Deep Learning for Knowledge Extraction and Integration to Build the Amazon Product Graph

Subhabrata Mukherjee
Knowledge Graph Conference, 5/2019
Product Graph vs. Knowledge Graph
Knowledge Graph Example for 2 Movies

- **Movie**: "Forrest Gump" (mid345)
  - **Starring**: Tom Hanks (mid127)
  - **Starring**: Julia Roberts (mid128)
- **Movie**: "Larry Crowne" (mid346)
  - **Starring**: Robin Wright Penn (mid129)
  - **Directed By**: Julia Roberts (mid128)

- **Person**: Robin Wright (mid127)
  - **Name**: "Robin Wright"
  - **Birth Date**: July 9th, 1956
- **Person**: Tom Hanks (mid128)
  - **Name**: "Tom Hanks"
- **Person**: Julia Roberts (mid129)
  - **Name**: "Julia Roberts"
Product Graph vs. Knowledge Graph

- **mid345**: "Forrest Gump"
- **mid127**: "Robin Wright"
- **mid128**: "Tom Hanks"
- **mid129**: "Julia Roberts"
- **mid567**: "Robin Wright Penn"
- **mid568**: "Forrest Gump"
- **mid569**: "Forrest Gump"
- **mid570**: "Forrest Gump"
- **mid571**: "Forrest Gump"
- **mid346**: "Larry Crowne"
- **mid345**: "Forrest Gump"
- **mid127**: "Robin Wright"
- **mid128**: "Tom Hanks"
- **mid129**: "Julia Roberts"
- **mid567**: "Robin Wright Penn"
- **mid568**: "Forrest Gump"
- **mid569**: "Forrest Gump"
- **mid570**: "Forrest Gump"
- **mid571**: "Forrest Gump"
Knowledge Graph vs. Product Graph

Generic KG

Movie, Music, Book, etc.

Product Graph

(Hardline, softline, consumables, etc.)
Another Example of Product Graph
Knowledge Extraction
### Motivation: Knowledge Extraction / Named Entity Recognition

<table>
<thead>
<tr>
<th>ASIN</th>
<th>Brand</th>
<th>Product Title</th>
<th>Flavor</th>
<th>Package Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>B01HGBAJEC</td>
<td>Nescafe Taster’s Choice</td>
<td>Instant Coffee Beverage, Hazelnut</td>
<td>16 - 0.1 oz packets (Pack of 8)</td>
<td></td>
</tr>
<tr>
<td>B01HGBAESS</td>
<td>Nescafe Taster’s Choice</td>
<td>Instant Coffee, French Roast</td>
<td>0.52 Oz (Pack of 12)</td>
<td></td>
</tr>
<tr>
<td>B01H3KK5H2</td>
<td>NESCAFÉ Rich</td>
<td>French Vanilla, Instant Coffee</td>
<td>100g Jar</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

- E-commerce retailers maintain product catalogs
- A lot of structured data on products buried in text descriptions in Catalog
OpenTag: Highlights

INPUT

- Attribute (e.g., flavor, ingredients, packageSize)
  - Product Profile (e.g., title, description, bullets)

OUTPUT

- Structured Triples (subject, predicate, object)

OpenTag: Open Attribute Value Extraction from Product Profiles, G. Zheng, S. Mukherjee, X.L. Dong, F. Li, KDD 2018

- B01B3ALHES, hasFlavor, {'berry'}
- B013RSC536, hasFlavor, {'bacon'}
- B0132D8WA6, hasFlavor, {'chicken'}
- B079YYFN9, hasFlavor, {'apricot', 'honey'}
- B0073ROTFK, hasIngredients, {'rice powder', 'cocoa mass', 'lecithin', 'ethically sourced palm oil', 'rice', 'cocoa', 'flavour', 'cocoa solids', 'sugar', 'orange', 'emulsifier', 'cocoa butter'}
- B005BXW9O6, hasIngredients, {'milk protein isolate', 'alkali', 'oil', 'artificial', 'maize', 'inulin', 'xanthan gum', 'lecithin', 'carrageenan', 'cellulose gum', 'cocoa', 'whey protein concentrate', 'whey protein', 'whey protein isolate', 'sodium citrate', ...}
- B071XP4XD6, hasPackageSize, {'5 oz'}
- B0716KF5VY, hasPackageSize, {'pack of 12'}
- B075QZBS3Z, hasPackageSize, {'60 packets'}
- B00TBC0R1A, hasPackageSize, {'8 ounce'}
Attribute Extraction as Sequence Tagging

