Issue No. 31 (vol.11, no. 1)

Winter 2002

ISSN 0965-5476

Published three times a year

MT News International

Newsletter of the International Association for Machine Translation

Special Feature: Speaking of MT

After 40 years of domination by rule-based design, empirical or data-driven MT systems are incubating in various parts of the world and a number will enter the market as products in the next year. Such systems derive at least some part of their capability by learning directly from text. Fluent Machines is one of these which has gotten some high profile press coverage—including Scientific American and Red Herring in 2002. MTNI's American regional editor David Clements spoke to Fluent Machines' Mike Steinbaum to get the real story. In the next issue, we'll look at the whole crop of upstart MT systems. –ed

Fluent Machines

by David Clements

Fluent Machines, a New York-based company founded in 2001, is developing what company literature calls "breakthrough technology" using elements of both example-based MT (EBMT) and Statistical MT. The company, founded by Israeli immigrant Eli Abir, has created a technology centering on Mr. Abir's theory of the "DNA of language," resulting in two patent-pending processes: the "Automated Cross Language Database Builder," and the "N-gram Connector." The database builder forms the core of the system's learning component. The n-gram connector is responsible for generation of natural language output. In addition to these two established components, a third technology, AIMT (for Artificial Intelligence) is being developed which will reduce the reliance on fully bilingual text sources for training and may even eliminate a need for them entirely. Fluent Machines claims its advanced processes will provide "a complete and comprehensive solution for achieving human-quality MT." While the two patent-pending processes have been tested operationally using English-French, English-Spanish, and English-Hebrew, Fluent Machines does not yet have any deliverable products, and testing has been done primarily on components. The company does not have translation samples to share or examples of translation tests. The Automated Cross-Language Database Builder is based on insights into natural language by Mr. Abir, and enables a computer to generate a database of translation pairs of n-grams without regard to the size of the n-gram. These translation pairs are automatically learned from previously translated written text. The system, through statistical analysis, looks at a complete document or combination of documents (e.g., books, articles, manuals, journals) in two or more languages and begins to distill the translation of all the component parts of the texts.

Although the Database Builder currently operates with "modest processing power and limited access to cross language texts," Fluent Machines has begun to build cross-language databases that it anticipates, within a year, will be the largest in existence. Fluent Machines is also developing (but has not yet tested) a method that is not dependent on parallel text to glean n-gram translations. The method is called "AIMT" because it focuses on the actual meaning of an n-gram. This method uses large monolingual source and target corpora in combination with existing translation methods or word-for-word dictionaries. The method requires more processing power than the database builder, but will enable the system to learn broader coverage of the language pairs being translated because it doesn't rely on parallel text, which is much less available.

The second of Fluent Machines' patent-pending processes is the N-gram Connector, which connects contiguous n-grams in a target language with (the company claims) "human-quality accuracy." N-grams will be connected only if the system knows with certainty that the connection will yield an accurate new word-string translation.

The company makes three interesting points about the N-gram Connector:

Each translation added to the database increases the number of word-strings that can be accurately translated in the future by a large multiple because all naturally connecting n-grams in the database can combine with the new database entry. This allows Fluent Machines to translate word-string combinations that the system has not yet encountered. The system can automatically build many new, longer word-strings each time a single new entry is added to the cross-language database.

The system's ability to lock word-strings together only at points where they organically fit is analogous to the process by which a strand of DNA replicates itself. It is this locking mechanism that allows the reproduction of an infinite number of variations from a finite set of building blocks.

Human editors can focus their review on the portion of the translated text highlighted by the system as "not approved" because the system can identify potentially incorrect portions of its translation.

The Database Builder is responsible for completeness, while the N-gram Connector is responsible for human-quality accuracy. Until a cross-language database is complete (or reaches critical mass), the system will continue to yield accurate, but not complete translations. That is, the N-gram Connector will produce translations only for the portions it is confident of. It may not produce a translation for the entire text.

At a time when many commercially available MT systems aim for mere "gisting," Fluent Machines maintains the goal of achieving human-quality MT, although company literature is careful to distinguish human-quality MT from "perfect translations." Fluent Machines has received the endorsement of Dr. Jaime Carbonell of Carnegie Mellon University. According to a 2002 report issued by Dr. Carbonell and distributed by the company, "The EliMT Method is clearly the most promising and theoretically important MT development in the past several years (and probably since the advent of MT itself). It is the one recent development with the greatest possibility of making a major advance in practical large scale diffusion of MT technology." EliMT is a term coined by Dr. Carbonell to refer to Eli Abir, the inventor of the Fluent Machines technology. Dr. Carbonell now serves on the Fluent Machines Board of Directors.

According to Mike Steinbaum, the company's COO, Fluent Machines currently has 13 employees. So-called "EliMT" is "memory and processing power intensive," and remaining tasks, besides parallel corpus acquisition, include combining the two processes discussed in this article, as well as increasing computational speed and efficiency. In addition, the company will continue the development of its AIMT methods. The Fluent Machines system, he says, "knows what it knows," and will produce "incomplete translations, but not wrong translations." The company expects that as it refines its algorithms and builds larger databases, its system will lessen the incompleteness of translations and approach ever closer to human quality, extending the range of the system beyond the common Euro-centric and English-centric pairs.

Fluent Machines is a subsidiary of Meaningful Machines, Inc., a technology development company. The sister company of Fluent Machines is Internet Driver, which has developed a patent-pending technology that "provides the Internet's missing piece of multi-lingual infrastructure, allowing users to access the entire existing Internet (domain names, email addresses and subsites) using any language's character set."

For more information on Meaningful Machines and its subsidiaries, visit the Website at www.meaningfulmachines.com, or contact Mike Steinbaum, COO, at mike@meaningfulmachines.com.