Statistical versus knowledge-based machine translation

The problem of translating between languages is both ancient, illustrated by the Tower of Babel in biblical times, and widespread, with over 1,000 different languages in use today. Researchers have been working on machine translation of languages for almost 50 years. While there have been some successes and a few commercial systems, high-quality, fully automatic machine translation remains an elusive goal. Not surprisingly, there is some disagreement about how best to proceed. On.one side, researchers working on knowledge-based approaches argue that to obtain high-quality translation requires considerable linguistic knowledge and large knowledge bases. On the other side, researchers working on statistical approaches argue that it is impractical to build large enough knowledge bases to make this feasible, but large corpora of translated text do exist that can be used to train a statistics-based system. In the middle are the hybrid approaches that attempt to combine the strengths of these approaches.

In this issue; we have gathered a number of distinguished researchers who are actively working on the problem of machine translation. Yorick Wilks, a professor of computer science at the University of Sheffield and director of the Institute of Language, Speech, and Hearing, identifies the reasons behind the success of the statistical approaches and argues why they must be integrated with knowledge-based approaches. Ken Church, a researcher at AT&T Research, describes the basic idea behind statistical approaches and claims that they will play an important role in building tools for machine translation. Sergei Nirenburg, a professor of computer science at New Mexico State University and director of the Computing Research Laboratory, maintains that machine translation is too complex to be handled by statistics and that a pure knowledge-based approach will eventually win out. Finally, Eduard Hovy, who leads the natural language project at the Information Sciences Institute and is a professor of computer science at the University of Southern California, argues that integrated approaches are inevitable and that precise nature of the integration will depend on the specific application task.

- Craig A.Knoblock, Editor

Machine translation: a hybrid view

Yorick Wilks, University of Sheffield

After only 35 years of effective machine translation R&D, I feel about its condition somewhat the way Mao Tse-Tung is said to have felt about the significance of the French Revolution after nearly 200 years: it's too early to tell.

The broad facts are apparent to anyone who reads the newspapers, and are therefore a potentially inconsistent set: MT works, in the sense that everyday MT systems at the Federal Translation Division in Dayton, Ohio, and at the European Commission in Luxembourg produce fully automatic translations that many people use with apparent benefit. Moreover, more than 6,000 MT systems have been sold in Japan alone. But, the failure of intellectual breakthroughs to produce indisputably high-quality, fully automatic MT is also apparent, which has led some to say it is impossible, a claim inconsistent with the first observations.

These simple statements could have been made 10 years ago. What has changed since then is twofold: first, the irruption into MT of a range of techniques from speech research, pioneered by IBM Laboratories, that claimed the way out of the deadlock was empirical, in particular statistical, methods that took as data very large text corpora. With these techniques, IBM argued that high-quality MT would be possible without recourse to linguistics, artificial intelligence, or even foreign language speakers. It was not a new claim, for King had made it in the fifties, but IBM reapplied speech algorithms (in particular, hidden Markov models) to execute the program.

The second response, one championed at the time by Martin Kay, was to argue that no theory, linguistic or otherwise, would deliver MT in the foreseeable future. So, the escape from the very same deadlock was to move to machine-assisted MT, which then spawned a score of systems, many now available, that would help users create a translation but that involved, or required, no large claims about automatic MT.

Both developments agreed that linguistic theory was not going to deliver a solution, nor was artificial intelligence. AI had argued since the mid-seventies that knowledge-based systems were the key to MT, as to everything else. They had failed, however, to deliver knowledge bases of sufficient size, and had left us with only plausible examples, as in "The soldiers fired at the women and I saw several fall," where we understand that the "several" is the women—not because of any linguistic selection rules or statistical regularities, but because of our knowledge of how the world works. But the knowledge banks did not appear. Doug Lenat with the CYC project at the Microelectronics and Computer Technology Corp. (MCC) is building a large formal-knowledge base, as is Sergei Nirenburg at New Mexico State University (NMSU), with an ontology of conceptual facts. However, these have not yet been brought into contact with large-scale problems, which was why some people took the statistical claims seriously.

