The Surprise Language Exercises

DOUGLAS W. OARD
University of Maryland

For ten days in March and twenty-nine days in June of 2003, sixteen teams in two nations sought to develop language technologies for two previously unanticipated languages; Cebuano and Hindi. This introduction to a pair of special issues explains the motivation for those exercises, the approaches that were tried, and some of the lessons that were learned.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing

General Terms: Algorithms, Design, Experimentation, Languages, Measurement

Additional Key Words and Phrases: Cross-language information retrieval, information extraction, summarization, machine translation, language parsing and understanding, text analysis

1. INTRODUCTION

In the 1966 "ALPAC Report" on computers in translation and linguistics, the US National Academy of Sciences recommended a retreat from the aggressive investments in machine translation that were being made at the time, and a greater focus on "basic developmental research in computer methods for handling language . . . as tools to help check proposed generalizations against data . . . " [Automatic Language Processing Advisory Committee: ALPAC 1966]. Two decades later, introduction of data-driven techniques for what has come to be known as "statistical machine translation" led to a fundamental breakthrough, opening the potential for rapid development of capable machine translation systems based on example translations previously produced by people. With new techniques came new questions, however. Could enough examples be found in whatever language pair was needed? If the examples that could be obtained required cleaning and conversion, how much of a delay would that introduce? If system development timelines were severely constrained, would the resulting translations be good enough to be useful for at least some purposes? In 2003, the US Defense Advanced Research Projects Agency (DARPA) set out to answer these questions with a "Surprise Language Exercise" (SLE).

The genesis of the SLE can be found in DARPA's Translingual Information Detection Extraction and Summarization (TIDES) program. TIDES was

This work was supported in part by DARPA cooperative agreement N66001-00-2-8910.

Author's address: College of Information Studies and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742; email: oard@umd.edu.

Table I. Participating Teams

	Resources	Translation	Detection	Extraction	Summarization
Alias-i	X			X	
BBN	X		X	X	X
CMU	X	X			
CUNY-Queens			X		
IBM	X	X	X		
Johns Hopkins	X	X			
K-A-T			X		
MITRE	X			X	
Navy-SPAWAR	X			X	
NYU	X			X	
UC Berkeley	X		X		
U Sheffield	X			X	
U Maryland	X		X		X
U Mass	X		X		
U Penn-LDC	X				
USC-ISI	X	X	X		X

created to capitalize on an increasingly robust set of language technologies, targeting four grand challenges that have the potential to transform access to information, regardless of the language in which it is expressed. Machine Translation (MT) clearly lies at the center of this effort; without it the information space would be balkanized by language. In many languages, we already have access to more information than we can keep up with; using translation to overcome language barriers merely exacerbates that problem. The goal of "detection" (the core technology that underlies both searching through existing information and filtering newly arriving information streams) is to cut that problem down to size. The results of detection can be used in two ways: by machines, or by people. Machine processing requires ways of recognizing specific types of data items in natural language texts; this is the role of "information extraction." People can make much more sophisticated use of information than can machines, but they are much more easily overloaded. Summarization seeks to improve human productivity by distilling the most salient content into concise summaries.

DARPA programs typically include a substantial focus on evaluation-guided research. In TIDES, each area sponsors one or more evaluation venue in which the effectiveness of the technology is the principal focus (e.g., the Document Understanding Conferences for summarization, and the Topic Detection and Tracking evaluations for detection). The SLE introduced a complementary focus on development timelines, seeking to produce the best possible capabilities in a constrained time frame. Language technologies are increasingly interdependent; a second useful characteristic of the SLE was that it resulted in a focused effort that brought together teams across a broad range of disciplines. Table I identifies the sixteen research teams from the United States and the United Kingdom that developed technology for at least one aspect of the SLE; several additional organizations participated as data providers.

2. ABOUT THE SPECIAL ISSUES

The sixteen papers in this pair of special issues convey many of the lessons learned in the SLE, but they cannot tell the entire story. Like the SLE itself, this special issue was produced in a remarkably short period; just four months elapsed between completion of the exercise and production of these issues. Experimentation and analysis with the data that was created will continue for some time, so the SLE story is in a sense still being written. Moreover, some teams that did very interesting work were unable to submit a paper because of other commitments during this period, and the story will not be complete until their voices can be added. But if we are to build on what we have learned, it is important that we provide venues for reflective thought about what has been achieved. Despite the tight production schedule, these papers have been carefully written, rigorously peer-reviewed, and extensively revised.

