TEXT ANNOTATION USING BACKGROUND KNOWLEDGE

Delia Sorina Rusu

PhD Thesis Defence
December 15, 2014

Jožef Stefan International Postgraduate School
Artificial Intelligence Laboratory, Jožef Stefan Institute
Overview

• Introduction

• Contributions
  • Contribution 1: Concept relatedness approaches
  • Contribution 2: Generic text annotation framework
  • Contribution 3: Applications

• Summary
Introduction
Text Annotation

or Text (Entity) Linking
or Word Sense Disambiguation

• Identifying the meaning of words in context
• Identifying the corresponding concept defined in background knowledge datasets
Linked Data as Background Knowledge

Linked Data: principles for publishing and interlinking structured data on the Web

- Text annotation solution which is not tailored to a specific type of structured data (i.e. specific ontology or knowledge base)
- We can annotate text with concepts which are interlinked
- Machine-readable representations of text at different levels of granularity
Slovenia has beautiful **mountains** and **lakes**
Slovenia has beautiful **mountains** and **lakes**

**WordNet**
- WN: mountain (land mass)
- WN: mountain (large amount)
- CYC: mountain (topographical feature)
- DB: mountain (landform)
- DB: mountain (band)
- DB: ...

**DBpedia**
- WN: lake (body of water)
- WN: lake (red pigment)
- WN: lake (any bright translucent pigment)
- CYC: lake (body of water)
- DB: lake (body of water)
- DB: lake (pigment)
- DB: ...
Propose, apply and evaluate a generic text annotation framework based on background knowledge datasets, relying on:

- datasets part of Linked Data, having common characteristics
- different types of information provided by background knowledge datasets
Scientific Contributions

1. Concept relatedness approaches that leverage concept definitions, the dataset structure and a hybrid approach

2. A modular and generic automatic text annotation framework based on concept relatedness

3. Applications: datasets part of Linked Data - WordNet, OpenCyc and DBpedia

Main scientific articles related to the thesis:

- Applied Ontology Journal, SCI IF 0.844 [1, 3]
- Language Resources and Evaluation, SCI IF 0.518 (under review) [1, 2, 3]
- + other 7 contributions (journal, conference, workshop)
Contribution 1: Concept Relatedness Approaches
Concept Relatedness Approaches

• Definition based: Extended Definition Vectors
• Structure based: Weighted Concept Paths
• Hybrid approach
Extended Definition Vectors

• Based on the Vector Space Model

• A generalisation of the Definition Vectors proposed in (Patwardhan, 2003), addressing the following limitations:
  
  1. the concept definitions are augmented with connected concepts which are directly linked
  
  2. all connected concept definitions are treated as being equally important
  
  3. cannot differentiate between connected concept definitions based on the type of relation
Slovenia has beautiful **mountains** and **lakes**.

WordNet

**WN: mountain** - a land mass that projects well above its surroundings; higher than a hill

**WN: lake** - a body of (usually fresh) water surrounded by land
Slovenia has beautiful mountains and lakes.

**WordNet**

- **WN: mountain** - a land mass that projects well above its surroundings; higher than a hill
- **WN: elevation** - a raised or elevated geological formation
- **WN: volcano** - a mountain formed by volcanic material
- **WN: lake** - a body of (usually fresh) water surrounded by land
- **WN: pond** - a small lake
- **WN: body of water** - the part of the earth’s surface covered with water
Slovenia has beautiful **mountains** and **lakes**.

**WordNet**

- **WN: mountain** - a land mass that projects well above its surroundings; higher than a hill
- **WN: elevation** - a raised or elevated geological formation
- **WN: geological formation** - (geology) the geological features of the earth
- **WN: volcano** - a mountain formed by volcanic material

- **WN: lake** - a body of (usually fresh) water surrounded by land
- **WN: body of water** - the part of the earth’s surface covered with water
- **WN: pond** - a small lake
- **WN: thing** - a separate and self-contained entity
- **WN: fishpond** - a freshwater pond with fish
Extended Definition Vectors

Solution: take into account the **weighted** contribution of each concept definition

- include an arbitrary number of connected concept definitions
- not all concepts are equally important
- assign different weights to different types of relations
Weighted Concept Paths

• **Concept weights**: facilitate distinguishing between abstract and specific concepts

• **Relation weights**: defined as a function of its two adjacent concepts

• **Concept relatedness**: determined based on the weighted concept path
Slovenia has beautiful mountains and lakes.

