THE MATECAT TOOL

M. Federico and **N. Bertoldi** and **M. Cettolo** and **M. Negri** and **M. Turchi**
Fondazione Bruno Kessler, Trento (Italy)
M. Trombetti and **A. Cattelan** and **A. Farina** and
D. Lupinetti and **A. Martines** and **A. Massidda**
Translated Srl, Roma (Italy)
H. Schwenk and **L. Barrault** and **F. Blain**
Université du Maine, Le Mans (France)
P. Koehn and **C. Buck** and **U. Germann**
The University of Edinburgh (United Kingdom)

www.matecat.com

Abstract

We present a new web-based CAT tool providing translators with a professional work environment, integrating translation memories, terminology bases, concordancers, and machine translation. The tool is completely developed as open source software and has been already successfully deployed for business, research and education. The MateCat Tool represents today probably the best available open source platform for investigating, integrating, and evaluating under realistic conditions the impact of new machine translation technology on human post-editing.

1 Introduction

The objective of MateCat[1] is to improve the integration of machine translation (MT) and human translation within the so-called computer aided translation (CAT) framework. CAT tools represent nowadays the dominant technology in the translation industry. They provide translators with text editors that can manage several document formats and suitably arrange their content into text *segments* ready to be translated. Most importantly, CAT tools provide access to translation memories (TMs), terminology databases, concordance tools and, more recently, to machine translation (MT) engines. A TM is basically a repository of translated segments. During translation, the CAT tool queries the TM to search for exact or fuzzy matches of the current source segment. These matches are proposed to the user as translation suggestions. Once a segment is translated, its source and target texts are added to the TM for future queries. The integration of suggestions from an MT engine as a complement to TM matches is motivated by recent studies (Federico et al., 2012; Green et al., 2013; Läubli et al., 2013), which have shown that post-editing MT suggestions can substantially improve the productivity of professional translators. MateCat leverages the growing interest and expectations in statistical MT by advancing the state-of-the-art along three directions:

- *Self-tuning MT*, i.e. methods to train statistical MT for specific domains or translation projects;
- *User adaptive MT*, i.e. methods to quickly adapt statistical MT from user corrections and feedback;
- *Informative MT*, i.e. supply more information to enhance users' productivity and work experience.

Research along these three directions has converged into a new generation CAT software, which is both an enterprise level translation workbench (currently used by several hundreds of professional translators) as well as an advanced research platform for integrating new MT functions, running post-editing

[1]MateCat, acronym of Machine Translation Enhanced Computer Assisted Translation, is a 3-year research project (11/2011-10/2014) funded by the European Commission under FP7 (grant agreement no 287688). The project consortium is led by FBK (Trento, Italy) and includes the University of Edinburgh (United Kingdom), Université du Maine (Le Mans, France), and Translated Srl (Rome, Italy).

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations, pages 129–132, Dublin, Ireland, August 23-29 2014.

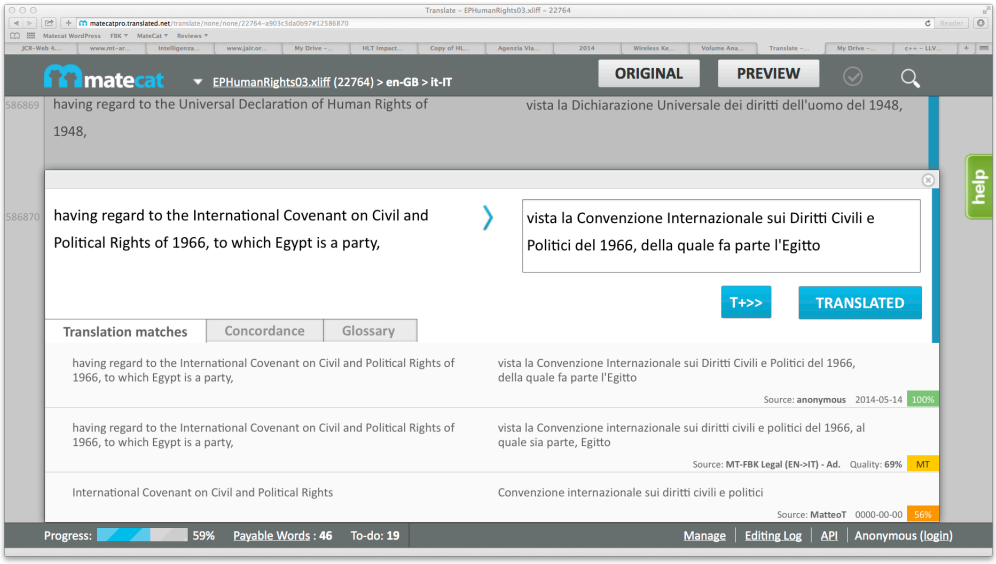

Figure 1: The MateCat Tool editing page.

experiments and measuring user productivity. The MateCat Tool, which is distributed under the LGPL open source license, combines features of the most advanced systems (either commercial, like the popular SDL Trados Workbench,[2] or free like OmegaT[3]) with new functionalities. These include: *i)* an advanced API for the Moses Toolkit,[4] customizable to languages and domains, *ii)* ease of use through a clean and intuitive web interface that enables the collaboration of multiple users on the same project, *iii)* concordancers, terminology databases and support for customizable quality estimation components and *iv)* advanced logging functionalities.

