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.
<table>
<thead>
<tr>
<th>Word</th>
<th>Related Words</th>
<th>Origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>alkaline battery</td>
<td>leclanché cell • alkaline • base forming</td>
<td>suggested</td>
</tr>
<tr>
<td>batteries</td>
<td>li-ion • battery • batteries</td>
<td>suggested</td>
</tr>
<tr>
<td>battery</td>
<td>li battery • t battery • stamp battery</td>
<td>suggested</td>
</tr>
<tr>
<td>li ion battery</td>
<td>li ion</td>
<td>suggested</td>
</tr>
<tr>
<td>lithium ion battery</td>
<td>lithium ion batteries • lithium polymer battery • rechargeable battery</td>
<td>suggested</td>
</tr>
<tr>
<td>mercury cell</td>
<td>standard cell • clark cell • electrolytic cell</td>
<td>suggested</td>
</tr>
<tr>
<td>nickel iron battery</td>
<td></td>
<td>suggested</td>
</tr>
<tr>
<td>oxyride batteries</td>
<td></td>
<td>suggested</td>
</tr>
<tr>
<td>re-chargeable</td>
<td>non-rechargeable • rechargeables • chargeable</td>
<td>suggested</td>
</tr>
<tr>
<td>rechargeable</td>
<td>chargers • rechargeable batteries • recharger</td>
<td>suggested</td>
</tr>
<tr>
<td>rechargeable</td>
<td>lithium • recharged • lithium-ion</td>
<td>suggested</td>
</tr>
<tr>
<td>rechargeable batteries</td>
<td>nihm batteries</td>
<td>suggested</td>
</tr>
<tr>
<td>rechargeable battery</td>
<td>nical • lead acid battery • rechargeable lithium ion battery</td>
<td>suggested</td>
</tr>
<tr>
<td>storage battery</td>
<td>accumulator • wet cell</td>
<td>suggested</td>
</tr>
</tbody>
</table>

Figure 1: Screenshot of the Vocab Expander, available at [https://vocab-expander.com](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/nicopopovic/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 (Rehůrek 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 $w_s \in W_s$ the top $k$ most similar words $w_{sim} \in W_{sim}$ are fetched along with the average pairwise similarity scores $P_{w_{sim},j,w_s}$ 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:

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

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 monolingual 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\textsuperscript{1} and SketchEngine\textsuperscript{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


