Entity Hierarchy Embedding

Zhiting Hu, Poyao Huang, Yuntian Deng, Yingkai Gao, Eric Xing
School of Computer Science
Carnegie Mellon University

Outline

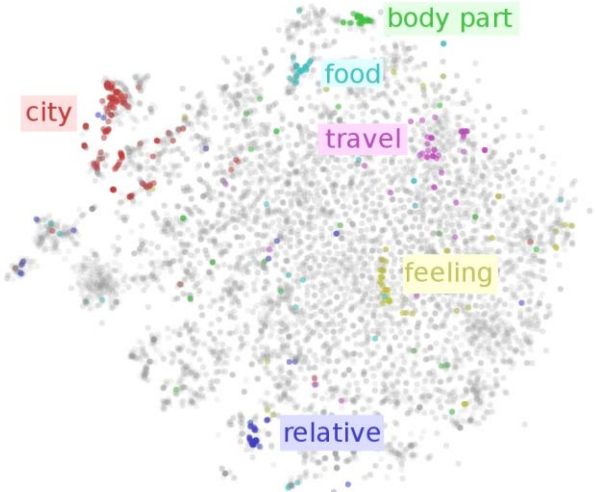
- Background
    - Distributed representation
- Entity hierarchy embedding
- Applications & Experiments
    - Entity linking
    - Entity search

Outline

- Background
  - Distributed representation
- Entity hierarchy embedding
- Applications & Experiments
  - Entity linking
  - Entity search

Distributed Representation

- Learn compact vectors (a.k.a. embedding) for
  - words [Mikolov et al., 2013, Bengio, et al. 2003, C&W, 2008]
  - phrases [Passos et al., 2014]
  - concepts [Hilland Korhonen, 2014]



http://colah.github.io/posts/2015-01-Visualizing-Representations/

Distributed Representation

- Learn compact vectors (a.k.a. embedding) for
  - words [Mikolov et al., 2013, Bengio, et al. 2003, C&W, 2008]
  - phrases [Passos et al., 2014]
  - concepts [Hilland Korhonen, 2014]

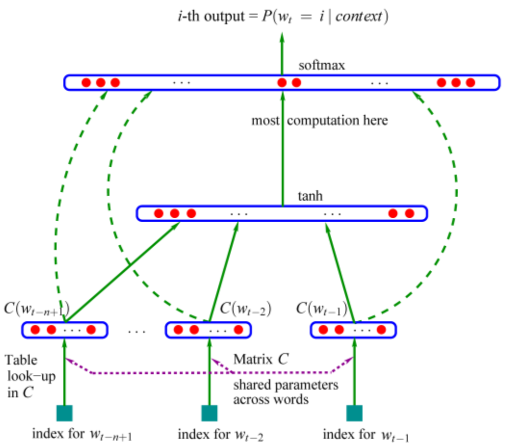
- Expected to capture semantic relatedness of the words/concepts

Distributed Representation

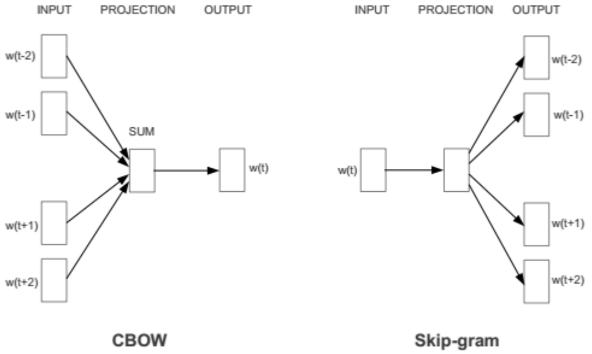
- Learn compact vectors (a.k.a. embedding) for
  - words [Mikolov et al., 2013, Bengio, et al. 2003, C&W, 2008]
  - phrases [Passos et al., 2014]
  - concepts [Hilland Korhonen, 2014]

- Expected to capture semantic relatedness of the words/concepts
- Widely used to improve performance
  - sentiment analysis [Tang et al., 2014], machine translation [Zhang et al., 2014], information retrieval [Clinchant and Perronnin, 2013], video understanding [Chang et al., 2015], etc.

