Beyond Bag of Words

CS6200
Information Retrieval
Bags of Words

• Most efficient (and still very effective) retrieval models treat words/terms as independent

• Generalized models based on (log-)linear combination of features of both query and document

\[ S_W(D;Q) = \sum_j w_j \cdot f_j(D,Q) \]
Term Dependence Models

- Full independence
- Sequential dependence
- Full dependence
- General dependence
Term Dependence Models

Sequential dependence

\[ S_W(D;Q) = \sum_j w_j \cdot f_j(D, Q) \]

#weight(0.8 #combine(president abraham lincoln))
0.1 #combine(#od:1(president abraham)
    #od:1(abraham lincoln))
0.1 #combine(#uw:8(president abraham)
    #uw:8(abraham lincoln)))
Term Dependence Models

Sequential dependence

\[ S_W(D;Q) = \sum_j w_j \cdot f_j(D, Q) \]

Weights of different bigrams are tied

#weight(0.8 #combine(president abraham lincoln))
0.1 #combine(#od:1(president abraham)
    #od:1( abraham lincoln))
0.1 #combine(#uw:8(president abraham)
    #uw:8( abraham lincoln)))
Term Dependence Models

Sequential dependence

Weights of different bigrams are tied

Therefore, estimate weights for classes of features, not for each individual bigram

\[ S_W(D;Q) = \sum_j w_j \cdot f_j(D,Q) \]
Term Dependence Models

\[ S_W(D;Q) = \sum_j w_j \cdot f_j(D, Q) \]

Full dependence

#weight(0.8 #combine(president abraham lincoln))
0.1 #combine(#od:1(president abraham)
    #od:1(abraham lincoln)
    #od:1(president abraham lincoln))
0.1 #combine(#uw:8(president abraham)
    #uw:8(abraham lincoln)
    #uw:8(president lincoln)
    #uw:12(president abraham lincoln)))
Term Dependence Models

Full dependence

\[ S_W(D;Q) = \sum_j w_j \cdot f_j(D, Q) \]

Features (with tied weights) for bigrams, trigrams, etc.

- `#weight(0.8 #combine(president abraham lincoln))`
- `0.1 #combine(#od:1(president abraham))`
- `0.1 #combine(#od:1(abraham lincoln))`
- `0.1 #combine(#od:1(president abraham lincoln))`
- `0.1 #combine(#uw:8(president abraham))`
- `0.1 #combine(#uw:8(abraham lincoln))`
- `0.1 #combine(#uw:8(president lincoln))`
- `0.1 #combine(#uw:12(president abraham lincoln)))`
Term Dependence Models

Unigram relevance model

Latent concept expansion

<table>
<thead>
<tr>
<th>1-word concepts</th>
<th>2-word concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>telescope</td>
<td>hubble telescope</td>
</tr>
<tr>
<td>hubble</td>
<td>space telescope</td>
</tr>
<tr>
<td>space</td>
<td>hubble space</td>
</tr>
<tr>
<td>mirror</td>
<td>telescope mirror</td>
</tr>
<tr>
<td>NASA</td>
<td>telescope hubble</td>
</tr>
<tr>
<td>launch</td>
<td>mirror telescope</td>
</tr>
<tr>
<td>astronomy</td>
<td>telescope NASA</td>
</tr>
<tr>
<td>shuttle</td>
<td>telescope space</td>
</tr>
<tr>
<td>test</td>
<td>hubble mirror</td>
</tr>
<tr>
<td>new</td>
<td>NASA hubble</td>
</tr>
<tr>
<td>discovery</td>
<td>telescope optical</td>
</tr>
<tr>
<td>time</td>
<td>telescope astronomy</td>
</tr>
<tr>
<td>universe</td>
<td>telescope optical</td>
</tr>
<tr>
<td>optical</td>
<td>telescope discovery</td>
</tr>
<tr>
<td>light</td>
<td>telescope shuttle</td>
</tr>
</tbody>
</table>
Syntactic Dependencies

Maximum directed spanning tree

I did not unfortunately receive an answer to this question
Syntactic Dependencies

Maximum directed spanning tree

I did not unfortunately receive an answer to this question
Syntactic Dependencies

I did not unfortunately receive an answer to this question

$\mathit{Maximum\ directed\ spanning\ tree}$

$n^{n-1} = 10^9 = 1$ billion possible trees!
Cross-Language Syntax

Target language

Auf diese Frage habe ich leider keine Antwort bekommen

I did not unfortunately receive an answer to this question

Source language
Cross-Language Syntax

Tschernobyl könnte dann etwas später an die Reihe kommen

Then we could deal with Chernobyl some time later
Cross-Language Syntax

Tschernobyl könnte dann etwas später an die Reihe kommen

Then we could deal with Chernobyl some time later
Cross-Language Syntax

Then we could deal with Chernobyl some time later.

