Knowledge Based Text Representation for Information Retrieval

Chenyan Xiong
cx@cs.cmu.edu
Language Technologies Institute
Carnegie Mellon University
Document Ranking, the core task of information retrieval

Query
Carnegie Mellon

Documents
Computer
Andrew
CMU
……..
……..
……..
……..

Representation

\( \vec{q} = \{101:1, 233:1\} \)

\( d_1 \):
\{101:10, 234:5, 456:3\…\}

\( d_2 \):
\{151:5, 233:5, 601:3\…\}

\( d_3 \):
\{456:30, 562:7, 233:1\…\}

Ranking Model
Carnegie Mellon

\( d_1: -3.56 \)
\( d_2: -5.12 \)
\( d_3: -13.2 \)

Searches for: carnegie mellon

1. Carnegie Mellon Today
   (chows609-vm0008-93-03237) http://www.carnegie Mellon.edu
   Carnegie Mellon Today...October 2008, Vol. 5 No. 4 Fall is always an exciting time of year for
   the returning students. This year we have some added...Despite the adversity, she kept moti-
   vated to Carnegie Mellon, where she accomplished what no other underdog has done here! At
   Just like us, the world’s greatest works of art eventually show their age, some sooner than oth-
   ers. By Paul Whitmore, we are part of a unique center that finds the best ways...
   (cached version)

2. Human Resources - Careers - Carnegie Mellon University
   (chows609-vm0005-12-36393) http://www.cmu.edu/jobs/
   Human Resources - Careers - Carnegie Mellon University Search Only Human Resources - C-
   arnegie Mellon University: Combining Arts and Technology to Create a
   Specially interested in joining the Carnegie Mellon team? We’re pleased to introduce you to our Univers-
   ityCareers@CarnegieMellon...
   (cached version)

3. carnegie - carnegie international co - carnegie center med
   (chows609-vm0001-77-20401) http://wincom.com/...html
   carnegie - carnegie international co - carnegie center merit award like carnegie stock fish earn
carnegie melIon davidish dealership carnegie deal carnegie southern california carnegie mille
   and southfield mi skelton and carnegie insurance carnegie electricity center carnegie capital a
   symptoms carnegie institute of technology in michigan mits carnegie mellon stresses bond carnegie
   and eric jobs carnegie how did...
   (cached version)
Information Retrieval: State-of-the-art

Representation

Ranking Model

Unsupervised
- Language Model
- BM25
- VSM
- DPH
- COOR

Learning to Rank
- Pointwise: LR, SVM…
- Pairwise: RankSVM…..
- Listwise: SVMMAP, ListMLE…
Information Retrieval: State-of-the-art

Representation

• Bag-of-Words
  • Represent text by its individual words
• State-of-the-art
• The backbone of many text related tasks

Ranking Model

Unsupervised

- Language Model
- VSM
- BM25
- DPH
- COOR

Learning to Rank

- Pointwise: LR, SVM…
- Pairwise: RankSVM…..
- Listwise: SVM-MAP, ListMLE…
Text Understanding of Human beings

Carnegie Mellon University is a private research university in Pittsburgh, Pennsylvania
Text Understanding of Human beings

Carnegie Mellon University is a private research university in Pittsburgh, Pennsylvania.
Text Understanding of Human beings

Good at CS

My heart is in the “work”

Located In

Carnegie Mellon University is a private research university in Pittsburgh, Pennsylvania

Won a lot Tony and Turing Award
Graph based storage of human knowledge:

Large Scale:
- DBpedia
- Freebase
- NELL

Semi-structured
Carefully Curated
External to IR systems

Carnegie Mellon University is a private research university in Pittsburgh, Pennsylvania.
Our Goal: Knowledge Based Text Representation (ReMAKE)

From bag-of-words

Query: Carnegie Mellon

- Vocabulary Mismatch
- Ambiguity
- Synonymy
- Shallow Text understanding

Documents:
Carnegie Mellon University is a private research university…

Count

Exact Match

{101:1, 233:1}

To Knowledge Based Representations

Query: Carnegie Mellon

Count

Inference

More Evidences

High Quality Domain Knowledge

Structured Representation

Deeper Text Understanding

Documents:
Carnegie Mellon University is a private research university…

Intelligent Match
Outline

Query Representation
• Query Expansion with Freebase [Term-based] (ICTIR 2015)
• EsdRank [Entity-based] (CIKM 2015)
• Learning to Find Entity [Related Entity Finding] (SIGIR 2016 Short)

Document Representation
• Bag-of-entities [Entity-based] (Under review)

