User Evaluation of Interactive Machine Translation Systems

Vicent Alabau, Luis A. Leiva, Daniel Ortiz-Martínez, Francisco Casacuberta
ITI/DSIC – Universitat Politècnica de València
{valabau,luileito,dortiz,fcn}@{iti,dsic}.upv.es

Abstract

Recent developments in search algorithms and software architecture have enabled multi-user web-based prototypes for Interactive Machine Translation (IMT), a technology that aims to assist, rather than replace, the human translator. Surprisingly, formal human evaluations of IMT systems are highly scarce in the literature. To this regard, we discuss experiences gained while testing IMT systems. We report the lessons learned from two user evaluations. Our results can provide researchers and practitioners with several guidelines towards the design of on-line IMT tools.

1 Introduction

Research in machine translation (MT) aims to develop computer systems which are able to translate documents without human intervention. However, current translation technology has not been able to deliver full automated error-free translations. Typical solutions to improve the quality of an MT system require manual post-editing. This serial process does not allow integrating the knowledge of the human translator into the system decisions.

One alternative to take advantage of the existing MT technologies is to apply the so-called interactive machine translation (IMT) paradigm (Langlais et al., 2002). The IMT paradigm adapts data driven MT techniques for its use in collaboration with human translators. Following these ideas, Barrachina et al. (2009) proposed a new approach to IMT, in which fully-fledged statistical MT systems are used to produce full target sentences hypotheses, or portions thereof, which can be accepted or amended by a human translator. Each corrected text segment is then used by the MT system as additional information to achieve improved suggestions. Figure 1 shows a minimal IMT session example.

© 2012 European Association for Machine Translation.

source: Para ver la lista de recursos
reference: To view a listing of resources

suggestion	**s**	*To view the resources list*
interaction	**p**	To view
	k	**a**
	s	*listing of resources*
accept	**p**	To view a listing of resources

Figure 1: An IMT session example, using only 1 key stroke (k) to achieve the reference sentence. Notice that the user submits partial sentences (**p**) to the system, which tries to complete them (**s**).

Following the IMT paradigm, recent developments in search algorithms and software architecture have allowed multi-user web-based translation prototypes. These systems have grown in features, e.g., allowing advanced multimodal interaction, which have also added extra complexity to the prototypes. Then, their effectiveness should be tested with respect to technology dissemination. While pure data-driven evaluations have already shown that IMT is a promising technology (Barrachina et al., 2009), surprisingly, formal human evaluations are highly scarce in the literature.

In this paper, we describe our experiences evaluating two IMT prototypes with real users: an initial, advanced version and a simplified but improved version. Our results identify important design issues, which open a discussion regarding how IMT systems should be deployed.

2 Related Work

Langlais et al. (2002) performed a human evaluation on their IMT prototype. They emulated a realistic working environment in which the users could obtain automatic completions for what they were typing. Users reported an improvement in performance; however, raw productivity decreased by 17%, although the users appreciated the tool and were confident to improve their productivity after proper training. That work was extended in the TT2 project (Casacuberta et al., 2009), where the

performance tended to increase as the participants grew accustomed to the system, over a 18-month period. A slightly different approach was studied in (Koehn, 2010). There, monolingual users evaluated a translation interface supporting IMT predictions and the so-called 'translation options'. When translating from undecipherable languages (as Chinese or Arabic for an English speaker), richer assistance improved user performance.

3 User Interfaces and Evaluation

Previous research on multimodal interfaces in natural language processing have shown a comprehensible tendency to choose an interactive collaborative environment over a manual system for non-expert computer users (Leiva et al., 2011). We followed this approach to build a prototype with an IMT backend. We will refer to this system as the advanced demonstrator (IMT-AD, Figure 2) since it implemented a number of complementary features, which conditioned the design of the interface; e.g., the use of one boxed text field per sentence word aimed to ease e-pen interaction.

