Time-Aware and Corpus-Specific Entity Relatedness

Nilamadhaba Mohapatra\textsuperscript{1,2}, Vasileios Iosifidis\textsuperscript{1}, Asif Ekbal\textsuperscript{2}, Stefan Dietze\textsuperscript{1}, Pavlos Fafalios

\textsuperscript{1}L3S Research Center, University of Hannover, Germany
\textsuperscript{2}Indian Institute of Technology, Patna, India
Introduction

• Entity Relatedness
  • Determining the degree of relatedness between two entities

• Useful in a variety of applications
  • IR
  • Search Recommendations
  • Entity Linking

• Approaches:
  • Structural similarity in graphs
  • Using lexical characteristics
  • Wikipedia-based entity distributions and embeddings
Introduction

• Temporality of entity context
  • Entity popularity often changes across time [Fang et al. 2014]
  • Entity recommendations are time-dependent [Zhang et al. 2016] [Tran et al. 2017]
  • Exploiting different KB versions can advance entity relatedness [Prangnawarat and Hayes, 2017]

• What about the corpus-context?
Introduction

• What about the corpus-context?

• Example:
  • Input entity: 2014 FIFA World Cup
  • Related entities:

<table>
<thead>
<tr>
<th>German news articles</th>
<th>Greek news articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germany national football team</td>
<td>Greece national football team</td>
</tr>
<tr>
<td>Argentina national football team</td>
<td>Costa Rica national football team</td>
</tr>
<tr>
<td>Mario Götze</td>
<td>Sokratis Papastathopoulos</td>
</tr>
</tbody>
</table>
Approach overview

• Entity relatedness depends on both:
  • Time-context
  • Corpus-context

• Method:
  • Train time- and corpus-specific word embeddings (using Word2Vec)
    • Not using general-purpose corpora like Wikipedia!
  • Exploit entity annotations for transforming word embeddings to entity embeddings
  • Relax the time boundaries (optionally)
Problem Modeling

• Corpus of documents D covering a time period T
  • E.g., German sport articles of 2015
• Entities E mentioned in the documents (persons, locations, events, ...)
  • Extracted using an entity linking system
  • Each entity is associated with a unique URI in a KB
Problem Modeling

• Modeling: ranking problem

1. Generate a list of candidate entities
   • Exploiting Wikipedia links, DBpedia, and entity co-occurrences in the corpus [Zhang et al. 2016]

2. Rank the candidate entities based on their relevance to the query entities
Time-Aware Word Vector Similarity

\[ T = [t_1, t_5] \]

Word2Vec Model 1

Word2Vec Model 2

Word2Vec Model 3

Word2Vec Model 4

Word2Vec Model 5

Dec'17

MODEL
Time-Aware Word Vector Similarity

$T = [t_1, t_5]$
Making the embeddings entity-aware

• Limitations of word embeddings:
  • Handling of multi-word entity names, e.g. “United Nations”
  • Same entity name may refer to different entities, e.g. “Kobe”

• Solution: exploit entity annotations
  • Each entity mention in a document is associated with a unique URI in a KB

• Replace entity mentions with unique IDs
  • As done in [Mikolov et al. 2013] for phrases

• Train Word2Vec models on the modified corpora

• Use entity IDs for computing entity relatedness

MAPPINGS:
3844878080843 ➔ https://en.wikipedia.org/wiki/Donald_Trump
8798729506352 ➔ https://en.wikipedia.org/wiki/Democratic_Party_(United_States)
1365985456623 ➔ https://en.wikipedia.org/wiki/Hillary_Clinton

3844878080843 was elected president in a surprise victory over 8798729506352 nominee 1365985456623.
Making the embeddings entity-aware

\[ T = [t1, t5] \]

- Word2Vec Model 1
- Word2Vec Model 2
- Word2Vec Model 3
- Word2Vec Model 4
- Word2Vec Model 5

Time-Aware and Corpus-Specific Entity Relatedness, DL4KGS@ESWC'18, June 2018
Relaxing the time boundaries

- Problem
  - Entity embeddings built on specific time periods (e.g., November 2017)
    - An important event related to the query entity may happened very close to the boundaries of the query time period
    - The query entity may correspond to an event spanning a longer time period
  - Two entities might be highly related some time before or after the query time period!