Proposed Research
• Hierarchal Learning to Rank for Multiple Representations
• Co-Learning: Entity Finding and Document Ranking
• Entity Graph Representation
QUERY REPRESENTATION
Query Representation with Knowledge Base

Entity plays a crucial role in query understanding

- >70% web (Bing) queries contain entities [Guo et al. 2009]
- >50% web (Yahoo!) queries are searching for entities [Pound et al. 2010]
- ~100% head and medium (AOL) queries contains Wikipedia entities
- Ambiguity frequently comes from entities [TREC Web Track]
Three Ways to Find Related Entities

Query Annotation
• Entities that appear in the query are useful
Three Ways to Find Related Entities

Entity Search

- Entities that are retrieved by the query are useful
Three Ways to Find Related Entities

Document Annotations

• Entities that frequently appear in the top retrieved documents are useful (Pseudo Relevance Feedback).
# Examples of Selected Entities

<table>
<thead>
<tr>
<th>Query: <strong>ESPN</strong></th>
<th>Entity Search</th>
<th>Document Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESPN</td>
<td>ESPN</td>
<td></td>
</tr>
<tr>
<td>Sports radio</td>
<td>ABC Sports</td>
<td></td>
</tr>
<tr>
<td>ESPN Radio</td>
<td>Bill Simmons</td>
<td></td>
</tr>
<tr>
<td>Sports Center</td>
<td>Sports Center</td>
<td></td>
</tr>
<tr>
<td>ESPN International</td>
<td>NBA</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query: <strong>Yahoo</strong></th>
<th>Entity Search</th>
<th>Document Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yahoo</td>
<td>Yahoo</td>
<td></td>
</tr>
<tr>
<td>Yahoo News</td>
<td>Microsoft</td>
<td></td>
</tr>
<tr>
<td>Yahoo Music</td>
<td>Jerry Yang</td>
<td></td>
</tr>
<tr>
<td>Yahoo Mail</td>
<td>Associated Press</td>
<td></td>
</tr>
<tr>
<td>Yahoo Messenger</td>
<td>YUI Library</td>
<td></td>
</tr>
</tbody>
</table>
Outline

Query Representation
• Query Expansion with Freebase [Term-based] (ICTIR 2015)
• EsdRank [Entity-based] (CIKM 2015)
• Learning to Find Entity [Related Entity Finding] (SIGIR 2016 Short)

Document Representation
• Bag-of-entities [Entity-based] (Under review)

Proposed Research
• Hierarchal Learning to Rank for Multiple Representations
• Co-Learning: Entity Finding and Document Ranking
• Entity Graph Representation
Query Expansion with Freebase

Term Based Query Representation from Knowledge Base
Expansion with Freebase in two steps:
1. Related entity selection
2. Expansion term selection

Two methods in each step:
- Select related entities by:
  - Entity search
  - Entity annotation in top retrieved docs
- Select expansion terms by:
  - PRF in related entities’ descriptions
  - Category similarity with query
Term Selection Method 1: PRF in Entities’ Descriptions

Entities’ descriptions as Pseudo Relevance Feedback documents

• PRF score of a term $t_i$ from entities’ descriptions:

$$s(t_i) = \sum_{e_k \in E} tf_{i,k} \times idf_i \times s(e_k)$$

Selected entities

Tf and idf in entity description

Score from entity selection
Term Selection Method 2: Category Similarity

Select terms by category similarities with query

Trained on entity descriptions in each category

Distribution on Freebase’s first level categories (77 total)

Negative JS divergence as term’s expansion score
Unsupervised Expansion Methods: Summary

**Entity Selection × Term Selection = Four unsupervised methods**

<table>
<thead>
<tr>
<th>Entity Search</th>
<th>Document Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pseudo Relevance Feedback</td>
<td>FbSearchPRF</td>
</tr>
<tr>
<td>Category Similarity</td>
<td>FbSearchCat</td>
</tr>
</tbody>
</table>

Two entity selection methods

Two term selection methods
Evaluation: Unsupervised Expansion

ClueWeb09 + TREC Queries

All our query expansion with Freebase methods outperform all baselines, on all metrics.

