DOI: 10.1145/1562764.1562798

BY FRANCISCO CASACUBERTA, JORGE CIVERA, ELSA CUBEL, ANTONIO L. LAGARDA, GUY LAPALME, ELLIOTT MACKLOVITCH, AND ENRIQUE VIDAL

Human Interaction For High-Quality Machine Translation

TRANSLATION FROM A SOURCE LANGUAGE INTO A target language has become a very important activity in recent years, both in official institutions (such as the United Nations and the EU, or in the parliaments of multilingual countries like Canada and Spain), as well as in the private sector (for example, to translate user's manuals or newspapers articles). Prestigious clients such as these cannot make do with approximate translations; for all kinds of reasons, ranging from the legal obligations to good marketing practice, they require target-language texts of the highest quality. The task of producing such high-quality translations is a demanding and time-consuming one that is generally conferred to expert human translators. The problem is that, with growing globalization, the demand for high-quality translation has been steadily increasing, to the point where there are just not enough qualified translators available today to satisfy it. This has dramatically raised the need for improved machine translation (MT) technologies.

The field of MT has undergone something of a revolution over the last 15 years, with the adoption of empirical, data-driven techniques originally inspired by the success of automatic speech recognition.[10] Given the requisite corpora, it is now possible to develop new MT systems in a fraction of the time and with much less effort than was previously required under the formerly dominant rule-based paradigm. As for the quality of the translations produced by this new generation of MT systems, there has also been considerable progress; generally speaking, however, it remains well below that of human translation. No one would seriously consider directly using the output of even the best of these systems to translate a CV or a corporate Web site, for example, without submitting the machine translation to a careful human revision. As a result, those who require publication-quality translation are forced to make a diffcult choice between systems that are fully automatic but whose output must be attentively post-edited, and computer-assisted translation systems (or CAT tools for short)[7] that allow for high quality but to the detriment of full automation.

Currently, the best known CAT tools are translation memory (TM) systems. These systems recycle sentences that have previously been translated, either within the current document or earlier in other documents. This is very useful for highly repetitive texts, but not of much help for the vast majority of texts composed of original materials.

Since TM systems were first introduced, very few other types of CAT tools have been forthcoming. Notable exceptions are the TransType system[6] and its successor TransType2 (TT2).[4] These systems represent a novel rework-ing of the old idea of interactive machine translation (IMT). Initial efforts on TransType are described in detail in Foster;[5,6] suffice it to say here the system's principal novelty lies in the fact the human-machine interaction focuses on the drafting of the target text, rather than on the disambiguation of the source text, as in all former IMT systems.

In the TT2 project, this idea was further developed. A full-fledged MT engine was embedded in an interactive editing environment and used to generate suggested completions of each target sentence being translated. These completions may be accepted or amended by the translator; but once validated, they are exploited by the MT engine to produce further, hopefully improved suggestions. This is in marked contrast with traditional MT, where typically the system is first used to produce a complete draft translation of a source text, which is then post-edited (corrected) offline by a human translator. TT2's interactive approach offers a significant advantage over traditional post-editing. In the latter paradigm, there is no way for the system, which is off-line, to benefit from the user's corrections; in TransType, just the opposite is true. As soon as the user begins to revise an incorrect segment, the system immediately responds to that new information by proposing an alternative completion to the target segment, which is compatible with the prefix that the user has input.

Another notable feature of the work described in this article is the importance accorded to a formal treatment of human-machine interaction, something that is seldom considered in the now-prevalent framework of statistical pattern recognition.

Interactive Machine Translation

We start with an illustrative example of how a TT2 IMT system works (see Figure 1) before presenting a more formal description.

Let us suppose that a source English sentence $\mathbf{s}$ = *"Click OK to close the print dialog"* is to be translated into a target Spanish sentence $\mathbf{t}$. Initially, with no user information ($\mathbf{t}_p = \lambda$), the system provides a complete translation suggestion ($\mathbf{t}_s$ = *"Haga clic para cerrar el diálogo de impresión"*).

From this translation, the user marks a prefix as correct ($\mathbf{a}$="*Haga clic*") and begins to type the rest of the target sentence. Depending on the system or the user's preferences, the new input $\mathbf{k}$ can be the next word or some letters from it (in our example $\mathbf{k}$ is the next correct word "*en*"). A new target prefix $\mathbf{t}_p$ is then defined by the previously validated prefix together with the new input the user has just typed ($\mathbf{t}_p$ = *"Haga clic en"*).

The system then generates a new suffix $\mathbf{t}_s$ to complete the translation: *ACEPTAR para cerrar el diálogo de impresión."* The interaction continues with a new validation followed, if necessary, by new input from the user, and so on, until such time as a complete and satisfactory translation is obtained.

