CERES: DISTANTLY SUPERVISED RELATION EXTRACTION FROM THE SEMI-STRUCTURED WEB

Colin Lockard, Xin Luna Dong, Arash Einolghozati, Prashant Shiralkar
IMDb has thousands of (similarly formatted) pages about movies.
Nigerian films

Twisted (Short film)

Year of production: 2014
Running Time: 2:12 mins
Written by: Daniel Ademinokan
Produced by: Daniel Ademinokan
Directed by: Daniel Ademinokan
Starring: Stella Damasus Rob Byrnes, Matt Meinsen and David Ademinokan

Not on IMDb!
Made In China

Banner: Maddock Films
Director: Mikhil Musale
Producer: Dinesh Vijan
Star: Rajkummar Rao, Mouni Roy, Boman Irani ... see full cast & crew

Not on IMDb!
"Do the Right Thing", film.director, "Spike Lee"

"Do the Right Thing", film.writer, "Spike Lee"

"Do the Right Thing", film.actor, "Danny Aiello"

"Do the Right Thing", film.actor, "Ossie Davis"

"Do the Right Thing", film.actor, "Ruby Dee"

"Do the Right Thing", film.release, "21 July 1989"

"Do the Right Thing", film.genre, "Comedy"

"Do the Right Thing", film.genre, "Drama"

"Do the Right Thing", film.runtime, "2h"

"Do the Right Thing", film.rating, "R"
An emotive journey of a former school teacher, who writes letters for illiterate people, and a young boy, whose mother has just died, as they search for the father he never knew.

**Central Station**

**Director:** Walter Salles

**Writers:** Marcos Bernstein, João Emanuel Carneiro

**Stars:** Fernanda Montenegro, Vinícius de Oliveira, Marília Pêra

On the hottest day of the year on a street in the Bedford-Stuyvesant section of Brooklyn, everyone's hate and bigotry smolders and builds until it explodes into violence.

**Do the Right Thing**

**Director:** Spike Lee

**Writer:** Spike Lee

**Stars:** Danny Aiello, Ossie Davis, Ruby Dee

On Amazon Prime Video:

Watch Now

From $2.99 (SD)
An emotive journey of a former school teacher, who writes letters for illiterate people, and a young boy, whose mother has just died, as they search for the father he never knew.

Director: Walter Salles
Writers: Marcos Bernstein, João Emanuel Carneiro | 1 more credit »
Stars: Fernanda Montenegro, Vinícius de Oliveira, Marília Pêra | See full cast & crew »
An emotive journey of a former school teacher, who writes letters for illiterate people, and a young boy, whose mother has just died, as they search for the father he never knew.

**Director:** Walter Salles

**Writer:** Marcos Bernstein, João Emanuel Carneiro | 1 more credit »

**Stars:** Fernanda Montenegro, Vinícius de Oliveira, Marília Pêra | See full cast & crew »

**Metascore**
From metacritic.com | 80

**Reviews**
261 user | 73 critic
Traditional approach: Wrapper Induction

Learn rules based on manually annotated pages.

Labor intensive: Need annotations for every site
Knowledge Vault @ Google showed big potential from distantly supervised DOM-tree extraction [Dong et al., KDD’14][Dong et al., VLDB’14]

<table>
<thead>
<tr>
<th></th>
<th>Accu</th>
<th>Accu (conf ≥ .7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TXT (301M)</td>
<td>0.36</td>
<td>0.52</td>
</tr>
<tr>
<td>DOM (1280M)</td>
<td>1.10M</td>
<td>1.1M</td>
</tr>
<tr>
<td>TBL (10M)</td>
<td>0.3M</td>
<td>0.3M</td>
</tr>
<tr>
<td>ANO (145M)</td>
<td>13K</td>
<td>13K</td>
</tr>
</tbody>
</table>

Can’t find new entities

But accuracy is still low
Can we automatically and accurately extract from semi-structured pages?
• Input:
  • Pages from a semi-structured website
  • Seed KB (and ontology)
• Output:
  • Newly extracted triples corresponding to the given ontology
  • Subject and object are strings
  • We do not address entity linkage or knowledge fusion
DISTANT SUPERVISION

- Assume joint mentions of subjects and objects indicate relation
- Use KB to automatically annotate data
- Use this as training data to learn classifier
Problem:
Webpages may mention thousands of entities, creating millions of pairs:
1. Computationally complex.
2. Spurious matches very likely
Automatic annotation is hard: Same person is writer and director
Same person also mentioned in recommendation
CERES OVERVIEW