Tschernobyl könnte dann etwas später an die Reihe kommen.
Then we could deal with Chernobyl some time later

Tschernobyl könnte dann etwas später an die Reihe kommen
Tschernobyl könnte dann etwas später an die Reihe kommen

Then we could deal with Chernobyl some time later
Cross-Language Syntax

Tschernobyl könnte dann etwas später an die Reihe kommen

Then we could deal with Chernobyl some time later

Structures not isomorphic:
use quasi-synchronous alignment features
Alignment Configurations
Alignment Configurations

sehe ---- see
ich --------- I

monotonic
(parent-child)
Alignment Configurations

monotonic (parent-child)

head swapping
Alignment Configurations

sehe ----> see
ich ----> I

schwimmt ----> likes

gern ----> swimming

Völkerrecht ----> law

international

monotonic
(parent-child)

head swapping

two-to-one
Alignment Configurations

- monotonic (parent-child)
- head swapping
- siblings
- two-to-one
Alignment Configurations

- sehe ----> see
- schwimmt ----> likes
- Völkerrecht ----> law

Monotonic (parent-child)

- ich -----> habe
- ich -----> gekauft
- ich -----> I

Head swapping

- schwimmt ----> swimming

- habe -----> bought

International

- gehabt -----> bought

- Wahlkampf -----> campaign

Two-to-one

- wegen -----> since

Grandparent-grandchild

- von -----> from

- 2003 -----> 2003
Alignment Configurations

- Monotonic (parent-child)
- Head swapping
- Two-to-one
- Siblings
- Grandparent-grandchild

Also C-command, descendant, “none of the above”
Quasi-Synchronous Dependence

- Syntactic models of document mismatch
  - Developed (by me) for MT (SMT 2006; EMNLP 2009)

Query

shih tzu health problems

Documents

... Find out all the serious health problems that face the Shih Tzu. ...
Quasi-Synchronous Dependence

- Syntactic models of document mismatch
  - Developed (by me) for MT (*SMT 2006; EMNLP 2009*)

Query

shih tzu health problems

Documents

... Find out all the serious health problems that face the Shih Tzu. ...

\[ P@10 \]

<table>
<thead>
<tr>
<th></th>
<th>Igram LM</th>
<th>SDM</th>
<th>LM+QG</th>
<th>SDM+QG</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.45</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.43</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.41</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.39</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Quasi-Synchronous Dependence

• Syntactic models of document mismatch
  ✤ Developed (by me) for MT (SMT 2006; EMNLP 2009)

... Find out all the serious health problems that face the Shih Tzu. ...

Query

Documents

shih tzu health problems

P@10

0.47

0.45

0.43

0.41

0.39

1gram LM

SDM

LM+QG

SDM+QG
Quasi-Synchronous Dependence

- Syntactic models of document mismatch
  * Developed (by me) for MT (SMT 2006; EMNLP 2009)

Query

... Find out all the serious health problems that face the Shih Tzu. ...

Documents

... Find out all the serious health problems that face the Shih Tzu. ...

- P@10
  - 0.39
  - 0.41
  - 0.43
  - 0.45
  - 0.47

<table>
<thead>
<tr>
<th>Models</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1gram LM</td>
<td>0.39</td>
</tr>
<tr>
<td>SDM</td>
<td>0.41</td>
</tr>
<tr>
<td>LM+QG</td>
<td>0.43</td>
</tr>
<tr>
<td>SDM+QG</td>
<td>0.47</td>
</tr>
</tbody>
</table>
Quasi-Synchronous Dependence

- Syntactic models of document mismatch
  - Developed (by me) for MT (SMT 2006; EMNLP 2009)

