Identifying Emerging Technologies in the Renewable Energy Sector revisited

Andrea Zielinski, Denilton Luiz Darold and Jiao Jiao

Fraunhofer Institute for Systems and Innovation Research (ISI), Karlsruhe, Germany

Introduction

Identifying emerging technologies for Science, Technology, and Innovation (STI) studies is a challenging task which – to some extend – can be supported by Text Mining and Natural Language Processing methods (Porter, 2020; Bongiorno, 2020), apart from bibliometric and network analysis.

The text mining software "Tools for Innovation Monitoring" (TIM) has been specifically designed for this task. It extracts a set of relevant keywords from a corpus of scientific abstracts and metadata (i.e., author keywords), patents and grant proposals based on two different count-based/TF-IDF methods. In a recent paper, Moro et al. (2020) compare the results from TIM to expert reviews in the fields of renewable energies and find that emerging technologies identified by the experts could also be retrieved by TIM, i.e. the expert keywords were also included in the 300 top-ranked keywords with a probability of 65 % (in the best case), but only 25 % (in the worst case), depending on the technology sector.

In this work, we benchmark the performance of TIM for a specific use case against more recent Text Mining methods in the area of Automatic Terminology Extraction (ATE) based on word embeddings (Cram et al., 2016). Our main contributions in this work focus on a) a procedure for applying term extraction algorithms to the dataset; b) a discussion on the differences between the static and dynamic ATE models for this type of data; and c) investigating and comparing the results, i.e. the extracted terms and their ranking.

While recent results suggest that ATE and trend detection can benefit from latent contextual information, it is still an open question what the best choice of the pre-trained representations is, and if this has an impact on the downstream task of finding emergent keywords.

Methodology

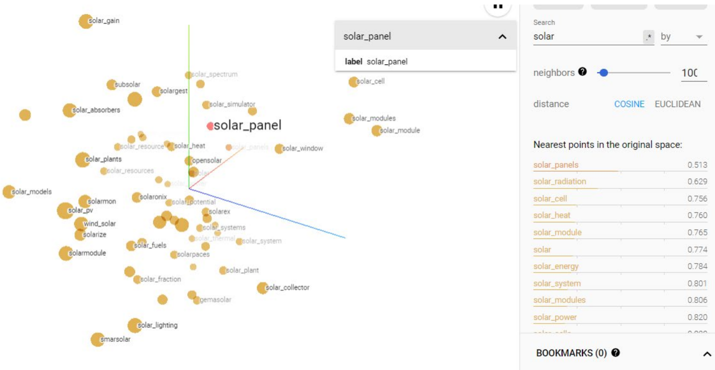
We compile a dataset, focusing on the renewable energy sector and re-use the setup as described in Moro et al. (2020) but rely only on Scopus abstracts which was found to be the most valuable source for finding emergent technologies. The use of distributed representations of words (Mikolov, 2014; Pennington, 2014) and transformer models (Devlin et al. 2018: Liu et al. 2019) have advanced the state-of-the-art performance across many tasks, including ATE (Terryn, 2020), because they can capture semantic variability. In our experiments, we will use *TermSuite* (Cram et al.) and a multi-label classifier for term prediction based on pre-trained transformers (Lang et al., 2021). To yield optimal performance, however, domain adaptation is required. To this end, off-the-shelf pre-trained embeddings of a closely-related domain are incorporated, e.g. *KNOWMAK* (Maynard, 2020), and new embeddings are trained on in-domain data. Note that clustering term variants (also known as term clumping) is largely facilitated this way, since modelling words in distributed spaces bring up semantically related keywords automatically. Based on ATE output, we experiment with regression models for modelling the time series data and the growth of scientific keywords (Asooja et al., 2016).

Another strain of work is devoted to diachronic data analysis and forecasting emerging trends based on dynamic word embeddings that can capture semantic variability over time. Dynamic embedding models also capture how the context of certain concepts might shift over time, exploring the similarities between pairs of keywords. We use the framework proposed in Dridi (2020) to detect emerging research trends in our dataset. The semantic change between keyword pairs is measured by means of cosine distance at different periods (Hamilton et al., 2016).

Preliminary Results

We evaluate the models on four datasets: a collection of 160,000 English Scopus abstracts, based on a database search using queries related to ‘renewable energies’. We used four specific Boolean search queries for the fields *solar photovoltaics (PV), wind power, ocean and tidal energy, hydropower*.

