Bulletin

SYSTRAN and the Reinvention of MT

Analysts: Mary Flanagan and Steve McClure

IDC Opinion

Can SYSTRAN's new technology extend the perception of machine translation (MT) as a tool for more than simple gisting?

Yes. With its newly architected system, its persuasive hold on the multilingual Web-browsing market, and its successful incursions into the customer service area, SYSTRAN is poised to achieve a level of success that has not yet been realized by any other MT company and perhaps to redefine the field of uses for automatic translation solutions. SYSTRAN will need to move cautiously, given MT's history, and avoid the obvious pitfalls of unrealistic expectations and inappropriate applications for the technology. Having demonstrated the ability to navigate these challenges successfully thus far, there is every reason to expect SYSTRAN will succeed.

Marketing Evolution, Technology Revolution

The conventional wisdom among marketers of machine translation (MT) products is that MT is a technology best suited to creating rapid, rough draft translations. This application, known as “gisting” utilizes translation software for multilingual information scanning in situations where the importance of quick availability of information exceeds the need for precise translation. Gisting is a concept that has been used both to educate users to the potential of MT and to evangelize one of its broadest uses. Even discussing MT for high-quality translation has become taboo among MT marketers who fear associating themselves with fly-by-night vendors of word lookup systems who claim to translate among hundreds of language pairs with high accuracy.

But just as the conventional wisdom has become established, it is now being challenged by the most conventional of MT companies, SYSTRAN. The idea that MT can be used for high-quality translation is not new. In fact, the earliest researchers of MT at Georgetown University and IBM believed that the translation problem could be solved fully and quickly. MT would replace human translators and deliver fully automatic, high-quality translations at a fraction of the time and cost. Their naivete and underestimation of the enormous complexity of modeling human language is understandable. But early results disappointed U.S. government funders, leading to a near cessation of MT research for almost 15 years.

SYSTRAN’s claims reflect the benefits of three decades of experience and a major technology rearchitecture.

SYSTRAN’s claim is less far-reaching than the claims of MT’s pioneers. More importantly, it reflects the benefits of three decades of experience and a major technology rearchitecture. Nonetheless, it is a bold and perhaps risky move to challenge the prevailing assumption about MT. It is all the more surprising coming from a traditionalist company like SYSTRAN, known for its caution and conservatism. The claim is not an empty one; SYSTRAN can back it up with metrics and benchmarking of translation results. Users are engaged in the deployment process to establish acceptance at each phase of implementation before going forward.

The SYSTRAN Story

SYSTRAN has the longest history of any MT developer in the world. The company was founded by Dr. Peter Toma, one of the scientists who worked on the original U.S. government MT projects at Georgetown University. SYSTRAN was founded in 1968, and has an R&D investment measured in thousands of person-years. The resulting system is the Goliath of the MT world. SYSTRAN's dictionaries are enormous and diverse, it offers more language pairs than any other MT system (35) and has broad coverage of grammatical patterns across many text styles.

SYSTRAN has managed to navigate the treacherous waters of Internet applications more successfully than any other MT provider.

With such a long history, it is not surprising that SYSTRAN is viewed both within and outside of the MT community as a stable, respectable technology but not as a cutting edge innovator. However, SYSTRAN's enduring image does not reflect its current reality. The system has undergone a comprehensive redesign to bring its code and linguistic resources into step with current computing standards and to add many new cutting-edge linguistic resources. Unlike many of its competitors, SYSTRAN has managed to navigate the treacherous waters of Internet applications more successfully than any other MT provider. SYSTRAN has focused on multilingual Web browsing, and more recently on Web-based customer support. The deployments at Alta Vista, Google, Autodesk and others have been successful, while other Internet translation applications have failed to generate adequate audience and revenue. The demise of e-lingo, Wholetree, Lernout & Hauspie, and Logos are cases in point. More importantly, the success of these applications was an important step in building general acceptance of MT as a useful tool for rapid gist translation.

