Evaluation of Translation Technology

Walter Daelemans
University of Antwerp

Véronique Hoste
University College Ghent/ University of Ghent

Lacking widely accepted and reliable evaluation measures, the evaluation of Machine Translation (MT) and translation tools remains an open issue. MT developers focus on automatic evaluation measures such as BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) which primarily count n-gram overlap with reference translations and which are only indirectly linked to translation usability and quality. Commercial translation tools such as translation memories and translation workbenches are widely used and their developers claim usefulness in terms of productivity, consistency or quality. However, these claims are rarely proven using objective comparative studies. This collection dissects the state of the art in translation technology and translation tool development and provides quantitative and qualitative answers to the question how useful translation technology is.

Evaluation of translation technology requires a multifaceted approach. It involves the evaluation of the textual output quality in terms of intelligibility, accuracy, fidelity to its source text, and appropriateness of style and register. But it also takes into account the usability of supportive tools for creating and updating dictionaries, for post-editing texts, for controlling the source language, for customization of documents, for extendibility to new languages and for domain adaptability, etc. Finally, evaluation involves contrasting the costs and benefits of translation technology with those of human translation performance.

This collection comprises 10 original contributions from researchers and developers in the field. The volume is divided into two parts. The first addresses evaluation of Machine Translation, the second evaluation of Translation Tools.

Part I opens with an invited position paper of Andy Way (*A critique of statistical machine translation*) in which he analyzes the divide between on the one hand the developers of Statistical Machine Translation (SMT) systems, and on the other hand translators. In spite of the technical success of SMT, with phrase-based SMT dominating research and development, translators largely ignore it. According to Andy Way, the reason for this is the fact that the approach is perceived as being extremely difficult to understand, as its proponents are not interested in addressing any community other than their own. After a fascinating account of the early history of

SMT, the author argues convincingly that SMT has much to learn from other paradigms, including more linguistically sophisticated ones. He also criticizes the danger of over-optimizing systems when using only automatic MT evaluation methods.

The topic of evaluation methodology is further taken up by Paula Estrella, Andrei Popescu-Belis, and Maghi King (*The FEMTI guidelines for contextual MT evaluation: principles and resources*) in their introduction to the Framework for the Evaluation of Machine Translation in ISLE (FEMTI). This methodology takes into account the context of the use of an MT system and is based on ISO/IEC standards and guidelines for software evaluation. The methodology provides support tools and helps users define contextual evaluation plans. Context in terms of tasks, users, and input characteristics indeed plays an all-important role in evaluation. The web-based FEMTI application allows evaluation experts to share and refine their knowledge about evaluation.

Despite the high correlations with human judgements (e.g. Zhang et al., 2004), automatic metrics such as BLEU and NIST do not necessarily result in an actual improvement in translation quality (Way, Callison-Burch et al., 2006). Furthermore, a limitation of current automatic scores developed within SMT is the fact that they give only a very general indication of translation quality. Both the article of Bogdan Babych and Anthony Hartley, and the contribution of Nora Aranberri-Monasterio and Sharon O'Brien focus on more fine-grained MT evaluation, aiming at a more thorough error analysis which can help MT developers to focus on problematic categories. Bogdan Babych and Anthony Hartley (*Automated error analysis for multiword expressions: using BLEU-type scores for automatic discovery of potential translation errors*) adapt the BLEU metric to allow for the detection of systematic mistranslations of multiword expressions (MWE), and also to create a priority list of problematic issues. Two aligned parallel corpora serve as the basis for their experiments and they experiment both with rule-based and statistical MT systems. They show that their approach allows for the discovery of poorly translated MWEs both on the source and target language side. Even more specific is the evaluation of output of rule-based MT systems when translating –ing forms by Nora Aranberri-Monasterio and Sharon O' Brien (*Evaluating RBMT output for –ing forms: a study of four target languages*). These forms have a reputation for being hard to translate into e.g. French, Spanish, German, and Japanese and are therefore frequently addressed in controlled language rules which seek to reduce the ambiguities in the source text in order to improve the machine translation output. For the evaluation of the translation quality of the -ing-form, the authors opted for a human evaluation and show that Systran, a rule-based MT system, obtains reasonable accuracy (over 70%) in translating this form. Due to the labour-intensive nature of human evaluation, they also assess the agreement between the human scores and automatic metrics such as NIST, GTM, etc. and show good correlations. The authors conclude on the basis of their experimental work that the problem of the -ing forms is overstated and explore a few possibilities for further improving these results.

