Learning Semantic Entity Representations with Knowledge Graph and Deep Neural Networks and its Application to Named Entity Disambiguation

Hongzhao Huang¹ and Larry Heck²

Computer Science Department, Rensselaer Polytechnic Institute¹

Microsoft Research²

{huangh9@rpi.edu, Larry.Heck@microsoft.com}

Specific Thanks Yelong Shen and Gustavo Abrego for the help on deep neural network related issues
Word Embeddings

- Standard word representation
  - “One-hot” representation
    - Microsoft [0, 0, 0, ..., 0, 1, 0, ..., 0]
- Neural word embeddings
  - Distributed representation
    - Microsoft [0.453, -0.292, 0.732, ..., -0.243]
  - Represent a word by its contextual surrounding words
    “You shall know a word by the company it keeps”
    (J. R. Firth 1957: 11)
  - government debt problems turning into banking crises as has happened in
  - saying that Europe needs unified banking regulation to replace the hodgepodge
  Examples from (Socher et al, NAACL2013 tutorial)
From Word Embeddings to Entity Embeddings

• How about entities?
  o Usually composed of multiple words
    • Microsoft Research, James Cameron, Atlanta Hawks
  o Entities play crucial role in many applications
    • Entity Linking, Relation Extraction, Question & Answering...

• Our goal
  o Learn task specific accurate semantic entity representations
How can we represent entities?

• How we learn about a new entity/concept?

James Cameron

Film director

James Francis Cameron is a Canadian film director, film producer, deep-sea explorer, screenwriter, and editor who has directed the two biggest box office films of all time. He first found success with the science-fiction hit The Terminator. Wikipedia

• <James Cameron, film director, Titanic>
• <James Cameron, won awards, Academy Award for Best Picture>

....
Semantic Knowledge Graphs (KGs)

- A graph composed of:
  - Nodes: uniquely identified entities or literals
  - Edges: semantic relations
    - E.g., film director, film producer, CEO of...
- Many rich and clean KGs
  - Satori, Google KG, Freebase, Dbpedia....
- Broad applications to natural language processing and spoken language understanding
  - E.g., Unsupervised semantic parsing (Heck et al, 2012)
    - Use KG to guide automatic labeling of training instances
- This work: encode world knowledge from KG to assist deep understanding and accurate semantic representations of entities
Semantic Knowledge Graphs: An Example

James Cameron

James Francis Cameron (born August 16, 1954) is a Canadian film director, film producer, deep-sea explorer, screenwriter, and editor who has directed the two biggest box office films of all time. He first found success with the science fiction hit The Terminator (1984). He then became a popular Hollywood director and was hired to write and direct Aliens (1986); three years later he followed up with The Abyss (1989).

[en.wikipedia.org](http://en.wikipedia.org)

Born: August 16, 1954
Height: 6' 2" (1.88 M)

Spouse:
- Suzy Amis (2000)
- Gale Anne Hurd (1985 - 1989)
- Sharon Williams (1978 - 1984)

Awards:
- Mainichi Film Concours Readers' Choice Award for Best Film 1999
- Academy Award for Best Film Editing 1997
- Academy Award for Best Director 1997
- Academy Award for Best Picture 1997

Films:
- Solartaxi: Around the World with the Sun (2010)
- Explorers: From the Titanic to the Moon (2006)
- Ray Harryhausen: Special Effects Titan (2011)
Named Entity Disambiguation (NED): Task Definition

- Disambiguate linkable mentions from a specific context to their referent entities in a Knowledge Base
  - A mention: a phrase referring to something in the world
    - Named entity (person, organization), object, event...
  - An entity: a page in a Knowledge Base

At a WH briefing here in Santiago, NSA spox Rhodes came with a litany of pushback on idea WH didn't consult with.
Entity Semantic Relatedness is Crucial for NED

- Stay up **Hawk Fans**. We are going through a **slump**, but we have to stay positive. Go **Hawks**!

- The most important feature used for NED
  - Non-collective approaches (Ferragina & Scaiella, 2010; Milne and Witten, 2008; Guo et al., 2013)
  - Collective Approaches (Cucerzan, 2007; Milne and Witten, 2008b; Kulkarni et al., 2009; Pennacchiotti and Pantel, 2009; Ferragina and Scaiella, 2010; Cucerzan, 2011; Guo et al., 2011; Han and Sun, 2011; Han et al., 2011; Ratinov et al., 2011; Chen and Ji, 2011; Kozareva et al., 2011; Shen et al., 2013; Liu et al., 2013, Huang et al., 2014)
The State-of-the-art Approaches for Entity Semantic Relatedness

