Translingual Information Detection Extraction and Summarization

TEAM TIDES

April 2003

Highlights:

- **Tech Note:** "A new training criterion for log-linear models directly optimizes the error rate on unseen test data."
- **System Note:** "Newsblaster demonstrates the robustness of current human language technology."

On the Inside:

Program Overview

Charles Wayne, cwayne@darpa.mil

Many good things have been happening. In December, TIDES started a major push to package and transition useful technology components. In January, we delivered the first component, Arabic-to-English machine translation from IBM. We will release additional components (averaging one per month) throughout 2003.

Total Information Awareness (TIA) system developers will combine these components in various ways (with one another and with other technology), conduct experiments to assess their utility, and give us constructive feedback. TIA is already using these and other TIDES technologies in their "wind" experiments and is praising our work.

We are also assembling a first generation translingual Text and Audio Processing (TAP-XL) system. Integrating synergistic components from five groups, TAP-XL will perform detection, extraction, summarization, and translation for English and Arabic speech and text. TAP-XL will evolve over time, integrating more components and another language.

We will use TAP-XL in a translingual Integrated Feasibility Experiment (IFE-XL) to help mature TIDES components (for TIA and other applications) and to help devise and demonstrate new concepts of operation. We ran a preliminary experiment in March, and we will conduct a full experiment with multiple scenarios in July.

To create powerful new capabilities, sites have been conducting wide-ranging research and collecting the linguistic resources needed to accelerate their research. Articles in this issue give technical insights into portions of our extraction, summarization, and translation work.

Objective evaluations are a key aspect of all TIDES research. We use the evaluations to focus attention on key technical challenges and to determine which approaches work (and how well they work). The schedule on the back page lists the formal evaluations and associated workshops planned for 2003. NIST oversees these evaluations and welcomes outside participants (who benefit from and enrich the process).

In March, we ran a dry run of the process for collecting resources rapidly for an unexpected language. We used the Philippine language Cebuano and got surprisingly good results. We will run a full surprise language experiment in June, using a different language and rapidly porting key detection, extraction, summarization, and translation algorithms to work on it. Two articles in this issue discuss that.

This summer's PI Meeting will address the full spectrum of TIDES activities – robust research and resource creation, formal performance evaluations, development and transition of technology components, experiments with those components, and surprise language experiments - plus potential follow-on program initiatives. This by-invitation-only meeting will be held 4-7 August at the Arden House Conference Center in Harriman, New York (pictured above).

Team TIDES is the quarterly newsletter of the TIDES program. The views expressed by the authors are their own, and do not necessarily reflect the views of their organizations or of DARPA. Comments and contributions should be directed to the editor.

Language Technology to the Rescue!

Donna Harman, donna.harman@nist.gov

Suppose that the United States had a sudden requirement to deal with a new "exotic" language, one with few existing resources. This has happened several times in the recent past, with languages such as Haitian Creole and Serbo-Croatian. That can create an urgent need for machine translation, cross-language searching, and additional language processing such as identification of proper names or summarization for more rapid access. The TIDES program is currently sponsoring research in these technologies for Chinese and Arabic, languages for which extensive resources have been assembled by the Linguistic Data Consortium (LDC). This June, a short experiment using a language with comparatively scarce resources will be tried. This whole exercise is viewed as practice for a real national emergency, with the goal of learning how to gather resources quickly, where the technical problems are likely to occur, and how the quality of the various items produced would affect a real problem.

DARPA will start the experiment by announcing the "surprise" language. Participating groups will then have one month to do the given tasks in that language. Results will be reported out at the August PI meeting, undoubtedly providing interesting lessons learned for the whole community.

There will be tasks in four research areas: detection, extraction, summarization, and machine translation. Each area is currently assembling teams of researchers, organized by the TIDES area coordinators, to tackle specific tasks within their area. The detection tasks will be cross-language information retrieval and TDT event tracking. The extraction group will work on name tagging, including tracking both within and across documents, and also date and time normalization. For summarization, the goal will be the creation of headline-length summaries in English, and the machine translation task will obviously be to translate documents.

The LDC will provide some basic resources (a 100K corpus of electronic data) early in the process, and will coordinate the resource sharing process. All participating groups will join in the resource gathering, since issues involving rapid location of resources are critical to the process. As practice for this resource gathering, the LDC led the resource gathering "dry run" that is described below.