Linguistics was in a far worse position than AI to weather the statistical onslaught: Noam Chomsky's only argument against early statistical claims was that "I saw a triangular whale" was enormously improbable, as a sentence, but nevertheless well-formed. For some reason no one can now remember, arguments of that quality succeeded in repressing empirical methods for 30 years, which explains in part why IBM's pioneering claims were a little like the fall of an intellectual Berlin Wall.

AI researchers who were hostile to linguistics, myself included, perhaps should have been more positive about the IBM

claims when they emerged: some of us had espoused symbolic theories of language that rested on quantifiable notions of the coherence or preference of linguistic items for each other. So, perhaps the statistical view was simply offering a data-gathering method for what we had claimed all along?

But IBM, and its imitators, did better than many expected. Its researchers could produce 50-plus percent of correctly translated sentences from unseen sentences in a trained corpus. To many onlookers that was a striking achievement. But they could not regularly beat Systran, the oldest and tiredest piece of MT software, the one that produces the daily translations at Dayton and Luxembourg.

The IBM researchers then backed away and began to argue that, even if they did need linguistic/AI information of a classic type to improve MT performance (such as lexicons, grammar rule sets, and morphologies), these too could be produced by empirical data-gathering methods and not intuition. In that, they were surely right. That fact constitutes my main argument for the future of hybrid systems for MT, ones that optimize by fusing the best of symbolic and statistical methods and data.

A moment's pause is in order to consider the Systran system, still the world's best performer on unseen text, despised by linguists and AI researchers alike until they needed it as a champion against the statisticians. The truth, of course, is that by dint of 30 years hard labor the Systran teams had produced by hand the large coded knowledge base needed for the symbolic AI approach to work!

Why did the statistical approach do as well as it did so quickly? The best explanation I know is revealing, and also cheering for the future of hybrid systems. Evaluation methods clearly showed that translation fidelity closely correlates with the intelligibility of the output text. Statistical models created a plausible model of generation intelligibility, based on *n*-gram models of plausible text sequence, and in doing so, dragged by the correlation a substantial amount of MT fidelity along with the intelligibility.

The moral here is clear: MT, like prophesy and necromancy, is easy, not hard. One can do some MT on any theory whatsoever, including word-for-word substitution. So, do not be seduced by the claims of theory—only by results. We now have two

Yorick Wilks is a professor of computer science at the University of Sheffield and director of the Institute of Language, Speech, and Hearing. He received his doctorate from Cambridge University in 1968 for work in computer programs that understand written English in terms of a the- later called preference semantics. His most recent books are *Artificial Believers* (with Afzal Ballim, Lawrence Erlbaum Associates, 1991) and Electric *Words: Dictionaries, Computers, and Meanings* (with Brian Slator and Louise Guthrie, MIT Press, 1995). He is a fellow of the American Association for Artificial Intelligence. Contact him at the University of Sheffield, Department of Computer Science, West Court, 2 Mappin St., Sheffield S1 4DT, England; yorick@cs.sheffield.ac.uk.

Kenneth W. Church has been working at AT&T Research on various problems in speech and language: speech recognition, text-to-speech, pronunciation of surnames, morphology, part-of-speech tagging, text analysis, spelling correction, alignment of bilingual texts, machine translation, tools for translators, information retrieval, and optical character recognition. He received his PhD from MIT in 1983 in computer science. Contact him at AT&T Research, Office 2B421,600 Mountain Ave., Murray Hill, NJ 07974; kwc@research.att.com.