The organization of the papers in these issues reflects five main themes of the TIDES program. Language resources are the nexus of all data-driven techniques, so our story naturally begins with the challenge of rapidly assembling the resources needed by the participating teams. Papers on translation and cross-language detection fill the remainder of this issue. The papers in the second special issue focus on using the results of translation and detection. That issue begins with papers extraction and summarization, and concludes with four papers that present work that cuts across these five areas.

The SLE evolved in two stages. In March of 2003, the Linguistic Data Consortium (LDC) organized a dry run that was principally intended to exercise the data collection methods that were being developed for the actual SLE. That dry run, using the Philippine language Cebuano, is described in the next section. Section 4 then describes the June 2003 SLE, for which the Indian language Hindi was chosen. Finally, Section 5 concludes this introduction to the pair of special issues with a brief look at what lies ahead for further work on this challenge.

3. THE CEBUANO DRY RUN

The Los Angeles Times reported that at about 5:20 P.M. on Tuesday March 4, 2003, a bomb concealed in a backpack exploded at the airport in Davao City, the second largest city in the Philippines. At least 23 people were reported dead, with more than 140 injured, and President Arroyo of the Philippines characterized the blast as a terrorist act. Twenty-four hours later, TIDES teams were notified that Cebuano, a language widely spoken in the region around Davao City, had been chosen for the SLE dry run; as it happened, March 5 had already been selected as the start date for that ten-day dry run.

Cebuano was in many ways a perfect choice as a surprise language; tell someone that you are working on Cebuano, and they almost invariably will ask “what’s that?” Unless, that is, they are Jan Edmund Carlsen, who had been working in Sweden for several years on a Cebuano-English dictionary. Around noon on March 5, he became a very popular fellow. Nine of the sixteen SLE teams participated in the dry run, with most contributing to the creation of language resources for common use, and many also using the opportunity to try out their technologies on Cebuano in a rapid-development scenario. This resulted in one MT system, two cross-language search systems, one information extraction system, and two systems for the generation of translated summaries. Most importantly, the dry run yielded four key lessons that shaped the community's planning for the June exercise:

Diversity. It is often said that two heads are better than one; in the dry run it became apparent that nine teams could simply generate more good ideas than any one team could have possibly produced. The production of "parallel text" tells that story well. The University of Maryland produced nearly a million words of verse-aligned parallel text the first day, by the simple expedient of obtaining a Cebuano bible and aligning the verse numbers with those in an English bible that was already at hand. The University of Southern California Information Sciences Institute (USC-ISI) hired native speakers of Cebuano to produce translations, producing several thousand words within a week. But it was not until the team at Carnegie Mellon University found the newsletter of the Philippine Communist Party on the Web in both Cebuano and English that a large amount of truly representative example translations became available.

Parallel Development. When time is available and effort must be limited, sequential investigation of promising alternatives typically yields the best return on investment. But when time is sharply limited and dependencies abound, parallel efforts dramatically improve the chances of success. Imperfect coordination in what was a somewhat chaotic start-up effort paradoxically provided to be beneficial; parallel development efforts evolved naturally when teams were not aware that someone else was attacking the same problem. In most cases, it turned out that one team chose an approach that resulted in rapid availability of a solution. It proved to be remarkably difficult to foresee that winner, however. This was most clear in the case of sentence alignment for bilingual "parallel text." A sophisticated alignment system that was available at one site seemed to offer the greatest promise, but adapting that system to the noisy Cebuano data took longer than anticipated. Initially, other teams had hesitated to waste effort on less sophisticated approaches; but as the end of the dry run approached, several parallel efforts were initiated. Some rapidly yielded usable results, results that could have been available several days earlier.

Formative Evaluation. Perhaps the biggest surprise from the dry run was the crucial role of formative evaluation. Some of the language resources that became available were excellent; others turned out to be nearly useless. Before the dry run, our goals for language resource creation were generally stated in terms of size (e.g., number of translation pairs in a translation lexicon). After the dry run, we thought more in terms of utility. This need led to some innovation in techniques for rapid creation of evaluation resources, notably for detection and translation. That experience proved to be very helpful during the June SLE.