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP@20</th>
<th>NDCG@20</th>
<th>ERR@20</th>
<th>Relative Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>IndriLm</td>
<td>0.357</td>
<td>0.147</td>
<td>0.116</td>
<td>NA</td>
</tr>
<tr>
<td>SDM</td>
<td>0.387†</td>
<td>0.166†</td>
<td>0.122†</td>
<td>5.52%</td>
</tr>
<tr>
<td>RmWiki</td>
<td>0.362</td>
<td>0.161†</td>
<td>0.114</td>
<td>−1.70%</td>
</tr>
<tr>
<td>SVMPRF</td>
<td>0.367</td>
<td>0.158†</td>
<td>0.125</td>
<td>8.00%</td>
</tr>
<tr>
<td>FbSearchPRF</td>
<td>0.436†</td>
<td>0.186†</td>
<td>0.152†</td>
<td>30.80%</td>
</tr>
<tr>
<td>FbSearchCat</td>
<td>0.421†</td>
<td>0.182†</td>
<td>0.144†</td>
<td>23.99%</td>
</tr>
<tr>
<td>FbFaccPRF</td>
<td>0.428†</td>
<td>0.184†</td>
<td>0.145†</td>
<td>24.71%</td>
</tr>
<tr>
<td>FbFaccCat</td>
<td>0.400†</td>
<td>0.173†</td>
<td>0.136†</td>
<td>17.25%</td>
</tr>
</tbody>
</table>

Combination of our methods provide further improvements.
More details in the paper.
Outline

Query Representation
• Query Expansion with Freebase [Term-based] (ICTIR 2015)
• EsdRank [Entity-based] (CIKM 2015)
• Learning to Find Entity [Related Entity Finding] (SIGIR 2016 Short)

Document Representation
• Bag-of-entities [Entity-based] (Under review)

Proposed Research
• Hierarchical Learning to Rank for Multiple Representations
• Co-Learning: Entity Finding and Document Ranking
• Entity Graph Representation
**EsdRank: Connecting Query and Document Through External Semi-structured Data**

Entity Based Query Representation from Knowledge Base
EsdRank: A General Framework to Use External Semi-Structured Data for Ranking

Entities as bridges between query and documents

Incorporate more information from knowledge bases as features
EsdRank: A General Framework to Use External Semi-Structured Data for Ranking

A novel ranking model: **Latent-ListMLE**
- Learns how to represent query and how to rank documents together
Latent-ListMLE: Connects q and docs through a latent layer

- For each rank position $i$:
  1. Sample object $o_i$ based on query: $p(o_i|q)$
  2. Sample document $d_i$ from unselected documents $S_i$: $p(d_i|o_i, S_i)$

For each ranking position

The probability/importance of latent objects to the query

ListMLE-like ranking generation, but from latent objects

The set of docs not ranked yet
Latent-ListMLE

The likelihood of a given rank list \( \vec{D} \):

\[
p(\vec{D}|q; w, \theta) = \prod_{i=1}^{n} \sum_{j=1}^{m} p(d_i|o_j, S_i)p(o_j|q)
\]

For each document For each latent object

- where:

\[
p(o_j|q) = \frac{\exp(\theta^T v_j)}{\sum_{k=1}^{m} \exp(\theta^T v_j)}
\]

Features

\[
p(d_i|o_j, S_i) = \frac{\exp(w^T u_{ij})}{\sum_{k=i}^{n} \exp(w^T u_{kj})}
\]

Parameters

Learning by maximizing the likelihood of best rank with EM
Relevant Entities: Three Ways to Find Related Entities

Candidate entities in the latent layer are selected from:

- **Query Annotation (AnnQ)**
  - Query: Obama Family Tree
  - Related Objects: Barack Obama, Family Tree

- **Entity Search (Osrch)**
  - Query: Obama Family Tree
  - External Data Index
  - Search: Barack Obama, Family Tree

- **Document Annotation (AnnD)**
  - Query: Obama Family Tree
  - Retrieved Documents: Family Tree, Michael Obama, US President
  - Related Objects: United States, White House

Diagram showing the interaction and selection of entities through query annotation, entity search, and document annotation.
Features

Query-Entity Features

• Textual similarities
  – BM25, language model, etc.
• Scores produced by related objects selectors
  – Annotation score
• Ontology overlap
  – In external data’s ontology
• Frequency of object in corpus annotation
  – “IDF” effect

Entity-Document Features

• Textual Similarities
  – BM25, language model, etc.
• Connection in the knowledge graph
  – Document reachable in 1-2 hops, etc.
• Ontology Overlap
  – In external data’s ontology
• Document quality
  – Spam score, URL length…
# Experiment Datasets

