# Integrating Online and Active Learning in a Computer-Assisted Translation Workbench

**Vicent Alabau** valabau@prhlt.upv.es
**Jesús González-Rubio** jegonzalez@prhlt.upv.es
**Daniel Ortiz-Martínez** dortiz@prhlt.upv.es
**Germán Sanchis-Trilles** gsanchis@dsic.upv.es
**Francisco Casacuberta** fcn@prhlt.upv.es
Departamento de Sistemas Informáticos y Computación, Universitat Politècnica de València
Camino de Vera s/n, 46021 Valencia (Spain)

**Mercedes García-Martínez** mgm.ibc@cbs.dk
**Bartolomé Mesa-Lao** bm.ibc@cbs.dk
**Dan Cheung Petersen** dcp.icb@cbs.dk
**Barbara Dragsted** bd.ibc@cbs.dk
**Michael Carl** mc.ibc@cbs.dk
Center for Research and Innovation in Translation and Translation Technology (CRITT)
Copenhagen Business School, Dalgas Have 15, 2000 Frederiksberg (Denmark)

**Abstract**

This paper describes a pilot study with a computed-assisted translation workbench aiming at testing the integration of online and active learning features. We investigate the effect of these features on translation productivity, using interactive translation prediction (ITP) as a baseline. User activity data were collected from five beta testers using key-logging and eye-tracking. User feedback was also collected at the end of the experiments in the form of retrospective think-aloud protocols. We found that OL performs better than ITP, especially in terms of translation speed. In addition, AL provides better translation quality than ITP for the same levels of user effort. We plan to incorporate these features in the final version of the workbench.

## 1 Introduction

The use of machine translation (MT) systems for the production of post-editing drafts has become a widespread practice in the industry. Many language service providers are now using port-editing workflows due to a greater availability of resources and tools for the development of MT systems, as well as a successful integration of MT systems in already well-established computer-assisted translation (CAT) workbenches.

This paper reports on the CAT workbench being developed within the CASMACAT project[1]. Among the different features implemented in the workbench, we will investigate the *interactive translation prediction* (ITP) approach (Langlais and Lapalme, 2002; Casacuberta et al., 2009; Barrachina et al., 2009). Within the ITP framework, a state-of-the-art statistical

[1]CASMACAT: *Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation*. Project co-funded by the European Union under the Seventh Framework Programme Project 287576 (ICT-2011.4.2).



*Proceedings of the Workshop on Interactive and Adaptive Machine Translation*, pages 1–8
AMTA Workshop. Vancouver, Canada. September 22, 2014

machine translation (SMT) system is used in the following way. For a given source sentence, the SMT system automatically generates an initial translation. A human translator then proofreads checks this machine generated translation, correcting the first error. The SMT system then proposes a new completion (or suffix), taking the user correction into account. These steps are repeated until the whole input sentence has been correctly translated.

The CASMACAT workbench further extends the ITP approach by introducing two new features, namely, online and active learning. These two new features are designed to allow the system to take further advantage from user feedback. Specifically, the SMT models are updated in real time from the target translations validated by the user, preventing the system from repeating errors in the translation of similar sentences. Despite the strong potential of these features to improve the user experience (Ortiz-Martínez et al., 2010; González-Rubio et al., 2012; Bertoldi et al., 2013; Denkowski et al., 2014), they are still not widely implemented in CAT systems. To the best of our knowledge, the only exception is (Ortiz-Martínez et al., 2011) where the authors describe the implementation of online learning within an ITP system.

The present study reports on the results and user evaluation of the CASMACAT workbench under three different conditions: 1) basic ITP, 2) ITP with online learning, and 3) ITP with active learning. The ultimate aim of testing these different configurations was to assess their potential in real world post-editing scenarios and decide which of them can be successfully integrated into the final prototype of the CASMACAT workbench for the benefit of the human translator.

## 2 Online and Active Learning for SMT

The proposed CAT workbench has been extended by incorporating online and active learning, which are targeted to optimizing the quality of the final translations and speeding the post-editing process by taking advantage of user feedback in real time.

