The Prague Bulletin of Mathematical Linguistics
NUMBER 108 JUNE 2017 97–108

Comparative Human and Automatic Evaluation of Glass-Box and Black-Box Approaches to Interactive Translation Prediction

Daniel Torregrosa, Juan Antonio Pérez-Ortiz, Mikel L. Forcada

Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, E-03690 Sant Vicent del Raspeig, Spain

Abstract

Interactive translation prediction (ITP) is a modality of computer-aided translation that assists professional translators by offering context-based computer-generated continuation suggestions as they type. While most state-of-the-art ITP systems follow a glass-box approach, meaning that they are tightly coupled to an adapted machine translation system, a black-box approach which does not need access to the inner workings of the bilingual resources used to generate the suggestions has been recently proposed in the literature: this new approach allows new sources of bilingual information to be included almost seamlessly. In this paper, we compare for the first time the glass-box and the black-box approaches by means of an automatic evaluation of translation tasks between related languages such as English–Spanish and unrelated ones such as Arabic–English and English–Chinese, showing that, with our setup, 20%–50% of keystrokes could be saved using either method and that the black-box approach outperformed the glass-box one in five out of six scenarios operating under similar conditions. We also performed a preliminary human evaluation of English to Spanish translation for both approaches. On average, the evaluators saved 10% keystrokes and were 4% faster with the black-box approach, and saved 15% keystrokes and were 12% slower with the glass-box one; but they could have saved 51% and 69% keystrokes respectively if they had used all the compatible suggestions. Users felt the suggestions helped them to translate faster and easier. All the tools used to perform the evaluation are available as free/open–source software.

1. Introduction

Translation technologies such as machine translation (MT) (Hutchins and Somers, 1992) or translation memories (TM) (Somers, 2003) are frequently used by professional

Corresponding author: dtorregrosa@dlsi.ua.es
Cite as: Daniel Torregrosa, Juan Antonio Pérez-Ortiz, Mikel L. Forcada. Comparative Human and Automatic Evaluation of Glass-Box and Black-Box Approaches to Interactive Translation Prediction. The Prague Bulletin of Mathematical Linguistics No. 108, 2017, pp. 97–108. doi: 10.1515/pralin-2017-0012.

translators to produce a first, usually inadequate suggestion of a target-language equivalent of a source-language sentence. The suggestion is then modified by the professional translator by rearranging or accepting parts of it, or by introducing new words when an appropriate equivalent fragment is not present; this can be perceived as a process in which the computer outputs the translation, and then the professional translator fixes the mistakes (if using MT) or the mismatches (if using TM). This paper focuses however on a different translation technology approach: *interactive translation prediction* (ITP), a human–computer collaborative approach in which computer-generated translation suggestions are offered as the professional translator carries out the translation of the source-language sentence.

The TransType project (Langlais et al., 2000), and its continuation, the TransType2 project (Macklovitch, 2006) were the pioneers of ITP. An automatic best-scenario evaluation with in-domain corpora (Barrachina et al., 2009) showed that it might theoretically be possible to save between 55% and 80% of the keystrokes in comparison with unassisted translation. A number of projects continued the research where TransType2 had left it off. Caitra (Koehn, 2009) is an ITP tool which uses both the phrase table and the decoder of a statistical machine translation (SMT) (Koehn, 2010) system to generate suggestions. Researchers at the Universitat Politècnica de València have also made significant improvements to ITP systems (Barrachina et al., 2009). The CASMACAT project (`casmacat.eu`) followed the same line of research, improving ITP using active and on-line learning (Alabau et al., 2014). More recent works use neural MT systems (NMT) to generate the suggestions, as the decoding procedure can easily be adapted to use a given prefix (Peris et al., 2016; Knowles and Koehn, 2016). All these systems follow a *glass-box* strategy: in the case of SMT, suggestions are obtained by means of a tightly coupled system that is modified or tailor-made to provide additional information such as word alignments, alternative translations, and scores or probabilities for the translation; NMT systems only need to be slightly modified. ITP systems can therefore exploit most (if not all) the information captured in the translation model to generate the ITP suggestions, but inherit common SMT and NMT requirements, such as their dependency on extensive parallel corpora. Integrating other resources (such as commercial, translation-as-a-service engines over which no control is available) as part of the ITP process would be almost impossible, as most of them would not be able to provide the additional information needed to generate the suggestions.

