CATaLog Online: A Web-based CAT Tool for Distributed Translation with Data Capture for APE and Translation Process Research

Santanu Pal[1], Sudip Kumar Naskar[2], Marcos Zampieri[1], Tapas Nayak[2], Josef van Genabith[1,4]
[1]Saarland University, Germany, [2]Jadavpur University, India,
[4]German Research Center for Artificial Intelligence (DFKI), Germany
{santanu.pal, marcos.zampieri, josef.vangenabith}@uni-saarland.de
tnk02.05@gmail.com, sudip.naskar@jdvu.ac.in

Abstract

We present a free web-based CAT tool called *CATaLog Online* which provides a novel and user-friendly online CAT environment for post-editors/translators. The goal is to support distributed translation where teams of translators work simultaneously on different sections of the same text, reduce post-editing time and effort, improve the post-editing experience and capture data for incremental MT/APE (automatic post-editing) and translation process research. The tool supports individual as well as batch mode file translation and provides translations from three engines – translation memory (TM), MT and APE. TM suggestions are color coded to accelerate the post-editing task. The users can integrate their personal TM/MT outputs. The tool remotely monitors and records post-editing activities generating an extensive range of post-editing logs. Compared with current state-of-the-art CAT tools, *CATaLog Online* provides an enhanced interface, an option to integrate APE and more informative logs to help translation process research.

1 Introduction

Machine translation (MT) technology has improved substantially over the past few decades. MT output is no longer used just for gisting but also for post-editing by professional translators as an important part of the translation workflow. Several studies confirm that post-editing MT output increases translators' productivity and improves translation consistency (Guerberof, 2009; Plitt and Masselot, 2010; Zampieri and Vela, 2014). Alongside classical TM matches, computer-aided translation (CAT) Tools that integrate MT and TM output are a trend in the translation and localization industries providing translators more useful suggestions. Another important trend is the development of web-based CAT tools which require no local software installation and allow teams of translators to work on the same project simultaneously (e.g., WordFast Anywhere[1], MateCat[2] (Federico et al., 2014), and Wordbee[3], Lilt[4] etc.).

This paper presents *CATaLog Online*, a web-based CAT tool that provides translators MT, TM and APE output and ensures data capture for APE development and translation process research. The MT and APE systems integrated in *CATaLog Online* are based on Pal et al. (2015) and Pal et al. (2016b), respectively. In this paper, we present the key features implemented in *CATaLog Online* and their importance to translation project managers, translators, and MT and APE developers. Compared to state-of-the-art CAT tools (e.g., MateCat, Lilt) *CATaLog Online* offers the following advantages: (i) color coded TM translation suggestions (highlighted TM source and corresponding target fragments are shown in the same interface), (ii) a wide range of editing logs, (iii) alignment between source, TM/MT/APE and the results of human PE, (iv) improved TM similarity measure and search technique (Pal et al., 2016a), and (v) additional translation option from APE which learns from human post-edited data.

The paper is organized as follows. Section 2 presents the desktop version of the *CATaLog* tool. Section 3 describes in detail the main functionalities of *CATaLog Online*. Section 4 outlines APE and translation

[1]https://www.freetm.com/
[2]https://www.matecat.com/
[3]http://www.wordbee.com/
[4]https://lilt.com/

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, pages 98–102, Osaka, Japan, December 11-17 2016.

process research with *CATaLog Online*. Section 5 concludes and provides avenues for improving the CAT tool further.

2 CATaLog

CATaLog (Nayek et al., 2015) is a TM-based CAT tool which provides core functionalities for *CATaLog Online*. What distinguishes *CATaLog* from existing TM-based CAT tools is a set of newly introduced features targeted towards improving post-editing experience in terms of both performance and productivity. These include an improved TM similarity measure, searching and a novel coloring scheme. The color coding introduced into *CATaLog* guides the user during the translation (or post-editing) process. The matching parts in the TM source matches, as well as their translations in the target, are displayed in green, while the non-matching parts in both the TM source and target suggestions are displayed in red. Unaligned words are shown in orange. Similarly, when the user clicks on one of the 5 TM suggestions to start the post-editing task, the corresponding matching and non-matching parts in the input segment are also displayed in green and red, respectively. The color coding scheme not only helps the user to choose the most suitable TM suggestion for post-editing, it also helps the user to identify which parts of a TM match require more post-editing effort and which fragments are reliable translations.