Two different scenarios

<table>
<thead>
<tr>
<th></th>
<th>Scenario 1</th>
<th>Scenario 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Corpus</strong></td>
<td>Web Page</td>
<td>Medical Abstracts</td>
</tr>
<tr>
<td></td>
<td>• ClueWeb09-B</td>
<td>• OSHUMED</td>
</tr>
<tr>
<td></td>
<td>• ClueWeb12-B13</td>
<td></td>
</tr>
<tr>
<td><strong>Query</strong></td>
<td>From Web Users (Bing’s)</td>
<td>From Experts</td>
</tr>
<tr>
<td></td>
<td>• TREC Web Track 2009-2013</td>
<td>• OHSUMED</td>
</tr>
<tr>
<td><strong>External Data</strong></td>
<td>Modern Knowledge Base</td>
<td>Controlled Vocabulary</td>
</tr>
<tr>
<td></td>
<td>• Freebase</td>
<td>• Medical Subject Headings (MeSH)</td>
</tr>
</tbody>
</table>

© 2016 Chenyan Xiong
Web Search Results: ClueWeb09-B

Baselines:
- Unsupervised Retrieval
- Supervised + Freebase
- Learning to Rank

Knowledge base helps!
- EsdRank-AnnQ (Query Annotation) outperforms all baselines, on all metrics, with large margins.

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP@100</th>
<th>NDCG@20</th>
<th>ERR@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDM</td>
<td>0.375</td>
<td>0.214</td>
<td>0.135</td>
</tr>
<tr>
<td>ESQFE</td>
<td>0.364</td>
<td>0.213</td>
<td>0.139</td>
</tr>
<tr>
<td>RankSVM</td>
<td>0.352↑,‡,§</td>
<td>0.193↑,‡</td>
<td>0.147</td>
</tr>
<tr>
<td>ListMLE</td>
<td>0.410↑,‡,§</td>
<td>0.221↑,‡</td>
<td>0.156↑,‡</td>
</tr>
<tr>
<td>EsdRank-AnnQ</td>
<td>0.437↑,‡,§,¶</td>
<td>0.237↑,‡,§,¶</td>
<td>0.190↑,‡,§,¶</td>
</tr>
<tr>
<td>EsdRank-0srch</td>
<td>0.397↑,‡,§,¶</td>
<td>0.219↑,‡,§,¶</td>
<td>0.155↑,‡,§,¶</td>
</tr>
<tr>
<td>EsdRank-AnnD</td>
<td>0.403↑,‡,§,¶</td>
<td>0.221↑,‡,§,¶</td>
<td>0.167↑,‡,¶,§</td>
</tr>
</tbody>
</table>

Similar results on ClueWeb12-B13
## Medical Search Results: OSHUMED

### Baselines:
- **Unsupervised Retrieval**
- **Learning to Rank**

### Consistent results on OSHUMED
- Same method
- Different corpus (smaller and cleaner)
- Different external data (MeSH)

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP@100</th>
<th>NDCG@20</th>
<th>ERR@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDM</td>
<td>0.413</td>
<td>0.332</td>
<td>0.453</td>
</tr>
<tr>
<td>RankSVM</td>
<td>0.396</td>
<td>0.336</td>
<td>0.453</td>
</tr>
<tr>
<td>ListMLE</td>
<td>0.414</td>
<td>0.343</td>
<td>0.483</td>
</tr>
<tr>
<td>EsdRank-AnnQ</td>
<td>0.428</td>
<td>0.355</td>
<td>0.511</td>
</tr>
<tr>
<td>EsdRank-Osrch</td>
<td>0.427</td>
<td>0.347</td>
<td>0.489</td>
</tr>
<tr>
<td>EsdRank-AnnD</td>
<td>0.416</td>
<td>0.342</td>
<td>0.500</td>
</tr>
</tbody>
</table>

© 2016 Chenyan Xiong
Outline

**Query Representation**
- Query Expansion with Freebase [Term-based] (ICTIR 2015)
- EsdRank [Entity-based] (CIKM 2015)
- Learning to Find Entity [Related Entity Finding] (SIGIR 2016 Short)

**Document Representation**
- Bag-of-entities [Entity-based] (Under review)

**Proposed Research**
- Hierarchal Learning to Rank for Multiple Representations
- Co-Learning: Entity Finding and Document Ranking
- Entity Graph Representation
Learning to Rank Related Entities

A Preliminary Study about Finding Better Entities
Learning to Rank Related Entities

Improve entity search to find better entities
Introduce LeToR to Entity Search

<table>
<thead>
<tr>
<th>Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virtual Document for An Entity</td>
</tr>
<tr>
<td>Name</td>
</tr>
<tr>
<td>Alias</td>
</tr>
<tr>
<td>Category  (Name of Types)</td>
</tr>
<tr>
<td>Attributes  (All Other Attributes)</td>
</tr>
<tr>
<td>Related Entity  (Connected Entities’ Names)</td>
</tr>
</tbody>
</table>