Unlike the previously described glass-box approach, Pérez-Ortiz et al.'s (2014) system follows a *black-box* strategy: suggestions are obtained by splitting the source-language sentence in all possible sub-segments up to a given number of words, querying any available bilingual resource capable of delivering one or more translations into the target language, and eventually offering some of these translated segments as suggestions as the translation is typed. These bilingual resources can be MT systems, but also translation memories, dictionaries, catalogs of bilingual phrases, or any combination of them. The performance of this approach has been explored using rule-

based MT systems (Pérez-Ortiz et al., 2014) and in-domain and out-of-domain SMT systems (Torregrosa et al., 2014); more recently, the performance of the method used for suggestion ranking and selection has been improved by replacing the heuristics used in the early black-box ITP papers (Pérez-Ortiz et al., 2014; Torregrosa et al., 2014) with a neural network working on a set of features extracted from the source sentence, from the current prefix of the target sentence, and from the sub-segments translated with the bilingual resources (Torregrosa et al., 2016). Black-box systems have no access to the internals of the bilingual resource and can only use an approximation of the knowledge contained in the resource by translating each word multiple times in different contexts, that is, as part of the different segments (this means more words are translated overall), but this allows the integration of new resources without modifying how the ITP system works; similarly, the resources used do not need to provide additional information or be modified in any way. This makes it possible to use any resource available to the professional translator in an almost seamless way.

ITP popularity is on the rise and some commercial translation memory systems already integrate some form of ITP as one of their basic features (such as SDL Trados AutoSuggest 2.0, `translationzone.com/products/trados-studio/autosuggest`), and new translation tools such as Lilt (Green et al., 2014) (`lilt.com`) focus on delivering glass-box ITP on a user-friendly computer-assisted translation (CAT) web tool.

A comparison between the glass-box and the black-box approaches is carried out for the first time in this paper, by performing both an extensive automatic evaluation and a preliminary human evaluation. We evaluate both approaches when translating between related language pairs, particularly English–Spanish, and between less related languages such as Arabic–English and Chinese–English. This will help us assess the validity of the approaches for translating between languages that do not share the same syntactical structure, that is, those exhibiting frequent crossed and long-range word-alignments.

The remainder of the paper is organized as follows. In Section 2 we introduce our experimental set-up, and describe the automatic evaluation along with the results. In Section 3 we describe the experimental set-up and the results of the human evaluation. Finally, in Section 4, we discuss the results and propose future lines of research.

2. Experimental setup

2.1. Software used

As glass-box ITP model we will use the free/open-source toolkit Thot (`daormar.github.io/thot`) (Ortiz-Martínez and Casacuberta, 2014), which provides SMT, and ITP as a particular case of SMT where the system is forced to constrain the translation to a given prefix. Thot's ITP generates a word graph with probabilities using a modified version of the SMT decoder, and searches for the most probable translation constrained by the already typed prefix according to the word graph; an error-

	In-domain			Development	Out-of-domain	
Thot	Test	-		Development	Train	-
Forecat	-	Train	Development	-		
Evaluation	Test	-				
Sentences	3 000	10 000	2 000	2 000	1 000 000†	Rest of sentences†

Table 1. Distribution of the corpora. The sentences follow the same order as in the original corpus, except for the sentences tagged with †, which are ordered according to the similarity score of the bitext domain adaptation procedure. The top 1 million sentences for the glass-box training set were selected after filtering with the preprocessing tools in Thot.

