MT-EQuAl: a Toolkit for Human Assessment of Machine Translation Output

Christian Girardi[(1)] **Luisa Bentivogli**[(1)]
Mohammad Amin Farajian[(1,2)] **Marcello Federico**[(1)]
[(1)] FBK - Fondazione Bruno Kessler, Trento, Italy
[(2)] University of Trento, Italy
{cgirardi,bentivo,farajian,federico}@fbk.eu

Abstract

MT-EQuAl (Machine Translation Errors, Quality, Alignment) is a toolkit for human assessment of Machine Translation (MT) output. MT-EQuAl implements three different tasks in an integrated environment: annotation of translation errors, translation quality rating (*e.g.* adequacy and fluency, relative ranking of alternative translations), and word alignment. The toolkit is web-based and multi-user, allowing large scale and remotely managed manual annotation projects. It incorporates a number of project management functions and sophisticated progress monitoring capabilities. The implemented evaluation tasks are configurable and can be adapted to several specific annotation needs. The toolkit is open source and released under Apache 2.0 license.

1 Introduction

It is widely recognized within the MT field that human evaluation can play a crucial role in improving MT technology. Despite the well-known difficulties in collecting human annotations (the process is time-consuming, costly and often subjective), state of the art MT research is now moving towards integrating as much as possible human quality assessment into the MT workflow. The most commonly used human evaluation methodologies are based on absolute adequacy and fluency scores, relative ranking of alternative MT outputs, and, more recently, human post-editing. Although very useful, these methods do not provide information about the specific problems of MT systems. To address this limitation, new approaches based on human error analysis have emerged, where annotators identify and classify translation errors thus giving precise indications about specific deficiencies of the evaluated MT systems. Given the outlined trend, it is of the utmost importance to make available to the MT community tools (i) able to support large-scale annotation projects involving a great variety of languages, (ii) addressing the most required MT assessment tasks, and (iii) designed in a way to reduce as much as possible the problems related to manual annotation. MT-EQuAl is a toolkit for the manual assessment of MT output, created with the aim of addressing the above requirements. The main characteristics of MT-EQuAl are the following:

- Web-based and multi-user: allows large-scale and remotely managed annotation projects. It incorporates project management functions and sophisticated progress monitoring capabilities.
- Three different MT assessment tasks in an integrated environment: annotation of translation errors, translation quality rating (*e.g.* adequacy, fluency, relative ranking), and word alignment. An integrated environment offering different tasks can address the needs of a higher number of potential users within the MT field.
- Highly configurable tasks: possibility to evaluate a single MT output as well as two or more automatic translations in parallel, which is useful if the purpose of the annotation is to compare MT systems. Furthermore, all tasks can be adapted to specific annotation needs (see Section 2).
- Fast and well-designed annotation interfaces: particular attention was paid to the usability of the interfaces, especially for the error annotation task where a lot of annotations, often overlapping and

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations, pages 120–123, Dublin, Ireland, August 23-29 2014.

covering long sequences of words, have to be made. A fast and easy-to-use interface can reduce the problems related to manual evaluation and ensure annotation speed and data quality.

- Open source: released under Apache 2.0 license at `http://github.com/hltfbk/mt-equal`.

We think that these features give to MT-EQuAl an added value with respect to other existing annotation tools which only partially fulfill the requirements illustrated above.

Over the years, various annotation tools with different characteristics have been made available for the assessment tasks offered by our toolkit. However, none of them incorporates all the features of MT-EQuAl: either the integration in a multi-task platform, or a web-based interface, or the implementation of the error annotation task which is the most needed to support the upcoming research. The most comparable tools to MT-EQuAl are PET (Aziz et al., 2012), COSTA (Chatzitheodorou and Chatzistamatis, 2013), TAUS DQF framework,[1] translate5,[2] Blast (Stymne, 2011), and Appraise (Federmann, 2012), since they all implement translation error annotation. These tools were created for different purposes and differ in various ways among each other and with respect to MT-EQuAl. All of them except Appraise do not support multiple MT outputs, and PET, COSTA, and Blast are stand-alone tools. From the error analysis point of view, their interfaces show different levels of flexibility. PET and COSTA permit only sentence-level annotation, which is not the suitable granularity for that kind of information. Appraise offers word-level annotation but displays the MT output word by word, which does not facilitate the annotator in getting a global view of the sentence and of the errors. Finally, the translate5 and Blast interfaces show the whole MT output and allow the annotator to mark the specific portion(s) of text where an error occurs. This type of annotation is the same implemented in MT-EQuAl. However, with respect to these tools, MT-EQuAl represents a step further as one of our main design goals was usability. The MT-EQuAl error analysis interface is simple and very intuitive, and offers visualization functions aimed at reducing annotators' cognitive load, so to enable them to focus on the task itself (see Section 2.2).

2 System Overview

MT-EQuAl is a web-based application implemented using PHP and JavaScript. It takes as input several UTF-8 encoded *csv* files: the source text, the reference translation (optional), and one file for each of the MT outputs to be evaluated. This allows the evaluation of single systems as well as the comparison of multiple systems. Each row in the input *csv* files contains one evaluation item, typically one sentence. In order to annotate translation errors, the sentences must be tokenized. To this purpose, the tool accepts input files already tokenized by the user or - if needed - it applies a simple tokenization based on spaces, punctuation, and other language-dependent rules (*e.g.* a character-based tokenization is applied to Chinese texts). The source and target languages must be declared in the *csv*, so that the tool can apply the most suitable text tokenization and visualisation (*e.g.* the text can be displayed left to right and viceversa). The annotated data can be exported both in *csv* and XML format.

