A hybrid approach for creating knowledge graphs: recognizing emerging technologies in Dutch companies

Bakker, R.M.; Boer, M.H.T. de; Meyer-Vitali, A.P.; Bakker, B.J.; Raaijmakers, S.A.

HHAI2022: Augmenting Human Intellect
S. Schlobach et al. (Eds.)

doi:10.3233/FAIA220226

A Hybrid Approach for Creating Knowledge Graphs: Recognizing Emerging Technologies in Dutch Companies

Roos M. BAKKER [a,1], Maaike H.T. DE BOER [a] and André P. MEYER-VITALI [c] and Babette J. BAKKER [b] and Stephan A. RAAIJMAKERS [a]

[a] *Data Science dep., TNO, The Netherlands*
[b] *Strategy & Policy dep., TNO, The Netherlands*
[c] *Agents and Simulated Reality dep., DFKI, Germany*

Keywords. Knowledge Graphs, Text Mining, Foresight, Hybrid AI

Automatically recognizing emerging technologies in textual data is a challenge. Text-mining solutions based on user-defined keywords may create selection bias and inability to discover new technologies. In this work, we propose a hybrid approach to answer the following research question: Can information extraction techniques be combined with domain expert knowledge to produce a knowledge graph that is representative for the domain and usable for classification of companies working on emerging technologies? We combine different methods of information extraction together with domain expert knowledge for the semi-automatic creation of a knowledge graph, and test this methodology in a trend recognition use case that was inspired and partially supported by the province of South Holland, CBS, and Innovatiespotter.

Emerging technologies can have a large economical, political, and societal impact. They can present both opportunities and threats for innovation to companies [1,2,3]. To gain insight in relevant emerging technologies for corporate foresight, commercial companies can use data to monitor innovation trends, such as Google Trends. Such applications often use keywords that are identified by a user [4]. This approach can lead to bias since the keywords are based on the knowledge of the user, and it will not be able to detect emerging topics since they are described with new keywords. These pitfalls can be tackled by automatically extracting relevant keywords or phrases from the field of interest. Such approaches have been used in the past as described by Mühlroth & Grottke [5], but one of their main findings is that there is still need for increasing the level of automation and using the human resources in a later stage of evaluation and decision making. The increasing of the level of automation can be done by using Open Information Extraction (OIE) techniques. Existing tools include TextRunner [6], ReVerb, and Graphene [7,8]. These tools do not only extract keywords, but often also provide relations between

[1] Corresponding Author: Roos M. Bakker, TNO - Netherlands Organisation for Applied Scientific Research, Anna van Buerenplein 1, 2595 DA, The Hague, The Netherlands; E-mail: roos.bakker@tno.nl.

them. Algorithms often used in OIE behind these tools include co-occurrences, OpenIE, dependency parsing, Hearst patterns and word2vec [9].

In this work, we automated the selection of relevant terms of a domain by extracting a knowledge graph from textual data, and use human evaluation to improve and evaluate it. We tested this process on a use case on recognizing trends on the fields of artificial intelligence (AI) and synthetic biology (SB). We divided these fields in subtopics such as Natural Language Processing and Machine Learning, and gathered representative papers as input for the five different algorithms mentioned above. We first extracted knowledge graphs from these topics, and then filtered out common words (e.g. *the*) and words that occur very often (e.g. *ideas, beliefs*) based on word frequencies. For the last step, a set of domain experts labeled non-relevant concepts and added missing concepts. Finally, the resulting graph was qualitatively evaluated by another domain expert.

The resulting knowledge graphs before evaluation of the domain expert showed large differences in the amount of concepts and relations. For instance, on average 964 concepts were extracted in papers on the field of AI using the Hearst patterns algorithm, where hypernym relations are extracted using the lexico-syntactic patterns introduced by Hearst [10]. With word2vec, only 40 concepts were extracted. The knowledge graphs about the field of SB produced less concepts for all algorithms, for instance the average number of concepts and relations for Hearst patterns was 161 and 100, and for word2vec 39 and 55 respectively. For word2vec, the domain experts needed to add many multi-word phrases such as *speech recognition*, since the algorithm only recognized single words. For Hearst patterns, many concepts were not specific enough and needed to be removed. This indicates that a higher amount of concepts extracted by the algorithms lead to more filtering required from the domain experts, whereas a lower number requires additions. The final versions were all evaluated as representative for the domain, and usable for a future rule-based classifier by domain experts.

Since all the graphs were evaluated as representative, we can conclude that our method is a promising approach. However, we see large differences in the number of concepts and relations that are extracted by both algorithms. The word2vec algorithm produced relevant concepts on a high level according to the domain experts, but combined phrases such as *neural networks* were missing and could also be added automatically in the future. Additionally, the domain expert had to filter out many concepts manually for the Hearst patterns algorithm, and for word2vec concepts had to be added. The filtering took less time, which indicates that a higher number of concepts and relations will reduce the manual work. We hypothesize that manually searching for trends is more labour intensive. It would be interesting for future work to compare a manual search to our method to measure the time reduction. In future work we plan on improving on these points and on using the knowledge graphs in classification tasks. We aim to classify companies by working on the selected subtopics, based on whether their website texts contain a minimum amount of concepts from the knowledge graphs. The domain expert turned out to be essential for the quality of the graph, therefore we plan on creating a more interactive setting where system and expert mutually improve the graph. It would also be interesting to extract the graphs over a longer period of time to measure the success of emerging technologies. Further future work lies in testing this approach with keyword-based methods, since keywords might be enough as input for a rule-based classifier. The combination of a keyword-based method together with the above approach may prove to be a further step towards the automation of extracting high quality knowledge graphs.

References

[1] Geurts A. A critical review of Alec Ross's The Industries of the Future. Technological Forecasting and Social Change. 2018;128(1):311-3.
[2] Cozzens S, Gatchair S, Kang J, Kim KS, Lee HJ, Ordóñez G, et al. Emerging technologies: quantitative identification and measurement. Technology Analysis & Strategic Management. 2010;22(3):361-76.
[3] Martin BR. Foresight in science and technology. Technology analysis & strategic management. 1995;7(2):139-68.
[4] Braaksma B, Daas P, Raaijmakers S, Geurts A, Meyer-Vitali A. AI-Supported Innovation Monitoring. In: International Workshop on the Foundations of Trustworthy AI Integrating Learning, Optimization and Reasoning. Springer; 2020. p. 220-6.
[5] Mühlroth C, Grottke M. A systematic literature review of mining weak signals and trends for corporate foresight. Journal of Business Economics. 2018;88(5):643-87.
[6] Yates A, et al. Textrunner: open information extraction on the web. In: Proc. of Human Language Technologies: The Annual Conf. of the N.-A. Chapter of the Association for Comp. Ling.: Demonstrations. Association for Computational Linguistics; 2007. p. 25-6.
[7] Niklaus C, Cetto M, Freitas A, Handschuh S. A Survey on Open Information Extraction. arXiv preprint arXiv:180605599. 2018.
[8] Glauber R, Claro DB. A systematic mapping study on open information extraction. Expert Systems with Applications. 2018;112:372-87.
[9] De Boer MH, Lu YJ, Zhang H, Schutte K, Ngo CW, Kraaij W. Semantic reasoning in zero example video event retrieval. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM). 2017;13(4):60.
[10] Hearst MA. Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th conference on Computational linguistics-Volume 2. Association for Computational Linguistics; 1992. p. 539-45.