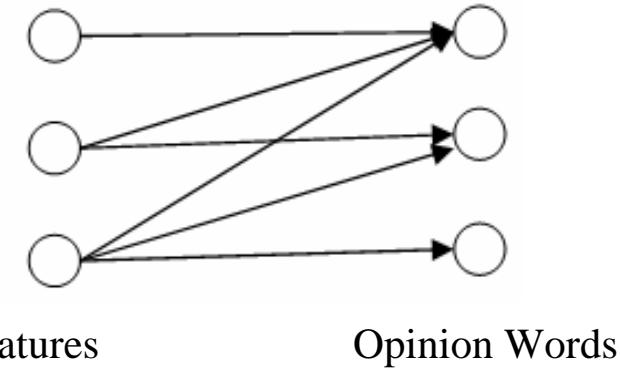


Extraction Rules	Examples
$\text{if } \exists(M, NP = f) \rightarrow po = M$	(expensive) scanner
$\text{if } \exists(S = f, P, O) \rightarrow po = O$	lamp has (problems)
$\text{if } \exists(S, P, O = f) \rightarrow po = P$	I (hate) this scanner
$\text{if } \exists(S = f, P, O) \rightarrow po = P$	program (crashed)

	Observations	Constraints	Outputs
R1 ₁	$S_{j0} \xrightarrow{} S_{j0}-Dep \xrightarrow{} S_{j0}$	$S_{j0} \in \{S\}$, $S_{j0}-Dep \in \{\text{CONJ}\}$, $POS(S_{j0}) \in \{JJ\}$	$s = S_{j0}$
R1 ₂	$S_i \xrightarrow{} S_r-Dep \xrightarrow{} H \xleftarrow{} S_j-Dep \xleftarrow{} S_j$	$S_i \in \{S\}$, $S_r-Dep == S_j-Dep$, $POS(S_j) \in \{JJ\}$	$s = S_j$
R2 ₁	$S \xrightarrow{} S-Dep \xrightarrow{} F$	$F \in \{F\}$, $S-Dep \in \{MR\}$, $POS(S) \in \{JJ\}$	$s = S$
R2 ₂	$S \xrightarrow{} S-Dep \xrightarrow{} H \xleftarrow{} F-Dep \xleftarrow{} F$	$F \in \{F\}$, $S/F-Dep \in \{MR\}$, $POS(S) \in \{JJ\}$	$s = S$
R3 ₁	$S \xrightarrow{} S-Dep \xrightarrow{} F$	$S \in \{S\}$, $S-Dep \in \{MR\}$, $POS(F) \in \{NN\}$	$f = F$
R3 ₂	$S \xrightarrow{} S-Dep \xrightarrow{} H \xleftarrow{} F-Dep \xleftarrow{} F$	$S \in \{S\}$, $S/F-Dep \in \{MR\}$, $POS(F) \in \{NN\}$	$f = F$
R4 ₁	$F_{j0} \xrightarrow{} F_{j0}-Dep \xrightarrow{} F_{j0}$	$F_{j0} \in \{F\}$, $F_{j0}-Dep \in \{\text{CONJ}\}$, $POS(F_{j0}) \in \{NN\}$	$f = F_{j0}$
R4 ₂	$F_i \xrightarrow{} F_r-Dep \xrightarrow{} H \xleftarrow{} F_j-Dep \xleftarrow{} F_j$	$F_i \in \{F\}$, $F_r-Dep == F_j-Dep$, $POS(F_j) \in \{NN\}$	$f = F_j$

Opinion Target Extraction: 监督与半监督 (4/4)

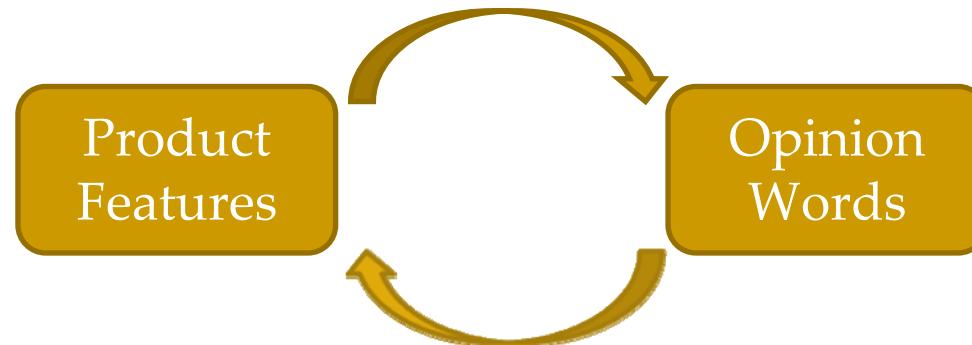
- 半监督学习方法 (Wang IJCNLP 2008, Zhu CIKM 2009)
 - 采用Bootstrapping策略
 - 使用少量标注的种子词
 - 利用属性词与评价此词之间的关联关系进行迭代抽取



- 监督学习方法 (Li Coling 2008, Zhao COAE 2008)
 - 看作序列标注任务
 - 利用CRFs进行标注

小结

- 基本思路
 - 迭代抽取



- 难点问题
 - “属性词-评价词”搭配关系的抽取
 - 商品名识别
 - 同义词问题
 - 体积、大小、尺寸.....
 - Implicit 属性词抽取
 - 太漂亮了 (外观)

Opinion Holder Extraction

- 基本思路(Kim AAAI 2005)
 - 命名实体识别
 - 人名、机构名
 - 句法结构特征
 - Convolution Kernel
 - 分类或者序列标注
 - SVM, Naïve Bayes, CRFs
 - 需要指代消解
 - “国家主席胡锦涛今天在钓鱼台国宾馆接见日本首相一行，在会谈中**他**表示.....”

内容

- Sentiment Identification
- Opinion Mining
- Opinion Retrieval
- Resources and Evaluations

Opinion Retrieval

□ 任务：

- 从海量文本中根据查询找到观点信息
- 根据主题相关度(topic relevance)与观点倾向性(opinion relevance)对于结果进行重排序
 - Topic relevance: traditional retrieval
 - Opinion relevance: opinion identification

□ 关键问题

- 找到Topic relevance score与Opinion relevance score的折中

Generative Model

- 基于词的观点检索模型 (Zhang SIGIR 2008)
 - 产生式模型

S: 观点信息(观点词)

$$p(d | q) \propto p(q | d)p(d)$$

主题相关

$$p(d | q, s) = \sum_i p(d | q, s_i)p(s_i, s)$$

$$= \frac{1}{|S|} \sum_i p(d | q, s_i)$$

Opinion
Relevance

$$\propto \frac{1}{|S|} \sum_i p(q, s_i | d)p(d)$$

Topic
Relevance

$$= \frac{1}{|S|} \sum_i p(s_i | d, q)p(q | d)p(d)$$

$$p(d | q, s) = (1 - \lambda)p(s | d, q) + \lambda p(q | d)p(d)$$

Unified Relevance Model

□ 查询词扩展(Huang CIKM 2009)

□ 查询中往往没有观点词

- 7.23事件
- 需要对于查询进行扩展(添加观点词)

□ 与查询独立的观点信息扩展

- 词典信息
- 标注的倾向性语料中进行统计

□ 与查询相关的观点信息扩展

- 从用户的反馈数据中得到

□ 混合模型

$$\begin{aligned}Score(D) = & \alpha \sum_{w \in Q} P(w|Q) \log P(w|D) + \\& + \beta \sum_{w \in OV_1} P(w|R_1) \log P(w|D) + \\& + (1 - \alpha - \beta) \sum_{w \in OV_2} P(w|R_2) \log P(w|D)\end{aligned}$$

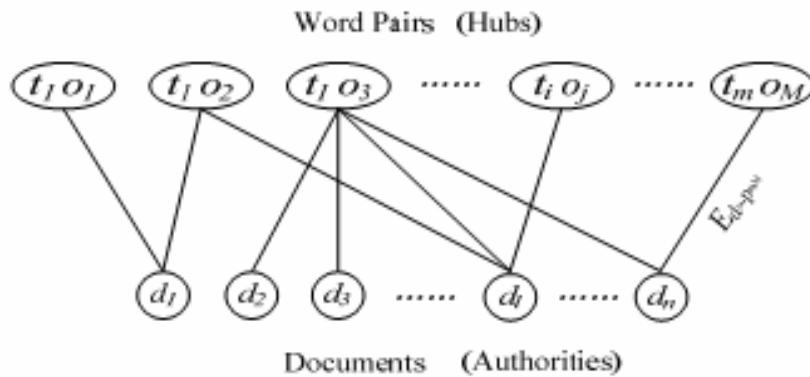
Topic
Relevance

Query
independent
sentiment
expansion

Query
dependent
sentiment
expansion

Sentence-based Opinion Retrieval

- 面向句子级观点检索文本表示 (Li ACL 2010)
 - 传统的词袋子模型不能很好表示文档中的观点信息
 - 利用topic-sentiment pair 表示每一个句子
 - 采用窗口共现策略抽取pair
 - 利用HITS算法来计算每个pair在篇章中的权重



Approach	COAE08			
	Evaluation metrics			
Run id	MAP	R-pre	bPref	P@10
IR	0.2797	0.3545	0.2474	0.4868
Doc	0.3316	0.3690	0.3030	0.6696
ROSC	0.3762	0.4321	0.4162	0.7089
Baseline	0.3774	0.4411	0.4198	0.6931
GORM	0.3978	0.4835	0.4265	0.7309

内容

- Sentiment Identification
- Opinion Mining
- Opinion Retrieval
- Resources and Evaluations
 - 资源：词典、语料
 - 评测：评测会议

Resources: Lexicon (1/2)

□ English

- General Inquier (<http://www.wjh.harvard.edu/~inquirer/>)
 - Manually labeled terms (positive, negative)
- SentiWordnet (<http://sentiwordnet.isti.cnr.it/>)
 - Extend from WordNet
 - Each synset is automatically labeled as P, N, O
- OpinionFinder's Subjectivety Lexicon
(<http://www.cs.pitt.edu/mpqa/>)
 - Subjective words provided by OpinionFinder
- Taboada and Grieve's Turney adjective list
 - Available through Yahoo SentimentAI group. 1700 words
- IBM Lexicon
 - 1,267 positive words and 1,701 negative words (Melville 2009)

Resources: Lexicon (2/2)

□ Chinese

- Hownet (http://www.keenage.com/html/e_index.html)
 - 正面情感、负面情感、正面评价、负面评价、程度级别、主张词语6个子集
- NTU Sentiment Lexicon
(<http://nlg18.csie.ntu.edu.tw:8080/opinion/userform.jsp>)
 - List the polarities of many Chinese words

Resource: Corpus (1/2)

□ English

- MPQA (<http://www.cs.pitt.edu/mpqa/databaserelease/>)
 - 535 news articles (subjective, objective; P,N,O)
- Movie review data (<http://www.cs.cornell.edu/people/pabo/movie-review-data/>)
 - IMDB
 - Document level 2000
 - Sentence level 5000
- Custom review data (<http://www.cs.uic.edu/~liub/FBS/CustomerReviewData.zip>)
 - Product reviews (Product features, P,N)
- Multi product reviews (<http://john.blitzer.com/software.html>)
 - Book, Electronic, Kitchen, DVD
 - 2000 in each domain
- TREC Blog corpus (<http://trec.nist.gov/>)
 - Blog data
 - 3,000,000 Webpages
- Multiple-aspect restaurant reviews
 - 4,488 reviews
 - Each review labeled as 1-5 stars

Resource: Corpus (2/2)

□ Chinese

□ NTCIR (<http://research.nii.ac.jp/ntcir/>)

□ Multilingual news articles

□ COAE商品属性语料

□ 口碑网, it168,

□ 494 document, 5 domains

□ 中文情感挖掘语料

□ Positive, Negative

□ 10,000

□ Zagibalov (<http://www.informatics.sussex.ac.uk/users/tz21/>)

□ Phone reviews

□ 1,158 positive and 1,159 negative

Evaluations

- TREC Blog Track (start from 2006)
 - ▣ Task: Opinion Retrieval and Polarity Identification
 - ▣ Corpus: 3,000,000 English webpages
- NTCIR
 - ▣ Task:
 - ▣ Topic Relevance
 - ▣ Opinion identification
 - ▣ Polarity Identification
 - ▣ Opinion Holder extraction
 - ▣ Opinion Target extraction
 - ▣ Corpus: news articles (English, Chinese, Japanese, Korea)
- Chinese (COAE 2008, 2009)
 - ▣ Task:
 - ▣ Words level (sub/obj, positive/negative)
 - ▣ Documents level (sub/obj, positive/negative)
 - ▣ Opinion Target extraction
 - ▣ Opinion Retrieval
 - ▣ Corpus: Chinese

目录

□ 第一部分：

- 我们为什么需要观点挖掘与倾向性分析？
- 什么是观点挖掘与倾向性分析？

□ 第二部分：

- 如何进行观点挖掘与倾向性分析？
 - 任务、方法、资源、评测

□ 第三部分：

- 问题与挑战

观点信息应该如何表示？(1/2)

- An *opinion* is a quintuple (Liu *Handbook in NLP*)
- $(o_j, f_{jk}, so_{ijkl}, h_i, t_l),$
- where
 - o_j is a target object.
 - f_{jk} is a feature of the object o_j .
 - so_{ijkl} is the sentiment value of the opinion. so_{ijkl} is +ve, -ve, or neu, or a more granular rating.
 - h_i is an opinion holder.
 - t_l is the time when the opinion is expressed.

观点信息应该如何表示？(2/2)

□ 结构化表示 (Wu EMNLP 2011)

- “Takes good picture during the daytime. Very poor picture quality on the night.”

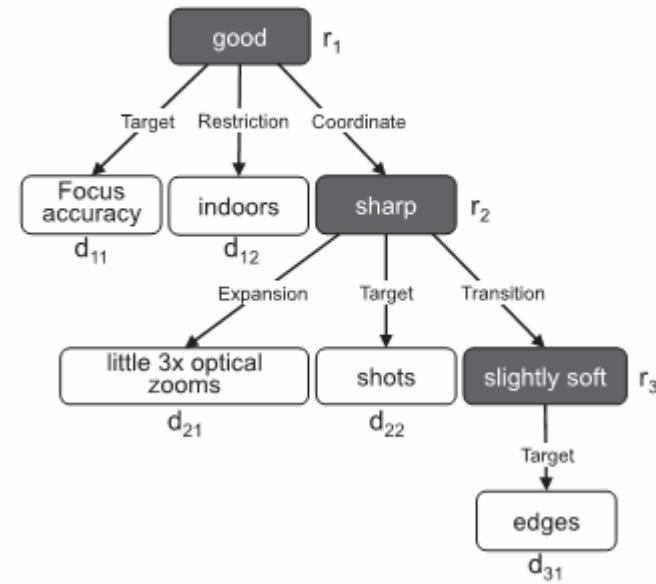
□ 观点内部的结构化表示

- Target, Restriction,...

□ 观点间的结构化表示

- Coordinate, Transition,...

Focus accuracy was good indoors, and although the little 3x optical zooms produced sharp shots, the edges were slightly soft on the Canon.



Challenge: Sentiment Identification(1/5)

□ Sentence Level

- 如何对于一个句子中的观点信息进行表示 ?
- BOW 模型 ? 句法结构 ?
- What is different with topic-based categorization

	Features	# of features	frequency or presence?	NB	ME	SVM
(1)	unigrams	16165	freq.	78.7	N/A	72.8
(2)	unigrams	"	pres.	81.0	80.4	82.9
(3)	unigrams+bigrams	32330	pres.	80.6	80.8	82.7
(4)	bigrams	16165	pres.	77.3	77.4	77.1
(5)	unigrams+POS	16695	pres.	81.5	80.4	81.9
(6)	adjectives	2633	pres.	77.0	77.7	75.1
(7)	top 2633 unigrams	2633	pres.	80.3	81.0	81.4
(8)	unigrams+position	22430	pres.	81.0	80.1	81.6

- 一些特定的句法现象
 - Polarity Shift (否定、转折.....)
 - Comparative Sentence
- 数据稀疏问题, 如何特征扩展
 - 微博、产品评论

Comparatives 类型	例子
Non-Gradable	A have feature F, but B doesn't have
Non-Equal Gradable	A is better than B
Equative	A is as good as B
Superlative	A is the best

Challenge: Sentiment Identification(2/5)

□ Word Level

□ 如何对于词的领域性进行区分

□ Independence word

□ 好、坏、轻松、迅速.....

□ Topic depended word

□ 大 (屏幕大, 体积大)

□ 高 (个子高, 温度高)

□ 名词、动词也具有倾向性

□ “这家餐馆不会再来了”

□ “坑爹啊”

□ “😊”

Challenge: Sentiment Identification(3/5)

- Feature (Aspect) Level

- Product Feature Extraction

- Explicit Feature: 这款手机的屏幕很漂亮。 (屏幕)
 - Implicit Feature: 这款手机太大了。 (体积)

- Feature Grouping

- 屏幕：LCD、屏幕、显示屏.....

- Feature-based Sentiment Identification

- Word matching + Word level sentiment identification
 - “**诺基亚5800个头很大**”。
 - 句法分析？面对口语化文本似乎力不从心

Challenge: Sentiment Identification(4/5)

- Document Level
- Sentiment Rating
 - 不用用户，不同尺度
- 多观点混合问题
 - 很难精确地确定篇章的倾向性由哪个句子决定
 - 不一定是多数制胜



口味: 2(好) 环境: 2(好) 服务: 2(好) 人均: ¥90(晚餐)

跟朋友聚餐来吃的，晚上来的，人很多，好在我们来得早，没有等位，点的是锅底一直不上，后来上了还上错了，好不容易才把我们要的端上来，味道们家的特色已经逐渐不明显了，就是羊肉还不错，比较新鲜。最后临走服务折，后来填完了说大不了折了，说送个大果盘，我们本来也不是想占这个便宜了，跟海底捞什么的确实没法比

“诺基亚5800屏幕很好，操作也很方便，通话质量也不错，外形还可以，但是电池太不行了，只能坚持一天，反正我觉得不值。”

Challenge: Sentiment Identification(4/5)

- 按照Aspect进行划分 (Yu ACL 2011)
 - 用户往往关注篇章中重要的Aspect
 - 方法：
 - 抽取Aspect
 - 计算Aspect的重要性，并进行排序
 - 利用重要的Aspect Words以及修饰词作为特征，训练分类器

<i>Our Method</i>
Usability
Apps
3G
Battery
Looking
Storage
Price
Software
Camera
Call quality

Data set	SVM + Boolean			SVM + tf			SVM + IA		
	P	R	F ₁	P	R	F ₁	P	R	F ₁
Canon EOS	0.689	0.663	0.676	0.679	0.654	0.666	0.704	0.721	0.713
Fujifilm	0.700	0.687	0.693	0.690	0.670	0.680	0.731	0.724	0.727
Panasonic	0.659	0.717	0.687	0.650	0.693	0.671	0.696	0.713	0.705
MacBook	0.744	0.700	0.721	0.768	0.675	0.718	0.790	0.717	0.752
Samsung	0.755	0.690	0.721	0.716	0.725	0.720	0.732	0.765	0.748
iPod Touch	0.686	0.746	0.714	0.718	0.667	0.691	0.749	0.726	0.737
Sony NWZ	0.719	0.652	0.684	0.665	0.646	0.655	0.732	0.684	0.707
BlackBerry	0.763	0.719	0.740	0.752	0.709	0.730	0.782	0.758	0.770
iPhone 3GS	0.777	0.775	0.776	0.772	0.762	0.767	0.820	0.788	0.804
Nokia 5800	0.755	0.836	0.793	0.744	0.815	0.778	0.805	0.821	0.813
Nokia N95	0.722	0.699	0.710	0.695	0.708	0.701	0.768	0.732	0.750

Challenge: Others

- 观点检索

- Re-Rank路线

- $\text{Score} = \text{lamada} * \text{TopicRelevance} + (1 - \text{lamada}) * \text{OpinionScore}$

- 能否有独立的模型框架

- 更宽广的应用：微博

- 观点Spam

- 内容、网页结构、用户行为

- 观点信息是动态变化的

- 时间、地点

- 观点分析与应用紧密结合

- 推荐系统

- 广告投放

- ...

Q&A

Thanks

Reference (1/5)

【Blitzer et al. 2007】 J. Blitzer, M. Dredze and F. Pereira. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL). pages 440-447. 2007.

【Dai et al. 2007】 Wenyuan Dai, Gui-Rong Xue, Qiang Yang and Yong Yu. Transferring Naïve Bayes Classifiers for Text Classification. In Proceedings. of AAAI. 2007.

【Du et al. 2010】 Weifu Du, Songbo Tan, Xueqi Cheng, Xiaochun Yun: Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon. WSDM 2010: 111-120

【Hassan et. al. 2010】 Ahmed Hassan, and Dragomir Radev. 2010. Identifying Text Polarity Using Random Walks. The 48th Annual Meeting of the Association for Computational Linguistics

【Hu et al. 2004】 M. Hu and B. Liu. Mining Opinion Features in Customer Reviews. In Proceedings of AAAI, 2004.

【Huang et al. 2009】 Xuanjing Huang and W. Bruce Croft. A Unified Relevance Model for Opinion Retrieval. In Proceedings of CIKM 2009.

【Kamps et al., 2004】 Jaap Kamps, Maarten Marx, Robert J. Mokken and Maarten de Rijke. Using WordNet to measure semantic orientation of adjectives. In Proc. of LREC'04, pp. 1115-1118, 2004.

Reference (2/5)

【Jiang et al. 2007】 Jin Jiang and ChengXiang Zhai. Instance Weighting for Domain Adaptation in NLP. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL), pages 264-271. 2007.

【Kim et al. 2005】 Soo-Min Kim and Eduard Hovy. Identifying Opinion Holders for Question Answering in Opinion Texts. 2005. *In Proceedings of AAAI-05 Workshop on Question Answering in Restricted Domains*

【Li et al. 2010】 Binyang Li, Lanjun Zhou, Shi Feng, Kam-Fai Wong, A Unified Graph Model for Sentence-based Opinion Retrieval, In Proceedings of ACL 2010

【Li et al. 2009】 Tao Li, Yi Zhang and Vikas Sindhwan. A Non-negative Matrix Tri-factorization Approach to Sentiment Classification with Lexical Prior Knowledge. In Proceedings of ACL. 2009.

【Li et al. 2009】 Shoushan Li, Rui Xia, Chengqing Zong, Chu-Ren Huang: A Framework of Feature Selection Methods for Text Categorization. ACL/AFNLP 2009: 692-700.

【Li et al. 2010】 Shoushan Li, Sophia Yat Mei Lee, Ying Chen, Chu-Ren Huang, Guodong Zhou: Sentiment Classification and Polarity Shifting. COLING 2010: 635-643

【Li et al. 2010】 Fangtao Li, Chao Han, Minlie Huang and Xiaoyan Zhu. Structure-Aware Review Mining and Summarization. In The 23rd International Conference on Computational Linguistics (COLING 2010),

Reference (3/5)

【Liu et al. 2009】 Kang Liu and Jun Zhao. Cross-Domain Sentiment Classification using a Two-Stage Method. In Proceedings of *the 18th ACM Conference on Information and Knowledge Management (CIKM)*. November 2-6, 2009, Hong Kong

【Lin et al. 2009】 Chenghua Lin and Yulan He. Joint Sentiment/Topic Model for Sentiment Analysis. In Proceedings of CIKM's 09. 2009

【Mao et al. 2007】 Y. Mao and G. Lebanon, Isotonic Conditional Random Fields and Local Sentiment Flow. Advances in Neural Information Processing Systems 19, 2007

【McDonald et al. 2007】 Ryan McDonald, Kerry Hannan and Tyler Neylon et al. Structured Models for Fine-to-Coarse Sentiment Analysis. In Proceedings of ACL, 2007, pp. 432-439.