Evaluation of the final results in July will concentrate on understanding where the successes are, where the major failures are, what kinds of research would have the biggest payoff in this very scarce resource environment, and maybe what kinds of tools and/or resources should be stockpiled by the government to handle such an emergency. For these reasons, the scale of the evaluation will be small; possibly too small to show significant differences between systems, but large enough to allow good failure analysis. Standard evaluation methodologies and metrics will be used.

The surprise language exercise has become an intriguing part of the TIDES program, offering a tantalizing peek at what the HLT community could do on short notice to help in any future language "emergency".

Surprise: It's Cebuano!

Doug Oard, oard@umd.edu

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 had been chosen for the surprise language data collection dry run that began on that date. The notification observed that Cebuano is spoken by 24% of the population of the Philippines, and that it is the lingua franca in the south Philippines, where the event occurred.

Cebuano was a surprise, but the dry run itself was not – the LDC had surveyed the available resources for a large number of languages, Cebuano among them. They had created a Web-based coordination infrastructure that would allow teams from across the program to both contribute to and benefit from the resource collection effort. And they had set up a "war room" to manage the collection effort.

All this planning paid off remarkably well. Over the next ten days, nine research teams on two continents staged the most extensive resource creation effort ever undertaken for a "low density" language. This resulted in collection of over one million words of parallel text from four types of sources (the Web, examples of usage from scanned bilingual dictionaries, commissioned human translations of news and editorials, and religious texts), translation lexicons from three types of sources (parallel text alignment, scanned bilingual dictionaries, and Web-accessible bilingual term lists), four types of taggers (named entity, part of speech, noun phrase, and

time expression), and two types of morphological analysis (simple stemming and full linguistic root/affix analysis). Five machine translation systems of two types (term-by-term gloss and statistical) were built, along with three cross-language retrieval systems of two types (batch and interactive), and two headline generation systems. Internal evaluations by participating teams indicated that some of these systems worked remarkably well; for example, the University of Maryland demonstrated the ability to find a single relevant Cebuano document using English queries at a (harmonic mean) rank of 7 and the University of Sheffield achieved an F measure above 0.7 for named entity tagging, both within the first three days!

What made this exceptional accomplishment possible? Teamwork, spade work, and grunt work. Of the three, teamwork was the key. No single team could have done this much in ten days; indeed, no team could even build all of the pieces they needed in that period. But after three years of working together in the program, the participating teams pulled together in remarkable ways. For example, Maryland found parallel text on the Web, NYU used that text to create the alignments necessary to train a statistical machine translation system, USC-ISI used those alignments to build that system and translate a collection of news stories, BBN used the resulting translations as a basis for cross-language headline generation, and the LDC provided the coordination infrastructure that made it all possible. All in ten days! Spade work was the second key, and two different types of spade work paid off. First, we had invested over the years in approaches that were well suited to this task, leveraging existing resources and limited human involvement to rapidly build capable systems. Just as importantly, we had invested in people; people with the expertise to apply and assess the tools that they had created. Finally, we succeeded through grunt work. One multi-site team invested 250 hours in the first three days alone.

So what have we learned? Perhaps the most obvious lessons were places that we can leverage a bit more spade work to achieve dramatic reductions in grunt work. Ideas here run the gamut, from streamlining coordination to fine-tuning individual tools. Another important lesson was the value of exchanging results from internal evaluations – without that, it is hard to focus the effort on the most promising resources. But perhaps the most important lesson was that we learned what we can accomplish as a team. On March 4, Cebuano was, without question a low-density language, with little in the way of computational resources and few research groups possessing the expertise needed to develop any. By March 14, that had changed forever.

Technical Note:

Log-linear Models and Minimum Classification Error Training for Statistical Machine Translation

Franz Josef Och, och@isi.edu

Recently, new automatic evaluation methods for machine translation such as the IBM BLEU score and the NIST variant of that measure have been suggested. Given a few hundred test sentences, these metrics often achieve an astonishing degree of correlation to human subjective evaluation of fluency and adequacy. Yet, the use of statistical techniques in machine translation often starts out with the simplifying (often implicit) assumption that the goal is to minimize the number of incorrectly translated sentences. Hence, there is a mismatch between the training criterion and the final evaluation criterion used to measure success in a task.