3.1 Evaluation of the Advanced Prototype

The goal of this evaluation was aimed to assess both qualitatively and quantitatively IMT-AD, and compare it to a state-of-the-art post-editing (PE) MT output. Translating from scratch was not considered since this practice is being increasingly displaced by assistive technologies. Indeed, PE of MT systems is found frequently in a professional translation workflow (TT2, 2001). Thus, in addition to IMT-AD, a post-editing version of the demonstrator (PE-AD) was developed to make a fair comparison with state-of-the-art PE systems. PE-AD used the same interface as IMT-AD, but the IMT engine was replaced by autocompletion-only capabilities as found in popular text editors.

Design Both systems were evaluated on the basis of the ISO 9241-11 standard (ergonomics of human-computer interaction). Three aspects were considered: efficiency, effectiveness, and user satisfaction. For the former, we computed the average time in seconds that took to complete each translation. For the second, we evaluated the BLEU against the reference and a crossed multi-BLEU among users' translations. For the latter, we adapted the system usability scale (SUS) questionnaire to score the user satisfaction, by asking 10 questions that users would assess in a 1–5 Likert scale (1:strongly disagree, 5:strongly agree), plus a text area to submit free-form comments.

Participants A group of 10 users (3 females) aged 26–43 from our research group volunteered to perform the evaluation as non-professional translators. All of them were proficient in Spanish and had an advanced knowledge of English. Although none had worked with IMT systems, all knew the basis of the IMT paradigm.

Apparatus Since participants were Spanish natives, we decided to perform translations from English to Spanish. We chose a medium-sized corpus, the EU corpus, typically used in IMT (Barrachina et al., 2009), which consists of legal documents. We built a glossary for each source word by using the 5-best target words from a word-based translation model. We expected this would cover the lack of knowledge for our non-expert translators towards this particular task. In addition, a set of 9 keyboard shortcuts was designed, aiming to simulate a real translation scenario, where the mouse is typically used sparingly. Furthermore, autocompletion was added to PE-AD, i.e., words with more than 3 characters were autocompleted using a task-dependent word list. In addition, IMT-AD was set up to predict at character level interactions. We disabled the complementary features to focus the evaluation on basic IMT.

Procedure Three disjoint sentence sets (C1, C2, C3) were randomly selected from the test dataset. Each set consisted of 20 sentence pairs and kept the sequentiality of the original text. Sentences longer that 40 words were discarded. C3 was used in a warm up session, where users gained experience with the IMT system (5–10 min per user on average) before carrying out the actual evaluation. Then, C1 and C2 were evaluated by two user groups (G1, G2) in a counterbalanced fashion: G1 evaluated C1 on PE-AD and C2 on IMT-AD, while G2 did C1 on IMT-AD and C2 in PE-AD.

Results Although the results were not conclusive (there were no statistical differences between groups), we observed some trends. First, the time spent (efficiency) per sentence on average in the IMT system was higher than in PE (67 vs. 62 s). However, the effectiveness was slightly higher for IMT in BLEU with respect to the reference (41.5 vs. 40.7) and with respect to a cross-validation with other user translations (78.9 vs. 77.4). This

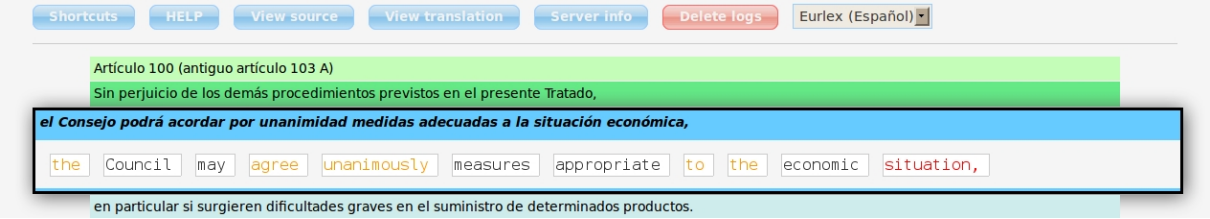

Figure 2: Detail of the advanced web-based interface with a boxed text field for each word.

