The Guide from MultiLingual Computing & Technology

LANGUAGE TECHNOLOGY

#65 Supplement

July/August 2004

contents

GUIDE GETTING STARTED

A Guide to Language Technology

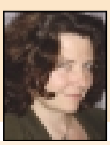
Like any technology in today's world, language technology (LT) continues to evolve and improve, but the variety and technology can be overwhelming. Chris Langewis gives us an introduction to LT, telling us what types of tools are available and how they are used.

Once a technology has come into general use, it becomes critical to measure its effectiveness — hence, the CESTA MT evaluation campaign. One of the outputs of this French Ministry of Research and Education project will be the creation of an MT evaluation toolkit for users and system developers.

An exciting area of LT is the application of speech technology in all sorts of devices other than computers. Bettina Hein writes about the smart avatars that are meeting all of us in our daily lives.

And, to help you sort this out, we have assembled a directory of LT products, a sample of which is on pages 14-15, with more information at www.multilingual.com/lt

— Donna Parrish, Publisher


Marianne Dabbadie


Bettina Hein


Chris Langewis


Widad Mustafa El Hadi


Ismaïl Timimi

authors

MARIANNE DABBADIE is a private company NLP expert for I-KM and MT researcher at the University of Lille 3 at the CERSATES Research unit (UMR 8529). She can be reached at mdabbadie@I-km.com

BETTINA HEIN is chief operating officer at SVOX AG in Zürich. She can be reached at hein@svox.com

CHRIS LANGEWIS, a long-time specialist in translation tools, teaches at the Monterey Institute of International Studies and has a consulting practice in LT and automated process management. He is a member of the *MultiLingual Computing & Technology* advisory board. He can be reached at langewis@langewis.com

WIDAD MUSTAFA EL HADI is a teacher at the University of Lille 3 in the CERSATES Research unit (UMR 8529). She can be reached at mustafa@univ-lille3.fr

ISMAÏL TIMIMI teaches at the University of Lille 3 in the CERSATES Research unit (UMR 8529). He can be reached at timimi@univ-lille3.fr

MultiLingual Computing & Technology

Editor-in-Chief, Publisher Donna Parrish
Managing Editor Laurel Wagers
Translation Department Editor Jim Healey
Copy Editor Cecilia Spence
Research Jerry Luther, David Shadbolt
News Kendra Gray, Becky Bennett
Illustrator Doug Jones
Production Sandy Compton
Photographer Brent Rosengrant

Editorial Board
Jeff Allen, Henri Broekmate, Bill Hall, Andres Heuberger, Chris Langewis, Ken Lunde, John O'Conner, Mandy Pet, Reinhard Schäler

Advertising Director Jennifer Del Carlo
Advertising Kevin Watson, Bonnie Merrell
Webmaster Aric Spence
Assistant Zabrielle Whittom
Intern Kyle Elsasser

Advertising: advertising@multilingual.com
www.multilingual.com/advertising
208-263-8178

Subscriptions, customer service, back issues:
subscriptions@multilingual.com
www.multilingual.com/subscribe
208-263-8178

Submissions: editor@multilingual.com
Editorial guidelines are available at
www.multilingual.com/editorialWriter

Reprints: reprints@multilingual.com
208-263-8178

This guide is published as a supplement to *MultiLingual Computing & Technology*, the magazine about language technology, localization, Web globalization and international software development.

MultiLingual Computing, Inc.
319 North First Avenue
Sandpoint, Idaho 83864 USA
208-263-8178 • Fax: 208-263-6310
info@multilingual.com
www.multilingual.com

CESTA: The European MT Evaluation Campaign

Marianne Dabbadie, Widad Mustafa El Hadi & Ismaïl Timimi

The Campagne d'Evaluation de Systèmes de Traduction Automatique (CESTA, Machine Translation Systems Evaluation Campaign) was approved and financed in 2002 by the French Ministry of Research and Education within the framework of the Technolangue call for projects and integrated to the EVALDA evaluation platform.

