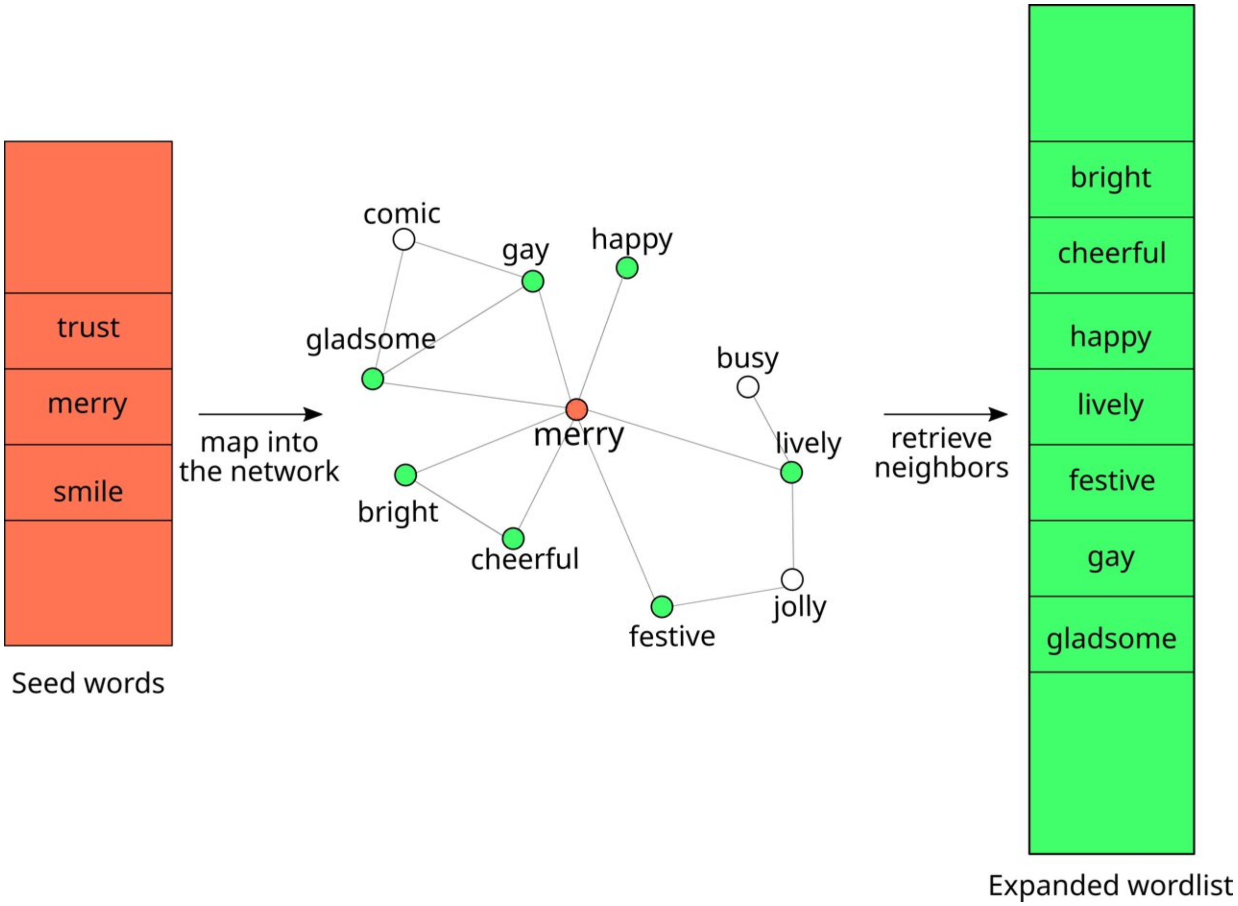
A new and better way to create word lists

March 13 2023



A short list of seed words (red, on the left) is expanded into a longer word list (green, on the right) by mapping the seed words onto a colexification network and retrieving the neighboring nodes. Credit: Complexity Science Hub

Word lists are the basis of so much research in so many fields. Researchers at the Complexity Science Hub have now developed an algorithm that can be applied to different languages and can expand word lists significantly better than others.

Many projects start with the creation of a word list, not only in companies when mind maps are created, but also in all areas of research. Imagine you want to find out on which days people are in a particularly good mood by analyzing Twitter postings. Just looking for the word "happy" wouldn't be enough.

Instead, you would have to use an algorithm that detects all tweets that indicate that someone is happy. "So the first step is to create a list of all the words that indicate just that. The whole research stands or falls on doing so," explains Anna Di Natale, a researcher at the Complexity Science Hub in Vienna. But how to come up with the most accurate, complete word lists possible?

A problem that concerns many

This widespread problem not only concerns opinion researchers who want to find out how politicians' statements are received by the public. Companies, too want to find out how their products are perceived through sentiment analysis.

To improve things, Di Natale has now developed a new method, called LEXpander, that outperforms previous algorithms in two different languages—German and English. Moreover, for the very first time ever, she has developed a way through which it is possible to compare different tools at all.

Improved performance

In comparison with four other algorithms for wordlist expansion

(WordNet, Empath 2.0, FastText and GloVe), LEXpander performed significantly better, especially in German. For example, the researchers found that LEXpander guesses 43% of words right when expanding an English word list for positive meaning. An existing popular model, FastText, in comparison, is right only 28% of the time.

Independence from the language itself

The reason is that this tool works language-independently. It is not based on one language, but on a so-called colexification network. This recognized linguistic concept resides on homonyms and polysemies, single words that have two or more distinct meanings. For example: the ancient Greek word φάρμακον (pharmacon) can mean medicine or poison. These are two different things, but thematically close. But there are others that don't suggest kinship—such as "bank" as a financial institution or the land alongside a river.

"If you collect them across many languages—and here we analyzed about 19 different languages—you can see connections between them," Di Natale says. The network is formed when these colexifications occur in several languages across different language families, creating connections.

This independence from the language itself allows LEXpander to achieve better results in different languages. "There are many methods developed for English. They work very well and quickly and everyone uses them. Trying to apply them to other languages works, but not as well as it might work if you had started developing a method for German or Italian," Di Natale explains.

Important for new topics like COVID

For many topics there are already good word lists. But for new topics—such as COVID—new ones must be created. Until now, they were usually created by hand during brainstorming among colleagues, and several tools were used to help. But until now there was no way to compare them.

Anna Di Natale and her team have now created this possibility and have also developed a new tool that performs better than the others. This can be an important cornerstone for many future research projects in various fields.

More information: Anna Di Natale et al, LEXpander: Applying colexification networks to automated lexicon expansion, *Behavior Research Methods* (2023). DOI: 10.3758/s13428-023-02063-y

Provided by Complexity Science Hub Vienna

Citation: A new and better way to create word lists (2023, March 13) retrieved 9 April 2024 from https://techxplore.com/news/2023-03-word.html