A rose is a rózsa is a ??: Image-search tool speaks hundreds of languages

12 September 2007

A search for the Zulu word "ifriji" using PanImages that selects matches in Japanese and Russian generates 472,000 images.

From the fall of the Tower of Babel to the Esperanto global language movement, many humans have dreamed of sharing a common tongue. Despite the Internet's promise of global communication, language barriers remain. Even pictures on the Web get lost in translation.

"Images are universal, but image search is not," said Oren Etzioni, a professor of computer science and engineering at the University of Washington. "A person who types his or her search in English won't find images tagged in Chinese, and a Dutch person won't find images tagged in English. We've created a collaborative tool that solves this problem."

A new multilingual search tool developed at the UW's Turing Center makes the universal appeal of pictures available to all. PanImages, presented today at the Machine Translation Summit in Copenhagen, Denmark, allows people to search for images on the Web using hundreds of languages.

Search engines such as Google look for images by detecting the search term in captions and other nearby text. But since the process looks for a string of letters, the results are limited to the seeker's mother tongue.

The new tool is named PanImages, from the Greek prefix, "pan," meaning whole or all-inclusive. It automatically translates the search term into about 300 other languages, suggests a few that might work and then displays images from Google and the online photo database Flickr.

PanImages promises to help people who speak languages that have a small Web presence. Imagine you are a Zulu speaker looking for a picture of a refrigerator, Etzioni said. You type the Zulu word for refrigerator ("ifriji") into an image search and get two results. The same search using PanImages generates 472,000 hits. In a test of so-called minor languages, PanImages was able to find 57 times more results, on average, than a Google image search.

"We want to serve the vast number of people who don't speak one of the major languages," Etzioni said. "As the Internet becomes more widely available outside of the major industrialized nations, it becomes increasingly important to serve people who don't speak English, French or Chinese."

Even people who speak these more common languages can benefit by switching electronic tongues. Words that have more than one meaning inevitably produce unwanted results. For instance, typing the word "spring" in an English-language image search generates diverse images: grassy meadows, metal coils and pictures from the town of Silver Spring, Md. If you want images of a metal spring, you might use PanImages and search for the more precise French word "ressort." If you want a picture of a rectangular bar and don't want businesses where patrons drink alcohol, you might try the Russian word "??????." Experiments showed that, for common languages, PanImages

nearly doubles the number of correct images on the first 15 pages of results.

PanImages' powerful brains were created by scanning more than 350 machine-readable online dictionaries. Some of these were "wiktionaries," online multilingual dictionaries written by volunteers. The PanImages software scans these dictionaries and uses an algorithm to check the accuracy of the results. It then assembles the results in a matrix that allows translation in combinations that may never have been attempted—for instance, from Gujarati to Lithuanian.

"It's an unprecedented lexical resource. The most distinguishing element is its ability to scale to such a broad set of languages," Etzioni said. "Our goal is to ultimately cover all the languages people are interested in."

Free online translation services used by Yahoo! and Google incorporate just one or two dozen common languages. In the United States, research on machine translation tends to focus on languages with military importance, such as Arabic and Chinese, Etzioni said. PanImages had 50 languages earlier this year and by June it incorporated 100 languages. It now includes some 300 languages, 2.5 million words and millions of individual translations.

PanImages also lets people instantly add new words or translations.

Future work on PanImages will scour more online dictionaries to expand the number of words and languages it can handle. Researchers also hope to translate the words used in tagging sites, such as del.icio.us, where visitors use single-word labels to describe the page's content.

"Our goal is to promote pan-lingual translation," said Etzioni. "With this first step, we've created a service we hope will be a handy tool."

To try PanImages, go to www.panimages.org .

Source: University of Washington

APA citation: A rose is a rózsa is a ??: Image-search tool speaks hundreds of languages (2007, September 12) retrieved 20 June 2019 from https://phys.org/news/2007-09-rose-rzsa-image-search-tool-hundreds.html