correction algorithm is used if the typed prefix is not in the word graph. As black-box ITP model we will use the free/open-source toolkit Forecat (Torregrosa et al., 2016) (`github.com/transducens/forecat`). Forecat creates a pool of suggestions by splitting the source sentence in all the sub-segments up to a given length L, then translating them using any available bilingual resource. A set of features extracted from the source sentence, from the current prefix of the target sentence, and from the translated sub-segments is used by a feedforward neural network to rank the viability of the suggestions that are compatible (if the last word of the already typed prefix is the prefix of a suggestion, the suggestion is compatible); the top M suggestions are then offered to the user. In order to perform a fair comparison unaffected by the quality of the translation models, Forecat will use the same Thot SMT system as bilingual resource for translating the sub-segments in our experiments.

2.2. Corpora and model training

Parts of the Arabic–English (ar–en), English–Chinese (en–zh) and English–Spanish (en–es) bitexts from the United Nations Parallel Corpus 1.0 (Ziemski et al., 2016) have been used to train Thot models and the Forecat neural network, as well as to provide a test set for the automatic evaluation. Due to processing resources and time limitations, we had to reduce the size of the corpora used to train Thot models; to this end, we used the bitext domain adaptation procedure described by Axelrod et al. (2011) as implemented in XenC (Rousseau, 2013). This technique minimizes the impact of reducing the size of the training set by keeping the sentences that are more similar to the ones in the test set. The distribution of the corpus is shown in Table 1.

Thot's training and development sets were lowercased to reduce data sparsity and tokenized; those sentence pairs that could hinder the training procedure, such as extremely long sentences (more than 80 words) or sentence pairs with disparate lengths, were removed using the preprocessing tools in Thot, as described in its manual (`daormar.github.io/thot/docsupport/thot_manual.pdf`); however, the Stanford Tokenizer (`nlp.stanford.edu/software/tokenizer.shtml`) was used for the tokeniza-

tion of Chinese, as Thot does not support this task. The Simplified Chinese corpus was transliterated to the corresponding sequences for the Pinyin input method using Python's `pinyin 0.4.0` (`pypi.python.org/pypi/pinyin`), as Simplified Chinese characters are seldom directly typed. Thot was compiled to use IBM2 alignment models, and the training procedure used the parameter values in the user manual; a trigram language model and a maximum phrase length of 10 tokens were used. The reader may refer to the paper by Ortiz-Martínez and Casacuberta (2014) for more information about Thot's architecture. The BLEU (Papineni et al., 2002) scores for the resulting models (computed using the Thot tookit over the evaluation set) are: 0.49 for en→es, 0.47 for es→en, 0.43 for en→ar, 0.33 for ar→en, 0.23 for en→zh and 0.19 for zh→en.[1]

The Forecat feedforward neural network had one unit per feature in the input layer, 128 units in a single hidden layer, all fully connected to the input layer, and a single output unit fully connected to the hidden layer; it has a relatively small number of parameters, in the order of magnitude of 10^4. The training was performed via backpropagation with a learning rate of 10^{-3}, using the mean squared error (MSE) as the error function to optimize and no momentum or regularization; each model was trained five times with different weight initializations, and the one that results in a lower MSE was used in both the automatic and human evaluations. The reader may refer to the paper by Torregrosa et al. (2016) for more information about Forecat's architecture and for a description of the features.[2]

2.3. Automatic evaluation

The automatic evaluation model is similar to the one described by Langlais et al. (2000). A reference translation T is provided to the automatic system, which proceeds to "type" it; after each character, the system evaluates all the suggestions offered and chooses the suggestion or suggestion prefix that locally saves the most keystrokes and exactly matches the following words in T. Accepted suggestions or prefixes need to be full-word translations: if the word of T currently being translated is "thesaurus", a suggestion "the" will not be accepted. Accepting a full suggestion costs one keystroke, and accepting a suggestion prefix costs one keystroke per word in the selected prefix plus one keystroke for accepting the prefix (simulating the behaviour of the interface the human translators use, as described in Section 3). In order to measure the performance, we use the keystroke ratio (KSR), the ratio between the actual number of keys pressed for typing the translation and the length of the translation in characters; lower KSR values mean the suggestions were more useful while typing T. The glass-box model always offers one suggestion that completes the translation, and the user

[1] Even though 1 million sentences are too few for SMT, each of the resulting models use around 6 GB of RAM when loaded into the ITP server, most of the 8 GB available in the system used for human evaluation.