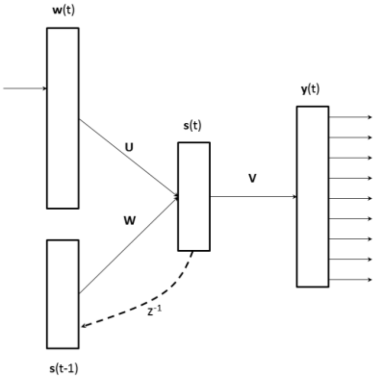
Distributed Representation



NNLM [Bengio, et al. 2003]

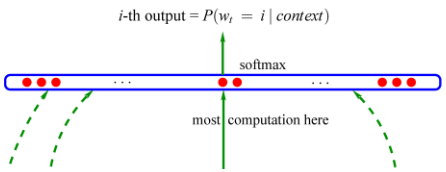


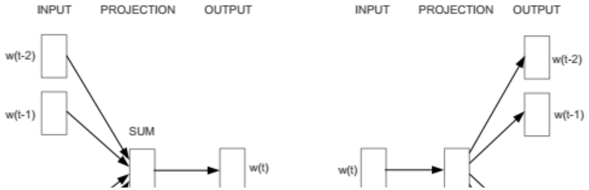
CBOW & Skip-gram
[Mikolov, et al. 2013]



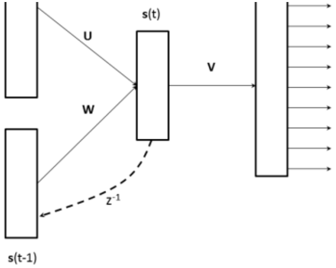
RNNLM [Mikolov, et al. 2010]

Distributed Representation





- Induce word/phrase embedding from *free text*
- Limited in utilizing *structured knowledge*



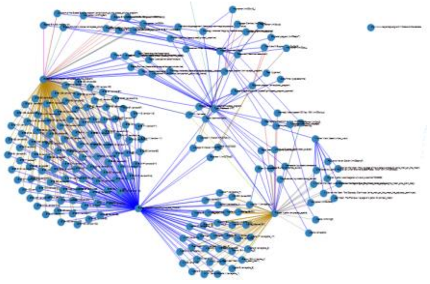
RNNLM [Mikolov, et al. 2010]

Structured Knowledge

- Knowledge bases
  - Wikipedia, Freebase, Dbpedia, …

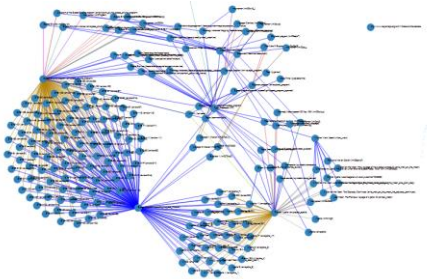
Structured Knowledge

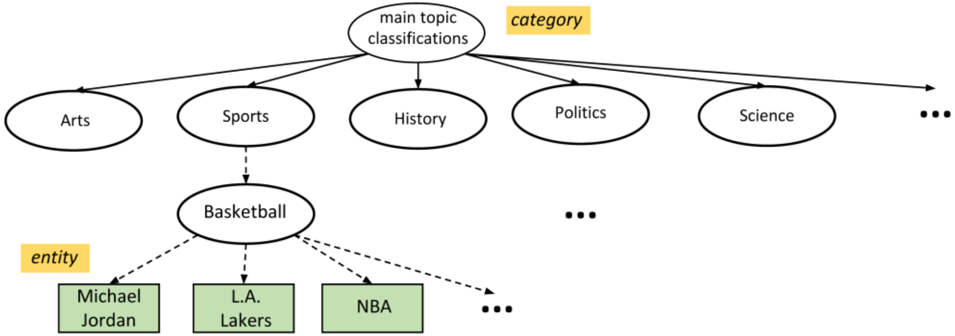
- Knowledge bases
  - Wikipedia, Freebase, Dbpedia, …
- Entities, relations
  - recent work, e.g., TransE [Bordes et al., 2011; Wang et al., 2014; Lin et al., 2015], learns entity vectors from the relational structure
  - usually does not incorporate text
  - lacks an explicit entity relatedness measure



