OPEN CERES: WHEN OPEN INFORMATION EXTRACTION MEETS THE SEMI-STRUCTURED WEB

Colin Lockard, Prashant Shiralkar, Xin Luna Dong
6 QUESTIONS YOU MIGHT HAVE

• What are semi-structured websites?
• Why do we want to extract from them?
• Why OpenIE?
• Is this problem hard?
• Well then how are you going to solve it?
• Does a dataset even exist to evaluate these extractions anyway?
6 QUESTIONS YOU MIGHT HAVE

• What are semi-structured websites?
  • Exactly what they sound like

• Why do we want to extract from them?
  • They have lots of information

• Why OpenIE?
  • They have LOTS of information

• Is this problem hard?
  • Yes.

• Well then how are you going to solve it?
  • By automatically creating training data

• Does a dataset even exist to evaluate these extractions anyway?
  • It does now.
When Harry Met Sally... (1989)

Harry and Sally have known each other for years, and are very good friends, but they fear sex would ruin the friendship.

Director: Rob Reiner
Writer: Nora Ephron
Stars: Billy Crystal, Meg Ryan, Carrie Fisher

Scary Good: IMDb's Guide to Horror

Can't get enough of movies and TV shows that scare up a good fright? Check out Scary Good, IMDb's Horror Entertainment Guide.
Semi-structured:

- Rich layout features
- Each page provides info about an entity
- Data represented as key-value pairs and lists
- Text fields consisting of just entity/attribute name
There are lots of semi-structured websites.
Semi-structured websites have lots of pages with very similar template structure.
Knowledge Vault @ Google found 4x more facts from semi-structured than unstructured text [Dong et al., KDD’14][Dong et al., VLDB’14]

And they have a lot of information
They might have TOO MUCH information

“Production Design? I don’t have a relation type for Production Design!”

Produced by
Nora Ephron
Steve Martin
Rob Reiner
Andrew Bergman
Jeffrey Stott ...
co-producer

Cinematography by
Barry Sonnenfeld ... director of photography

Film Editing by
Robert Leighton

Casting By
Janet Hirshenson
Jane Jenkins

Production Design by
Jane Musky
When Harry Met Sally... (1989)

Harry and Sally have known each other for years, and are very good friends, but they fear sex would ruin the friendship.

Director: Rob Reiner
Writer: Nora Ephron
Stars: Billy Crystal, Meg Ryan, Carrie Fisher
See full cast & crew

Relation types defined in ontology

(film.written_by, “Nora Ephron”)

(film.actor, “Meg Ryan”)
On 10 semi-structured movie websites, the IMDb ontology covered only 7% of relations.
OPEN INFORMATION EXTRACTION

• Instead of relying on ontology for relation type, just extract relation string from page
• Widely explored in unstructured text
• Not explored on semi-structured websites

PROBLEM DEFINITION

• Input:
  • Pages from a semi-structured website
  • Seed KB (and ontology)

• Output:
  • A set of triples \((s, r, o)\), where
    • \(s\) is a string corresponding to the subject (page topic entity),
    • \(o\) is a string corresponding to the object
    • \(r\) is a string indicating the relation/predicate
“When Harry Met Sally”

(“Writer:”, “Nora Ephron”)

(“Stars:”, “Meg Ryan”)

FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE | SHARE

When Harry Met Sally... (1989)

7.6/10 160/124

Nominated for 1 Oscar. Another 4 wins & 17 nominations. See more awards »

Watch Now

From $2.99 (SD) on Amazon Video
OpenIE triples can be used for:
- KB Completion
  - See our paper “OpenKI: Integrating Open Information Extraction and Knowledge Bases with Relation Inference” by Dongxu Zhang et al!
- Question Answering
- Better KG Embeddings
- Ontology Discovery
- Fact Checking

To make use of all facts on these sites, we need Open Information Extraction
WHAT ARE THE CHALLENGES?
WHY IS THIS HARD?

- Too much variety in expression for rules
- Need to consider **pairs** of text fields
- No labeled data
  - Lots of websites, would need lots of annotations
- Traditional DOM (HTML) features too limited to find new relation types
Semi-structured website W

Knowledge base (KB)

Identify candidate (predicate, object) pairs

Distantly supervised annotation of topic entity and objects of KB predicates

Training Data Creation

Identify predicate strings for KB predicates

Label propagation of (predicate, object) pairs creates training data for KB and open predicates

Supervised learning

DOM node classifier
Step 1: Identify candidate pairs

- Use site-wide statistics
  - Relation string must be more common than object string
  - Relation and object should be “nearby” in visual or DOM distance
- Reduces # pairs from quadratic to quasilinear.
Step 2:
• Distantly supervised annotation of facts in existing knowledge base