### 2.1 Online Learning

*Online learning* (OL) allows us to efficiently re-estimate the parameters of the SMT model with the new translations generated by the user (Ortiz-Martínez et al., 2010). As a result, the SMT system is able to learn from the translation edits of the user preventing further errors in the machine generated translations.

Conventional batch learning techniques establish a strict separation between model training and the subsequent use of the estimated parameters for prediction. As a result, SMT systems implementing batch learning require to retrain the whole corpus whenever a new training example is available, spending days or even weeks of computation depending on the size of the training set. In contrast, OL techniques process the training examples one at a time or in small batches. This approach allows the re-estimation of the parameters of an SMT model in constant time, whatever the number of training examples previously processed is.

The application of OL to the SMT framework requires the definition of incremental update rules for the statistical models involved in the translation process. For this purpose, first it is necessary to identify a set of sufficient statistics for such models. A sufficient statistic for a statistical model is a statistic that captures all the information that is relevant to estimate this model. If the estimation of the statistical model does not require the use of the EM algorithm (Dempster et al., 1977), e.g. language models, then it is generally easy to incrementally extend the model given a new training sample. By contrast, if the EM algorithm is required, e.g. alignment models, the estimation procedure has to be modified, since the conventional EM algorithm is designed for its use in batch learning scenarios. To address this problem, we implement the incremental version of the EM algorithm defined in (Neal and Hinton, 1999).

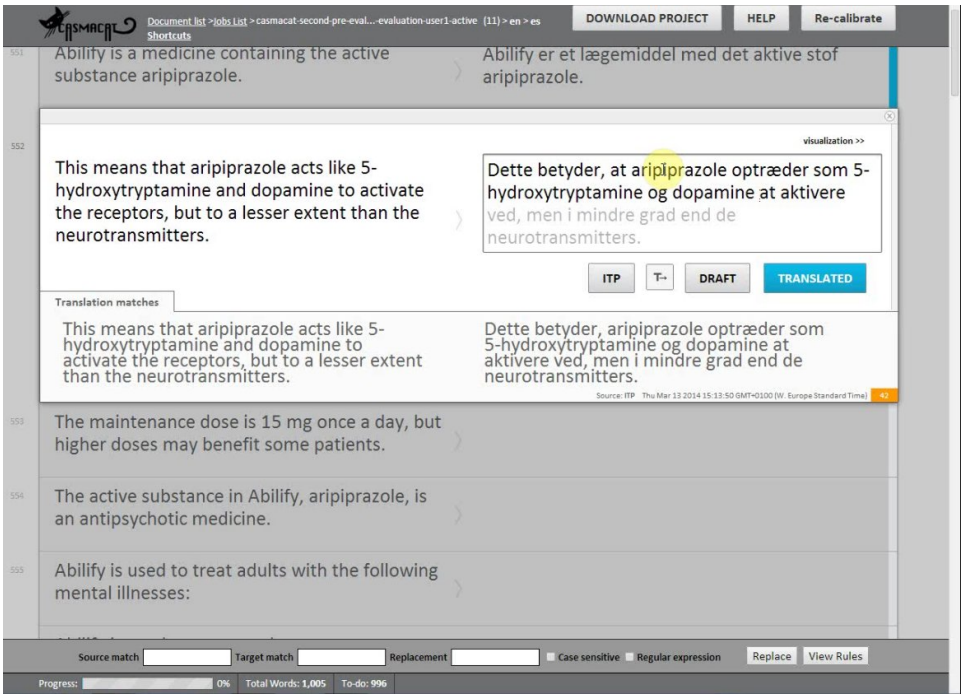


Figure 1: Screenshot of the CASMACAT workbench.

### 2.2 Active Learning

*Active learning* (AL) applied to ITP aims at optimizing the quality of the final translation as a whole when the available resources, (e.g. manpower, time, money, etc.) are limited (González-Rubio and Casacuberta, 2014). In this case, the user is asked to post-edit only a subset of the worst machine generated translations while the system returns SMT outputs for the rest of the sentences. Moreover, each time the user translates a sentence, we feed the newly generated translation example to the SMT model.