Sergei Nirenburg is the director of the Computing Research Laboratory and a professor of computer science at New Mexico State University. He received his PhD in linguistics from the Hebrew University of Jerusalem and his MSc in computational linguistics from Kharkov State University. His research interests include knowledge-based, example-based, and multiengine machine translation; computational semantics; computational lexicography; natural language analysis and generation; knowledge acquisition; intelligent interfaces; planning; and cognitive modeling. He has served since 1987 as editor-in-chief of *Machine Translation.* Contact him at New Mexico State Univ., Computing Research Lab, New Science Hall, Rm. 286, Stewart St., Las Cruces, NM88003; sergei@crl.nmsu.edu.

Eduard H. Hovy is a project leader at the Information Sciences Institute and research assistant professor of computer science at the University of Southern,California. He currently leads the natural language project at ISI, project conducting research in machine translation, natural language text planning and generation, the construction of large knowledge bases and lexicons, automated topic extraction and summarization, medical informatics, and multimedia presentation planning. He received his PhD in computer science from Yale in 1987 and is vice president of the Association of Machine Translation in the Americas (AMTA). He also serves on the executive board of the Association for Computational Linguistics. (ACL). Contact him at USC/ISI, 4676 Admiralty Way, Marina del Rey, CA 90292-6696; hovy@isi.edu

CraigA: Knoblock is a senior research scientist at the University of Southern California's Information Sciences Institute and a research assistant professor in USCs Computer Science Department He received his BS in computer science from Syracuse University in 1984, and both his MS and PhD in computer science from Carnegie Mellon in 1988 and 1991. His current research interests are in developing and applying planning, machine learning, and knowledge representation techniques to the problem of information gathering and integration. Contact him at USC/ISI, 4676 Admiralty Way, Marina del Rey, CA 90292-6696; knoblock@isi.edu

competing paradigms, symbolic and statistical, each armed with a set of rock-solid examples and arguments, but neither able to beat Systran unaided.

The mass of active MT work in Japan has also, I believe, come up with a cookbook of useful heuristic hints: work with lexicon structures not syntax; preprocess difficult structures in advance of MT input; do not think of MT as an absolute self-contained task but as a component technology that links into and subdivides into a range of related office tasks such as information extraction, word processing, and teaching systems.

The last seems too simple. It is correct but ignores the historic position of MT, the oldest linguistic and AI task, one with a substantial evaluation methodology, so that any natural language processor (NLP) or linguistic theory can still be reliably tested within it. The consequence of these observations is that hybrid cooperative methods are the only way forward in MT, even though for now they may be pursued separately as grammars are extracted empirically from texts and texts are automatically sense-tagged. Work also progresses in parallel on the development of ontologies and knowledge bases. They will meet up again, for neither can do without the other, and all attempts to prove the self-sufficiency or autonomy of each have failed and will probably continue to do so.

Statistical MT ≠ stone soup

Kenneth W. Church, AT&T Research

There is a considerable history of statistical/empirical approaches to machine translation, starting with Warren Weaver[1] and the Georgetown system in the 1950s and 1960s.[2] The Georgetown system eventually became known as Systran, and is still one of the more successful systems on the market. Statistical/empirical approaches lost favor when Chomsky and others pointed out some of their limitations in the late 1950s. It is difficult, for example, to capture long-distance constraints such as subject-verb agreement with trigrams—sequences of three words. Increasing the window size to four or five words does little to address the fundamental issue. The constraint between the subject and the verb ought to be expressed in terms of subjects and verbs, and not in terms of words.

Despite these limitations, though, there has been a resurgence of interest in 1950s-style empirical and statistical methods in a variety of applications of natural language processing, including MT. The reasons for this resurgence are difficult to pin down. Some point to massive quantities of online text (corpus data), while others point to improvements in computer technology. In my more cynical moments, I wonder if the never-ending cycle from empiricism to rationalism and back again is just an artifact of human nature. Maybe it is inevitable that students revolt against their teachers. As Mark Twain put it, grandparents and grandchildren have a natural alliance; they have a common enemy.