	PE-AD	IMT-AD
Avg. time (s)	62 ($SD = 51$)	67 ($SD = 65$)
BLEU	40.7 (13.4)	41.5 (13.5)
Crossed BLEU	77.4 (4.5)	78.9 (4.8)
Global Satisfaction	2.5(1.2)	2.1(1.2)

Table 1: Summary of the results for the first test.

suggested that the IMT system helped to achieve more consistent and standardized translations.

Finally, users perceived the PE system more satisfactorily than the IMT system, although the global scores were 2.5 for PE and 2.1 for IMT, which suggested that users were not comfortable with none of the systems. IMT failed to succeed in questions regarding the system being easy to use, consistent, and reliable. This was corroborated by the submitted comments. Users complained about having too many shortcuts and available edit operations, some operations not working as expected, the word-box based interface, and some annoying common mistakes in the predictions of the IMT engine (e.g., inserting a whitespace instead of completing a word, which would be interpreted as two different words). One user stated that the PE system "was much better than the [IMT] predictive tool". Regarding PE, users mainly questioned the usefulness of the autocompletion feature.

3.2 Simplified Web Based Prototype

The results from the first evaluation were quite disappointing. Not only participants took more time to complete the evaluation with IMT-AD, but they also perceived that IMT-AD was more cumbersome and unreliable than PE-AD. However, we still observed that IMT-AD had been occasionally beneficial, and probably the bloated UI was the cause for IMT to fail. Thus, we developed a simplified version of the original prototype (Figure 3).

Design In this case, the word-box based interface was changed to a simple text area. In addition, the edit operations were simplified to allow only word substitutions and single-click rejections. Besides, we expected that the simplification of the interface logic would reduce some of the programming bugs that bothered users in the first evaluation. The PE interface was simplified in the same way. Furthermore, the autocompletion feature was improved to support n-grams of arbitrary length.

Participants Fifteen participants aged 23–34 from university English courses (levels B2 and C1 from the Common European Framework of Reference for Languages) were paid to perform the evaluation (5 € each). A special price of 20 € was given to the participant who would contribute with the most useful comments about both prototypes. It was found that, following this method, participants were more verbose when providing feedback.

Apparatus In this case, a different set of sentences (C1′, C2′, C3′) was randomly extracted from the EU corpus.

Procedure To avoid the bias regarding which system was being used, sentences were presented in random order, and the type of system was hidden to the participants. As a consequence, users could not evaluate each system independently. Therefore, a reduced questionnaire with just two questions was shown on a per-sentence basis. **Q1** asked if the system suggestions were useful. **Q2** asked if the system was cumbersome to use. A text area for free-form comments was also included.

Results Still with no statistical significance, we found that the IMT prototype was perceived now better than PE. First, interacting with IMT was more efficient than with PE on average (55 s vs. 69 s). The number of interactions was also lower (79 vs. 94). Concerning user satisfaction, the IMT system was perceived as more helpful (3.5 vs. 3.1) but also more cumbersome (3.1 vs. 2.9). However, in this case the differences were narrower. On the

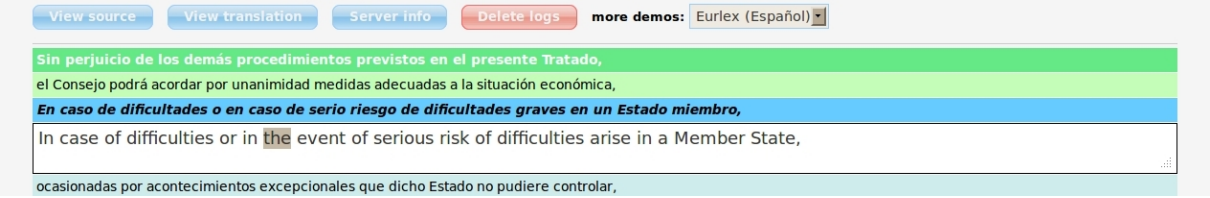

Figure 3: Detail of the simplified web-based interface.