[2] The specific feature that takes the value of the starting letter (f_{26} in the paper by Torregrosa et al. (2016)) of the suggestion has been reworked for the en→ar task: rather than using the English alphabet, it uses the Arabic one; all the diacritics of the starting letter of the suggestion are removed.

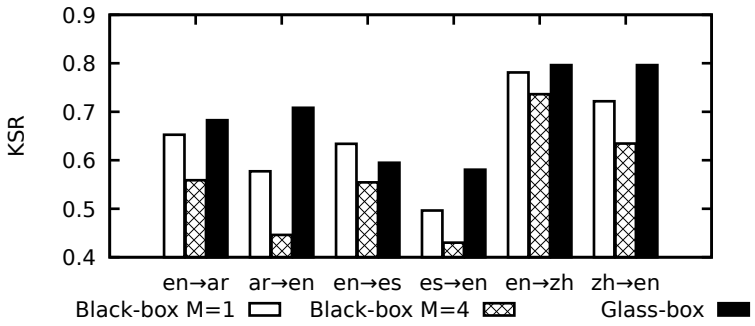

Figure 1. KSR values for the automatic evaluation. All differences between the values are statistically significant ($p \leq 0.05$).

can accept the full suggestion or a prefix of it; the suggestion will therefore be longer at the start of the task, and will shorten as the translation gets carried out. The average length over the 6 different translation tasks of the glass-box model suggestions offered during the automatic evaluation was of 20 words. The black-box model offers at most one suggestion ($M = 1$; if no suggestion is compatible with the typed prefix nothing is offered) with a maximum source sub-segment length of $L = 4$; the final length of the suggestion depends on the language pair and the words being translated. On average, the black-box model offered 2.3 words, or 1.4 words if we also consider the steps in which no suggestion is offered. The results obtained when allowing the black-box model to show up to 4 suggestions ($M = 4$) will also be shown, as this is the value used during the human evaluation; the black-box model with $M = 4$ shows on average 7.5 words (combining the length of the up to 4 suggestions), or 5 words if we also consider those steps where no suggestion is available. In both cases, the user can accept a full suggestion or a prefix of one of them.

2.4. Results of the automatic evaluation

We have performed extensive automatic evaluation for all six language pairs with both the black-box and the glass-box approaches, using all the sentences in the evaluation set described in 2.2. We tested the statistical significance of the results of the different models using paired bootstrap resampling (Koehn, 2004) with 1000 iterations and $p \leq 0.05$. The results of the automatic evaluation are shown in Figure 1. The black-box system using $M = 4$ outperformed the glass-box strategy by a wide margin, even when it had no access to all the information contained in the SMT system and, on average, showed less than half the words to the user as explained in the previous section; the black-box system with $M = 1$ still outperformed the glass-box system for every task but en→es, and showed on average less than a fourth of the words of

the glass-box approach. The black-box and glass-box approaches have closer performances when translating from English, as the corpora was originally written in English then translated; for en→es, the translation process is simpler and the glass-box method offers better suggestions.

3. Human evaluation

3.1. Experimental setup

We performed a human evaluation in order to compare the black-box and the glass-box approaches. To this end, both Forecat and Thot have been integrated into the open-source TM tool OmegaT (`omegat.org`) as plugins `github.com/dtr5/Forecat-OmegaT`, `github.com/dtr5/thot-omegat`). We used a preexistent plug-in to log user actions non-obstructively (`github.com/mespla/OmegaT-SessionLog`). No translation memory was used during testing. The suggestions (either the single sentence completion suggestion offered by the glass-box strategy or the up to $M = 4$ suggestions offered by the black-box strategy) are offered to the users in a drop-down list as they type the translation; these suggestions can then be accepted by selecting them using the arrow keys and pressing the enter key, by using a hot-key combination (`Alt+p`, with p being the position on the list) or with the mouse. Another hot-key (`Tab`) is used to select a prefix of the current suggestion, word by word. All the actions performed by the human translator, such as typing one character or selecting a full suggestion either with the mouse or the keyboard cost one keystroke, but selecting the prefix of a suggestion has a cost of one keystroke per word (`Tab` has to be pressed once per word) in the selected prefix plus one additional keystroke for accepting the prefix.