This AL framework has several potential advantages over conventional ITP technology. On the one hand, asking the user to only translate a subset of the sentences allows us to limit the amount of effort to be invested in the translation process and, by focusing human effort in those sentences for which the investment of user effort is estimated to be more profitable, we also maximize the utility of each user interaction. On the other hand, the underlying SMT model is continually updated with new examples which allows the system to learn new translations and to adapt its outputs to match the preferences of the user. As a result, the subsequent machine generated translations will be closer to those preferred by the user thus reducing the human effort required to translate them. Additionally, all these technicalities are transparent to the user who interacts with the system in the same way she does with a conventional ITP system.

An important practical challenge is the strict bound to the response time imposed by the interaction with the user. This fact constraints the models and techniques that can be used to implement AL. Particularly, we select which sentences should be post-edited by the user according to a sentence-level quality measure based on statistical lexicons (González-Rubio et al., 2012) and, given a new translation example, the parameters of the SMT model are re-estimated via the OL techniques described above.

| | Native Danish Speaker | Professional translator |
|---|---|---|
| U0 | yes | no |
| U1 | yes | yes |
| U2 | no | yes |
| U3 | yes | yes |
| U4 | yes | yes |

Table 1: Profile of the users in the pilot study.

## 3 CASMACAT Workbench

CASMACAT is a CAT workbench developed on top of the MATECAT post-editing interface (Bertoldi et al., 2012). The user is presented with a GUI in which the left-hand window displays the source text while the right-hand one contains the target text. Texts are split into segments (corresponding to sentences and headings in the text) so that the translator post-edits one translation segment at a time. The user can see several segments on the screen at the same time and can scroll back and forth to choose which segment to translate. The workbench contains a fully-fledged MT engine with interactivity which can search for alternative translations whilst the user is post-editing the machine translation. The SMT engine providing the above mentioned functionalities has been implemented using the Thot toolkit (Ortiz-Martínez and Casacuberta, 2014). Figure 1 shows a screenshot of the CASMACAT workbench.

Moreover, the workbench includes facilities for logging system configuration and user activity data including keystrokes and gaze obtained using an eye-tracking device.

## 4 Experimental design

The main goal of this pilot study was to assess and compare OL and AL against conventional ITP. To analyze the results, we used the following measures of the translation process:

• **Speed:** total number of words translated divided by time in minutes.

• **Effort:** total number of edits done by the user divided by the number of translated words.

The source texts were extracted from the EMEA corpus (Tiedemann, 2009). A group of five users volunteered to perform the evaluation of the system post-editing from English into Danish. Table 1 summarizes the profile of the users. According to the professional experience of the users, we carried out two different experiments:

**First experiment:** U0 post-edited three comparable texts with 55 segments each (843 words, 803 words, and 1,005 words). Each text was translated using a different condition, i.e. ITP, ITP with OL, or ITP with AL.

**Second experiment:** Four users (U1 to U4) were asked to post-edit the same source text (the one with 1,005 words in the first experiment), each user in a different condition. In this case we maintain constant the translation task and compare results from different users.

| **U0** | ITP | OL |
|---|---|---|
| Words translated | 843 | 803 |
| Words/min. | 14.1 | 16.4 |
| Keystrokes/word | 2.3 | 2.3 |

Table 2: First experiment: ITP vs. OL results.

| | **U1** | **U2** | **U3** |
|---|---|---|---|
| Native | Yes | No | Yes |
| Condition | ITP | OL | OL |
| Words/minute | 15.2 | 40.2 | 18.0 |
| Keystrokes/word | 2.9 | 0.6 | 1.8 |

Table 3: Second experiment: ITP vs. OL results.

## 5 Results

### 5.1 User activity data

First we will present the results comparing conventional ITP and ITP with OL. In both conditions, users post-edited all the sentences in the corpus. Table 2 shows ITP and OL results for the first experiment in which U0 post-edited different texts under the three conditions. Table 3 shows the corresponding results for the second experiment, where the same text (1,005 words) was post-edited by different users under one condition each.