SYSTRAN's rebirth may be a unique event in the MT industry. Mature MT systems are tremendously complex with deeply intradependent code. As a result, modernizing these systems is often prohibitively costly. Quick fixes tend to have limited impact, and regression is a constant danger because of the system's complexity and dependencies within the code. The corporate culture of an established company can also serve as a barrier to significant change. Entrenched patterns of problem solving and utilization of staff tend to preserve the status quo. In sum, transforming an old MT system can seem as elusive a goal as finding the fountain of youth. Yet, SYSTRAN seems to have done just that with its newly released and redesigned system.

From Old World to New World

Building the Team

The redesign plan was a shared vision of SYSTRAN CEO Dimitris Sabatakakis and Pierre-Yves Foucou, the company's CTO. An entire development team was assembled in Paris, 6,000 miles away from SYSTRAN's original development site in La Jolla, California. Although much of SYSTRAN's development historically was conducted in the La Jolla (and more recently San Diego), the true command and control center for the company is now in Paris where Sabatakakis resides.

Staff functions in San Diego have been gradually shifted toward engineering aspects of system development, though a number of long-time SYSTRAN linguists still work at the San Diego site. Linguistic development for the new generation system, however, is being carried out primarily in Paris. The team of software engineers and computational linguists is led by Foucou, a computational linguist and former professor of linguistics and computer science at the University of Paris. The development effort is part of the European Union's Matchpad project, a European Union–funded effort to extend availability of MT systems for language pairs involving Polish and Hungarian.

Foucou's charter was to reengineer SYSTRAN toward several ambitious goals: to improve maintenance and scalability in the face of ever-growing linguistic resources, to improve efficiency of access to linguistic resources, to increase modularity, and to support emerging exchange standards.

Along with the growing MT opportunity, components of SYSTRAN's technology may be reusable for applications such as content management and multilingual indexing.

Sabatakakis also recognized that the growing presence of MT on the Internet would create requirements for many additional language pairs, and that they would be required quickly. The typical development cycles of 2–5 years for a new language pair wouldn't suffice for fast-paced Internet deployments. Making the development of new languages a quicker process would be essential to keep SYSTRAN competitive. And along with the growing MT opportunity, components of SYSTRAN's technology might be reusable for other natural language applications such as content management and multilingual indexing *if* SYSTRAN's resources were modular enough to extract and adapt. The resulting system entails changes to all of these areas and is without doubt the most ambitious redesign of an MT system ever undertaken.

The New SYSTRAN Technology

SYSTRAN's redesign is comprehensive. It dramatically alters the system's basic structure and characteristics and introduces many new components as well. The extensive list of linguistic resources available within the SYSTRAN product allows vast customization potential. Users can customize their SYSTRAN application according to their text quality and type, their computing environment, required languages, and numerous other variables. Customization is essential to producing high-quality machine translation results. Simply used out of the box, MT software tends to have limited success because its knowledge bases are not equipped with the terminology and information needed for the subject area.

Modularity

The redesign has modularized the code so that the output of each module is independent and can be used for external purposes as well as for input to the subsequent module.

Modularity contributes to ease of maintenance and reusability of resources and was thus an essential goal for SYSTRAN, whose linguistic resources are extensive. The redesign has modularized the code so that the output of each module is independent and can be used for external purposes as well as for input to the subsequent module.

SYSTRAN's former approach was monolithic, with a unique program for each language pair. The rearchitecture of components has created independent modules with more complex relationships. Modules exchange information to build the most relevant context. Another advantage of this approach is that modules of different generations can coexist. The newer modules have been designed to access SYSTRAN's knowledge bases, while the older modules apply very refined grammatical phenomena.

Foucou adds:

> SYSTRAN has more resources than common hardware can support. MT design needs to integrate this constraint to optimize embedded knowledge to meet linguistics requirements. We compile resources into finite-state data structures to maximize efficiency. Future MT technology will have to aggregate multiple components into a multiagent architecture that is able to compute parallel results and find the most relevant translation among dozens or hundreds of alternatives.