- Consider the Word2Vec models before & after the model of the query time period
  - But with smaller weight
  - This can increase the ranking of an important candidate entity that co-occurs frequently with a query entity some time before or after the query time period
    - But can also decrease the ranking of an entity co-occurring with the query entity during the query time period
Relaxing the time boundaries

\[ T = [t1, t5] \]

- **Word2Vec Model 1**: 10%
- **Word2Vec Model 2**: 80%
- **Word2Vec Model 3**: 10%
- **Word2Vec Model 4**:
- **Word2Vec Model 5**:

Dec'17

3844878080843

5465462136453

7634723647924

1387093820878

Time-Aware and Corpus-Specific Entity Relatedness, DL4KG5@ESWC’18, June 2018
Evaluation

• Objective:
  • Evaluate the effectiveness of the proposed approach
  • Compare it with similar but time and entity agnostic models

• Dataset and Ground Truth:
  • Provided in [Zhang et al. 2016]
  • Different task: Time-aware entity recommendation (for keyword queries)
  • Candidate entities and relevance judgements for 22 keyword queries (July 2014 – January 2015)
    • Each query corresponds to a particular date range (month)
  • Adapted for our problem (time-aware entity relatedness)
    • Keyword query $\rightarrow$ query entities
    • Remove query entities from list of candidate entities
  • Example: “Tour de France Nibali” (07/2014)
    • Query entities: {Wikipedia:Tour_de_France, Wikipedia:Vincenzo_Nibali}
Evaluation

• Setup
  • 7 CBOW models (one per month)
    • Using default Word2Vec setting (300 dimensions, 5 words window size, 5 min word count)
  • Compare our approach on ranking the candidate entities with two baselines:
    • 1) entity-agnostic, time agnostic
    • 2) entity-agnostic, time-aware
Evaluation

• Results

<table>
<thead>
<tr>
<th>nDCG@k</th>
<th>Time+Entity Agnostic</th>
<th>Entity Agnostic</th>
<th>Time+Entity Aware</th>
</tr>
</thead>
<tbody>
<tr>
<td>k=5</td>
<td>0.3210</td>
<td>0.3653</td>
<td>0.4999 †</td>
</tr>
<tr>
<td>k=10</td>
<td>0.3748</td>
<td>0.4113 †</td>
<td>0.5402 †</td>
</tr>
<tr>
<td>k=20</td>
<td>0.4546</td>
<td>0.4971 †</td>
<td>0.6115 †</td>
</tr>
<tr>
<td>k=30</td>
<td>0.5092</td>
<td>0.5704 †</td>
<td>0.6562 †</td>
</tr>
</tbody>
</table>
Evaluation

• Relaxing the time boundaries

<table>
<thead>
<tr>
<th>nDCG@k</th>
<th>$w_1 = 1.0$</th>
<th>$w_1 = 0.9$</th>
<th>$w_1 = 0.8$</th>
<th>$w_1 = 0.7$</th>
<th>$w_1 = 0.6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>k=5</td>
<td>0.4999</td>
<td><strong>0.5017</strong></td>
<td>0.4990</td>
<td>0.4933</td>
<td>0.4890</td>
</tr>
<tr>
<td>k=10</td>
<td><strong>0.5402</strong></td>
<td>0.5332</td>
<td>0.5358</td>
<td>0.5296</td>
<td>0.5291</td>
</tr>
<tr>
<td>k=20</td>
<td><strong>0.6115</strong></td>
<td>0.6039</td>
<td>0.5971</td>
<td>0.5932</td>
<td>0.5893</td>
</tr>
<tr>
<td>k=30</td>
<td><strong>0.6562</strong></td>
<td>0.6517</td>
<td>0.6451</td>
<td>0.6403</td>
<td>0.6371</td>
</tr>
</tbody>
</table>

• Examples of positive impact
  • “2014 FIFA World Cup” (July 2014): nDCG@5 increases from 0.45 to 0.51 ($w_1=0.8$)
  • “Tim Cook” (October 2014): nDCG@5 increases from 0.58 to 0.62 ($w_1=0.8$)
Conclusion

• Flexible model for entity relatedness
  • Considers the underlying corpus
  • Time- and Entity-aware
  • Outperforms similar but time and entity agnostic models

• Future work
  • Support of arbitrary time intervals (join results of several models)
  • Identify cases where time boundaries relaxation should be applied
  • Extensive evaluation using a variety of corpora (of different contexts and time periods)
Thank you

Comments / Questions?

Indian Institute of Technology Patna

L3S Research Center, Leibniz Universität Hannover

ALEXANDRIA Project (ERC Nr. 339233)