3 CATaLog Online

CATaLog Online provides a novel and user-friendly online CAT environment for post-editors and translators to reduce post-editing time and effort and improve the post-editing experience. The basic TM fuctionalities in *CATaLog Online* follow *CATaLog*'s color coding scheme. *CATaLog Online* is a freeware software that can be used through a web browser (works best in Mozilla Firefox) and requires only a simple registration. The tool remotely monitors and records translator/post-editor activities generating a wide range of post-editing logs (cf. Section 3.5) that are a fundamental source of information for APE and translation process research. *CATaLog Online*, produces multiple translation options for an uploaded input text file. It is a language independent tool that enables users to upload their own translation memories.

On the main user interface[5], users can translate a single segment after choosing the source language and the target language (cf. "Quick Translation" in the main interface). The suggested translations are generated by three different engines: MT, TM and APE. The TM output is color coded. Unlike other existing CAT tools, *CATaLog Online* provides many facilities including file translation, CAT tool environment, user management, project management, translation data capture, TM/MT and APE support, as well as distributed translation, where teams of translators working on the same job, etc.

3.1 File Translation

CATaLog Online provides facilities for batch mode file translation[6], i.e., a user can input a source file. The *CATaLog Online* batch mode file translation option provides a post-editing environment which allows the user to post-edit the selected translation from among the three translation suggestions (MT, TM and APE). The user has to choose the source–target language pair and upload a text file which contains a set of source segments. The tool translates this text file at the back end by creating a project and then assigns a unique job identification number (Job ID) to the user which is displayed on the large red button in the interface (cf. Figure 3). Each project/job is associated with a unique job URL. The user can either keep this Job ID for future reference or directly go to the job page by clicking on the recent Job ID (i.e., the red button marked with the Job ID). To recover a project/job, the translator has to search the project/job using the corresponding Job ID (cf. Figure 3). The File translation interface provides on-the-fly user guidance regarding the "usage" and "tool functionality" in terms of message services.

[5] http://santanu.appling.uni-saarland.de/CATaLog/

[6] http://santanu.appling.uni-saarland.de/CATaLog/GeustTranslation.jsp

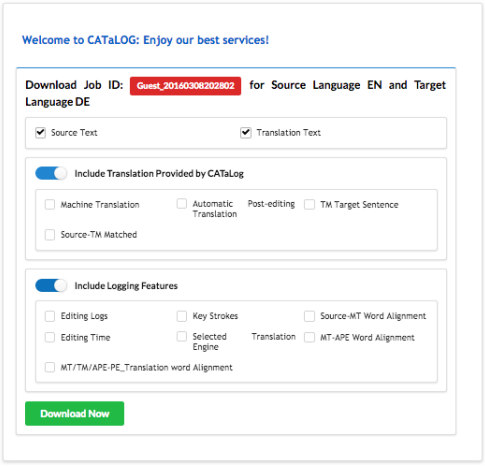

Figure 1: Job download interface

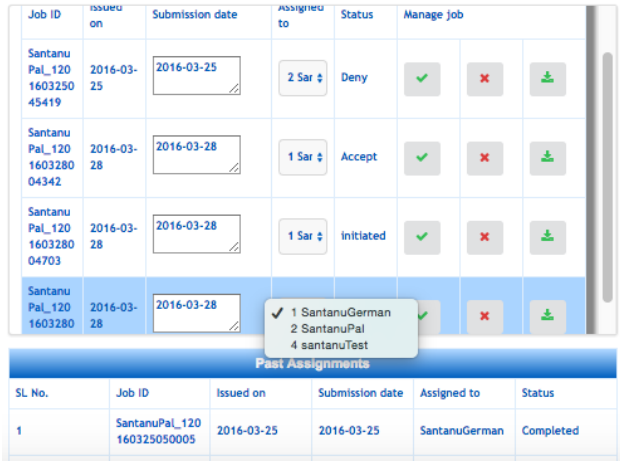

Figure 2: Project Management interface for PM

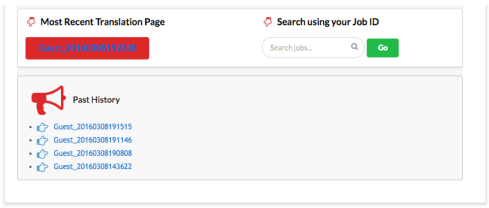

Figure 3: Job search interface

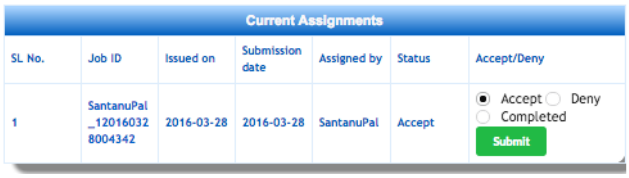

Figure 4: Project Management interface for translators

3.2 CAT Tool

The CAT Tool interface[7] is similar to the File Translation interface described in Section 3.1, however, it differs in terms of features and functionalities. Users can upload their own translation memories as tab separated text files. The tool is language agnostic and allows the user to upload files in any language. Users have full freedom to use MT translations generated by their own MT systems or third party MT engines (up to two alternatives are supported in the current version. Additionally, the tool provides color coded translations from the back end TM. When uploading finishes, the system provides a unique Job ID; the functionality is similar to that described in Section 3.1.