We have selected the first 20 English sentences with lengths between 15 and 25 words in the English–Spanish test set: this range of lengths excludes those sentences that are too long to be easily understood by non-native speakers and those so short that are hard to translate isolated from their context or do not present any kind of challenge to the translators. The sentences were arranged in 4 blocks SB_1–SB_4 of 5 sentences each, and the blocks were distributed so that each block was translated by two users under each modality. The 4 blocks were presented to the 8 users using 4 different modalities: the *induction task* let them familiarize with the interface and both suggestion models; the *unassisted task* offered no suggestions whatsoever; the *black-box task* used the black-box model, offering up to 4 suggestions ranked using the best neural network configuration, and the *glass-box task* used the glass-box model, offering a sentence-completion suggestion using the typed prefix as a constraint.

All eight test subjects U_1–U_8 were computer science researchers currently working in our university as technical or research staff. All of them except for U_5 claimed to be experienced typists. All of them are native Spanish speakers, and self-assessed themselves to have an R2/R2+ level (limited working proficiency) of English in the Interagency Language Roundtable scale for reading (a proficiency scale available

SB_1	Time	Tc	Tc/s	KS	KS/s	KSR	ESR	SB_2	Time	Tc	Tc/s	KS	KS/s	KSR	ESR
U_1								U_1	528	637	1.21	996	1.89	1.56	–
U_2	996	666	0.67	996	1.00	1.50	–	U_2	**626**	**636**	**1.02**	**686**	**1.10**	**1.08**	**0.71**
U_3	**524**	**603**	**1.15**	**830**	**1.58**	**1.38**	**0.74**	U_3	*576*	*570*	*0.99*	*537*	*0.93*	*0.94*	*0.75*
U_4	*715*	*567*	*0.79*	*747*	*1.04*	*1.32*	*0.68*	U_4							
U_5								U_5	477	677	1.42	690	1.45	1.02	–
U_6	687	736	1.07	996	1.45	1.35	–	U_6	**642**	**631**	**0.98**	**686**	**1.07**	**1.09**	**0.67**
U_7	**468**	**604**	**1.29**	**583**	**1.25**	**0.97**	**0.76**	U_7	*466*	*547*	*1.17*	*548*	*1.18*	*1.00*	*0.65*
U_8	*602*	*581*	*0.97*	*717*	*1.19*	*1.23*	*0.70*	U_8							

SB_3	Time	Tc	Tc/s	KS	KS/s	KSR	ESR	SB_4	Time	Tc	Tc/s	KS	KS/s	KSR	ESR
U_1	**613**	**677**	**1.10**	**686**	**1.12**	**1.01**	**0.62**	U_1	*513*	*615*	*1.20*	*819*	*1.60*	*1.33*	*0.49*
U_2	*732*	*618*	*0.84*	*819*	*1.12*	*1.33*	*0.68*	U_2							
U_3								U_3	298	646	2.17	765	2.57	1.18	–
U_4	668	606	0.91	782	1.17	1.29	–	U_4	**479**	**612**	**1.28**	**661**	**1.38**	**1.08**	**0.69**
U_5	**542**	**639**	**1.18**	**686**	**1.26**	**1.07**	**0.65**	U_5	*525*	*595*	*1.13*	*819*	*1.56*	*1.38*	*0.67*
U_6	*605*	*635*	*1.05*	*819*	*1.35*	*1.29*	*0.77*	U_6							
U_7								U_7	396	660	1.67	681	1.72	1.03	–
U_8	595	644	1.08	783	1.32	1.22	–	U_8	**392**	**647**	**1.65**	**807**	**2.06**	**1.25**	**0.66**