	PE-BD	**IMT-BD**
Avg. time (s)	69 ($SD = 42$)	55 ($SD = 37$)
No. interactions	94 (60)	79 (55)
Q1 (Likert scale)	3.1 (1.2)	3.5 (1.1)
Q2 (Likert scale)	2.9 (1.2)	3.1 (1.3)

Table 2: Summary of results for the second test.

other hand, IMT received 16 positive comments whereas PE received only 5. Regarding negative comments, the counts were 35 (IMT) and 31 (PE). While the number of negative comments is similar, there was an important difference regarding the positive ones. Finally, the users' complaints of the IMT system can be summarized in the following items: *a)* system suggestions changed too often, offering very different solutions; *b)* while correcting one mistake, subsequent words that were correct were changed by a worse suggestion; *c)* system suggestions did not keep gender, number, and time concordance; *d)* if the user goes back in the sentence and performs a correction, parts of the sentence already corrected were not preserved on subsequent system suggestions.

4 Discussion and Conclusions

Our initial UI performed poorly when tested with real users. However, when the UI design was adapted to the users' expectations, the results were encouraging. Note that in both cases the same IMT engine was evaluated under the hood. This fact remarks the importance of the UI design when evaluating a highly interactive system as IMT is.

The literature had reported good experimental results in simulated-user scenarios, where IMT is focused on optimizing some automatic metric. However, user productivity is strongly related to how the user interacts with the system and other UI concerns. For instance, a suggestion that changes on every key stroke might obtain better automatic results, whereas the user productivity decreases because of the cognitive effort needed to process those changes. Therefore, a new methodology is required for optimizing interactive systems (like IMT) towards the user.

In sum, the following issues should be addressed in an IMT system: *1)* user corrections should not be modified, since that causes frustration; *2)* system suggestions should not change dramatically between interactions, in order to avoid confusing the user; *3)* the system should propose a new suggestion only when it is sure that it improves the previous one.

We hope these considerations will reduce the gap between translators and researchers needs, so that future developments can have an impact on the translation industry.

Acknowledgments

This research has received funding from the EC's 7th Framework Programme (FP7/2007-2013) under grant agreement No. 287576 - CasMaCat, and from the Spanish MEC/MICINN under the MIPRCV project (CSD2007-00018). We would also like to thank the participants and the Centro de Lenguas at the UPV.

References

Barrachina, S., O. Bender, F. Casacuberta, J. Civera, E. Cubel, S. Khadivi, A. L. Lagarda, H. Ney, J. Tomás, and E. Vidal. 2009. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

Casacuberta, F., J. Civera, E. Cubel, A. L. Lagarda, G. Lapalme, E. Macklovitch, and E. Vidal. 2009. Human interaction for high quality machine translation. *Communications of the ACM*, 52(10):135–138.

Koehn, P. 2010. Enabling Monolingual Translators: Post-Editing vs. Options. In *Proc. ACL-HLT*.

Langlais, P., G. Lapalme, and M. Loranger. 2002. TRANSTYPE: Development-Evaluation Cycles to Boost Translator's Productivity. *Machine Translation*, 15(4):77–98.

Leiva, L. A., V. Romero, A. H. Toselli, and E. Vidal. 2011. Evaluating an Interactive-Predictive Paradigm on Handwriting Transcription: A Case Study and Lessons Learned. In *Proc. COMPSAC*.

TT2. 2001. TransType2 - Computer Assisted Translation. Project Technical Annex. Information Society Technologies (IST) Programme, IST-2001-32091.