VOCAB-EXPANDER: A System for Creating Domain-Specific Vocabularies Based on Word Embeddings

Michael Färber
Karlsruhe Institute of Technology (KIT)
Institute AIFB
Karlsruhe, Germany
michael.faerber@kit.edu

Nicholas Popovic
Karlsruhe Institute of Technology (KIT)
Institute AIFB
Karlsruhe, Germany
popovic@kit.edu

Abstract

In this paper, we propose VOCAB-EXPANDER at https://vocab-expander.com, an online tool that enables end-users (e.g., technology scouts) to create and expand a vocabulary of their domain of interest. It utilizes an ensemble of state-of-the-art word embedding techniques based on web text and ConceptNet, a common-sense knowledge base, to suggest related terms for already given terms. The system has an easy-to-use interface that allows users to quickly confirm or reject term suggestions. VOCAB-EXPANDER offers a variety of potential use cases, such as improving concept-based information retrieval in technology and innovation management, enhancing communication and collaboration within organizations or interdisciplinary projects, and creating vocabularies for specific courses in education.

1 Introduction

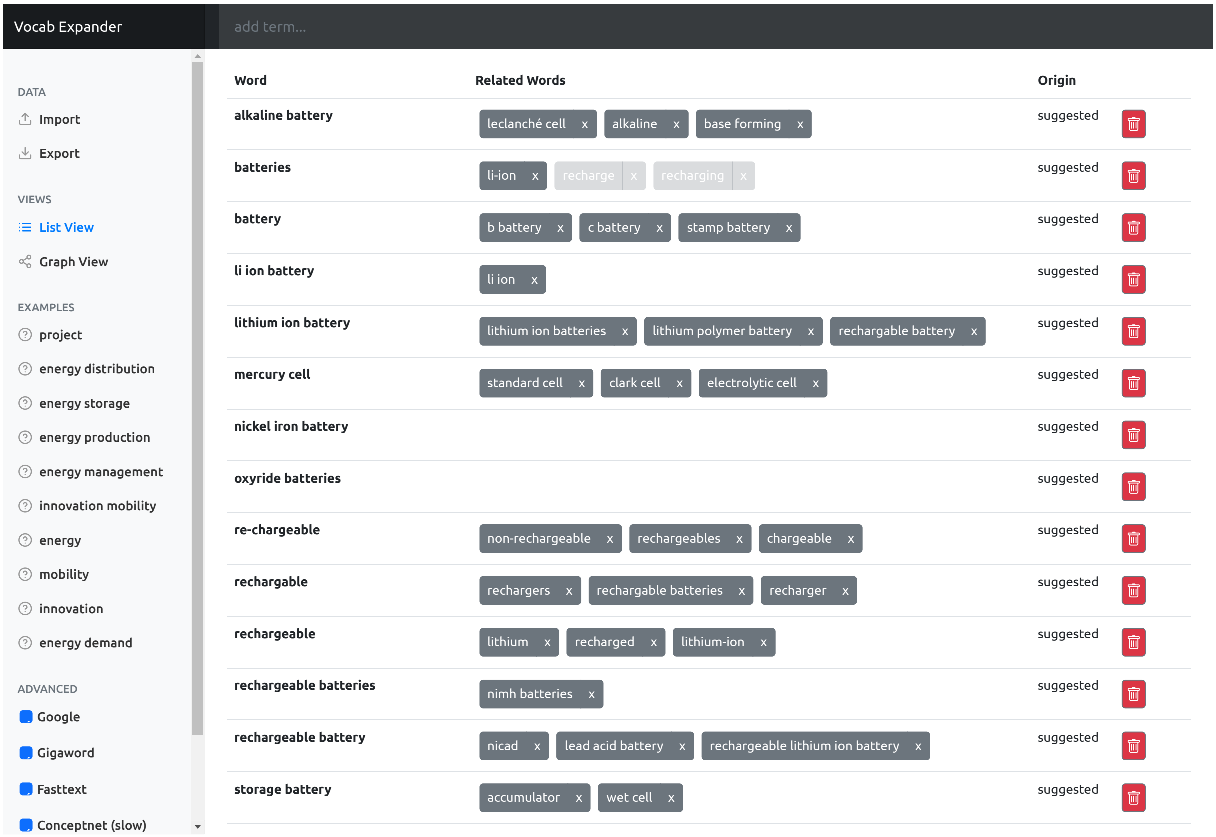
Motivation. In many scenarios, it is necessary to create an ontology or other formal model of a domain of interest from scratch. For instance, in the field of technology and innovation management, technology scouts and other end-users without technical skills often use a list of terms for continuously retrieving and scanning texts from different media sources (e.g., news articles, social media, publications, patents) in order to become aware of novel relevant technologies and to create and populate profiles of technologies and actors within a particular domain, such as Smart Cities. However, coming up with such a vocabulary is typically highly time-consuming and costly due to the domain-specificity (i.e., non-experts have no starting point what to add), the complexity of correctly defining the scope of the domain (e.g., Smart Cities can range from Smart Home to energy efficiency to security), the ambiguity of natural language (i.e., the meaning of terms may vary depending on the context in which they are used), and the emergence of new terms over time.