"Existing translations contain more solutions to more translation problems than any other existing resource."—Pierre Isabelle

Peter Brown et al.[3] are credited with reviving interest in statistical MT. Their work is based on Shannon's noisy-channel model. Imagine a noisy channel, such as a noisy telephone, or a speech recognition machine that almost hears. A sequence of good text (*I*) goes into the channel, and a sequence of corrupted text *(O)* comes out the other end.

$I \rightarrow$ Noisy channel $\rightarrow O$

How can an automatic procedure recover the good input text, *I*, from the corrupted output, *O*? In principle, one can recover the most likely input, *I*, by hypothesizing all possible input texts, I', and selecting the input text with the highest score, $\Pr(I'|O)$. Probability estimates are obtained by computing various statistics over a large sample of text such as a few years of the Associated Press newswire.

Translation doesn't exactly fit into the noisy channel model. Brown et al. assume that a French sentence, *F,* is just a noisy version of an English sentence, *E.* In this way, they view French-to-English translation as the task of recovering the "underlying" English sentence from the "observed" French sentence.

$E \rightarrow$ Noisy channel $\rightarrow F$

Conceptually, their translation program searches the space of all possible English sentences for the sentence *E* that maximizes $\Pr(E|F)$. Their probability estimates are based on large samples of Canadian parliamentary debates, which are published in both English and French.

This approach is extremely controversial. On the surface, it would appear to be fundamentally flawed for reasons pointed out by Chomsky and others in the late 1950s. How can a (purely) statistical approach handle subject-verb agreement? Morphology? In many cases, Brown et al. have adopted solutions to these problems that look remarkably "linguistic," leading Yorick Wilks to charge that their approach is just stone soup. They talk a lot about the statistics, but we "know" that the linguistics is doing the bulk of the work.

There has been a lot of rhetoric on both sides. Who knows whether statistics are more important than linguistics or vice versa? I must say that I find the debate somewhat tiresome. Neither approach has made much progress; we are still a long ways from Yehoshua Bar-Hillel's ultimate goal: fully-automatic high-quality translation (FAHQT). Perhaps the statistical/empirical approach is a step in the right direction, and perhaps not.

But either way, the statistical approach is producing a very interesting by-product: alignment programs that figure out which parts of a translation correspond to which parts of the original. These programs are being used in translation reuse. Many large jobs (such as manuals) are updated on a regular basis and don't change all that much from one version to another. Translation reuse tools make it easy to translate just the "diffs," rather than the entire job. There is a significant niche market for translation reuse. Reuse could easily be a bigger money-maker than MT. At best, MT might be able to speed up a translator by a factor of two, whereas translation reuse can achieve much larger speedups if there aren't too many "diffs."

Alignment programs are also being used to produce just-in-time glossaries. Terminology is a major bottleneck for translators. How would Microsoft, or some other soft-

ware vendor, want the term "dialog box" to be translated in their manuals? Technical terms such as "dialog box" are difficult for translators because they are generally not as familiar with the subject domain as either the author of the source text or the reader of the target text. In the past, translators had to read a lot of background material in both the source and target languages until they mastered the terminology in both languages, an extremely labor-intensive process.

Parallel texts could be used to help translators overcome their lack of domain expertise by providing them with the ability to search previously translated documents for examples of potentially difficult terminology and see how they were translated in the past.

"Existing translations contain more solutions to more translation problems than any other existing resource."[4]

In this way, the statistical approach is producing a set of useful terminology and reuse tools. Unlike traditional MT, these tools do not attempt to compete with the human at what the human does best (translating the easy vocabulary and the easy grammar), but complement the human in areas where they know they need help (difficult vocabulary and reuse). The tools approach was proposed by Martin Kay 15 years ago in "The Proper Place of Men and Machines in Language Translation."[5] In contrast with fully-automatic MT and largely automatic approaches such as machine-assisted translation followed by post-editing, Kay advocated the more modest goal of building tools that human translators would want to use.

It would be ironic if statistical MT ended up producing a toolbench that isn't statistical and isn't MT. But at least it isn't stone soup. . . .