Finite State Technology

Perhaps the most distinctive change is the effective use of finite state technology at many levels of the system. The hallmark of finite state technology is efficiency, and, thus, the approach is used in applications such as indexing documents for search and retrieval and spelling checkers because it can provide a constant access time to any record in a database regardless of the size of the resource. Finite state methods allow SYSTRAN to maintain its high performance despite its enormous linguistic and lexical resources. SYSTRAN has embedded finite state technology at a number of levels of its system, including morphology, conceptual description, and transfer dictionary encoding.

Dictionary Access

SYSTRAN's exhaustive dictionaries are one of its most valuable assets.

SYSTRAN's exhaustive dictionaries are one of its most valuable assets. But managing million-entry knowledge bases poses challenges for scaling, access, and management of duplication. An extremely robust database is needed to accommodate its dictionaries, which average 200,000 entries for European languages, while allowing for continued growth. Compounding this challenge is the fact that the system must be able to support thousands of lookups per second as the translation program iteratively analyzes word forms and attempts to locate the root form in the dictionary.

Duplication is also a problem because SYSTRAN, like most MT systems, uses bilingual dictionaries. For example, the English-to-French system will have a very similar but not identical dictionary to the English-to-Spanish system. Most dictionary entries have multiple targets in different languages. The dictionaries are compiled at runtime to minimize the demand on hardware resources. With 35 language pairs, the amount of duplication across dictionaries is enormous.

Having duplication can also create consistency problems when the same English term is coded with different grammatical tags in two different bilingual dictionaries. SYSTRAN's rearchitecture attacks the scaling, access, and duplication problems by introducing monolingual dictionaries. The monolingual dictionaries are maintained in addition to bilingual dictionaries and contain both simple and compound entries. The monolingual dictionary factorizes complex entries to a single access point via the headword. For example, “pilote de course automobile” (race car driver) is indexed on “pilote.” At the subsequent level of description, only the additional information is encoded, reducing redundancy. The second-level dictionary is also generated at runtime from the first structure, improving efficiency.

Declarativity

Perhaps the most sweeping change to SYSTRAN’s code is its conversion to a declarative system. Declarative programming is an innovation of the past decade and has largely replaced procedural programming, in which each minute step of the programming task is explicitly specified in the code. In a declarative system, the developer specifies the intended results of a programming task, typically using a graphical formalism that serves as a shorthand for describing the linguistic phenomenon. The details of how the task is conducted are implicit — they are defined by the system using the tools and resources that are made available to it. Nonetheless, the declarative approach cannot solve all linguistic processing problems. Some complex or idiosyncratic structures still require special processing.

Implicit Transfer

Transfer is a stage in the machine translation process in which the results of the analysis of the source language sentence are reordered according to a set of rules that embodies the structural relationship between the source and target language syntax. Transfer is a step carried out in so-called “transfer MT systems” such as SYSTRAN (other MT methods exist, but are beyond the scope of this bulletin). SYSTRAN has introduced implicit transfer methods into the redesigned system to simplify and speed the transfer process. The motivation for this is that some types of local expressions and verbal constructs have unique and complex internal structures and, thus, are hard to describe using transfer rules. Implicit transfer establishes parallel source and target descriptions for these phenomena, then aligns and generates a correct syntactic structure in the target based on the target description.

Exchange Format

Until recently, MT vendors had little interest in standardization. Their systems each utilized unique methods of description for language phenomena. This information was carefully protected and treated as proprietary trade secrets. As natural language applications become more numerous and diverse, the need for standardization is becoming evident, both as a way to facilitate exchange among natural language applications and as part of the gradual mainstreaming of MT within the software world.

SYSTRAN is developing a filter that provides full support of XML exchange format. The task is not a simple one because it requires defactorization of graphed entries and explication of implicit transfer patterns. Preserving the organization of the information is one of the biggest challenges. Nevertheless, SYSTRAN needs to forge ahead with this effort to enable its resources to be exported and to permit importing of external resources, such as glossaries, into the system.