Current Situation. So far, domain experts (e.g., technology scouts in technology and innovation management) still rely heavily on domain expert knowledge (de Weck, 2022). Several tools for modeling a domain of interest exist, including Protegé (Musen, 2015) and D-Terminer (Rigouts Terryn et al., 2022). However, these tools are often considered as "too heavy" for creating only a domain-specific vocabulary instead of an ontology with a specific data model and standardizations (e.g., W3C RDF, OWL). Furthermore, these tools are typically designed to support the modeling process (e.g., based on an existing text corpus), but do not suggest directly related terms for given terms.

Contributions. In this paper, we propose the system VOCAB-EXPANDER, available online at https://vocab-expander.com, that enables end-users without technical skills to create and expand a vocabulary of their domain of interest. The system utilizes an ensemble of state-of-the-art word embedding techniques to suggest related terms for already given terms. In addition to word embedding models based on web text, the system also incorporates embeddings based on ConceptNet, a common-sense knowledge base. The system is equipped with an easy-to-use interface that allows end-users to quickly confirm or reject term suggestions. The ranking of the suggested terms is based on the number of links they possess to other terms within the vocabulary. The created vocabulary can be listed as a table and visualized as a graph (see Figures 1 and 2). We also provide an import and export functionality for the vocabularies.

Use Cases. Our tool offers a variety of potential use cases. For tasks such as technology and innovation management, it can be used to improve concept-based information retrieval by utilizing the created domain-specific vocabulary as search terms.

Proceedings of Recent Advances in Natural Language Processing, pages 331–335
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_037



Word	Related Words	Origin
alkaline battery	leclanché cell x alkaline x base forming x	suggested
batteries	li-ion x recharge x recharging x	suggested
battery	b battery x c battery x stamp battery x	suggested
li ion battery	li ion x	suggested
lithium ion battery	lithium ion batteries x lithium polymer battery x rechargable battery x	suggested
mercury cell	standard cell x clark cell x electrolytic cell x	suggested
nickel iron battery		suggested
oxyride batteries		suggested
re-chargeable	non-rechargeable x rechargeables x chargeable x	suggested
rechargable	rechargers x rechargable batteries x recharger x	suggested
rechargeable	lithium x recharged x lithium-ion x	suggested
rechargeable batteries	nimh batteries x	suggested
rechargeable battery	nicad x lead acid battery x rechargeable lithium ion battery x	suggested
storage battery	accumulator x wet cell x	suggested

Figure 1: Screenshot of the VOCAB-EXPANDER, available at https://vocab-expander.com.