LeToR Models
- RankSVM
- Coor-Ascent

Ranking Features
- LM
- SDM
- BM25
- Cosine
- Boolean

Nikita Zhiltsov et al. SIGIR 2015
© 2016 Chenyan Xiong
Evaluation Results

Experiments on DBpedia search [Balog 2013]

- Four query types (Keywords and Questions), 485 queries

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP@100</th>
<th>P@10</th>
<th>P@20</th>
<th>NDCG@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDM–CA</td>
<td>0.192</td>
<td>0.198</td>
<td>0.155</td>
<td>0.294</td>
</tr>
<tr>
<td>MLM–CA</td>
<td>0.196</td>
<td>0.206</td>
<td>0.157</td>
<td>0.297</td>
</tr>
<tr>
<td><strong>FSDM</strong></td>
<td>0.231$^\dagger\ddagger$</td>
<td>0.231$^\dagger\ddagger$</td>
<td>0.179$^\dagger\ddagger$</td>
<td>0.339$^\dagger\ddagger$</td>
</tr>
<tr>
<td><strong>RankSVM</strong></td>
<td>0.246$^{†††}$</td>
<td>0.251$^{†††}$</td>
<td>0.193$^{†††}$</td>
<td>0.362$^{†††}$</td>
</tr>
<tr>
<td><strong>Coor–Ascent</strong></td>
<td>0.245$^{†††}$</td>
<td>0.248$^{†††}$</td>
<td>0.189$^{†††}$</td>
<td>0.358$^{†††}$</td>
</tr>
</tbody>
</table>

Prior best: Fielded SDM

New state-of-the-arts in Entity Search
Summary: Query Representation

Query Representation

• Query Expansion with Freebase [Term-based] (ICTIR 2015)
  – Enrich query with knowledge base terms
  – State-of-the-art in unsupervised ranking

• EsdRank [Entity-based] (CIKM 2015)
  – Learning to build entity representation and to rank together
  – State-of-the-art in supervised ranking

• Learning to Find Entity [Related Entity Finding] (SIGIR 2016 Short)
  – Start using supervisions in finding better entities
  – A preliminary work; state-of-the-art in entity search
Outline

Query Representation
• Query Expansion with Freebase [Term-based] (ICTIR 2015)
• EsdRank [Entity-based] (CIKM 2015)
• Learning to Find Entity [Related Entity Finding] (SIGIR 2016 Short)

Document Representation
• Bag-of-entities [Entity-based] (Under review)

Proposed Research
• Hierarchal Learning to Rank for Multiple Representations
• Co-Learning: Entity Finding and Document Ranking
• Entity Graph Representation
Bag-of-Entities Representation for Ranking

Moving to the Entity Space
## Classic Controlled Vocabulary Based IR

<table>
<thead>
<tr>
<th></th>
<th>Controlled Vocabularies</th>
<th>Bag-of-Words</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Information Unit</strong></td>
<td>Predefined Term</td>
<td>Word</td>
</tr>
<tr>
<td><strong>Structure</strong></td>
<td>Ontology</td>
<td>Flat</td>
</tr>
<tr>
<td><strong>Text Understanding</strong></td>
<td>Domain Expertise</td>
<td>Shallow</td>
</tr>
<tr>
<td><strong>Cost</strong></td>
<td>Manual Efforts</td>
<td>Automatic</td>
</tr>
<tr>
<td><strong>Coverage</strong></td>
<td>Small</td>
<td>Full</td>
</tr>
<tr>
<td><strong>Usage</strong></td>
<td>Specific Domain</td>
<td>General Domain</td>
</tr>
</tbody>
</table>
## From Controlled Vocabulary to Knowledge Base

<table>
<thead>
<tr>
<th></th>
<th>Controlled Vocabularies</th>
<th>Bag-of-Words</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Information Unit</strong></td>
<td>Predefined Term</td>
<td>+ Word</td>
</tr>
<tr>
<td><strong>Structure</strong></td>
<td>Ontology</td>
<td>+ Flat</td>
</tr>
<tr>
<td><strong>Text Understanding</strong></td>
<td>Domain Expertise</td>
<td>+ Shallow</td>
</tr>
<tr>
<td><strong>Cost</strong></td>
<td>Manual Efforts</td>
<td>- Automatic</td>
</tr>
<tr>
<td><strong>Coverage</strong></td>
<td>Small</td>
<td>- Full</td>
</tr>
<tr>
<td><strong>Usage</strong></td>
<td>Specific Domain</td>
<td>- General Domain</td>
</tr>
</tbody>
</table>

Knowledge Base:

- Automatic Entity Linking
- Large Scale Modern KBs
- Our Approaches Work for General Domain Queries
It is time to revisit the classic controlled vocabularies based approach, with modern knowledge bases.