- **B**: Beginning of attribute value
- **I**: Inside of attribute value
- **O**: Outside of attribute value
- **E**: End of attribute value

$x = \{w_1, w_2, ..., w_n\}$ input sequence

$y = \{t_1, t_2, ..., t_n\}$ tagging decision

{beef meal} → Flavor Extractions → {ranch raise lamb}

INPUT

beef

meal

&

ranch

raised

lamb

recipe

OUTPUT

$y$

$X$

$w_1$

$w_2$

$w_3$

$w_4$

$w_5$

$w_6$

$w_7$
OpenTag: Neural Network Architecture

CRF

Attention

Bi-LSTM

Word Embedding

ranch  raised  beef  flavor
Char CNN for noisy text

Distantly Supervised Knowledge Extraction from Product Profiles, S. Mukherjee and X.L. Dong, Submitted
OpenTag Extraction from Product Profiles

- BiLSTM+CRF+Attention obtains best results
- Extraction on new values is comparable to already known values
# Production Impact (*Sample*)

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Previous Coverage of Existing Production System (%)</th>
<th>OpenTag Coverage (%)</th>
<th><em>Increase</em> in Coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute_1</td>
<td>23</td>
<td>78</td>
<td>53</td>
</tr>
<tr>
<td>Attribute_2</td>
<td>21</td>
<td>72</td>
<td>45</td>
</tr>
<tr>
<td>Attribute_3</td>
<td>&lt; 1</td>
<td>56</td>
<td>50</td>
</tr>
<tr>
<td>Attribute_4</td>
<td>&lt; 1</td>
<td>49</td>
<td>48</td>
</tr>
</tbody>
</table>

- Live on Alexa Shopping, being deployed in Amazon Search and Browse
OpenTag with Active Learning

With active learning, we can obtain high precision and recall with 3.3x reduction in human annotation efforts.
OpenTag Extraction from Product Profiles

Interpretability via attention
OpenTag Word Embeddings: Semantically related words come closer in the embedding space.
Knowledge Cleaning
Solution 1: Human-in-the-loop to Clean-up Noisy Training Labels

• Input *negative* keywords
  • E.g., negative *flavor* values {pink, christmas, coffee, ...}

• Remove *semantically similar* values from training labels using KMeans clustering and Glove

OpenTag performance improves 3x with clean training labels
Solution 2: Embedding-based Cleaning of Extractions

- Filtering based on OpenTag embeddings of attribute values

Taking Flavor as an example

Score = \cos(a, b)

Mean Value Embedding

OpenTag Value Embedding

French Roast

Vanilla

Roasted Almond

Hershey’s
Knowledge Integration
**Top Gun**

From Wikipedia, the free encyclopedia

This article is about the film. For other uses, see Top Gun (disambiguation).

*Top Gun* is a 1986 American action drama film directed by Tony Scott, and produced by Don Simpson and Jerry Bruckheimer, in association with Paramount Pictures. The screenplay was written by Jim Cash and Jack Epps Jr., and was inspired by an article titled "Top Guns" published in California magazine three years earlier. The film stars Tom Cruise, Kelly McGillis, Val Kilmer, Anthony Edwards, and Tom Skerritt. Cruise plays Lieutenant Pete "Maverick" Mitchell, a young naval aviator aboard the aircraft carrier USS Enterprise. He and his Radar Intercept Officer, Nick "Goose" Bradshaw (Edwards) are given the chance to train at the US Navy's Fighter Weapons School at Naval Air Station Miramar in San Diego, California.
**Top Gun**

From Wikipedia, the free encyclopedia

This article is about the film. For other uses, see Top Gun (disambiguation).

*Top Gun* is a 1986 American action drama film directed by Tony Scott, and produced by Don Simpson and Jerry Bruckheimer, in association with Paramount Pictures. The screenplay was written by Jim Cash and Jack Epps Jr., and was inspired by an article titled "Top Guns" published in *California* magazine three years earlier. The film stars Tom Cruise, Kelly McGillis, Val Kilmer, Anthony Edwards, and Tom Skerritt. Cruise plays Lieutenant Pete "Maverick" Mitchell, a young naval aviator aboard the aircraft carrier USS Enterprise. He and his Radar Intercept Officer, Nick "Goose" Bradshaw (Edwards) are given the chance to train at the US Navy's Fighter Weapons School at Naval Air Station Miramar in San Diego, California.

