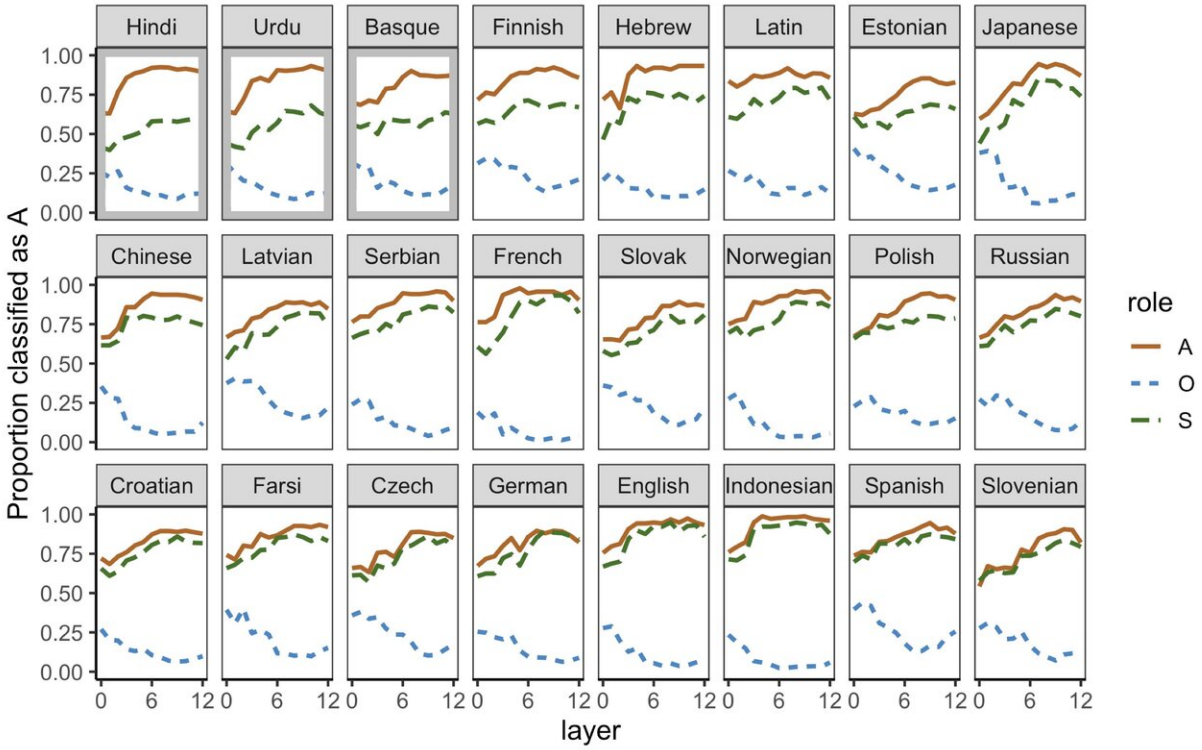
Researchers examine how multilingual BERT models encode grammatical features

February 22 2021, by Ingrid Fadelli



For each layer (x-axis), the proportion of the time that the researchers predict that a noun is a subject(A), separated by grammatical role. In higher layers, intransitive subjects (S) are mostly classified as subjects (A). When the source language is Basque (ergative) or Hindi or Urdu (split-ergative) S is less likely to pattern with A. The figure is ordered by how close the S line is to A, and ergative and split-ergative languages are highlighted with a gray box. Credit: Papadimitriou et al.

Over the past few decades, researchers have developed deep neural network-based models that can complete a broad range of tasks. Some of these techniques are specifically designed to process and generate coherent texts in multiple languages, translate texts, answer questions about a text and create summaries of news articles or other online content.

Deep learning systems with linguistic capabilities are already widely available, for instance, in the form of applications for real-time translation, text analysis tools and virtual assistants such as Siri, Alexa, Bixby, Google Assistant and Cortana. Some of these systems use a specific deep-learning model released by Google called Multilingual BERT (mBERT). This model was trained on approximately 100 languages simultaneously. This allows it to complete a variety of language tasks, for instance, translating content from one language to another.

Users can interact with systems based on mBERT in a multitude of languages, ranging from English, Spanish and French to Basque and Indonesian. While the mBERT model has been found to perform well on many language tasks, how it encodes language-related information and makes its predictions is still poorly understood.

Researchers at Stanford University, University of California, Irvine and University of California, Santa Barbara have recently carried out a study aimed at better understanding how mBERT-based techniques work and how they encode grammatical features. Their paper, whose lead author is Isabel Papadimitriou, a graduate student in computer science at Stanford, is set to be presented at the computational linguistics conference EACL. The paper offers valuable insight into the underpinnings of these commonly used models and how they analyze language when completing various tasks.

"Models like Multilingual BERT are very powerful, but, unlike pre-trained deep learning models, it's not obvious what information they actually contain, even to their creators," Kyle Mahowald, a linguist at University of California, Santa Barbara and one of the senior researchers who supervised the study, told TechXplore. "That's because the models are trained, not programmed; thus, they learn parameters through a training process on enormous amounts of data."

Essentially, the mBERT model represents texts as a series of vectors, each of which consists of thousands of numbers. Every vector corresponds to a word, while the relationships between words are encoded as geometrical relations in high-dimensional space.