References

1. W. Weaver, "Translation," in *Machine Translation of Languages,* W. Locke and A. Booth, eds., MIT Press, Cambridge, Mass., 1955.

2. B. Henisz-Dostert, R. Macdonald, and M. Zarechnak, eds. *Machine Translation,* Mouton Publishers, The Hague, The Netherlands, 1979.

3. P. Brown et al., "A Statistical Approach to Machine Translation," *Computational Linguistics,* Vol.16, No. 2, 1990, pp. 79-85.

4. P. Isabelle, "Bi-Textual Aids for Translators," *Proc. Eighth Ann. Conf. UW Centre for the New OED and Text Research,* UW Centre for the New OED and Text Research, University of Waterloo, Waterloo, Ontario, Canada, 1992.

5. M. Kay, "The Proper Place of Men and Machines in Language Translation," unpublished manuscript, Xerox, Palo Alto, Calif., 1990.

The inflexible fickleness of fashion

Sergei Nirenburg, New Mexico State Univ.

Machine translation has been a fashionable field for at least 40 years of its 50-year history. The reasons for this vary from R&D glory to commercial payoff. Over the years, researchers have used an impressive variety of methods as the basis for translation programs. The problem, however, has proved so complex that the quality of the result has not correlated significantly with the method chosen. Rather, it typically correlated with the amount of descriptive work on language that was carried out.

Of course, MT research has brought about significant side benefits. Entire scientific fields have emerged largely because of MT efforts. Witness the nascence of computational linguistics. Often, MT served as an application of choice for various workers to test and attempt to corroborate their theories of language and human thinking capacity. Characteristically, the Eurotra project's final report listed as its major success the creation of computational-linguistic infrastructure in the countries of the European Community, deemphasizing the fact that no realistic MT system was built under its auspices. Many factors contributed to the lack of engineering achievement in this project, among them the relative lack of emphasis in Eurotra on actual description and system building, with preference given to designing detailed formal specifications of (largely syntactic) levels of analysis and their corresponding formalisms.

Is the Eurotra case prototypical for the entire field of MT? One problem with the field has been that the descriptive work is, frankly, rather monotonous and boring. This is why attempts were made either to make it less boring (by adding an independently motivated theoretical angle to the descriptive work), or to try to avoid it altogether.

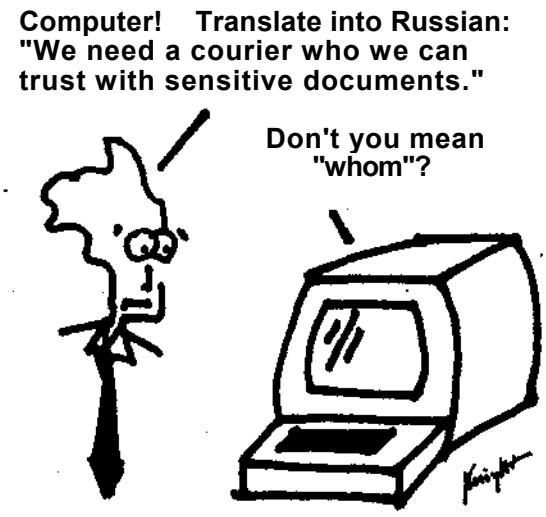
The latter objective manifested itself in

- attempts to use AI learning techniques or more practical semiautomatic procedures for knowledge acquisition, and
- the application of statistical methods for establishing cross-linguistic correspondences in lieu of language-description work.

The former solution emerged in viewing MT as a testbed for one's favorite linguistic or computational-linguistic theories, such as the currently fashionable "principle-based" approach to syntax. MT is indeed a tempting avenue of computational inquiry into modeling human mental and language processes, and a number of approaches to NLP in AI dabbled in MT as a potential application. Knowledge-based MT is a direct offshoot of the AI tradition.

The most remarkable feature of the statistical methods in MT is that they are not at all specific to their subject matter: the same techniques applied to processing language could be and are used, for example, in the studies of the human genome.