• (Milne and Witten, 2008): Wikipedia Link-based unsupervised method
  
  \[ SR(c_i, c_j) = 1 - \frac{\log \max(|C_i|, |C_j|) - \log |C_i \cap C_j|}{\log(|C|) - \log \min(|C_i|, |C_j|)} \]

  - \( C_i \): the set of incoming links to \( i \)

• Supervised Method (Ceccarelli et al., 2013)
  
  - Formulate as a learning-to-rank problem
  - Explore a set of link-based features

**Limitation I**: Ignore the world knowledge from the rich Knowledge Graphs

**Limitation II**: What if we donot have anchor links?

Boot used "imperialism" to describe United States policy, not only in this embrace of empire is made by others neoconservatives, including Fox and Mark Steyn. It is also made by some liberal hawks, such as polit
Our Approach

- Learn entity representations with supervised DNN and KG
  - Non-linear DNN proven to have more expressive power than the linear models
  - Directly to optimize parameters for semantic relatedness
- The DNN-based Semantic Similarity Model (DSSM) (Huang et al, 2013)

\[
\text{relatedness}(C_1, C_2) = \cosine(y \downarrow C_1, y \downarrow C_2)
\]

Semantic space

Deep non-linear projections

Feature vector

## Encoding Knowledge from Knowledge Graph

<table>
<thead>
<tr>
<th>Knowledge</th>
<th>Representation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Letter tri-gram vector</td>
<td><code>dog = &lt;#do, dog, og#&gt; &lt;0,...,1,1,...,0,1,...,0&gt;</code></td>
</tr>
<tr>
<td>Entity Type</td>
<td>1-of-V vector</td>
<td><code>&lt;0,...,0,...,1,...,0,...&gt;</code></td>
</tr>
<tr>
<td>Subgraph</td>
<td>1-of-V vector for relation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Letter tri-gram for entities</td>
<td></td>
</tr>
</tbody>
</table>
Unsupervised Collective Disambiguation with Graph Regularization

- Perform collective disambiguation for a set of topically-related tweets simultaneously.

  - Handle information shortage and noiseness problems.
  - Easily collect a set of topically-related tweets (e.g., via social networks).

Accuracy = 0.25, tweets are short and noisy, can not provide rich context information.

Underlining concepts are referent concepts.
Graph Construction Over Multiple Tweets

- Each node is a pair of mention and entity candidates
  - Entity candidates are retrieved based on anchor links in Wikipedia
- An edge is created for two nodes if
  - Two mentions are relevant
    - Detect with meta path
  - And two entities are semantically related
    - Cosine similarity over semantic entity embeddings
  - Similarity is used as the edge weight

![Graph Diagram]

- Florida Gators
- Florida Gators men's basketball
- Slump
- Slump (sports)
- Slump (geology)
- Atlanta Hawks
- Kemba Walker
- Milwaukee Bucks
- Milwaukee Bucks
- Kemba Walker

<table>
<thead>
<tr>
<th>Mention 1</th>
<th>Mention 2</th>
<th>Edge Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Florida Gators men's basketball</td>
<td>Kemba Walker, Kemba Walker</td>
<td>0.821</td>
</tr>
<tr>
<td>Florida Gators men's basketball</td>
<td>Milwaukee Bucks</td>
<td>0.578</td>
</tr>
<tr>
<td>Slump, Slump (sports)</td>
<td>Kemba Walker, Kemba Walker</td>
<td>0.625</td>
</tr>
<tr>
<td>Slump, Slump (sports)</td>
<td>Atlanta Hawks</td>
<td>0.524</td>
</tr>
<tr>
<td>Kemba Walker, Kemba Walker</td>
<td>Milwaukee Bucks</td>
<td>0.252</td>
</tr>
<tr>
<td>Kemba Walker, Kemba Walker</td>
<td>Atlanta Hawks</td>
<td>0.325</td>
</tr>
<tr>
<td>Kemba Walker, Kemba Walker</td>
<td>Florida Gators men's basketball</td>
<td>0.245</td>
</tr>
<tr>
<td>Kemba Walker, Kemba Walker</td>
<td>Florida Gators men's basketball</td>
<td>0.02</td>
</tr>
</tbody>
</table>
Relevant Mention Detection: Meta Path

- A meta-path is a path defined over a network and composed of a sequence of relations between different object types (Sun et al., 2011)
  - Each meta path represents a semantic relation
- Meta paths between mention and mention
  - M-T-M
  - M-T-U-T-M-M
  - M-T-H-T-M
  - M-T-U-T-M-T-H-T-M
  - M-T-H-T-M-T-U-T-M

M: mention, T: tweet, U: user, H: hashtag

- Two mentions are considered as relevant if there exist at least one meta path between them
Unsupervised Graph Regularization

