Issue No. 32 (vol.11, no. 2)

Spring 2003

ISSN 0965-5476

Published three times a year

MT News International

Newsletter of the International Association for Machine Translation

Data-Driven MT Grows Up

LG

Following last year's AMTA conference (October, 2002, Tiburon, California), it became clear that a number of data-driven MT companies were on the verge of commercializing new machine translation systems. In parallel with this, established developers in the U.S., Japan, and Taiwan were starting over with new hybrid engines slated to replace their earlier symbolic engines. Early this year, we began contacting the upstarts to find out what we could expect from them and when. We asked each developer questions, including "what is your core approach to machine translation" and "what sets it apart from other machine translation systems". Two surprising pieces of information emerged:

1: There are quite a number of upstarts! We found 6 startups with the primary mission of commercializing data-driven MT systems (Aixplain, AKS, Huajian, Language Weaver, Linear-B, Verbalis). A number of other companies with broader offerings (or projected offerings) have entered the ring with data-driven MT systems (Meaningful Machines, Microsoft, Morphologic, Symbionautics). And of course a number of the existing MT developers are actively introducing data-driven methods into existing rule-based products (Systran), or working on completely new products with more of a data-driven core (Bowne Global Solutions (the Barcelona system), and BehaviorTrans in Taiwan).

2: The designs of the systems challenge the distinction established in 1992, between the empiricist and rationalist approaches to MT. The rationalists, we understood, built systems with hand-crafted rules. The empiricists used automated learning techniques to extract information from linguistic data, primarily existing translations. But in the current group, there are example-based systems where the examples are all hand-crafted from existing translations (Verbalis, Oki, Mor-phoLogic, IBM Japan). The research community may object to the use of the term "example-based" in this case, but we have used the descriptions provided by developers. In addition, many of the systems claim to be hybrids of various sorts (Aixplain, Microsoft, Oki, Sehda, Huajian and Linear-B).

We have not attempted to evaluate developers' claims regarding the type of system being developed. Although the MT research community seems to have a clear understanding of the necessary and sufficient conditions for considering a system "statistical" or "example-based," those outside the ACL/AAAI research community have adopted and applied the labels more loosely.

Because of space limitations, the results of the survey will be presented as a two or three part series. Readers are invited to suggest systems that should be covered in this series.

Company: Aixplain

Aixplain's core approach to MT is Statistical and symbolic MT (a complex hybrid system). Its target market is Business clients, with a "B-2-B" approach.

What sets your system apart from other MT systems?

-very high flexibility
-very low Time-To-Market
-high quality of translations within specific domains
-adaptive HLT system

Aixplain AG
Monheimsallee 22
52074 Aachen, Germany
Tel: +49.241.18927-0
www.aixplain.de

Company: IBM Japan

IBM Japan's core approach to machine translation is Pattern-based. Usually empirical approaches include example-based approach and statistical approach. In this sense, IBM has not yet released any empirical-based commercial MT system. They report their target market or application to be Web browsing.

But, in a broad sense, the pattern-based approach which IBM Japan developed can be considered as one of the empirical approaches. IBM Japan's "pattern-based approach" uses a large number of translation patterns, each of which is a pair of source lexicalized CFG rule and a target lexicalized CFG rule. It, given a source sentence, analyzes it by using CFG parsing method with source side rules of translation patterns, and generates a target structure by synchronized derivation mechanism using translation patterns whose source rules are used for parsing. We collected about more than 10,000 translation patterns. For technical details, please see the following paper: Takeda, K., "Pattern-Based Context-Free Grammars for Machine Translation," Proc. of 34th ACL, pp. 144--151, June 1996.

What sets your system apart from other MT systems? The pattern-based approach allows users to add phrasal-level translation knowledge as well as word-level knowledge. For instance, a user can register a Japanese translation for a phrase "hit a big shot."