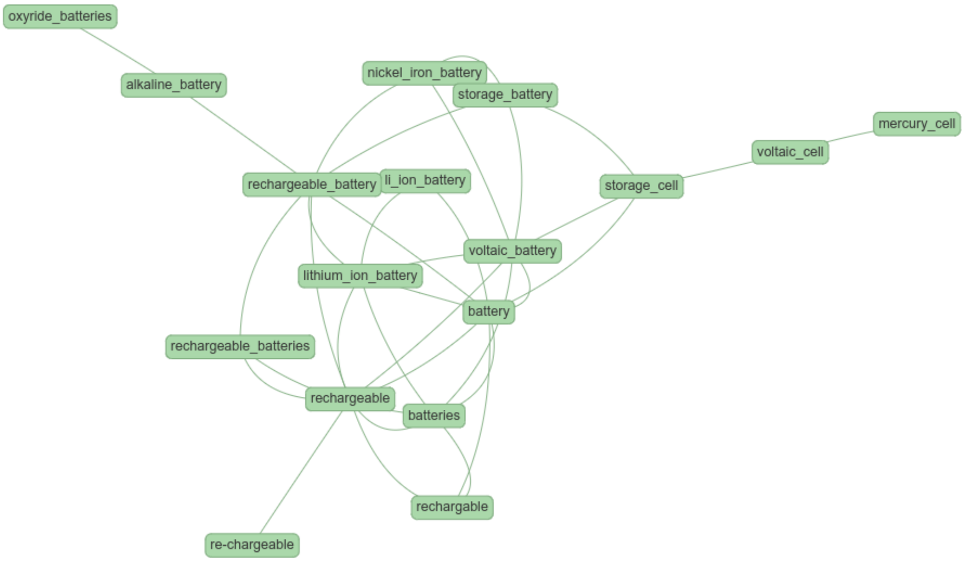


Figure 2: Screenshot

Additionally, the created vocabulary can serve as a basis to enhance communication and collaboration within organizations or interdisciplinary projects by ensuring the use of consistent terminology among all involved parties. In the field of education, our tool allows for the creation of vocabularies for specific courses or subjects, ensuring that all relevant terms within a field or subject are covered. Overall, our system provides a valuable solution for creating and maintaining domain-specific vocabularies, which can be used in various fields to improve information retrieval, human communication, and natural language processing.

Provisioning. The source code of our system is publicly available on GitHub (`https://github.com/nicpopovic/VocabExpander`) under the MIT License, making it easy to reuse and adapt for a wide range of use cases.

2 System Design

The system utilizes an ensemble E of state-of-the-art pre-trained word embedding models available in *gensim* (Řehůřek and Sojka, 2010) to suggest related terms for already given terms. Specifically, the user can choose one or several of the following models: (1) *word2vec-google-news-300* (Mikolov et al., 2013), (2) *glove-wiki-gigaword-300* (Pennington et al., 2014), (3) *fasttext-wiki-news-subwords-300* (Mikolov et al., 2018), (4) *conceptnet-numberbatch-17-06-300* (Speer et al., 2017).

Words $w \in W$ are categorized into 3 categories, accepted words W_a, rejected words W_r, and suggested words W_s. Initially, a user adds one or more words to W_a. For each word $w_a \in W_a$ the top k most similar words $w_{sim} \in W_{sim}$ according to each embedding model $e \in E$ are fetched along with the average pairwise similarity scores $P_{w_{sim,j},w_i}$ across E. $w_{sim} \notin W$ are added to W_s. Next, we calculate a score S_{w_s} for each suggested word $w_s \in W_s$ by aggregating similarity scores to accepted words and subtracting weighted similarity scores to rejected words:

$$S_{w_{s,i}} = \sum_{w_{a,j} \in W_a} P_{w_{s,i},w_{a,j}} - \lambda \sum_{w_{r,k} \in W_r} P_{w_{s,i},w_{r,k}}$$

where $\lambda = 0.5$. Suggested words are then associated with the accepted word with which they have the highest pairwise similarity and ordered according to their score $S_{w_{s,i}}$.

The system's frontend, presented in Figure 1, displays the list view of accepted words and the corresponding suggested words. The list view showcases the three highest-ranked suggestions for a selected accepted word. If the score of a suggested word falls below a pre-determined threshold, a lower opacity indicates this. Users can quickly accept a suggested word by clicking on it or reject it by clicking the "x" button next to it. Additionally, a graph view is available as shown in Figure 2, allowing users to visualize the similarity scores between accepted words. The user interface also includes import and export buttons in the top left corner, enabling the import and export of vocabulary lists.

3 Related Work

Ontology Engineering and Ontology Learning. Various methods have been proposed for constructing an ontology for a specific domain in a manual, semi-automated, or automated way (Hazman et al., 2011). Automated methods typically involve extracting concepts and relations between them from domain-specific text corpora provided by the user (Elnagar et al., 2020). In contrast to them, our approach does not rely on the availability of a large text corpus; instead, we enable users (domain experts as well as newcomers) to independently explore and discover related concepts from scratch. Furthermore, ontology learning (Buitelaar et al., 2005) typically includes additional processing steps, which are out of our scope, such as clustering the concepts with identical or similar meanings and assigning unique identifiers to concepts.

Automated Term Extraction. Research on automatically extracting terms from text corpora, such as named entities, has been performed extensively. Early approaches on automatic term extraction combined linguistic hints, e.g., part-of-speech patterns, with statistical metrics for calculating the termhood and unithood (Kageura and Umino, 1996), which allows to quantify to which degree the candidate term is related to the domain. Rule-based approaches have been used through many years (e.g., Daille (1994); Drouin (2003)) and are still popular nowadays (Kosa et al., 2020). Machine learning-based approaches for automated term extraction utilize, among other things, external data sets and web search (Ramisch et al., 2010) and word embeddings (Wang et al., 2016; Amjadian et al., 2018). Newest approaches are also based on language models (e.g., (Gao and Yuan, 2019; Lang et al., 2021)), but require, as many other approaches, more context than a few keywords as input as for our system.

