Exploring Semantic Similarity and Relatedness Measures for Semantic Search

Nitish Aggarwal
Supervised by Dr. Paul Buitelaar
DERI, NUI Galway
firstname.lastname@deri.org

Tuesday, 23 Apr, 2013
UNLP PhD day
Overview

- Problem Space
- Cross-Lingual Search
- Semantic Similarity/Relatedness
- Knowledge Linking/Identification
- Work in Progress
The newly ordained Bishop of Limerick took inspiration from U2 singer Bono when speaking about the need to connect with those living on the outskirts of society. More than 1,500 people packed St John’s Cathedral in Limerick on Sunday for the ordination of Fr Brendan Leahy as the 47th Bishop of Limerick, and the first to be ordained since 1974.
Wer ist die Tochter von Bill Clinton verheiratet?@de
Cross-Lingual NL-Querying

Wer ist die Tochter von Bill Clinton verheiratet?@de

Linguistic Analysis Entity Identification CL-Similarity and Relatedness

Marc Mezvinsky

WIKIPEDIA

English

Deutsch

Français

Svenska

Polski

Español

Italiano

Nederlands

Português

Enabling Networked Knowledge
Cross-Lingual NL-Querying over Web of Data

Mit wem is die Tochter von Bill Clinton verheiratet?

Linguistic Analyzer

Von(Tochter, Bill Clinton)
subj(Verheiraten, Tochter)

Ordered list of appeared entities in NL-Query

Entity Search

- index look up
- resolve wikiRedirects

dbpedia: Bill_Clinton

dbo:birthDate

Tochter

verheiraten

dbo:child
dbo:relative
dbo:spouse
Cross-Lingual NL-Querying over Web of Data

Linguistic Analyzer

Von(Tochter, Bill Clinton)
subj(Verheiraten, Tochter)

Entity Search

- index look up
- resolve wikiRedirects

Ordered list of appeared entities in NL-Query

dbpedia: Bill_Clinton

CL-Semantic Similarity and Relatedness

dbpedia: Bill_Clinton

dbo:child

dbpedia: Chelsea_Clinton

dbo:relative

dbo:spouse

dbo:birthDate

dbo:Person
“Cross-Lingual Semantic Search by Improving Semantic Similarity and Relatedness Measures” at ISWC-DC-2012
DERI NL-queries

- 50 German NL-queries (translated and selected from 100 NL-queries QALD-2)
- Performed three types of vocabulary matching

<table>
<thead>
<tr>
<th></th>
<th>Average Precision</th>
<th>Average Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation</td>
<td>0.217</td>
<td>0.24</td>
<td>0.228</td>
</tr>
<tr>
<td>Translation with ESA</td>
<td>0.34</td>
<td>0.386</td>
<td>0.361</td>
</tr>
<tr>
<td>CL-ESA</td>
<td>0.459</td>
<td>0.506</td>
<td>0.481</td>
</tr>
</tbody>
</table>

"Cross-Lingual Natural Language Querying over Web of Data" at NLDB-2013
Entity Search

Natural Language Query (NL-Query)

Who is the daughter of Bill Clinton married to?@en

Short paragraphs/News article

The newly ordained Bishop of [Limerick] took inspiration from [U2] singer Bono when speaking about the need to connect with those living on the outskirts of society. More than 1,500 people packed St John's Cathedral in Limerick on Sunday for the ordination of Fr Brendan Leahy as the 47th Bishop of Limerick, and...
Semantic Relatedness

- **Structured-based**
  - WordNet Similarity
    - Information found in the lexical database
      - Limited to predefined lexical structure
    - Combine this with corpus based relatedness

- **Corpus-based**
  - High dimensional distributional vector
    - Latent Semantic Space
      - Unsupervised latent/hidden topics
    - Explicit Semantic Space
      - Explicitly defined concepts/topics (e.g. Wikipedia article)
Explicit Semantic Analysis