We trained Word2Vec embeddings on our dataset and assessed their quality intrinsically



solar_panel

label solar_panel

Search solar by

neighbors 100

distance COSINE EUCLIDEAN

Nearest points in the original space:

solar_panels	0.513
solar_radiation	0.629
solar_cell	0.756
solar_heat	0.760
solar_module	0.765
solar	0.774
solar_energy	0.784
solar_system	0.801
solar_modules	0.806
solar_power	0.820

BOOKMARKS (0)

Figure 1 shows the word embedding of ‘solar panel’ trained on Scopus 2014-2020

The entire workflow is as follows:

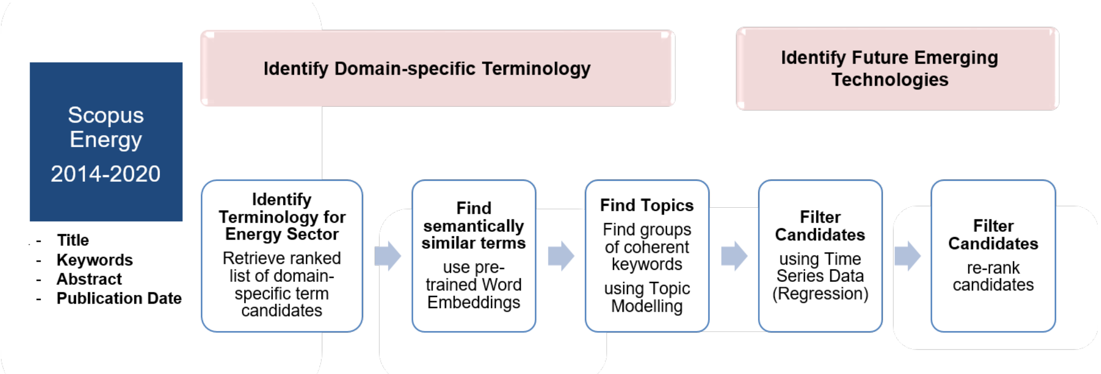

Scopus Energy 2014-2020
- Title
- Keywords
- Abstract
- Publication Date
Identify Domain-specific Terminology
Identify Future Emerging Technologies
Identify Terminology for Energy Sector
Retrieve ranked list of domain-specific term candidates
Find semantically similar terms
use pre-trained Word Embeddings
Find Topics
Find groups of coherent keywords
using Topic Modelling
Filter Candidates
using Time Series Data (Regression)
Filter Candidates
re-rank candidates

Figure 2. The Workflow for our Text Mining Pipeline

Conclusion and Outlook

Our main contributions in this work focus on a) a procedure for applying term extraction algorithms to the dataset; b) a discussion on the differences between the static and dynamic ATE models for this type of data; and c) investigating and comparing the results, i.e. the extracted terms and their ranking.

References

Asooja, K., Bordea, G., Vulcu, G., & Buitelaar, P. (2016). Forecasting emerging trends from scientific literature. In *Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 417-420).

Bongiorno, D. L., Prakasan, N., Truswell, J., Posadowski, M., & Walsh, J. (2020). AiCE: automating horizon scanning for the detection of emerging technologies. In *2020 IEEE SSCI* (pp. 1751-1756). IEEE.

Cram, D., & Daille, B. (2016). Terminology Extraction with Term Variant Detection. *ACL*. TermSuite https://termsuite.github.io/

Devlin J, Chang M, Lee K, Toutanova K. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv

Dridi, A., Gaber, M. M., Azad, R. M. A., & Bhogal, J. (2019). Leap2trend: A temporal word embedding approach for instant detection of emerging scientific trends. *IEEE Access*, *7*, 176414-176428.

Hamilton, W.L., Leskovec, J., Dan Jurafsky, D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *ACL*.

Lang, Christian et al. (2021) "Transforming Term Extraction: Transformer-Based Approaches to Multilingual Term Extraction Across Domains." *FINDINGS ACL*.

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv

Maynard, D., Lepori, B., Petrak, J., Song, X., & Laredo, P. (2020). Using ontologies to map between research data and policymakers' presumptions: the experience of the KNOWMAK project. *Scientometrics*, *125*(2), 1275-1290.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *NIPS*.

Moro, A., Joanny, G., & Moretti, C. (2020). Emerging technologies in the renewable energy sector: A comparison of expert review with a text mining software. *Futures, 117*, 102511.

Pennington, J., Socher, R., & Manning, C.D. (2014). Glove: Global Vectors for Word Representation. *EMNLP*.

Porter, A. L., Chiavetta, D., & Newman, N. C. (2020). Measuring tech emergence: A contest. *Technological Forecasting and Social Change*, *159*, 120176.

Terryn, A.R., Hoste, V., Drouin, P., & Lefever, E. (2020). TermEval 2020: Shared Task on Automatic Term Extraction Using the Annotated Corpora for Term Extraction Research (ACTER) Dataset. *COMPUTERM*.