Table 2. Performance of the users with the different sentence blocks for the unassisted task (in regular typeface), the black box task (in bold) and the glass box task (in italics). The rows corresponding to the induction task are blank, as those results are not relevant.

at `govtilr.org/skills/ILRscale4.htm`). None of them had any kind of translation education or was familiar with the domain of the corpora. All of them resorted to using Google translate (`translate.google.com`) to look up the translation of single words or short phrases, except for U_1, who used the online version of the Cambridge English dictionary (`dictionary.cambridge.org`), and U_7, who preferred Linguee (`linguee.com`). Most users consulted domain-specific terms such as "guidelines", "compliance" or "interim". They were supervised during the test, and encouraged to ask as many questions as they needed to and experiment with the different suggestion systems, but only during the induction task. The instructions included all the ways they could use the suggestions and stressed that users were not obliged to accept one of the suggestions offered, but that they should also avoid ignoring them altogether.

3.2. Results of the human evaluation

We measured the time, the size in characters of the translations (Tc) and the number of keystrokes (KS), and calculated the translation speed (Tc/s), the number of keystrokes per second (KS/s) and the keystroke ratio (KSR=KS/Tc). We also calculated the emulated KSR (ESR) by performing the automatic evaluation described in Subsection 2.3 using the same conditions as the human test and the generated translations as references. The results of the human evaluation are shown in Table 2; an analysis of the differences in translation speeds and KSR of each method and user is shown in Figure 2. Only U_2 managed to translate both faster and with less effort

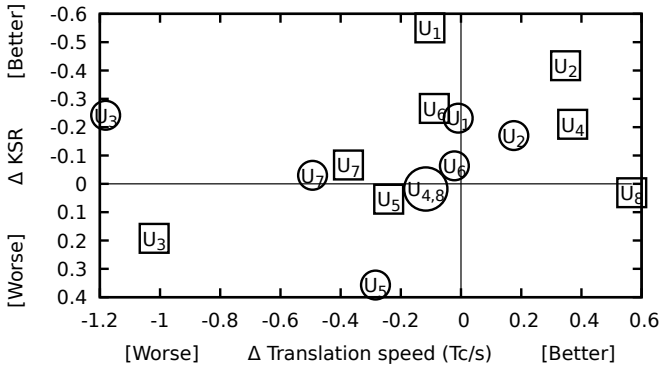

Figure 2. Absolute increase of KSR and Tc/s of the glass-box (○) and the black-box (□) tasks against the unassisted task. U_4 and U_8 got grouped because they attained very similar performances with the glass-box system.

with both techniques; U_4 managed to do so only with the black-box method. On average, when compared to the unassisted task, the evaluators saved 10% keystrokes and were 4% faster with the black-box approach, and saved 15% keystrokes and were 12% slower with the glass-box one; black-box suggestions proved therefore less useful but they allowed translators to perform faster. ESR values show that they could have theoretically saved 51% and 69% respectively if they had used the compatible suggestions. Users only had a few minutes to familiarize with the techniques, and it is expected that the translation speed will rise and the gap between the KSR and the ESR will close (but not completely, as part of this margin can be explained by user mistakes and rethought translations) as users get more and more familiar with the technology; a recent study by Autodesk (`langtech.autodesk.com/productivity.html`) considers experience the most single important factor in translation productivity.

After the tests, users were asked to sort the tasks according to their perceived speed of translation and ease of translation. U_1, U_4 and U_8 perceived the black-box system as faster and more helpful than the rest; the rest preferred the glass-box system; U_4 thought the glass-box system led to faster translations, yet it made the translation task harder than without assistance; finally, U_5 thought the black-box system made the task both harder and slower. Users' perceptions strongly contrasted with the measurements: only U_2 was faster with both methods compared to unassisted translation (0.67 Tc/s), though glass-box (0.84 Tc/s) was incorrectly perceived to be faster than black-box (1.02 Tc/s); and U_4 correctly ranked black-box (1.28 Tc/s) as the fastest task.