This type of interactive translation process can be easily formalized within the elegant statistical framework for machine translation first pioneered by Brown et al.[2] In this framework, translations are generated on the basis of statistical and information-theoretic models whose parameters are automatically derived ("trained") from the analysis of bilingual text corpora.

More formally, we are given a sentence s in a source language and the system has to find a best translation in a target language. Using statistical decision theory, a best translation is a target-language sentence, $\hat{t}$, which is most probable given the source sentence:

(1)

$$\hat{t} = \arg\max_{t} \Pr(\mathbf{t} \mid s) = \arg\max_{t} \Pr(\mathbf{s}, \mathbf{t}).$$

Different models have been proposed to approach one or the other of these probabilistic distributions, from *statistical (word or phrase-based) alignment models* (SAM)[2] for the conditional distribution, to *stochastic finite-state transducers* (SFST)[3] for the joint distribution. In the TT2 project, both SFST and SAM were deployed, although in this article we focus on the results obtained with SFST. In this case, the translation of a new source sentence, as given by equation (1), is carried out by searching for an optimal path in a weighted graph representing all possible translations of the source sentence.[3] SFST lend themselves well to the real-time requirements of IMT.

In the TT2 project, we developed and tested translation models for English, Spanish, French, and German (with English as the pivot). Needless to say, given the requisite training corpora, the formalism can also be extended to other languages, although translation results generally tend to be poorer between languages belonging to different families, such as Arabic or Chinese.

In the IMT framework, we need to take into account the corrections provided by the translator in the form of a *validated translation prefix*, $\mathbf{t}_p$. Conse-

Figure 1.

ITER-0	$\mathbf{t}_p$	λ
ITER-1	$\hat{\mathbf{t}}_s$	*Haga clic para cerrar el diálogo de impresión*
	a	Haga clic
	k	**en**
	$\mathbf{t}_p$	Haga clic en
ITER-2	$\hat{\mathbf{t}}_s$	*ACEPTAR para cerrar el diálogo de impresión*
	a	*ACEPTAR para cerrar el*
	k	**cuadro**
	$\mathbf{t}_p$	Haga clic en ACEPTAR para cerrar el cuadro
FINAL	$\hat{\mathbf{t}}_s$	*de diálogo de impresión*
	a	*de diálogo de impresión*
	k	#
	t	Haga clic en ACEPTAR para cerrar el cuadro de diálogo de impresión

An example of keyboard interaction with a system like TT2 used to translate the English sentence "Click OK to close the print dialog" into Spanish. System suggestions are printed in italics (blue) and user input in boldface type¬writer font (red). In the final line, text that was typed by the user is underlined.

quently, rather than a full translation, the system must produce a target sentence suffix, $\mathbf{t}_s$ that best completes the user prefix (see Figure 1). The problem stated in equation (1) therefore needs to be reformulated as follows:

(2)

$$\hat{\mathbf{t}}_s = \underset{\mathbf{t}_s}{\operatorname{argmax}} \Pr(\mathbf{t}_s \,|\, \mathbf{s}, \mathbf{t}_p) = \underset{\mathbf{t}_s}{\operatorname{argmax}} \Pr(\mathbf{s}, \mathbf{t}_p\mathbf{t}_s).$$

Since $\mathbf{t}_p\mathbf{t}_s = \mathbf{t}$, the same models as for equation (1) can be used in the IMT case, but now $\mathbf{t}_p$ is given and the search problem needs to be modified to operate over the set of suffixes that complete the given user prefix.[1]

In the first iteration, the system can actually generate a *word-graph* which represents a huge subset of all the possible translations of the source sentence. In each successive human-machine iteration, the corresponding consolidated prefix $\mathbf{t}_p$ constrains the search space to the subset of paths in the word graph whose prefix matches the $\mathbf{t}_p$ provided by the user. Note that tp may not actually be found in the word graph, in which case an error-correcting matching technique must be used.[1]

System Evaluation

One of the reasons that MT evaluation poses a challenging problem is the absence of a unique gold standard to which system translations can be compared. The same sentence can often be translated in different ways, all of which convey the same meaning. In contrast, this problem does not exist in other fields like speech recognition or text categorization. This peculiarity of MT (and IMT alike) has sparked some original research on the development of automatic and manual evaluation metrics. Automatic metrics, based on bilingual corpora, are particularly useful in providing rapid and inexpensive feedback about the performance of the system during its development phase; but if the goal is to assess the anticipated impact of an MT or a CAT system on its intended end-users, nothing can replace a bona fide usability study.

Corpora. Statistical MT is based on the "Rosetta Stone" approach to translation, which is to say that the sole source of translation knowledge is a set of bilingual sentences. It is therefore not surprising that translation quality should be cor-related with the amount of available bilingual training data. Depending on the particular language pairs, large parallel corpora can sometimes be obtained from international organizations or governments, although their compilation and preprocessing usually demand a non-negligible amount of work.