"Because these models do so well in dealing with human language, we know that these vectors of numbers must represent linguistic knowledge," Mahowald said. "But how do they encode this information, and is it anything like the way that knowledge is represented in the human brain? Our work is part of this effort to understand the ways in which deep neural models of language represent and use linguistic information."

Understanding how mBERT models encode language is not so different from trying to understand how humans process it. Therefore, the team behind the recent study was composed of both computer scientists and linguists. Their main objective was to determine whether mBERT vector models actually contain information about some of the deeper aspects of human language and its structure. More specifically, they wanted to determine whether these models autonomously uncovered the generalizations that several decades of research in linguistics have identified as particularly useful for language analysis.

"This is a particularly exciting time to be studying computational linguistics," said Richard Futrell, a language scientist at University of

California, Irvine and another of the project's senior advisors. "For years, linguists have talked about ideas like 'semantic space," thinking of the meanings of words and phrases as points in some space, but it was all somewhat vague and impressionistic. Now, these theories have been made completely precise: We actually have a model where the meaning of a word is a point in space, and that model really does behave in a way that suggests it understands (some of) human language."

To process human languages, mBERT models and other deep-learning-based frameworks for language analysis may have actually re-discovered theories devised by linguistics researchers after deeply analyzing human languages. Alternatively, they might base their predictions on entirely new language theories or rules. Mahowald and his colleagues wanted to explore both these possibilities further, as understanding how these computational techniques encode language could have important implications for research in both computer science and linguistics.

"Understanding how these models work (i.e., what information they have learned and how they use it) is not just scientifically fascinating, it's also practically critical if we want to develop AI systems that we can use and trust," Futrell said. "If we don't know what a language model knows, then we can't trust that it will do the right thing (i.e., that its translations will be correct, that its summaries will be accurate) and we also can't trust that it hasn't learned undesirable things like race or gender bias."

As mBERT models are generally trained on datasets compiled by humans, they might pick up some of the mistakes that humans commonly make when tackling language-related problems. The study carried out by the multi-disciplinary team could play a part in uncovering some of these mistakes and other errors that AI tools make when analyzing language. Firstly, the researchers set out to investigate how mBERT models represent the difference between subjects and objects across different languages (i.e., who is doing what and to whom/what).

"When a sentence is entered into mBERT, each word gets a vector representation," Mahowald said. "We built a new model (much smaller than mBERT) which we then ask: if we give you a word vector from mBERT, can you tell us if it's a subject or an object? That is, here is the representation of the word 'dog." Can you tell us if that usage of 'dog' was the subject of a sentence, as in "The dog chased the cat?" or the object of a sentence, as in "The cat chased the dog?'"

One might assume that subject and object relations are delineated in all languages and that they are represented in similar ways. However, there are actually huge differences in what constitutes a subject and object in different languages. Papadimitriou and her colleagues tried to leverage these differences to gain a better understanding of how mBERT models process sentences.

"If you speak a language like English, it might seem obvious that the word 'dog' in "The dog chased the cat' is playing a similar role to the word 'dog' in "The dog ran,'" Papadimitriou said. "In the first case, the verb has an object ('cat'), and in the second case it has no object; but in both cases, 'dog' is the subject, the agent, the doer and in the first sentence 'cat' is the object—the thing that is having something done to it. However, that is not the case in all languages."

English and most languages spoken in Europe have a structure known as nominative alignment, which clearly characterizes subjects and objects in sentences. On the other hand, some languages, including Basque, Hindi and Georgian, use a structure known as ergative alignment. In ergative alignment, the subject in a sentence with no object (e.g., the word 'dog' in the sentence 'the dog ran') is treated more like an object, in the sense that it follows the grammatical structure used for objects.

"The main goal of our work was to test whether Multilingual BERT understands this idea of alignment, ergative or nominative,"

Papadimitriou said. "In other words, we asked: Does Multilingual BERT understand, on a deep level, (1) what constitutes the agent and the patient of a verb, and (2) how different languages carve up that space into subjects and objects? It turns out that mBERT, which is trained on about 100 languages at once, is aware of these distinctions in linguistically interesting ways."

The findings offer new and interesting insights into how mBERT models and perhaps other computational models for language analysis represent grammatical information. Interestingly, the model examined by the researchers, which was based on mBERT vector representations, was also found to make consistent errors that could be aligned with those made by humans who are processing language.

"Across languages, our model was more likely to incorrectly call a subject an object when that subject was an inanimate noun, meaning a noun which is not a human or an animal," Papadimitriou said. "This is because most doers in sentences tend to be animate nouns: humans or animals. In fact, some linguists think that subjecthood is actually on a spectrum. Subjects that are human are more 'subject-y' than subjects that are animals, and subjects that are animals are more subject-y than subjects that are neither humans nor animals, and this is exactly what our model seems to find in mBERT."

Overall, the study suggests that mBERT models identify subject and objects in sentences and represent the relationship between the two in ways that are aligned with existing linguistics literature. In the future, this important finding could help computer scientists to gain a better understanding of how deep-learning techniques designed to process human language work, helping them to improve their performance further.

"We now hope to continue exploring the ways in which deep neural

models of language represent linguistic categories, like subject and object, in their continuous vector spaces," Mahowald said. "Specifically, we think that work in linguistics, which seeks to characterize roles like subject and object not as discrete categories but as a set of features, could inform the way that we think of these models and what they are doing."

More information: Deep subjecthood: higher-order grammatical features in multilingual BERT. arXIv: 2101.11043 [cs.CL]. arxiv.org/abs/2101.11043