In France, EVALDA is the new evaluation platform, a joint venture between the French Ministry of Research and Technology and European Language Resources Association (ELRA), Paris, France. The CERSATES research unit at the University of Lille (France) is the CESTA project leader and head of the CESTA scientific committee.

The campaign's aims are twofold: to provide an evaluation of commercial machine translation (MT) systems and also to work collectively on setting a new reusable MT evaluation protocol. The resulting protocol is to be user oriented and also to account for the necessity to use semantic metrics in order to make available a high-quality reusable MT protocol to system providers.

User-oriented evaluations. CESTA is a scientific campaign that refers to the state of the art in the field of MT systems evaluation. It is grounded in particular on an enhancement of the Defense Advanced Research Projects Agency (DARPA) MT evaluation campaign (1992-1994), the IBM NIST BLEU metric and the FEMTI taxonomy developed within the framework of the ISLE project. The second DARPA campaign, making use of the IBM BLEU metric, is mentioned in the CESTA protocol.

An approach based on use cases. ISO 14598 directives for evaluators put forth as a prerequisite for systems development the detailed identification of user needs that ought to be specified through the use-case document. Moreover, conducting a full evaluation process involves going through the establishment of an evaluation requirements document. The ISO 14598 document specifies that quality requirements should be identified "according to user needs, application area and experience, software integrity and experience, regulations, law, required standards, etc."

The evaluation specification document is created using the Software Requirement Specifications (SRS) and the use-case document. The CESTA protocol relies on a use case that refers to a translation need grounded on basic syntactic correctness and the simple understanding of a text, as required by information watch tasks, for example, and excludes making a direct use of the text for post-editing purposes.

Object of the campaign. The object of the CESTA campaign is to evaluate technologies together with metrics, that is, to contribute to the setting of a state of the art within the field of MT systems evaluation.

The campaign will last three years, starting from January 2003. The European experts on the board are members of the CESTA scientific committee and have been working together in order to determine the protocol to use for the campaign. Six systems are being evaluated. Five of these systems are commercial MT systems; the sixth is a prototype developed at the University of Montreal by the RALI research center. Two runs will be carried out. For industrial reasons, systems will be made anonymous.

Campaign organization and schedule. Two campaigns are being organized. The first is organized using a system's default dictionary. After systems terminological adaptation, a second campaign will be organized.

Evaluation is carried out on text rather than on sentences. Text approximate length will be 400 words. The language pair referred to as the "major language pair" uses French as source language and English as target language.

Before the second campaign takes place, the systems will have to go through terminological adaptation. Since the second series of tests are being carried out on a thematically homogeneous corpus, only the thematic domain will be communicated to participants. For thematic adaptation and in order to avoid system optimization after the first series of tests, a new domain-specific 200,000-word hiding corpus will be used. The second run will start during the year 2005. Organizers have committed themselves not to publish the results between the two campaigns, and a workshop dedicated to participants will be organized between the two campaigns. After the second run, an additional three-month period will be necessary to carry out result analysis and prepare data publication and workshop organization. After result analysis and final report redaction, a public workshop will be organized. The results will be disseminated and subject to publication at the end of 2005.

The CESTA scientific committee also decided in parallel with the two campaigns to evaluate systems' capacity to process formatted texts including images and HTML tags.

Participants who do not wish to participate to this additional test have informed the scientific committee.

Contrastive evaluation. What is different about CESTA? One of the particularities of the CESTA protocol is to provide a meta-evaluation of the automated metrics used for the campaign — a kind of state of the art of evaluation metrics. The robustness of the metrics will be tested on minor language pairs through a contrastive evaluation against human judgment.

The scientific committee has decided to use Arabic>French as a minor language pair. Evaluation on the minor language pair will be performed directly on two systems and by using English as a pivotal language on the other systems. Translation through a pivotal language will then be Arabic>English>French.