<table>
<thead>
<tr>
<th>Information Unit</th>
<th>Controlled Vocabularies</th>
<th>Bag-of-Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predefined Term</td>
<td>+</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Text Understanding</th>
<th>Domain Expertise</th>
<th>Shallow</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cost</th>
<th>Manual Efforts</th>
<th>Automatic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coverage</th>
<th>Small</th>
<th>Full</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Usage</th>
<th>Specific Domain</th>
<th>General Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

Knowledge Base:

- **Automatic Entity Linking**
- **Large Scale Modern KBs**
- **Our Approaches Work for General Domain Queries**
Bag-of-Entities Representation

Represent a document by its entity annotations

Document

Carnegie Mellon University is a private research university in Pittsburgh, Pennsylvania

Entity Linking

Three Linking Systems:
- Google FACC1, High Precision
- CMNS, High Recall
- TagMe, Balanced

Two Ranking Models
- Coordinate Match (COOR)
- Entity Frequency (EF)

Bag-of-Entities

CMU University
Pittsburgh PA
Evaluation: Entity Linking Coverage

Coverage on ClueWeb09 Queries and Documents

- The annotation is no longer sparse

<table>
<thead>
<tr>
<th></th>
<th>Query</th>
<th>Document</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Freq</td>
<td>Dens</td>
</tr>
<tr>
<td>FACC1</td>
<td>0.42</td>
<td>0.20</td>
</tr>
<tr>
<td>TagMe</td>
<td>1.54</td>
<td>0.70</td>
</tr>
<tr>
<td>CMNS</td>
<td>1.50</td>
<td>0.69</td>
</tr>
</tbody>
</table>

Results on ClueWeb12 are similar.
Evaluation: Entity Linking

Entity linking’s Accuracy on Queries
• Ground truth: manual labels [Dalton 2014] and [Liu 2015]

50% ~ 60% Accuracy, Good Enough?

<table>
<thead>
<tr>
<th></th>
<th>ClueWeb09 Query</th>
<th></th>
<th>ClueWeb12 Query</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec</td>
<td>Rec</td>
<td>F1</td>
<td>Prec</td>
</tr>
<tr>
<td>FACC1</td>
<td>0.274</td>
<td>0.236</td>
<td>0.254</td>
<td>NA</td>
</tr>
<tr>
<td>TagMe</td>
<td>0.581</td>
<td>0.597</td>
<td>0.589</td>
<td>0.460</td>
</tr>
<tr>
<td>CMNS</td>
<td>0.577</td>
<td>0.596</td>
<td>0.587</td>
<td>0.485</td>
</tr>
</tbody>
</table>
## Evaluation: BOE Ranking ClueWeb09

### Ranking Performance of BOE, ClueWeb09

<table>
<thead>
<tr>
<th>Model</th>
<th>NDCG@20</th>
<th>ERR@20</th>
<th>Win/Tie/Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lm</td>
<td>0.176</td>
<td>-12.92%</td>
<td>39/88/71</td>
</tr>
<tr>
<td>SDM</td>
<td>0.202†</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>FACC1-COOR</td>
<td>0.173</td>
<td>-14.16%</td>
<td>64/56/78</td>
</tr>
<tr>
<td>FACC1-EF</td>
<td>0.167</td>
<td>-17.32%</td>
<td>63/51/84</td>
</tr>
<tr>
<td>TagMe-COOR</td>
<td>0.211†</td>
<td>4.55%</td>
<td>108/35/55</td>
</tr>
<tr>
<td>TagMe-EF</td>
<td>0.229†,‡</td>
<td>13.71%</td>
<td>96/24/78</td>
</tr>
<tr>
<td>CMNS-COOR</td>
<td>0.210†</td>
<td>4.08%</td>
<td>105/37/56</td>
</tr>
<tr>
<td>CMNS-EF</td>
<td>0.216†</td>
<td>6.97%</td>
<td>97/22/79</td>
</tr>
</tbody>
</table>

Results on ClueWeb12 are similar.