【Mei et al. 2007】 Qiaozhu Mei, Xu Ling, et al. Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs. In Proceedings of WWW 2007.

【Melville et al. 2009】 Prem Melville, Wojciech Gryc and Richard D. Lawrence. Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification. In Proceedings of KDD. 2009.

【Pang et al. 2004】 Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In Proceedings of the Association of Computational Linguistics (ACL).

【Pang et al. 2002】 Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of EMNLP 2002, pp.79-86.

Reference (4/5)

【Pan et al. 2010】 Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang and Zheng Chen. Cross-Domain Sentiment Classification via Spectral Feature Alignment. In Proceedings of the 19th International World Wide Web Conference (WWW-10). Raleigh, NC, USA. April 26-30, 2010. Pages 751-760.

【Popescu et al. 2005】 Popescu A. M. and Etzioni O. Extracting Product Features ad Opinion Reviews. In Proceedings of EMNLP'05, 2005.

【Qiu et al. 2009】 L. Qiu, Weishi Zhang, Changjian Hu and Kai Zhao. SELC: A Self-Supervised for Sentiment Classification. In Proceedings of CIKM, 2009.

【Qiu et al. 2009】 Guang Qiu, Bing Liu, Jiajun Bu, Chun Chen: Expanding Domain Sentiment Lexicon through Double Propagation. IJCAI 2009: 1199-1204

【Turney et al. 2002】 Peter Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of ACL. 2002.

【Wan et al. 2009】 Xiaojun Wan. Co-Training for Cross-Lingual Sentiment Classification. In Proceedings of ACL-IJCNLP, pages 235-243, 2009.

【Wan et al. 2008】 Xiaojun Wan. Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis. In Proceedings of EMNLP, pages 553-561. 2008.

【Wang et al. 2008】 Bo Wang, Houfeng Wang: A Cross-Inducing Method for Bootstrapping Product Features and Opinion Words. In Proceedings of 2008 International Conference on Natural Language Processing (IJCNLP 2008), India

Reference (5/5)

【Webie et al. 2005】 Janyce Webie, Theresa Wilson and Claire Cardie. Annotating expressions of opinions and emotions in Proceedings of language. Language Resources and Evaluation 2005

【Zagibalov et al. 2008】 Taras Zagibalov. and John Carroll. Automatic seed word selection for unsupervised sentiment classification of Chinese text. In Proceedings of The 22nd International Conference on Computational Linguistics (COLING), 2008. Manchester, UK.

【Zhang et al. 2008】 Min Zhang and Xingyao Ye. A Generative Model to Unify Topic Relevance and Lexicon-based Sentiment for Opinion Retrieval. In Proceedings of SIGIR, pp. 411-418, 2008.

【Zhao et al. 2008】 Jun Zhao, Kang Liu and Gen Wang. Adding Redundant Features for CRFs-based Sentence Sentiment Classification. In Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP). October 25-27, 2008, Hawaii

【Zhu et al. 2009】 Jingbo Zhu, Huizhen Wang, Benjamin Tsou and Muhua Zhu. 2009. Multi-aspect opinion polling from textual reviews, In Proceedings of CIKM'09, short session, pp1799-1802

第三课

问答系统

中国科学院自动化研究所
模式识别国家重点实验室

概述

- 问答系统概述
 - 研究背景
 - 发展历史
- 问答式检索系统
- 社区问答系统
- 问题与挑战

问答系统概述：定义

□ Marybur, AAAI 2003

- Question Answering (QA) is an interactive human computer process that encompasses understanding a user information need, typically expresses in a natural language query; retrieving relevant documents, data, or knowledge from selected sources; extracting, qualifying and prioritizing available answers from these sources; and presenting and explaining responses on an effective manner.

□ 定义

- 输入：自然语言的问句，而非关键词的组合

谁获得1987年的诺贝尔文学奖？

- 输出：直接答案，而非文档集合

约瑟夫•布罗茨基

QA研究对于自然语言理解的意义

- 美国认知心理学家G.M.Ulson认为，判别计算机理解自然语言的4个标准是：QA、Summarization、Paraphrase和MT。计算机只要达到以上标准之一，就认为它理解了自然语言
- 自然语言理解在词语层面、句子层面、篇章层面、篇章之间、语言之间的基本问题，在QA中都会出现。另一方面，QA和信息检索密切相关，信息检索中的基本问题在QA中也同样存在
- QA是自然语言理解研究人员追求的目标之一，它的研究会带动自然语言理解的发展
- QA研究本身也有重要应用价值

概述

□ 问答系统概述

□ 研究背景

□ 发展历史

□ 问答式检索系统

□ 社区问答系统

□ 问题与挑战

问答系统概述：发展历史(1/7)

- 1950年，A. M. Turning 提出“图灵测试”
 - 一个人在不接触对方的情况下，通过一种特殊的方式，和对方进行一系列问答。如果在相当长时间内，他无法根据这些问题判断对方是人还是计算机，那么，就可以认为这个计算机具有同人相当的智力，即这台计算机是能思维的。

理想模型

问答系统概述：发展历史(2/7)

□ 基于知识推理的问答系统

- 主要特点：答案或者从知识库中检索得到，或者在知识库上经过推理得到
- 代表性系统：
 - 自然语言界面的专家系统：如：MIT开发的数学符号运算系统MACSYMA (1960年左右)
 - 基于本体的问答系统：如：陆汝钤院士主持开发的Pangu (2000年)

对话者：“动物园有一头黑熊死了。”

计算机：“黑熊是怎么死的？”

对话者：“据说黑熊是吃塑料袋死的。”

计算机：“准是哪个不文明的游客投进去的。”

问答系统概述：发展历史(3/7)

□ 基于知识推理的问答系统

□ 优点：

- 性能良好，对于用户提出的许多问题，回答准确，具有一定的推理能力

□ 缺点：

- 人工建设知识库非常困难，知识库规模有限，领域有限
- 如果用户的问题超出了知识库的范围，系统性能很差

问答系统概述：发展历史(4/7)

□ 问答式检索系统

- 互联网技术的发展，使得人们可以方便地从网络上获取信息
- 搜索引擎为人们的信息获取提供了可能，但搜索引擎无法清楚表达人们的信息需求意图，返回的信息太多
- 为了克服搜索引擎的不足，问答式检索系统应运而生
- 主要特点：利用信息检索以及浅层自然语言处理技术从大规模文本库或者网页库中抽取出答案
- 代表性系统：
 - MIT开发的Start(<http://start.csail.mit.edu/>)
 - Umass开发的QuASM(<http://nyc.lti.cs.cmu.edu/IRLab/11-743s04/>)
 - Microsoft开发的Encarta(<http://encarta.msn.com/>)

问答系统概述：发展历史(5/7)

□ 问答式检索系统

□ 优点：

- 相对于基于知识推理的问答系统而言：不受知识库规模限制，不受领域限制，更加接近真实应用需求
- 相对于传统的搜索引擎而言：问答式检索系统接受的是自然语言形式的提问，由于自然语言处理技术的应用，对用户意图的把握更加准确，呈现给用户的答案更加准确

□ 缺点：

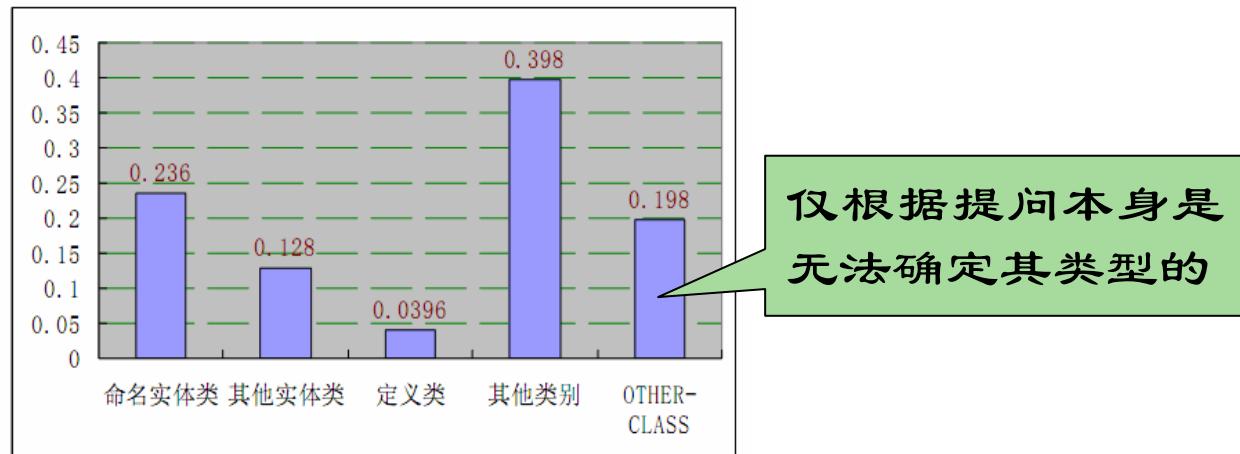
- 目前问答式检索系统仅能处理有限的简单问题，如Factoid问题、List问题等
- 而用户的提问是复杂多样的

问答系统概述：发展历史(6/7)

□ 问答式检索系统

□ 缺点：

- 答案类型集中于实体、定义等有限类别，后来TAC-QA扩展了观点类别，但是很多用户需求是没有明显类别的，目前的问答式检索系统还很难处理



(吴友政 博士论文 2006)

问答系统概述：发展历史(6/7)

□ 社区问答系统：

- 随着Web2.0 的兴起，基于用户生成内容(User-Generated Content, UGC)的互联网服务越来越流行，社区问答系统应运而生，为问答系统的发展注入了新的活力和生机
- 用户可以提出任何类型的问题，也可以回答其它用户的问题，通过问答方式来满足人们的信息查找和知识分享需求
- 代表性系统：
 - 英文：Yahoo! Answers等
 - 中文：百度知道、新浪爱问等

问答系统概述：发展历史(7/7)

	知识推理系统	问答式检索	社区问答
数据源规模	小规模	大规模	大规模
数据源质量	没噪	噪声较小	有噪
查询类型	受限制	受限制	不受限制
社会性	无	无	有
典型应用	专家系统	事实性问答	一般性问答

小结

- 在知识工程、互联网以及自然语言处理技术的推动下，问答系统取得了一定进展
- 数据源：从限定领域向开放领域发展，从小规模向海量发展
- 提问类型：从受限类型向开放类型发展
- 参与者：领域专家到普通用户，门槛越来越低
- 知识共享：从个人独享到大众知识分享

概述

- 问答系统概述
- 问答式检索系统
 - 问答式检索评测
 - 问答式检索方法
 - 问答式检索应用：Watson
- 社区问答系统
- 问题与挑战

问答式检索评测

- 目前的评测都是面向大规模文本库的开放域问答式检索
- 评测现状
 - 美国：TREC (Text Retrieval Conference)
 - NIST, DARPA
 - 日本：NTCIR (NII-NACSIS Test Collection for IR Systems) Project
 - 欧洲：CLEF (Cross-Language Evaluation Forum)
 - 欧洲语言，单语言，跨语言
 - 中国：有一些问答资源，但是还没有组织过评测

TREC—QA Track评测任务(1/2)

- 1999年TREC-8到2007年TREC-16， QA评测进行了9届，评测任务主要包括：
 - Factoid任务：测试系统对基于事实、有简短答案的提问的处理能力。而那些需要总结、概括的提问不在测试之列
 - Where is Belize located ? (✓)
 - 如何办理出国手续？如何赚钱？ (✗)
 - List任务：要求系统列出满足条件的几个答案
 - Name 22 cities that have a subway system.
 - List the names of chewing gums.
 - Definition任务：要求系统给出某个概念、术语或现象的定义和解释
 - What is SARS ?

TREC—QA Track评测任务(2/2)

- Context任务：测试系统对相关联的系列提问的处理能力，即对提问 i 的回答依赖于对提问 j 的理解
 - a. 佛罗伦萨的哪家博物馆在1993年遭到炸弹的摧毁？
 - b. 这次爆炸发生在那一天？
 - c. 有多少人在这次爆炸中受伤？
- Passage任务：对答案的要求偏低，不需要系统给出精确答案，只要给出包含答案的一个字符序列
 - 中华人民共和国是什么时候成立的？
 - 自从1949年10月1日中华人民共和国成立以来至1994年底止，我国已经同世界上的约160个国家建立了外交关系，而且还同更多的国家和地区发展了经济贸易关系和文化往来

TREC-QA评测（评测指标）

- TREC QA Track的评测指标主要有平均排序倒数（Mean Reciprocal Rank，简称MRR）、准确率（Accuracy）、CWS（Confidence Weighted Score）等
- MRR：
$$MRR = \sum_{i=1}^N \frac{1}{\text{标准答案在系统给出的排序结果中的位置}}$$
 - 如果标准答案存在于系统给出的排序结果中的多个位置，以排序最高的位置计算；如果标准答案不在系统给出的排序结果中，得0分。N表示测试集中的提问个数
- CWS：
$$CWS = \frac{1}{N} \sum_{i=1}^N \frac{\text{前}i\text{个提问中被正确回答的提问数}}{i}$$
 - N表示测试集中的提问个数

小结

- TREC-QA评测的难度逐渐增加，涉及到的自然语言处理技术也越来越复杂
- TREC-QA评测的任务类型非常有限。据统计，仅占提问总数的40%左右(吴友政 博士论文 2006)
- TREC-QA评测从英语逐渐向多语言发展

仅占40%左右



仅根据提问本身是无法分类的

概述

- 问答系统概述
- 问答式检索系统
 - 问答式检索评测
 - 问答式检索方法
 - 问答式检索应用：Watson
- 社区问答系统
- 问题与挑战

问答式检索方法

- 信息检索 + 信息抽取
- 信息检索 + 模式匹配
- 信息检索 + 自然语言处理技术
- 基于统计翻译模型的问答技术

信息检索 + 信息抽取

- **方法描述：**从问句中提取关键词语，用信息检索的方法找出包含候选答案的段落或句子，然后基于问答类型用信息抽取方法从这些段落和句子中找出答案
 - **检索过程：**段落或者句子级排序，利用不同类型关键词的加权组合
 - **答案抽取过程：**根据问答类型从排序后的段落或句子中抽取答案
- **特点：**
 - **优点：**技术相对成熟，易于开发
 - **缺点：**准确率一般，不能推理
- **代表系统：**新加坡国立大学Hui Yang的系统(Yang TREC 2002)

信息检索 + 模式匹配(1/2)

□ 方法描述：

- 基本思想：对于某些提问类型（某人的出生日期、原名、别称等），问句和包含答案的句子之间存在一定的答案模式，该方法在信息检索的基础上根据这种模式找出答案。因此如何自动获取某些类型提问的尽可能多的答案模式是其中的关键技术
- 例如，询问“某人生日年月日”类提问的部分答案模式如下：
 - 1.0 <NAME> (<ANSWER> -)
 - 0.85 <NAME> was born on <ANSWER>
 - 0.6 <NAME> was born in <ANSWER>
 - 0.59 <NAME> was born <ANSWER>
 - 0.53 <ANSWER> <NAME> was born
 - 0.50 - <NAME> (<ANSWER>
 - 0.36 <NAME> (<ANSWER> -

信息检索 + 模式匹配(2/2)

□ 包括两阶段的任务：

- 离线阶段：获取答案模式
- 在线阶段：首先判断当前提问属于哪一类，然后使用这类提问的所有模式来抽取候选答案

□ 模式获取方法：

- 表层字符串匹配(Ravichandran ACL 2002)
- 深层句法分析(Lin NLE 2001)
- 人们的注意力从原来的基于深层文本分析方法转移到基于字符的表层的文本分析技术上

□ 特点：

- 优点：对于某些类型的问题(如生日问题等)效果良好
- 缺点：无法表达长距离、复杂关系，没有推理能力

□ 代表系统：俄罗斯Martin Soubbotion等人研发的系统 (Soubbotion TREC 2002)

信息检索+自然语言处理技术(1/2)

□ 方法描述：

- 对问句和答案句进行浅层分析，获得句子的浅层句法、语义表示，作为对前两种方法的补充和改进

□ 涉及到的自然语言处理技术主要包括：

- 命名实体识别技术(Ravichandran ACL 2002)
- 句法分析技术(Lin NLE 2001)
- 逻辑表示(Harabagiu TREC 2000; Moldovan ACL 2001; Pasca ACL 2001)
- 复述关系(Duclay EACL 2003)

...

信息检索+自然语言处理技术(2/2)

□ 特点：

- 优点：能够从句法、语义的角度解析答案
 - 缺点：技术还不成熟
- 代表系统：美国Language Computer Corporation公司 Sanda Harabagiu 等人研发的系统(Harabagiu TREC 2000)，该系统在TREC QA Track 评测中获得好成绩，且具有较大的领先优势

基于统计翻译模型的问答技术

□ 方法描述：

- 把提问句看作答案句在同一语言内的一种翻译

□ 特点：

- 过分依赖于训练集

□ 代表性工作：

- Berger SIGIR 2000;
- Echihabi ACL 2003;
- Murdock SIGIR Workshop 2004

四类问答技术的比较分析(1/2)

- **基于信息检索和信息抽取的问答技术**: 相对简单，容易实现。但它以基于关键词的检索技术(或称为词袋检索技术)为重点，只考虑离散的词，不考虑词之间的关系。因此无法从句法关系和语义关系的角度解释系统给出的答案，也无法回答需要推理的提问
- **基于模式匹配的问答技术**: 虽然对于某些类型提问(如定义，出生日期提问等)有良好的性能，但无法找到所有提问的答案模式，长距离模式和表达复杂关系的模式的获取也很困难，同样无法实现推理

四类问答技术的比较分析(2/2)

- **基于自然语言处理的问答技术**：可以对提问和答案文本进行一定程度的句法和语义分析，从而实现推理。但目前自然语言处理技术还不成熟，除一些浅层的技术（汉语分词、命名实体识别、词性标注等）外，其他技术还没有达到实用程度。所以这种技术的作用还有限，只能作为对前两种方法有效补充
- **基于统计翻译模型的问答技术**：在很大程度上依赖训练语料的规模和质量，而对于开放域问答系统，这种大规模训练语料的获取是非常困难的

小结

- 每种方法都有自身的优缺点，需要综合各种方法，可能是未来发展的方向
- 随着自然语言处理技术的发展，自然语言处理技术将会在问答式检索系统中得到更加广泛的应用

概述

- 问答系统概述
- 问答式检索系统
 - 问答式检索评测
 - 问答式检索方法
 - 问答式检索应用：Watson
- 社区问答系统
- 问题与挑战

问答式检索应用：Watson(1/6)

- 沃森 (Watson)：2011年，IBM研发的计算机“沃森”正在美国智力竞赛节目《危险边缘Jeopardy!》中上演“人机问答大战”，战胜人类选手
- DeepQA 问答系统是 Watson 实现的核心。



问答式检索应用：Watson(2/6)

□ 强大的硬件平台

- 90台IBM服务器、360个CPU组成，每个CPU主频可达4.1GHz
- 15TB内存、每秒可进行80万亿次运算
- Linux操作系统
- 分布式计算



问答式检索应用：Watson(3/6)

□ 强大的知识资源

- 存储了大量图书、新闻和电影剧本资料、辞海、文选和《世界图书百科全书》(World Book Encyclopedia)等海量的资料
- 每当读完问题的提示后，沃森就在不到三秒钟的时间里对资源库"挖地三尺"，在长达2亿页的漫漫资料里展开搜索

问答式检索应用：Watson(4/6)

□ 自然语言处理、信息抽取和知识工程技术

- 统计机器学习理论、**句法分析**(意大利特兰托大学 Giuseppe Riccardi和Alessandro Moschitti领导的团队)
 - 基于句法和语义结构的文本表示，对 IBM 沃森系统进行优化，并将基于Kernel的文本分析方法应用到IBM 沃森系统中
- **主题分析** (美国卡内基梅隆大学计算机科学学院语言技术研究所教授 Eric Myberg领导的团队)
 - 确定问题主题，根据主题确定给定主题的最佳文本资源
- **信息抽取**(美国麻省理工学院计算机科学及人工智能实验室首席研究科学家 Boris Katz领导的团队)
 - 对象-属性-值数据模型，该模型支持对半结构化数据源中的信息进行有效的检索，以回答自然语言问题

问答式检索应用：Watson(5/6)

- 自然语言处理、信息抽取和知识工程技术
 - 大规模知识库集成和知识推理技术(美国南加州大学南加州大学维特比工程学院信息科学研究所人类语言技术部门主任Eduard Hovy博士领导的团队)
 - 将大量不同来源的资料转化为该系统的知识资源，并利用这些知识进行推理，以发现答案的矛盾和差异之处
 - 信息检索技术(马萨诸塞大学安姆斯特分校 James Allan 教授领导的团队)
 - 问答技术的关键步骤：寻找和检索最可能包含准确答案的文字。然后，该系统的深度语言处理功能对返回的信息进行分析，以便在文字中寻找实际答案
 - 知识工程(纽约州立大学阿尔巴尼分校的 Tomek Strzalkowski 教授领导的团队)
 - 如何将多源异构知识统一集成和表示，方便后续的推理

问答式检索应用：Watson(6/6)

□ 沃森存在的问题

- Watson主要针对Jeopardy竞赛，回答的问题类型只限于简单的实体类问题，即所回答的答案大多数是实体或者是词语级别
- 对于答案不能用实体或者是词语表示的提问，例如描述一个过程的、表示评价的提问等，Watson处理得很少
- 从其官方的资料和比赛中的表现分析，Watson对于需要推理的复杂问题依然不能很好处理

小结

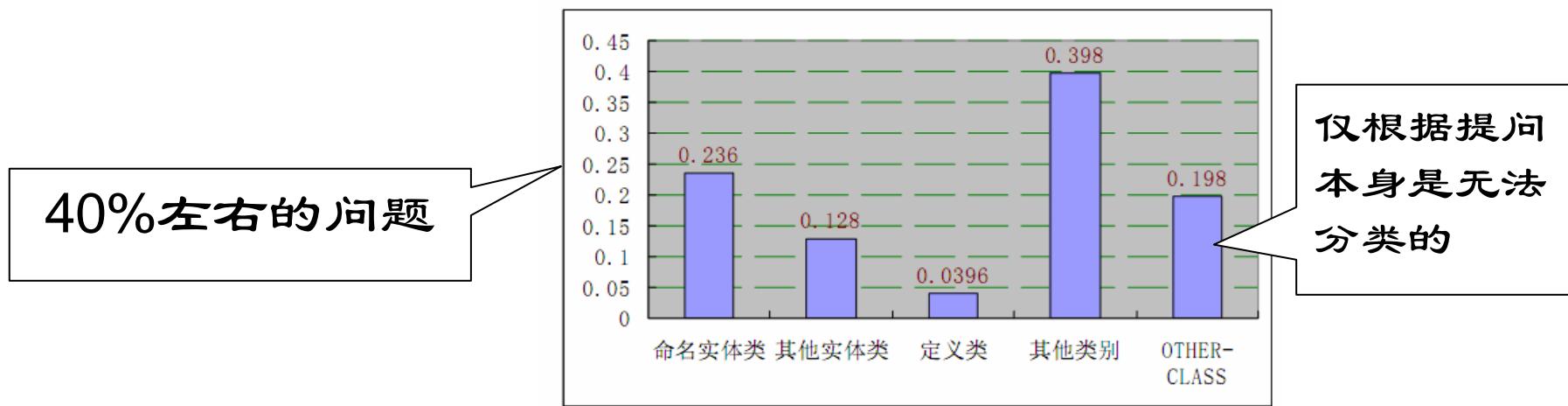
- 沃森(Watson)的强大问题回答能力体现：
 - 强大的硬件平台
 - 强大的知识资源
 - 自然语言处理、信息抽取和知识工程技术
 - 从沃森(Watson)官方公开的资料，涉及到的自然语言处理技术：
 - 句法分析
 - 信息抽取
 - 知识库构建
 - 知识推理
- 多种自然语言处理技术的集成和综合
 - 探索基于知识推理的问答与问答式检索系统的综合？