Organizers are, of course, aware of the potential loss of quality provoked by the use of a pivotal language. But evaluation carried out on the minor language pair through a pivotal system will not be used to evaluate these systems themselves, but to measure metric robustness. During the tests of the first campaign, the French>English system obtaining the best ranking will be selected to be used as a pivotal system for metrics meta-evaluation.

Test tool and corpora. The required material is a set of corpora and a test tool that will be implemented according to metrics requirements and under the responsibility of CESTA organizers.

The evaluation corpus is composed of 50 texts, each 400 words long, to be translated twice, considering that a translation already exists in the original corpus. The different corpora are provided by ELRA. The masking corpus has 250,000 words and must be thematically homogeneous. Three human translations will be used for each of the 50 source texts. CESTA relies on common user use cases, and evaluation is not made in order to obtain a ready-to-publish target language translation, but rather to provide a foreign user simple access to information within the limits of basic grammatical correctness.

BLEU, BLANC and ROUGE. CESTA is based on an original protocol that aims at providing a state of the art in the field of MT evaluation. The metrics used for CESTA are referred to as BLEU, BLANC and ROUGE. Two of the metrics have already been tested: the IBM BLEU protocol and BLANC, a metric derived from a study presented at the LREC 2002 conference. We only take into account a part of the protocol described in that paper, the X score, that corresponds to grammatical correctness. The third metric, ROUGE, is a research experimental protocol developed at the University of Leeds in the United Kingdom.

In BLANC, six systems were submitted to evaluation: Candide (CD), Globalink (GL), MetalSystem (MS), Reverso (RV), Systran (SY) and XS (XS). Each of the systems was due to translate 100 source texts ranging from 250 to 300 words each. A corpus of 600 translations is thus produced — a corpus of six translations being produced automatically for each of the source texts. According to the protocol initiated by White and Forner, 2001, these series were then ranked by medium adequacy score. Every five series, a series is extracted from the whole, packs of 20 series of target translations being thus obtained and submitted to human evaluators.

Each evaluator read 10 series of six translations (60 texts). Each of these series was then read by six different evaluators who did not know that the texts were translated automatically.

Human judgment that ranks from best to worst corresponds in reality to a set of the fluency, adequacy and informativeness criteria that can be attributed to the texts translated automatically. Two scores were generated automatically — the X-score (a syntactic score) and the D-score (a semantic score). Only the X-score is referred to as the BLANC metric. The D-score, remaining unstable, had to be submitted to further study, the ROUGE metric being now a result of its reformulation.

ROUGE is an original metric based on semantic correctness. The original idea on which this protocol is based relies on the fact that MT evaluation metrics are based on comparing the distribution of statistically significant words in corpora of MT output and in human reference translation corpora.

The method used to measure MT quality is a statistical model for MT output corpora and for a parallel corpus of human translations. Each statistically significant word is highlighted in the corpus. A statistical significance score is given for each highlighted word. Then statistical models for MT target texts and human translations are compared, special attention being paid to words that are automatically marked as significant in MT outputs, whereas they do not appear to be marked as significant in human translations. These words are considered to be "over generated." The same operation is then carried out on "under-generated words." A third operation consists in the marking of the words equally marked as significant by the MT systems and the human translations. The overall difference is then calculated for each pair of texts in the corpora. Three measures specifying differences in statistical models for MT and human translations are then implemented. The first aims at avoiding "over generation"; the second aims at avoiding "under generation"; and the last is a combination of these two measures. The average scores for each of the MT systems are then computed.

Next step. CESTA results will be published in a final report. At the end of the campaign, a final workshop will be organized.

It is important to note that the CESTA campaign aims at ensuring protocol reusability to the originality of a protocol relying on three different types of measures carried out in parallel with a meta evaluation of the metrics. One of the outputs of the campaign will be the creation of an MT evaluation toolkit that will be put at users' and system developers' disposal at a reasonable price. It is expected that the toolkit will be available from ELRA. Ω

This article, expanded to include references and a diagram of the CESTA evaluation process, may be found at www.multilingual .com/dabbadieHadiTimimi65.htm