Patent Citation Prediction with Seq2Seq

John Knowles, Sam Premutico
{jkn0wles, samprem} @ stanford.edu

Problem

Offer citation recommendations for patents in the US Patent Office
Explore embeddings over longer text documents
Designed as a multi-part system, utilizing NLP encodings, clustering techniques and supervised learning

Task

Represent document in vectorized form
Document pruning to last 1000 words
Vectorized into 256 components
Recommend citations using vector forms
Framed as an clustering task

Data

train/test/dev : 75,000/5,000/5,000 patents

Approach

Data / Task

Task

Represent document in vectorized form
Documents pruned to last 1000 words
Vectorized into 256 components
Recommend citations using vector forms
Framed as an clustering task

Data

train/test/dev : 75,000/5,000/5,000 patents

Table 1: MRR Results Across K (cluster size)

<table>
<thead>
<tr>
<th>Model</th>
<th>K=50</th>
<th>K=100</th>
<th>K=300</th>
<th>K=500</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF Baseline</td>
<td>0.003</td>
<td>0.011</td>
<td>0.008</td>
<td>0.014</td>
</tr>
<tr>
<td>GloVe Baseline</td>
<td>0.014</td>
<td>0.016</td>
<td>0.016</td>
<td>0.018</td>
</tr>
<tr>
<td>Cluster Distance</td>
<td>0.172</td>
<td>0.172</td>
<td>0.174</td>
<td>0.175</td>
</tr>
<tr>
<td>Cluster + Predict</td>
<td>0.164</td>
<td>0.169</td>
<td>0.185</td>
<td>0.185</td>
</tr>
</tbody>
</table>

MRR (mean reciprocal rank): gives a score to an ordered list of possible responses to a query

Analysis

Autoencoder

- Performed poorly on the text reconstruction task (repetitions)
- Due to large inputs, we were not optimizing over text reconstruction, but rather quality of embeddings

Example Autoencoder Output:

“... application application liquid thermal ...”

System

- Lower MRR patent recommendations contained large number of low degree seed patents. MRR was better overall across high-degree patents
- Prediction network did not boost results much beyond a simple clustering and results with no clustering performed poorly
- Able to significantly improve on baselines, which suggests that feature extraction from autoencoder provided meaningful representations

Results

Conclusion

Large search spaces are difficult
Autoencoders provide a general way to extract meaningful feature representations
NLP modules can be used in contexts of larger systems

Future Work

Explore deeper neural architectures and different input representations for the autoencoder
Train over larger subset of the patent network