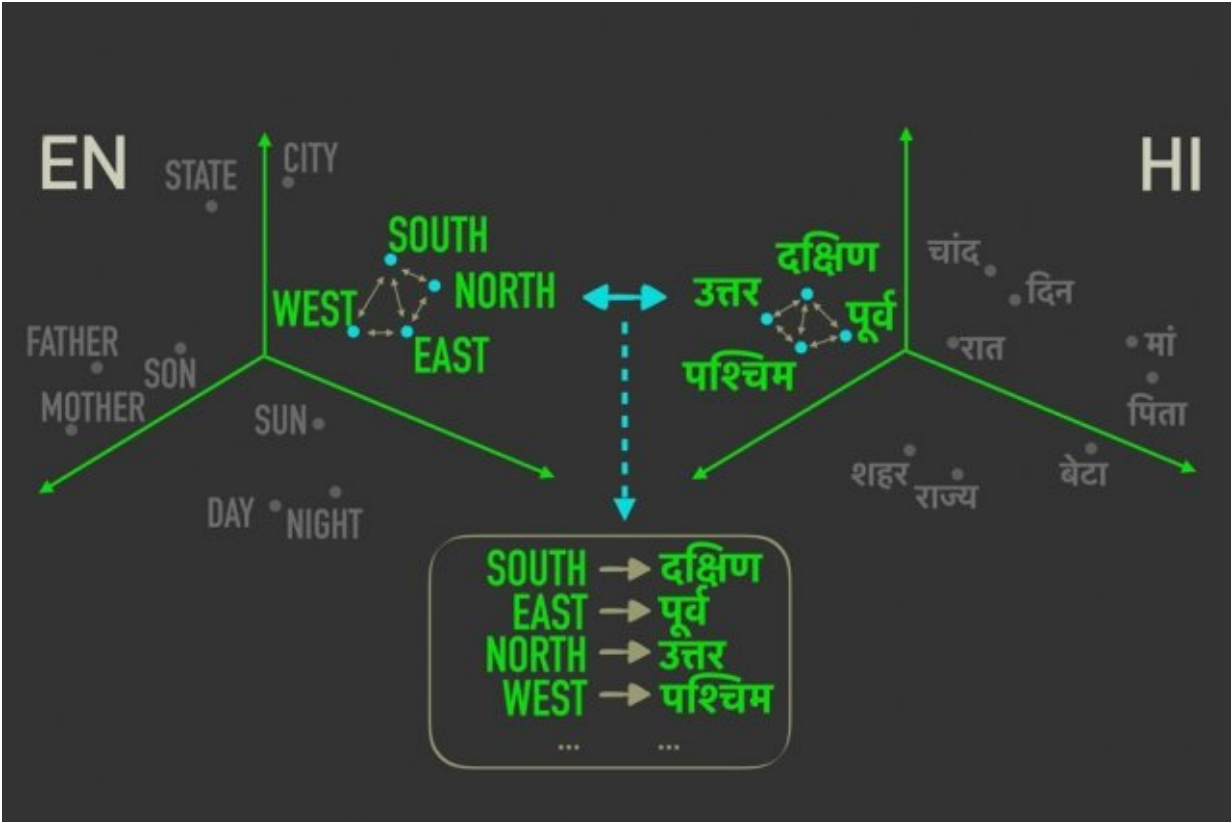


Model paves way for faster, more efficient translations of more languages

October 31 2018, by Rob Matheson



The new model measures distances between words with similar meanings in “word embeddings,” and then aligns the words in both embeddings that are most closely correlated by relative distances, meaning they’re most likely to be direct translations of one another. Credit: Massachusetts Institute of Technology

MIT researchers have developed a novel "unsupervised" language

translation model—meaning it runs without the need for human annotations and guidance—that could lead to faster, more efficient computer-based translations of far more languages.

Translation systems from Google, Facebook, and Amazon require training models to look for patterns in millions of documents—such as legal and political documents, or news articles—that have been translated into various languages by humans. Given new words in one language, they can then find the matching words and phrases in the other language.

But this translational data is time consuming and difficult to gather, and simply may not exist for many of the 7,000 languages spoken worldwide. Recently, researchers have been developing "monolingual" models that make translations between texts in two languages, but without direct translational information between the two.

In a paper being presented this week at the Conference on Empirical Methods in Natural Language Processing, researchers from MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) describe a model that runs faster and more efficiently than these monolingual models.