2 The MateCat Tool in a Nutshell

Overview. The MateCat Tool runs as a web-server accessible through Chrome, Firefox and Safari. The CAT web-server connects with other services via open APIs: the TM server MyMemory[5], the commercial Google Translate (GT) MT server, and a list of Moses-based servers specified in a configuration file. While MyMemory's and GT's servers are always running and available, customized Moses servers have to be first installed and set-up. Communication with the Moses servers extends the GT API in order to support self-tuning, user-adaptive and informative MT functions. The natively supported document format of MateCat Tool is XLIFF,[6] although its configuration file makes it possible to specify external file converters. The tool supports Unicode (UTF-8) encoding, including non latin alphabets and right-to-left languages, and handles texts embedding mark-up tags.

How it works. The tool is intended both for individual translators or managers of translation projects involving one or more translators. A translation project starts by uploading one or more documents and specifying the desired translation direction. Then the user can optionally select a MT engine from an available list and/or a new or existing private TM in MyMemory, by specifying its private key. Notice that the public MyMemory TM and the GT MT services are assumed by default. The following step is the *volume analysis* of the document, which reports statistics about the words to be actually translated based on the coverage provided by the TM. At this stage, long documents can be also split into smaller portions to be for instance assigned to different translators or translated at different times. The following step starts the actual translation process by opening the editing window. All source segments of the

[2] http://www.translationzone.com/
[3] http://www.omegat.org/
[4] http://www.statmt.org/moses/
[5] http://mymemory.translated.net
[6] http://docs.oasis-open.org/xliff/v1.2/os/xliff-core.html

document and their corresponding target segments are arranged side-by-side on the screen. By selecting one segment, an editing pane opens (Figure 1) including an editable field that is initialized with the best available suggestion or with the last post-edit. Translation hints are shown right below together with their origin (MT or TM). Their ranking is based on the TM match score or the MT confidence score. MT hints with no confidence score are assigned a default score. Tag consistency is automatically checked during translation and warnings are possibly shown in the editing window. An interesting feature of the MateCat Tool is that each translation project is uniquely identified by its URL page which also includes the currently edited segment. This permits for instance more users to simultaneously access and work on the same project. Moreover, to support simultaneous team work on the same project, translators can mark the status (*draft*, *translated*, *approved*, *rejected*) of each segment with a corresponding color (see Figure 1, right blue bar). The user interface is enriched with search and replace functions, a progress report at the bottom of the page, and several shortcut commands for the skilled users. Finally, the tool embeds a concordance tool to search for terms in the TM, and a glossary where each user can upload, query and update her terminology base. Users with a Google account can access a project management page which permits then to manage all their projects, including storage, deletion, and access to the editing page.

MT support. The tool supports Moses-based servers able to provide an enhanced CAT-MT communication. In particular, the GT API is augmented with feedback information provided to the MT engine every time a segment is post-edited as well as enriched MT output, including confidence scores, word lattices, etc. The developed MT server supports multi-threading to serve multiple translators, properly handles text segments including tags, and instantly adapts from the post-edits performed by each user (Bertoldi et al., 2013).

Edit Log. During post-editing the tool collects timing information for each segment, which is updated every time the segment is opened and closed. Moreover, for each segment, information is collected about the generated suggestions and the one that has actually been post-edited. This information is accessible at any time through a link in the Editing Page, named Editing Log. The Editing Log page (Figure 2) shows a summary of the overall editing performed so far on the project, such as the average translation speed and post-editing effort and the percentage of top suggestions coming from MT or the TM. Moreover, for each segment, sorted from the slowest to the fastest in terms of translation speed, detailed statistics about the performed edit operations are reported. This information, with even more details, can be also downloaded as a CSV file to perform a more detailed post-editing analysis. While the information shown in the Edit Log page is very useful to monitor progress of a translation project in real time, the CSV file is a fundamental source of information for detailed productivity analyses once the project is ended.

