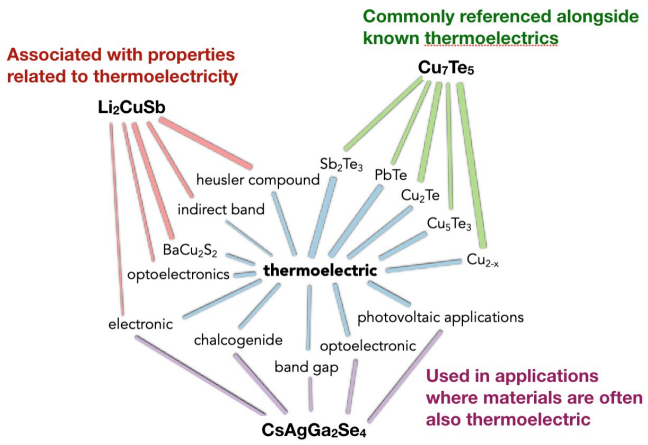
With little training, machine-learning algorithms can uncover hidden scientific knowledge

3 July 2019



Berkeley Lab researchers found that text mining of materials science abstracts could turn up novel thermoelectric materials. Credit: Berkeley Lab

Sure, computers can be used to play grandmaster-level chess (chess_computer), but can they make scientific discoveries? Researchers at the U.S. Department of Energy's Lawrence Berkeley National Laboratory (Berkeley Lab) have shown that an algorithm with no training in materials science can scan the text of millions of papers and uncover new scientific knowledge.

A team led by Anubhav Jain, a scientist in Berkeley Lab's Energy Storage & Distributed Resources Division, collected 3.3 million abstracts of published materials science papers and fed them into an algorithm called Word2vec. By analyzing relationships between words the algorithm was able to predict discoveries of new thermoelectric materials years in advance and suggest as-yet unknown materials as candidates for thermoelectric materials.

"Without telling it anything about materials science, it learned concepts like the periodic table and the crystal structure of metals," said Jain. "That hinted at the potential of the technique. But probably the most interesting thing we figured out is, you can use this algorithm to address gaps in materials research, things that people should study but haven't studied so far."

The findings were published July 3 in the journal *Nature*. The lead author of the study, "Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature," is Vahe Tshitoyan, a Berkeley Lab postdoctoral fellow now working at Google. Along with Jain, Berkeley Lab scientists Kristin Persson and Gerbrand Ceder helped lead the study.

"The paper establishes that text mining of scientific literature can uncover hidden knowledge, and that pure text-based extraction can establish basic scientific knowledge," said Ceder, who also has an appointment at UC Berkeley's Department of Materials Science and Engineering.

Tshitoyan said the project was motivated by the difficulty making sense of the overwhelming amount of published studies. "In every research field there's 100 years of past research literature, and every week dozens more studies come out," he said. "A researcher can access only fraction of that. We thought, can machine learning do something to make use of all this collective knowledge in an unsupervised manner—without needing guidance from human researchers?"

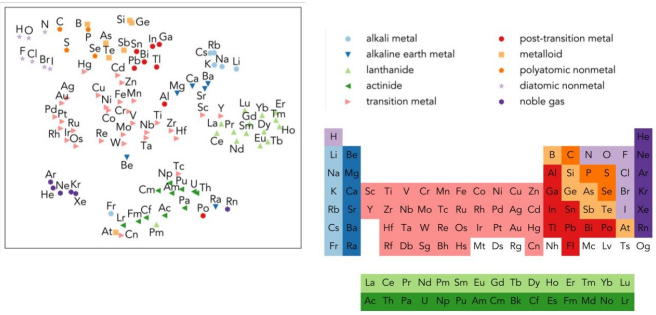
'King—queen + man = ?'

The team collected the 3.3 million abstracts from papers published in more than 1,000 journals between 1922 and 2018. Word2vec took each of

the approximately 500,000 distinct words in those abstracts and turned each into a 200-dimensional vector, or an array of 200 numbers.

"What's important is not each number, but using the numbers to see how words are related to one another," said Jain, who leads a group working on discovery and design of new materials for energy applications using a mix of theory, computation, and data mining. "For example you can subtract vectors using standard vector math. Other researchers have shown that if you train the algorithm on nonscientific text sources and take the vector that results from 'king minus queen,' you get the same result as 'man minus woman.' It figures out the relationship without you telling it anything."

Similarly, when trained on materials science text, the algorithm was able to learn the meaning of scientific terms and concepts such as the crystal structure of metals based simply on the positions of the words in the abstracts and their co-occurrence with other words. For example, just as it could solve the equation "king—queen + man," it could figure out that for the equation "ferromagnetic—NiFe + IrMn" the answer would be "antiferromagnetic."



Mendeleev's periodic table is on the right. Word2vec's representation of the elements, projected onto two dimensions, is on the left. Credit: Berkeley Lab

Word2vec was even able to learn the relationships between elements on the periodic table when the vector for each chemical element was projected onto two dimensions.

Predicting discoveries years in advance

So if Word2vec is so smart, could it predict novel thermoelectric materials? A good thermoelectric material can efficiently convert heat to electricity and is made of materials that are safe, abundant and easy to produce.

The Berkeley Lab team took the top thermoelectric candidates suggested by the algorithm, which ranked each compound by the similarity of its word vector to that of the word "thermoelectric." Then they ran calculations to verify the algorithm's predictions.

Of the top 10 predictions, they found all had computed power factors slightly higher than the average of known thermoelectrics; the top three candidates had power factors at above the 95th percentile of known thermoelectrics.

Next they tested if the algorithm could perform experiments "in the past" by giving it abstracts only up to, say, the year 2000. Again, of the top predictions, a significant number turned up in later studies—four times more than if materials had just been chosen at random. For example, three of the top five predictions trained using data up to the year 2008 have since been discovered and the remaining two contain rare or toxic elements.

The results were surprising. "I honestly didn't expect the algorithm to be so predictive of future results," Jain said. "I had thought maybe the algorithm could be descriptive of what people had done before but not come up with these different connections. I was pretty surprised when I saw not only the predictions but also the reasoning behind the predictions, things like the half-Heusler structure, which is a really hot crystal structure for thermoelectrics these days."

He added: "This study shows that if this algorithm were in place earlier, some materials could have conceivably been discovered years in advance." Along with the study the researchers are releasing the top 50 thermoelectric materials predicted by the algorithm. They'll also be releasing the word embeddings needed for people to make their own applications if they want to search on, say, a better

topological insulator material.

Up next, Jain said the team is working on a smarter, more powerful search engine, allowing researchers to search abstracts in a more useful way.

The study was funded by Toyota Research Institute. Other study co-authors are Berkeley Lab researchers John Dagdelen, Leigh Weston, Alexander Dunn, and Ziqin Rong, and UC Berkeley researcher Olga Kononova.

More information: Unsupervised word embeddings capture latent knowledge from materials science literature, *Nature* (2019). DOI: 10.1038/s41586-019-1335-8 , nature.com/articles/s41586-019-1335-8

Provided by Lawrence Berkeley National Laboratory