概述

- 问答系统概述
- 问答式检索系统
- 社区问答系统
 - 研究背景
 - 主要任务和方法
- 问题与挑战

传统问答式检索系统的问题

- 目前问答式检索系统仅能处理有限的简单的问题，如Factoid问题、List问题等
 - 而用户的提问是复杂多样的
- 答案类型集中于实体、定义等优先类别，后来TAC-QA扩展了观点类别，但是很多用户需求是没有明显类别的，目前的问答式检索系统还很难处理



社区问答的出现

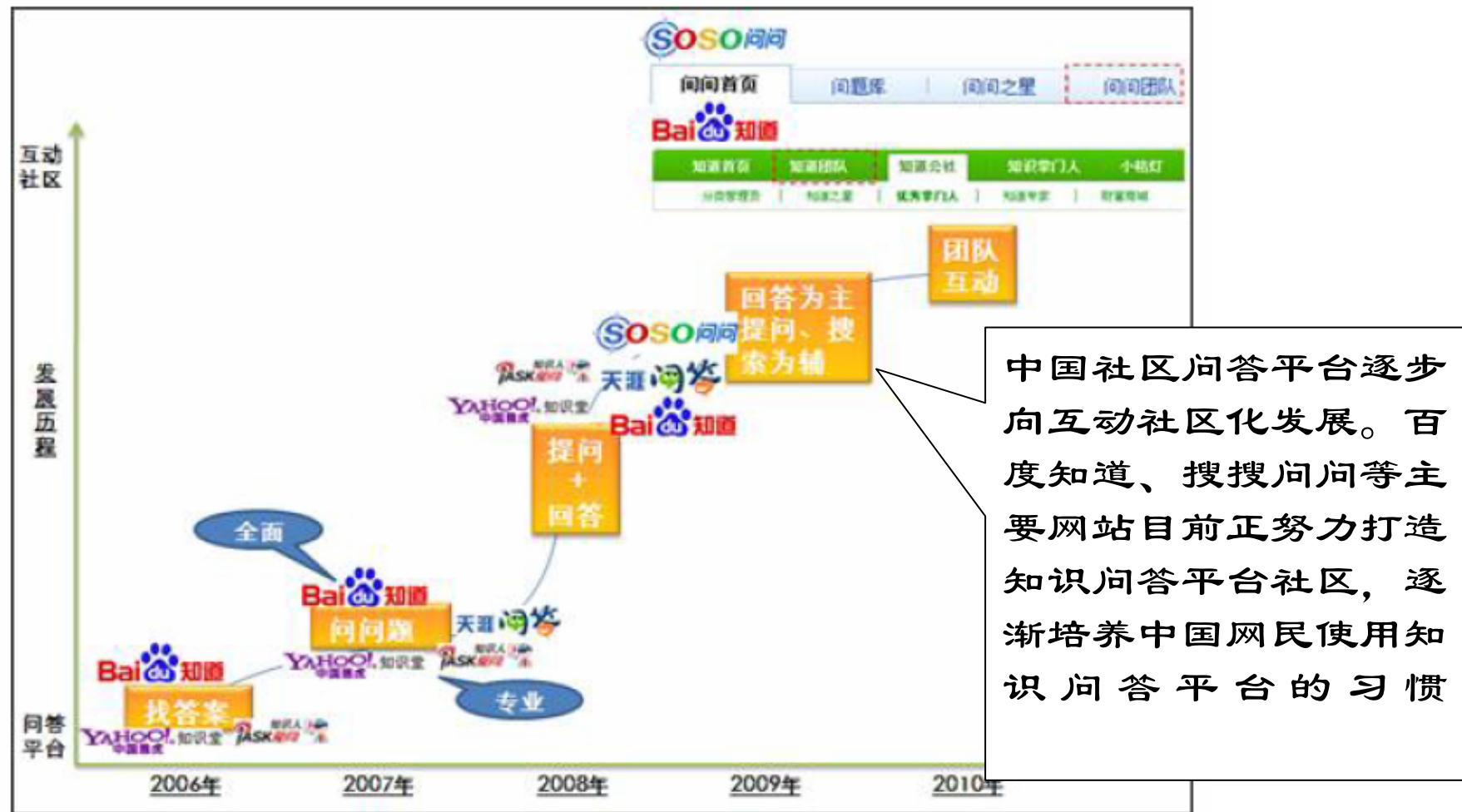
- 随着Web2.0 的兴起，基于用户生成内容的互联网服务越来越流行，社区问答服务的应运而生，为人们快速获取信息提供了更方便的选择
- 用户可以提出任何类型的问题，也可以回答其它用户的问题，区别于传统问答式检索系统仅能回答有限类型的问题
- 问题的类型和答案的类型不限定

社区问答系统的定义

- **社区问答系统**：指用户之间通过提出和回答问题的方式共享和积累知识，从而提供知识交流与信息服务的社会化系统(刘明荣 博士论文 2010)
- 从功能角度来看，社区问答系统包括三方面内容：
 - 用户通过自然语言问句方式表述信息需求
 - 用户通过回答问题方式响应其他用户的信息需求
 - 大量用户通过共同参与问答系统而形成了一个社区

网站名称	发布时间(年)	网址(http)
新浪爱问知识人	2004	iask.sina.com.cn
百度知道	2005	zhidao.baidu.com
雅虎知识堂	2006	ks.yahoo.com.cn
天涯问答	2007	wenda.tianya.cn
Naver	2002	kin.naver.com
WikiAnswers	2002	wiki.answers.com
Answerbag	2003	www.answerbag.com
Yahoo! Answers	2005	answers.yahoo.com

著名的社区问答网站（张中峰 博士论文 2011）



中国社区问答平台发展趋势(张中峰 博士论文 2011)

小结

- 社区问答系统作为社会媒体中一种，为人们的信息需求和知识分享提供了便利的条件
- 信息爆炸式的增长以及人们对信息需求的多样性，是社区问答系统迅速发展的重要因素
- 社区问答系统结合了搜索引擎和问答式检索系统的优点，又有自身的特点。对社区问答系统的研究，无论是从促进自然语言处理和信息检索技术发展的角度，还是从实际应用需求来看，都具有十分重要的价值

概述

- 问答系统概述
- 问答式检索系统
- 社区问答系统
 - 研究背景
 - 主要任务
- 问题与挑战

主要任务

□ 预处理

- 问题分类(很少)
- 作弊检测(很少)

□ 与回答新提交问题相关的研究

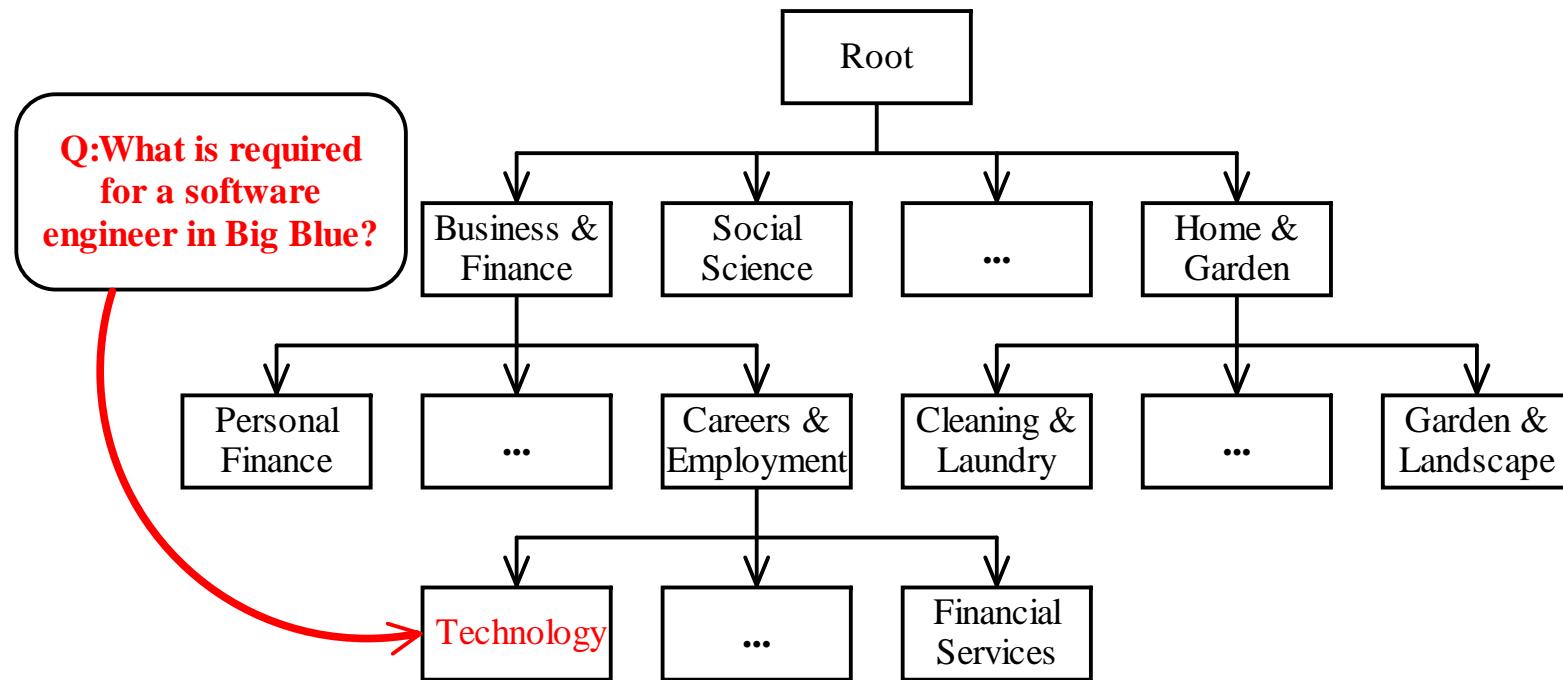
- 相似问题检索(很多)
- 答案质量评估(较多)
- 专家用户推荐及最佳回答者推荐(较少)
- 问句的主客观判断(很少)

□ 与用户体验相关的研究

- 用户满意度预测(较少)
- 潜在好友推荐(几乎没有)
- 用户社区结构挖掘(几乎没有)
- 热点话题检测(几乎没有)

问题分类(CAI, CIKM 2011) (1/9)

□ 任务：将用户提问自动分到社区问答系统对应的类别体系中



Yahoo! Answers 的类别体系结构

问题分类(CAI, CIKM 2011) (2/9)

□ 难点：

- 类别多且分布不均衡(Yahoo! Answers中类别：1,535)
- 用户的提问往往很短，包含的信息少

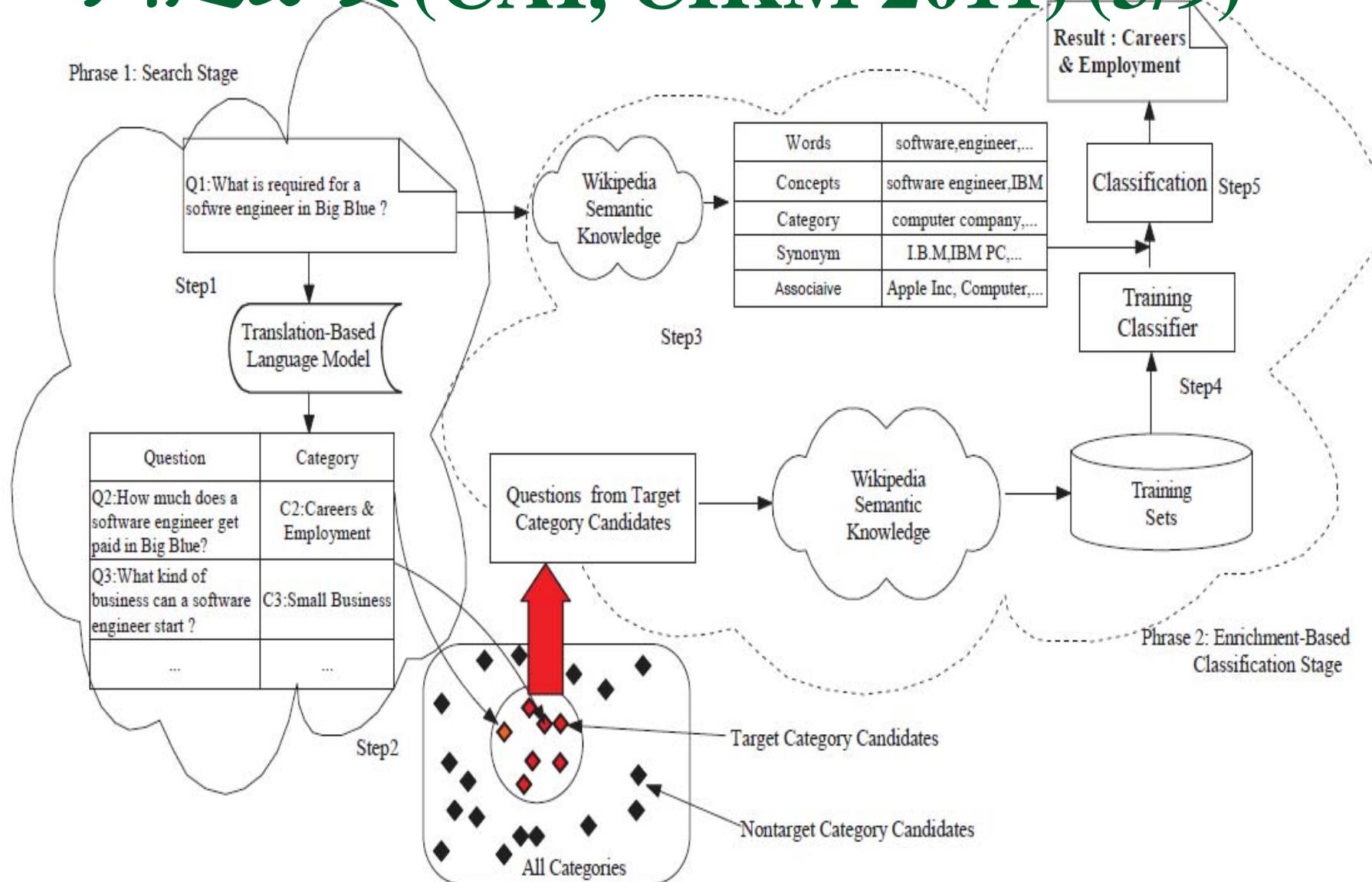
□ 问题：

- 准确率很低，导致马太效应
- 数据稀疏

用户的提问一般很短，
比如在Yahoo! Answers
中用户提问的平均长度
为11.2个词

借用传统文本分类和问
答式检索中问题分类的
方法，无法取得满意的
结果

问题分类(CAI, CIKM 2011) (3/9)



问题分类(CAI, CIKM 2011) (4/9)

Phrase 1: Search Stage

Result : Careers & Employment

基本出发点：

- 对于一个特定问题，先从完整的类别体系中找出一部分类别，作为问题分类的候选类别集合
- 如果用户的提问与历史问答对中的问题在语义上越相似，那么这些历史问题的类别很可能是该问题的候选类别

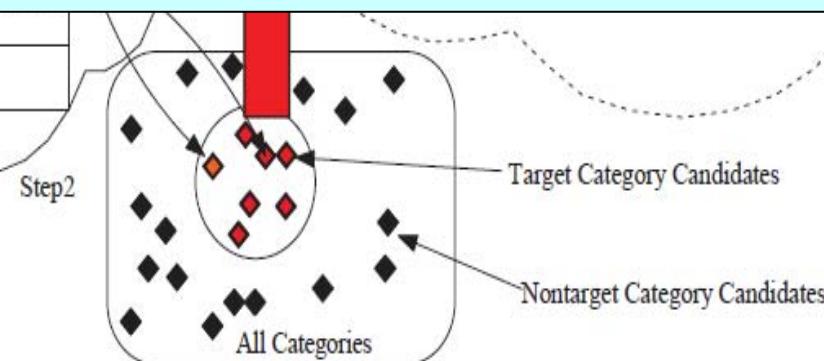
将一个大分类问题转化成了一个规模适度的分类问题

Q2:H
softw
paid it

Q3:W
busin

engineer start ?

Phrase 2: Enrichment-Based Classification Stage



问题分类(CAI, CIKM 2011) (5/9)

□ 剪枝技术：利用基于翻译的问题检索技术，将与待分类问题相关的问题从问题集合中检索出来，在训练分类器阶段仅考虑相关的问题集合，从而将一个大分类问题转化成了一个规模适度的分类问题

解决第一个难点问题

□ 查询扩展技术：利用Wikipedia对问题进行扩展，扩展的过程考虑了Wikipedia的上下位关系以及链接信息，有效缓解了数据稀疏性的问题

解决第二个难点问题

问题分类(CAI, CIKM 2011) (6/9)

#	Methods	Accuracy (%)
1	<i>BOW</i>	37.75
2	<i>Search_BOW</i>	46.90

剪枝技术的作用

问题分类(CAI, CIKM 2011) (7/9)

#	Methods	Strategy	Accuracy (%)
1	<i>Search_BOW</i>	-	46.90
2	<i>Search_BOW_SA</i>	Synonyms	43.88
		A_5	50.12
		A_{10}	49.91
		A_{15}	47.11
		A_{20}	46.35

利用关联关系进行扩展的作用

问题分类(CAI, CIKM 2011) (8/9)

#	Methods	Strategy	Accuracy (%)
1	<i>Search_BOW</i>	-	46.90
2	<i>Search_BOW_HR</i>	H_1	49.48
		H_2	48.64
		H_3	46.91
		H_4	45.83
		H_5	44.27

利用上下位关系进行扩展的作用

问题分类(CAI, CIKM 2011) (9/9)

#	Methods	Strategy	Accuracy (%)
1	<i>Search_BOW</i>	-	46.90
2	<i>Search_BOW_COB</i>	H_1, A_5	52.20

剪枝、关联扩展和上下位
扩展技术的作用

小结

- 问题分类是社区问答系统的关键因素，分类的好坏直接影响后续的问题检索
- 大规模问题分类是一个挑战性的问题，目前还很少有公开发表的工作，尚处在探索阶段

主要任务

□ 预处理

- 问题分类(很少)
- 作弊检测(很少)

□ 与解决新提交问题相关的研究

- 相似问题检索(很多)
- 答案质量评估(较多)
- 专家用户推荐及最佳回答者推荐(较少)
- 问句的主客观判断(很少)

□ 与用户体验相关的研究

- 用户满意度预测(较少)
- 潜在好友推荐(几乎没有)
- 用户社区挖掘(几乎没有)
- 热点话题检测(几乎没有)

相似问题检索(1/2)

- **任务**: 是一种基于大规模用户产生的问答数据集提供的信息检索服务，即从数据集中找出与用户提问相似的问题，这些相似问题的答案可以作为用户提问问题的答案候选

Search

what is the best way to prevent stuffy nose? Search Y! Answers

Sort by: [Relevance](#) | [Newest](#) | [Most Answers](#)

Free Symptom Checker Sponsor Results

Enter Your Symptoms Or Choose From A List To Find Out...
SymptomChecker.AARP.org

H Best way to prevent a sore throat/cold?
... to get a lil sore and my nose is gettn a little **stuffy** I was just wondering what can I eat/drink/do so that I don...sore throat or a cold.. Besides goin to the dr office.
★ In Infectious Diseases - Asked by John Stiller - 2 answers - 2 months ago

H whats the best way to get rid of a cold?
...night), a slight cough and a **stuffy nose**. i cant get a cold because of school. how do i **prevent** it from getting worse? i have **been** taking lots...echinacea, cold meds, etc..) **what are** some good **ways** to get rid of a cold?
★ In Infectious Diseases - Asked by Aly kat - 1 answer - 2 months ago

H WHAT IS THE BEST WAY TO..?
I got a **stuffy nose** just like 2 hours ago how can i **prevent** it from getting worse?
★ In Other - General Health Care - Asked by irockmiami - 3 answers - 8 months ago

用户查询

返回的相关问题

相似问题检索(2/2)

- **核心**: 计算两个问题的相似度
- **挑战**: 问题一般较短, 包含的信息很少, 词汇鸿沟问题很严重
- **主要方法有**:
 - 传统的信息检索模型, 比如VSM, BM25等
 - 语言模型(Zhai SIGIR 2001; Jeon CIKM 2005)
 - 基于词的翻译模型(Jeon CIKM 2005; Xue SIGIR 2008)
 - 基于短语的翻译模型(Zhou ACL 2011)
 - 其它(Duan ACL 2008; Wang SIGIR 2009; Cao WWW 2010)

语言模型(LM)(1/3)

□ Jeon CIKM 2005:

- 每个历史问答对对应一个统计语言模型
- 一个查询可以看作是由问答对的语言模型抽样产生的一个样本
- 可以根据每个问答对的语言模型抽样生成查询的概率来对其排序，概率值越大，该问答对就越满足要求

语言模型(LM)(2/3)

$$Score(\mathbf{q}, D) = \prod_{w \in \mathbf{q}} (1 - \lambda)P_{ml}(w|D) + \lambda P_{ml}(w|C)$$

一元语言模型

整个文档集上的平滑

$$P_{ml}(w|D) = \frac{\#(w, D)}{|D|}, \quad P_{ml}(w|C) = \frac{\#(w, C)}{|C|}$$

极大似然估计

Q表示一个查询，D表示一个问答对，C表示问答对的集合

语言模型(LM)(3/3)

□ 存在的问题：

- 词汇鸿沟问题：在社区问答系统中，由于问题一般较短，词汇鸿沟的现象很严重
- 词汇鸿沟：意义相关表述不同的词汇
比如：减肥与瘦身

查询：如何减肥？

相似问句：怎样瘦身？

不相似问句：减肥有好处么？

不相似问句：减肥有副作用吗？

基于词的翻译模型(WTM)(1/4)

□ Jeon CIKM 2005:

- 为了解决词汇鸿沟问题, Jeon 提出了基于词的翻译模型
- **基本思想:** 如果查询和候选问答对中的两个词不匹配, 可以利用统计翻译模型找到一些在语义上相关的词

和语言模型的方法类似, 只是把语言模型换成翻译模型

$$Score(\mathbf{q}, D) = \prod_{w \in \mathbf{q}} (1 - \lambda) P_{tr}(w|D) + \lambda P_{ml}(w|C)$$

翻译模型

查询中的词

$$P_{tr}(w|D) = \sum_{t \in D} P(w|t) P_{ml}(t|D), \quad P_{ml}(t|D) = \frac{\#(t, D)}{|D|}$$

问答对中的词

翻译概率

核心: 利用大量的问答对作为平行语料来训练问题和答案中的词的翻译模型, 获得翻译概率

基于词的翻译模型(WTM)(2/4)

- Xue SIGIR 2008:
 - Xue扩展了Jeon (CIKM 2005) 的工作，将语言模型与基于词的翻译模型做了一个线性组合

$$Score(\mathbf{q}, D) = \prod_{w \in \mathbf{q}} (1 - \lambda)P_{mx}(w|D) + \lambda P_{ml}(w|C)$$

$$P_{mx}(w|D) = \alpha \sum_{t \in D} P(w|t) P_{ml}(t|D) + (1-\alpha) P_{ml}(w|D)$$

翻译模型 语言模型

基于词的翻译模型(WTM)(3/4)

Source	everest			xp			search		
TTable	$P(A Q)$	$P(Q A)$	P_{pool}	$P(A Q)$	$P(Q A)$	P_{pool}	$P(A Q)$	$P(Q A)$	P_{pool}
1	everest	mountain	everest	xp	xp	xp	search	search	search
2	29,035	tallest	mountain	drive	window	window	google	information	google
3	ft	everest	tallest	install	computer	install	page	website	information
4	mount	highest	29,035	click	system	drive	list	free	internet
5	8,850	mt	highest	system	pc	computer	engine	info	website
6	feet	discover	mt	window	version	system	internet	internet	web
7	measure	hillary	ft	computer	edition	click	click	web	list
8	expedition	edmund	measure	pc	install	pc	web	address	free
9	height	mountin	feet	program	software	program	information	picture	info
10	nepal	biggest	mount	microsoft	98	microsoft	result	online	page

基于词的翻译模型得到的相关词

基于词的翻译模型(WTM)(4/4)

□ 存在的不足：

- 在翻译的过程中，没有考虑上下文信息

比如：java program中，基于词的翻译模型很容易将java翻译成coffee, island等词

□ 解决的办法：

- 基于短语的机器方法是否可以借用？
- 基于短语的翻译方法在**机器翻译**(Och, 2002; Koehn et al., 2003)、**拼写检查**(Sun et al., 2010)、**网络搜索**(Gao et al., 2010)

基于短语的翻译模型(PTM) (1/3)

□ Zhou ACL 2011:

- 为了解决基于词的翻译模型的不足, Zhou利用基于短语的翻译模型来找到一些在语义上相关的短语, 减少基于词翻译歧义带来的错误
- 算法的基本思想:
 - 短语切分: 将历史问答对D分割成一系列的短语E
 - 短语翻译: 将短语E逐个翻译得到一系列相关的短语F
 - 短语调序: 对所有的短语F做调序生成最终的查询q

基于短语的翻译模型(PTM) (2/3)

$$P(\mathbf{q}|D) \propto \sum_{(E,F,M) \in B(D,\mathbf{q})} P(E|D) \times P(F|D, E) \times P(M|D, E, F)$$

短语切分概率 短语翻译概率 短语调序概率

↓ 假设短语切分服从均匀分布

$$P(\mathbf{q}|D) \propto \sum_{\substack{(E,F,M) \in \\ B(D,\mathbf{q})}} P(F|D, E) \cdot P(M|D, E, F)$$

具体计算方法见论文

基于短语的翻译模型(PTM) (3/3)

□ 实验结果：

#	Methods	Trans Prob	MAP
1	Jeon et al. (2005)	P_{pool}	0.289
2	TransLM	P_{pool}	0.324
3	Xue et al. (2008)	P_{pool}	0.352
4	P-Trans ($\mu_1 = 1, l = 5$)	P_{pool}	0.366
5	P-Trans ($l = 5$)	P_{pool}	0.391

在问题检索中，基于短语的翻译模型优于
基于词的翻译模型

其它

- Duan (ACL 2008)考虑问题的结构信息: topic和focus
- Wang (SIGIR 2009)提出了一种基于句法树匹配的方法用于相似问题检索， 目标仍然是解决词汇鸿沟问题
- Cao (WWW 2010)将Category信息融入到检索模型中， 使问题检索的性能得到了显著提升

...