The current R&D-oriented MT approaches, whether rule-based or statistical



or hybrid, are based on "imported" ideas. Also, the best systems on the market cannot boast much of technological or scientific advances. Instead, they rely on brawn: huge, hand-crafted dictionaries and grammars, and a plethora of specialized translation routines. All of us are curious to see how well the R&D approaches will work once sufficient resources are allocated for one or more of them to reach the status of a product. The question is: What kind of imported techniques show the most promise? The answer is not obvious and is determined by sociological (read: the vagaries of funding) as well as scientific and technological trends.

The major scientific (or methodological) trend in the field is experimenting with how well the statistics-oriented methods will advance the state of the art in MT without the need for massive, manual knowledge acquisition.

The major technological trend in the field is looking for the best ways to mix the statistical and the rule-based methods. I have been an early advocate of mixing such methods at the level of their final results, an approach called multiengine MT. Other approaches seek a more involved interaction, with statistics used not only during the process of MT but also to support development of background resources—dictionaries and grammars.

The major sociological trend, at least in the US, is the emphasis on a regimen of evaluations and competitions among MT (and, more broadly, NLP) systems. This promotes rigor and discipline as well as conformity and search for local solutions, which are not necessarily the most promising ones in the long run. Approaches that show a steady improvement are rewarded. Approaches with long gestation periods are punished.

Emphasis on mixed approaches is, for nonstatisticians, a rearguard regrouping action, while for statisticians—witness the evolution of the claims and practices of the Candide IBM MT group—it is a search for any avenue for improving the rather modest results.

The knowledge- and linguistics-based methods will do well to regroup and concentrate on those tasks and situations in which statistical approaches fail to deliver. One must remember, however, the lesson of computer chess: at present, the best chess-playing systems are not terribly knowledgeable about chess strategy and tactics, but they consistently beat AI-based programs and compete on equal terms with grandmasters. The $64,000 question is: How much more complex is human translation ability compared with the human chess-playing ability? That is, for how long will there be an opportunity to study language use through MT? If statistical methods succeed, rule-based MT may go the way of the AI-based chess programs.

My opinion is that MT is too complex for the current statistical processing methods to handle, even though these methods do not aspire to building representational models of human language capacity and rely only on the input-output behavior of such models (in MT, a text and its translation). In the final analysis, the open-endedness of language will become the stumbling block for these methods conceptually, just as, logistically, the chronic shortage of resources (bilingual corpora) may precipitate the swing of the pendulum of MT R&D fashion back to the mentalist camp from its current behaviorist direction.

How long will this take? If history is any guide, such swings come roughly every 30 years: Mentalism was in scientific ascendancy between 1960 and 1990, while behaviorism reigned, at least in the US, for about 30 years prior to that. Of course, we cannot be certain that we are witnessing this pendulum swing and not some other, unconnected development. Time will show.

A more intriguing thought is that, just possibly, the rule-based/corpus-based dichotomy is not as important as we think. Maybe the real problem of MT as technology is that AI researchers do not generally understand how difficult the problem actually is. The confident claims, made by newcomers to MT (including me some 15 years ago), help stoke the high expectations of getting the desired result with a modest expenditure. At the current level of MT R&D, either we should lower the expectations or significantly extend the time scale for getting results. For best results, we might need to fund a language description effort of truly Tower of Babel proportions.

Deepening wisdom or compromised principles?—the hybridization of statistical and symbolic MT systems

Eduard H. Hovy, USC/ISI

In late 1991, at the outset of ARPA's MT program in the US, the Statistics Wars were getting into full swing. On the one side stood the Candide system (built at IBM, New York), untouched by any taint of symbolic or linguistic methodology or knowledge; on the other, in statisticsless purity, stood the Interlingual systems out of which Pangloss would be built (by a collaboration of the Center for Machine Translation at Carnegie Mellon University, the Computing Research Laboratory at NMSU, and the Information Sciences Institute of the University of Southern California). A third system, Lingstat, refused to enter the Wars, opting instead to mix and match whatever techniques would help in the next evaluation.