Query

Documents

... Find out all the serious health problems that face the Shih Tzu. ...

```
Query: shih tzu health problems

Documents: (Content not shown)

... Find out all the serious health problems that face the Shih Tzu. ...
```

```
P@10

<table>
<thead>
<tr>
<th>Model</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1gram LM</td>
<td>0.39</td>
</tr>
<tr>
<td>SDM</td>
<td></td>
</tr>
<tr>
<td>LM+QG</td>
<td>0.47</td>
</tr>
<tr>
<td>SDM+QG</td>
<td></td>
</tr>
</tbody>
</table>
```
Quasi-Synchronous Dependence

- Syntactic models of document mismatch
  - Developed (by me) for MT (*SMT 2006; EMNLP 2009*)

Query

- Find out all the serious health problems that face the Shih Tzu...
Query:
biomedical research and technology

Top Ranked Results:
minneapolis research signs inc.
syntex
california institute of technology
massachusetts institute of technology
therapeutic products

Top 50 entity results for **Tycho Brahe**

<table>
<thead>
<tr>
<th>Entity/Description</th>
<th>Type</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tycho Ottesen Brahe</td>
<td>PERSON</td>
<td>(Dec 14, 1546 - Oct 24, 1601)</td>
</tr>
<tr>
<td>Tycho</td>
<td>MISC</td>
<td></td>
</tr>
<tr>
<td>Brahe</td>
<td>MISC</td>
<td></td>
</tr>
<tr>
<td>Uranienborg</td>
<td>MISC</td>
<td></td>
</tr>
<tr>
<td>Johannes Kepler</td>
<td>PERSON</td>
<td>(Dec 27, 1571 - Nov 15, 1630)</td>
</tr>
<tr>
<td>Hveen</td>
<td>MISC</td>
<td></td>
</tr>
<tr>
<td>Erra Pater</td>
<td>MISC</td>
<td></td>
</tr>
<tr>
<td>Nicolaus Copernicus</td>
<td>PERSON</td>
<td>(Feb 19, 1473 - May 24, 1543)</td>
</tr>
</tbody>
</table>

Top 50 entity results for **Oneida**

<table>
<thead>
<tr>
<th>Entity/Description</th>
<th>Type</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oneida, New York</td>
<td>LOCATION</td>
<td>(Longitude: -75</td>
</tr>
<tr>
<td>Onondagas</td>
<td>MISC</td>
<td></td>
</tr>
<tr>
<td>Mohawks</td>
<td>MISC</td>
<td></td>
</tr>
<tr>
<td>Cayugas</td>
<td>MISC</td>
<td></td>
</tr>
<tr>
<td>Senecas</td>
<td>MISC</td>
<td></td>
</tr>
<tr>
<td>Oneida Community</td>
<td>MISC</td>
<td></td>
</tr>
<tr>
<td>Oneida County</td>
<td>LOCATION</td>
<td>(Longitude: -75</td>
</tr>
<tr>
<td>Tuscaroras</td>
<td>MISC</td>
<td></td>
</tr>
<tr>
<td>Oneida Castle, New York</td>
<td>LOCATION</td>
<td>(Longitude: -75 43.0783333333;</td>
</tr>
<tr>
<td>Oneida Conference</td>
<td>ORGANIZATION</td>
<td></td>
</tr>
<tr>
<td>Oneida Indians</td>
<td>MISC</td>
<td></td>
</tr>
<tr>
<td>Mohicans</td>
<td>MISC</td>
<td></td>
</tr>
<tr>
<td>Mohawk</td>
<td>MISC</td>
<td></td>
</tr>
</tbody>
</table>
Entity Search

Joint distribution over queries and entities

\[ P(e, q) = \sum_{d \in D} P(e, q|d) P(d) \]

Factorized dist’n

\[ P(e, q|d) = P(q|e, d)P(e|d) \]

Probability by proximity

\[ P(q|e, d) = \frac{1}{Z} \sum_{i=1}^{N} \delta_d(i, q)k(q, e) \]

Gaussian proximity kernel

\[ \exp -||q - e||^2 / 2\sigma^2 \]
Question Answering

• Human vs. machine: Compare search engine queries with, e.g., StackOverflow

• Current QA systems perform simple “factoid” retrieval

  • Who invented the paper clip?
  • Where is the Valley of the Kings?
  • When was the last major eruption of Mt. St. Helens?
Question Answering
Question Answering

Example: Indri system used as first stage of IBM Watson
Example: Indri system used as first stage of IBM Watson