小结：相似问题检索

- 相似问题检索的核心是计算两个问题之间的相似度
 - 从文本计算层面：难点是如何解决词汇鸿沟的问题
 - 答案的质量：相似问题检索中需要考虑答案的质量（两个问题虽然语义上相关，但答案的质量不高，也不符合用户的要求）

主要任务

□ 预处理

- 问题分类(很少)
- 作弊检测(很少)

□ 与解决新提交问题相关的研究

- 相似性问题检索(很多)
- 答案质量评估(较多)
- 专家用户推荐及最佳回答者推荐(较少)
- 问句的主客观判断(很少)

□ 与用户体验相关的研究

- 用户满意度预测(较少)
- 潜在好友推荐(几乎没有)
- 用户社区挖掘(几乎没有)
- 热点话题检测(几乎没有)

答案质量评估(1/5)

- 社区问答作为一种社会媒体，数据是由用户自动生成的，造成质量差异的主要因素有：
 - 用户水平的高低
 - 自然语言表述上的多样性
- 通过对答案质量的自动分析，将全部候选答案按质量高低排序后展现给用户，可以节省用户的浏览时间，增强用户满意度

答案质量评估(2/5)

高质量

问题：谁比我帅？	问题：网上买手机应该注意什么？
答案：哈哈，我啊	答案：你买之前一定问好，有没有保修，保修凭证是什么。快递是什么，几天到。还有就是看店铺了，最好找皇冠以上店铺，问清楚，你要买的翻新机有什么问题，可能有什么问题。
问题：What can I do tonight?	问题：How to get from north london to thorpe park?
Sleep!!!	答案：Take the Underground to Waterloo Station. From Waterloo take an SWT Reading or Windsor train to Staines Station. They run four times per hour and the journey takes 35 minutes.

低质量

答案质量评估(3/5)

- 主要方法：

- 采用统计机器学习方法，比如分类或回归等

- 核心：如何选取特征

- 文本特征
 - 非文本特征

答案质量评估(4/5)

- Jeon SIGIR 2006:
 - 采用非文本化特征对答案质量预测：
 - 回答者的问题采纳率
 - 答案长度(也可以称为是文本化的特征)
 - 用户推荐的次数
 - 页面点击次数
 - ...
 - 采用最大熵分类器作预测

答案质量评估(5/5)

□ Agichtein WSDM 2008:

- 通过组合不同类型的文本化和非文本化特征, 以Yahoo! Answers作为实验数据, 采用随机梯度增强树(stochastic gradient boosted trees)作为分类框架, 找出社会化媒体中的高质量内容
- 采用的特征:
 - 内容特征
 - 用户之间的关系
 - 使用特征(点击次数)

特征	因素	维度	代表工作
文本	答案	准确性	Harper 2008; Blooma 2008
		完整性	Blooma 2008
		可读性	Zhu 2009
		合理性	Blooma 2008
		相关性	Blooma 2008; Zhu 2009
		新颖性	Zhu 2009
		内容长度	Jeon 2006; Lee 2007
非文本	提问者	权威性	Blooma 2008
	回答者	权威性	Jeon 2006; Bian 2009
	问题	主题类别	Jeon 2006
	答案	时间、位置等	Jeon 2006; Bian 2009
	问答页面	点击信息等	Jeon 2006; Agichtein 2008

小结：答案质量评估

- 质量评估问题是一个非常主观的工作
- 目前的主流方法是采用统计机器学习方法，将其看成是一个分类或回归的问题，关键是如何选择特征
- 难点：主观性较强，难以评测或对比

主要任务

□ 预处理

- 问题分类(很少)
- 作弊检测(很少)

□ 与解决新提交问题相关的研究

- 相似性问题检索(很多)
- 答案质量评估(较多)
- 专家用户推荐及最佳回答者推荐(较少)
- 问句的主客观判断(很少)

□ 与用户体验相关的研究

- 用户满意度预测(较少)
- 潜在好友推荐(几乎没有)
- 用户社区挖掘(几乎没有)
- 热点话题检测(几乎没有)

专家用户发现及最佳回答者推荐

□ 专家用户发现（回答的准确）

- 在社区问答系统中，专家用户给出的答案更有可能被提问者或者其他投票者选为最佳答案

□ 最佳回答者推荐（回答的既准确又及时）

- 在社区问答系统中，对于用户提出的问题能够给予及时准确答案的用户
 - 要求用户具有领域专业知识(用户的兴趣)
 - 要求用户经常上线(很活跃)
- ...
- 专家用户不一定是最佳回答者；最佳回答者一般是专家用户

专家用户发现(1/2)

- 任务：在一个社区问答系统中，找出哪些用户相对更加权威
 - 用户回答问题的情况
 - 用户提问的情况
- 这里“专家用户”是个相对概念，有时候也称为“权威用户”

专家用户发现(2/2)

□ 主要方法：

- 基于图结构的专家用户发现(Jurczyk CIKM 2007)
- 基于启发式的专家用户发现(Bouguessa KDD 2008)
- 其它: (Liu CIKM 2005; Pal CIKM 2010; Kao SAC 2010)

基于图结构的专家用户发现(1/3)

□ Jurczyk CIKM 2007:

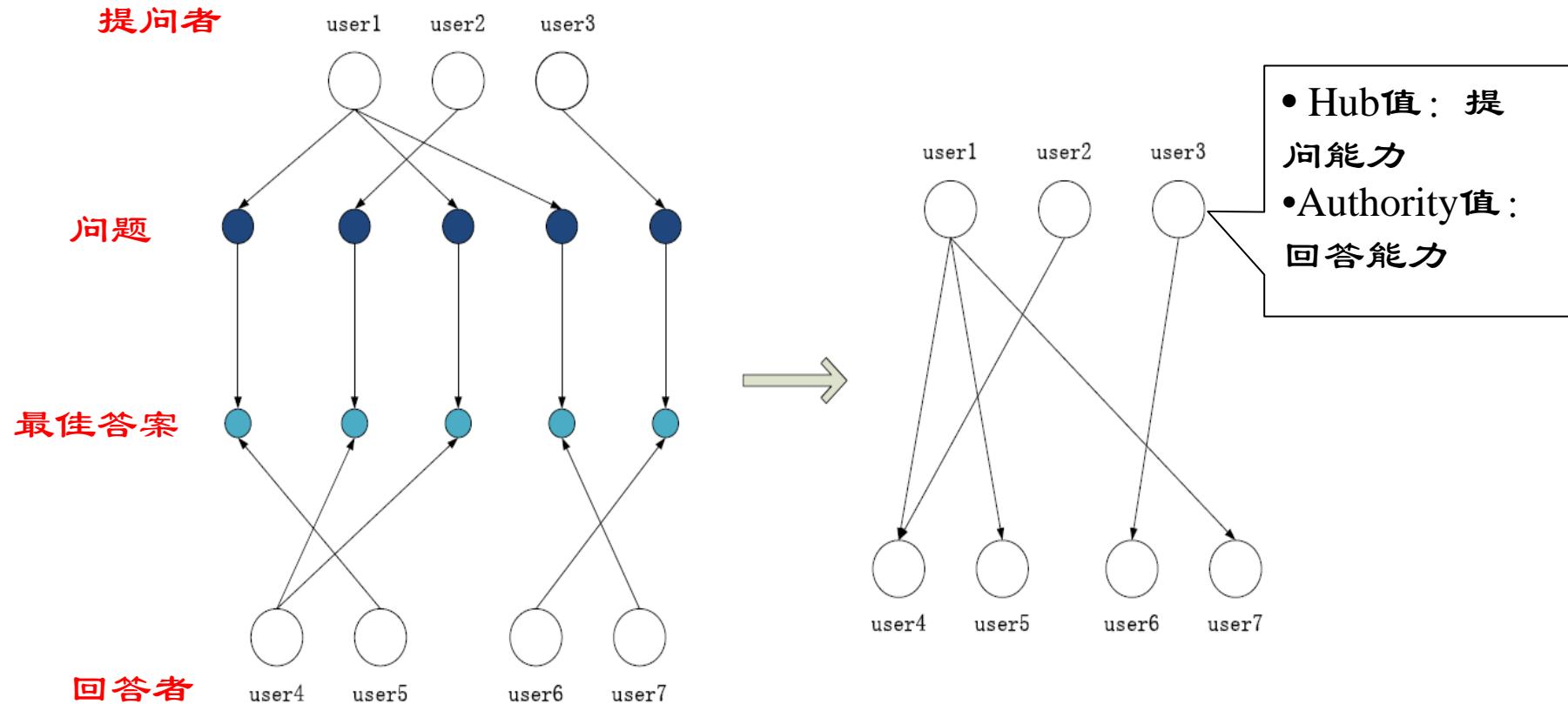
□ 出发点：

- 用户回答的问题越多，越有可能成为专家用户
- 用户提出的问题越多，越有可能成为专家用户

□ 方法：采用类似网页排序任务中的PageRank算法和HITS算法

- 通过问答建立用户之间关系，如果一个用户对另一个用户提出的问题做了最佳回答，那么这两个用户之间就存在一条有向边
- 由于一个用户既可以提出问题又回答许多问题，整个图呈网状结构
- 对每个顶点计算Hub值（出度）和Authority值（入度），分别对应用户的提问能力和回答能力

基于图结构的专家用户发现(2/3)



基于图结构的专家用户发现(3/3)

- 顶点的Hub值为其所有指向它的顶点的Authority值的加权和

$$Hub(i) = \sum_{j=1, \dots, N} w(i, j) Authority(j)$$

- 顶点的Authority值为其所有指向它的顶点的Hub值的加权和

$$Authority(j) = \sum_{i=1, \dots, M} w(i, j) Hub(i)$$

基于启发式的专家用户发现(1/2)

- Bouguessa KDD 2008:
 - 出发点：用户提供的最佳答案数越多，其权威度就越高
 - 方法：
 - 根据提问和回答的关系，建立用户之间的联系，构成一张有向图
 - 通过计算顶点的入度(指向顶点的加权边的和)表示问答系统中用户的权威度

基于启发式的专家用户发现(2/2)

- 衡量用户的权威度：

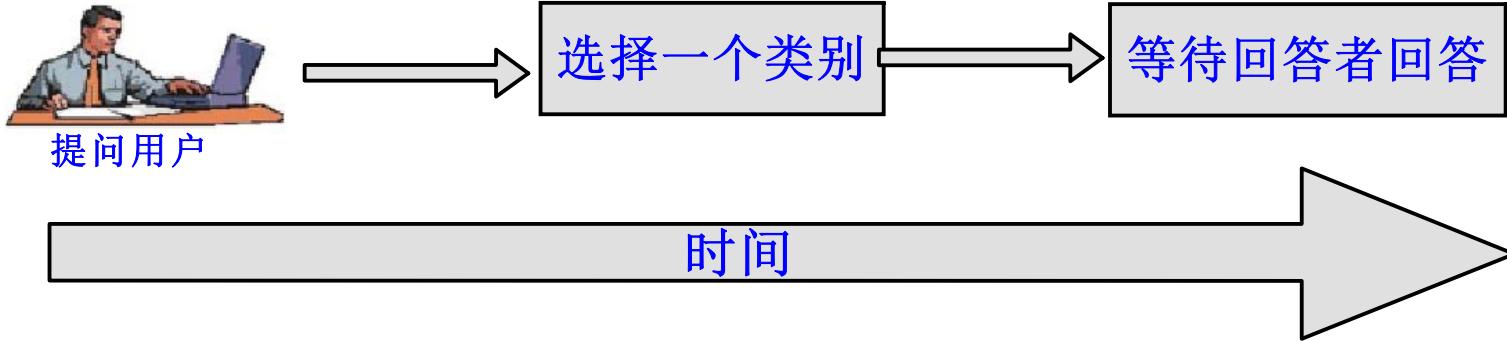
$$authority(u) = \log(1 + numans_u)$$

- 其中 $numans_u$ 表示用户 U 在问答系统提供的最大答案数
- 使用对数值是为了平滑其对结果的影响程度
- 与 Jurczyk CIKM 2007 相比，该研究只考虑出度信息，没考虑入度信息，反而取得了较好的结果

小结

- 专家用户发现是社区问答系统中的一个挑战性难题，目前的研究工作主要是将该问题转化成一个二分图问题，利用图算法计算节点的权重大大小（权威度大小），只考虑了用户的提问能力和回答能力
- 此外，用户的rating信息对专家用户发现可能是有帮助的，难点是如何解决数据的稀疏性(目前社区问答系统中仅有30%的问题有rating信息(Adamic WWW 2008))

最佳回答者推荐

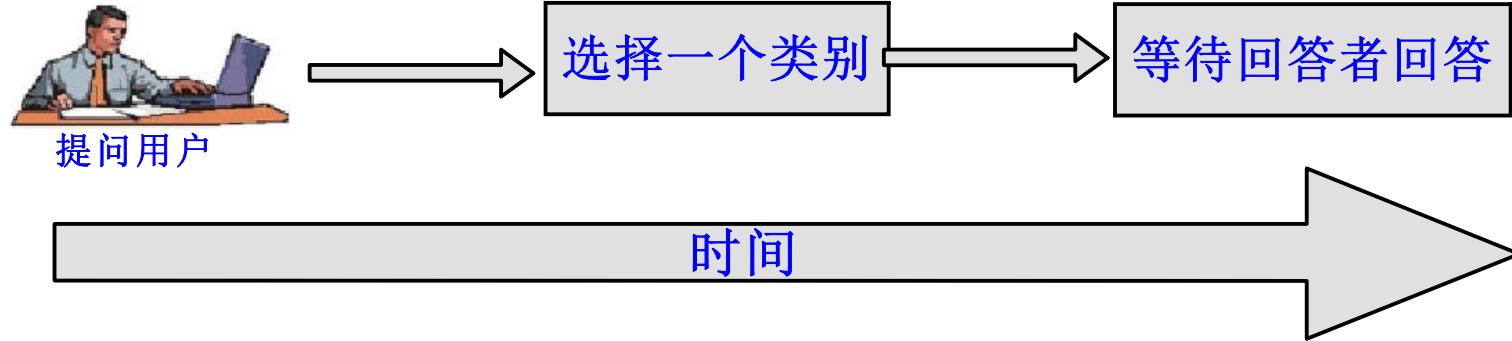


社区问答系统中用户参与的过程

□为什么要这么做（从提问者角度）

- 在当前的问答系统中，提问者必须被动等待其他用户访问系统、浏览其提出的问题、然后提供问题的答案
- 提问者从提出问题到问题获得解答可能需要等待几分钟、几个小时甚至几天的时间，依赖于回答者的参与程度
- 提问者可能希望系统中的专家用户回答他们的问题，因为专家提供的答案为正确的可能性相对更高

最佳回答者推荐 (1/4)



社区问答系统中用户参与的过程

□ 为什么要这么做（从回答者角度）

- 一个回答者可能对某个特定问题非常了解，但是他可能没有访问到该问题，或者该问题被大量其它问题所淹没
- 某个用户对问题做了回答可能是因为他碰巧看到了这个问题，并不一定是该问题的最佳回答者

最佳回答者推荐 (2/4)

- 最佳回答者推荐考虑的因素
 - 用户兴趣建模（回答者，内容角度）
 - 基于用户资料（每个用户提供的最佳答案的集合），建立用户和已回答问题之间的关系
 - 用户的行为信息（回答者，行为角度）
 - 用户的权威度（专家用户发现）
 - 用户的活跃度

最佳回答者推荐 (3/4)

□ 一般性框架：

- 对给定的新的提问 q , 从用户集 $\mathcal{U} = \{u_1, u_2, \dots, u_M\}$ 中寻找一个用户 u_i , 使得 u_i 对 q 提供最佳答案的可能性最大

$$\arg \max_i P(u_i|q), i = 1, 2, \dots, M$$

□ 利用贝叶斯公式：

$$P(u|q) = \frac{P(u)P(q|u)}{P(q)}$$

用户的 behavior 信息 用户兴趣模型
常数

最佳回答者推荐 (4/4)

□ 用户兴趣建模：

- 基于语言模型的方法(Li CIKM 2010)
- 基于类别语言模型的方法(Li CIKM 2011)
- 基于聚类的方法(Zhou ICDE 2009)
- 基于主题模型的方法(Liu WAIM 2010)

□ 用户的行为信息 (刘明荣 博士论文 2010)：

- 用户的权威度
- 用户的活跃度

用户兴趣建模：基于语言模型的方法

- Li CIKM 2010:

- 基于语言模型的用户兴趣建模：

$$P(q|u) = P(q|q_u)$$

用户提供最佳答
案的问题集合

$$P(q|q_u) = \prod_{w \in q} P(w|q_u)$$

- 由于用户提供最佳答案的问题数相对较少，导致数据稀疏：

$$P(w|q_u) = (1-\lambda)P_{ml}(w|q_u) + \lambda P_{ml}(w|Coll)$$

Jelinek-Mercer 平滑

用户兴趣建模：基于类别语言模型方法

- Li CIKM 2011 : (利用问题所在的类别，是社区问答系统已有的类别体系中的类别)
- 由于各个类别包含的问题数分布均衡，为了提高推荐的性能 提出基于类别的语言模型

$$P(q|u) \downarrow P(q, c|u) = P(q|c, u) \underbrace{P(c|u)}_{\text{极大似然估计}}$$

$$P(q|c, u) = P(q|c, q_u) = \prod_{w \in q} P(w|q_{uc})$$

$$P(w|q_{uc}) = (1 - \lambda)P_{ml}(w|q_{uc}) + \lambda P_{ml}(w|Coll)$$

用户兴趣建模：基于聚类的方法

- Zhou ICDE 2009 : (问题自动聚类的类别)
 - 对用户提出的所有问题按照不同的主题进行聚类，利用主题信息对用户推荐行为做约束

$$P(q|u) = \sum_{Cluster} \prod_{w \in q} P(w|\theta_{Cluster})^{n(w,q)} \boxed{con(Cluster, u)}$$

用户对类的贡献

参数估计的方法见论文

用户兴趣建模：基于主题模型的方法

- Liu WAIM 2010: (LDA对问题进行聚类)

- 将用户推荐问题转化成问题检索问题，为了克服词汇鸿沟的问题，将主题模型与语言模型进行线性组合

$$P(q|q_u) = \prod_{w \in q} P(w|q_u)$$

解决词汇鸿沟

$$P(w|q_u) = \delta P_{LDA}(w|q_u) + (1 - \delta)P_{LM}(w|q_u)$$

$$P_{LDA}(w|\hat{\theta}, \hat{\gamma}, q_u) = \sum_{z=1}^Z P(w|z, \hat{\gamma})P(z|\hat{\theta}, q_u)$$

参数估计

用户的行为信息(1/2)

□ 用户活跃度的表现形式

- 从提问者的角度来看，当其提出一个新的问题后，除了期望能够得到正确的答案，同时通常会期望该问题被解决的时间延迟越短越好
- 从回答者的角度看，系统的用户活跃度变化非常大，有些用户在一段很长的时间内都是活跃的，而另一些用户可能只是偶尔会进入社区，甚至很长时间都不提供答案

用户的行为信息(2/2)

- 用户的活跃度受很多因素影响，很难建立长期的用户活跃度模型(刘明荣 博士论文 2010)

$$activity(u) = \exp^{-(t_q - t_u)}$$

从提问发生到用户提供答案的时间间隔

- 其中, t_q 是用户提交问题的时间
- t_u 是用户u最近一次提供答案的时间

- 用户行为信息：

权威度

活跃度

$$P(u) = authority(u) \times activity(u)$$

最佳回答者推荐

□ 用户推荐的预测：

$$P(u|q) \propto P(u)P(q|u)$$

$$\propto [authority(u) \times activity(u)] \times [\prod_{w,w \in q} P_X(w|\theta_u)^{n(w,q)}]$$

用户行为信息

用户兴趣模型

小结

□ 专家用户发现

- 将其转化成一个图算法问题。利用二分图的相关算法计算每个节点的重要程度，目前尚处在探索阶段

□ 最佳回答者推荐

- 影响专家推荐的因素主要包括：
 - 用户兴趣建模：
 - 目前主要根据用户的历史提问和回答来建模
 - 用户行为建模：
 - 影响用户行为的因素很多，也很难把握。如何描述和计算用户的
行为是一个挑战性的工作，目前尚处在研究探索阶段

概述

□ 问答系统概述

- 研究背景
- 发展历史

□ 问答式检索系统

□ 社区问答系统

□ 问题与挑战

问题与挑战(1/3)