The evaluation presented here was carried out on the so-called Xerox corpora,[4] comprised of user manuals for Xerox printers and photocopiers. In each case, English was the source language of the manual and the reference translations into French, Spanish, and German were kindly provided by the company's language services. For each language pair, about 50,000 sentences and their corresponding translations were used to train a translation model, while 1,000 sentences were reserved for the automatic evaluation of the IMT system.

Automatic evaluation. We compared the translation of the source test sentences produced by our translation engine with the corresponding target reference sentence and then com-puted evaluation figures, as described below. The aim was to *estimate* the effort that a human translator would require to produce a correct translation using the output of the TT2 system. In order to estimate this effort, we define the ratio between the number of *keystrokes* needed to achieve the reference target sentence and the number of characters in the reference sentence. Basically, this figure boils down to the ratio between the number of characters a translator would need to type with and without a IMT system. To this end, the target translation that a real user would have in mind when translating a sentence is simulated by the single reference translation.

On the test corpus, key-stroke ratios as low as 20-25% were obtained using our SFST-based suffix-predictive IMT system to translate between English and Spanish.[1] In the other language pairs involving French and German, the estimated key-stroke ratios were somewhat higher (approximately 45%), which presumably reflects a greater variability of style in the Xerox translations for these languages.

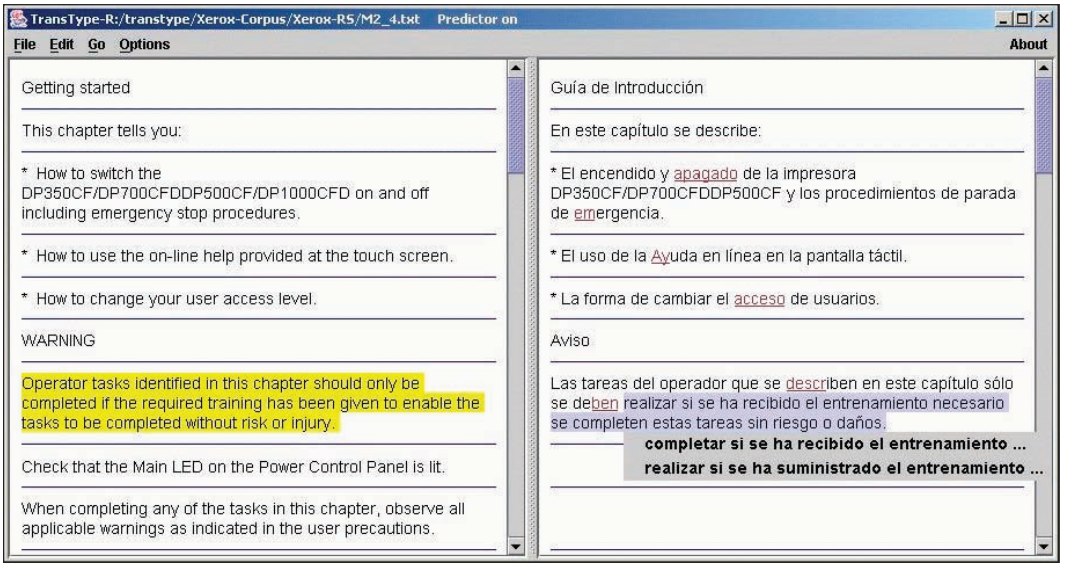
Human evaluation. The results of the automatic evaluation metrics discussed above were intended to give us a rough idea of how the system could be expected to perform when used by real translators. The obvious next step was to assess this behavior un-der laboratory-controlled, though realistic working conditions. One of the more intuitive metrics that has been proposed for evaluating IMT systems[8] is to measure the overall time required to translate a test corpus, including the time it takes the user to read and evaluate the system's proposed translations, in addition to all her interactions with the CAT system. Hence, in our user trials, we equipped TranTypes GUI with a system clock, which allowed us to precisely measure the time it took the trial participants to complete the translations, both with and without the benefit of the system's proposed completions. The participants in these user trials were six professional translators, recruited from the two translation agencies that participated in the TT2 project. A snapshot of a typical TT2 session is shown in Figure 2.

Productivity results. Five rounds of user trials were organized during the final eighteen months of the TT2 project. The first rounds were essentially intended to train the participants on the new system and to provide the developers with important feedback on its user interface—a critical point in an interactive system. The last three rounds were more production-oriented, and saw the participants working with the system for ten consecutive half-day sessions. The texts used for these trials were all drawn from the Xerox corpus described here.

In order to adequately assess the contribution of the system's proposed completions, each trial round included at least one *dry-run* session, during which the participants were asked to translate a chapter of the test corpus on their own, such as using the same text editor but without the benefit of the system's predictions. These dry-run sessions provided us with baseline productivity figures against which we could then compare the participants' productivity on the same technical manuals but translated with the help of the system's proposed completions.

