CONVERTING HINDI TO ENGLISH:
A GRADUATE STUDENT IS MAKING HEADLINES
by KIMBERLY BURTNYK

Thanks to a project worthy of Hollywood spy movies, the secret world of US military intelligence and espionage changed dramatically this past June.

On the first of the month, the Defense Advance Research Project Agency launched the Surprise Language Project. The task: develop an automated system that translates electronic Hindi documents into English, thus aiding intelligence gathering.

Sound simple? Think again. Though similar to Urdu and other Middle Eastern languages, Hindi is poorly understood in the West. Further, Hindi websites employ several different methods to encode the language's characters making automated translation a significant challenge. But still, in just 29 days, the 12 participating institutions around the world completed the task successfully.

One important player in the collaboration was 24-year-old University of Southern California Doctoral student Liang Zhou. The Chinese native developed "Headline Generation" software, a critical component as headlines allow for the rapid evaluation of a document's strategic importance.

Using artificial intelligence, grammar and statistics, the software condenses pages of translated text into a few descriptive words. It does this by counting the number of times key words are used in a document, noting where those words appeared in the text and judging their importance to the document's meaning. After considering all the combinations, the program spits out a "headline" in just mere seconds.

Before working on this project, Zhou had experience with a similar branch of computer science called Summarization. As the name implies, in
summarization, a document is scanned and a content summary is returned. Meanwhile, headline generation reduces a summary to a mere handful of words making it a more difficult task.

"A lot of people are doing summarization, but not a whole lot of people are doing headline generation," said Zhou, who moved to the United States nine years ago. "Language is hard to imitate, and grammar is hard to learn," especially for computers, she explained.

The tools created by the Surprise Language Project may also have commercial applications, from cataloging on-line library resources to academic research. The technology could revolutionize electronic and foreign language communication. "If my system works well [for Hindi], we can say it can work for all different types of languages," Zhou explained, making automatic electronic translation from one language to another routine.

Zhou is just one of three women a part of the 11-member USC team involved in this research, but she said she doesn't think about being in the minority. "We all work well together. We're interested in what we do, and the work," Zhou said.