Domain Adaptive Relation Extraction for Big Text Data Analytics

Feiyu Xu
Outline

- Introduction to relation extraction and its applications
- Motivation of domain adaptation in big text data analytics
- Solutions
- Conclusion and future work
What is Relation Extraction

- Given an **unstructured** text, a relation extraction (RE) tool should be able to automatically recognize and extract relations among the relevant entities or concepts that are salient to the user's needs.

**Nobel Peace Prize**

Barack Obama

2009

RE

- Walter Kohn, Nobel, Chemistry, 1998
- J.M. Coetzee, Nobel, Literature, 2003
- Barack Obama, Nobel, Peace, 2009

**Linguistic Patterns:**

- `<prize>` be awarded to `<person>`
- `<person>` win `<prize>` in `<year>`
- ......
Mitten in der Euro-Krise geht **Altkanzler Helmut Kohl** mit **Angela Merkel** äußerst hart ins Gericht

-- Welt online, 25.08.2011
General application task 1:

✩ Information access for information finder
mapping unstructured textual queries of users to more structured formal query for search and answer engines
General application task 2:

- Information acquisition for information provider
- Extract structured information from big amount free texts to construct knowledge bases
Acquisition of Social Network of Pop Stars from Web

Social Network of “Madonna” (Depth = 1)
Social Network of “Madonna” (Depth = 3)
General application task 3: Big Data Analytics

- Enabling the linking between structured and unstructured data
  - Large-scale information monitoring
  - Analytics: analyses of areas, markets, trends
  - Watch: Scanning for relevant new developments
Example: Network of Innovation Keyplayers

AIR LIQUIDE

All Patents in Database:
- METHOD FOR LASER WELDING USING A NOZZLE CAPABLE OF STABILISING THE KEYHOLE (2006-09-03)
- Laser Beam Welding Method with a Metal Vapour Capillary Formation Control (2009-05-28)
  Laser beam welding method with a metal vapour capillary formation control (2008-10-22)
  Process for laser-ARC hybrid welding aluminium metal workpieces (2008-01-17)
  Laser arc hybrid welding method for surface coated metal parts, the surface coating containing aluminium (2008-01-16)
  Laser/MIG hybrid welding process with a high wire speed (2007-10-30)
  Laser or hybrid arc-laser welding with formation of a plasma on the back side (2006-11-29)
  LASER-MIG HYBRID WELDING PROCESS WITH A HIGH WIRE SPEED (2006-05-19)
  Hybrid laser/Metal in Gas welding with a elevated welding and filler wire supply speeds and a high welding current, quality for codewelding
Outline

- Introduction to relation extraction and its applications
- Motivation of domain adaptation in big text data analytics
- Solutions
- Conclusion and future work
Text Analytics for Big Textual Data

- Three main features of big data
  - *Volume*: large-scale in volume
  - *Variety*: with respect to heterogeneous domains and formats
  - *Velocity*: because of its rapid and steady growing.

- Requirements of text analytics technologies for big data
  - *efficient*
  - *robust*
  - *scalable*
  - *domain-adaptive*
Domain Adaptation is Essential for Big Data!

- Among the three big data features, **variety** and **velocity** are even more challenging than the sheer size **volume**
Reasons:

- New domains have been constantly emerging, rapidly growing in size.

- Domains can differ in
  - **topics** (e.g., medicine, chemistry or mechanics)
  - **genres** (e.g., news, novels, blogs, scientific publications or patents)
  - **targets** (e.g., different relations such as marriage, person-parent relation, disease-symptom relation)
  - **data internal properties** (e.g., size or redundancy or connectivity).

- Systems, methods or strategies developed or trained for so-called general purpose or one specific domain can often not be directly taken over by other domains, because
  - each domain needs its own domain knowledge and
  - each application data has its own special properties.
Relevant Strategies for Domain Adaptation

- Minimally dependent on the labeled training data
  - Minimally or weakly supervised machine learning methods

- Strategies for
  - confidence estimation of automatically learned information and knowledge
  - filtering of irrelevant and wrong information

- Domain adaptation of generic systems for specific applications
Outline

- Introduction to relation extraction and its applications
- Motivation of domain adaptation in big text data analytics
- Solutions
- Conclusion and future work
Our solutions (1)

- minimally supervised and distantly supervised automatic learning of domain-specific grammar-based pattern rules for n-ary RE: DARE and Web-DARE Systems


  - Hans Uszkoreit, Feiyu Xu, Hong Li. “Analysis and Improvement of Minimally Supervised Machine Learning for Relation Extraction”. In NLDB 2009.

