Language Technologies for Scalable Digital Libraries

Douglas W. Oard
College of Information Studies / Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742 USA
oard@umd.edu http://www.glue.umd.edu/~oard

Abstract

Developers of language technology now routinely build scalable systems that can learn from examples, but we have yet to fully embrace these new capabilities to address the challenge of scalability in digital libraries. This paper relates some lessons learned from a recent experience with building language technology for Hindi, using those lessons to suggest implications for future digital library research.

Introduction

It is often said that quantity has a quality all its own.[1] Perhaps the most obvious example of this aphorism is the World Wide Web. From WebCrawler, to Lycos, to AltaVista, to Google, the search engine that indexes the most content has rapidly come to dominate the marketplace. In digital library research, we strive to achieve quality in many ways, including (among others) selection, organization, and access. It is therefore natural to ask about the role of quantity in achieving the quality that we seek in each of these tasks. That is the overarching theme of this paper. Given the venue of this conference, it might be useful to illustrate my points with examples from our recent experience with Hindi.

The Surprise Language Exercise

In the United States of America, many government and industrial concerns sponsor basic research on language technology. For the past four years, the largest such program has been TIDES: Translingual Information Detection, Extraction, and Summarization. One goal of TIDES has been to create responsive language technologies; tools that can be rapidly adapted to meet unforeseen needs. In June of 2003, sixteen research teams worked together in an unprecedented 29-day "Surprise Language Exercise" to see how well they could apply their technology to an unanticipated language—Hindi.

From the perspective of language engineering, the results were impressive. At the start of the exercise, we were unable to identify any broad coverage Hindi-to-English machine translation systems. Five were built in a month. At the start of the exercise, none of the major international Web search engines indexed Hindi, in part because of the proliferation of encoding schemes. Seven research teams built information retrieval systems that could accommodate multiple Hindi sources in a month, and three of these teams developed fully functional interactive systems. Notably, all three could be used without knowing Hindi—queries could be posed in English, and results could be displayed in English. This was accomplished by integrating search technology with machine translation, automated summarization, and automatic extraction of specific types of information (e.g., names and dates).

[1] The quote is first attributed to Lenin, apparently not in the context of digital libraries.

Two factors were responsible for these remarkable achievements: (1) fifteen years of research on techniques for using examples to transform language in useful ways; and (2) the ease with which the needed examples were assembled. This combination has the potential to revolutionize the way we think about language technology for digital libraries.

Machines that Learn

We often speak of automated techniques for language processing as being based on "natural language understanding"—of course, machines cannot actually "understand" anything. Rather, what we generally mean when we say this is that our machines represent things in a way that reflects the way humans understand things. They then use these representations to generate results that humans find useful. Expert systems are perhaps the best known example of such an approach, but many language technologies can be cast in this framework as well. For example, one common approach to machine translation is to analyze each word to find the root form (and, for polysemous words, the intended sense of that root), then to select an appropriate root in the other language, and finally to generate an appropriate word considering factors such as tense, gender and number. This approach works to some degree, but it suffers from two major limitations: (1) doing it well requires a lot of effort, and (2) the effort ultimately yields diminishing returns, often long before the results are satisfactory.

As is often the case, the key to getting beyond these limitations is to look at the problem from a different perspective. Instead of building representations that reflect the way people think, the trick is to build representations that leverage what machines do best. The goal remains the same: to transform language in ways that people find useful. But by moving away from too strong a focus on how people think, we gain the same kind of advantages that airplanes gain by moving from wings that flap to engines that rotate. Flapping wings are fine for birds, but airplanes can fly far faster, farther, higher and more reliably with rotating jet engines.

The central tenet of "statistical natural language processing" is that machines learn better from examples than from explicit instructions. To see how this might be the case, consider a simple ay of doing machine translation. Imagine for a moment that your machine can remember every Hindi sentence that has ever been translated into English. If someone now says something in Hindi that they would like translated, you could simply look up the Hindi sentence and produce whichever of the previously created translations bets matches the context (based on words in the preceding and following sentences, perhaps). If you have never seen the sentence before, you can try stringing together translations of shorter segments (phrases, individual words, syllables, etc.) in ways that reflect the examples you have seen. Here is some output from a Hindi-to-English machine translation system that was built using an approach like that (and just few million words of example translations) at the University of Southern California Information Sciences Institute:

> "warner brothers picture studio said that harry potter of other picture the box office on rash caused great. the company said that by the end of the foreign markets in the picture of tickets on sale of the people 75 billion rupees expenses made. it is so far the most thick earnings to the films in a has become. harry potter's first picture of the name was harry potter-stone and recent film is the name of-harry potter-effect chamber hnton."

Clearly, that's not very good English. But the system that produced it was built using only example translations that could be found or created in 29 days. Careful evaluations of similar techniques for translating Chinese and Arabic into English indicate that systems built in this way are already doing as well as the best available systems that were built the "old" (i.e., human-inspired) way. And there is a broad base of experience that indicates that more examples will likely yield even better accuracy. Similar results have been seen for automated summarization, automatic recognition of entities and relationships in text, and automatic assignment of category labels to text. So we now face a new challenge: where to find the examples from which our systems can learn?

