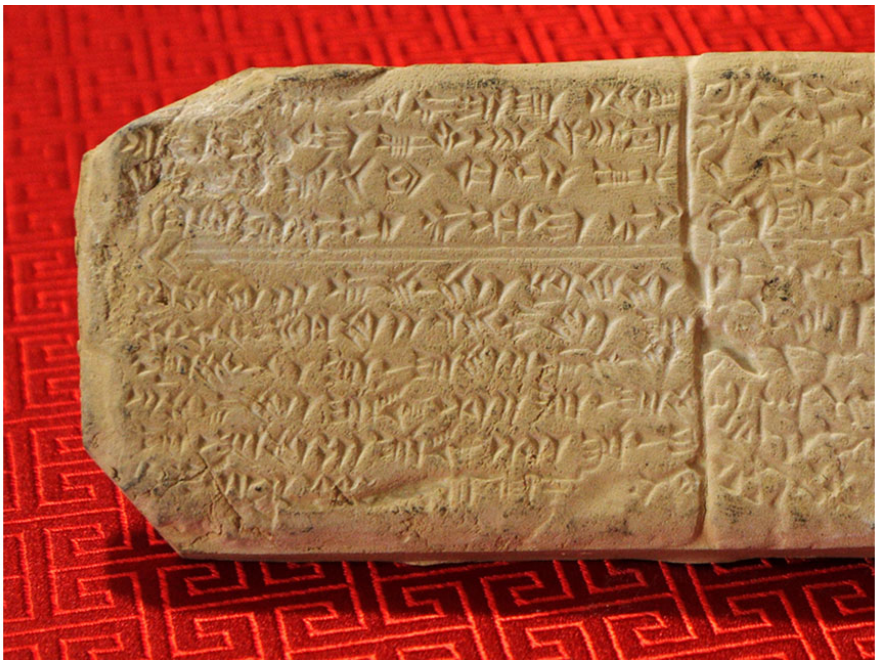
Translating lost languages using machine learning ARTDAILY – 25 October, 2020



In developing a system to help decipher lost languages, MIT researchers studied the language of Ugaritic, which is related to Hebrew and has previously been analyzed and deciphered by linguists. Image courtesy: S.R.K. Branavan.

CAMBRIDGE, MASS.- Recent research suggests that most languages that have ever existed are no longer spoken. Dozens of these dead languages are also considered to be lost, or "undeciphered" — that is, we don't know enough about their grammar, vocabulary, or syntax to be able to actually understand their texts.

Lost languages are more than a mere academic curiosity; without them, we miss an entire body of knowledge about the people who spoke them. Unfortunately, most of them have such minimal records that scientists can't decipher them by using machine-translation algorithms like Google Translate. Some don't have a well-researched "relative" language to be compared to, and often lack traditional dividers like white space and punctuation.
(To illustrate, imaginetryingtodecipheraforeignlanguagewrittenlikethis.)

However, researchers at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) recently made a major development in this area: a new system that has been shown to be able to automatically decipher a lost language, without needing advanced knowledge of its relation to other languages. They also showed that their system can itself determine relationships between languages, and they used it to corroborate recent scholarship suggesting that the language of Iberian is not actually related to Basque.

The team's ultimate goal is for the system to be able to decipher lost languages that have eluded linguists for decades, using just a few thousand words.

Spearheaded by MIT Professor Regina Barzilay, the system relies on several principles grounded in insights f fact that languages generally only evolve in certain predictable ways. For instance, while a given language rarely adds or deletes an entire sound, certain sound substitutions are likely to occur. A word with a "p" in the parent language may change into a "b" in the descendant language, but changing to a "k" is less likely

due to the significant pronunciation gap.

By incorporating these and other linguistic constraints, Barzilay and MIT PhD student Jiaming Luo developed a decipherment algorithm that can handle the vast space of possible transformations and the scarcity of a guiding signal in the input. The algorithm learns to embed language sounds into a multidimensional space where differences in pronunciation are reflected in the distance between corresponding vectors. This design enables them to capture pertinent patterns of language change and express them as computational constraints. The resulting model can segment words in an ancient language and map them to counterparts in a related language.
The project builds on a paper Barzilay and Luo wrote last year that deciphered the dead languages of Ugaritic and Linear B, the latter of which had previously taken decades for humans to decode. However, a key difference with that project was that the team knew that these languages were related to early forms of Hebrew and Greek, respectively.

With the new system, the relationship between languages is inferred by the algorithm. This question is one of the biggest challenges in decipherment. In the case of Linear B, it took several decades to discover the correct known descendant. For Iberian, the scholars still cannot agree on the related language: Some argue for Basque, while others refute this hypothesis and claim that Iberian doesn't relate to any known language.

The proposed algorithm can assess the proximity between two languages; in fact, when tested on known languages, it can even accurately identify language families. The team applied their algorithm to Iberian considering Basque, as well as less-likely candidates from Romance, Germanic, Turkic, and Uralic families. Iberian than other languages, they were still too different to be considered related.

In future work, the team hopes to expand their work beyond the act of connecting texts to related words in a known language — an approach referred to as "cognate-based decipherment." This paradigm assumes that example of Iberian shows that this is not always the case. The team's new approach would involve identifying semantic meaning of the words, even if they don't know how to read them.

"For instance, we may identify all the references to people or locations in the document which can then be further investigated in light of the known historical evidence," says Barzilay. "These methods of 'entity recognition' are commonly used in various text processing applications today and are highly accurate, but the key research question is whether the task is feasible without any training data in the ancient language."

The project was supported, in part, by the Intelligence Advanced Research Projects Activity (IARPA).