3.3 Project Management and Distributed Translation

The *CATaLog Online* project management system supports basic project management activities. A registered Project manager (PM) creates a translation project for a specific language pair by uploading a source file. Once a project/Job has been created, a Job Id appears in a row of the job assignment table. Additional information is associated with the Job Id, including issue date, submission date, available translators for that particular language pair, etc. The PM can review the job and assign translation sub-jobs to any of the available translators supporting concurrent distributed translation management including submission deadlines (cf. Figure 2).

As soon as the PM assigns a sub-job to a particular registered translator, the translator can see and review that job. The interface provides three options to the translator by which the translator can set the status of his/her activity for that particular job. A translator can either delete the assigned job from his/her profile by setting a "Deny" status or can accept it by setting the "Accept" status (cf. Figure 4). After finishing a translation task, the translator sets the corresponding job status as "Completed" which is directly updated in the PM's job status where the PM can see the completed and pending jobs. Finally, after reviewing, the PM can download the completed job and deliver it to the client.

[7] http://santanu.appling.uni-saarland.de/CATaLog/CATTool.jsp

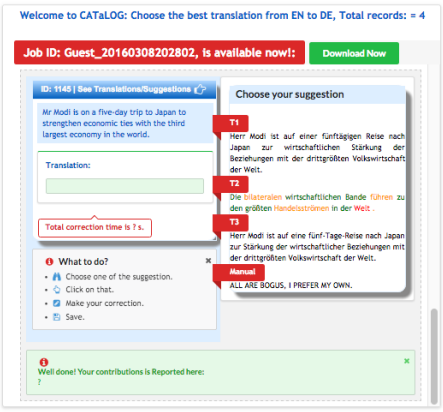

Figure 5: Job interface

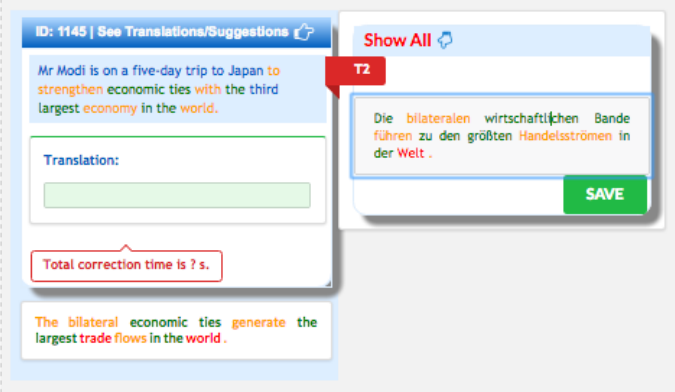

Figure 6: Job interface of TM selection

3.4 Job Management

A job is created when the PM or a guest user uploads a source file. The job interface provides three different translation alternatives for each source segment (cf. Figure 5).The TM translation alternative is color coded. The other two outputs are from MT and APE engines provided by *CATaLog Online* (cf. Section 3.1) or the uploaded third party MT engine outputs (cf. Section 3.2). As shown in Figure 5, source segments are listed in the blue panel on the left and the corresponding translation suggestions appear on the right panel upon clicking a link shown above the source segment. The translator chooses one of these suggestions and post-edits it. Figure 6 shows the interface when the translator selects the TM suggestion. The final translation appears in the green panel on the left when the translator presses the "Save" button. The editing time (in seconds) is also shown below the final translation panel. After finishing each translation, an editing summary shows the number of editing operations performed by the translator. *CATaLog Online* provides an on-the-fly editing guide throughout the translation process. In case of re-editing a translation, the previously stored final translation is shown as the first translation suggestion in the suggestion panel.