Part I closes with yet another perspective on the evaluation of Machine Translation: recipient evaluation. This study is another nice application of the context-based evaluation of MT. In order to determine the usefulness of MT as a cost-effective way of providing more material in the language of minorities, Lynne Bowker (*Can Machine Translation meet the needs of official language minority communities in Canada? A recipient evaluation.*) investigates the reception of MT in the Canadian context where bilingualism is officially legislated. The reception of MT output by the two studied Official Language Minority Communities (OLMCs) was investigated by presenting four translation versions, viz. human translations and raw, rapidly post-edited and maximally post-edited MT output to members of the two OLMCs. Bowker's study reveals that whereas (rapidly and maximally post-edited) MT output could be acceptable for information assimilation in cases where there is a lack of ability to understand the source text, only high-quality translations are acceptable for information dissemination where translation is seen as a means for preserving or promoting a culture. Another interesting finding was that the 'average' recipients are more open to MT output than language professionals.

Part II of this volume addresses the evaluation of computer-aided translation tools (see e.g. Bowker, 2002 for an introduction). These tools include Translation Memories (TM), (bilingual) terminology management software, monolingual authoring tools (spelling, grammar, style checking), workflow management tools etc. A first question to be answered is whether current state of the art tools are perceived as useful by translators, and how they can be improved. Iulia Mihalache (*Social and economic actors in the evaluation of translation technologies. Creating meaning and value when designing, developing and using translation technologies*) discusses the advantages for companies as well as for translators of encouraging public evaluation of tools in on-line communities, and develops evaluation criteria from the perspective of translators communities, making use of different technology adoption models. She also discusses the 'how' of evaluation: a more complete understanding of translation technologies evaluation criteria is obtained if translators' attitudes, perceptions and behaviours related to technologies are studied in a multidisciplinary way from sociological, economic, psychological, and cultural perspectives. Alberto Fernández Costales (*The role of computer assisted translation in the field of software localization*) analyzes the effectiveness of computer assisted translation tools in Localization, the adaptation of a product to a particular locale. By empirically studying the usability and reliability of a particular tool (Passolo) for localizing a program, insight is provided into how translation tools can alleviate some of the challenges of localization. Besides improving text consistency and terminological coherence (but see Miguel Jiménez-Crespo's paper for contradictory results), the main advantage is that these tools can save time, and thereby improve the productivity of localization experts.

Possible improvements in current Translation Memory technology are studied in the article of Lieve Macken (*In search of recurrent units of translation*). Translation Memories are currently sentence-based. This means that new text to be translated can only be matched with sentence-like segments, leading to limited recall in many cases. However, the number of matches can be increased if input is allowed to match sub-sentential segments. In a series of experiments, the degree of repetitiveness of different text types is compared, and performance of a sentential Translation Memory system is compared with a sub-sentential one. The results show that whereas sub-sentential memory systems are certainly a move in the right direction, they also sometimes lead to distracting translation suggestions. For solving the latter problem, better word alignment algorithms are necessary.

TM tools have changed the nature of translation by imposing a number of technological constraints that can in principle lead to either positive results (increased consistency) or negative results (increased decontextualization). Miguel Jiménez-Crespo (*The effect of translation memory tools in translated web texts: evidence from a comparative product-based study*) provides an empirical study on the often-debated question whether TMs improve or degrade translation quality. In a corpus-based study of 40,000 original and localized Spanish websites, he shows that the localized texts (translated using TMs) show higher numbers of inconsistencies at the typographic, lexical, and syntactic levels than spontaneously produced, non-translated texts, and therefore lead to lower levels of quality. While this article does not provide the last word in this discussion, it paves the way to interesting follow-up studies controlling for different variables that may influence the difference observed.

Acknowledgements

The authors would like to take this opportunity to thank all the authors for their contributions. The final contributions have undergone a detailed review followed by a thorough revision step. Our sincere thanks also go to the reviewers who helped us to assure the highest level of quality for this publication: Joost Buysschaert, Gloria Corpas Pastor, Alian Desilets, Andreas Eisele, Frederico Gaspari, David Farwell, Eva Forsbom, Johann Haller, David Langlois, Lieve Macken, Karolina Owczarzak, Jörg Tiedemann, Harold Somers. We also thank Aline Remael for her advice throughout the publication process and for some of the final formal editing with Jeremy Schreiber.

Bibliography

Bowker, L. (2002). Computer Aided Translation Technology: A Practical Introduction, University of Ottawa Press, Ottawa, Canada.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the Role of Bleu in Machine Translation Research. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (pp.249-256).Association for Computational Linguistics. Trento, Italy.

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics. *Proceedings of the Second Human Language Technologies Conference (HLT)* (pp.138-145). Morgan Kaufmann. San Diego, USA.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.J. (2002). BLEU: a method for automatic evaluation of Machine Translation. *Proceedings of the 40th Annual Metting of the Association for Computational Linguistics (ACL)* (pp.311-318). Association for Computational Linguistics. Philadelphia, USA.

Zhang, Y., Vogel, S. and Waibel, A. (2004). Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.(pp.2051-2055). European Language Resources Association. Lisbon, Portugal.