It can be seen that OL significantly improved translation speed (about 2.5 more words translated per minute). Regarding the number of keystrokes, results are not consistent: no significant difference was found in the first experiment for the two conditions while it was significantly better for OL in the second experiment. The anomalous results for U2 can be explained by the different profile of the user (i.e. U2 was not a native speaker of Danish).

Regarding the results for ITP with AL against conventional ITP, the users were asked to post-edit the segments according to the quality of the SMT output. That is, users post-edited first the segments for which the machine generated translations were considered to be worst. It is important to note that since the user did not post-edit all machine generated translations (just the ones with the worst quality), the final target text was a mixture of automatic and human post-edited translations. In a second phase, we computed the quality (BLEU) of the output translations and the effort invested (keystrokes per post-edited word) as a function of the number $n$ of automatic translations post-edited by the user. We ranged $n$ between zero and $55$, the number of segments in the text. Figure 2 shows the improvement in translation quality with respect to SMT as a function of the effort invested by U0. Similar results were obtained when comparing U1 versus U4 in the second experiment. Results show that for the same amount of effort, AL provides a larger increase in translation quality as compared to conventional ITP.

### 5.2 User feedback

User feedback was collected after each post-editing session in the form of retrospective think-aloud protocols. The post-editing process was recorded in the form of screen capture video and then replayed to the users in order to elicit their actions and feelings as they went about with the post-editing tasks. Below, we include some of the comments and ideas provided by the users.

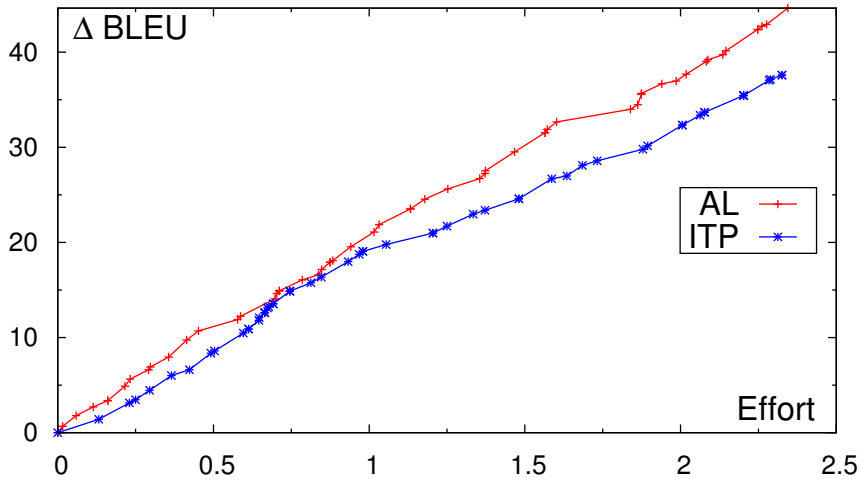


Figure 2: First experiment: improvement in translation quality with respect to SMT as a function of the human effort (keystrokes/word) invested by U0.