Finally, they were asked to provide some open feedback. U_4 strongly disliked the OmegaT tool. Most users were slightly annoyed by the unassisted block after experimenting with the induction block; some of them also said that the unassisted block had the harder sentences to translate, even when the sentences themselves were dif-

ferent from user to user. As none of them are professional translators, most of them expressed that the first full-sentence suggestion that the glass-box system gave them was very useful for planning the translation, but some complained those suggestions were too long and unwieldy. Some users complained about suggestions being offered too often, specially when none of them were useful. Some users praised the tool as they were able to operate it using only the keyboard; they all are experienced coders and most work in environments operable without a mouse. However, none of them used the `Alt+p` option for accepting specific suggestions from the drop-down list. The option for using the prefix of a suggestion by pressing `Tab` was neglected until they reached the glass-box block, as the suggestions were too long to be useful as a whole, but some had an adequate prefix.

4. Conclusions and future work

Interactive translation prediction (ITP) is a computer-assisted translation modality that focuses on offering translation suggestions as the translation is carried out. The automatic evaluation performed on this paper shows that 20%–50% of keystrokes can potentially be saved compared to unassisted translation using either the black-box or the glass-box approaches, regardless of whether the translation task is for related languages such as en–es or more unrelated ones such as ar–en or en–zh. The comparison between the black-box and the glass-box approaches shows that under these particular conditions, the black-box approach consistently outperforms the glass-box one in all but one translation task (en→es), even when the black-box approach does not have access to the internal information of the SMT model and shows to the user less than a fourth of the words offered by the glass-box model. Exhaustive analysis under different conditions needs to be carried out to identify when each system is useful and which one performs the best. Once these conditions are known, a hybrid strategy that chooses the best approach for each task could be devised. Also, even when the black-box strategy shows less words, we do not know the effect this has on the user; a detailed study about the cost of showing words and how many of them the users read before accepting or rejecting the suggestions has to be carried out.

In the human evaluation for en→es, test subjects mostly agreed in that both methods were useful, but were also divided when choosing which system was better for performing the translations; five of them preferred the glass-box approach and three preferred the black-box approach. Only one user managed to save keystrokes and be faster with both approaches. On average, the evaluators saved 10% keystrokes and were 4% faster with the black-box approach, and saved 15% keystrokes and were 12% slower with the glass-box one, but they could have saved 51% and 69% respectively if they used the compatible suggestions; as the users get more comfortable with the tool, the translation speed and the keystroke savings may both improve. Our preliminary human tests can be used to give an indication of how each system performs, but they suffer of two limitations: the size of the task and the profile of the users. A more

extensive evaluation with professional translators, translation students, or both will be carried out to explore the influence of different parameters and translation tasks. One common user complaint was that suggestions were being offered too often. Both models can be improved so they can assess the quality of the suggestions and offer only those that surpass some threshold. The detailed logs of the human evaluation sessions could also be used to tune the automatic evaluation strategies so they better reflect how users interacted with both approaches.

Finally, all the software used in this work is available under a free/open-source license. OmegaT users can now integrate both black-box and glass-box ITP and benefit from the performance improvements; using the plugins as inspiration, developers of other CAT tools can also integrate them into their tools.

Acknowledgments: Work partially funded by the Generalitat Valenciana through grant ACIF/2014/365, the Spanish government through project EFFORTUNE (TIN2015-69632-R), and by the Government of the Republic of Kazakhstan.

Bibliography

Alabau, Vicent, Jesús González-Rubio, Daniel Ortiz-Martínez, Germán Sanchis-Trilles, Francisco Casacuberta, M García-Martínez, Bartolome Mesa-Lao, Dan Cheung Petersen, Barbara Dragsted, and Michael Carl. Integrating online and active learning in a computer-assisted translation workbench. In *Proceedings of the First Workshop on Interactive and Adaptive Statistical Machine Translation, page to appear*, pages 1–8, 2014.

