Technology Scouting as a Service (TSaaS)

Tim Schopf, 24.06.2021, sebis day

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
www.matthes.in.tum.de
Enable engineers to find matching solutions for their technical challenges at the push of a button.
The AI algorithm **continuously searches** the internet for new technologies.

The AI algorithm **understands the problem** of an engineer.

Via the ROKIN platform, we inform the engineer at **the push of a button** about **all technologies relevant to him**.
Try TechMonitor for 1 month for free

Two steps to your TechMonitor:

1. DESCRIBE YOUR TOPIC
   Select your field in the form.
   Describe the topic you would like to receive information on with at least 3 keywords.

2. GET YOUR TECHMONITOR
   You will receive matching articles once a week to the email address you provided.
   The free trial month expires automatically after 30 days.
   You will then receive an email in which you can continue to subscribe to the TechMonitor if you would like to.

https://en.rokin.tech/techmonitor
Idea:
- Word Embeddings are vector representation of words that are supposed to encode their semantic meaning
- Pre-trained Word Embeddings like Google’s Word2Vec are trained on huge corpora (~ 100 billion words)
- However, they are trained on domain-unspecific corpora (often news articles and Wikipedia)
- In domains with a lot of technical jargon (like engineering), they might not be able to represent meaning

Research question:
- Can a domain-specific Word Embedding model (even if trained on a smaller data set) outperform a larger non-specific model in domain-specific tasks?
The Language of Engineering
Training a Domain-Specific Word Embedding Model for Engineering

Data Set

- More than 100 engineering trade publications in English in the domain of mechanical and electrical engineering
- Focus mainly on topics such as robotics, automation, 3D printing or augmented reality, but also on more economical aspects such as investments, mergers and acquisitions or personnel changes in companies
- Roughly 600,000 articles published between 1969 and 2020

Model Training

- Gensim Word2Vec algorithm
- Vocabulary of over 1.1 mio words
Conclusion

- Training domain-specific embedding models can
  
  - improve the semantic representation of technical terms within the vector space and
  
  - improve the results of domain-specific classification tasks,
  
  - even if the model was trained on a smaller data set than a general purpose model.
Problem:
- Crawler provided us with ~600,000 unlabeled documents
- ROKIN wants the documents labeled according to their pre-defined topics

- Manual labeling is not possible considering the huge amount of documents

→ How to classify huge amounts of documents unsupervised?
**Lbl2Vec**

An Embedding-Based Approach for Unsupervised Document Retrieval on Predefined Topics

**Idea:**

- Learn jointly embedded semantic representations of words and documents
- Learn label embedding from predefined topic description keywords
- Assign class of most similar label embedding to each document
Lbl2Vec
An Embedding-Based Approach for Unsupervised Document Retrieval on Predefined Topics

Classification Evaluation

<table>
<thead>
<tr>
<th>Classification Method</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised baseline</td>
<td>76.6</td>
</tr>
<tr>
<td>Lbl2Vec</td>
<td>82.7</td>
</tr>
<tr>
<td>Supervised Naive Bayes</td>
<td>89.8</td>
</tr>
</tbody>
</table>

Conclusion

- Lbl2Vec can
  - create better representations of predefined topics than standard modeling approaches,
  - yield better unsupervised document classification results than previous approaches,
  - but providing labels for each document is paramount for highly accurate classification results.
Technology Scouting as a Service
Research Team

Further research:

- Evaluation of different approaches to train BERT for classification in the engineering domain
- Evaluation of semantic linking capabilities between engineering specific word embeddings in English and German
- Classification of new technologies in engineering articles with neural networks
- Information extraction of technologies, products, product properties, and companies from engineering articles
- And much more ...

Prof. Dr. Florian Matthes
matthes@tum.de

Daniel Braun
daniel.braun@tum.de

Alexandra Klymenko
alexandra.klymenko@tum.de

Tim Schopf
tim.schopf@tum.de

Thomas Kinkeldei
thomas.kinkeldei@rokin.tech
M.Sc.
Tim Schopf
Research Associate

Technische Universität München
Faculty of Informatics
Chair of Software Engineering for Business Information Systems

Boltzmannstraße 3
85748 Garching bei München

Tel  +49 89 289-17105
Fax  +49 89 289-17136

tim.schopf@tum.de
www.matthes.in.tum.de