9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

LEXpander: applying colexification networks to automated lexicon expansion

Keywords: colexification, semantic network, word list expansion, text analysis, benchmark

Extended Abstract

Thematic word lists or lexica, that is lists collecting words that deal with a chosen topic, are at the base of most text analysis applications. Be it to collect text snippets that address a specific topic (as suicide [8]) or to run text analysis on the texts themselves (like collective emotion [9]), word lists are ubiquitous and their usage is very common in Computational Social Science. Usually, word lists are hand crafted by researchers, adapted to a novel setting from previously published resources, for example through translation, or created with the help of automatic word list expansion algorithms. Despite the importance of the problem of their creation, previous work has not systematically assessed the quality of word list expansion methods nor compared existent tools against each other in a comprehensive way. Our work is the first to provide a benchmark to evaluate and compare word list expansion methods. This novel framework takes into account more than 70 different thematic word lists and can be used to test algorithms developed for expanding word lists in languages different from English. Additionally, we propose a new word list expansion algorithm based on a semantic network, LEXpander, and prove that it outperforms previous approaches in two different languages, English and German. LEXpander is available as a free, interactive web app where everyone can expand their word lists with just a few clicks.

LEXpander deploys a colexification network, that is the network built from colexification records. A colexification is a linguistic phenomenon that occurs when one language uses the same word to convey two different meanings [6]. This phenomenon is thought be related to semantic similarity, that is colexification instances that occur in multiple languages indicate similarity between the two meanings [6], as suggested by previous results in the realm of affective meaning [3]. LEXpander deploys this property to expand thematic word lists. In particular, colexifications can be shaped in the form of a network, where nodes represent concepts and edges link concepts that are colexified by a certain number of languages [7]. LEXpander maps seed words given as input onto the colexification network and retrieves their neighboring words, which will form the expanded word list (see Figure 1 for a graphical representation). One of the advantages of using a colexification network for this task is that its structure is independent from language, as the phenomenon of colexification is cross-linguistic. That is, the structure of the colexification network is super-lingual and is independent from the single language chosen. As a consequence, the same algorithm can be used to expand word lists in languages different from English, as we showcase with German thematic word lists.

In this work, we compare LEXpander with other lexicon expansion algorithms in a novel benchmark. In this comparison framework, we consider thematic word lists from LIWC in English [11] and German [14] as ground truth. We select random words from the ground truth word lists and expand them with the word list expansion algorithms, computing then the precision, recall and F_1 of the expanded word lists against the ground truth ones. Using this benchmark, we show that the network-based algorithm LEXpander outperforms other, widely spread methods: neural word embeddings as FastText [1] and GloVe [12] trained on English and German

textual data, other semantic networks like WordNet [10] and its German counterpart, OdeNet [13], and Empath 2.0, the re-implementation of a popular word list expansion method [4], in the expansion of both English and German word lists. For example, LEXpander achieves a mean F_1 score of 0.15 when expanding 30% randomly selected words from the English ground truth word lists, while the best method we compared it with, Empath 2.0, yields to a score of 0.12 (see Figure 2). We also manually annotate the English expanded word lists for negative and positive emotions to assess whether the lexicon expansion algorithms do propose words that fit the topic chosen but are not included in the relative LIWC word list, finding that LEXpander has the best performance. Moreover, we perform a sentiment analysis exercise on English texts from traditional press (the Brown [5] and the COHA corpora [2]) and online communication (text snippets from Twitter and Reddit) and find that the word lists produced by LEXpander achieve the best or is tied with the best performances on all databases.
The results of our study confirm the potential of the usage of linguistic resources and network science to address NLP problems and improve methods of text mining and analysis in behavioral research. Moreover, we show that our approach can be applied to German and potentially to other languages for which traditionally the existing resources are limited in comparison to English.

References

[1] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the ACL*, 5, 2017.

[2] M. Davies. Expanding horizons in historical linguistics with the 400-million word corpus of historical american english. *Corpora*, 7(2), 2012.

[3] A. Di Natale, M. Pellert, and D. Garcia. Colexification networks encode affective meaning. *Affective Science*, 2(2), 2021.

[4] E. Fast, B. Chen, and M. S. Bernstein. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016.

[5] W. N. Francis and H. Kucera. Brown corpus manual. *Letters to the Editor*, 5(2), 1979.

[6] A. François. Semantic maps and the typology of colexification. *From polysemy to semantic change: Towards a typology of lexical semantic associations*, (106), 2008.

[7] J.-M. List, S. J. Greenhill, C. Anderson, T. Mayer, T. Tresoldi, and R. Forkel. Clics2: An improved database of cross-linguistic colexifications assembling lexical data with the help of cross-linguistic data formats. *Linguistic Typology*, 22(2), 2018.

[8] H. Metzler, H. Baginski, T. Niederkrotenthaler, and D. Garcia. Detecting potentially harmful and protective suicide-related content on twitter: machine learning approach. *Journal of medical internet research*, 24(8), 2022.

[9] H. Metzler, B. Rimé, M. Pellert, T. Niederkrotenthaler, A. Di Natale, and D. Garcia. Collective emotions during the covid-19 outbreak. *Emotion*, 2022.

[10] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11), 1995.

[11] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015.

[12] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[13] M. Siegel and F. Bond. Odenet: Compiling a germanwordnet from other resources. In *Proceedings of the 11th Global Wordnet Conference*, 2021.