Axelrod, Amittai, Xiaodong He, and Jianfeng Gao. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 355–362. Association for Computational Linguistics, 2011.

Barrachina, Sergio, Oliver Bender, Francisco Casacuberta, Jorge Civera, Elsa Cubel, Shahram Khadivi, Antonio Lagarda, Hermann Ney, Jesús Tomás, Enrique Vidal, and Juan-Miguel Vilar. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35 (1):3–28, 2009.

Green, Spence, Jason Chuang, Jeffrey Heer, and Christopher D Manning. Predictive Translation Memory: A mixed-initiative system for human language translation. In *Proceedings of the 27th annual ACM Symposium on User Interface Software and Technology*, pages 177–187, 2014.

Hutchins, W. John and Harold L. Somers. *An introduction to machine translation*. Academic Press, 1992. ISBN 9780123628305.

Knowles, Rebecca and Philipp Koehn. Neural Interactive Translation Prediction. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*, volume 1: MT researchers track, pages 107–120, 2016.

Koehn, Philipp. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the conference on Empirical Methods on Natural Language Processing (EMNLP 2004)*, pages 388–395, 2004.

Koehn, Philipp. A web-based interactive computer aided translation tool. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 17–20, 2009.

Koehn, Philipp. *Statistical Machine Translation*. Cambridge University Press, 2010. ISBN 0521874157, 9780521874151.

Langlais, Philippe, Sébastien Sauvé, George Foster, Elliott Macklovitch, and Guy Lapalme. Evaluation of TransType, a computer-aided translation typing system: a comparison of a theoretical-and a user-oriented evaluation procedures. In *Conference on Language Resources and Evaluation (LREC)*, pages 641–648, 2000.

Macklovitch, Elliott. TransType2: The last word. In *Proceedings of the 5th International Conference on Languages Resources and Evaluation (LREC 06)*, pages 167–172, 2006.

Ortiz-Martínez, Daniel and Francisco Casacuberta. The New Thot Toolkit for Fully Automatic and Interactive Statistical Machine Translation. In *Proc. of the European Association for Computational Linguistics (EACL): System Demonstrations*, pages 45–48, Gothenburg, Sweden, April 2014.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

Pérez-Ortiz, Juan Antonio, Daniel Torregrosa, and Mikel L. Forcada. Black-box integration of heterogeneous bilingual resources into an interactive translation system. *EACL 2014 Workshop on Humans and Computer-assisted Translation*, pages 57–65, 2014.

Peris, Álvaro, Miguel Domingo, and Francisco Casacuberta. Interactive neural machine translation. *Computer Speech & Language*, 2016.

Rousseau, Anthony. Xenc: An open-source tool for data selection in natural language processing. *The Prague Bulletin of Mathematical Linguistics*, 100:73–82, 2013.

Somers, Harold L. *Computers and Translation: A Translator's Guide*. Benjamins translation library. John Benjamins Publishing Company, 2003. ISBN 9789027216403.

Torregrosa, Daniel, Mikel L. Forcada, and Juan A. Pérez-Ortiz. An open-source web-based tool for resource-agnostic interactive translation prediction. *The Prague Bulletin of Mathematical Linguistics*, 102(1):69–80, 2014.

Torregrosa, Daniel, Mikel L. Forcada, and Juan A. Pérez-Ortiz. Ranking suggestions for black-box interactive translation prediction systems with multilayer perceptrons. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*, volume 1: MT researchers track, pages 65–78, 2016.

Ziemski, Michał, Marcin Junczys-Dowmunt, and Bruno Pouliquen. The United Nations Parallel Corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3530–3534, Paris, France, 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

Address for correspondence:
Daniel Torregrosa
dtorregrosa@dlsi.ua.es
Universitat d'Alacant, E-03690 Sant Vicent del Raspeig, Spain