Structured Knowledge

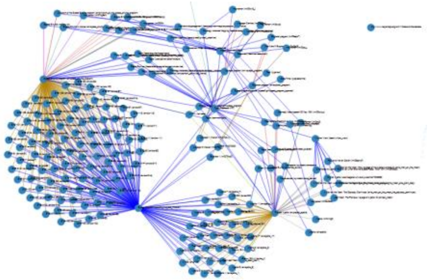
- Knowledge bases
  - Wikipedia, Freebase, Dbpedia, …
- Entities, relations
  - recent work, e.g., TransE [Bordes et al., 2011; Wang et al., 2014; Lin et al., 2015], learns entity vectors from the relational structure
  - usually does not incorporate text
  - lacks an explicit entity relatedness measure
- Entity hierarchies



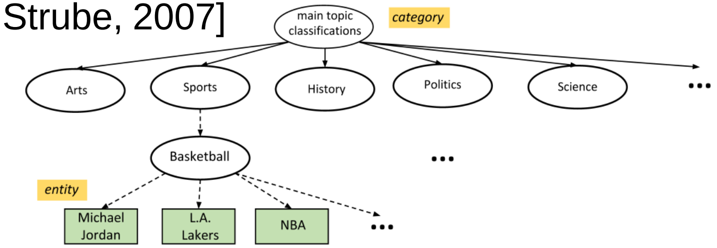


Structured Knowledge

- Knowledge bases
  - Wikipedia, Freebase, Dbpedia, …
- Entities, relations
  - recent work, e.g., TransE [Bordes et al., 2011; Wang et al., 2014; Lin et al., 2015], learns entity vectors from the relational structure
  - usually does not incorporate text
  - lacks an explicit entity relatedness measure



- Entity hierarchies
  - encode rich knowledge on entity relatedness
  - heuristic use: hand-crafted features [Ponzetto & Strube, 2007]
  - few distributed representation has incorporated hierarchical knowledge



This work: entity hierarchy embedding

- Integrates *hierarchical structure* from KBs into distributed representation learning
- Develops a principled optimization-based framework
  - incorporating both free text and hierarchical structure
  - efficient to handle large complex hierarchies

Outline

- ❑ Background
  - ❑ Distributed representation
- ❑ Entity hierarchy embedding
- ❑ Applications & Experiments
  - ❑ Entity linking
  - ❑ Entity search

Recap: skip-gram word embedding

- Objective: find a representation for each word that is useful for predicting its context

Apple released their first Apple Watch update.

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t)$$

$$p(w_C|w_T) = \frac{\exp\left\{v_{w_C}^{\top\prime} v_{w_T}\right\}}{\sum_{w\in\mathcal{V}}\exp\left\{v_w^{\top\prime} v_{w_T}\right\}}$$

Recap: skip-gram word embedding

- Objective: find a representation for each word that is useful for predicting its context

Apple released their first Apple Watch update.

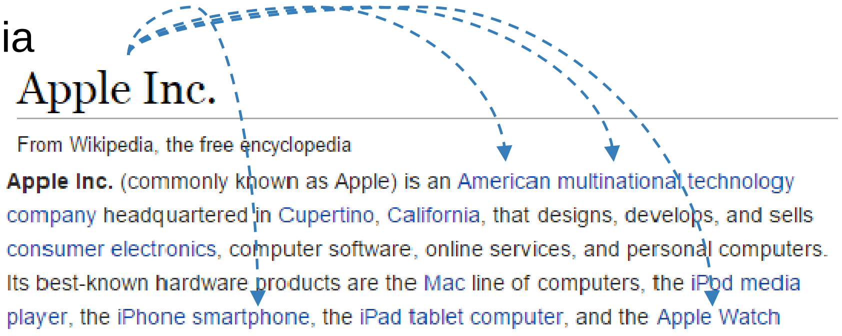
1) Context of a word
   - words surrounding the target word
2) Similarity measure of context prediction
   - inner-product

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t)$$

$$p(w_C|w_T) = \frac{\exp\{v_{w_C}^{\top\prime} v_{w_T}\}}{\sum_{w\in\mathcal{V}}\exp\{v_{w}^{\top\prime} v_{w_T}\}}$$

Entity hierarchy embedding

- Objective: find a representation for each entity that is useful for predicting its context
- Entity: each corresponds to an encyclopedia article in KB (e.g. Wikipedia)



