CAVA: An open source R toolkit for dictionary Coherence, Adaptation, Validation, and Analysis

Wouter van Atteveldt; Dafne van Kuppevelt; Kasper Welbers

Dictionaries remain an important tool for computational communication research (Boumans & Trilling 2016; Praet et al, 2021; De Vries, 2021). Simply put, in dictionary analysis, concepts are traced using a list of indicator words (or search terms, keywords). At its simplest, such dictionaries consist of lists of single words, but they can also include wildcards, multi-word terms, and/or boolean constructs of words that must or should not co-occur. Even studies which rely on machine learning or manual annotation for the actual coding often use (simply) keyword searches to define the population of interest.

The main benefits of dictionaries are their ease of use and transparency. Especially compared to machine learning (ML) models, it is easy to inspect, understand, and apply a dictionary. There are two main drawbacks, however. First, the very simplicity of dictionaries means that in general ML will outperform dictionaries on most serious tasks. Second, it is not trivial to properly construct a dictionary, validate it, and adapt it to the task at hand (if needed). In fact, even though it is often repeated that validation and domain adaptation are critical for successfully using dictionaries (e.g. Van Atteveldt et al., 2021), very often dictionaries are still applied with little scrutiny or adaptation. We believe that this is partly due to lacking tooling for working with dictionaries in a good way. There are many tools that make it really easy to apply an existing dictionary (e.g. quanteda; Benoit et al., 2018), but do not propel researchers to improve and validate the dictionary.

To remedy this, we present **CAVA:** dictionary Coherence, Augmentation, Validation, and Analysis. This is an open-source R package for dealing with dictionaries that helps the researcher inspect, improve, adapt, and validate their dictionaries.

Functionality:

CAVA combines word embedding models and corpus analysis to offer the following main functionality:

- **Semantic coherence:** Using CAVA, the researcher can inspect the semantic coherence of a dictionary: which terms belong together based on their semantic values according to the word embedding? This can be shown as a distance-from-centroid statistic and visualized as a semantic graph (Figure 1). This shows which different meaning clusters are present in a dictionary, and allows the researcher to identify terms that seemingly do not belong to the dictionary. For example, Figure 1 shows the dictionary created by the wildcard expansion of “eco*”, where terms related to “ecology” are clearly separate from terms related to “economy”
- **Domain adaptation:** Off-the-shelf dictionaries are often applied outside of the domain or genre for which they were created, and moreover word use in a domain can also change over time: “COVID” would not have been included in a dictionary for health issues created before 2019. By computed corpus statistics for a given target corpus, the researcher can quickly check the most frequent words to see if they are correctly classified. Moreover, suggestions for terms to be added or removed can be automatically

included for the researcher to consider. These suggestions can stem from the embedding model, or from comparing matching to non-matching documents or by using machine learning with matching documents as training data (cf. Hopkins & King, 2010).

- **Validation:** Given a target corpus and a dictionary, an open source in-browser annotation tool ([authors] 2022) can be launched to quickly do systematic validation of the results (see Figure 2). This can either be done directly from R by the researcher themself, or by posting the annotation job to a backend or crowd-coding platform depending on the required scale. The results are then automatically read back into R for analysis.
- **Analysis:** From the validation results, three important analyses can be performed automatically. First, the precision and recall (and sensitivity and specificity) of the dictionary is computed, including computed confidence intervals. This helps estimate not just how good the dictionary is, but also whether the validation was sufficiently thorough (which can be problematic especially for rare cases). Second, based on the validation suggestions for further improvement can be given, leading to an iterative process of improvement (assuming that a held-out sample is used for validating the end result). Finally, the substantive outcomes on the target corpus can be calculated, including estimated measurement and sampling errors.

The tool is designed to interoperate with existing packages such as *quanteda*, *tidytext*, and *corpustools*, making it easy to fit into your existing workflow. By making it much easier (and more visually pleasing) to work with dictionaries, we hope to improve the accessibility and especially the quality of dictionary analyses in communication science.

Validation

The validity and efficacy of dictionary expansion are tested on multiple existing datasets: multilingual political manifestos (using the CMP data and the Pimpo immigration dataset; Lehman & Zobel 2018), Dutch and English newspaper articles (using the VU Election Study and Comparative Agendas Project datasets[1]), and government policy documents on COVID (Coronanet; Cheng et al., 2020). In all cases, we started from a small (existing or crafted) dictionary of seed terms, and then used the method to expand these terms (either automatically or with minimal human intervention). This is then tested against the manual annotations of the texts as present in these data sets.

Results show that dictionary expansion generally provides a strong increase of recall at the expense of precision (as expected), but still have a significant positive effect on F1-score in most cases, depending strongly on the initial (seed) dictionary. In most cases there is a clear ceiling effect for recall, so even after adding all terms semantically close to the seed dictionary, some texts remain unfound.

[1] See https://www.comparativeagendas.net/ and https://github.com/vupolcom/VU-Election-Study

Technical Details

Word embeddings provide the main functionality for dictionary coherence and expansion in this tools (Mikolov et al, 2013). Word embeddings are essentially numerical representations of the meaning of words as distilled from millions or even billions of uncoded texts. The end result is that words which have similar representations (such as 'economy' and 'business') also have similar meaning. This effectively allows us to use the linguistic 'knowledge' distilled from these uncoded texts to find words that are semantically close to the existing dictionary, which can be seen as good candidates for inclusion. Reversely, terms in the dictionary that are semantically distant from existing terms (e.g. because they were included accidentally) are good candidates for further scrutiny. Technically, the distance between two words is defined as the cosine distance between their embedding vectors. This can be computed efficiently in R by using matrix multiplication of the centroid vector of the existing dictionary with the embedding matrix.