---

**Extracted (String) Triples**

- ("Top Gun", "Director", "Tony Scott")
- ("Top Gun", "was written by", "Jim Cash")
- ("Top Gun", "Stars", "Tom Cruise")
- (Top Gun, "Runtime", "1h 50min")
- (Top Gun, "scheduled for release on", "16 May 1986")

---

OpenIE Extractions from Semi-structured Data

OpenIE Extractions from Text
Top Gun

From Wikipedia, the free encyclopedia

This article is about the film. For other uses, see Top Gun (disambiguation).

Top Gun is a 1986 American action-drama film directed by Tony Scott, and produced by Don Simpson and Jerry Bruckheimer, in association with Paramount Pictures. The screenplay was written by Jim Cash and Jack Epps Jr., and was inspired by an article titled "Top Guns" published in California magazine three years earlier. The film stars Tom Cruise, Kelly McGillis, Val Kilmer, Anthony Edwards, and Tom Skerritt. Cruise plays Lieutenant Pete "Maverick" Mitchell, a young naval aviator aboard the aircraft carrier USS Enterprise. He and his Radar Intercept Officer, Nick "Goose" Bradshaw (Edwards) are given the chance to train at the US Navy's Fighter Weapons School at Naval Air Station Miramar in San Diego, California.

Extracted (String) Triples

- ("Top Gun", "Director", "Tony Scott")
- ("Top Gun", "was written by", "Jim Cash")
- ("Top Gun", "Stars", "Tom Cruise")
- (Top Gun, "Runtime", "1h 50min")
- (Top Gun, "scheduled for release on", "16 May 1986")

Integration with KG
Joint Inference over OpenIE and KG

“Cast and Crew”

“worked in the movie”

@film.actor  @film.director

@film.director
Neighborhood Entity Encoding

- Model entity as a composition of its neighboring **relation embeddings** in KB + OpenIE
  - Can discover new entities never seen before!

```
OpenIE

KG

A
“Cast and Crew”

“worked in the movie”

B

E

D

C

@film.actor  @film.director

@film.director

@film.director
```
Joint Relation Inference with Dual Attention

\[ \text{Score}(\text{"life of Pi"}, @\text{film.directed_by,"Ang Lee"}) = \alpha_1f(\text{"Life of Pi"} @\text{film.directed_by,"Ang Lee"}) + \alpha_2f(\text{"Life of Pi"} @\text{film.directed_by,"Ang Lee"}) + \alpha_3f(\text{"Aug Lee"} @\text{film.directed_by,"Ang Lee"}) \]

+ Dual attention mechanism to focus on important neighboring relations

## OpenKI Results for Integrating OpenIE and KB

<table>
<thead>
<tr>
<th></th>
<th>ReVerb + Freebase</th>
<th>ReVerb + Freebase (/film)</th>
<th>Ceres + Freebase (/film)</th>
<th>Ceres + IMDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMI</td>
<td>0.41</td>
<td>0.30</td>
<td>0.51</td>
<td>0.66</td>
</tr>
<tr>
<td>Entity model</td>
<td>0.22</td>
<td>0.16</td>
<td>0.43</td>
<td>0.51</td>
</tr>
<tr>
<td>OpenKI</td>
<td>0.51</td>
<td>0.37</td>
<td>0.66</td>
<td>0.81</td>
</tr>
</tbody>
</table>

- **Neighborhood relation encoding with attention performs the best.**
- **Best performance on semi-structured data + domain-specific KB.**
Unsupervised Clustering of OpenIE Predicates
How to Get to the Next Level of Success?

One vertical,
A few sources
How to Get to the Next Level of Success?

One vertical, A few sources

Thousands-to-millions of sources
How to Get to the Next Level of Success?

One vertical, A few sources

Hierarchy of thousands of types

Thousands-to-millions of sources
How to Get to the Next Level of Success?

One vertical, A few sources

Effective search, mining and analysis

Hierarchy of thousands of types

Thousands-to-millions of sources

Big challenge: Limited training labels for large-scale, rich data
How to Get to the Next Level of Success?

- **Challenges:** Limited training labels for large-scale, rich data
- **Solution:** Learning with limited labels
  - Active learning
  - Weak learning (e.g., distance supervision, data programming)
  - Semi-supervised learning (e.g., graph-based learning)
  - Transfer learning
  - One/few-shot learning
Take Aways

- It is necessary to be open: discover new entities, relations and types
- Neural networks outperform traditional methods with more data and (clean) labels
- A rich KG harnesses power from all kinds of data formats and structures
- Learning with limited labels is the key to scale

Reference: To know more about *Building a Broad Knowledge Graph for Products* refer [http://lunadong.com/talks/BG.pdf](http://lunadong.com/talks/BG.pdf)
Thank You!