What kind of answer is the user expecting?
<table>
<thead>
<tr>
<th>Question</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>What do you call a group of geese?</td>
<td>Animal</td>
</tr>
<tr>
<td>Who was Monet?</td>
<td>Biography</td>
</tr>
<tr>
<td>How many types of lemurs are there?</td>
<td>Cardinal</td>
</tr>
<tr>
<td>What is the effect of acid rain?</td>
<td>Cause/Effect</td>
</tr>
<tr>
<td>What is the street address of the White House?</td>
<td>Contact Info</td>
</tr>
<tr>
<td>Boxing Day is celebrated on what day?</td>
<td>Date</td>
</tr>
<tr>
<td>What is sake?</td>
<td>Definition</td>
</tr>
<tr>
<td>What is another name for nearsightedness?</td>
<td>Disease</td>
</tr>
<tr>
<td>What was the famous battle in 1836 between Texas and Mexico?</td>
<td>Event</td>
</tr>
<tr>
<td>What is the tallest building in Japan?</td>
<td>Facility</td>
</tr>
<tr>
<td>What type of bridge is the Golden Gate Bridge?</td>
<td>Facility Description</td>
</tr>
<tr>
<td>What is the most popular sport in Japan?</td>
<td>Game</td>
</tr>
<tr>
<td>What is the capital of Sri Lanka?</td>
<td>Geo-Political Entity</td>
</tr>
<tr>
<td>Name a Gaelic language.</td>
<td>Language</td>
</tr>
<tr>
<td>What is the world’s highest peak?</td>
<td>Location</td>
</tr>
</tbody>
</table>
Answer Categorization

How much money does the Sultan of Brunei have?  
Jackson Pollock is of what nationality?  
Who manufactures Magic Chef appliances?  
What kind of sports team is the Buffalo Sabres?  
What color is yak milk?  
How much of an apple is water?  
Who was the first Russian astronaut to walk in space?  
What is Australia's national flower?  
What is the most heavily caffeinated soft drink?  
What does the Peugeot company manufacture?  
How far away is the moon?  
Why can’t ostriches fly?  
What metal has the highest melting point?  
What time of day did Emperor Hirohito die?  
What does your spleen do?  
What is the best-selling book of all time?  

Money  Nationality  Organization  Org. Description  Other  Percent  Person  Plant  Product  Product Description  Quantity  Reason  Substance  Time  Use  Work of Art
OCR Transcripts

*Original:
The fishing supplier had many items in stock, including a large variety of tropical fish and aquariums of all sizes.

*OCR:
The fishing supplier had many items in stock, including a large variety of tropical fish and aquariums of all sizes~

*Original:
* This work was carried out under the sponsorship of National Science Foundation Grants NSF-GN-380 (Studies in Indexing Depth and Retrieval Effectiveness) and NSF-GN-482 (Requirements Study for Future Catalogs).

*OCR:
This work was carried out under the sp011J!0rship 01 NationLUL1 Setenee Foundation Orant. NSF-ON-580 (Studl .. In Indexing Depth and Retrieval Efliccth"ene&&) and NSF-ON-482 (Requirements Study lor Future ‘Catalogs)•
ASR Transcripts

Transcript:
French prosecutors are investigating former Chilean strongman Augusto Pinochet. The French justice minister may seek his extradition from Britain. Three French families whose relatives disappeared in Chile have filed a Complaint charging Pinochet with crimes against humanity. The national court in Spain has ruled crimes committed by the Pinochet regime fall under Spanish jurisdiction.

Speech recognizer output:
french prosecutors are investigating former chilean strongman of coastal fish today the french justice minister may seek his extradition from britain three french families whose relatives disappeared until i have filed a complaint charging tenants say with crimes against humanity the national court in spain has ruled crimes committed by the tennessee with james all under spanish jurisdiction
Image Tagging

cherry blossom obama

President Obama visits the Cherry Blossom Festival? Well, not really—his motorcade passed by the Tidal Basin en route to Ft. McNair for his big Libya speech. A few lucky tourists got a chance to see it whizz by.

Tags

- washington dc
- district of columbia
- 15th street sw
- tidal basin
- 2011 cherry blossom festival
- bureau of printing and engraving
- department of the treasury
- president barack obama
- motorcade
- presidential limousine
- aka the beast
- dcist
- POTUS
- cadillac
Image Annotation

- Offline annotation
- Online pseudo-relevance feedback
  - Retrieve images by text annotation
  - Estimate text/image relevance model
- Rerank more images