**U1 (native speaker and professional translator) observations on post-editing through ITP.**

> "Compared with editing in a non-interactive setting, the interactive translation mode was generally quite a different experience from a users point of view. It was necessary to 'unlearn' some of the editing processes normally carried out during revision of human or machine translation, such as highlighting words or segments and overwriting them with improved alternatives, and reading and planning a whole sentence before making corrections. This lead to a very different editing process, which required some getting used to and caused a good deal of frustration at first. However, after some time and practice, and 'unlearning' of old habits, efficiency improvements kicked in, but only to the extent that the dynamic changes were appropriate, which was not always the case. Thus, the problems experienced when working in the interactive mode were generally associated more with the quality of some of the dynamic corrections made by the system and less with the interactive mode as such.
>
> On the positive side, the grammatical corrections generally worked well. For example, when the definite article ('det'/'den'/'de' in Danish) was inserted (by the user) before a pre-modifying adjective, the system automatically added the inflection -e to the adjective, which is the correct form in Danish. Also, when a noun was written as an alternative to the original MT solution, the original noun was automatically removed, which saved the user the delete action and thus improved efficiency.
>
> On the negative side, dynamic corrections at the lexical level were not always appropriate. For example, when adding the morpheme 'op-' to the Danish noun 'løsning' to arrive at the Danish word for 'dissolution' ('opløsning'), rather than 'solution' ('løsning'), the system suggested 'opfølgning' ('follow-up'). This inappropriate dynamic correction then had to be revised by deleting 'følgning' and reinserting 'løsning', which lead to decreased efficiency in the post-editing process.
>
> The gray/black distinction to differentiate between edited and non-edited text worked well for me. It was easy to keep track of already accepted text and output that was yet to be checked."

**U0 (native speaker and non professional translator) observations on ITP with AL.**

> "The use of AL features while post-editing helped me a lot especially when using a more technical vocabulary. The interactivity seems faster and easier to recall completely different words, but it is quite the opposite when it comes to introduce small grammatical chances, such as word endings in Danish. I think that I would need more hours interacting with the system to make the most of it, but it is a nice feature when the system is able to remember my word preferences to help me improving my productivity and consistency overall."

## 6 Conclusions

We have presented the results of a pilot study concerning the implementation of OL and AL within a CAT workbench. We have reported both quantitative results measuring the efficiency of the translation process, and qualitative results in form of the comments and observations provided by different users of the workbench. Both configurations according to the feedback provided and the measurements registered have proven to be useful when integrated in the workbench. These results must be interpreted cautiously because of the small number of users involved in the study. Nevertheless, given that OL yielded the best productivity results in this pilot study, it will be the feature finally included in future versions of the workbench.

## Acknowledgments

Work supported by EU's $7^{\text{th}}$ Framework Programme (FP7/2007-2013) under grant agreement 287576 (CASMACAT).

## References

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

Bertoldi, N., Cattelan, A., and Federico, M. (2012). Machine translation enhanced computer assisted translation. First report on lab and field tests.

Bertoldi, N., Cettolo, M., and Federico, M. (2013). Cache-based online adaptation for machine translation enhanced computer assisted translation. In *Proc. MT Summit*, pages 35–42.

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

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society.*, 39(1):1–38.

Denkowski, M., Dyer, C., and Lavie, A. (2014). Learning from post-editing: Online model adaptation for statistical machine translation. In *Proc. EACL*, pages 395–404, Gothenburg, Sweden. Association for Computational Linguistics.

González-Rubio, J. and Casacuberta, F. (2014). Cost-sensitive active learning for computer-assisted translation. *Pattern Recognition Letters*, 37:124–134.

González-Rubio, J., Ortiz-Martínez, D., and Casacuberta, F. (2012). Active learning for interactive machine translation. In *Proc. EACL*, pages 245–254.

Langlais, P. and Lapalme, G. (2002). TransType: development-evaluation cycles to boost translator's productivity. *Machine Translation*, 17(2):77–98.

Neal, R. and Hinton, G. (1999). A view of the EM algorithm that justifies incremental, sparse, and other variants. *Learning in graphical models*, pages 355–368.

Ortiz-Martínez, D. and Casacuberta, F. (2014). The new thot toolkit for fully automatic and interactive statistical machine translation. In *Proc. EACL*, pages 45–48.

Ortiz-Martínez, D., García-Varea, I., and Casacuberta, F. (2010). Online learning for interactive statistical machine translation. In *Proc. NAACL-HLT*, pages 546–554.

Ortiz-Martínez, D., Leiva, L. A., Alabau, V., García-Varea, I., and Casacuberta, F. (2011). An interactive machine translation system with online learning. In *ACL (System Demonstrations)*, pages 68–73.

Tiedemann, J. (2009). News from opus - a collection of multilingual parallel corpora with tools and interfaces. In *Proc. RANLP*, volume V, pages 237–248.