Five years and four evaluations later, the picture looks completely different. All three systems, as well as a new system called Japangloss, built at USC/ISI, ended up employing methods from both statistics and linguistics. Although fundamental differences remain, it is informative for all future MT systems (and, in fact, for other NLP systems as well) to identify what parts of the systems tend toward linguistics, what parts toward statistics, and why this should be so.

Hybridization: natural and inevitable

Departing from your principles is hard. The frequency and volume with which the principles of statistical and symbolic MT were repeated during the Statistics Wars

means impressive-sounding arguments will continue for a while. (The muted tone of my fellow contributors is rather surprising.) But it is perhaps more instructive to consider what happens in practice. Baldly stated, if you want to build a non-toy MT system—a system with more than approximately 5,000 lexical items—that handles previously unseen input robustly, you always end up including some statistics-based modules or knowledge. And if you want a reasonable level of output quality, you always end up including some symbolic/linguistically motivated knowledge or modularization.

It is not hard to see why this should be so. Consider first the hybridization of the statistical approach. A system built on purely statistical grounds starts without any knowledge about language. This means that the fact that a word has five letters, ends with a "t", is capitalized, or reliably pairs up with the words "the" or "a", are all equally significant. The system has to sift through millions of combinations, looking for valid correspondences that hold between the source and target languages.

Some correspondences are easy to find—word *X* in Spanish is word *Y* in English—while others are not. As language speakers, we use abstract types such as word classes (nouns, verbs, and so on) to find general, very powerful correspondences. For example, we know that capitalization is important precisely because in many languages it signals proper nouns, which translate differently than other words. We also know that words that combine with "the" and "a" are common nouns, so they usually appear also in plural form, can affect verb forms, and so forth. If we pre-inform a statistical system with this kind of knowledge, the system can save enormous amounts of processing time by focusing its search to find regularities where they are likely to be found.

Simply building into the initial statistical model the idea of word classes is a big step away from pure language-independent statistics, a step toward symbolic/linguistic knowledge. For statistical systems, the impetus has always been a drive toward quality—coverage and robustness the systems already have. But we gain increased quality only by using increasingly specific rules, and inevitably these rules involve abstractions based on linguistic patterns. The questions facing statistical system builders are: Which phenomena should we abstract over, and what kinds of symbol systems should we create for them? Every time a new phenomenon is identified as a bottleneck or as problematic, the very acts of describing the phenomenon, defining it, and creating a set of symbols to represent its abstractions are symbolic (in both senses of the word!). The benefits: decreased learning time and more powerful rules, hence improved translation output quality. The picture is inverted on the symbolic/linguistic side. Here the system is designed to use a great deal of knowledge about lexical features, grammatical word classes, and even perhaps semantic knowledge, as in the case of Pangloss. But this knowledge must be built into the system. Lexicons of words and rules of grammar, acquired by human labor, are expensive to compile and slow to accrue. Where a statistical system can sift through thousands of bilingual word correspondences an hour, a human cannot build more than a handful of detailed lexical items or grammar rules in that time. Even if the human's work is perfect and complete, the fact that one needs at least 120,000 words to cover a significant portion of a language such as English or Spanish means that it takes years for a group of lexicogrammarians to develop an adequate MT system. Generally, in the real world, the oldest systems are still the best.

Even if the human's work is perfect and complete, the fact that one needs at least 120,000 words to cover a significant portion of a language such as English or Spanish means that it takes years for a group of lexicogrammarians to develop an adequate MT system.

But ARPA had only three years' funding for MT. And the ARPA MT program became increasingly ambitious, from initially calling for high-quality translations in only a limited domain (necessitating a small but detailed lexicon), to ultimately requiring the systems to handle unrestricted newspaper text. Over the four years of the program, ARPA held four formal evaluations, which used various scales to compare translations produced by research systems, several commercial systems, and human experts.

