Type inference in Dbpedia from free text

Ondřej Zamazal
Tomáš Kliegr
Václav Zeman

University of Economics, Prague
Motivation

- New DBpedia resource types extraction
- Type/hypernym discovering from resources without infobox or mapping rules
  
  | dbpedia:Dublin_City_Cup | rdf:type | dbpedia-owl:Event | #new type |
  
- Find more specific DBpedia types
  
  | dbpedia:Angela_Merkel | rdf:type | dbpedia-owl:Person | #mapping-based type |
  | dbpedia:Angela_Merkel | rdf:type | dbpedia-owl:Chancellor | #more specific discovered type |

- Extract types/hypernyms from plain text
- Provide datasets with newly discovered DBpedia types and evaluation
- Provide a hypernym discovery tool for local DBpedia chapters to acquire new types
LHD 1.0: identifies the hypernym word from the first sentence of the short abstract and maps it to a DBpedia resource
  ◦ Language dependent process
  ◦ POS Tagger and user-defined grammar are required
LHD 2.0 – STI classifier: inferring DBpedia ontology types based on co-occurrence of LHD 1.0 and DBpedia types
hSVM classifier: inferring types based on machine learning over bag of words
  ◦ Language independent approach
Evaluation on a large crowdsourced dataset
Extract a hypernym word from the first sentence
- TreeTagger + Gate
  Havel was a Czech playwright, essayist, poet, dissident and politician.

Map the hypernym word to a DBpedia resource
- dbp:Playwright

Get their DBpedia types along with frequency
- Comedian:1, ... OfficeHolder:7, ..., Writer:266, Artist:277, Person:521

Remove the supertypes, balancing specificity and reliability
- Comedian:1, ... OfficeHolder:7, ..., Writer:266, Artist:277, Person:521

Select the type with highest support
- dbpedia-owl:Writer

Language dependent approach !!!
Each resource is represented with two bag-of-words models
- BOW1 based on short abstract
- BOW2 based on article categories

Classification models are trained by SVM with mapping-based resources

3 layers of SVM classifiers = 58 SVM classifiers
Each SVM model returns an assignment confidence
Supertype confidence is propagated to its subtypes as well
We define 3 strategies to determine the final type
- \( \alpha \) strategy assigns with maximum confidence from all specific types (leaf types)
- \( \beta \) strategy assigns type with maximum confidence from all types
- \( \gamma \) strategy combines \( \alpha \) and \( \beta \) strategies

Language independent approach !!!
Evaluation on a crowdsourced dataset
Each entity was typically annotated by three to four annotators
1021 entities with a DBpedia type in the gold standard
Evaluation of
1. STI – language dependent approach
2. hSVM – language independent approach
3. STI + hSVM fusion
4. Mapping–based types dataset
5. SDType – Mapping–based types (Heuristic) dataset
   - heuristic link–based type inference mechanism
# Results

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Acc\text{Exact}</th>
<th>Acc\text{Supertypes}</th>
<th>Acc\text{[Sub\mid Super]types}</th>
</tr>
</thead>
<tbody>
<tr>
<td>STI</td>
<td>.375</td>
<td>.456</td>
<td>.643</td>
</tr>
<tr>
<td>hSVM</td>
<td>.350</td>
<td>.725</td>
<td>.768</td>
</tr>
<tr>
<td>STI+hSVM</td>
<td>.468</td>
<td>.831</td>
<td>.856</td>
</tr>
<tr>
<td>SDType</td>
<td>.337</td>
<td>.706</td>
<td>.775</td>
</tr>
<tr>
<td>DBpedia (mapping-based)</td>
<td>.374</td>
<td>.856</td>
<td>.920</td>
</tr>
</tbody>
</table>
Conclusion

- **STI/hSVM vs SDType**
  - Type intersection is not so large
  - STI/hSVM is more specific
  - Use both STI/hSVM and SDType to discover new DBpedia types

- **STI** – language dependent (now only for English, German and Dutch)

- **hSVM** – language independent (available for all DBpedia chapters)

- The newest tool and datasets will be available very soon; publication is under review

- Watch our website
Thank you for your attention

http://ner.vse.cz/datasets/linkedhyponyms/

vaclav.zeman@vse.cz