NLP Components

Although many of SYSTRAN's natural language processing components are shared by all of its language systems, their modularity allows the user to create a customized environment best suited to the translation need, audience, and text type. The components include the following:

- Document filter for separating text and formatting codes
- Encoding and character set converter for interpreting common character encoding formats
- Language recognizer for identifying the source language of the text
- Preprocessor for identifying document types, such as chat, email, or structured text
- Spell checker to perform spelling correction for misspelled items (Misspellings would otherwise go unrecognized by the system, sometimes resulting in adverse impact to translation.)
- Sentence segmenter for dividing the text into sentences
- Word delimiter to identify word boundaries for languages where blank spaces are not inserted between words
- Lemmatizer, a tool for identifying and creating the variant forms of a word (e.g., develop, developing, and developed)
- Part-of-speech tagger for identifying the grammatical function of each word in the sentence (e.g., noun, verb, and adjective)
- Text synthesizer for production of the correct word forms in the target language
- Semantic domain recognizer to identify the subject area of the text so that appropriate knowledge bases can be employed

SYSTRAN's resources also include a tool set designed originally for quality assurance tasks. The tools are useful in the deployment process to assess quality levels and determine the characteristics of the source texts. The tool set includes a concordancer, terminology extractor, and tools for measuring the quality and consistency of translations and of custom resources, such as dictionaries.

Risks and Possibilities

To remain a leader, SYSTRAN will need to preserve both its strong output quality and high-speed performance after the redesign.

To remain a leader, SYSTRAN will need to preserve both its strong output quality and high-speed performance after the redesign. Although this has always been true, it will be all the more critical now that SYSTRAN's marketing message is beginning to target higher-quality applications.

Early results suggest that the system operates more efficiently than before and it produces greater throughput without loss of output quality. Preserving these advantages is important because SYSTRAN is facing emerging competition. Example-based and hybrid systems are in the works at a number of universities in the United States, Europe, and Asia, including USC's Information Sciences Institute and New York University.

These systems can have shorter development times than traditional approaches to MT because translation rules are generated automatically based on analysis of bilingual corpora. Although their development timetables are shorter, they are by no means short, and building an example-based MT system has its own unique set of pitfalls, such as the difficulty of finding aligned bilingual corpora from which to draw the examples.

The development timetable for new systems is shorter. It is by no means a trivial task, and SYSTRAN's position as an entrenched leader will not be easy to upset.

Creating a robust, high-performance, fault-tolerant translation system to compete with SYSTRAN also requires substantial engineering resources. So, while the development timetable for new systems is shorter, it is by no means a trivial task, and SYSTRAN's position as an entrenched leader will not be easy to upset.

Another risk for SYSTRAN is the continued bad press that MT receives when linguistically naive users deploy MT applications. The level of education about MT's capabilities is very low in the United States, although it is somewhat better in Europe and Asia. The typical American user has little acquaintance with translation software, other languages, or the challenges and issues of translation. This leads to a tendency to oversimplify the translation task and assume that it can be performed perfectly by MT. This assumption, in turn, leads to failed deployments because the MT system cannot meet the expectations of the user.

In some environments, naivete has been coupled with intentional derailing of MT by human translators. Although many human translators have come to understand that MT does not compete with them, purists continue to point to MT's obvious foibles as evidence that it is not usable for any translation requirement. Combined with unrealistic expectations, this viewpoint is almost always lethal for MT applications. Prevention is the key, and SYSTRAN, as well as other MT developers, will have to continue to work hard to educate users and tune its translation services to the particular needs of its customer before the service is released for production.

MT companies have reduced the marketing hype that fueled unrealistic expectations, however, they still persist and only broad use and familiarity with MT will completely do away with them. Herein lies the catch-22 of MT — customers need to use MT to truly understand its value, but they must first understand what it does in order to use it successfully. Gradually, this barricade is eroding, and continued successful deployments of MT will eventually unleash the tantalizing and enormous potential that everyone in the industry can see but none have yet realized.

The new SYSTRAN is a robust, modular natural language analysis and generation system with deep lexical development in many domains and a highly customizable set of natural language processing tools.