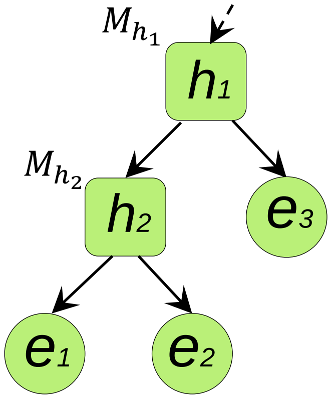
1) Context of an entity
   - entities occurs in its encyclopedia article
   - entity annotations are readily available
2) Similarity measure of context prediction
   - incorporates entity hierarchy

$$p(e_C|e_T) = \frac{\exp\left\{-d\left(e_T, e_C\right)\right\}}{\sum_{e\in\mathcal{E}} \exp\left\{-d\left(e_T, e\right)\right\}}$$

Incorporating hierarchy

$$p(e_C|e_T) = \frac{\exp\{-d(e_T, e_C)\}}{\sum_{e\in\mathcal{E}} \exp\{-d(e_T, e)\}}$$

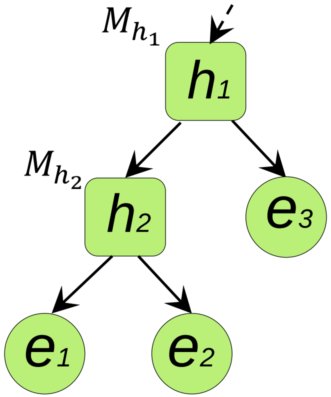
- Distance metric learning and aggregation



Incorporating hierarchy

$$p(e_C|e_T) = \frac{\exp\{-d(e_T, e_C)\}}{\sum_{e \in \mathcal{E}} \exp\{-d(e_T, e)\}}$$

- Distance metric learning and aggregation
  - associate a separate distance metric $M_h \in R^{n \times n}$ (n: dimension of the embedding) with each category node h
  - measure the distance between two entities under some *aggregated* distance metric



Incorporating hierarchy

$$p(e_C|e_T) = \frac{\exp\{-d(e_T, e_C)\}}{\sum_{e \in \mathcal{E}} \exp\{-d(e_T, e)\}}$$

- Distance metric learning and aggregation
  - associate a separate distance metric $M_h \in R^{n \times n}$ (n: dimension of the embedding) with each category node h
  - measure the distance between two entities under some *aggregated* distance metric

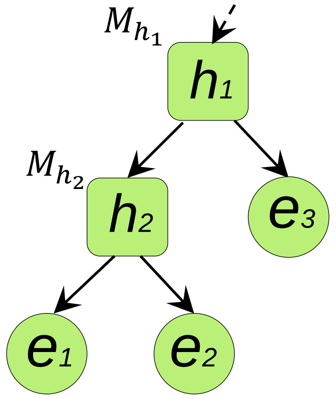
Mahalanobis distance

$$d(e_1, e_2) = \left(v_{e_1} - \bar{v}_{e_2}\right)^T M_{e_1,e_2} (v_{e_1} - \bar{v}_{e_2})$$

$$d(e_1, e_3) = \left(v_{e_1} - \bar{v}_{e_3}\right)^T M_{e_1,e_3} (v_{e_1} - \bar{v}_{e_3})$$

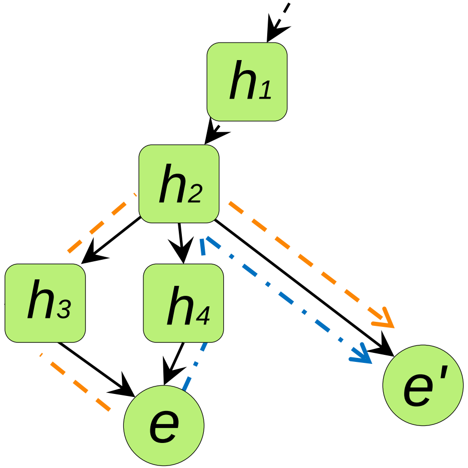
$v_e: entity\ vector\ as\ a\ target$

$\bar{v}_e: entity\ vector\ as\ a\ context$



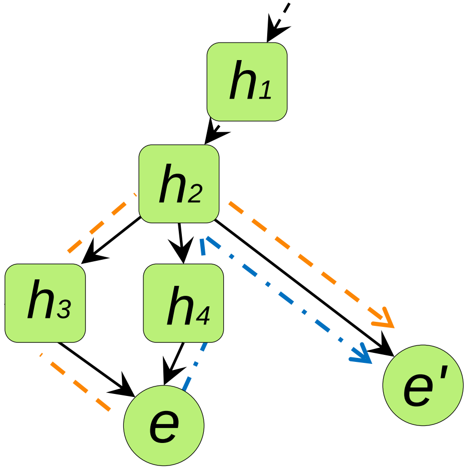
Metric aggregation

- Given two entities e and e' , $M_{e,e'} \in R^{n \times n}$
- A naïve approach
  - $M_{e,e'} := \sum_{h \in P_{e,e'}} M_h$
  - $P_{e,e'}$: path between e and e' in the hierarchy
- Problem
  - entity hierarchy usually has complex DAG structure
  - many paths between two entities
  - use only the shortest path?
    - ignore other related category nodes
    - fail to capture the full aspects of entity relatedness