BOE Works Well with Simple Ranking Models

Standard BOW Ranking (Indri)
Summary: Current Progress

Query Representation
• Effective term based and entity based representation from knowledge bases
• A preliminary study about finding better entities

Documents Representation
• BOE work shows the potential of entity based document representation

Conclusion
• Knowledge bases have the power in search
• IR is ready to move from the word space to the knowledge space
Summary: Challenges

Query Representation
• The quality of related entities
  – The bottleneck of all our systems

Document Representation
• The uncertainties from entity representations
  – Brings new noises to ranking systems
• The heterogeneity of the entity space
  – Ranking intuitions of words may not be suitable for entities

Knowledge base is much richer than individual terms/entities
• Connections and relationships between entities
• How to use them?
PROPOSED RESEARCH
Outline: Proposed Research

Proposed Research
• Hierarchal Learning to Rank for Multiple Representations
  – Handle the uncertainties and incorporate more evidence from entities
• Co-Learning: Entity Finding and Document Ranking
  – Find better relevant entities
• Entity Graph Representation
  – Utilize entity connections
Hierarchical Ranking with Multiple Representations

Two Layer Attention Based Ranking Model

- To incorporate more evidence, and to handle BOE’s uncertainty
- Learning to rank and combine multiple representations jointly

1st Layer: Individual Ranking Models
- BOE Uncertainty Features
- Entity Features
- LeToR with More Evidence from Entities

2nd Layer: Attend the more confident one
- BOW Uncertainty Features
- Word Features
- Classic LeToR

Final Ranking

$f(q, d)$
Co-Learning: Ranking and Entity Finding

Multi-task learning for entity finding and document ranking

- Finding relevant entities towards best ranking performance
- Seek supervision from both entity side and document side

Query: Carnegie Mellon

Documents:
- Carnegie Mellon Today
- Human Resources - Carnegie - Carnegie Mellon University
- Carnegie - Carnegie International Co - Carnegie Center Merit Award

Entities:
- CMU
- Computer
- Pittsburgh
- Tartans

LeToR with document labels

Propagate information round with document-entity connections

LeToR with entity labels
Entity Graph Representation

From Individual Entities

Carnegie Mellon University (CMU), is a private research university in Pittsburgh, Pennsylvania. Founded in 1900 by Andrew Carnegie as the Carnegie Technical Schools......

To Connected Entity Graph

Carnegie Mellon University

CMU

University

Pittsburgh

Entities

Alias

Located In

Connections

Founded By

Use to be

Terms

Research

Andrew Carnegie

Carnegie Technical Schools

For Better Ranking

Promote Central Entities:

Carnegie Mellon University

Structural Ranking

Query:

CMU Founder

Document:

Alignment

CMU

Founded By

Andrew Carnegie
## Entity Graph Representation

### Entity Connections
- Closed Form Relationship
- Open Information Extraction
- Relationship Language Model

### Inference Methods
- CRF Topic Modeling
- Learn from Key Phrases
- Learn from Manual Labels

### Ranking Methods
- BOE with Better Node Weights
- Graph Ranking Models
- Graph Embedding
Query Representation:

Co-Learning: Entity Finding and Document Ranking

08/2016

11/2016

03/2017

01/2018

Document Representation:

Hierarchical Learning to Rank with Multiple Representations

Entity Graph Representation
- Node Based Inference
- Explore Different Connections
- Ranking Models

Dissertation Writing & Job Hunting

06/2018

Planed Graduation Date
Conclusion

Query Representation:

- Query
- Carnegie Mellon

Documents

- Document
- Carnegie Mellon University is a private research university in Pittsburgh, Pennsylvania

March Towards more *Intelligent*, *Semantic* and *Structured* Information Retrieval

- State-of-the-art in Unsupervised Ranking
  - [ICTIR 2015]
- State-of-the-art in Supervised Ranking
  - [CIKM 2015]
- State-of-the-art in Entity Search
  - [SIGIR 2016]

Co-Learning: Entity Finding and Document Ranking

Hierarchical Learning to Rank with Multiple Representations

Entity Graph Representation

Boolean Retrieval works well already

© 2016 Chenyan Xiong
QUESTION & ANSWERING
EsdRank: More Details
Latent-ListMLE: Learning

Learn parameters by Maximize the likelihood of best rank \( \overrightarrow{D^*} \):

\[
w^*, \theta^* = \arg \max_{w, \theta} \log p(\overrightarrow{D^*} | q, w, \theta) \\
= \arg \max_{w, \theta} \sum_{i=1}^{n} \log \sum_{j=1}^{m} p(d_i | o_j, S_i) p(o_j | q)
\]

EM method is used to solve this problem:

• E step: find the posterior of latent objects
  – “Propagate document level training data to latent layer”
• M step: maximize the expected complete likelihood
  – “Fit the parameters on current belief”
• Iterate until converge
Latent-ListMLE: Learning
E-Step