Although the risks that SYSTRAN faces are serious, the possibilities are compelling. The new SYSTRAN is a robust, modular natural language analysis and generation system with deep lexical development in many domains and a highly customizable set of natural language processing tools. While MT has been on the fringes of success for many years, less comprehensive natural language applications are seeing some real interest and success. Natural language techniques, such as morphological analysis, semantic networks, noun phrase identification, and text normalization, have been introduced into content management, search, and cross-lingual information retrieval with success. SYSTRAN's linguistic resources are unparalleled among commercial MT systems. The company can make incursions into these other areas of text analysis if its resources are modular, efficient, and exchangeable.

The content management industry is ripe for machine translation as the balance of languages on the Internet shifts away from English as the majority.

In particular, the content management industry is ripe for machine translation as the balance of languages on the Internet shifts away from English as the majority. In fact, IDC has recently published a study (*Internet Commerce Market Model* version 7.3, 2002) demonstrating that Internet users in Western Europe now surpass the number of U.S. users (see Figure 1). For information-based businesses, content is a corporate asset that must be carefully managed. SYSTRAN customizes its technology for content management to help customers structure their content, distribute it more broadly, create abstracts, and manage terminology. The MT market is maturing but slowly, and it may turn out that content management applications are the growth engine for SYSTRAN in the near term.

A Changing Competitive Arena

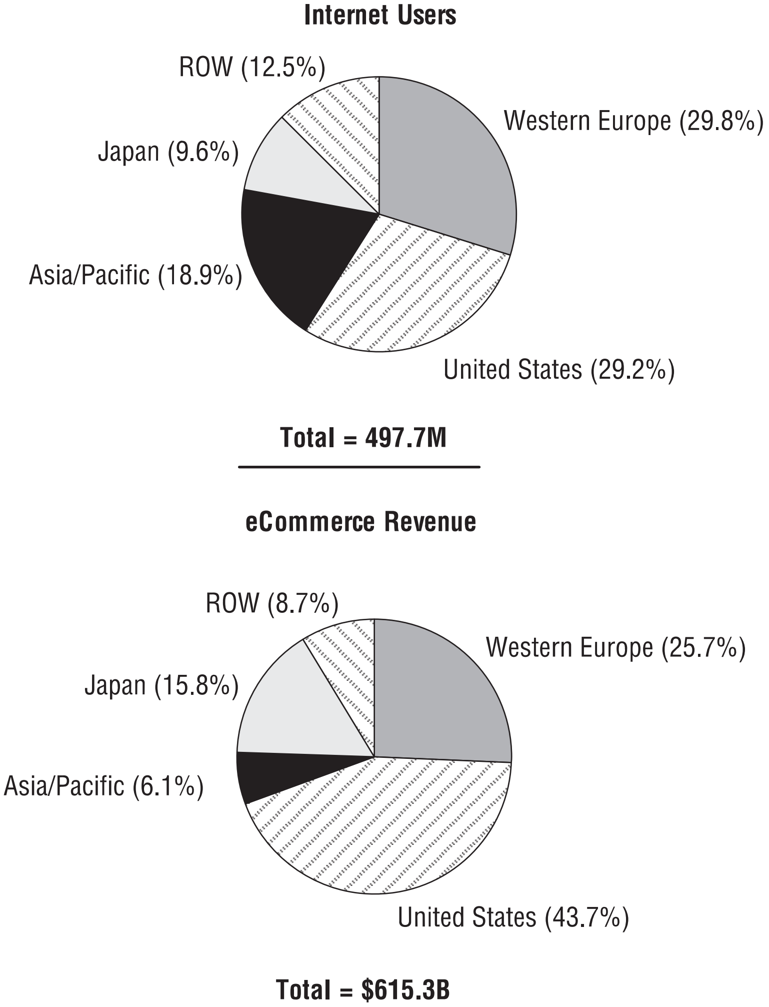
The 18-month period between mid-2000 and the present has been more eventful for MT than the entire previous decade.

The 18-month period between mid-2000 and the present has been more eventful for MT than the entire previous decade. The collapse of Lernout & Hauspie, its spin-off of Sail Labs, the series of failed acquisition attempts of the Barcelona technology, SYSTRAN's Internet deployments, the release of IBM's WebSphere translation server, the demise of Logos, and the acquisition of MT by localization companies such as SDL and Bowne Global Solutions are only a few of the events of the period. The landscape of the MT world is radically changed after decades of stability and, in some cases, stagnation.