Metric aggregation

- Given two entities e and e' , $M_{e,e'} \in R^{n \times n}$
- A naïve approach
  - $M_{e,e'} := \sum_{h \in P_{e,e'}} M_h$
  - $P_{e,e'}$: path between e and e' in the hierarchy
- Problem
  - entity hierarchy usually has complex DAG structure
  - many paths between two entities
  - use only the shortest path?
    - ignore other related category nodes
    - fail to capture the full aspects of entity relatedness
- An ideal scheme
  - taking into account all possible paths/related categories between two entities
  - efficient to handle large complex hierarchy



Metric aggregation (cont.)

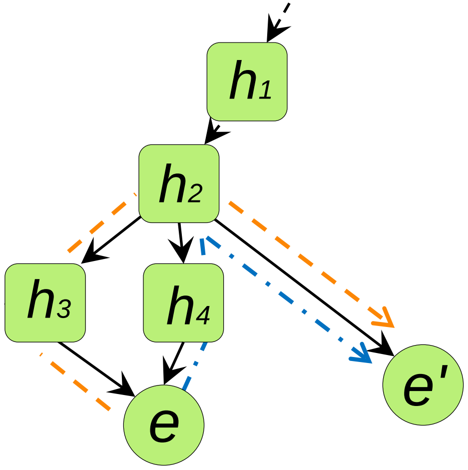
- Extend $P_{e,e'}$:
  - the set of all category nodes in any of the $e \rightarrow e'$ paths
- Aggregated metric:

$$M_{e,e'} = \gamma_{e,e'} \sum_{h \in \mathcal{P}_{e,e'}} \pi_{ee',h} M_h$$

scaling factor, ∝ distance between the least common ancestor and e/e'

$\sum_{h \in \mathcal{P}_{e,e'}} \pi_{ee',h} = 1$

- balance the size of P across different entity pairs
- $\pi_{ee',h}$ ∝ distance between h and e/e'



Metric aggregation (cont.)

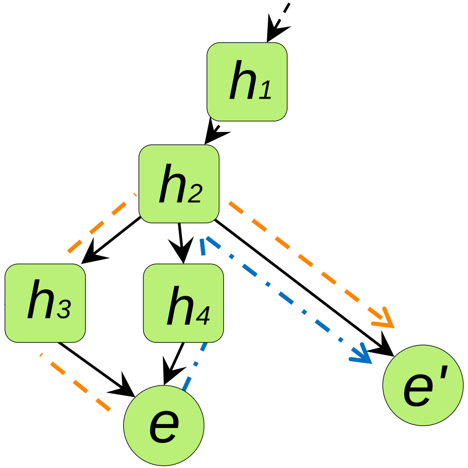
- Extend $P_{e,e'}$:
  - the set of all category nodes in any of the $e \rightarrow e'$ paths
- Aggregated metric:

$$M_{e,e'} = \gamma_{e,e'} \sum_{h \in \mathcal{P}_{e,e'}} \pi_{ee',h} M_h$$

scaling factor, ∝ distance between the least common ancestor and e/e'