- 问答系统已经取得了一定的进展：
 - 问题的分析
 - 自然语言处理技术的发展
 - 检索技术
 - 检索模型的发展
 - 理论完备的语言模型
 - 单语言翻译模型
 - 知识库的构建
 - 由限定领域知识库向开放领域知识库发展
 - 由规模较小的单一知识库向网络知识库发展

问题与挑战(2/3)

□ 尚未解决的问题：

- 问题分析
 - 对复杂问题的处理尚处在探索阶段
- 检索技术
 - 目前的方法主要集中如何计算相似度，无论是哪种方法都存在一定的不足，检索结果还不理想
- 知识库的构建
 - 规模还较小，尚无法覆盖所有的问题
 - 网络知识库的质量不高

问题与挑战(3/3)

□ 从沃森(Watson)的成功看问答系统的发展：

- 强大的硬件平台
- 强大的知识资源
- 深层的自然语言处理技术

各种资源和方法的综合

□ 问答系统的未来

- 基于知识推理的问答系统
- 问答式检索系统
- 社区问答系统

各种资源和方法的综合

参考文献

- E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In WSDM 2008.
- L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge sharing and Yahoo! Answers: everyone knows something. In WWW 2008.
- J. Bian, Y. Liu, D. Zhou, E. Agichtein, and H. Zha. Learning to recognize reliable users and content in social media with couple mutual reinforcement. In WWW 2009.
- M. Bouguessa, B. Dumoulin, and S. Wang. Identifying authoritative actors in question-answering forums: the case of Yahoo! Answers. In KDD 2008
- A. Berger and R. Caruana, and D. Cohn, D. Freitag, and V. Mittal. Bridging the lexical chasm: statistical approach to answer finding. In SIGIR 2000.
- M. Blooma, A. Chuu, and D. Goh. A predicting framework for retrieving the best answer. In SAC 2008.
- L. Cai, G. Zhou, K. Zhao, and J. Zhao. Large-scale question classification in cQA by leveraging wikipedia semantic knowledge. In CIKM 2011.
- F. Duclay and F. Yvon. Learning paraphrases to improve a question answering system. In EACL 2002
- J. Gao, X. He, and J. Nie. Clickthrough-based translation models for web search: from word models to phrase models. In CIKM 2010.
- F. Harper, D. Raban, S. Rafaeli, and J. Konstan. Predictors of answer quality in online Q&A sites. In SIGCHI 2008
- F. Harper, D. Raban, S. Rafaeli, and J. Konstan. Predictors of answer quality in online Q&A sites. In SIGCHI 2008

参考文献

- P. Jurczyk and E. Agichtein. Discovering authorities in question answer communities by using link analysis. In CIKM 2007.
- J. Jeon, and W. B. Croft, and J. H. Lee. Finding similar questions in large question and answer archives. In CIKM 2005.
- J. Jeon, W. B. Croft, J. H. Lee, and S. Park. A framework to predict the quality of answers with non-textual features. In SIGIR 2006.
- P. Koehn. F. Och, and D. Marcu. Statistical phrase-based translation. In NAACL 2003.
- W. Kao, D. Liu, and S. Wang. Expert finding in question-answering websites: a novel hybrid approach. In CIKM 2010.
- M. Liu, Y. Liu, and Q. Yang. Predicting question answerers for new questions in community question answering. In WAIM, 2010.
- Baichuan Li, Irwin King, Michael R. Lyu. Question routing in community question answering: putting category in its place. In CIKM 2011.
- J. -T Lee, S. -B. Kim, Y. Song, and H. Rim. Bridging lexical gaps between queries and questions on large online Q&A collections with compact translation models. In EMNLP 2008.
- J. T. Lee, Y. I. Song, and H. C. Rim. Predicting the quality of answers using surface linguistics features. In ALPWI 2007.
- X. Liu, W. Croft, and M. Koll. Finding experts in community-based question-answering services. In CIKM 2005
- D. K. Lin and P. Pantel. Discovery of inference rules for question-answering. In NLE 2001.
- R. Mihalcea and P. Tarau. TextRank: Bringing order into text. In EMNLP.

参考文献

- F. Och. Statistical machine translation: from single word models to alignment templates. Ph.D thesis, 2002.
- A. Pal and J. A. Konstan. Expert identification in community question answering: exploring question selection bias. In CIKM 2010.
- X. Sun, J. Gao, D. Micol, and C. Quirk. Learning phrase-based spelling error models from clickthrough data. In ACL 2010.
- X. Xue, J. Jeon, and W. B. Croft. Retrieval models for question and answer archives. In SIGIR 2008
- H. Yang and T.-S. Chua. The integration of lexical knowledge and external resources for question answering. In TREC 2002.
- G. Zhou, L. Cai, J. Zhao, and K. Liu. Phrase-based translation model for question retrieval in community question answer. In ACL-HLT 2011.
- C. Zhai and J. Laffery. 2001. A study of smooth methods for language models applied to ad hoc information retrieval. In SIGIR 2001.
- Y. Zhou, G. Cong, B. Cui, C. S. Jensen, and J. Yao. Routing questions to the right users in online communities. In ICDE, 2009
- Z. M. Zhu, D. Bernhard, and I. Gurevych. A multi-demensional model for assessing the quality of answers in social Q&A sites. In IQ 2009
- 刘明荣. 协作式问答系统关键技术研究. 中国科学院自动化研究所博士论文, 2010.

参考文献

吴友政. 问答系统关键技术研究. 中国科学院自动化研究所博士论文. 2006.

张中峰. 社区问答系统中主题及用户社区挖掘的关键技术研究. 中国科学院自动化研究所博士论文. 2011.

Q&A

Thanks

第四课

依存句法分析

中国科学院自动化研究所

模式识别国家重点实验室

纲要

- 第一部分：概述
- 第二部分：方法
 - 基于动态规划的方法
 - 基于决策的方法
 - 基于融合的方法
 - 扩展性工作
- 第三部分：问题与挑战

概述：定义

- 依存句法的基本思想：句法结构由词汇组成，词汇之间由二元非对称关系连接起来，这些关系叫作依存关系(Lucien Tesniere, 1959)
- 依存句法结构的定义(Nivre, 2005):

The sentence is an *organized whole*, the constituent elements of which are *words*. Every word that belongs to a sentence ceases by itself to be isolated as in the dictionary. Between the word and its neighbors, the mind perceives *connections*, the totality of which forms the structure of the sentence. The structural connections establish *dependency* relations between the words. Each connection in principle unites a *superior* term and an *inferior* term. The superior term receives the name *governor*. The inferior term receives the name *subordinate*.

概述：定义

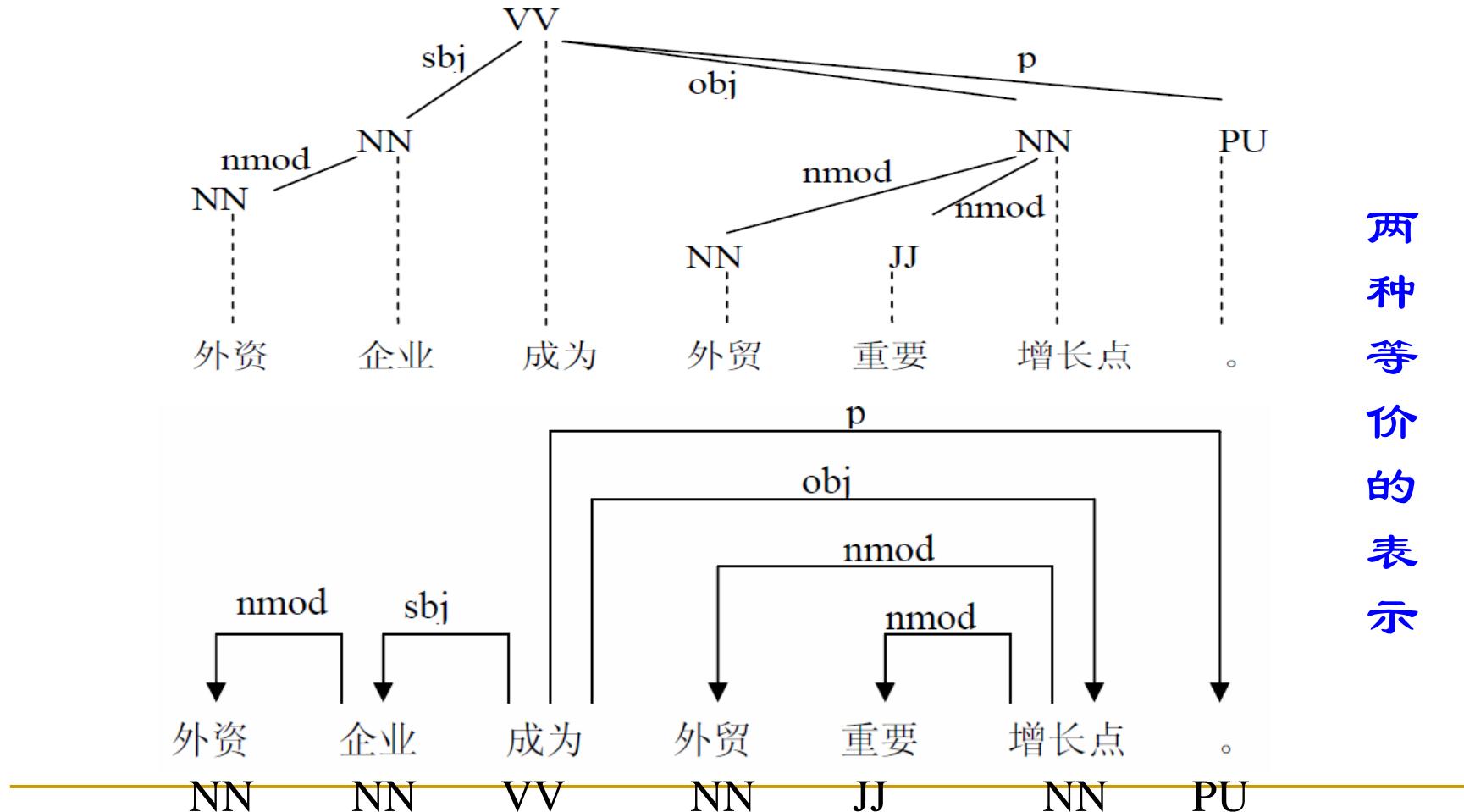
- 依存句法的基本思想：句法结构由词汇组成，词汇之间由二元非对称关系连接起来，这些关系叫作依存关系(Lucien Tesniere, 1959)
- 依存句法结构的定义(Nivre, 2005):

The sentence is a well-organized whole, whose elements are words. The parts connected by the structure establish dependencies between words. Every part connected by the structure is原则上将一个上级词与一个下级词联系起来. The upper word is called the governor and the lower word is called the subordinate. The governor receives the name *governor*. The inferior term receives the name *subordinate*.

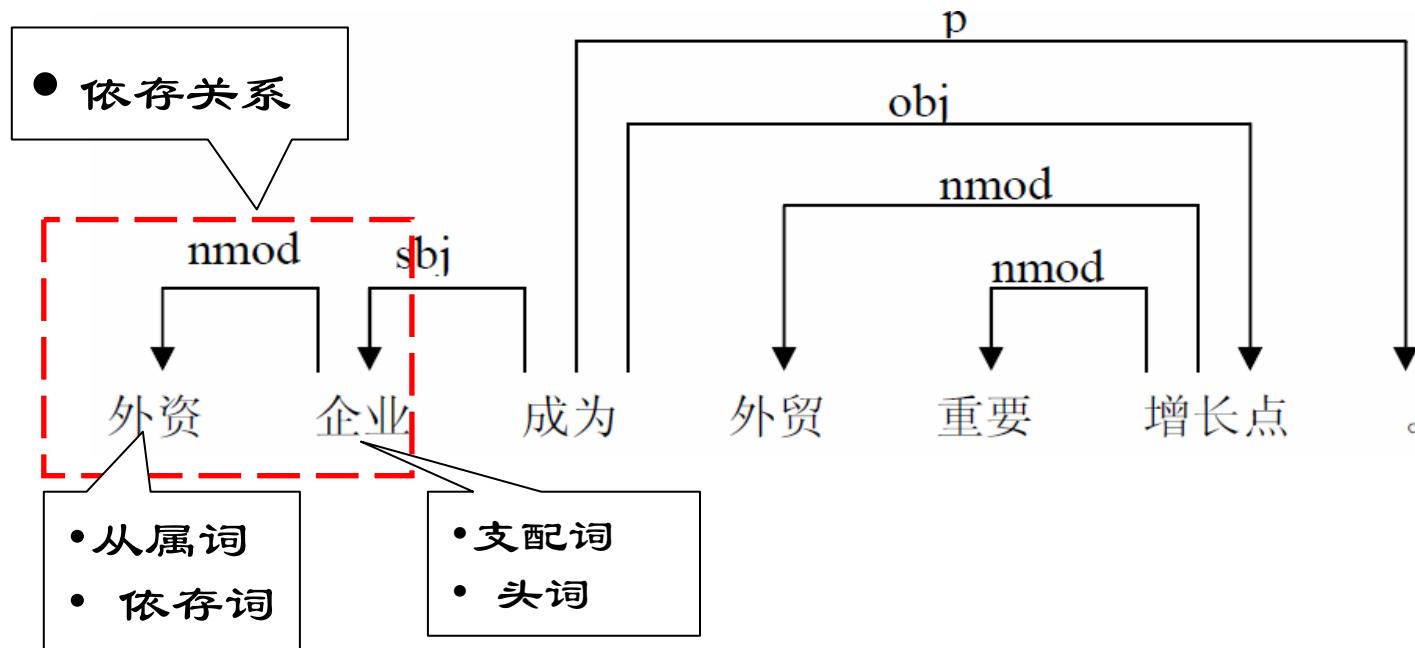
- 句子是一个有组织的整体，其组成元素是词
- 结构中相连的部分建立了词之间的依存关系
- 每一个相连的部分原则上将一个上级词与一个下级词联系起来
- 上级词称为支配词，下级词称为从属词

概述：依存句法结构的表示——依存图

□ 例句：外资企业成为外贸重要增长点。



概述：依存关系



概述：依存图的形式定义

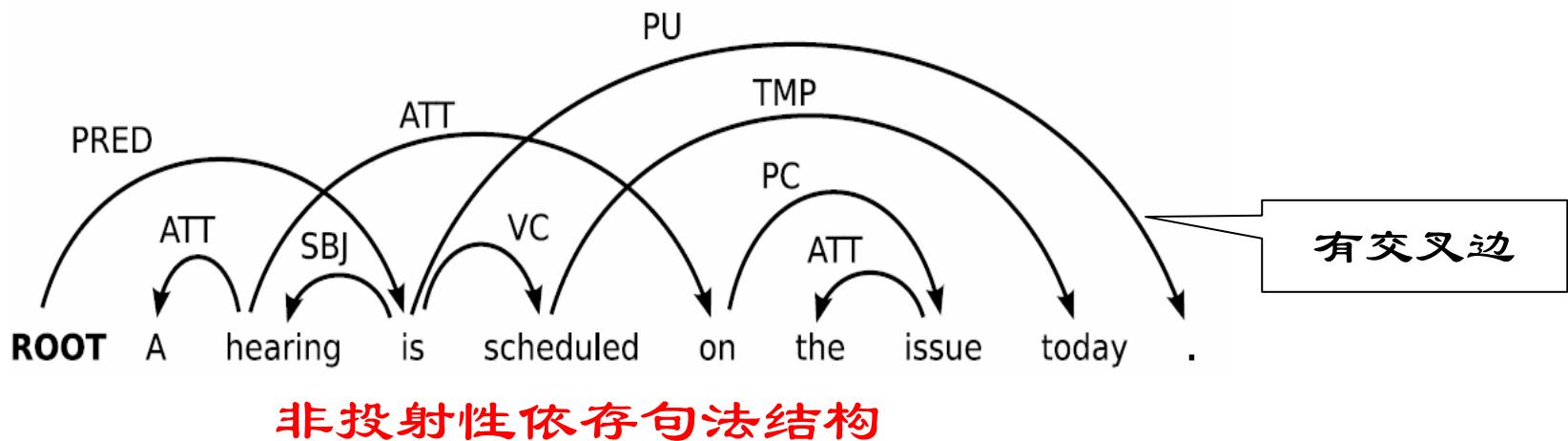
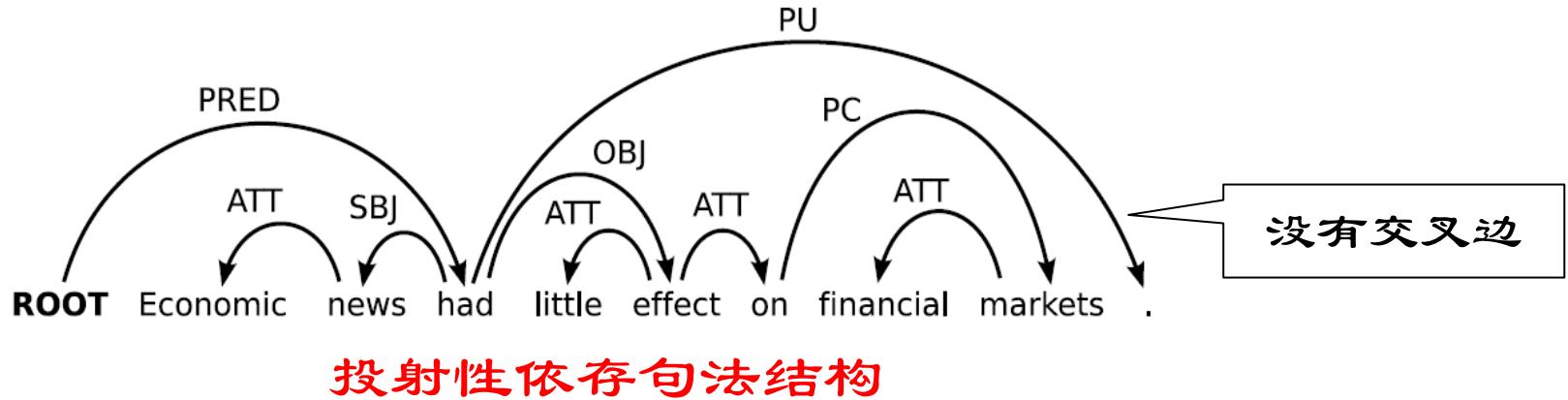
□ 有向图 $G=(V, E, L)$: 结点、边、标记

- $V=\{1, \dots, n\}$ 表示节点的集合
- $E \subseteq V \times V$ 表示边的集合, e 关于 V 的非对称关系 $<$
- 带有标记的依存图是指依存边 E 上带有依存类型标记 L

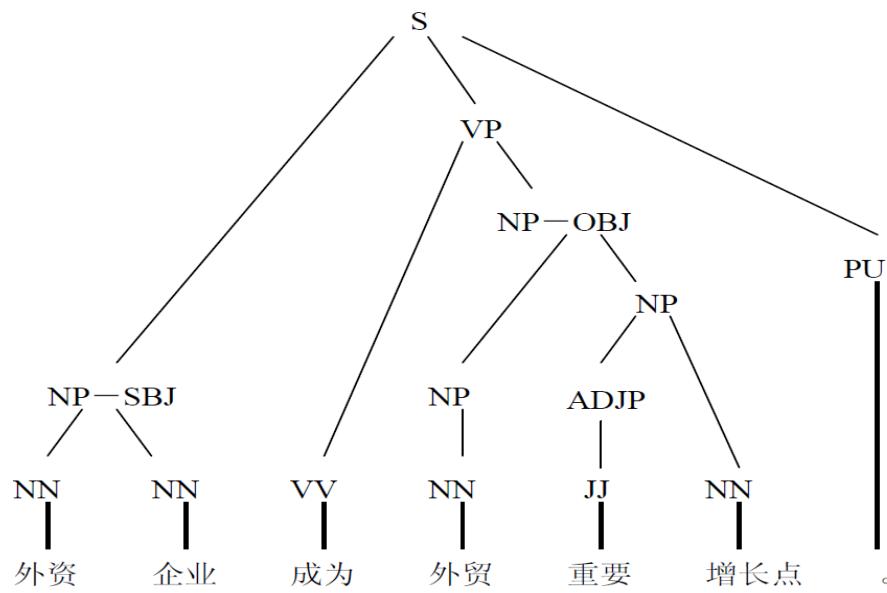
依存句法分析概述：依存图

- 依存图G满足以下四个基本属性：
 - **连通性**：任何两个节点都是相互可达的
 - **无环**：G中不包含任何环
 - **每个节点只有一个头词**：任何一个节点不存在两个或两个以上的父节点
 - **投射性**：G中不包含任何交叉边（在汉语和英语中投射性现象很少，但在德语和荷兰语中非投射现象较多）

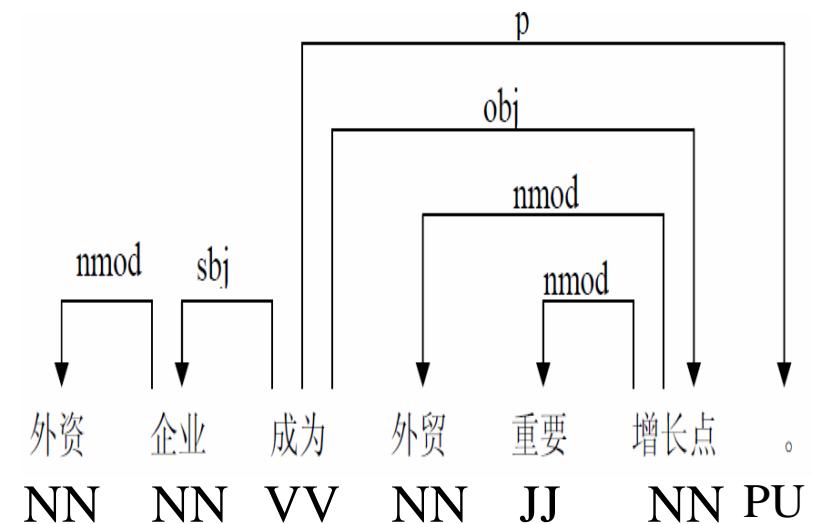
投射性 vs. 非投射性依存句法结构



依存语法 vs. 短语结构语法(1/2)



短语结构语法



依存语法

依存语法 vs. 短语结构语法(2/2)

- 简单，没有非终结符
- 依存语法关注句中词与词之间的关系，短语结构语法关注成分结构
- 短语结构语法可以很容易地转化为依存语法；反之，则较为困难
- 依存句法分析的时间复杂度较低 (Eisner, 1996; McDonald, 2006)

评测

□ CoNLL 2006任务：

- 多语言的依存句法分析：要求参与者在13种不同的语言上训练单一的依存句法分析器，用以比较单一依存句法分析器在多语言上的推广能力

□ CoNLL 2007任务：

- 继续对多语言依存句法分析进行评测
- 依存句法分析的领域自适应：源领域是WSJ标注语料，目标领域是生物医学摘要、化学摘要、父母子女对话语料

评测

- CoNLL 2008任务：

- 依存句法分析和语义角色标注的联合学习
 - 评测只考虑英文

- CoNLL 2009任务：

- 继续依存句法分析和语义角色标注的联合学习
 - 评测考虑7种语言

评价指标

- 无标记依存正确率(UAS): 所有词中找到正确的头词所占的百分比，对于没有头词的根节点，只要根节点是对的，也将这个根节点算作其中(Nivre et al., 2004)
- 根正确率(RA): 所有句子中找到正确根的句子所占的百分比(Yamada and Matsumoto, 2003)
- 完全匹配率(CM): 所有句子中无标记依存结构完全正确的句子所占的百分比(Yamada and Matsumoto, 2003)
- 带标记依存正确率(LAS): 所有词中找到正确的头词并分配到正确标记的词所占的百分比，对于没有头词的根节点，只要根节点是对的，也将这个根节点算作其中(Nivre et al., 2004)
- 标记正确率(LA): 所有词中依存标记正确的词所占的百分比，只要根节点是对的，也将这个根节点算作其中(Nivre et al., 2004)