Demo Systems for Automated Term Extraction. Rigouts Terryn et al. (2022) proposed D-Terminer, a running system for monolignual and bilingual automatic term extraction. In contrast to us, they focus on multiple languages, and use a text corpus as input for the system. Additionally, TermoStat (Drouin, 2003) and TerMine (Frantzi et al., 2000) are examples of online systems for term extraction and rely on rule-based hybrid approaches. Finally, MultiTerm Extract[1] and SketchEngine[2] are available commercial systems.

4 Conclusion

In this paper, we proposed VOCAB-EXPANDER, an online tool that enables end-users to create and expand a vocabulary of their domain of interest. It uses state-of-the-art word embedding techniques based on web text as well as ConceptNet, a common-sense knowledge base, to suggest related terms for already given terms. The system can be used for a variety of purposes such as improving information retrieval, communication and collaboration, creating vocabularies for education, and fine-tuning language models in natural language processing.

For the future, we will allow for the integration of domain-specific text corpora (e.g., provided by the domain experts) and provide a functionality to see to which degree the suggested terms occur in the text corpora. Furthermore, we plan to evaluate the performance of VOCAB-EXPANDER by means of user studies in different domains and applications.

Acknowledgments

This work was partially supported by the German Federal Ministry of Education and Research (BMBF) as part of the project IIDI (01IS21026D).

References

Ehsan Amjadian, Diana Inkpen, T Sima Paribakht, and Farahnaz Faez. 2018. Distributed specificity for automatic terminology extraction. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 24(1):23–40.

Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini. 2005. Ontology learning from text: An overview. *Ontology learning from text: Methods, evaluation and applications*, 123.

Béatrice Daille. 1994. *Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques*. Ph.D. thesis, Paris 7.

Patrick Drouin. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–115.

Samaa Elnagar, Victoria Y. Yoon, and Manoj A. Thomas. 2020. An automatic ontology generation framework with an organizational perspective. In *53rd Hawaii International Conference on System Sciences, HICSS 2020, Maui, Hawaii, USA, January 7-10, 2020*, pages 1–10. ScholarSpace.

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms:. the c-value/nc-value method. *International journal on digital libraries*, 3(2):115–130.

Yuze Gao and Yu Yuan. 2019. Feature-less end-to-end nested term extraction. In *CCF international conference on natural language processing and Chinese computing*, pages 607–616. Springer.

Maryam Hazman, Samhaa R El-Beltagy, and Ahmed Rafea. 2011. A survey of ontology learning approaches. *International Journal of Computer Applications*, 22(9):36–43.

Kyo Kageura and Bin Umino. 1996. Methods of automatic term recognition: A review. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 3(2):259–289.

Victoria Kosa, David Chaves-Fraga, Hennadii Dobrovolskyi, and Vadim Ermolayev. 2020. Optimized term extraction method based on computing merged partial c-values. In *Information and Communication Technologies in Education, Research, and Industrial Applications: 15th International Conference, ICTERI 2019, Kherson, Ukraine, June 12–15, 2019, Revised Selected Papers*, pages 24–49. Springer.

Christian Lang, Lennart Wachowiak, Barbara Heinisch, and Dagmar Gromann. 2021. Transforming term extraction: Transformer-based approaches to multilingual term extraction across domains. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3607–3620.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[1] https://www.trados.com/de/products/multiterm-desktop/

[2] https://sketchengine.eu

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Mark A. Musen. 2015. The protégé project: a look back and a look forward. *AI Matters*, 1(4):4–12.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. Mwetoolkit: a framework for multiword expression identification. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Ayla Rigouts Terryn, Veronique Hoste, and Els Lefever. 2022. D-terminer: Online demo for monolingual and bilingual automatic term extraction. In *Proceedings of the TERM21 Workshop*, pages 33–40. Language Resources and Evaluation Conference (LREC 2022).

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. pages 4444–4451.

Rui Wang, Wei Liu, and Chris McDonald. 2016. Featureless domain-specific term extraction with minimal labelled data. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 103–112.

Olivier L de Weck. 2022. Technology scouting. In *Technology Roadmapping and Development: A Quantitative Approach to the Management of Technology*, pages 395–424. Springer.