3.5 Editing Log

For a given input segment, the post-editor edits the best translation suggestion which may contain errors. The system records the user activities such as key strokes, cursor positions, text selection and mouse clicks. The tool provides analytical summaries of post-editing activities during translation and presents well structured XML formatted logs which can be customized according to the user's choice, e.g., the user can download the entire logs or some specific logs for a particular translation job (cf. Figure 1). The tool also provides word alignment which is also a part of the XML logs. *CATaLog Online* records word alignments between source–MT, MT–APE and source–HPE (human PE). The source–MT and MT–APE word alignments are established based on the decoding traces. The MT–HPE and APE–HPE alignments are recorded from the keystroke logs based on whether the user edits the MT output or APE output. Finally, the source–HPE alignments are generated by combining the transitive links between source–MT, MT–APE and APE–HPE in case of editing on the APE output or as the combination of source–MT and MT–HPE. These alignments and post-editing information are beneficial for translation process research.

4 APE and Translation Process Research using CATaLog Online

The post-editing logs collected during the translation process are a valuable source of information for translation process research as well as APE research and development. These user activity data logs not only help to assess the performance and understand the behavior of the translators, they also provide crucial information about cognitive aspects of post-editing. The logs can be used to model APE to improve quality and productivity.

User Perspective: *CATaLog Online* generates a summary for every completed translation task which

includes translator productivity in terms of number of words translated per minute and time taken per word. From the logs it is also possible to generate a report on translator style and behavior which can include, e.g., number of keystrokes per (effective) character editing, repetitive typing, preference for certain function words, etc.

Research Perspective: *CATaLog Online* records word alignments between source–MT, MT–APE, source–APE and source–HPE. These alignments and related post-editing information are beneficial for incremental MT/APE. Moreover, the source–HPE word alignments gathered by the tool can serve as a potential source for terminology extraction.

5 Conclusions and Future Work

CATaLog Online is a novel and user-friendly online CAT tool offering new features developed with the objective of improving translation productivity and experience. The tool provides a wide range of logs and data which serve as important information to translation process researchers, MT developers, and APE developers. The success of the two editions of the APE shared task in WMT (Bojar et al., 2016) indicate that APE is one of the important directions that research in MT is moving to. Post-editing tools, such as CATaLog Online, are able to provide crucial information for APE development. We would like to further expand and improve the tool by including additional features, e.g., interactive translation prediction in the form of on-the-fly translation suggestion, terminology extraction, option for compiling corpora, auto-suggestion for words, on-click pop-up terminology view, etc. Finally, we would like to model user behaviour and implement incremental MT/APE using the edit logs provided by the tool.

Acknowledgments

Santanu Pal is supported by the People Programme (Marie Curie Actions) of the EU Framework Programme (FP7/2007-2013) under REA grant agreement no 317471. Sudip Kumar Naskar is supported by Media Lab Asia, DeitY, Government of India, under the Young Faculty Research Fellowship of the Visvesvaraya PhD Scheme for Electronics & IT. Josef van Genabith is supported by funding from the EU Horizon 2020 research and innovation programme under grant agreement no 645452 (QT21).

References

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 Conference on Machine Translation (WMT16). In *Proceedings of WMT*.

Marcello Federico, Nicola Bertoldi, Mauro Cettolo, Matteo Negri, Marco Turchi, Marco Trombetti, Alessandro Cattelan, Antonio Farina, Domenico Lupinetti, Andrea Martines, et al. 2014. The Matecat Tool. In *Proceedings of COLING*.

Ana Guerberof. 2009. Productivity and Quality in the Post-editing of Outputs from Translation Memories and Machine Translation. *Localisation Focus*, 7(1):133–140.

Tapas Nayek, Sudip Kumar Naskar, Santanu Pal, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2015. CATaLog: New Approaches to TM and Post Editing Interfaces. In *Proceedings of the NLP4TM Workshop*.

Santanu Pal, Sudip Naskar, and Josef van Genabith. 2015. UdS-Sant: English–German Hybrid Machine Translation System. In *Proceedings of WMT*.

Santanu Pal, Marcos Zampieri, Sudip Kumar Naskar, Tapas Nayak, Mihaela Vela, and Josef van Genabith. 2016a. CATaLog Online: Porting a Post-editing Tool to the Web. In *Proceedings of LREC*.

Santanu Pal, Marcos Zampieri, and Josef van Genabith. 2016b. USAAR: An Operation Sequential Model for Automatic Statistical Post-Editing. In *Proceedings of WMT*.

Mirko Plitt and François Masselot. 2010. A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *The Prague Bulletin of Mathematical Linguistics*, 93:7–16.

Marcos Zampieri and Mihaela Vela. 2014. Quantifying the Influence of MT Output in the Translators' Performance: A Case Study in Technical Translation. In *Proceedings of the HaCat Workshop*.