In this note, we describe a framework for statistical machine translation based on log-linear models that uses alternative training criteria. Preliminary empirical results obtained with this framework are promising.

A very general and convenient approach for statistical machine translation is provided by structuring the translation probability using a log-linear model that combines a set of feature functions.

A standard training criterion for log-linear models is MMI (maximum mutual information), which can be derived from the maximum entropy principle. MMI tries to obtain a probability distribution that optimally explains the given training data. This optimization problem has some nice properties: there is one global optimum, and it can be solved, for example, with gradient descent algorithms. Yet the ultimate goal is to obtain good translation quality on unseen test data, and it is unclear whether this approach yields parameters that are optimal with respect to translation quality.

For simplicity, we assume here that we can measure the number of errors in sentence **e** by comparing it with a reference sentence **r** using a function E(**r**,**e**). We assume that the number of errors for a set of sentences is obtained by summing the errors for the individual sentences. We also refine this approach to handle BLEU and NIST scores by accumulating the relevant statistics for computing those scores (n-gram precision, translation

Training Criterion	Evaluation Criterion mWER	BLEU	NIST
MMI	68.3%	11.3%	5.86
mWER	**68.3**%	13.5%	6.28
BLEU	76.1%	**17.2%**	6.66
NIST	73.3%	16.4%	**6.80**
95% conf	2.7%	0.8%	0.12

Effects of different criteria. Better results correspond to larger BLEU and NIST scores and to smaller mWER.

length and reference length) instead of error counts.

Our goal is to train the model parameters to obtain a minimal error count on a representative corpus. The direct optimization of such an error count is not easy to handle because it is not a continuous function, so gradient descent methods are not appropriate. Moreover, the objective function has many local minima, so even if we manage to solve the optimization problem, we might face the problem of overfitting to the training data.

To be able to compute a gradient, it is possible to develop a smoothed error count criterion, a common approach in the speech community. Powell's algorithm can be used, together with a grid-based line optimization method. To avoid finding a poor local optimum, we start from different initial parameter values. A major problem of the standard approach is that grid-based line optimization can be hard to adjust so that both good performance and efficient search are guaranteed. If a fine-grained grid is used, then the algorithm is slow. If a large grid is used, then the optimal solution might be missed. We use a new algorithm, described elsewhere, for efficient line optimization of the unsmoothed error count using a log-linear model that is guaranteed to find the optimal solution.

The table above shows results for the 2002 TIDES Chinese-English small data track task, using a training corpus of about 100,000 words and a small lexicon provided by the LDC. For optimizing the parameters of the log-linear model, an additional development corpus has been used; translation results are reported on an independent test corpus. The feature functions contain a word-based lexicon, alignment templates (a phrase-based lexicon), an alignment model, various language models and a sentence length model.

In addition to BLEU and NIST scores, the table also shows multi-reference word error rate (mWER), which computes the number of insertions, deletions and substitutions needed to transform the system's translation to any of the reference translations, then divides that sum by the number of reference words. The best system in the official 2002 evaluation obtained a NIST score of 6.14 on this task.

We observe that if we choose a certain error criterion for evaluation, we obtain the best results using the same criterion for training. The differences can be quite large: If we optimize with respect to word error rate, the mWER results are 7.8% or 5.0% (absolute) better than if we optimize with respect to BLEU or NIST. The BLEU or NIST measures seem to behave more similarly, but we see the same overall effect of the training criterion on the final results. The MMI training criterion produces significantly worse results for all evaluation criteria except mWER.

We looked in more detail at a new training criterion for log-linear models that directly optimizes error rate on unseen test data. We showed that better quality measured by these automatic evaluation criteria on unseen test data is obtained. Note, that this approach can be applied to any evaluation criterion. Hence, if an improved automatic evaluation criterion is developed that has an even better correlation with human judgments, we can plug this alternative criterion directly into the training procedure and optimize the model parameters for it. This means that improved translation evaluation measures could lead directly to improved machine translation quality. Of course, the approach presented here places a high premium on the fidelity of the measure being optimized. It might happen that by directly optimizing an error measure in the way described above, weaknesses in the measure might be exploited that could yield better scores without improved translation quality. Hence, this approach poses new challenges for developers of automatic evaluation criteria.