Dr. Hideo Watanabe
Group Leader of Intelligent Information Human Interaction Technology, S&S, Tokyo Research Laboratory
IBM Japan
Tel: +81-46-215-4561

Company: Fluent Machines

The Fluent Machines approach to machine

Data-Driven MT Companies at a Glance

Company: Aixplain
Founded: 2001
Inventor: Hassan Sawaf, Hermann Ney and Franz Josef Och
CEO/President: Hassan Sawaf (CEO), **h.sawaf@aixplain.de**
Customer/Investor contact: Chafik Moalem, **c.moalem@aixplain.de**
Company size: 22 people
First Product deployment: October 2001

Company: IBM Japan
When did IBM Japan start developing empirically-based MT systems? 1990
When did IBM Japan start its NLP R&D group?
-Early 80s (MT research), 1970s (general NLP research)
Who is the inventor of the empirical technology that has been commercialized?
-Koichi Takeda
Leader of the MT development group/NLP R&D Group at IBM Japan:
-Dr. Hideo Watanabe, **hiwat@jp.ibm.com**
Sales Information (in Japanese): **www-6.ibm.com/jp/software/internet/king/**
First sale or deployment: "Internet King of Translation" released by IBM Japan in 1996.

Company: Fluent Machines
Founded: Fluent Machines, a subsidiary of Meaningful Machines, was founded in 2001.
Inventor: Eli Abir, Inventor and Chief Architect
CEO and Chairman: Steve Klein
President: David Miller
Customer/Investor contact: Michael Steinbaum, COO, **mike@meaningfulmachines.com**
Company size: 11 employees
First product sale/deployment: fourth quarter of 2003 (projected)

Company: Linear-B
Founded: 2002
Inventors: Colin Bannard and Chris Callison-Burch
Customer/Investor Contact: Colin Bannard, **colin@linearb.co.uk**
First product deployment/sale: end of 2003 (projected)

More DDMT systems next issue!

translation incorporates several purely empirical natural language learning processes that are new to the fields of machine translation and NLP.

Fluent Machines expects their first applications to target major European languages, followed by major Asian languages.

What sets your system apart from other MT systems? Eli Abir, inventor and architect of Fluent Machines' technology, had important insights into the way different languages represent the same universal ideas. Mr. Abir leveraged those insights into four novel and unique processes: (i) the first process uses previously translated parallel text to automatically build large cross-language databases of basic word-string combinations, (ii) the second process leverages known word-string translations between language pairs to discern translations between different language pairs, (iii) the third process determines semantic equivalents in both the source and target languages, and (iv) the fourth process links together target language word-string translations created by any of the other processes to produce translated text.

Furthermore, Fluent Machines leverages a patent-pending technology that truly understands wide-ranging natural language through a process that (1) uncovers language patterns found in written text (any language and across all domains) and (2) focuses on concepts regardless of the words used to express them. This enables the Fluent Machines system to understand single and multi-word concepts, which gives tremendous power and flexibility to the MT system.

1450 Broadway, 40 th Floor
New York, New York 10018
Tel: 212-716-0070
info@fluentmachines.com

Company: Linear-B

Linear B's core approach to MT is a combination of machine-learning and example-based methods. Their target market/application is robust translation of web content and email.

What sets this system apart from other MT systems? Our system has been designed from the outset to deal with, and indeed utilize, the diverse nature of web data and the specific needs of web users in a more satisfactory fashion than the one-size-fits-all approach of most systems. It employs innovative techniques for dealing with the non-stationary nature of language, and crucially for acquiring knowledge from previously neglected kinds of data, meaning that it is able to offer the advantages of robustness and low cost that are associated with data-driven systems for a larger number of language pairs than have previously been possible.

Linear B Ltd.
Edinburgh Technology Transfer Centre
King's Buildings
Edinburgh EH9 3JL, Scotland
Tel: +44 131 472 4816
Linearb.co.uk
Continued in next issue...