英语依存句法分析现状 (段湘煜, 2008)

分析器	UAS	CM
Charniak	92.2	45.2
Collins	91.5	43.3
Bikel	91.4	42.6
McDonald and Pereira	91.5	42.1
McDonald et al.	91.0	37.5
Eisner	86.9	-
Yamada	90.4	38.4
Nivre	89.4	36.4

汉语依存句法分析现状(段湘煜, 2008)

	UAS	RA	CM
动作短语	84.36	73.70	32.70
动作链	84.05	73.39	32.34
Yamada	82.82	70.13	30.39
Nivre	82.69	68.19	29.82
dbparser	80.13	70.09	27.56
MSTParser ₁	81.26	68.20	25.72
MSTParser ₂	82.26	69.36	28.23

小结

- 依存句法分析关注词与词之间的关系，方便融合词义信息和词语关联信息，使其成为句法分析领域的一个重要方向
- 从依存句法分析的评测来看，评测的任务越来越复杂，与句义分析结合
- 从依存句法分析的结果来看，依存句法分析在准确率和效率上还远远无法满足真实应用的需要
- 从CoNLL评测看自然语言处理的发展，研究深度从词汇层→句法层→语义层(即词法分析、句法分析、语义角色标注)，但归根结底落实在词间关系上

纲要

- 第一部分：概述
- 第二部分：方法
 - 基于动态规划的方法
 - 基于决策的方法
 - 扩展性工作
- 第三部分：问题与挑战

依存句法分析方法

□ 任务：

- 输入：分词和词性标注好的句子
- 输出：句法依存图，满足：
 - 连通性
 - 无环性
 - 单一头词

依存句法分析方法

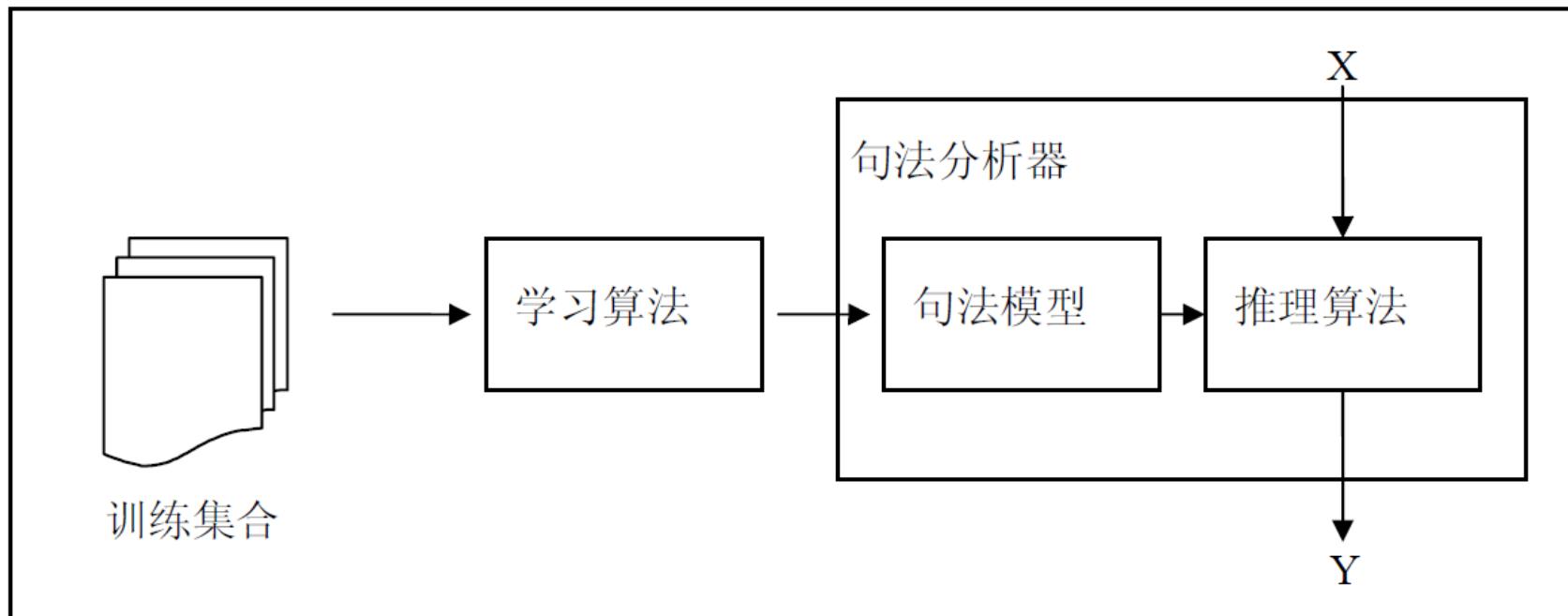
□ 基于语法驱动的方法(规则方法)

- 人工编写规则费时费力
- 人工编写规则覆盖度低
- 处理歧义问题效果不佳

□ 基于数据驱动的方法(统计方法)

- 不需要语法规则，任何一种句法结构都有可能存在
- 句法结构是由统计模型来决定的
- 能够比较灵活地处理歧义结构

基于数据驱动依存句法分析的框架 (McDonald, 2006; 段湘煜, 2008)



句法分析的一般性框架，其中 X 是输入的句子， Y 是输出的与 X 对应的句法树

基于数据驱动方法的分类

□ 根据依存分析过程分解方式：

□ 基于动态规划的方法

- 直接对依存图进行分解

□ 基于决策的方法

- 将依存分析过程分解成决策序列

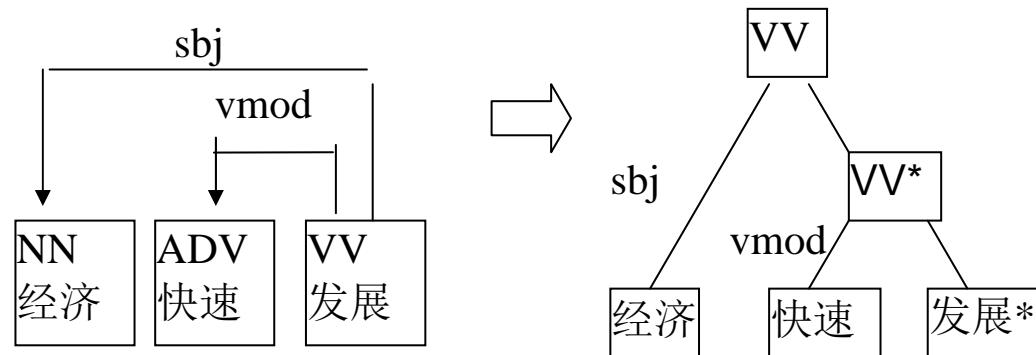
□ 基于约束满足的方法

- 将依存分析过程分解成去除不满足约束的过程

- 只在德语依存分析中使用过

基于动态规划的方法

- 早期方法：将依存图中的节点看作短语结构中的节点，从而可以应用上下文无关文法中成熟的CKY算法(Gaifman, 1965)



- 时间复杂度为 $O(n^5)$
- 双词汇语法的方法 (Bilexical Grammar)
 - 产生式方法(Eisner, 1996)
 - 判别式方法(McDonald et al., 2005; McDonald, 2006)
 - 时间复杂度为 $O(n^3)$

生成式方法(Eisner, 1996)

- **基本思想**：采用联合概率模型生成一系列依存句法树并赋予其概率分值，然后采用相关算法找到概率打分最高的分析结果作为最后的输出
- Eisner提出了三种生成式的概率依存模型
 - 模型A：二元亲和词汇模型(Bigram Lexical Affinities)
 - 模型B：优先选择模型(Selectional Preferences)
 - 模型C：递归生成模型(Recursive Generation)

模型A：二元亲和词汇模型

- 该模型利用一个三元马尔可夫模型进行词性标注，并确定任意一个词对是否是一个依存对

$$\Pr(\text{words}, \text{tags}, \text{links}) = \Pr(\text{words}, \text{tags}) \cdot \Pr(\text{link presences and absences} \mid \text{words}, \text{tags})$$
$$\approx \prod_{1 \leq i \leq n} \Pr(\text{tword}(i) \mid \text{tword}(i+1), \text{tword}(i+2)) \cdot \prod_{1 \leq i, j \leq n} \Pr(L_{ij} \mid \text{tword}(i), \text{tword}(j))$$

依存关系

两个词之间是否有依存关系

- 由于对交叉依存、多支配词等现象没有加以限制，这个模型是有疏漏的，可能有多个父节点，而违背单一头词这个约束

模型B: 优先选择模型

- 与模型A相比，不再对进行所有的词对进行穷举，而是利用每个词的优先选择信息($\text{preference}(i)$)，限制为每个词只选择一个父节点，因此不会有多个头词的问题

$$\begin{aligned} \Pr(\text{words}, \text{tags}, \text{links}) &= \Pr(\text{words}, \text{tags}) \cdot \Pr(\text{preferences} \mid \text{words}, \text{tags}) \\ &\approx \prod_{1 \leq i \leq n} \Pr(\text{tword}(i) \mid \text{tword}(i+1), \text{tword}(i+2)) \cdot \prod_{1 \leq i \leq n} \Pr(\text{preferences}(i) \mid \text{tword}(i)) \end{aligned}$$

词 i 优先选择的词，使之与词 i 构成依存关系

- 但是，该模型可能会违背无环的约束，因此也是有疏漏的

模型C: 递归生成模型

- 与模型A和B相比，该模型中每一个词生成自身的所有子节点，而不是像模型A一样对所有词对进行穷举，或者像模型B一样为每个词做优先选择（有可能形成环），因此该模型不再是疏漏的
- 在该模型中使用了两个马尔可夫链：左依存节点链和右依存节点链。词、词性、依存边的联合概率：

$$P(tw(1), \dots, tw(n), links) = \prod_{i=1}^n P(lc(i) | tw(i)) P(rc(i) | tw(i))$$

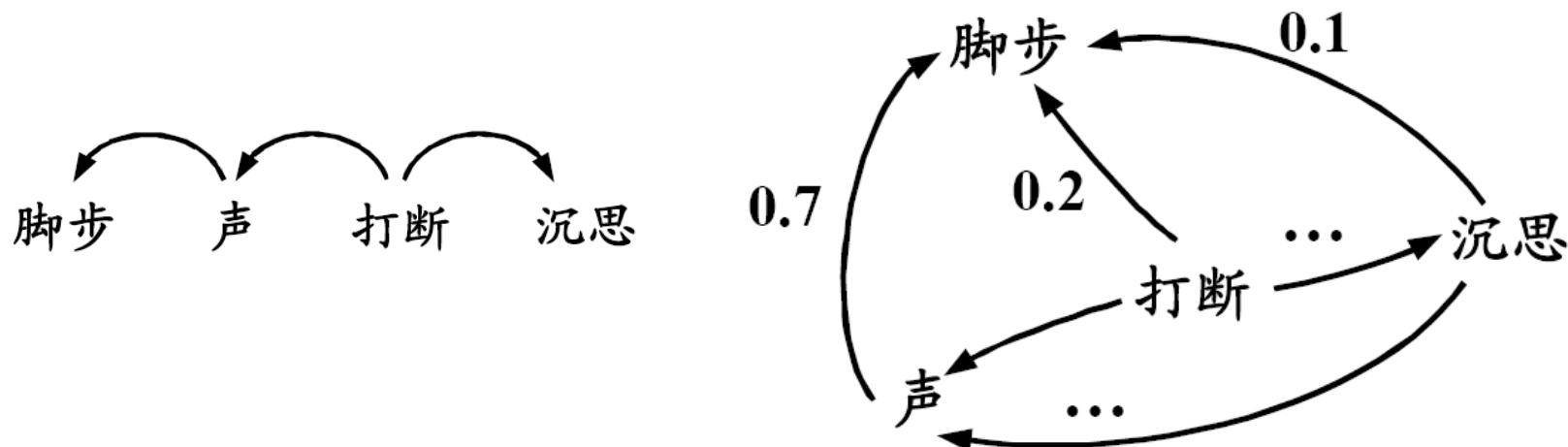
- 其中 $tw(i)$ 是指第 i 个带有词性标记的词， $lc(i)$ 和 $rc(i)$ 分别是第 i 个词的所有左子节点和所有右子节点

小结

- Eisner提出了生成式的概率模型，使得可以灵活地应用多种动态规划算法，取得了同短语结构句法分析可比的效果 (Eisner, 1996)
- 由于考虑了依存句法特性，利用跨度(span)代替子树 (subtree)，不再用词汇化的上下文无关文法常用的子树分析形式，算法的时间复杂度减少为 $O(n^3)$

判别式方法(McDonald et al., 2005) (1/2)

□ 基本思想：将依存分析看作是在一个依存图上寻找最大生成树(MST)的问题，该生成树满足上述三个约束条件：连通、单一父节点、无环



判别式方法(McDonald et al., 2005) (2/2)

- 对于一个经过分词和词性标注的句子：
 - 模型：利用给定的训练数据，学习一个全局最优的依存句法分析模型用来描述和刻画整棵句法树
 - 特征：词、词性及其组合
 - 训练算法：在线学习算法
 - 解码算法：CKY动态规划算法
 - 代表性成果：Eisner96, McDonald05, McDonald06, Carreras07, Wang et al.,07

判别式方法:模型(1/2)



X: 一个输入句子

Y: 一个候选依存句法树

$x_i \rightarrow x_j$: 从词*i*到词*j*的一条依存边

$\Phi(X)$: 对于输入句子X, 可能的句法树集合

基于边分解的模型
(Eisner 1996)

$$Y^* = \underset{Y \in \Phi(X)}{\operatorname{argmax}} score(Y | X)$$

$$= \underset{Y \in \Phi(X)}{\operatorname{argmax}} \sum_{(x_i \rightarrow x_j) \in Y} score(x_i \rightarrow x_j)$$

判别式方法:模型(2/2)

$$Y^* = \arg \max_{Y \in \Phi(X)} \sum_{(x_i \rightarrow x_j) \in Y} score(x_i \rightarrow x_j)$$

标准的线性分类器

$$score(x_i \rightarrow x_j) = \hat{f}(x_i, x_j) \cdot w$$

特征向量

特征向量的权重

三个关键部分：

- 特征选择：如何选择特征
- 参数学习（训练）：如何在训练集中学习到合适的权重
- 解码：如何找到打分最高的依存树

判别式方法: 特征



- 对词对(saw, duck), 特征主要有:

(saw, duck)=1

POS (saw, duck): (VBD, NN)=1

Dist (saw, duck)=2

Bigram 特征

距离特征

POS (saw, her, duck): (VBD, PRP, NN)=1

POS (I, saw, her, duck): (PRP, VBD, PRP, NN)=1

POS (saw, her, duck, with): (VBD, PRP, NN, IN)=1

上下文特征

(saw)=1

POS (saw): (VBD)=1

(duck)=1

POS (duck): (NN)=1

Unigram 特征

判别式方法:训练算法(1/2)

- 最大生成树的训练就是寻找使正确依存树得分最高的w
- McDonald等人采用在线学习(Online Learning)的方法训练w

Training data: $T = \{(x_t, y_t)\}_{t=1}^T$

1. $\mathbf{w}^{(0)} = 0; \mathbf{v} = 0; i = 0$

2. for $n : 1..N$

3. for $t : 1..T$

4. $\mathbf{w}^{(i+1)} = \text{update } \mathbf{w}^{(i)} \text{ according to instance } (x_t, y_t)$

5. $\mathbf{v} = \mathbf{v} + \mathbf{w}^{(i+1)}$

6. $i = i + 1$

7. $\mathbf{w} = \mathbf{v}/(N * T)$

在线学习算法的核心部分：下一页

权重向量取平均，可以避免过拟合

判别式方法:训练算法(2/2)

- 在McDonald等人的工作中，在线学习算法是边缘注入松弛算法MIRA
- 在每一步更新中，MIRA尝试在使正确依存树与错误依存树之间的分差大于某一个损失函数的条件下，使新的权重向量 $\mathbf{w}^{(i+1)}$ 尽可能靠近旧的权重向量 $\mathbf{w}^{(i)}$ ：

$$\min(\|\mathbf{w}^{(i+1)} - \mathbf{w}^{(i)}\|)$$

$$s.t. \ score(x_i, y_i) - score(x_i, y) \leq L(y_i, y)$$

$$\forall y \in k-best-trees(x_i)$$

正确依存树与错误依存树之间的损失函数，通常采用Hamming Loss

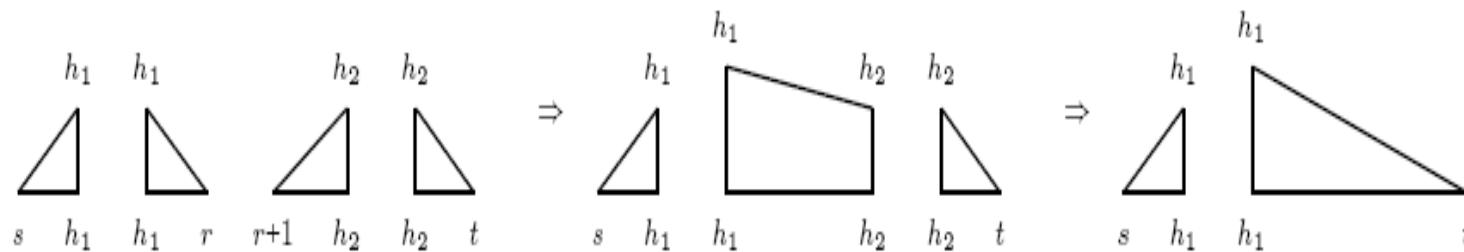
由于依存分析是复杂的结构预测问题，所有错误的依存树的得分难以全部计算，采用替代的方法

判别式方法:解码算法(Eisner, 1996)

- 完整依存结构与非完整依存结构:



- 算法图例:



- 在第一步，所有的依存结构是完整的；在第二步，将建立 h_2 依存于 h_1 的非完整结构；在最后，将建立完整依存结构
- 大依存结构通过自底向上的方式由小依存结构逐步建成
- 类似于CKY的动态规划方法

生成式方法 vs. 判别式方法(段湘煜, 2008)

□ 相同点：

- 均采用动态规划算法
- 全局搜索，得到的结果都是全局最优的

□ 不同点：

- 生成式方法采用联合概率模型，在进行概率乘积分解时做了近似假设和估计
- 判别式方法采用条件概率，具有更好的可操作性和可计算性，使得诸多机器学习算法得以应用
- 生成式方法算法比判别式方法复杂度较高，效率更低

小结：基于动态规划的方法

□ 优点：

- 可以计算整体依存树的得分，具有全局视角

□ 缺点：

- 由于受到计算复杂度的限制，可以考虑的特征结构受限

纲要

- 第一部分：概述
- 第二部分：方法
 - 基于动态规划的方法
 - 基于决策的方法
 - 扩展性工作
- 第三部分：问题与挑战

基于决策的方法

- 将分析过程看成是分析动作序列
 - ▣ 用类别有限的分析动作建立词与词之间的依存关系
 - ▣ 通过训练关于分析动作的分类器进行依存分析
- 代表方法
 - ▣ Covington (2001)
 - ▣ Yamada和Matsumoto (2003)
 - ▣ Nivre和Scholz (2004)

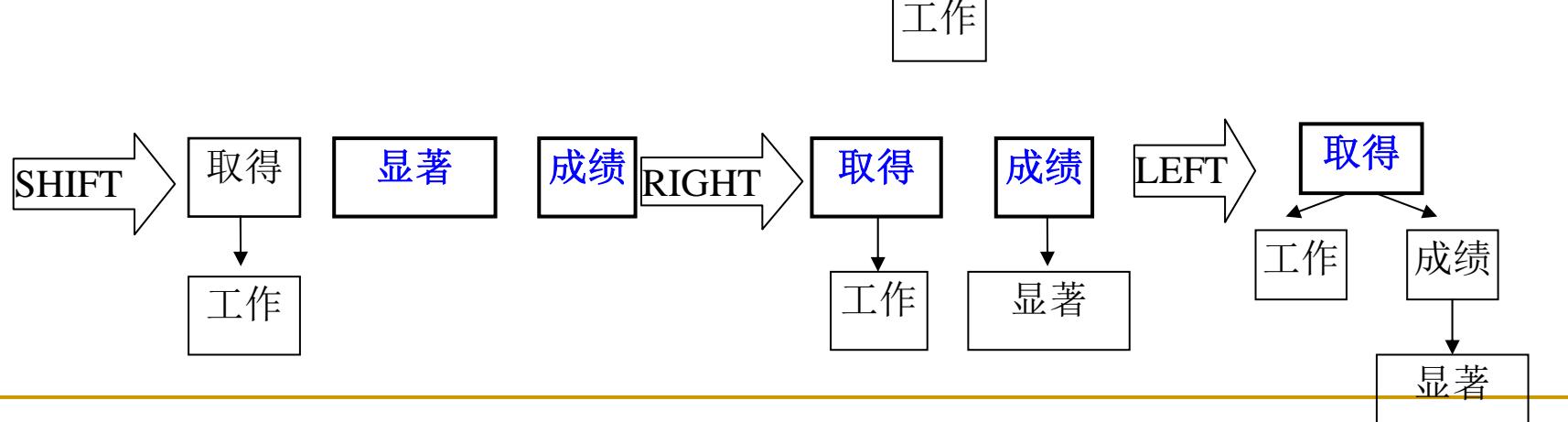
Covington(2001)方法：基本思想

- 决策的过程从句子的左端开始，逐个接受句子中的词，并尝试连接每个词与先前的所有词并将其作为头词或依存词
- 该算法是从左到右的穷尽式搜索算法，分析过程较为低效

Yamada和Matsumoto (2003)：示例

- 与Covington(2001)方法的穷尽式搜索不一样，该方法只关注当前的两个焦点词，效率更高
- 分析过程图例

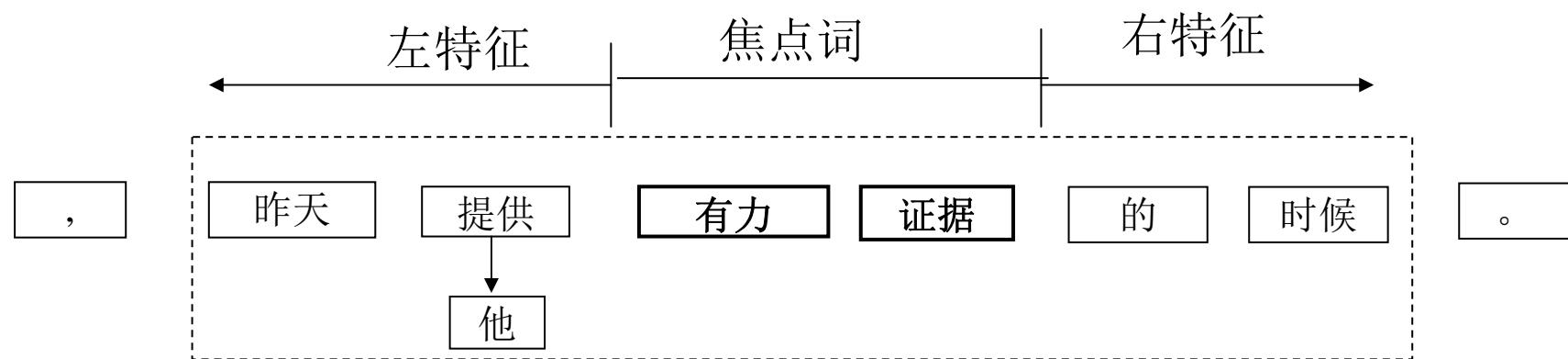
例句：工作取得显著成绩



Yamada和Matsumoto (2003)：分析动作

- LEFT：建立右焦点词依存于左焦点词的依存关系
- RIGHT：建立左焦点词依存于右焦点词的依存关系
- SHIFT：不建立依存关系，只转移句法分析的焦点，即新的左焦点词是原来的右焦点词，依此类推