Finding Examples

Finding the needed examples was the central challenge in the June 2003 Surprise Language Exercise. This is a classic make-vs.-buy tradeoff; examples created by hand can better reflect the way language is transformed in the intended application, but far larger quantities could be assembled rapidly by using existing examples. The full story of the Surprise Language Exercise is an interesting one, but it has already been told well elsewhere (Oard 2003). So I'll focus here on some of the insights that we gained into this make-vs.-buy tradeoff. Focusing again on translation, we learned of four existing efforts to collect the kinds of Hindi-English translation examples that we needed. Each of the projects had obtained examples from many sources, and the key to using them together was to convert the text from each source into a common encoding. This proved to be far more difficult for Hindi than for any other language with which we have experience; one participant was heard to remark that it seemed that no two sources used the same encoding! While that's a bit of an overstatement, the lesson here was that encoding conversion was a prerequisite to use of this sort of "found data." Another limitation of the examples that we found was that few of the examples were from news sources. Some news is indeed translated, but many of the sites that offer news in two languages find it more useful to write separate articles in each language (in part because the background and the interests of the audiences may differ). Our goal was to do well at translating news, so this limitation needed to be overcome.

The key to rapidly creating a sufficient quantity of examples is to minimize the "overhead" effort needed to organize the translation activity. The central idea in this case turned out to be automated evaluation. We started by hiring a few translators locally. Their translations were then used by a research team at the Johns Hopkins University as references to assess the quality of additional translations produced by a large number of volunteers. These volunteers were assembled using grassroots marketing—prizes were automatically awarded for the best translations, and people were encouraged to tell their friends, family and associates about the opportunity to participate. We observed rapid exponential growth in the number of volunteer translators, and in a week we had about as many example translations from our volunteers as we did from any other single source (Yarowsky 2003). Importantly, every one of these examples was a news article.

Language Technology for Digital Libraries

The key question that remains to be addressed is how systems that learn from examples can be used to build digital libraries. The potential of machine translation systems to facilitate management of multilingual content is obvious. As the example above illustrates, however, present systems provide far better support for the task of recognizing a document's topic than they do for supporting a nuanced understanding of its content. As we become adept at assembling progressively larger sets of example translations, we can look forward to improvements in this regard.

There are, however, several situations in which example-based techniques already offer excellent potential. Perhaps the best known is automatic text classification, in which example-based techniques can leverage an initial manual content annotation effort to produce automated assistants that can suggest annotations for previously unseen content. Unlike earlier rule-based machine-assisted indexing systems, these new systems benefit from experience, learning as human operators correct their mistakes. Similar techniques can be used to recognize names in running text and to help with name authority control.

The confluence of these technologies provides some particularly interesting opportunities. Because these systems are built using a common framework, it is relatively easy to build components that work well together. For example, we are now able to build cross-language search systems that do about as well as a comparable system for searching in a single language. Text classification and translation can be combined in a similar way; systems that classify Arabic and Chinese texts based on English examples have already been demonstrated. Another combination that has already been tried is recognition of proper names in French documents, combining a named entity recognition system

trained from English examples with a set of English and French translations in which no named entities had been annotated. All of these techniques work reasonably well because the redundancy that is naturally present in our use of language is easily exploited using statistical techniques.

Looking Forward

Digital libraries necessarily include a strong focus on the management of digital content, just as traditional libraries have focused for generations on the management of content in physical forms. Much of the digital content that we manage includes human language, whether expressed as character-coded electronic text, scanned versions of printed or handwritten words, or digital representations of human speech. Language technology is therefore potentially of great value for the management of digital content. This comes as no surprise, of course. Digital libraries today make good use of what we know about searching large collections, and techniques such as machine-assisted indexing are employed increasingly often as we strive to extend our reach to progressively larger collections. But we are on the verge of a new era, one in which our machines will learn from what we do and then apply those capabilities to enable the management of digital content at a far larger scale than we could ever hope to do ourselves.

What does this tell us about the future of digital library research? First, we need more people in our community that understand and appreciate the revolutionary potential of the emerging technologies that learn from examples. Much of what is known about this has been learned in relatively narrow domains (e.g., translating news), and the ways in which success is measured is often not grounded in the needs of a specific application. We need people that will apply these tools to challenging content management problems. Fortunately, we have a long tradition of recognizing emerging technologies and applying them to content management. So the digital library community is the natural place for these threads to come together.

The other important implication of these new capabilities is that scalable solutions bring unique advantages. Systems that learn from examples naturally do better with more examples, and with their assistance we are able to extend our scope to progressively larger applications. This, in turn, results in more examples from which we can learn. As we learn to design systems that exploit this virtuous cycle, we will create capabilities that we can only dream of today. Which allows us to end where we began, by observing that quantity has a quality all its own.

Acknowledgments

The word “we” throughout must be read as including the entire TIDES community, plus the broad range of organizations that created and contributed the resources that we all used. I am particularly indebted to the teams at the University of Southern California Information Sciences Institute, where I was visiting on sabbatical as this remarkable event unfolded, and at the University of Maryland. Meriting special mention are Franz Josef Och and Daqing He, both of whom helped immensely to shape my perspective on these questions. This work was supported in part by DARPA cooperative agreement N660010028910.

References

Oard, D. W., 2003. **The Surprise Language Exercises**, *ACM Transactions on Asian Language Information Processing*, special issue on the Surprise Language Exercises, 2(2-3), to appear.

Yarowsky, D., 2003. **Scalable Elicitation of Training Data for Machine Translation**, *Team TIDES*, November, pp. 3-4, to appear at http://language.cnri.reston.va.us/TeamTIDES.html.