EXCERPT

ESTHER DYSON'S MONTHLY REPORT Release 1.0®

VOLUME 22, NO. 3 | 21 MARCH 2004 | www.edventure.com

The Big Picture: In Focus

The following is excerpted from the March 2004 issue of Release 1.0.

Language Weaver: We know what you mean

BY CHRISTINA KOUKKOS

While the language for much international business-to-business communication is English, companies also need to be able to communicate with customers and employees in their local languages. Language Weaver provides a platform for that communication with a software system that applies statistical analysis to translation.

Language Weaver takes a novel approach to machine translation, but one with some background in the research community, says CEO C. Bryce Benjamin: "People try to teach [translation] software grammatical rules and dictionary definitions. Concept-ually, that sounds easy. But when you think of how words are used in different contexts and parts of speech, it gets more complicated." Thus the bizarre and sometimes hilarious results from rules-based translators.

Founders Kevin Knight and Daniel Marcu, researchers at the University of Southern California, decided to forget the rules. Instead, they built a system that *learns* to translate by example rather than by rules. The system runs a statistical analysis on large collections of the customer's already-translated documents – translated archives, standard glossaries, etc. – and establishes a large set of probable word and word-phrase correlations across the two languages. It runs another analysis on documents in the *target* language in order to determine how words are usually strung together and to produce natural-sounding translations.

The resulting translation mappings are used by Language Weaver to translate new documents for that customer. And because the mappings are specific to the documents of that customer, they mimic style and idiom. "As it turns out, you end up with a higher quality of output," says Benjamin. In one test case from Arabic to English, LW translations were two to three times higher in quality versus competitors (measured in terms of Bleu score, a statistical measurement that compares the accuracy of a machine

{ *continued on page 2* }

edventure
The conversation starts here.

translation with the human translation of the same document). In a test from French to English, the quality was 50 percent higher.

Of course, we suggest, the customer would need to have a large enough set of translations. "You're right," says Benjamin. "It only works if the system has the right terminology in the first place. But most [multinational] corporations already have large stockpiles of already-translated documents."

Target customers include corporate communications departments, facilitators of cross-market e-mail and chat, translation companies and any organization that wants all or part of a worldwide news feed – to assess competitors, gauge political risk and understand a foreign customer base. Language Weaver's current customers are "mostly in the government sector," i.e. the intelligence community. The company is running pilots with a number of corporate clients, including high-tech and translation-services companies.

Pricing for server licenses depends on the language pair. A perpetual license for a Latin-based European language to English costs $18,000 per CPU for one-way (e.g. French to English) and $25,000 for two-way capability. Lower-density or more difficult languages, such as Chinese, cost about $50,000 one-way/$75,000 two-way for a 3-year subscription. LW "teaches" the software the customer's custom terminology, with semiannual "retraining." It plans to build a feedback loop that will allow customers to retrain their software on their own.

Current supported languages include two-way French to Arabic, and one-way (to English) from Chinese, Hindi and Somali. The company plans to have English-to-Chinese and bidirectional Spanish capability by the end of the second quarter. But, says Benjamin, as long as the customer has enough data (and a budget), LW can create a translation system for any two languages within a month or two.

Release 1.0® (ISSN 1047-935X) is published monthly except for a combined July/August issue by EDventure Holdings Inc., 104 Fifth Avenue, New York, NY 10011-6987; 1 (212) 924-8800; fax, 1 (212) 924-0240; www.edventure.com. It covers the worlds of information technology and the Internet, including wireless communications, security, business models, online services, tracking systems, identity management and other unpredictable topics. . .and the policy issues they raise.

EDITOR: Esther Dyson (edyson@edventure.com)

PUBLISHER: Daphne Kis (daphne@edventure.com)

MANAGING EDITOR: Christina Koukkos (christina@edventure.com)

CONTRIBUTING WRITERS: Dan Gillmor (dan@gillmor.com), John Hagel (john@johnhagel.com), Steven Johnson (stevenberlinjohnson@earthlink.net), Clay Shirky (clay@shirky.com), Jeff Ubois (jeff@ubois.com), Dave Weinberger (self@evident.com)

CIRCULATION MANAGER: Brodie Crawford (brodie@edventure.com)

SYSTEMS MANAGER: Geoff Clarke (geoff@edventure.com)

CONSULTING EDITOR: Bill Kutik (bill@kutik.com)