[14] M. Wolf, A. B. Horn, M. R. Mehl, S. Haug, J. W. Pennebaker, and H. Kordy. Computergestützte quantitative textanalyse: äquivalenz und robustheit der deutschen version des linguistic inquiry and word count. *Diagnostica*, 54(2), 2008.

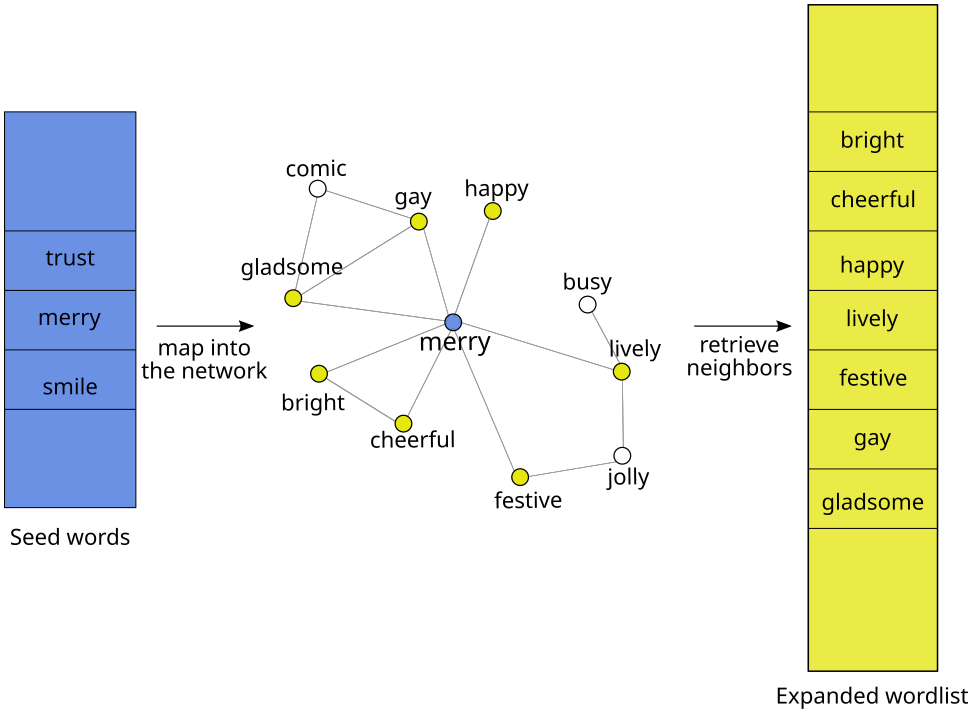

Figure 1: Representation of the word list expansion algorithm LEXpander: seed words (blue, on the left) are mapped onto a colexification network (center) and the neighbors of each word are retrieved to constitute the expanded word list (yellow, on the right). Here we represent the case of the word 'merry' in a word list for positive emotion.

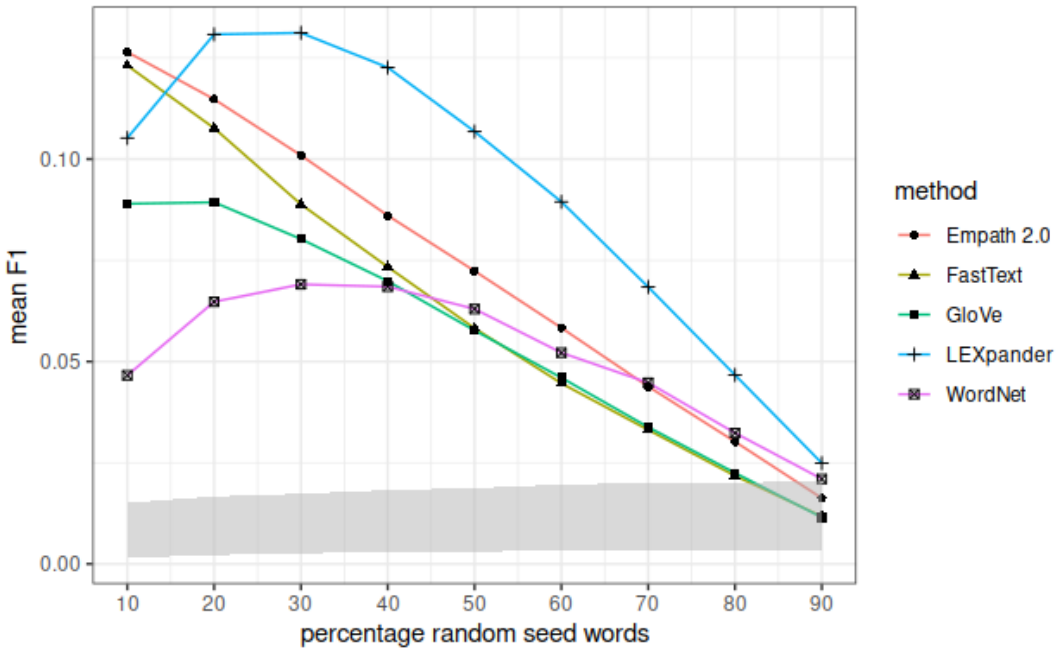

Figure 2: Mean F_1 scores of the expansion of English word lists as a function of the fraction of words randomly selected as seed words from a ground truth word list. The mean of the F_1 scores is computed over 73 different thematic word lists. LEXpander outperforms the other methods when selecting more than 10% words as seed words.