Our solutions (2)

- Various filtering and confidence estimation methods for high-performance and large-scale relation extraction
  - Sebastian Krause, Hong Li, Hans Uszkoreit, Feiyu Xu, “Large-Scale Learning of Relation-Extraction Rules with Distant Supervision from the Web”. In Proceedings of the 11th International Semantic Web Conference (ISWC 2012)
Our solutions (3)

- **Automatic adaptation and improvement of generic parsing results for specific domains**
Automatic generation of domain-specific linguistic knowledge resources


- Open source: sargraph.dfki.de
Web-DARE
Distant-supervised Web-scale RE
Web-DARE: Distant Supervision based RE

- Large number of RE rules are automatically learned by using Freebase as seed knowledge and Web as training corpus

- Goal:
  - covering most linguistic variants for expressing a relation
  - thus solving the notorious long-tail problem of real-world NL applications
Data Set

- rules learned for 39 relations
  - n-ary relations \( n \geq 2 \)
- three domains: business, awards and people
- 2.8 million relation instances retrieved from Freebase as seed
- 20 million web documents as training corpus
Example in Nobel Prize Award Domain

- Seed example

<Mohamed ElBaradei/Person, Nobel/Prize, Peace/Area, 2005/Year>

- Sentence matched with the seed

Mohamed ElBaradei won the 2005 Nobel Prize for Peace on Friday...
Dependency Parse Result

Subject: Person

Object: Prize

- lex-mod: Year
- lex-mod: PrizeName
- mod: for
  pcomp-n: Area

win
Bottom Up Rule Learning

Rule (1)

Subject: Person

Object: Prize

win

lex-mod: Year  lex-mod: PrizeName

mod: for

pcomp-n: Area
Bottom Up Rule Learning

Subject: Person

win

Object: Prize

lex-mod: Year, lex-mod: PrizeName

mod: for

pcomp-n: Area

Rule (1)

Rule (2)
Bottom Up Rule Learning

Rule (1)

Rule (2)

Rule (3)
Some Statistics of Web-DARE Rules

<table>
<thead>
<tr>
<th>Relation</th>
<th># Sentences used</th>
<th># Sentences w/ a learned rule</th>
<th># Rules</th>
<th># Rules w/o duplicates</th>
</tr>
</thead>
<tbody>
<tr>
<td>award nomination</td>
<td>13,966</td>
<td>13,149</td>
<td>23,987</td>
<td>7,800</td>
</tr>
<tr>
<td>award honor</td>
<td>50,550</td>
<td>49,001</td>
<td>106,550</td>
<td>40,578</td>
</tr>
<tr>
<td>hall of fame induction</td>
<td>31,244</td>
<td>28,278</td>
<td>44,920</td>
<td>17,450</td>
</tr>
<tr>
<td>organization relationship</td>
<td>46,331</td>
<td>42,824</td>
<td>60,379</td>
<td>28,903</td>
</tr>
<tr>
<td>acquisition</td>
<td>63,967</td>
<td>60,903</td>
<td>96,747</td>
<td>50,544</td>
</tr>
<tr>
<td>organization merger</td>
<td>2,996</td>
<td>1,521</td>
<td>3,243</td>
<td>1,758</td>
</tr>
<tr>
<td>company name change</td>
<td>9,433</td>
<td>9,132</td>
<td>15,619</td>
<td>6,910</td>
</tr>
<tr>
<td>spin off</td>
<td>5,247</td>
<td>5,094</td>
<td>8,319</td>
<td>4,798</td>
</tr>
<tr>
<td>marriage</td>
<td>342,895</td>
<td>335,313</td>
<td>557,478</td>
<td>176,949</td>
</tr>
<tr>
<td>sibling relationship</td>
<td>167,611</td>
<td>160,893</td>
<td>255,788</td>
<td>69,596</td>
</tr>
<tr>
<td>romantic relationship</td>
<td>155,335</td>
<td>152,878</td>
<td>229,393</td>
<td>74,895</td>
</tr>
<tr>
<td>person parent</td>
<td>192,610</td>
<td>186,834</td>
<td>390,878</td>
<td>119,238</td>
</tr>
<tr>
<td>average of 39 relations</td>
<td>66,545</td>
<td>66,509</td>
<td>109,435</td>
<td>41,620</td>
</tr>
</tbody>
</table>
Problems of Large-Scale Approach