Coordination. We started the dry run with a carefully thought-out process for supporting interaction between participating teams. Over the next ten days, it became clear that we would need to rethink our approach before June. We had started the March dry run with one way of doing each thing that we needed to do (a user-editable web site to share data, and an email reflector to permit notification about newly available data). We finished June with a far more robust set of overlapping data paths that included locally and centrally maintained web sites, frequent teleconferences, and email archives for ready reference to previously seen notices.

In the end, the dry run achieved far more than anyone had envisioned. The processes envisioned for June were indeed exercised, and improvements were made where needed. But the most striking lesson that we learned was how tractable a surprise language could be. In ten days, the participating teams had built working systems for every one of the four main tasks of TIDES.

4. THE HINDI SURPRISE LANGUAGE EXERCISE

On June 2, DARPA announced Hindi as the surprise language. The name of the exercise was well chosen; Hindi proved to be a source of many surprises. Cebuano is written in a western script, and Cebuano documents often contain “loan words” from English that simplify every cross-language task. Not so for Hindi; Hindi is written in the Devanagari script, and the proliferation of Devanagari encodings makes it nearly impossible to use documents obtained from one source together with documents obtained from somewhere else. There was no shortage of Hindi text; Hindi is spoken by about one-tenth of the world’s population. But for two weeks, longer than the entire Cebuano dry run, this encoding problem prevented development of useful systems.

In retrospect, this should not have been a surprise. The world’s largest search engine indexes documents in 35 languages; Hindi is not one of them. And, at the time we started, no broad-coverage Hindi-to-English MT system existed. In 29 days, SLE participants developed 4 Hindi-to-English MT systems, 7 search systems capable of finding Hindi documents, 5 information extraction systems for Hindi and 2 systems to produce English summaries of Hindi documents. What makes the SLE all the more remarkable is that almost all of this happened in the two weeks that remained after the encoding problem for Hindi was solved.

Cracking those codes is perhaps the greatest story from the SLE that has not yet been completely told. Five teams attacked the problem, using three basic approaches: brute force, font analysis, and cryptanalysis. Ultimately, font analysis proved to be successful. Every encoding must be renderable as Devanagari characters that a reader of Hindi would recognize. The font used to display the text thus contains the key to converting any encoding to a standard form; once the details of this were worked out, Johns Hopkins University produced a toolkit that made it possible for someone proficient in Hindi to build a converter from a new encoding in just a few hours.

The scale of the Hindi effort dwarfed the earlier Cebuano dry run. At its peak, at least 100 people were working on Hindi across the sixteen sites; 56 of them have authored papers in these special issues. Some important innovations resulted, including a system for automated elicitation of translation examples using volunteers from around the world [Yarowsky 2003], coupling of statistical machine translation with a far broader range of detection and summarization systems than had previously been tried, and creation of what we believe to be the world's first system that can automatically find answers to English questions in Hindi documents.

At the end of June, the US National Institute of Standards and Technology worked with the LDC to conduct evaluations that approximated those typically performed as part of TIDES. Although time and resource constraints required some compromises, the results of that evaluation process have served to better ground many of the lessons that we have learned from the experience. Many of the papers in these issues include results from those evaluations.

5. CONCLUSION

Of course, Cebuano and Hindi provide just two data points in what is admittedly a large space of possible surprise languages. Moreover, the a surprise language exercise is but one way that we can learn how to prepare for these kinds of challenges. Now that we have some idea what we can accomplish and where the hard spots are, we can focus our efforts where we see the greatest potential for advances. This will undoubtedly be more efficient than making a comprehensive attack at the problem every time would be. Ultimately, however, we will probably need to come back to do this again, if for no other reason than to validate what we have learned. When we do, we will look back to our first experiences here in the TIDES Surprise Language Exercises, and to the articles in these special issues, for the insights needed to design an experience that will test our abilities and challenge our imagination.

ACKNOWLEDGMENTS

The author is grateful to Mike Maxwell, Donna Harman, and Charles Wayne for their leadership on the surprise language exercises and to the 52 anonymous reviewers that made this special issue possible.

REFERENCES

Automatic Language Processing Advisory Committee: ALPAC. 1966. Language and machines: Computers in translation and linguistics.

Yarowsky, D. 2003. Scalable elicitation of training data for machine translation. *Team TIDES*. http://language.cnri.reston.va.us/TeamTIDES.html.

Received October 2003; revised October 2003; accepted October 2003