3 Applications.

The MateCat Tool has been exploited by the MateCat project to investigate new MT functions (Bertoldi et al., 2013; Cettolo et al., 2013; Turchi et al., 2013; Turchi et al., 2014) and to evaluate them in a real professional setting, in which translators have at disposal all the sources of information they are used to work with. Moreover, taking advantage of its flexibility and ease of use, the tool has been recently exploited for data collection and education purposes (a course on CAT technology for students in translation studies). An initial version of the tool has also been leveraged by the Casmacat project[7] to create a workbench (Alabau et al., 2013), particularly suitable for investigating advanced interaction modalities such as interactive MT, eye tracking, and handwritten input. Currently the tool is employed by Translated for their internal translation projects and is being tested by several international companies, both language service providers and IT companies. This has made possible to collect continuous feedback from hundreds of translators, which besides helping us to improve the robustness of the tool is also influencing the way new MT functions will be integrated to supply the best help to the final user.

[7] http://www.casmacat.eu

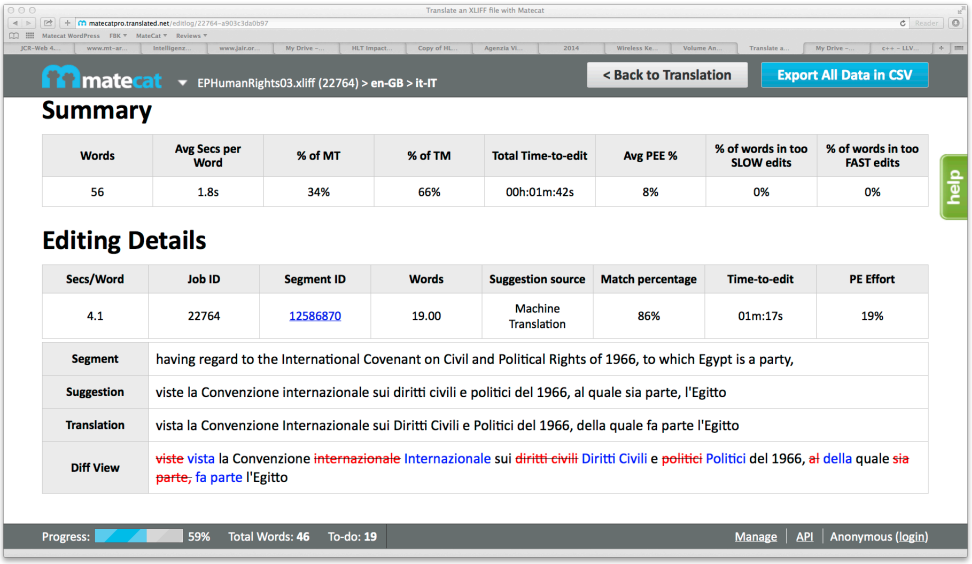

Figure 2: The MateCat Tool edit log page.

References

Vicent Alabau, Ragnar Bonk, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes Garca-Martínez, Jesús González, Philipp Koehn, Luis Leiva, Bartolomé Mesa-Lao, Daniel Oriz, Hervé Saint-Amand, Germán Sanchis, and Chara Tsiukala. 2013. Advanced computer aided translation with a web-based workbench. In *Proceedings of Workshop on Post-editing Technology and Practice*, pages 55–62.

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2013. Cache-based Online Adaptation for Machine Translation Enhanced Computer Assisted Translation. In *Proceedings of the MT Summit XIV*, pages 35–42, Nice, France, September.

Mauro Cettolo, Christophe Servan, Nicola Bertoldi, Marcello Federico, Loïc Barrault, and Holger Schwenk. 2013. Issues in Incremental Adaptation of Statistical MT from Human Post-edits. In *Proceedings of the MT Summit XIV Workshop on Post-editing Technology and Practice (WPTP-2)*, pages 111–118, Nice, France, September.

Marcello Federico, Alessandro Cattelan, and Marco Trombetti. 2012. Measuring user productivity in machine translation enhanced computer assisted translation. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)*.

Spence Green, Jeffrey Heer, and Christopher D Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 439–448. ACM.

Samuel Läubli, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow, and Martin Volk. 2013. Assessing Post-Editing Efficiency in a Realistic Translation Environment. In Michel Simard Sharon O'Brien and Lucia Specia (eds.), editors, *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*, pages 83–91, Nice, France.

Marco Turchi, Matteo Negri, and Marcello Federico. 2013. Coping with the subjectivity of human judgements in MT quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 240–251, Sofia, Bulgaria, August. Association for Computational Linguistics.

Marco Turchi, Antonios Anastasopoulos, José G.C. de Souza, and Matteo Negri. 2014. Adaptive Quality Estimation for Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL '14)*. Association for Computational Linguistics.