Entity Type Prediction in Knowledge Graphs using Embeddings

Russa Biswas, Radina Sofronova, Mehwish Alam, and Harald Sack

02.06.2020
Motivation

What are the types of the following entities?

- Violin: Instrument
- Lisbon: City
- Yellow billed duck: Bird
Motivation

What are the types of the following entities?

Violin

Lisbon

Yellow billed duck

Instrument

City

Bird
## Motivation

Coarse grained Type information in DBpedia at a glance

<table>
<thead>
<tr>
<th>Classes</th>
<th>#Total entities</th>
<th>Percentage of entities with more fine grained type</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbo:Person</td>
<td>1,818,072</td>
<td>36.6%</td>
</tr>
<tr>
<td>dbo:Scientist</td>
<td>25,760</td>
<td>3.5%</td>
</tr>
<tr>
<td>dbo:Settlement</td>
<td>581,293</td>
<td>68.3%</td>
</tr>
<tr>
<td>dbo:Company</td>
<td>109,629</td>
<td>13.9%</td>
</tr>
</tbody>
</table>
Related Work

Heuristic Approach
- SDType ISWC 2013

Classification
- SLCN WIMS 2016

Tensor Decomposition
- TD WWW 2018

Embeddings
- APE ACL 2018

Proposed Model

RDF triples

Incoming & Outgoing links

RDF Triples and entity descriptions

RDF Triples and entity descriptions

Unsupervised Approach

Supervised Approach

RDF Triples
Proposed Approach

- Three different word embedding models are used to model the KGs
  - Word2Vec
  - FastText
  - GloVe

- Entity Typing is done based on these 3 word embedding models separately and are compared against each other.
Embeddings


**Sentences**

**words**

Embedding models are trained on triples with **Object Properties**

- **Word2Vec:** Continuous Bag of Words (CBOW) approach is used.
- **FastText:** Continuous Bag of Words (CBOW) approach is used.
- **GloVe:** Word co-occurrence matrix is used to learn the model.
Pipeline of the Unsupervised Approach

Unsupervised approach is based on the vector similarity between the class vector and entity vector.

1. Calculate Class Vector
2. Cosine Similarity (C, test entity)
3. Cosine Similarity ($C_{\text{Hierarchy}}$, test entity)
4. Rank the classes based on Cosine similarity
5. Return Top1 and Top3 classes based on the rank
Generation of Class Vectors

Entities from class
dbo: Scientist

Generated Vectors

dbr:Albert_Einstein

dbr:Marie_Curie

dbr:Pierre_Curie

Class Vector of
dbo:Class C

\[
\text{Sum of entity vectors of } C = \frac{\text{Total no. of entities in } C}{...}
\]
Supervised Approach - 1D CNN

Input

Convolutional Layers

MaxPooling

Dense Layers

Output

Length = 100

Entity Vector

Feature Maps

DBpedia Classes
Datasets

1. **Dataset 1:**
   a. 59 less popular classes with the following characteristics:
      i. **15 classes** that have less than 500 entities per class,
      ii. **20 classes** that have entities between 500 and 1000 entities per class,
      iii. **24 classes** have more than 1000 entities per class,
      iv. **Max. no. of entities per class in this dataset is 500**, and
      v. **Min. no. of entities per class in this dataset is 276.**

2. **Dataset 2:** 86 classes with 2k entities per class.

3. **Dataset 3:** 81 classes with 4k entities per class.
## Results

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Models (Results in Percentage Accuracy)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Word2Vec</td>
</tr>
<tr>
<td></td>
<td>Vector Similarity</td>
</tr>
<tr>
<td></td>
<td>Hits@3</td>
</tr>
<tr>
<td>Dataset 1</td>
<td>47.83</td>
</tr>
<tr>
<td>Dataset 2</td>
<td>58</td>
</tr>
<tr>
<td>Dataset 3</td>
<td>58</td>
</tr>
</tbody>
</table>
**Results - Comparison with SDType**

**Test Dataset:** The common entities between our dataset and the entities for which SDType model[1] predicts a change are considered.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>#Test Entities</th>
<th>SDType</th>
<th>Vector Similarity (Accuracy in Percentage)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Word2Vec</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Hits@1</td>
</tr>
<tr>
<td>Dataset 1</td>
<td>7425</td>
<td>83.35</td>
<td>32</td>
</tr>
<tr>
<td>Dataset 2</td>
<td>57467</td>
<td>80.43</td>
<td>46.94</td>
</tr>
<tr>
<td>Dataset 3</td>
<td>109948</td>
<td>81.22</td>
<td>48.21</td>
</tr>
</tbody>
</table>

http://wiki.dbpedia.org/services-resources/documentation/datasets#InstanceTypesSdtypeddbo

Russa Biswas et al. Entity Type Prediction in KGs using Embeddings. DL4KG Workshop @ ESWC 2020.
Conclusion and Future Work

- **Word embeddings** when applied on the Knowledge Graphs can be efficiently used for the task of Entity Type Prediction.

- **Word2Vec proves** to be the best word embedding approach out of the three word embedding approaches used in KGs.

- Supervised approach, **1D CNN works better than the unsupervised approach** for the task.

- In Future Work, more information from the DBpedia such as Datatype properties are to be explored for the type prediction task.
References

Literature:


Images:

https://en.wikipedia.org/wiki/Lisbon#/media/File:Montagem_de_Lisboa.png
https://en.wikipedia.org/wiki/Violin#/media/File:Violin_VL100.png
Thank you

Homepage: https://www.fiz-karlsruhe.de/en/forschung/lebenslauf-und-publikationen-russa-biswas
Email id: russa.biswas@fiz-karlsruhe.de
Twitter: @russa.biswas