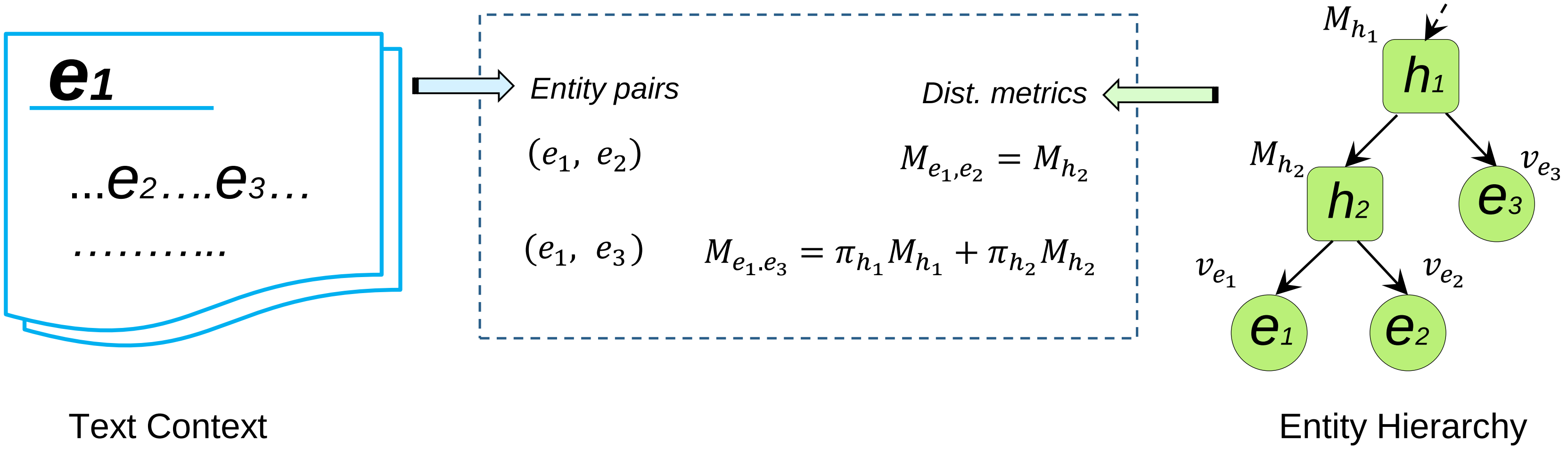
$$\sum_{h \in \mathcal{P}_{e,e'}} \pi_{ee',h} = 1$$

- balance the size of P across different entity pairs
- $\pi_{ee',h}$ ∝ distance between h and e/e'



- Develop an efficient algorithm to find $\{P_{e,e'}, \pi_{ee',\cdot}, \gamma_{e,e'}\}$
  - time complexity $O(\#child\ of\ two\ entities'common\ ancestors)$ *(Theorem 1)*

Summing up



$$p(e_C|e_T) = \frac{\exp\left\{-d\left(e_T, e_C\right)\right\}}{\sum_{e\in\mathcal{E}} \exp\left\{-d\left(e_T, e\right)\right\}}$$

$$\mathcal{L} = \frac{1}{|\mathcal{D}|} \sum_{(e_T, e_C)\in\mathcal{D}} \log p(e_C|e_T)$$

Outline

- ❑ Background
  - ❑ Distributed representation
- ❑ Entity hierarchy embedding
- ❑ Applications & Experiments
  - ❑ Entity linking
  - ❑ Entity search

Experiments

Training data:

- Wikipedia entities and categories
- 4.1M entities, 0.8M categories, 12 layers
- 87.6M entity pairs extracted from Wikipedia text corpus

- 100-dim entity vectors
- 100x100-dim category distance metrics (restricted to be diagonal)

Entity Linking

- link surface forms (mentions) of entities in a document to entities in a reference KB
- "Apple released an operating system Lion": *Apple Inc.* & *Mac OS X Lion*
- Intuition: entities in a document tend to be semantically related

entity assignments and mentions in a document

relatedness of e_{m_i} to other entities in the document

$$P(\mathcal{A}|\mathcal{M}) \propto \prod_{i=1}^{M} P(e_{m_i}|m_i) \sum_{\substack{j=1 \\ j\neq i}}^{M} \frac{1}{d\left(e_{m_i}, e_{m_j}\right) + \epsilon}$$

mention-to-entity compatibility score, $\propto$ frequency that m_i refer to e_{m_i} in Wikipedia

Results

- Dataset: IITB (http://www.cse.iitb.ac.in/soumen/doc/CSAW/Annot)
  - ~100 docs, 17K mentions
  - we use only the mentions whose referent entities are contained in Wikipedia (i.e., excludes NIL)

Methods	Precision	Recall	F1
CSAW	0.65	0.74	0.69
Entity-TM	0.81	0.80	0.80
Ours-NoH	0.78	0.85	0.81
Ours	**0.87**	**0.94**	**0.90**