The results varied from one round to the next, but, generally speaking, productivity tended to increase over the 18-month period, as the participants grew accustomed to translating with this new tool. On some rounds, partic-

Figure 2.



Snapshot of a TT2 working session (English to Spanish translation). Text in normal typeface derives from TT2's predictions, while text manually typed by the translator is shown in light (red and underlined) font. In the last translation segment, the current best TT2 prediction appears inline in shaded text; the popup shows two alternative completions.

ularly near the end of the project, the users registered some very substantial productivity gains; on the penultimate round, for example, the six participants bettered their dry-run productivity on that round by an average of almost 30% using IMT SFST models (similar productivity gains were achieved using other MT models). On the final round, however, similar gains were all but precluded owing to the inadvertent selection of a particularly easy text for the dry-run. (For full details on the TT2 user trials see Macklovitch.[9]) Overall, it seems fair to conclude that a suffix-predictive IMT system like TT2 can allow translators to increase their productivity while maintaining high-quality; and while the productivity gains afforded by this approach may not be spectacular, they are certainly substantial.

Conclusion

Our approach could be called *human-centered machine translation*, and by this we mean not just that the human translator remains in the production loop, but that he or she is at the very center of a process that aims to produce high-quality, automated translations. As developers of CAT technology, we take the kind of criticisms expressed by the participants in our user trials very seriously. Hence, a major component of our future work on interactive MT will be to study their principal complaint regarding the system's inability to learn from the revisions they made to its output, in order to improve the quality of subsequent predictions.

Furthermore, user behavior has suggested that productivity can be significantly improved by allowing interaction modalities other than the keyboard and mouse. In this direction, multi-modal systems involving the use of speech interaction are proposed and studied in Vidal et al.[11] with encouraging results.

References

1. Barrachina, S., Bender, et al. Statistical approaches to computer-assisted translation. *Computational Linguistics 35*, 8 (2009), 3-28.
2. Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics 19*, 2 (1993), 263–310.
3. Casacuberta, F., and Vidal, E. Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics 30*, 2 (2004), 205–225.
4. Esteban, J., Lorenzo, J., Valderrabanos, A. S., and Lapalme, G. TransType2 - -An innovative computer-assisted translation system. In *The Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics*, (Barcelona, Spain, July 2004), 94-97.
5. Foster, G. Text Prediction for Translators. PhD thesis, Université de Montreal, May 2002.
6. Foster, G., Isabelle, P., and Plamondon, P. Target-text mediated interactive machine translation. *Machine Translation 12*, 1–2, (1997), 175–194.
7. Isabelle, P. and Church, K. Special issue on new tools for human translators. *Machine Translation 12*, 1–2, 1997.
8. King, M., Popescu-Belis, A., and Hovy, E. FEMTI: creating and using a framework for MT evaluation. In *Proceedings of the Machine Translation Summit IX*, 224–231, (Sept 2003), New Orleans.
9. Macklovitch, E. TransType2: The last word. In *Proceedings of the 5th International Conference on Languages Resources and Evaluation*, (May 2006, Genoa, Italy), 167-172.
10. Ney, H. One decade of statistical machine translation: 1996-2005. In *Proceedings of the MT Summit X*, (Sept. 2005, Phuket, Thailand), 12-17.
11. Vidal, E., Casacuberta, F., Rodríguez, L, Civera, J., and Martínez, C. Computer-assisted translation using speech recognition. *IEEE Transactions on Speech and Audio Processing, 14*, 3, (2006), 941–951.

Francisco Casacuberta (fcn@iti.upv.es) is a professor of Computer Science at Universidad Politécnica de Valencia, Spain and coordinator of the Master in Artificial Intelligence, Pattern Recognition and Digital Image.

Jorge Civera (jorcisai@iti.upv.es) is an assistant professor of Computer Science at the Universidad Politécnica de Valencia, Spain.

Elsa Cubel (ecubel@iti.upv.es) is a Ph.D. candidate in the Instituto Tecnológico de Informática at the Universidad Politécnica de Valencia, Spain.

Antonio L. Lagarda (alagarda@iti.upv.es) is a Ph.D. candidate in the Instituto Tecnológico de Informática at the Universidad Politécnica de Valencia, Spain.

Guy Lapalme (lapalme@iro.umontreal.ca) is professor of Computer Science at Université de Montréal and a co-founder of the RALI (rali.iro.umontreal.ca) specialized in natural language processing.

Elliott Macklovitch (macklovi@iro.umontreal.ca) is Coordinator of the RALI Laboratory at Université de Montréal, and former President of the Association for Machine Translation in the Americas (AMTA).

Enrique Vidal (evidal@iti.upv.es) is a Professor of Computer Science at the Universidad Politécnica de Valencia and leader of one of the development teams of the TransType-2 project.