- The model (Adapted from Zhu et.al, 2003)
  \[ Q(y) = \mu \sum_{i=l+1}^{n} (y_i - y_i^0)^2 + \frac{1}{2} \sum_{i,j} W_{ij} (y_i - y_j)^2. \]

- Initial ranking score
  - prior popularity and context similarity
  - \( y_i \): the final ranking score of node \( i \)
  - \( y_i^0 \): the initial ranking score of node \( i \)
  - \( W \): weight matrix of the graph
Data and Scoring Metric

- **Data**
  - A public data set includes 502 messages from 28 users (Meiji et al., 2012)
  - A Wikipedia dump on May 3, 2013

- **Scoring Metric**
  - Accuracy on top ranked entity candidates
Models for Comparison

- TagMe: an *unsupervised* model based on prior popularity and semantic relatedness of a single message (Ferragina and Scaiella, 2010)
- Meij: the state-of-the-art *supervised* approach based on the random forest model (Meij et al., 2012)
- GraphRegu: our proposed *unsupervised* graph regularization model
## Overall Performance

- Our methods are unsupervised

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>TagMe (unsupervised)</td>
<td>61.9%</td>
</tr>
<tr>
<td>Meiji (5 fold cross-validation)</td>
<td>68.4%</td>
</tr>
<tr>
<td>GraphRegu + (Milne and Witten, 2008)</td>
<td>64.3%</td>
</tr>
</tbody>
</table>
Overall Performance (con’t)

- Encode Knowledge from contextual descriptions

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>TagMe (unsupervised)</td>
<td>61.9%</td>
</tr>
<tr>
<td>Meiji (5 fold cross-validation)</td>
<td>68.4%</td>
</tr>
<tr>
<td>GraphRegu + (Milne and Witten, 2008)</td>
<td>64.3%</td>
</tr>
<tr>
<td>GraphRegu + DSSM + Description</td>
<td>71.8%</td>
</tr>
</tbody>
</table>

- 26% error rate reduction over TagMe
- 21% error rate reduction over the standard method to compute semantic relatedness (Milne and Witten, 2008)
### Overall Performance

- Encode Knowledge from structured KG

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>TagMe (unsupervised)</td>
<td>61.9%</td>
</tr>
<tr>
<td>Meiji (5 fold cross-validation)</td>
<td>68.4%</td>
</tr>
<tr>
<td>GraphRegu + (Milne and Witten, 2008)</td>
<td>64.3%</td>
</tr>
<tr>
<td>GraphRegu + DSSM + Subgraph (Entity)</td>
<td>68.2%</td>
</tr>
<tr>
<td>GraphRegu + DSSM + Subgraph (Relation + Entity)</td>
<td>70.0%</td>
</tr>
<tr>
<td>GraphRegu + DSSM + Subgraph (Relation + Entity) + Entity Type</td>
<td>70.9%</td>
</tr>
</tbody>
</table>

- **23.6%** error rate reduction over TagMe
- **18.5%** error rate reduction over the standard method to compute semantic relatedness (Milne and Witten, 2008)
## Overall Performance

- Encode all Knowledge from KG

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>TagMe (unsupervised)</td>
<td>61.9%</td>
</tr>
<tr>
<td>Meiji (5 fold cross-validation)</td>
<td>68.4%</td>
</tr>
<tr>
<td>GraphRegu + (Milne and Witten, 2008)</td>
<td>64.3%</td>
</tr>
<tr>
<td>GraphRegu + DSSM + Description</td>
<td>71.8%</td>
</tr>
<tr>
<td>GraphRegu + DSSM + Subgraph (Entity)</td>
<td>68.2%</td>
</tr>
<tr>
<td>GraphRegu + DSSM + Subgraph (Relation + Entity)</td>
<td>70.0%</td>
</tr>
<tr>
<td>GraphRegu + DSSM + Subgraph (Relation + Entity) + Entity Type</td>
<td>70.9%</td>
</tr>
<tr>
<td>GraphRegu + DSSM + Description + Subgraph (Relation + Entity) + Entity Type</td>
<td>71.9%</td>
</tr>
</tbody>
</table>
Conclusions and Future work

• We propose to learn **deep semantic entity embeddings** with **supervised DNN** and **Knowledge Graph**
  - Significantly outperform the standard approach for named entity disambiguation

• Future Work
  - Encode semantic meta-paths from Knowledge Graph into DNN
    • To capture the semantic meaning of knowledge
  - Learn entity embedding with Knowledge Graph for other tasks
    • E.g., Question & Answering
Thank You !!!

Any Questions/Comments?

We will release the embedding for the whole Wikipedia Concepts Soon!!!