Princeton International

int'l

Fall 2021

also inside:

Lost in Translation

Summer in Seoul

On Top of the World

Reimagining International

Lost in Translation

A course teaches the fundamentals of machine learning, computer science and linguistics

By Samir Patel

When Srinivas Bangalore, visiting lecturer in Princeton Institute for International and Regional Studies (PIIRS) and the Program in Translation and Intercultural Communication, began teaching "Introduction to Machine Translation" in 2008, it seemed like an advanced concept for undergraduate students. The use of software to process, translate or generate language appeared in the voice recognition systems behind customer service calls, but Google Translate — now a mainstay of language learning and cross-cultural communication — was still a few years away from wide release. "The whole idea that some device is going to understand ideas expressed in one language and convey them in another language was exotic," Bangalore says.

Bangalore has now taught the course for more than a decade, during which time the exotic went mainstream. Today, we think nothing of pointing our phones at a sign in an unfamiliar language and seeing it magically translated on screen, or calling out a question to an empty room and expecting an answer. Machine translation makes all of that possible.

This popular course — cross-listed in the departments of computer science and linguistics, and the Program in Translation and Intercultural Communication — provides students the basics of natural language processing, or

converting spoken or written language into a computer-friendly representation. In the simplest version: Words and meanings are the data, syntax and sentence structure are the rules. From this foundation, students are set up to process these diverse representations using linguistic rules, statistical models or neural networks to translate into another language or even create a new and unexpected output. The students develop their own projects, and for many that's a favorite aspect of the class. As machine translation has advanced, so has their ambition and creativity. "Over the years, the course has sort of meandered in various different directions as the field itself has evolved," Bangalore says. "It's not necessarily one language to another."

Dora Demszky '17, now working on a Ph.D. in computational linguistics at Stanford University, took the course in 2015. Always fascinated by linguistics — she brought the International Linguistics Olympiad to her native Hungary as a high school student — she savored this first exposure to applying computer science to language. "I like to do… hands-on things," she says, "and this definitely enabled me to do that."

Demszky and fellow students Sarah Herrmann '17 and Quinlan Shen '17 tried to use machine translation to create poetry — specifically the deceptively complex form of haiku. But it wasn't as simple as feeding a computer some basic rules. They had to teach it about syllables, parts of speech and how words are related to one another (i.e., bloom, petal, flower). The intriguing results illustrate what the team wrote in their final presentation: "Semantics is hard."

infinite oceans
each seen throbbing in wardens
amorous dark capes

But the class and project helped set Demszky on her academic path, and she's now using these ideas to study and improve educational materials and approaches. "It was very influential for me," she says. For example, she's used natural language processing to analyze the content related to gender, race and ethnicity in the American history textbooks used in Texas. Another project broke down spoken language to understand discussion strategies that teachers can use to help math students excel.

Joomy Korkut, a native of Turkey who is currently a fourth-year Ph.D. student in computer science at Princeton, took Bangalore's class in spring 2019. Korkut had long been interested in language, and the language barrier he had encountered when he moved to the United States at age 19 only increased that interest.

As a precocious high school student — much like Demszky — Korkut had considered a career as a historian of late Ottoman Turkey. The project in Bangalore's class provided a chance to bring several of his academic interests together to attempt to translate Ottoman Turkish to Modern Turkish.

"Ottoman Turkish is essentially the same language as Turkish, with a different alphabet and very different spelling rules," Korkut says. "In 1928, Turkey switched from the Arabic alphabet to the Latin alphabet, with additional letters for the extra sounds. The Arabic alphabet was poor in vowels; a reader had to infer the vowel sounds when reading a text. Turkish depends on these vowel sounds a lot, so often there are ambiguities. Solving these ambiguities and reconstructing the suffixes with the right vowel harmony rules was the primary challenge."

After the class, Korkut put a draft of his final paper on his website, and others in the field have reached out about it. "This was a hobby for me but it makes me jubilant to see it taken seriously," he says.

Over the years Bangalore has overseen many other fascinating projects, from sentiment analysis of how people feel about portraits in museums to the rhythm of classical Latin poetry. "I ask them to be open with their thought process, and coming from industry, I see that as a valuable sort of skill to develop for research," he says. "Some students really thrive and really branch out in areas that I wouldn't even have anticipated."

Bangalore, professionally, has come to a similar place with respect to human intervention. Today, he leads a team that works on voice recognition systems — for AT&T, Apple and Nike, among others — that know when to ask a human for help. But the callers will never know a human stepped in, since they've never left the automated voice system. "It's like the Wizard of Oz kind of idea," he says. "It feels like the technology is getting it right, but in reality, there are humans helping the technology."

Even as AI becomes more prevalent in the field, humans still need to be involved in the machine translation process in some way, Bangalore believes. A course like his gives them the foundation to be the wizards behind the curtain.