The model leverages a metric in statistics, called Gromov-Wasserstein distance, that essentially measures distances between points in one computational space and matches them to similarly distanced points in another space. They apply that technique to "word embeddings" of two languages, which are words represented as vectors—basically, arrays of numbers—with words of similar meanings clustered closer together. In doing so, the model quickly aligns the words, or vectors, in both embeddings that are most closely correlated by relative distances, meaning they're likely to be direct translations.

In experiments, the researchers' model performed as accurately as state-

of-the-art monolingual models—and sometimes more accurately—but much more quickly and using only a fraction of the computation power.

"The model sees the words in the two languages as sets of vectors, and maps [those vectors] from one set to the other by essentially preserving relationships," says the paper's co-author Tommi Jaakkola, a CSAIL researcher and the Thomas Siebel Professor in the Department of Electrical Engineering and Computer Science and the Institute for Data, Systems, and Society. "The approach could help translate low-resource languages or dialects, so long as they come with enough monolingual content."

The model represents a step toward one of the major goals of machine translation, which is fully unsupervised word alignment, says first author David Alvarez-Melis, a CSAIL Ph.D. student: "If you don't have any data that matches two languages ... you can map two languages and, using these distance measurements, align them."

Relationships matter most

Aligning word embeddings for unsupervised machine translation isn't a new concept. Recent work trains neural networks to match vectors directly in word embeddings, or matrices, from two languages together. But these methods require a lot of tweaking during training to get the alignments exactly right, which is inefficient and time consuming.

Measuring and matching vectors based on relational distances, on the other hand, is a far more efficient method that doesn't require much fine-tuning. No matter where word vectors fall in a given matrix, the relationship between the words, meaning their distances, will remain the same. For instance, the vector for "father" may fall in completely different areas in two matrices. But vectors for "father" and "mother" will most likely always be close together.

"Those distances are invariant," Alvarez-Melis says. "By looking at distance, and not the absolute positions of vectors, then you can skip the alignment and go directly to matching the correspondences between vectors."

That's where Gromov-Wasserstein comes in handy. The technique has been used in computer science for, say, helping align image pixels in graphic design. But the metric seemed "tailor made" for word alignment, Alvarez-Melis says: "If there are points, or words, that are close together in one space, Gromov-Wasserstein is automatically going to try to find the corresponding cluster of points in the other space."

For training and testing, the researchers used a dataset of publicly available word embeddings, called FASTTEXT, with 110 language pairs. In these embeddings, and others, words that appear more and more frequently in similar contexts have closely matching vectors. "Mother" and "father" will usually be close together but both farther away from, say, "house."

Providing a "soft translation"

The model notes vectors that are closely related yet different from the others, and assigns a probability that similarly distanced vectors in the other embedding will correspond. It's kind of like a "soft translation," Alvarez-Melis says, "because instead of just returning a single word translation, it tells you 'this vector, or word, has a strong correspondence with this word, or words, in the other language.'"

An example would be in the months of the year, which appear closely together in many languages. The model will see a cluster of 12 vectors that are clustered in one embedding and a remarkably similar cluster in the other embedding. "The model doesn't know these are months," Alvarez-Melis says. "It just knows there is a cluster of 12 points that

aligns with a cluster of 12 points in the other [language](#), but they're different to the rest of the [words](#), so they probably go together well. By finding these correspondences for each word, it then aligns the whole space simultaneously."

The researchers hope the work serves as a "feasibility check," Jaakkola says, to apply Gromov-Wasserstein method to machine-translation systems to run faster, more efficiently, and gain access to many more languages.

Additionally, a possible perk of the model is that it automatically produces a value that can be interpreted as quantifying, on a numerical scale, the similarity between languages. This may be useful for linguistics studies, the researchers say. The [model](#) calculates how distant all vectors are from one another in two embeddings, which depends on sentence structure and other factors. If vectors are all really close, they'll score closer to 0, and the farther apart they are, the higher the score. Similar Romance languages such as French and Italian, for instance, score close to 1, while classic Chinese scores between 6 and 9 with other major languages.

"This gives you a nice, simple number for how similar languages are ... and can be used to draw insights about the relationships between languages," Alvarez-Melis says.

More information: David Alvarez-Melis, Tommi S. Jaakkola. Gromov-Wasserstein Alignment of Word Embedding Spaces. arXiv:1809.00013 [cs.CL]. arxiv.org/abs/1809.00013

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology

Citation: Model paves way for faster, more efficient translations of more languages (2018, October 31) retrieved 17 April 2024 from <https://techxplore.com/news/2018-10-paves-faster-efficient-languages.html>

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.