Contribution 1: CERES Annotator

Contribution 2: CERES Extractor
<table>
<thead>
<tr>
<th></th>
<th>No manual annotations</th>
<th>Finds new entities</th>
<th>Accurate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Supervised Wrapper</strong></td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Induction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Knowledge Vault</strong></td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td><strong>CERES</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Use both local (on a single page) and global (site-wide patterns) information
Local: The topic entity should be associated with a large number of entities on the page.
Global: The topic entity’s name should be in a consistent location on each page.
• Annotate known facts found on page

• If ambiguous:
  • Local: Objects of same predicate should be in same section of page
  • Global: Predicates should be in similar location on all pages. Cluster all potential mentions of a relation across site, choose most common location.
• Probabilistic classifier
• Robust to noise in training data
  • (compared to wrapper induction)
MODEL & FEATURES

- **Multi-class logistic regression model**
  - Input: Featurized DOM node
  - Output: Relation label (or “None”)

- **Features based on Vertex (Gulhane et al, ICDE 2011)**
  - Tag, ID, Class of ancestors/siblings
  - DOM path to template strings

- **Important to limit # of features to prevent overfitting to noise in training data**
Experiments show CERES is competitive with state-of-the-art supervised extractors.
• Baselines:
  • CERES-Baseline: Naïve Distant Supervision
  • CERES-Topic: Uses topic identification, but does not resolve ambiguous objects
• SWDE dataset (Hao et al, SIGIR 2011)
• 10 sites in each of 4 domains (Book, Movie, NBA, University)
SWDE

<table>
<thead>
<tr>
<th>System</th>
<th>Manual Labels</th>
<th>Movie</th>
<th>NBA Player</th>
<th>University</th>
<th>Book</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hao et al. [19]</td>
<td>yes</td>
<td>0.79</td>
<td>0.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>XTPath [7]</td>
<td>yes</td>
<td>0.94</td>
<td>0.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BigGrams [26]</td>
<td>yes</td>
<td>0.74</td>
<td>0.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LODIE-Ideal [15]</td>
<td>no</td>
<td>0.86</td>
<td>0.9</td>
<td>0.91</td>
<td>0.78</td>
</tr>
<tr>
<td>LODIE-LOD [15]</td>
<td>no</td>
<td>0.76</td>
<td>0.87&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.91</td>
<td>0.78</td>
</tr>
<tr>
<td>RR+WADaR [29]</td>
<td>no</td>
<td>0.73</td>
<td>0.80</td>
<td>0.79</td>
<td>0.70</td>
</tr>
<tr>
<td>RR+WADaR 2 [30]</td>
<td>no</td>
<td>0.75</td>
<td>0.91</td>
<td>0.79</td>
<td>0.71</td>
</tr>
<tr>
<td>WEIR [4]</td>
<td>no</td>
<td>0.93</td>
<td>0.89</td>
<td>0.97</td>
<td>0.91</td>
</tr>
<tr>
<td>Vertex++</td>
<td>yes</td>
<td>0.90</td>
<td>0.97</td>
<td>1.00</td>
<td>0.94</td>
</tr>
<tr>
<td>CERES-Baseline</td>
<td>no</td>
<td>NA&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0.78</td>
<td>0.72</td>
<td>0.27</td>
</tr>
<tr>
<td>CERES-Topic</td>
<td>no</td>
<td>0.99&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.97</td>
<td>0.96</td>
<td>0.72</td>
</tr>
<tr>
<td>CERES-Full</td>
<td>no</td>
<td>0.99&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.98</td>
<td>0.94</td>
<td>0.76</td>
</tr>
</tbody>
</table>

State-of-the-art results in two verticals. (better than supervised systems!)
**SWDE**