Table 1: Entity linking performance

Entity Search

- Query: a natural language question Q and one or more desired entity categories C
  - $Q =$"films directed by Akira Kurosawa", $C =${Japanese films}
- Retrieve a list of relevant entities in response to the query

Our method:

- Identify referent entities of the mentions in Q
  - Film, Akira Kurosawa
  - augment the short query text with background knowledge
- Find the most related entities within the categories in C

Results

Dataset: INEX 2009 entity ranking track
(http://www.inex.otago.ac.nz/tracks/entityranking/entity-ranking.asp)

- 55 queries

Methods	Precision@10	Precision@R
Balog	0.18	0.16
K&K	0.31	0.28
Chen	0.55	0.42
Ours	**0.57**	**0.46**

Table 2: Entity search performance.

Qualitative analysis

- Entity vectors
- Most relevant entities in a given category

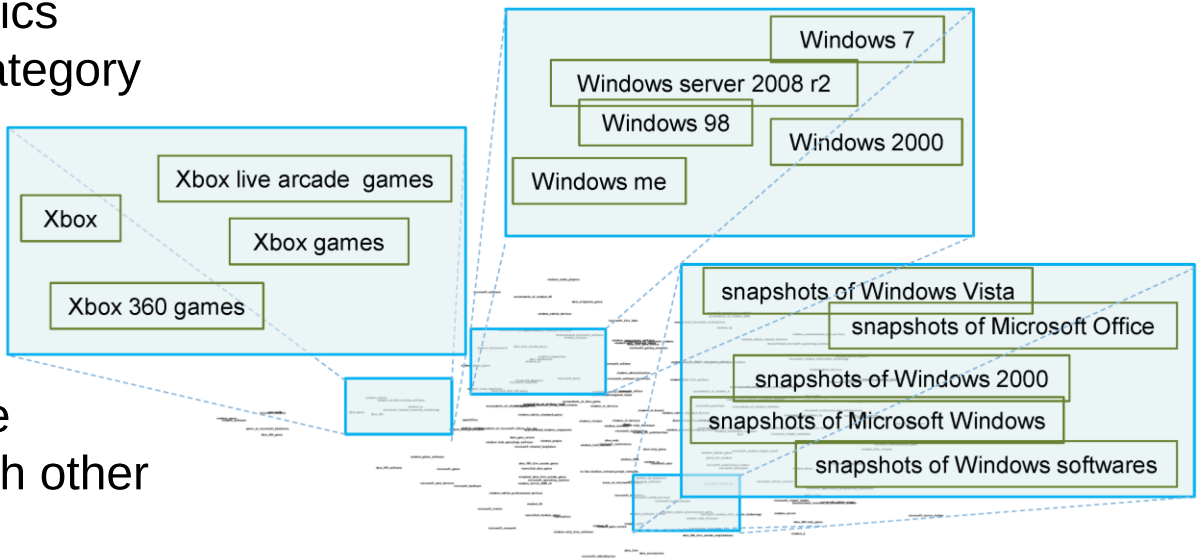
- Applications in semantic search, recommendation, knowledge base completion, …

Target entity	Most related entities	
black hole	**overall:** faster-than-light event horizon white hole time dilation	**American films:** Hidden Universe 3D Hubble (film) Quantum Quest Particle Fever
Youtube	**overall:** Instagram Twitter Facebook Dipdive	**Chinese websites:** Tudou 56.com Youku YinYueTai
Harvard University	**overall:** Yale University University of Pennsylvania Princeton University Swarthmore College	**businesspeople in software:** Jack Dangermond Bill Gates Scott McNealy Marc Chardon
X-Men: Days of Future Past (film)	**overall:** Marvel Studios X-Men: The Last Stand X2 (film) Man of Steel (film)	**children's television series:** Ben 10: Race Against Time Kim Possible: A Sitch in Time Ben 10: Alien Force Star Wars: The Clone Wars

Table 3: Most related entities under specific categories. "Overall" represents the most general category that includes all the entities.

Qualitative analysis

- Category distance metrics
- Subcategories of the category ``Microsoft''



- Relevant categories are embedded close to each other

Conclusion

- Incorporate hierarchical knowledge in distributed representation learning
  - exploit both text context and entity hierarchy
  - distance metric learning and aggregation
  - efficient algorithm for aggregation

- Improve entity linking and entity search

- Promising qualitative results

Future work
- Incorporate other sources of knowledge

Thanks!