WordNet:
- **WN: mountain** - a land mass that projects well above its surroundings; higher than a hill
- **WN: volcano** - a mountain formed by volcanic material
- **WN: geological formation** - (geology) the geological features of the earth
- **WN: elevation** - a raised or elevated geological formation
- **WN: pond** - a small lake
- **WN: body of water** - the part of the earth's surface covered with water …
- **WN: lake** - a body of (usually fresh) water surrounded by land
- **WN: thing** - a separate and self-contained entity
- **WN: elevation** - a raised or elevated geological formation
- **WN: geological formation** - (geology) the geological features of the earth
- **WN: mountain** - a land mass that projects well above its surroundings; higher than a hill
- **WN: volcano** - a mountain formed by volcanic material
- **WN: pond** - a small lake
- **WN: body of water** - the part of the earth's surface covered with water …
- **WN: lake** - a body of (usually fresh) water surrounded by land
- **WN: thing** - a separate and self-contained entity

more abstract | more specific
Hybrid Approach

• Weights the contribution of the definition-based relatedness and the structure-based relatedness
Contribution 2: Text Annotation Framework
Text Annotation Framework

Plain Text

Text Annotation Module
- Text Pre-processing
- Candidate Concept Identification
- Candidate Concept Ranking
- Concept Annotation

Relatedness Module
- Definition-based Relatedness
- Structure-based Relatedness
- Hybrid Approach
Text Annotation Algorithm
Text Annotation Algorithm

[Diagram showing the steps of the text annotation algorithm with symbols and variables]
Text Annotation Algorithm
Text Annotation Algorithm
Contribution 3: Applications: WordNet, OpenCyc, DBpedia
WordNet Evaluation

Concept Relatedness

<table>
<thead>
<tr>
<th>Measures used in the evaluation</th>
<th>MC - WordNet</th>
<th>RG - WordNet</th>
<th>WordSim - WordNet</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition</strong></td>
<td>0.87</td>
<td>0.82</td>
<td>0.69</td>
</tr>
<tr>
<td><strong>Structure</strong></td>
<td>0.84</td>
<td>0.86</td>
<td>0.67</td>
</tr>
<tr>
<td><strong>Hybrid</strong></td>
<td><strong>0.88</strong></td>
<td><strong>0.86</strong></td>
<td><strong>0.72</strong></td>
</tr>
<tr>
<td>Yang Powers</td>
<td>0.76</td>
<td>0.78</td>
<td>0.63</td>
</tr>
<tr>
<td>Banerjee Pedersen</td>
<td>0.76</td>
<td>0.69</td>
<td>0.46</td>
</tr>
<tr>
<td>Partwardhan Pedersen</td>
<td>0.88</td>
<td>0.81</td>
<td>0.55</td>
</tr>
</tbody>
</table>
WordNet Evaluation
Text Annotation

Annotation results for all parts of speech (F-measure)

Window size

Avg - WeightedConceptPath  Med - WeightedConceptPath  Max - WeightedConceptPath
Avg - WordNet Definition  Med - WordNet Definition  Max - WordNet Definition
Avg - Hybrid  Med - Hybrid  Max - WordNet Definition
# OpenCyc Evaluation

## Concept Relatedness

<table>
<thead>
<tr>
<th>Measures used in the evaluation</th>
<th>MC - WordNet</th>
<th>RG - WordNet</th>
<th>WordSim - WordNet</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition</strong></td>
<td>0.48</td>
<td>0.34</td>
<td>0.20</td>
</tr>
<tr>
<td><strong>Structure</strong></td>
<td>0.70</td>
<td>0.71</td>
<td>0.42</td>
</tr>
<tr>
<td>Moore et al.</td>
<td>0.65</td>
<td>0.56</td>
<td>0.36</td>
</tr>
<tr>
<td>Leacock Chodorow</td>
<td>0.59</td>
<td>0.30</td>
<td>0.24</td>
</tr>
<tr>
<td>Wu Palmer</td>
<td>0.55</td>
<td>0.39</td>
<td>0.29</td>
</tr>
</tbody>
</table>
### DBpedia Evaluation

#### Concept Relatedness

<table>
<thead>
<tr>
<th>Measures used in the evaluation</th>
<th>MC - WordNet</th>
<th>RG - WordNet</th>
<th>WordSim - WordNet</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition</strong></td>
<td>0.88</td>
<td>0.81</td>
<td>0.56</td>
</tr>
<tr>
<td><strong>Structure</strong></td>
<td>0.84</td>
<td>0.82</td>
<td>0.59</td>
</tr>
<tr>
<td><strong>Hybrid</strong></td>
<td><strong>0.92</strong></td>
<td><strong>0.88</strong></td>
<td><strong>0.65</strong></td>
</tr>
<tr>
<td>Moore et al.</td>
<td>0.84</td>
<td>0.82</td>
<td>0.46</td>
</tr>
<tr>
<td>Shortest Path Unit W.</td>
<td>0.82</td>
<td>0.79</td>
<td>0.42</td>
</tr>
<tr>
<td>Adapted Google Dist.</td>
<td>0.59</td>
<td>0.49</td>
<td>0.53</td>
</tr>
</tbody>
</table>
DBpedia Evaluation
Text Annotation