E step, find the posterior of latent variable given current $w, \theta$:

$$\pi(o_j | d_i, q) = p(o_j | d_i, q; \theta_{old}, w_{old}) = \frac{1}{Z} p(d_i | o_j, S_i; w_{old})p(o_j | q; \theta_{old})$$

- Posterior of $o_j$
- Normalizer
- Defined in the generative process
Latent-ListMLE: Learning M-Step

M step, find \( w^*, \theta^* \) that maximizes:

\[
E_{\pi(\hat{o} | D^*, q)} \log p(D^*, \hat{O} | q; w, \theta)
= \sum_{i=1}^{n} \sum_{j=1}^{m} \pi(o_j | d_i, q) \log p(d_i | o_j, S_i)p(o_j | q)
\]

Summations only outside log

Consider \( o_j \) as observed with posterior probability \( \pi(o_j | d_i, q) \)
**Web Search Results: ClueWeb12-B13**

**Baselines:**
- Unsupervised Retrieval
- Supervised + Freebase
- Learning to Rank

- Work well on ClueWeb12-B13
  - AnnQ outperforms all baselines, too
  - Fewer training queries, less improvements
    - We have a more complex model

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP@100</th>
<th>NDCG@20</th>
<th>ERR@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDM</td>
<td>0.293 (−26.19%)</td>
<td>0.126 (−27.53%)</td>
<td>0.091 (−25.66%)</td>
</tr>
<tr>
<td>EQFE</td>
<td>0.319 (−19.74%)</td>
<td>0.146 (−16.02%)</td>
<td>0.106 (−13.61%)</td>
</tr>
<tr>
<td>RankSVM</td>
<td>0.383† (−3.66%)</td>
<td>0.161† (−7.29%)</td>
<td>0.114 (−7.48%)</td>
</tr>
<tr>
<td>ListMLE</td>
<td>0.397†,‡,¶</td>
<td>0.174†,‡,¶</td>
<td>0.123†</td>
</tr>
<tr>
<td>EsdRank−AnnQ</td>
<td>0.410†,‡,¶ (3.12%)</td>
<td>0.181†,‡,¶ (3.78%)</td>
<td>0.133†,‡,¶,∥ (8.39%)</td>
</tr>
<tr>
<td>EsdRank−Osrch</td>
<td>0.391†,‡,¶,∥ (−1.51%)</td>
<td>0.167† (−3.88%)</td>
<td>0.121† (−1.26%)</td>
</tr>
<tr>
<td>EsdRank−AnnD</td>
<td>0.440†,‡,¶,∥ (10.69%)</td>
<td>0.186†,‡,¶,∥ (7.15%)</td>
<td>0.132†,¶,∥ (7.67%)</td>
</tr>
</tbody>
</table>
Entity Search: More Details
Prior State-of-the-Art

Fielded Documents + SDM

RDF of an Entity

Name
Alias
Types
Location
Contains

Virtual Document for the Entity

Name
Alias
Category
((Name of Types)
Attributes
(All Other Attributes)
Related Entity
(Connected Entities’ Names)

Nikita Zhiltsov et al. SIGIR 2015
BOE: More Details
Error Analysis

Defining Factor: Uncertainties from Automatic Entity Linking

<table>
<thead>
<tr>
<th></th>
<th>NDCG@20</th>
<th>ERR@20</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TagMe-COOR</strong></td>
<td>0.214†‡</td>
<td>14.56%</td>
</tr>
<tr>
<td><strong>TagMe-EF</strong></td>
<td>0.243†‡</td>
<td>30.43%</td>
</tr>
<tr>
<td><strong>CMNS-COOR</strong></td>
<td>0.211†</td>
<td>4.28%</td>
</tr>
<tr>
<td><strong>CMNS-EF</strong></td>
<td>0.240†‡</td>
<td>18.44%</td>
</tr>
</tbody>
</table>

When the query annotation is correct, huge improvements

<table>
<thead>
<tr>
<th></th>
<th>NDCG@20</th>
<th>ERR@20</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TagMe-COOR</strong></td>
<td>0.200†</td>
<td>-3.28%</td>
</tr>
<tr>
<td><strong>TagMe-EF</strong></td>
<td>0.209</td>
<td>0.65%</td>
</tr>
<tr>
<td><strong>CMNS-COOR</strong></td>
<td>0.201†</td>
<td>3.90%</td>
</tr>
<tr>
<td><strong>CMNS-EF</strong></td>
<td>0.185</td>
<td>-4.09%</td>
</tr>
</tbody>
</table>

When the query annotation is wrong, may hurt