Taxonomy Construction Using Syntactic Contextual Evidence

Luu Anh Tuan\(^1\), Jung-jae Kim\(^1\), Ng See Kiong\(^2\)

\(^1\)School of Computer Engineering, Nanyang Technological University, Singapore
\(^2\)Institute for Infocomm Research, A*STAR, Singapore
Outline

• Introduction
• Related work
• Methodology
• Experiments
• Conclusion and future work
Taxonomy

• Useful for many areas:
  • question answering
  • document clustering

• Some available hand-crafted taxonomies: WordNet, OpenCyc, Freebase
  • time-consuming
  • more general, less specific

→ demand for constructing taxonomies for new domains
Outline

• Introduction
• Related work
• Methodology
• Experiments
• Conclusion and future work
Taxonomic relation identification

• Statistical approach:
  • Co-occurrence analysis (Budanitsky, 1999), term subsumption (Fotzo, 2004), clustering (Wong, 2007).
  • Less accurate, heavily depend on feature types and dataset

• Linguistic approach:
  • Hand-written patterns: (Kozareva, 2010), (Wentao, 2012)
  • Automatic bootstrapping: (Girju, 2003), (Velardi, 2012)
  • Lack of contextual analysis across sentences → low coverage
Our contribution

• Propose syntactic contextual subsumption method:
  • Utilize contextual information of terms in syntactic structures by evidence from the Web
  • Infer taxonomic relations between terms in different sentences

• Introduce graph-based algorithm for taxonomy induction:
  • Utilize the evidence scores of edges
  • Base on graph’s topological properties
Outline

- Introduction
- Related work
- Methodology
- Experiments
- Conclusion and future work
Workflow

- Term extraction and filtering
- Taxonomic relation identification
- Taxonomy induction
Term extraction and filtering

- Term extraction:
  - Apply Stanford parser \(\rightarrow\) extract all noun phrases
  - Remove determiners, do lemmatization

- Term filtering:
  - TF-IDF
  - Domain relevance, domain consensus (Navigli and Velardi, 2004)

\[
TS(t, D) = \alpha \times \text{TFIDF}(t, D) + \beta \times \text{DR}(t, D) + \gamma \times \text{DC}(t, D)
\]
Taxonomic relation identification

- Combine three methods:
  - Syntactic contextual subsumption
  - String inclusion with WordNet
  - Lexical-syntactic pattern matching
Syntactic contextual subsumption (SCS)

• Find relations across different sentences

• Utilize syntactic structure (Subject, Verb, Object)

• Observation 1: (terrorist, attack, people), (terrorist, attack, American)

  → people ≫ American

• But from (animal, eat, meat) and (animal, eat, grass)?
Observation 2:

\[ \rightarrow s_1 \gg s_2 \]

- \( S(\text{animal, eat}) = \{\text{meat, wild boar, deer, buffalo, grass, potato, insects}\} \)
- \( S(\text{tiger, eat}) = \{\text{meat, wild boar, deer, buffalo}\} \)

\[ \rightarrow \text{animal} \gg \text{tiger} \]
Syntactic contextual subsumption (SCS)

- For terms $s_1$, $s_2$:
  - Find most common relation $v$ between $s_1$ and $s_2$. Suppose $s_1$ and $s_2$ are both subjects
  - Submit query “$s_1 \; v$” to search engine, collect first 1000 results, find
    \[ S(s_1, v) = \{ o \mid \exists (s_1, v, o) \} \]
  - Similar for $S(s_2, v)$
  - Calculate:

\[ \text{Score}_{SCS}(s_1, s_2) = \left[ \frac{|S(s_1, v) \cap S(s_2, v)|}{|S(s_2, v)|} + \left( 1 - \frac{|S(s_1, v) \cap S(s_2, v)|}{|S(s_1, v)|} \right) \right] \times \log(|S(s_1, v)| + |S(s_2, v)|) \]
String inclusion with WordNet (SIWN)

SIWN method:

\[ t_1 = w_{11} w_{12} w_{13} \]
\[ t_2 = w_{21} w_{22} w_{23} w_{24} w_{25} \]

\( \gg \) or \( \approx \)

\( t_1 \gg t_2 \)
\( \gg: \) is hypernym of

“suicide attack” \( \gg \) “self-destruction bombing”

- attack \( \gg \) bombing
- suicide \( \approx \) self-destruction

\[ Score_{SIWN}(t_1, t_2) = \begin{cases} 
1 & \text{if } t_1 \gg t_2 \text{ via SIWN} \\
0 & \text{otherwise}
\end{cases} \]
Lexical-syntactic pattern (LSP)

- Use following patterns to query on Google:
  
  "t₁ such as t₂"
  "t₁, including t₂"
  "t₂ is [a|an] t₁"
  "t₂ is a [kind|type] of t₁"
  "t₂, [and|or] other t₁"

\[
\text{Score}_{LSP}(t₁, t₂) = \frac{\log(WH(t₁, t₂))}{1 + \log(WH(t₂, t₁))}
\]
Combined method

\[ \text{Score}(t_1, t_2) = \alpha \times \text{Score}_{\text{SIWN}}(t_1, t_2) \]
\[ + \beta \times \text{Score}_{\text{LSP}}(t_1, t_2) \]
\[ + \gamma \times \text{Score}_{\text{SCS}}(t_1, t_2) \]
Taxonomy induction

- Step 1: Initial hypernym graph with a ROOT node
- Step 2:
  \[
  w(e(t_1, t_2)) = \begin{cases} 
  1 & \text{if } t_1 = \text{ROOT} \\
  \text{Score}(t_1, t_2) & \text{otherwise}
  \end{cases}
  \]
- Step 3: apply Edmonds’ algorithm to find maximum optimum branching of weighted directed graph
Taxonomy induction
Outline

