The Language Technology Ecosystem

Richard L. Sites, Google, Inc.

Abstract

As information dissemination moves from traditional speech and written media to computers and the Web, language technology is a key to improved education, improved social contact, and improved commerce. Populations that cannot access information in their own languages are left out. Conversely, Internet companies that invest in the language ecosystem enable development and also increase their user population by millions of people.

The ecosystem is more than just input keyboards and fonts for various scripts; it includes user interfaces in hundreds of languages, the ability to identify, parse, format and edit meaningful text in those languages, especially names, dates, currency, addresses, and telephone numbers. It includes translation between languages, and it includes open tools to let thousands of creative people add to the ecosystem. Speech-to-text gives access to information for people who cannot type or are illiterate or use mobile phones or cannot hear the sound tracks in videos. Text-to-speech gives access to people who are blind, illiterate, or use mobile phones. Optical character recognition allows access to scanned literature from all over the world.

We will survey Google's efforts in language technology, beyond the 47 languages that cover 99% of the current Web text, then give a brief explanation of one of those -- detecting over 200 different languages on Web pages: Afar, Bangla, Coptic, Dzongha, ... through Yoruba, Zulu.

Biographical sketch

Dick Sites is a Senior Staff Engineer at Google, where he has worked for 7 years. He previously worked at Adobe Systems, Digital Equipment Corporation, Hewlett-Packard, Burroughs, and IBM. His accomplishments include co-architecting the DEC Alpha computers, advancing the art of binary translation for computer executables, and building various computer performance monitoring and tracing tools at the above companies. He also taught Computer Science for four years at University of California/San Diego. Most recently he has been working on Unicode text processing and on CPU and network performance analysis. Dr. Sites holds a PhD degree in Computer Science from Stanford and a BS degree in Mathematics from MIT. He also attended the Master's program in Computer Science at University of North Carolina. He holds 34 patents and is a member of the US National Academy of Engineering.