Pressure increased on Pangloss, the symbolic/linguistic system, to expand its lexicon and grammar dramatically. The only way to respond was to automate: decrease the amount of information for each lexical item (because this usually requires human analysis), and acquire the lexical items and grammar patterns by machine. This step immediately introduced statistics-like processing into Pangloss.

Until they mature, symbolic systems thus respond mainly to a drive toward coverage and robustness. Especially in the face of increasingly challenging evaluations, symbolic system researchers begin to develop general rules to avoid catastrophic failure whenever the system encounters input for which specific rules have not yet been built. Such general rules usually provide not only the correct output for any input but a list of possible outputs for a general class of inputs. These outputs, which are correct at a certain level of generality, are filtered to select the best altemative(s). But what filter? When the task/evaluations prohibit human intervention, the filter must be automatic, and thus requires reliability indicators. By the twin moves of computing reliability numbers and extracting information from resources (semi-) automatically, symbolic system builders take their inevitable steps toward statistics. The benefits: a greatly expanded lexicon and more grammatical coverage, hence translation in larger domains.

Once begun, the process of hybridization continued for Candide and Pangloss (Lingstat and Japangloss, a sibling of Pangloss, were hybrids from the outset).

The future

It is possible of course to argue for a reduced role for statistics. In his (uncharacteristically subdued) article, Ken Church says that the primary value of statistics-based MT is to provide a basis for the construction of tools to assist translators. In this he aims low but hits a mark. Still, the eventual future of MT lies not with (semi-) professional translators, but in systems that work for *everyone,* and hence require more knowledge than pure statistics-based tools provide.

Equally uncommon is that Sergei Nirenburg writes with a muted pen as well. Still, his barb is there, just below the surface: the implication that statistical methods by themselves are not useful for anything at all in MT. But what about the "massive field work" Nirenburg identifies? Anyone who does such work without harnessing statistical methods is surely missing the boat.

Ultimately, the goal remains: fully automated, high-quality translation of non-toy domains. In systems of the future, what components will tend toward statistical, and what toward symbolic, solutions? Given the large variety of phenomena inherent in language, it is highly unlikely that a single method exists to optimally handle a given phenomenon—either in the data/rule collection stage or in the data/rule application (translation) stage. In general, symbolic approaches function better on phenomena exhibiting regular linguistic behavior, while statistical approaches handle phenomena with little regular behavior, such as lexically anchored phrases. What constitutes "sufficient" regularity is a matter of both linguistic sophistication and patience, and is often legitimately answered differently by different people.

While it is clear by now that some system modules are best approached under one paradigm or another, it is a relatively safe bet that others are genuinely hermaphroditic, and that their best design and deployment will be determined by the system's eventual use in the world. Thus we can expect all future non-toy MT systems to be hybrids. Just as today we use limousines, trucks, passenger cars, trolleys, and bulldozers, each for its own purpose, tomorrow we will develop different kinds of MT systems from different configurations of statistical and symbolic/linguistic modules, each system best suited to a different kind of MT application.

Scanning the range of MT applications, one can identify niches of optimum MT functionality, which provide clearly identifiable MT research and development goal Major applications include

- *assimilation tasks* (such as scan translations of foreign documents and newspapers): lower-quality, broad domains—primarily statistical technology.
- *dissemination tasks* (such as translations of manuals and business letters): higher quality, limited domains—primarily symbolic technology.
- *narrowband communication* (such as e-mail translation): medium quality, medium domain—highly hybridized technology.

Toward the end of his position statement, Yorick Wilks points out a fact worth remembering: it's easy to build MT theories, but not easy to get results. In this regard, statistics-based systems are currently in a better position than symbolic ones because they emphasize evaluation to drive research. But in the long term, despite Wilks's somewhat pessimistic view, large enough knowledge bases will exist to ma the symbolic and linguistic generalization of central importance.