PHYS.ORG

Grammatical habits in written English reveal linguistic features of non-native speakers' languages

July 22 2014, by Larry Hardesty

Computer scientists at MIT and Israel's Technion have discovered an unexpected source of information about the world's languages: the habits of native speakers of those languages when writing in English.

The work could enable computers chewing through relatively accessible documents to approximate data that might take trained linguists months in the field to collect. But that data could in turn lead to better computational tools.

"These [linguistic] features that our system is learning are of course, on one hand, of nice theoretical interest for linguists," says Boris Katz, a principal research scientist at MIT's Computer Science and Artificial Intelligence Laboratory and one of the leaders of the new work. "But on the other, they're beginning to be used more and more often in applications. Everybody's very interested in building computational tools for world languages, but in order to build them, you need these features. So we may be able to do much more than just learn linguistic features. … These features could be extremely valuable for creating better parsers, better speech-recognizers, better natural-language translators, and so forth."

In fact, Katz explains, the researchers' theoretical discovery resulted from their work on a practical application: About a year ago, Katz proposed to one of his students, Yevgeni Berzak, that he try to write an algorithm that could automatically determine the native language of someone writing in English. The hope was to develop grammar-correcting software that could be tailored to a user's specific linguistic background.

Family resemblance

With help from Katz and from Roi Reichart, an engineering professor at the Technion who was a postdoc at MIT, Berzak built a system that combed through more than 1,000 English-language essays written by native speakers of 14 different languages. First, it analyzed the parts of speech of the words in every sentence of every essay and the relationships between them. Then it looked for patterns in those relationships that correlated with the writers' native languages.

Like most machine-learning classification algorithms, Berzak's assigned probabilities to its inferences. It might conclude, for instance, that a particular essay had a 51 percent chance of having been written by a native Russian speaker, a 33 percent chance of having been written by a native Polish speaker, and only a 16 percent chance of having been written by a native Japanese speaker.

In analyzing the results of their experiments, Berzak, Katz, and Reichart noticed a remarkable thing: The algorithm's probability estimates provided a quantitative measure of how closely related any two languages were; Russian speakers' syntactic patterns, for instance, were more similar to those of Polish speakers than to those of Japanese speakers.

When they used that measure to create a family tree of the 14 languages in their data set, it was almost identical to a family tree generated from data amassed by linguists. The nine languages that are in the Indo-European family, for instance, were clearly distinct from the five that aren't, and the Romance languages and the Slavic languages were more similar to each other than they were to the other Indo-European languages.

What's your type?

"The striking thing about this tree is that our system inferred it without having seen a single word in any of these languages," Berzak says. "We essentially get the similarity structure for free. Now we can take it one step further and use this tree to predict typological features of a language for which we have no linguistic knowledge."

By "typological features," Berzak means the types of syntactic patterns that linguists use to characterize languages—things like the typical order of subject, object, and verb; how negations are formed; or whether nouns take articles. A widely used online linguistic database called the World Atlas of Language Structures (WALS) identifies nearly 200 such features and includes data on more than 2,000 languages.

But, Berzak says, for some of those languages, WALS includes only a handful of typological features; the others just haven't been determined yet. Even widely studied European languages may have dozens of missing entries in the WALS database. At the time of his study, Berzak points out, only 14 percent of the entries in WALS had been filled in.

The new system could help fill in the gaps. In work presented last month at the Conference on Computational Natural Language Learning, Berzak, Katz, and Reichart ran a series of experiments that examined each of the 14 languages of the essays they'd analyzed, trying to predict its typological features from those of the other 13 languages, based solely on the similarity scores produced by the system. On average, those predictions were about 72 percent accurate.

Branching out

The 14 languages of the researchers' initial experiments were the ones for which an adequate number of essays—an average of 88 each—were publicly available. But Katz is confident that given enough training data, the system would perform just as well on other languages. Berzak points out that the African language Tswana, which has only five entries in WALS, nonetheless has 6 million speakers worldwide. It shouldn't be too difficult, Berzak argues, to track down more English-language essays by native Tswana speakers.

Provided by Massachusetts Institute of Technology

Citation: Grammatical habits in written English reveal linguistic features of non-native speakers' languages (2014, July 22) retrieved 9 April 2024 from https://phys.org/news/2014-07-grammatical-habits-written-english-reveal.html