- Introduction
- Related work
- Methodology
- Experiments
- Conclusion and future work
Constructing new taxonomies

- **Terrorism domain:**
  - 104 reports of the US state department “Patterns of Global Terrorism (1991-2002)"
  - Each report ~1,500 words

- **Artificial Intelligence (AI) domain:**
  - 4,119 papers extracted
    - the IJCAI proceedings from 1969 to 2011
    - the ACL archives from 1979 to 2010
**Taxonomy construction**

- Compare constructed AI taxonomy with that of (Velardi et al., 2012)

<table>
<thead>
<tr>
<th></th>
<th>Our system</th>
<th>Velardi’s system</th>
</tr>
</thead>
<tbody>
<tr>
<td>#vertex</td>
<td>1839</td>
<td>1675</td>
</tr>
<tr>
<td>#edge</td>
<td>1838</td>
<td>1674</td>
</tr>
<tr>
<td>Average depth</td>
<td>6.2</td>
<td>6</td>
</tr>
<tr>
<td>Max depth</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Term coverage</td>
<td>83%</td>
<td>76%</td>
</tr>
</tbody>
</table>
### Taxonomy construction

- Number of taxonomic relations extracted by different methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Terrorism domain</th>
<th>AI domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCS</td>
<td>484</td>
<td>1308</td>
</tr>
<tr>
<td>SIWN</td>
<td>301</td>
<td>984</td>
</tr>
<tr>
<td>LSP</td>
<td>527</td>
<td>1537</td>
</tr>
<tr>
<td>SIWN + LSP</td>
<td>711</td>
<td>2203</td>
</tr>
<tr>
<td>SCS + SIWN + LSP</td>
<td>976</td>
<td>3122</td>
</tr>
</tbody>
</table>
Taxonomy construction

- Estimated precision of taxonomic relation identification methods in 100 random extracted relations

<table>
<thead>
<tr>
<th>Percentage of correct relations</th>
<th>Terrorism domain</th>
<th>AI domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCS</td>
<td>91%</td>
<td>88%</td>
</tr>
<tr>
<td>SIWN</td>
<td>96%</td>
<td>91%</td>
</tr>
<tr>
<td>LSP</td>
<td>93%</td>
<td>93%</td>
</tr>
<tr>
<td>SCS + SIWN + LSP</td>
<td>92%</td>
<td>90%</td>
</tr>
</tbody>
</table>
Evaluate against WordNet

- Three domains: Animals, Plants and Vehicles:
  - Use the bootstrapping algorithm described in (Kozareva, 2008)
- Compare the results with (Kozareva, 2010) and (Navigli, 2011)

<table>
<thead>
<tr>
<th></th>
<th>Animals domain</th>
<th></th>
<th>Plants domain</th>
<th></th>
<th>Vehicles domain</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Our</td>
<td>Kozareva</td>
<td>Navigli</td>
<td>Our</td>
<td>Kozareva</td>
<td>Navigli</td>
</tr>
<tr>
<td>Term coverage</td>
<td>96%</td>
<td>N.A.</td>
<td>94%</td>
<td>98%</td>
<td>N.A.</td>
<td>97%</td>
</tr>
<tr>
<td>Precision</td>
<td>95%</td>
<td>98%</td>
<td>97%</td>
<td>95%</td>
<td>97%</td>
<td>97%</td>
</tr>
<tr>
<td>Recall</td>
<td>56%</td>
<td>38%</td>
<td>44%</td>
<td>53%</td>
<td>39%</td>
<td>38%</td>
</tr>
<tr>
<td>F-measure</td>
<td>71%</td>
<td>55%</td>
<td>61%</td>
<td>68%</td>
<td>56%</td>
<td>55%</td>
</tr>
</tbody>
</table>
Syntactic structures

- Comparison of three syntactic structures: **S-V-O (Subject-Verb-Object)**, **N-P-N (Noun-Preposition-Noun)** and **N-A-N (Noun-Adjective-Noun)**

<table>
<thead>
<tr>
<th></th>
<th>S-V-O</th>
<th>N-P-N</th>
<th>N-A-N</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Animals domain</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>95%</td>
<td>68%</td>
<td>72%</td>
</tr>
<tr>
<td>Recall</td>
<td>56%</td>
<td>52%</td>
<td>47%</td>
</tr>
<tr>
<td>F-measure</td>
<td>71%</td>
<td>59%</td>
<td>57%</td>
</tr>
<tr>
<td><strong>Plants domain</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>95%</td>
<td>63%</td>
<td>66%</td>
</tr>
<tr>
<td>Recall</td>
<td>53%</td>
<td>41%</td>
<td>43%</td>
</tr>
<tr>
<td>F-measure</td>
<td>68%</td>
<td>50%</td>
<td>52%</td>
</tr>
<tr>
<td><strong>Vehicles domain</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>93%</td>
<td>59%</td>
<td>60%</td>
</tr>
<tr>
<td>Recall</td>
<td>69%</td>
<td>45%</td>
<td>48%</td>
</tr>
<tr>
<td>F-measure</td>
<td>79%</td>
<td>51%</td>
<td>53%</td>
</tr>
</tbody>
</table>
Dataset link

Outline

- Introduction
- Related work
- Architecture
- Experiments
- Conclusion and future work
Conclusion

• Proposed a novel method of identifying taxonomic relations using contextual evidence from syntactic structure and Web data

• Presented a graph-based algorithm to induce an optimal taxonomy from a given taxonomic relation set

• Generally achieve better performance than the state-of-the-art methods
Future work

• Build the probabilistic model for taxonomy
• Consider the time stamp of information
• Apply to other domains and integrate into other frameworks such as ontology learning or topic identification
THANK YOU

Q & A
References