Many tasks in natural language processing, for instance summarization, have evaluation criteria that go beyond simply counting the number of wrong system decisions, and the framework presented here might yield improved systems for those tasks as well.

System Note:

News on Newsblaster

Kathleen McKeown, kathy@cs.columbia.edu

Newsblaster is a system developed at Columbia to provide news updates on a daily basis; it crawls news sites, categorizes stories into six broad topics, groups stories on the same event, and generates a summary of each event. We began running Newsblaster on a daily basis right after 9/11. .Newsblaster generates a Web page for browsing news drawn from the Internet once daily. On a typical day, it generates summaries on event clusters containing anywhere between two and 100 articles.

As a deployed system, Newsblaster demonstrates

the robustness of current human language technologies such as summarization and topic tracking. We have also integrated Newsblaster in TAP-XL to support an Integrated Feasibility Experiment (IFE).

Newsblaster serves as a research environment in which we can explore many new problems. Currently, we are exploring multilingual summarization, where sources are drawn from multiple languages and a summary is generated in English. We have a prototype version of Newsblaster that creates multilingual clusters and produces a page of English summaries, and we are working on generating summaries that update the user on what has happened since the last time he saw news on an event. Another focus of current work is creating a capability to edit generated summaries to improve fluency and accuracy. For example, we are developing techniques to edit references to people, improving coherence while ensuring that those references are correct. Editing will be particularly important as we add multilingual capabilities, given the errors inherent in machine translation.

To move Newsblaster into a multilingual environment, we must be able to extract the article text from Web pages in multiple languages. For example, a recent Web page from the New York Times contained over 70 kilobytes, but the actual article text on that page amounted to less than 7 kilobytes.

Our previous approach to extracting article text used hard-coded regular expressions that were hand-tailored to specific Web sites. It was not easy to adapt this approach to new sites, especially those written in foreign languages. For each new site, a human would need to examine examples of the HTML code for the site and write regular expressions to extract only the text of interest.

Our recent work on article extraction addresses both research needs in multilingual summarization and scaling the system so that new sites can be easily added. The current version of Newsblaster incorporates a new "Article Extraction" module which uses machine learning techniques to identify the article text. That module parses each crawled HTML file into blocks of text based on HTML markup. A set of 34 features are then computed for each block based on simple surface characteristics such as the fraction of the text that is punctuation or the number of HTML links found in the block. Training data for the system is generated with a user interface that supports human annotation. The approach has been trained and tested on Japanese and Russian sites, with performance comparable to English.

In order to facilitate tracing the development of news stories of particular interest, Newsblaster will soon have the ability to track events across multiple days. We have performed a detailed evaluation of algorithms for this task and an investigation to determine

the set of features to be used for document representation.

We have augmented approaches to tracking events that were originally developed for the Topic Detection and Tracking (TDT) evaluations. We represent an event not as a single document cluster, but as a set of document clusters, with links across clusters representing consecutive days. With clusters as nodes and links as edges, this induces a graph-like structure, allowing the system to show how an event may split into several sub-events as time passes. We are investigating the form that a user interface for tracking events across days should take.

We are also developing a module that will evaluate articles as they are assigned to existing clusters to determine if they contain any important new information, or developments central to the event being followed. For example, on one day, news organizations reported that a suspect was arrested in the shooting of two Americans in Kuwait; the next day, they reported that the man was identified and that he confessed. We have an early prototype of a summarizer that can highlight such developments in ongoing events.

In earlier versions of Newsblaster, the daily news update could take up to 12 hours to compute. One current focus of our work is on improvements in efficiency. Through rewrites of the code and parallelization, we have more than doubled efficiency. We are working towards real-time interaction using optimization techniques drawn from the database community.

Newsblaster illustrates the multiple benefits that can result from deploying language technologies in large-scale applications. Scalable solutions place a premium on efficiency, the need for robust capabilities illuminates important new lines of research, and the interaction with real users fosters both requirements discovery and, ultimately, adoption of our technology by those whom it is designed to serve.