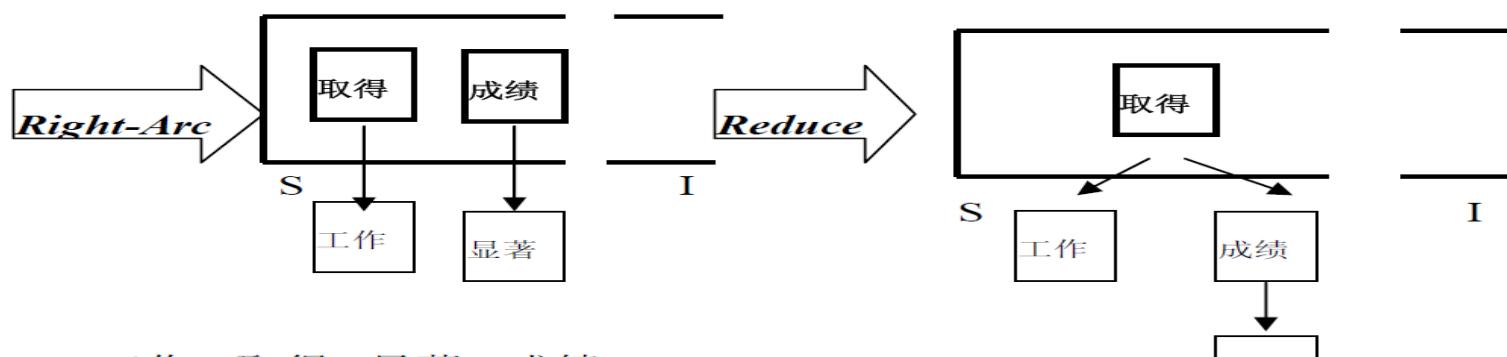
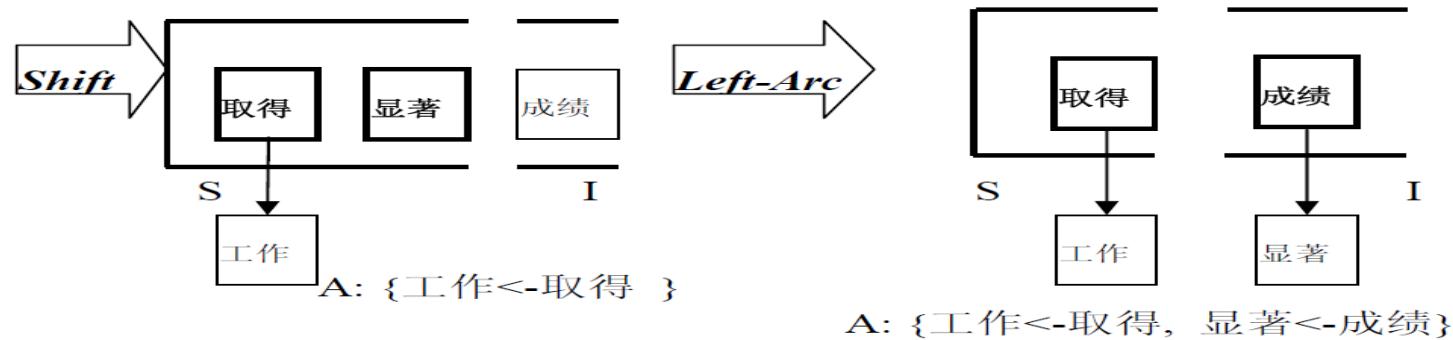
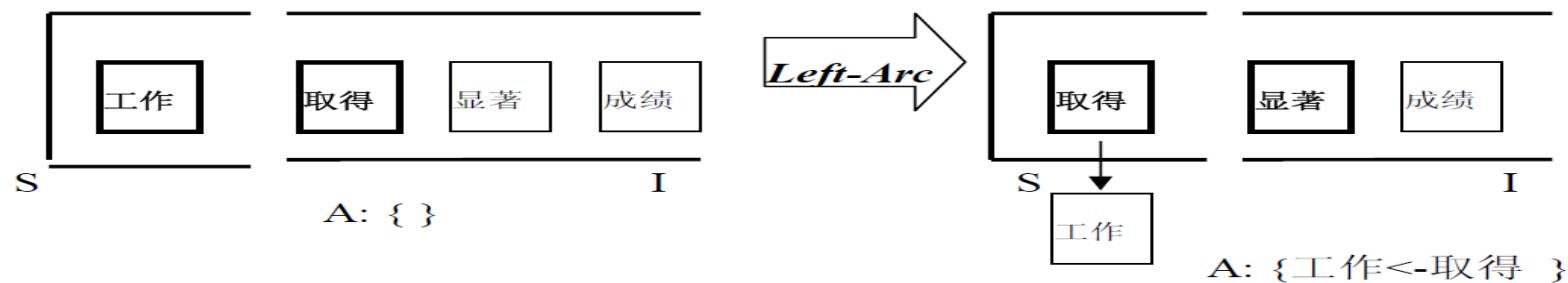
Yamada和Matsumoto (2003)：特征



- 特征在矩形框中：焦点词、左右上下文的词、词性及其组合关系
- 对应的分析动作是 *RIGHT*

Nivre和Scholz (2004)：特点

- 与Yamada和Matsumoto (2003)的方法没有太大的差异，其区别在于数据结构的定义和分析动作不一样
- 数据结构：依存句法分析器主要由一个三元组 $\langle S, I, A \rangle$ 构成，其中S表示一个栈结构，I表示剩余输入词序列，A表示在当前分析状态下所得到的依存关系集合
- 四个分析动作：Left-arc, right-arc, reduce, shift



A: {工作<-取得, 显著<-成绩,
取得->成绩}

A: {工作<-取得, 显著<-成绩,
取得->成绩}

Nivre和Scholz (2004)：分析过程

- 初始化

$S = \text{empty}$, $A = \text{句法树}$, $Q = \text{输入句子}$

- 终止状态

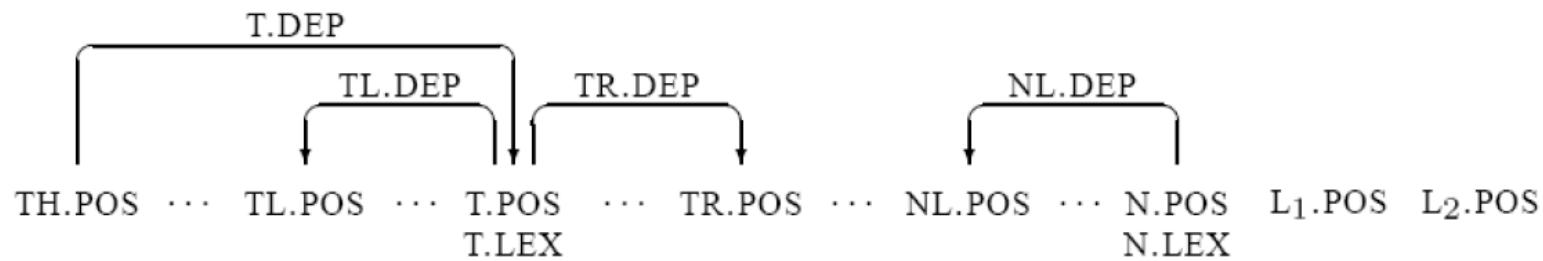
$S = [\text{root}]$, $A = \text{最终的句法树}$, $Q = []$

- 每次选择一个分析动作

- 分析动作是由分类器来决定的

Nivre和Scholz (2004)：特征

□ 特征模板：



T = Top of the stack

N = Next input token

TL = Leftmost dependent of T

TR = Rightmost dependent of T

NL = Leftmost dependent of N

L_i = Next plus i input token

X.LEX = Word form of X

X.POS = Part-of-speech of X

X.DEP = Dependency type of X

基于动态规划 vs. 基于决策的方法

(鉴萍, 2010)

The diagram features a green oval at the top right containing the text "两个极端" (Two Extremes). Two green arrows point downwards from this oval to the comparison table, one pointing to the left column and one pointing to the right column.

	基于动态规划的方法 (判别式方法)	基于决策的方法
分析单元	整棵依存树	一个词对
训练和解码	穷尽的、全局的	贪婪的、局部的
特征	受限制	丰富
分析模型	$\hat{G} = \arg \max_{G \in D(G_x)} \sum_{(i,j,l) \in A} s(i, j, l)$	$\hat{t} = \arg \max_{t \in T} s(c, t)$

纲要

- 第一部分：概述
- 第二部分：方法
 - 基于动态规划的方法
 - 基于决策的方法
 - **基于融合的方法**
 - 扩展性工作
- 第三部分：问题与挑战

基于融合的方法

- 基于搜索策略的融合方法 (Duan et al., ECML-PKDD 2007)
- 基于特征的融合方法 (Nivre and McDonald, ACL 2008)
- 基于模型的融合方法(Zhang and Clark, EMNLP 2008)

融合了基于动态规划方法和基于决策方法的优点

基于搜索策略融合的方法(Duan et al., ECML-PKDD 2007) (1/2)

- 目标：解决基于决策的方法的贪婪性
- 基本思想：将整个决策式依存句法分析过程看作是马尔科夫链。每一步分析中有若干个候选分析动作。句法分析的目标是在马尔科夫假设下寻找最有可能的分析动作序列
- 优点：该模型既可以利用丰富的上下文特征，又从全局的视角对决策动作建模
- 算法复杂度介于决策式方法和动态规划方法之间

基于搜索策略融合的方法(Duan et al., ECML-PKDD 2007) (2/2)

□ 中文实验结果：

	UAS	RA	CM
动作短语	84.36	73.70	32.70
动作链	84.05	73.39	32.34
Yamada	82.82	70.13	30.39
Nivre	82.69	68.19	29.82
dbparser	80.13	70.09	27.56
MSTParser ₁	81.26	68.20	25.72
MSTParser ₂	82.26	69.36	28.23

基于特征的融合方法 (Nivre and McDonald, ACL 2008) (1/3)

- 不同的句法分析器产生不同的错误(McDonald and Nivre, 2007)
- 基本思想：
 - 第一种策略：将基于动态规划的依存句法分析器(MSTParser)的输出作为基于决策的依存句法分析器(MaltParser)的特征-Malt_{MST}
 - 第二种策略：利用基于决策的依存句法分析器(MaltParser)的输出作为基于动态规划的依存句法分析器(MSTParser)的特征-MST_{Malt}

基于特征的融合方法 (2/3)

□ 特征

MST_{Malt} – defined over (i, j, l) ($*$ = any label/node)

Is $(i, j, *)$ in G_x^{Malt} ?

Is (i, j, l) in G_x^{Malt} ?

Is $(i, j, *)$ not in G_x^{Malt} ?

Is (i, j, l) not in G_x^{Malt} ?

Identity of l' such that $(*, j, l')$ is in G_x^{Malt} ?

Identity of l' such that (i, j, l') is in G_x^{Malt} ?

Malt_{MST} – defined over (c, t) ($*$ = any label/node)

Is $(\sigma_c^0, \beta_c^0, *)$ in G_x^{MST} ?

Is $(\beta_c^0, \sigma_c^0, *)$ in G_x^{MST} ?

Head direction for σ_c^0 in G_x^{MST} (left/right/ROOT)

Head direction for β_c^0 in G_x^{MST} (left/right/ROOT)

Identity of l such that $(*, \sigma_c^0, l)$ is in G_x^{MST} ?

Identity of l such that $(*, \beta_c^0, l)$ is in G_x^{MST} ?

基于特征的融合方法 (3/3)

□ 实验结果

Language	MST	MST _{Malt}	Malt	Malt _{MST}
Arabic	66.91	68.64 (+1.73)	66.71	67.80 (+1.09)
Bulgarian	87.57	89.05 (+1.48)	87.41	88.59 (+1.18)
Chinese	85.90	88.43 (+2.53)	86.92	87.44 (+0.52)
Czech	80.18	82.26 (+2.08)	78.42	81.18 (+2.76)
Danish	84.79	86.67 (+1.88)	84.77	85.43 (+0.66)
Dutch	79.19	81.63 (+2.44)	78.59	79.91 (+1.32)
German	87.34	88.46 (+1.12)	85.82	87.66 (+1.84)
Japanese	90.71	91.43 (+0.72)	91.65	92.20 (+0.55)
Portuguese	86.82	87.50 (+0.68)	87.60	88.64 (+1.04)
Slovene	73.44	75.94 (+2.50)	70.30	74.24 (+3.94)
Spanish	82.25	83.99 (+1.74)	81.29	82.41 (+1.12)
Swedish	82.55	84.66 (+2.11)	84.58	84.31 (-0.27)
Turkish	63.19	64.29 (+1.10)	65.58	66.28 (+0.70)
Average	80.83	82.53 (+1.70)	80.74	82.01 (+1.27)

基于模型的融合方法(Zhang and Clark, EMNLP 2008) (1/2)

- 将动态规划方法和决策方法进行加权组合

$$Score_{GRAPH}(parse) = \sum_{feature \in parse} feature \times weight(feature)$$

$$\begin{aligned} & Score_{TRANSITION}(parse) \\ &= \sum_{action \in parse} Score(action) \\ &= \sum_{action \in parse} \sum_{feature \in status \text{ for } action} feature \times weight(feature) \end{aligned}$$

$$Score_{COMBINED}(parse) = Score_{GRAPH}(parse) + Score_{TRANSITION}(parse)$$

线性加权组合

基于模型的融合方法 (Zhang and Clark, EMNLP 2008) (2/2)

□ 中文实验结果：

	Non-root	Root	Comp.
Graph [MA]	83.86	71.38	29.82
Duan 2007	84.36	73.70	32.70
Transition	84.69	76.73	32.79
Combined [TM]	86.13	77.04	35.25
Combined [TMA]	86.21	76.26	34.41

小结：基于融合的方法

- 融合了基于动态规划方法和基于决策方法的优点
- 既可以利用丰富的上下文特征，又从全局视角对分析动作建模
- 准确率：改进
- 效率：
 - 基于搜索策略的融合方法介于两者之间
 - 基于特征的融合方法复杂度要高
 - 基于模型的融合方法复杂度要高

纲要

- 第一部分：概述
- 第二部分：方法
 - 基于动态规划的方法
 - 基于决策的方法
 - 基于融合的方法
 - 扩展性工作
- 第三部分：问题与挑战

扩展性工作

- 目前依存句法分析的性能(无论是动态规划的方法,决策的方法或融合的方法)受限于树库规模的大小

- 宾州树库

- 450万词
 - 20万句子
 - 人工标注

- 宾州中文树库的规模更小

- 规模有限

- 标注费时费力

- 未标注文本

- 新闻文本
 - 维基百科
 - Web资源

- 海量
 - 容易获得

有监督

半监督

扩展性工作：基于半监督的方法

□ 基于聚类的方法

- 对词汇泛化(Koo et al., 2008; Chen et al., 2009; Agirre et al., 2011)
- 对词性细化(Zhou et al., 2011)

□ 双语句法结构映射的方法

- 整棵树结构映射(Ganchev et al., 2009; Hwa et al., 2005; Jiang et al., 2009; Smith and Eisner, 2009)
- 词对依存关系映射(Jiang et al., 2010)

□ 基于Web词语选择关系的方法

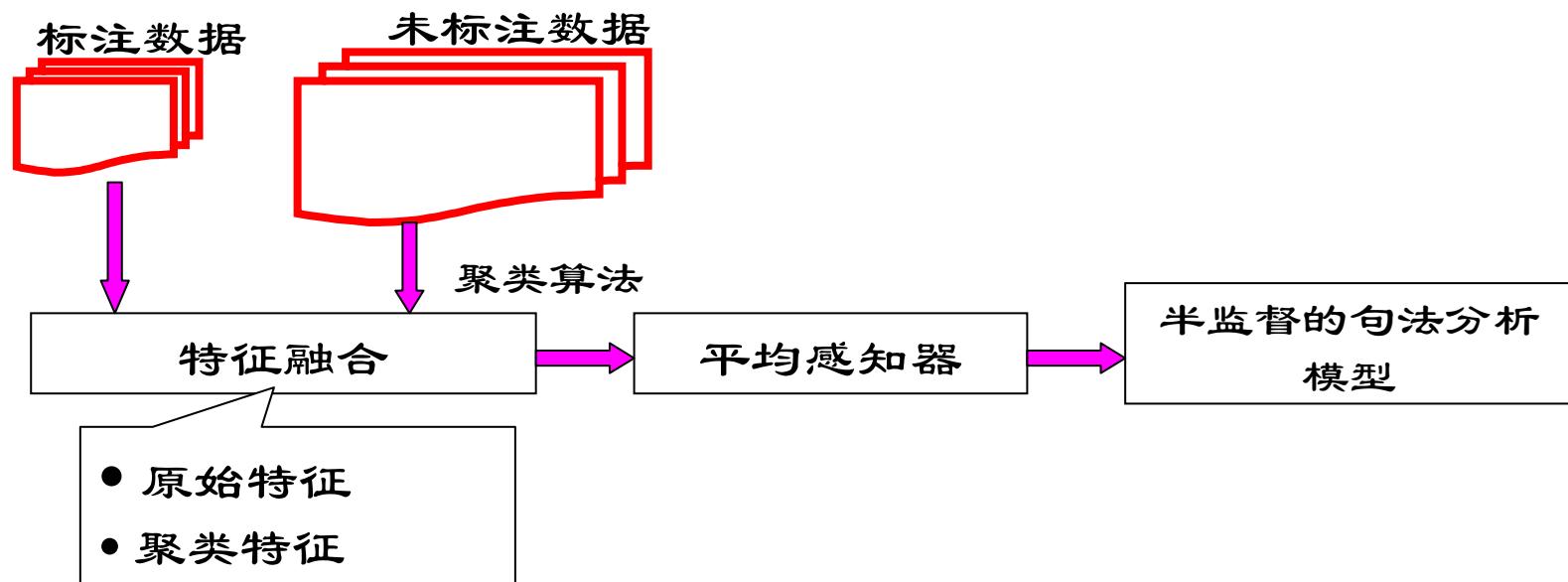
- Bansal and Klein (2011)
- Zhou et al. (2011)

基于聚类的方法

- 解决依存句法结构的歧义主要依赖两类特征：
 - 词汇
 - 词性
- 由于树库的规模有限，用词汇来区分句法结构的歧义，词汇表达过细，导致数据稀疏
- 用词性来区分句法结构的歧义，词性表达过粗，无法区分特定的歧义
- 能不能引入一种中间实体，这种实体的表达粒度介于词性和词汇之间？比词性更细，比词汇更粗

基于聚类的方法：对词汇泛化(1/3)

- Koo et al., ACL 2008:
 - 该方法从词汇层面入手，既然词汇表达会导致数据稀疏性的问题，采用词聚类的方法，引入词聚类特征，有效缓解了数据的稀疏性问题



基于聚类的方法：对词汇泛化(2/3)

Koo et al., (2008): 英语的实验结果：

Sec	dep1	dep1c	MD1	dep2	dep2c	MD2
00	90.48	91.57 (+1.09)	—	91.76	92.77 (+1.01)	—
01	91.31	92.43 (+1.12)	—	92.46	93.34 (+0.88)	—
23	90.84	92.23 (+1.39)	90.9	92.02	93.16 (+1.14)	91.5
24	89.67	91.30 (+1.63)	—	90.92	91.85 (+0.93)	—

基于聚类的方法：对词汇泛化(3/3)

□ Chen et al., EMNLP 2009:

□ 基本思想：

- 依存句法分析在短距离的依存关系上准确率很高，可以用现有的依存分析器分析大规模语料，得到长度为1或2的大规模依存对
- 在此基础上，对大规模依存对按照频率聚类，从而缓解数据的稀疏性
- 对句子进行在线依存分析时，将以上聚类的大规模依存对特征加入到原有特征中，从而提高分析性能

□ 在英语和汉语上均取得了不错的效果

□ Agirre et al., ACL 2011 :

□ 基本思想：

- WordNet中定义了词汇的同义概念，如果两个词属于同义词，那么它们就属于一个语义类，提取基于类的特征

基于聚类的方法：对词性细化

- Zhou et al., IJCNLP 2011:
- 词性的表达过粗，采用HowNet中定义的上下位关系，将每一类词性(NN, VV等)进行细分，提高特征对特定句法歧义的区分能力

NN		VV	
NN-InstitutePlace	企业(enterprise) 公司(company)	VV-event	猜到(guess) 预见(foresee)
NN-aValue	经济(economy) 国际(international)	VV-aValue	小心(care) 可以(can)
NN-organization	国家(country) 政府(government)	VV-SelfMoveInDirection	进行(conduct) 扩散(spread)
NN-event	发展(developing) 合作(cooperation)	VV-change	增长(increase) 涨价(deform)
NN-human	记者(reporter) 专家(expert)	VV-attribute	简称(abbreviation) 库容(storage capacity)
NN-affairs	贸易(trading) 金融(financial)	VV-entity	经历(experience) 考虑(consider)
NN-mental	情绪(mood) 感受(feelings)	VV-AlterRelation	围困(siege) 脱离(separate)
NN-entity	后者(latter) 机会(opportunity)	VV-AlterPossession	借用(borrow) 购进(buy)
NN-artifact	棉花(cotton) 维生素(vitamin)	VV-AlterPhysical	建造(build) 制成(make)
...
AD		JJ	
AD-aValue	以后(after) 唯有(only)	JJ-aValue	共同(together) 特别(special)
AD-event	还(also) 不管(no matter)	JJ-event	继续(continue) 相对(relatively)
...