Annotation results for all words (F-measure)

Window size

Avg - WeightedConceptPath - Categ
Max - WeightedConceptPath - Categ
Med - WeightedConceptPath - Categ
Avg - DBpedia Definition
Max - DBpedia Definition
Med - DBpedia Definition
Avg - Hybrid
Max - Hybrid
Med - Hybrid
Evaluation Conclusions

Concept Relatedness

- Consistent results across different knowledge bases: concept relatedness and text annotation

- Best results (which improve over SOTA approaches) using a **hybrid approach** for both WordNet and DBpedia

- in the absence of concept definitions the background knowledge dataset structure provides useful information for the relatedness and annotation tasks
Evaluation Conclusions
Text Annotation

• No additional domain-specific information, yet results comparable with SOTA systems

• The proposed approaches can be extended to other background knowledge datasets with similar properties
Summary

Three major contributions

1. Concept relatedness approaches that leverage concept definitions, the dataset structure and a hybrid approach
   [Rusu et al. Measuring concept similarity in ontologies using weighted concept paths. *Applied Ontology*]

2. A modular and generic automatic text annotation framework based on concept relatedness

3. Applications: datasets part of Linked Data - WordNet, OpenCyc and DBpedia
   [Rusu et al. Measuring concept similarity in ontologies using weighted concept paths. *Applied Ontology*]
   [Štajner et al. A service oriented framework for natural language text enrichment. *Informatica Slovenia*]

+ 6 conference and workshop papers
Future Work

• Evaluate the text annotation framework using smaller, domain-specific ontologies or knowledge bases

• Use a combination of datasets published as Linked Data as input

• Test the annotation framework on multilingual knowledge bases (e.g. BabelNet, DBpedia, WikiData)

• Test on real-world applications (e.g. event detection)

[Rusu et al. Unsupervised techniques for extracting and clustering complex events in news. The Second Workshop on EVENTS. ACL. 2014]
Thank You!
Extended Definition Vector

$c$ concept in the background knowledge dataset

t_i term vector corresponding to the definition of $c$

$C(c)$ centroid for the concept $c$

\[
C(c) = \frac{\sum_{i=1}^{n} \alpha_i \frac{t_i}{\|t_i\|_2}}{\sum_{i=1}^{n} \alpha_i}
\]

$ED(c)$ extended definition vector for $c$

\[
ED(c) = \frac{C(c)}{\|C(c)\|_2}
\]
Weighted Concept Path

$CW(c)$  concept weight for the concept  $c$

$CW(c) = \log Degree(c)$

$RW(c_i, c_j)$  relation weight between the concepts  $c_i$  and  $c_j$

$RW(c_i, c_j) = \max(CW(c_i), CW(c_j))$

$DS(c_i, c_j)$  distance between concepts  $c_i$  and  $c_j$

$NDS(c_i, c_j)$  normalized distance between concepts  $c_i$  and  $c_j$

$DS(c_i, c_j) = \text{ShortestWeightedPath}(c_i, c_j)$

$R(c_i, c_j)$  relatedness between concepts  $c_i$  and  $c_j$

$R(c_i, c_j) = 1 - NDS(c_i, c_j)$
OpenCyc Evaluation
Concept Relatedness

<table>
<thead>
<tr>
<th>Measures used in the evaluation</th>
<th>Modified DBI</th>
<th>INTRA cluster distance</th>
<th>INTER cluster distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure</td>
<td>1.36</td>
<td>0.34</td>
<td>0.56</td>
</tr>
<tr>
<td>Definition</td>
<td>1.62</td>
<td>0.58</td>
<td>0.81</td>
</tr>
<tr>
<td>Moore et al.</td>
<td>1.41</td>
<td>0.36</td>
<td>0.59</td>
</tr>
<tr>
<td>Leacock Chodorow</td>
<td>1.66</td>
<td>0.23</td>
<td>0.33</td>
</tr>
<tr>
<td>Wu Palmer</td>
<td>1.61</td>
<td>0.12</td>
<td>0.16</td>
</tr>
</tbody>
</table>