Extraction Update

Stephanie Strassel, strassel@ldc.upenn.edu

TIDES Extraction expands in 2003 with new data, new languages, and new tasks. Extraction is the key link between natural language text and structured applications such as data integration and pattern discovery. The Linguistic Data Consortium (LDC) is creating annotated corpora to support TIDES Extraction, in collaboration with the Automatic Content Extraction (ACE) program. This will result in two important types of resources: training data to enable research, and test data for an evaluation this Fall.

Existing ACE corpora will be augmented with new English data for both Entity Detection and Tracking (EDT) and Relation Detection and Characterization (RDC). Chinese and Arabic corpora are also being created for a new multilingual evaluation in 2003. In parallel with this data generation effort, the TIDES Extraction, ACE, and DARPA Evidence Extraction and Link Detection (EELD) communities are working with the LDC to extend ACE's two core tasks in important new ways.

For the EDT task, the current set of five entities (*Persons*, *Organizations*, *Locations*, *Facilities* and *GeoPolitical Entities*) will expand to include *Vehicles*, *Weapons* and *Substances*. Subtypes will be added to allow finer characterization of entities; a *Vehicle* might be further classified as *Land*, *Air* or *Water*, for example. The RDC task will be refined to highlight relations that are of particular interest and to allow finer categorization of some types of relations.

The newest challenge for ACE is recognition and detection of atomic events. This new research task will require systems to extract the specific "text mention" of an event along with associated entities and temporal information. Systems will also be required to characterize the event according to a set of common types. Targeted types include *Transport* (with subtypes *Purchase*, *Movement*, and *Communication*), *Crime*, *Creation* and *Use*. For example, in the sentence "George Bush and Tony Blair held high-level talks in Northern Ireland last Thursday," systems must recognize the specific text mention of the event (held high-level talks), the associated entities (*Person* entities George Bush and Tony Blair), the location (*GeoPolitical Entity* Northern Ireland), the time (last Thursday), and the event type (*Transport*, with subtype *Communication)*.

With both new tasks and new data, 2003 promises to be a great year for Extraction!

Web Links

TIDES	http://www.darpa.mil/iao/TIDES.htm
ACE	http://www.nist.gov/speech/tests/ace/
DUC	http://duc.nist.gov
LDC	http://www.ldc.upenn.edu/Projects/TIDES
MT Eval	http://www.nist.gov/speech/tests/mt/
TDT	http://www.nist.gov/TDT
TREC	http://trec.nist.gov
Newsblaster	http://newsblaster.cs.columbia.edu

Editor: Doug Oard oard@umd.edu
Layout: Erika Barragan-Nunez (USC-ISI)

TIDES Evaluations

Tasks	Test data to Sites	Outputs to NIST	Workshop
Detection			
Topic Detection & Tracking	2 Sep	3 Oct	17-18 Nov
New Event Detection (Arabic, Chinese, English)			
Event Clustering (Arabic, Chinese, English)			
Event Tracking (Arabic, Chinese, English)			
Story Link Detection (Arabic, Chinese, English)			
Text Segmentation (Arabic, Chinese, English)			
High Accuracy Retrieval of Documents	15 Jul	1 Sep	18-21 Nov
Ranked Document/Passage Retrieval (English)			
Extraction			
Automatic Content Extraction	15 Sep	19 Sep	21-23 Oct
Entity Detection & Tracking (Arabic, Chinese, English)			
Relation Detection/Characterization (Chinese, English)			
Summarization			
Document Understanding	31 Jan	17 Feb	31 May-1 Jun
Headlines for single docs			
100 words for 10 docs for a TDT topic			
100 words for 10 docs from a viewpoint			
100 words for 10 docs for a TREC topic			
Translation			
Machine Translation	5 May	9 May	21-22 Jul
Arabic to English			
Chinese to English			
Surprise Language			
Data collection and system development	1-22 Jun		4-7 Aug
Evaluation	23 Jun	30 Jun	4-7 Aug
Detection			
Cross-Language Retrieval (from English queries)			
Event Tracking (from surprise language examples)			
Extraction			
Name Tagging & Tracking (within/across documents)			
Date & Time Normalization			
Summarization			
Single Document Headlines (into English)			
Translation			
Machine Translation (into English)			