Competitive with top unsupervised systems in a third

<table>
<thead>
<tr>
<th>Method</th>
<th>Manual Labels</th>
<th>Movie</th>
<th>NBA Player</th>
<th>University</th>
<th>Book</th>
</tr>
</thead>
<tbody>
<tr>
<td>BigGrams [26]</td>
<td>yes</td>
<td>0.79</td>
<td>0.82</td>
<td>0.83</td>
<td>0.86</td>
</tr>
<tr>
<td>LODIE-Ideal [15]</td>
<td>yes</td>
<td>0.74</td>
<td>0.90</td>
<td>0.79</td>
<td>0.78</td>
</tr>
<tr>
<td>LODIE-LOD [15]</td>
<td>no</td>
<td>0.76</td>
<td>0.87 (^a)</td>
<td>0.91 (^a)</td>
<td>0.78</td>
</tr>
<tr>
<td>RR+WADaR [29]</td>
<td>no</td>
<td>0.73</td>
<td>0.80</td>
<td>0.79</td>
<td>0.70</td>
</tr>
<tr>
<td>RR+WADaR 2 [30]</td>
<td>no</td>
<td>0.75</td>
<td>0.91</td>
<td>0.79</td>
<td>0.71</td>
</tr>
<tr>
<td>WEIR [4]</td>
<td>no</td>
<td>0.93</td>
<td>0.89</td>
<td>0.97</td>
<td>0.91</td>
</tr>
</tbody>
</table>

- \(^a\) indicates the best performance in its category.

<table>
<thead>
<tr>
<th>Method</th>
<th>Manual Labels</th>
<th>Movie</th>
<th>NBA Player</th>
<th>University</th>
<th>Book</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vertex++</td>
<td>yes</td>
<td>0.90</td>
<td>0.97</td>
<td>1.00</td>
<td>0.94</td>
</tr>
<tr>
<td>CERES-Baseline</td>
<td>no</td>
<td>NA (^b)</td>
<td>0.78</td>
<td>0.72</td>
<td>0.27</td>
</tr>
<tr>
<td>CERES-Topic</td>
<td>no</td>
<td>0.99 (^a)</td>
<td>0.97</td>
<td>0.96</td>
<td>0.72</td>
</tr>
<tr>
<td>CERES-Full</td>
<td>no</td>
<td>0.99 (^a)</td>
<td>0.98</td>
<td>0.94</td>
<td>0.76</td>
</tr>
</tbody>
</table>
### SWDE

<table>
<thead>
<tr>
<th>System</th>
<th>Manual Labels</th>
<th>Movie</th>
<th>NBA Player</th>
<th>University</th>
<th>Book</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hao et al. [19]</td>
<td></td>
<td>0.83</td>
<td>0.86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>XTPath [7]</td>
<td></td>
<td>0.98</td>
<td>0.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BigGrams [26]</td>
<td></td>
<td>0.79</td>
<td>0.78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LODIE-Ideal [15]</td>
<td></td>
<td>0.96</td>
<td>0.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LODIE-LOD [15]</td>
<td>no</td>
<td>0.76</td>
<td>0.87</td>
<td>0.91 (a)</td>
<td>0.78</td>
</tr>
<tr>
<td>RR+WADaR [29]</td>
<td></td>
<td>0.79</td>
<td>0.70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RR+WADaR 2 [30]</td>
<td></td>
<td>0.79</td>
<td>0.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WEIR [4]</td>
<td></td>
<td>0.97</td>
<td>0.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vertex++</td>
<td></td>
<td>1.00</td>
<td>0.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CERES-Baseline</td>
<td>no</td>
<td>NA(^b)</td>
<td>0.78</td>
<td>0.72</td>
<td>0.27</td>
</tr>
<tr>
<td>CERES-Topic</td>
<td>no</td>
<td>0.99 (a)</td>
<td>0.97</td>
<td>0.96</td>
<td>0.72</td>
</tr>
<tr>
<td>CERES-Full</td>
<td>no</td>
<td>0.99 (a)</td>
<td>0.98</td>
<td>0.94</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Weak on one vertical. Why?

Only a few overlapping entities between websites and seed KB
ABOUT 10 ANNOTATED PAGES ARE NEEDED TO LEARN
COMMONCRAWL MOVIE WEBSITES

• 33 websites, 400,000 pages (not all semi-structured)
• Mostly long-tail (foreign, documentary, animated)
• 7 languages
COMMONCRAWL MOVIE WEBSITES

• 90% precision (compared to ~63% with Knowledge Vault)
• 1.25 million extractions
• Extracted 2.6 new entities for every annotated entity
• Automatic extraction from semi-structured pages is practical
  • If you have existing KB you can:
    • Identify new entities
    • Extract new facts with high-precision
• Probabilistic classifiers are effective at semi-structured extraction
  • Allow for precision/recall tradeoff
**FUTURE WORK**

- **The Dream:** Fully automatic web-scale extraction
- Need good ways to identify websites to target
- Open Information Extraction
- Discover new predicates
- Human-level webpage understanding

I don’t speak Korean, but I can tell this is a key-value pair. Can a computer?