<table>
<thead>
<tr>
<th></th>
<th>Concept_1</th>
<th>Concept_2</th>
<th>Concept_3</th>
<th>Concept_4</th>
<th>Concept_5</th>
</tr>
</thead>
<tbody>
<tr>
<td>term_1</td>
<td>W_11</td>
<td>W_12</td>
<td>W_13</td>
<td>W_14</td>
<td>W_15</td>
</tr>
<tr>
<td>term_2</td>
<td>W_21</td>
<td>W_22</td>
<td>W_23</td>
<td>W_24</td>
<td>W_25</td>
</tr>
</tbody>
</table>

- Explicit concepts can be defined by Wikipedia articles
  - E.g. term *tiger* is associated with
    - *Balinese Tiger*: 0.5916
    - *Kyukyoku Tiger*: 0.5421
    - *Celtic Tiger Live*: 0.5130
    - *Paper tiger*: 0.4337

ESA as GVSM

\[ \kappa(x, y) = xK y^T \]
Issues with ESA

- Non-uniform vocabulary distribution
  - Significant difference between the number of concepts associated with one term, against the another.
    - Cup: 250 concepts, Tableware: 4 concepts
- Considers orthogonality between concepts
  - Different concepts cover similar subject
- Ambiguity in building high dimensional vector for word
  - Cup: Cup (unit), World_Cup etc.
- **Pseudo Query Expansion**
  - Finding more concepts by extending the query with definitions of the concepts obtained by searching only the term

- **Considering concepts similarities**
  - \[ \kappa_{ESA}(x, y) = x_C K_C y_C^T = x(WK_CW^T)y^T \]
  - In our experiments, \( K_c \) is the standard cosine kernel, semantic Gaussian kernel, and semantic LSA kernel

- **Context Enrichment for word relatedness**
  - Provide context (Synonyms/definition) to build the vector
Results: Correlated topics & Pseudo Query Expansion

<table>
<thead>
<tr>
<th>WordSim353</th>
<th>Spearman</th>
<th>Pearson</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESA</td>
<td>0.6481</td>
<td>0.5233</td>
</tr>
<tr>
<td>ESA+LSA</td>
<td>0.6648</td>
<td>0.5854</td>
</tr>
<tr>
<td>ESA PQE+LSA</td>
<td>0.6636</td>
<td>0.5719</td>
</tr>
<tr>
<td>Combi*</td>
<td>0.6735</td>
<td>0.5621</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sentence</th>
<th>MSRpar</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spearman</td>
</tr>
<tr>
<td>Text</td>
<td>0.4200</td>
</tr>
<tr>
<td>ESA</td>
<td>0.2982</td>
</tr>
<tr>
<td>ESA+LSA</td>
<td>0.4219</td>
</tr>
<tr>
<td>ESA PQE+LSA</td>
<td>0.4144</td>
</tr>
<tr>
<td>Text + ESA+LSA</td>
<td>0.4450</td>
</tr>
<tr>
<td></td>
<td>0.7981</td>
</tr>
</tbody>
</table>

"Improving ESA with Documents Similarity" at ECIR-2013

*Combination of the two best kernels above, $K = \alpha K_1 + (1 - \alpha) K_2$, the tuned value of $\alpha$ is applied to sentence data to combine the plain textual similarity and ESA kernels (Text + ESA+LSA).
Results: Context Enrichment

Submitting to CIKM-2013
“The Effects of Different Context Enrichments on Semantic Relatedness using ESA”
Work Status

- **Work in progress**
  - Context Enrichment in ESA (CIKM-2013)
  - Keyword Semantic Search Visualization Demo (ISWC-2013)
  - Demo on ESA with its variants

- **Next Plan**
  - Entity Relatedness for Disambiguation
  - Explicit Topic Tagging/Identification
  - Semantic Relatedness using DBpedia and ESA
Thanks