双语句法结构映射的方法

- 虽然人工标注大规模树库费时费力，但往往可以比较容易获得其与树库规模较大语言的双语平行语料
 - 例如：汉语依存分析时，汉语依存树库少，不够训练，但是存在大量汉英平行语料库，而英语的依存分析的性能较高，可以利用英语依存分析器对平行语料中的英文部分进行自动分析，得到英文依存树库；再将英文依存结构投射到中文端
- 投射方法
 - 整棵树映射
 - 词对依存关系映射
- 其实质仍然是解决数据稀疏性问题

双语句法结构映射的方法：整棵树 结构映射(1/3)

- **基本思想**：借助词语对齐信息，将源语言依存树中词间的依存关系直接投射到未经句法分析的目标语言语句
- **难点问题**：直接映射过程受到词语对齐错误和语言异构性的差异很大，经常导致映射的依存边产生错误，甚至找不到完整的映射树(Hwa et al., 2005)
- **解决方法**：
 - Hwa et al. (2005)和Ganchev et al. (2009)用一些过滤规则减少映射产生的噪音，并手工设计一些规则以解决语言异构的问题
 - Smith and Eisner (2009)借助准同步语法改进依存映射，取得了一定的性能提升

双语句法结构映射的方法：整棵树 结构映射(1/3)

□ 基本思想：借助词语对齐信息，将源语言依存树中词间的依存关系直接投射到未经句法分析的目标语言语句

□ 难点问题：直接映射的差异很大，经常导致完整的映射树(Hw)

- 人工设计的语言学规则在质量上受语言学家水平的制约
- 这些规则在数量上不可能做到太大，对真实语言现象的覆盖面较窄

□ 解决方法：

- Hwa et al. (2005)和Ganchev et al. (2009)用一些过滤规则减少映射产生的噪音，并手工设计一些规则以解决语言异构的问题
- Smith and Eisner (2009)借助准同步语法改进依存映射，取得了一定的性能提升

受制于准同步语法的表达能力，仅从缓解语言异构性的角度出发，并未考虑词语对齐错误和以及如何利用对齐信息

双语句法结构映射的方法：整棵树 结构映射(3/3)

- Jiang et al. (2009)提出了一种具有容错性的依存句法映射算法
- 映射过程采用类似于句法分析的动态规划过程，并非直接投射
- 构造了所有可能双语词对之间的对齐概率二维表，减少了单一词语对齐结果可能产生的错误（多个候选代替唯一候选）

双语句法结构映射的方法：词对依存关系映射

- Jiang et al. (2010)提出了一种非同构的依存句法映射方法
- 基本思想：该方法不需要将完整的依存结构映射过来，映射过程只需要得到一部分置信度高依存边
 - 源语言端依存对的置信度
 - 对齐的置信度
- 过程：给定一个词语对齐且源语言端经过依存分析的双语语料，源语言语句中词对之间的依存关系可以投射到目标语言的词对上

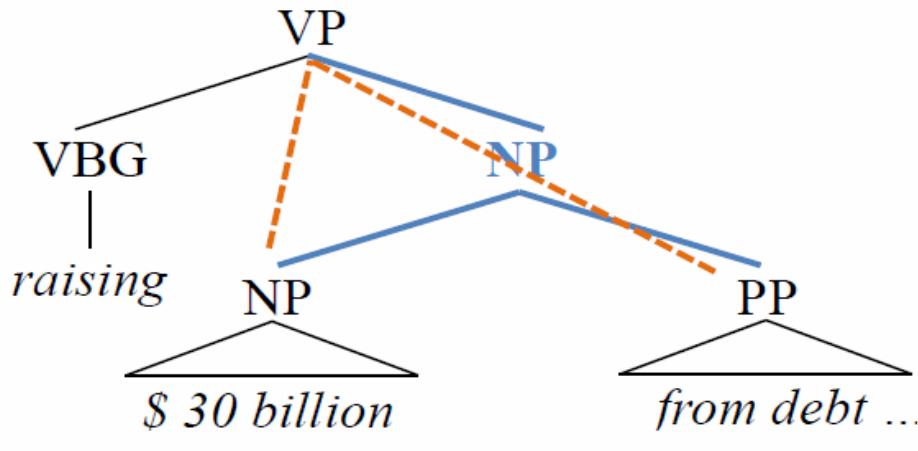
基于Web词语选择关系的方法(1/3)

- 句法分析的错误有很大一部分来源于词语错误的附着问题
 - 介词短语的附着
 - 副词的附着
 - 名词短语内部结构歧义
- 处理这些特定的错误，仅仅依靠有限规模的训练数据是难以解决的。这些歧义往往与特定的词语组合关系密切相关

基于Web词语选择关系的方法(2/3)

- 特定词语组合关系：词汇的语义选择限制关系
 - Resnik (1993)提出了基于类的选择限制关系，主要探讨的是动词与宾语之间的关系
 - 一些研究扩展了基于类的选择限制关系—基于词汇的选择限制关系
 - 应用于句法分析(Bansal and Klein, 2011; Zhou et al., 2011)
 - 解决特定结构的词汇歧义，比如介词附着问题

基于Web词语选择关系的方法(3/3)



- 介词短语的附着歧义
- 目前句法分析器输出的结果：from debt 附着到 \$ 30 billion
- 正确的结果：from debt 附着到 raising

□ 从Web海量数据中挖掘词语关联信息：

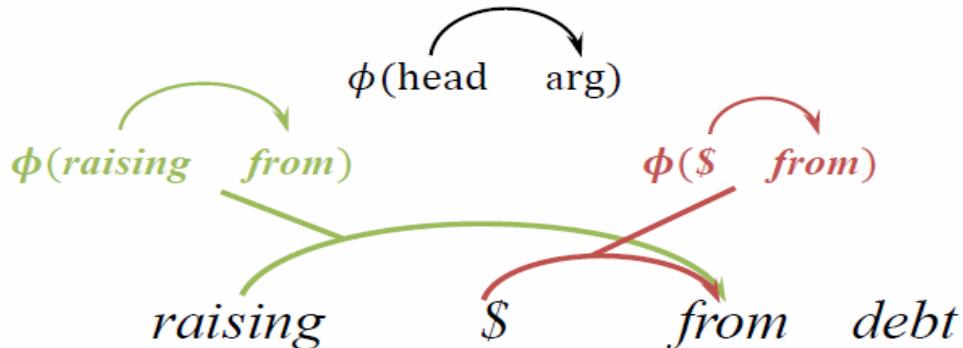
count(raising from)>count(\$ x billion from)

提示信息：from 附着于raising 更合理

基于Web词语选择关系的方法： Bansal and Klein (2011) (1/6)

- Web 数据
 - Google n-grams corpus
- Web 特征
 - 二元词汇关联
 - 复述特征（具有相似搭配模式的词语组合，例如 raise ... from 和 lower... with）

基于Web词语选择关系的方法： Bansal and Klein (2011) (2/6)



- $h = \text{raising}$, $a = \text{from}$
- $h = \$$, $a = \text{from}$

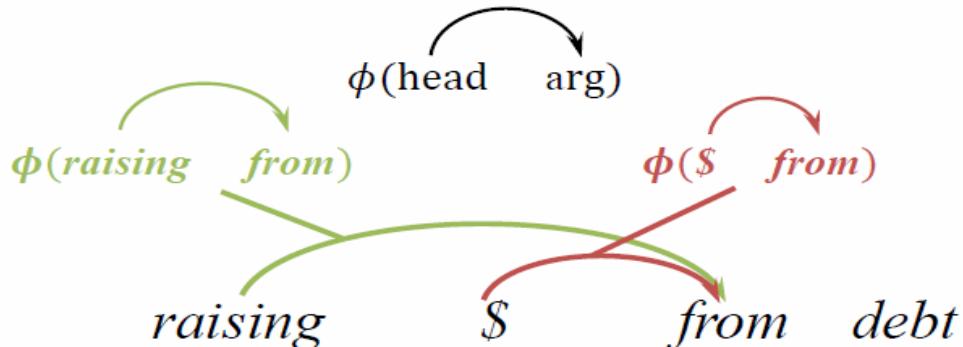
特征： $\text{POS}(h) \wedge \text{POS}(a) \wedge b$

其中： $b = \text{floor}(\log(\text{count})/5) * 5$

二元词汇关联

特征

基于Web词语选择关系的方法： Bansal and Klein (2011) (3/6)



- h=raising, a=from
- h= \$, a=from

特征： PARA[^]POS(h)[^]POS(a)[^]c[^]p[^]dir

复述特征

PARA[^]VBG[^]IN[^]it[^]Middle[^]→

在Google n-grams corpus中，查询(raising, c, from)，
选择出现次数最多的词c=it

基于Web词语选择关系的方法： Bansal and Klein (2011) (4/6)

□ 在测试阶段如何利用上述特征：

□ 如果存在 $h = \text{lowering}$, $a = \text{with}$

□ 就可以构造特征：

PARA^AVBG^BIN^Cit^DMiddle^E→

因为在训练数据中，我们已经知道 from 依存于 raising，所以在这里复述特征就暗示了 with 依存于 lowering

基于Web词语选择关系的方法： Bansal and Klein (2011) (5/6)

实验结果：

	Order 2	+ Web features	% Error Redn.
Dev (sec 22)	92.1	92.7	7.6%
Test (sec 23)	91.4	92.0	7.0%

基于Web词语选择关系的方法： Bansal and Klein (2011) (6/6)

实验结果：

Arg Tag	# Attach	Baseline	This Work	% ER
NN	5725	5387	5429	12.4
NNP	4043	3780	3804	9.1
IN	4026	3416	3490	12.1
DT	3511	3424	3429	5.8
NNS	2504	2319	2348	15.7
JJ	2472	2310	2329	11.7
CD	1845	1739	1738	-0.9
VBD	1705	1571	1580	6.7
RB	1308	1097	1100	1.4
CC	1000	855	854	-0.7
VB	983	940	945	11.6
TO	868	761	776	14.0
VBN	850	776	786	13.5
VBZ	705	633	629	-5.6
PRP	612	603	606	33.3

整体性能提升了，但是某些类的结果反而下降了，可能是噪音问题

基于Web词语选择关系的方法：

Zhou et al. (2011) (1/6)

- Web数据

- Google n-grams corpus
- Google hits

- 特征

- 二元词汇关联
- 三元词汇关联
- 介词附着歧义

基于Web词语选择关系的方法：

Zhou et al. (2011) (2/6)

□ 特征

□ 二元词汇关联

□ 三元词汇关联

□ 介词附着歧义

$$\text{PMI}(x, y) = \log \frac{p("x y")}{p("x")p("y")}$$

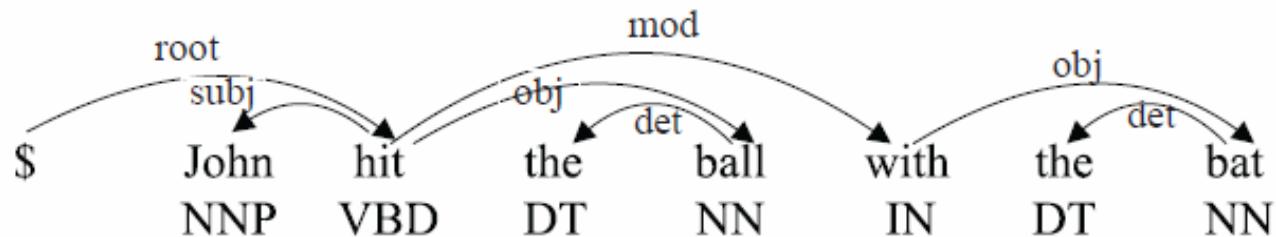
P("x y")词频可以通过
Google hits 和 Google n-
grams corpus 获得

$$\text{PMI}(x, y, z) = \log \frac{p("x y z")}{p("x y")p("y z")}$$

$$\text{PMI}_{IN}(x, z) = \log \frac{p("x IN z")}{p(x)}$$

基于Web词语选择关系的方法： Zhou et al. (2011) (3/6)

□ 例子



PMI("hit with")

$x_i\text{-word}=\text{"hit"}, x_j\text{-word}=\text{"with"}, \text{PMI}(\text{"hit with"})$

$x_i\text{-word}=\text{"hit"}, x_j\text{-word}=\text{"with"}, x_j\text{-pos}=\text{"IN"}, \text{PMI}(\text{"hit with"})$

$x_i\text{-word}=\text{"hit"}, x_i\text{-pos}=\text{"VBD"}, x_j\text{-word}=\text{"with"}, \text{PMI}(\text{"hit with"})$

$x_i\text{-word}=\text{"hit"}, b\text{-word}=\text{"ball"}, x_j\text{-word}=\text{"with"}, \text{PMI}(\text{"hit with"})$

$x_i\text{-word}=\text{"hit"}, x_j\text{-word}=\text{"with"}, \text{PMI}(\text{"hit with"}), \text{dir=R}, \text{dist}=3$

...

基于Web词语选择关系的方法： Zhou et al. (2011) (4/6)

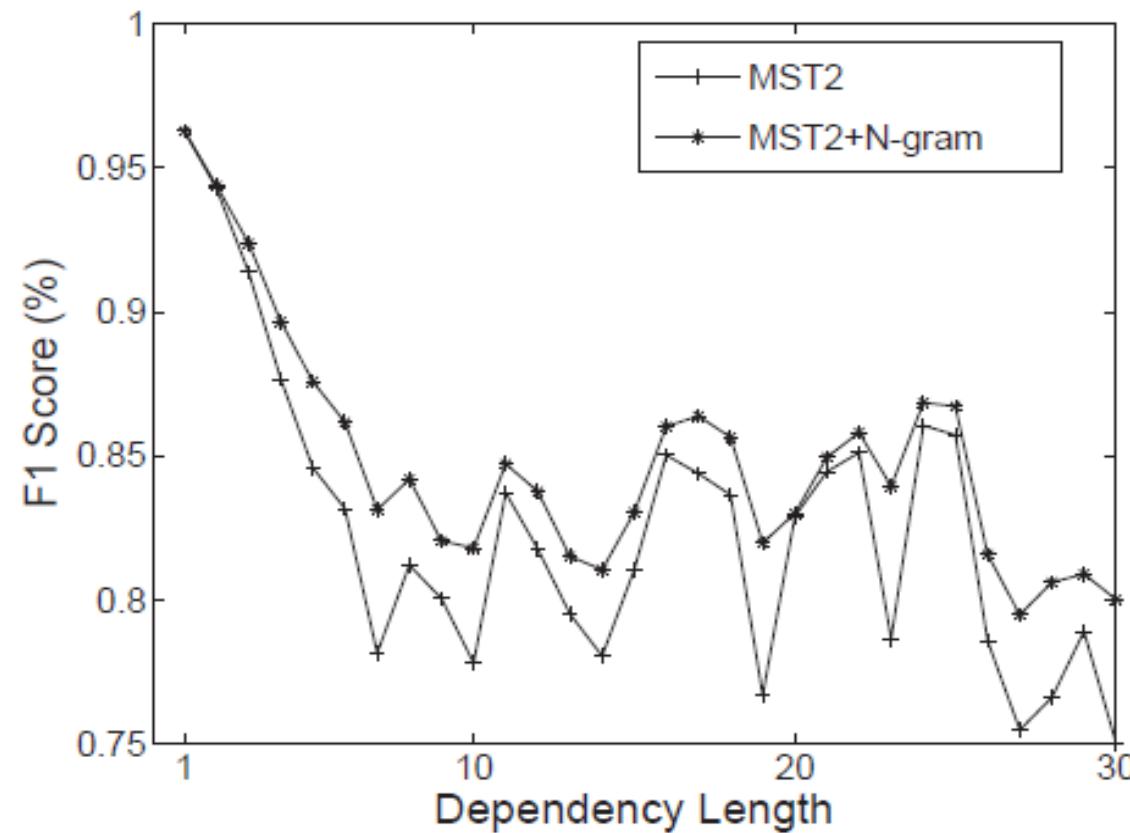
英文的实验结果：

Sec	dep1	+hits	+V1	dep2	+hits	+V1	dep1-L	+hits-L	+V1-L	dep2-L	+hits-L	+V1-L
00	90.39	90.94	90.91	91.56	92.16	92.16	90.11	90.69	90.67	91.94	92.47	92.42
01	91.01	91.60	91.60	92.27	92.89	92.86	90.77	91.39	91.39	91.81	92.38	92.37
23	90.82	91.46	91.39	91.98	92.64	92.59	90.30	90.98	90.92	91.24	91.83	91.77
24	89.53	90.15	90.13	90.81	91.44	91.41	89.42	90.03	90.02	90.30	90.91	90.89

基于Web词语选择关系的方法：

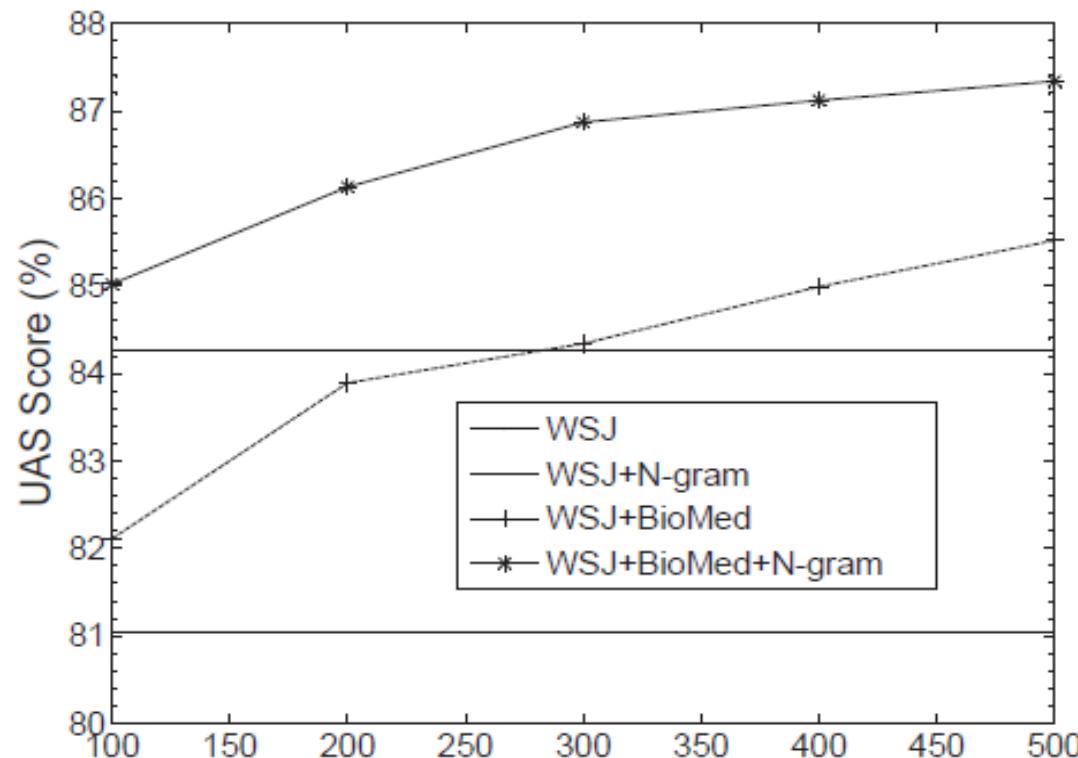
Zhou et al. (2011) (5/6)

依存长度 vs. F-score



基于Web词语选择关系的方法： Zhou et al. (2011) (6/6)

在生物医学领域上测试：



小结：扩展性工作

□ 优点：

- 利用Web海量数据获取词语间的选择关系，简单直观
- 为依存句法分析的**领域自适应**提供新的思路
- 为处理**长距离的依存关系**提供新的思路

□ 缺点：

- Web数据易引入噪声

纲要

- 第一部分：概述
- 第二部分：方法
 - 基于动态规划的方法
 - 基于决策的方法
 - 基于融合的方法
 - 扩展性工作
- 第三部分：问题与挑战

问题与挑战(1/4)

- 依存句法分析的准确度
 - 目前仍然无法满足实际应用需求
- 依存句法分析的评价指标是否合理
 - 目前依存句法分析评价主要借助于CoNLL评测指标
 - 能不能设计更加灵活的评价机制
- 依存句法分析的领域自适应问题
 - 如何设计高鲁棒性的依存句法分析器

问题与挑战(2/4)

□ 依存句法分析的速度

- ▣ 目前依存句法的速度较慢，无法适应大规模的真实应用
- ▣ 鉴萍(2010)针对依存句法分析的速度做了系统分析
- ▣ 依存句法分析的准确度与分析速度往往是相互牵制的，在实际应用中需要针对不同任务做适当的平衡
- ▣ Huang and Sagae(2010)提出了一种时间复杂度为线性的动态规划方法，是句法分析器在速度和精度上的折中

问题与挑战(3/4)

- 面向网络信息处理的依存句法分析
 - 依存句法分析一直局限在规模较小的语料上，随着互联网技术的发展，网络上有海量的文本资源，能不能开发面向网络应用需求的依存句法分析器
 - 面向信息抽取去设计
 - 面向社区问答系统去设计

问题与挑战(4/4)

□ 依存句法分析模型

- 目前依存句法的输入是需要经过分词和词性标注好的句子，分词和词性标注的错误会导致依存句法的错误传递
 - 能不能设计一个统一的模型，将分词、词性标注和依存句法分析融入一个框架中，减少错误传递
- 在基于动态规划的方法中，探索利用更复杂的特征
 - Koo and Collins(ACL 2010):
 - Third-order dependency parser
 - Hayashi et al. (EMNLP 2011):
 - Third-order variational reranking

参考文献

- M. Bansal and D. Klein. 2011. Web-scale Features for Full-Scale Parsing. In ACL-HLT.
- W. Chen, J. Kazama, K. Uchimoto and K. Torisawa. 2009. Improving Dependency Parsing with Subtrees from Auto-Parsed Data. In EMNLP.
- J. Eisner. 1996. Three new probabilistic models for dependency parsing: an exploration. In COLING.
- K. Ganchev, J. Gillenwater, and B. Taskar. Dependency grammar induction via bitext projection constraints. In ACL, 2009.
- J. Hall, J. Nivre and J. Nilsson. 2006. Discriminative classifier for deterministic dependency parsing. In ACL.
- L. Huang and K. Sagae. 2011. Dynamic programming for linear-time incremental parsing. In ACL
- L. Huang, W. Jiang and Q. Liu. 2009. Bilingually-constrained (monolingual) shift-reduce parsing. In EMNLP.
- R. Hwa, P. Resnik, A. Weinberg, C. Cabezas, and O. Kolak. Bootstrapping parsers via syntactic projection across parallel texts. In NLE, 2005.
- K. Hayashi, T. Watanabe, M. Asahara, and Y. Matsumoto. Third-order variational reranking on packed-shared dependency forests. In EMNLP.
- T. Koo and M. Collins. 2010. Efficient third-order dependency parsers. In ACL.
- D. Klein and C. Manning. Corpus based induction of syntactic structures: Models of dependency and constituency. In ACL, 2004.

参考文献

- T. Koo, X. Carreras and M. Collins. 2008. Simple semi-supervised dependency parsing. In ACL.
- R. McDonald, K. Crammer, and F. Pereira. 2005. On-line large-margin training of dependency parsers. In ACL.
- R. McDonald and J. Nivre. 2007. Characterizing the errors of data-driven dependency parsing models. IN EMNLP.
- J. Nivre and R. McDonald. 2008. Integrating graph-based and transition-based dependency parsing. In ACL.
- J. Nivre and R. McDonald. 2008. Integrating graph-based and transition-based dependency parsers. In ACL.
- J. Nivre and M. Scholz. Deterministic dependency parsing of english text. In COLING, 2004.
- W. Jiang and Q. Liu. Dependency parsing and projection based on word-pair classification. In ACL, 2010.
- Y. Zhang and S. Clark. 2008. A tale of two parsers: investigating and combining graph-based and transition-based dependency parsing using beam-search. In EMNLP.
- H. Zhao, Y. Song, C. Kit and G. Zhou. 2009. Cross language dependency parsing using a bilingual lexicon. In ACL.
- G. Zhou, J. Zhao, K. Liu and L. Cai. 2011. Exploiting web-derived selectional preference to improve statistical dependency parsing. In ACL-HLT.
- 段湘煜. 基于分析动作建模的依存句法分析. 中国科学院自动化研究所博士论文, 2008年
- 鉴萍. 依存句法分析方法研究与系统实现. 中国科学院自动化研究所博士论文, 2010年
- 宗成庆. 统计自然语言理解. 清华大学出版社, 2008年.

Q&A

Thanks

主要内容总结

□ 信息抽取

- 实体识别与抽取、实体消歧、关系抽取

□ 观点挖掘和倾向性分析

- 观点倾向性分析、观点信息抽取、观点检索

□ 问答系统

- 问答式检索、社区问答

□ 依存句法分析

- 基于动态规划的方法、基于决策的方法、基于融合的方法、扩展性工作

结束语

- 自然语言处理和信息抽取技术将在互联网应用中得到广泛应用
- 互联网也将为自然语言处理和信息抽取技术提供越来越多的资源和技术创新的源泉

感谢

- 国家自然科学基金项目：61070106, 60875041,
60673042, 60372016
- 863项目：2006AA01Z144

Thanks!