The second source of information for dictionary expansion and coherence is the target corpus (if given). First, effort to judge expansion candidates is reduced by including only words that exist within the target corpus and matching the frequency of words in the target corpus to the existing dictionary so they can be rank ordered. Second, expansion candidates can also be retrieved directly from the target corpus by finding words that occur more in documents that are found than in documents that are not found (using a simple chi-squared distribution). Corpus analysis and embeddings can be combined by computing the embedding vectors of documents (doc2vec; Le and Mikolov 2014) and looking for documents that are semantically similar to documents found by the dictionary, and treating words that occur in these documents as expansion candidates.

Figures

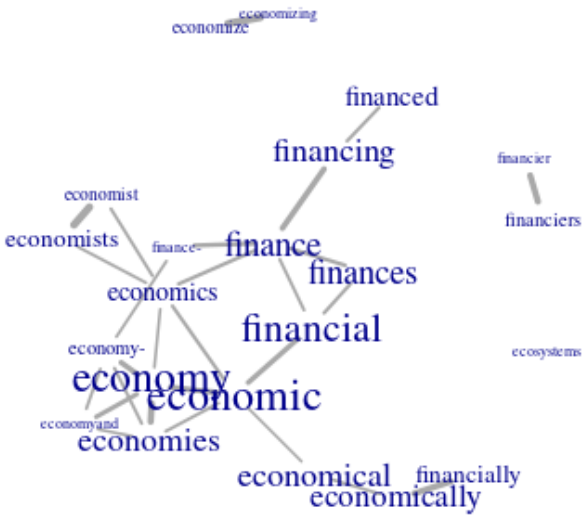

Figure 1: Lexical cohesion of a dictionary created by wildcard expansion of "eco finan*"*

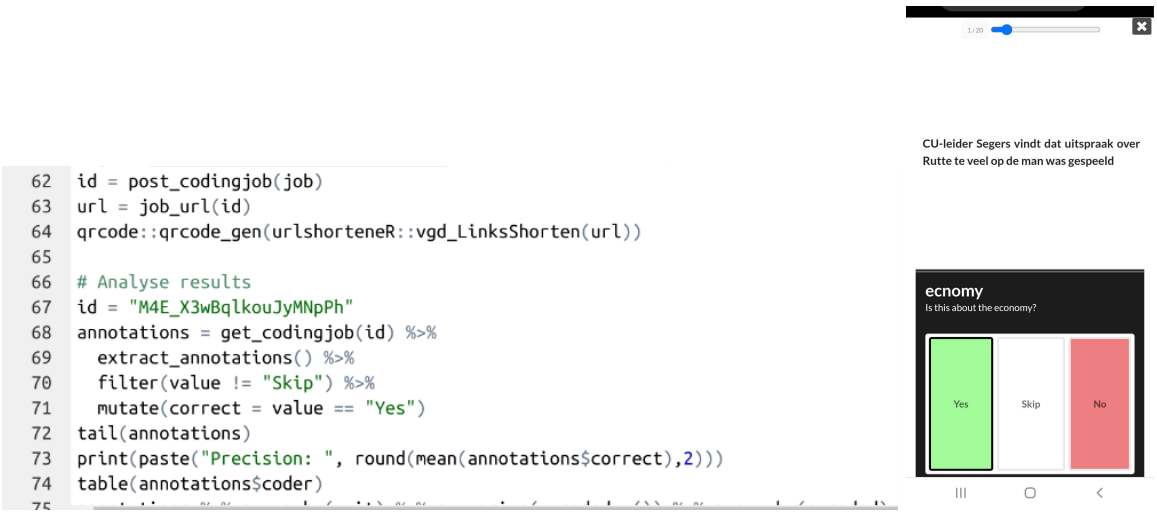

Figure 2: Using a web-based annotation interface to validate a dictionary from R

References

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30).

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. Digital journalism, 4(1), 8-23.

Cheng, Cindy, Joan Barceló, Allison Hartnett, Robert Kubinec, and Luca Messerschmidt. 2020. COVID-19 Government Response Event Dataset (CoronaNet v1.0). https://www.coronanet-project.org

De Vries, E. (2021), The Sentiment is in the Details: A Language-agnostic Approach to Dictionary Expansion and Sentence-level Sentiment Analysis in News Media. To be published in *Computational Communication Research*, https://osf.io/preprints/socarxiv/8y3jq

Le, Q. and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning* (ICML 2014), pages 1188–1196, Beijing, China.

Pola Lehmann, Malisa Zobel (2018): Positions and Saliency of Immigration in Party Manifestos. A Novel Dataset using Crowd Coding. European Journal of Political Research 57 (4). https://doi.org/10.1111/1475-6765.12266)

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* 26 (pp. 3111–3119). Curran Associates, Inc.

Praet, S., Van Aelst, P., Daelemans, W., Walgrave, S., Kreutz, T., Peeters, J., & Martens, D. (2021). Comparing automated content analysis methods to distinguish issue communication by political parties on Twitter. *Computational Communication Research*, 3(2), 195-219. https://doi.org/10.5117/CCR2021.2.004.PRAE

van Atteveldt, W., van der Velden, M. A., & Boukes, M. (2021). The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. *Communication Methods and Measures*, 15(2), 121-140.