SYSTRAN's rearchitecture introduces another change to the MT landscape. Although none of the previous generation of MT systems has completely disappeared, the lineup of MT systems that are still being marketed for enterprise, Internet, and retail applications has been reduced at least temporarily because MT systems are being acquired by globalization and localization companies.

The sale of the Transcend technology to SDL in February was a bellwether of things to come. The Barcelona system was acquired by Bowne Global Solutions, and Lionbridge has partnered with Sail Labs to deploy, develop, and comarket NLP technologies and services. SDL and Bowne Global Solutions appear to have plans for continued marketing of systems to external users, although not necessarily in the retail "shrink-wrap" market. Lionbridge's intentions are less clear.

Figure 1
Worldwide Internet Users and eCommerce Revenue, 2001



Source: IDC's *Internet Commerce Market Model* version 7.3, 2002

However, few localization companies have the specialized staff to develop and maintain MT systems. Although some of the staff of the former MT vendors may move with the technology, a slow start seems likely since the acquiring companies will need time to understand and assimilate the new technology, and define new business goals.

The buy-up of systems by localization companies leaves just three independent MT developer/vendors as potential competitors for SYSTRAN: IBM, Sail Labs, and LogoVista. IBM's WebSphere Translation Server, released in January 2001, is the result of many years of linguistic research at the company's Watson Laboratories. With 12 language pairs and a robust architecture, the system is a respectable competitor, though its linguistic resources do not yet match SYSTRAN's. However, IBM has concentrated exclusively on the Enterprise model, marketing its technology for in-house use by corporations with internal translation needs.

Sail Labs' business is primarily Europe centric. The company has substantial linguistic technology and a robust MT product, Comprendium, but most of its revenue is from consulting. IDC expects that to change as Sail Labs refocuses its business, now that it has freed itself of its ties to Lernout & Hauspie.

The LogoVista system for Japanese and Spanish is developed by Language Engineering Corporation (LEC). Several other licensed language pairs are marketed under the LogoVista name. LogoVista is a popular and respected system in Japan, where it is the market leader for Web-browsing applications. The technology has had a lower profile in the United States due largely to its focus on Japanese. LogoVista's recent licensing of several additional European and Asian language pairs will expand its presence in the United States and Europe. Although the company has been most successful in the Web-browsing application, it offers enterprise and desktop translators as well. LogoVista has a modern code base and produces high-quality MT. However, the company will have to rely on the licensors of its European language pairs to make innovations to the translation technology. It will be interesting to observe what market niches the company pursues beyond multilingual Web browsing.

In addition to the independents, there are numerous emerging systems, many are university based. None are positioned to unseat SYSTRAN at the moment, though some promising technologies using example-based and hybrid techniques are approaching commercialization.

Conclusion

With its newly architected system, its persuasive hold on the multilingual Web-browsing market, and its successful incursions into the customer service area, SYSTRAN is poised to achieve a level of success that has not yet been realized by any other MT company and perhaps to redefine the field of uses for automatic translation solutions.

SYSTRAN's stability as a 35-year-old independent MT developer can be leveraged in the current environment of upheaval among MT companies. The company has an opportunity to secure a position in more niches while its competitors adjust to their various transitions. Doing so, however, will require adequate staffing resources and a tolerance for risk that is not characteristic of the company historically. But, the fact that SYSTRAN opted to incur the risk and cost of modernizing its system suggests that the outlook within the company is as altered as its technology. Regardless of whether it expands the focus, SYSTRAN can become a very successful MT provider simply by owning the two market niches it already is in.

SYSTRAN will need to move cautiously, given MT's history, and avoid the obvious pitfalls of unrealistic expectations and inappropriate applications for the technology. Having demonstrated the ability to navigate these challenges successfully thus far, there is every reason to expect SYSTRAN will succeed.

Document #: 26459
Publication Date: January 2002
Published Under Services: eBusiness Globalization Support Applications; Content Management and Retrieval Software