(“Tape”, film.genre, “Drama”)

(“Tape”, film.actor, “Uma Thurman”)

(“Tape”, film.director, “Richard Linklater”)

Step 3:

- Use lexicon or embedding similarity to find relation string
- Annotation now defines a pair of text fields
Step 4:
- Identify pairs of text fields similar to our seed pairs

A screenshot shows:

- ("Tape", "Genres", "Drama")
- ("Tape", "Cast", "Uma Thurman")
- ("Tape", "Director", "Richard Linklater")
• Use **visual features** to understand similarity
  • Font
  • Font size
  • Bold/Italic
  • Color
  • Horizontal and vertical location
  • Horizontal/Vertical distance between relation and object

• **Graph-based label propagation**
Candidate pairs become nodes in graph.

Edges weighted by similarity metric based on visual and layout features.
SIMILARITY GRAPH CONSTRUCTION

Similar:
Relations have same font and size and are vertically aligned
SIMILARITY GRAPH CONSTRUCTION

Similar:
Objects have same font and size
SIMILARITY GRAPH CONSTRUCTION

Different:
Horizontal, vertical difference between relation and object
These pairs will have a higher similarity score.
• Multi-Rank Walk algorithm
• Uses Personalized PageRank to score similarity to seed nodes

TRAINING DATA CREATION BY LABEL PROPAGATION

- Multi-Rank Walk algorithm\(^1\)
- Uses Personalized PageRank to score similarity to seed nodes

---

• We now have training data for **new relation types** for **some** pages of a website
• Learn a logistic regression model and apply it to other pages of the site
DATASET AND EXPERIMENTS
EXPANDED SWDE LABEL SET

• Existing SWDE\(^3\) dataset labels 3-4 predicates on a set of 80 websites
  • All English language

• We created OpenIE labels for 21 of these sites

• 5 – 272 relation types per website
  • \(~800,000\) labels

• Released at homes.cs.washington.edu/~lockardc/expanded_swde.html

---

\(^3\) Hao, Qiang, Rui Cai, Yanwei Pang and Lei Zhang. “From one tree to a forest: a unified solution for structured web data extraction.” SIGIR (2011).
## EXPERIMENTS

<table>
<thead>
<tr>
<th>System</th>
<th>Movie</th>
<th></th>
<th>NBA</th>
<th></th>
<th>University</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>P</td>
<td>R</td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>WEIR (Bronzi et al., 2013)</td>
<td>0.23</td>
<td>0.17</td>
<td>0.08</td>
<td>0.17</td>
<td>0.13</td>
<td>0.18</td>
</tr>
<tr>
<td>Colon Baseline</td>
<td>0.63</td>
<td>0.21</td>
<td>0.51</td>
<td>0.33</td>
<td>0.46</td>
<td>0.31</td>
</tr>
<tr>
<td>OpenCeres</td>
<td>0.77</td>
<td>0.68</td>
<td>0.74</td>
<td>0.48</td>
<td>0.65</td>
<td>0.29</td>
</tr>
</tbody>
</table>

Fairly high precision
EXAMPLES OF DISCOVERED RELATION TYPES

Movie
Seed: Director, Writer, Producer, Actor, Release Date
New: Country, Filmed In, Language, MPAA Rating, Set In, Reviewed by, Studio, Metascore, Box Office, Distributor, Tagline, Budget, Sound Mix

NBA Player
Seed: Height, Weight, Team
New: Birth Date, Birth Place, Salary, Age, Experience

University
Seed: Phone Number, Web address, Type (public/private)
New: Calendar System, Enrollment, Highest Degree, Local Area, Student Services, President

4x increase in relation types in Movie/NBA

23x increase in relation types in University!
VARYING RELATION TYPES IN SEED

Able to reach 50% recall when only 10% of predicates are in seed set!
FALSE NEGATIVES

Hard to find this section since format is different from rest of page
Noble Lee Lester

Best known as a Supporting Actor based on a credit in that role in 1 film, with $15,432,314 worldwide aggregate box office (rank #42,574)

Best-Known Acting Roles: Media Jackel (The Bonfire of the Vanities)

Most productive collaborators: Tom Hanks, Brian De Palma, Bruce Willis, Michael Christofer, Melanie Griffith

(FALSE POSITIVES)

("Supporting Actor", "Tom Hanks")

"Supporting Actor" has similar format as true relation types
CONCLUSIONS

• Semi-structured OpenIE can yield millions of facts for QA, KB Completion, and more.
• OpenCeres can discover lots of new relation types.

Use our dataset to build/evaluate even better systems!

homes.cs.washington.edu/~lockardc/expanded_swde.html