- Very low precision
  - a lot of noisy rules
  - many rules are learned from more than one relation
Euler Diagram for Four People-Relations

- Rules of marriage: 160,853
- Rules of person parent: 107,381
- Rules of romantic relationship: 64,515
- Rules of sibling relationship: 61,176
Various Filtering Strategies for High-Performance Web-Scale RE
Frequency-Driven Rule Filters

- Merged Filter:

\[ \text{valid}^R_m(r) = \text{valid}^R_{freq}(r) \land \text{valid}^R_{inter}(r) \]

1) **absolute frequency filtering**: a threshold to exclude rules with low occurrence
Rule Frequency Driven Filters

- Merged Filter:

\[
valid^R_m(r) = valid^R_{\text{freq}}(r) \land valid^R_{\text{inter}}(r)
\]

1) **absolute frequency filtering**: a threshold to exclude rules with low occurrence

2) **inter-relation filter** (Overlap Filter – FO Filter):
   - based on mutual exclusiveness of relations with similar entity-type signatures.
   - a rule is only valid for a relation, if its relative frequency is higher than any other relations with similar entity type signatures.

\[
valid^R_{\text{inter}}(r) =
\begin{cases} 
  \text{true} & \text{if } \forall R' \in \mathbb{R}\setminus\{R\} : rf_{r,R} > rf_{r,R'} \\
  \text{false} & \text{otherwise}
\end{cases}
\]
Weakness of Filtering with Frequency

- Undetected low-quality patterns:
  - high frequency in target relation, low frequency in coupled relations
Weakness of Filtering with Rule Frequency

- Undetected low-quality patterns:
  - high frequency in target relation, low frequency in coupled relations
Weakness of Filtering with Rule Frequency

- Undetected low-quality patterns:
  - high frequency in target relation, low frequency in coupled relations

- Erroneously-deleted good patterns:
  - infrequent patterns

```
PERSON  Verb: “meet”  PERSON

PERSON  Noun: “widower”  PERSON
```
Lexical Semantics can help!

Relation-specific lexical semantic graphs

World Wide Web

Candidate RE Patterns

Unsupervised Classification

High-quality RE Patterns
Automatic learning of relation-specific lexical semantic network

Generic Lexical Semantic Network (BabelNet)

automatically learned unfiltered RE rules and their mentions
Automatic learning of relation-specific lexical semantic network

Generic Lexical Semantic Network (BabelNet)

automatically learned unfiltered RE rules and their mentions

Word Sense Disambiguation
The Relation-Specific Semantic Graph

An excerpt of the semantic graph for the relation *marriage*
Extrinsic Eval. – Web-DARE

- Relative Recall
- Baseline
- S-Filter (WordNet)
- FO-Filter
- S-Filter (BabelNet)
Parse-Reranking for Domain-adaptive RE
## Error Types of Extracted Wrong Instances

<table>
<thead>
<tr>
<th>Content Modality</th>
<th>Named Entity Recognition (NER)</th>
<th>Parsing</th>
<th>NER &amp; Parsing</th>
<th>DARE Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.8% 17.6%</td>
<td>5.9%</td>
<td>38.2%</td>
<td>11.8%</td>
<td>14.7%</td>
</tr>
</tbody>
</table>
Egyptian scientist Ahmed Zewail won the 1999 Nobel Prize for chemistry.
Egyptian scientist Ahmed Zewail won the 1999 Nobel Prize for chemistry.
Reranking Architecture

N-best parses → DARE → RE rules with confidence values

Parse Reranking

Reranked parses
Baseline: before Re-ranking

- **Best reading**: high precision, low recall, low F-measure
- **500 readings**: lower precision, higher recall, higher F-measure
After Re-Ranking:

- Re-ranked top readings match more sentence mentions containing RE instances
- Improvements of Recall and F-Measure
Conclusion

- The performance of large-scale RE for each application is dependent on the performance of domain-adaptation methods

- Three original contributions (among others):
  - Extension of relation extraction to n-ary relations
  - Semantic filtering with large lexical knowledge bases
  - Parser improvement for the specific RE task by reranking

- For our work we received a Google Focused Research Award
Planned Future Work

- Immediate next step of big text data analytics is to integrate the existing NLP and IE components into big data analytics platforms.

- Entity linking and RE will play an essential role for semantic interoperability between structured and unstructured data.

- Extension and Application of our IE technologies to the new Smart Data projects:
  - Smart Data Web: Industry 4.0
  - Smart Data for Mobility: Mobility