As regards data storage, all recorded information is permanently stored in a MySQL database. An interesting feature is that immediate persistence of data is achieved without an explicit action by the user to save the data, since every annotation is immediately sent to the server and stored in the database. Finally - being a web-based application - if the server encounters some problems, annotation is blocked and the user is notified with a warning message.

The MT-EQuAl front-end is composed of a project management interface and three annotation interfaces, one for each evaluation task.

2.1 Project Management Interface

The various project management functions implemented in the tool are accessible to the project manager through an interface which is composed of four tabs:

- **Task**. In this tab the project manager creates the task and sets its specific features. For the error analysis task, a default error typology - based on (Vilar et al., 2006) - is available, but any alternative

[1] `https://evaluation.taus.net/tools`

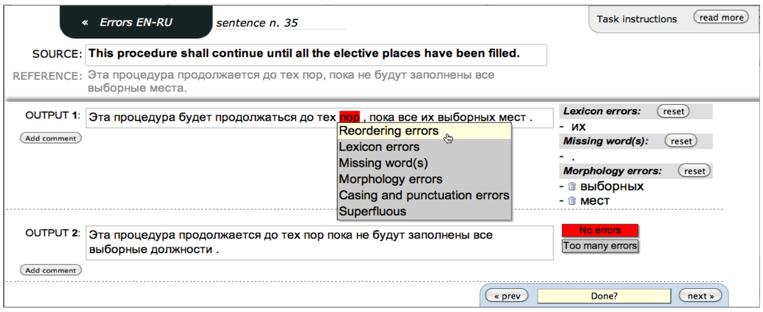
[2] `http://www.translate5.net`

tagset can be adopted. In the rating task it is possible to decide the number of points in the rating scale, while in the alignment task the number of alignment types (*e.g.* sure, possible) can be set.

- **Data**. In this tab the project manager can import the input files and apply the tokenization module if desired. Moreover, a table summarizing the data stored in the database for each task is displayed.
- **Users**. In this tab the project manager can create accounts for users and assign them to different tasks. Each user will see only the task(s) s/he has been assigned to. Users do not see other users' annotations unless they are working in "revision mode", where an existing annotation is presented for revision.
- **Annotation**. This tab contains the progress monitoring and export functions. As regards progress monitoring, a report containing real-time information about the progress of the annotation is displayed both at the task level and the user level. Moreover, the project manager can monitor user activity through the visualization of the remote client interface in read-only mode. This feature is particularly useful as it addresses the typical problems related to training and supervision of remote annotators. Regarding annotation export, data can be exported (i) for all the tasks, (ii) for each single task, and (iii) for each user. Furthermore, the annotations carried out by a user can be directly copied into another user account for revision.

2.2 Error Annotation Interface

The error annotation interface requires the annotator to identify the type of errors present in the MT output, according to the adopted error typology, and to mark their position in the text. As shown in Figure 1, the annotator is presented with the source sentence, a reference translation (optional) and the MT output(s) to be analyzed. Two buttons allow the annotator to mark the MT output as containing "*no errors*" or "*too many errors*". In order to annotate the errors, the annotator selects with the mouse the word(s) to be annotated. The selected word(s) are highlighted and, by right-clicking, the error typology menu is displayed and the suitable error type can be chosen. It is possible to annotate single words (including punctuation), spaces (*e.g.* to indicate the correct place for missing words in the candidate translation), and sequences of words (very useful especially for reordering problems which can involve entire portions of the sentence). The annotated errors are listed at the right of the corresponding sentences, subdivided by error type. If the mouse hovers over a given error instance, the corresponding word(s) appear underlined in the text. It is possible to delete single error instances (by clicking on the bin icon) or all the errors of a give type (by clicking on the "reset" button).



sentence n. 35

SOURCE: This procedure shall continue until all the elective places have been filled.

REFERENCE: Эта процедура продолжается до тех пор, пока не будут заполнены все выборные места.

OUTPUT 1: Эта процедура будет продолжаться до тех пор , пока все их выборных мест .

Lexicon errors:
- их

Missing word(s):
- .

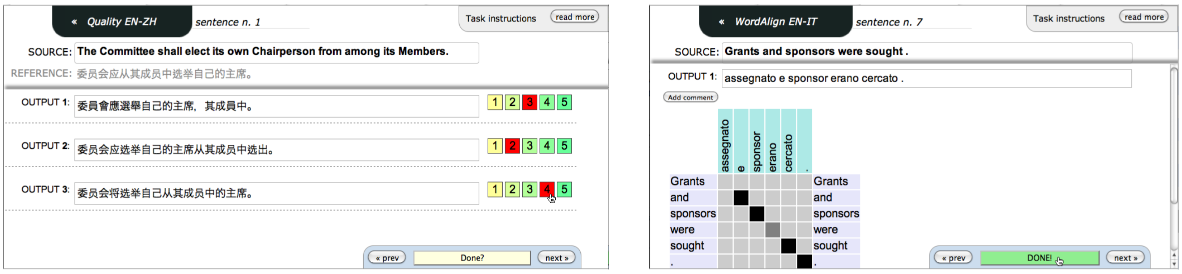
Morphology errors:
- выборных
- мест

OUTPUT 2: Эта процедура продолжается до тех пор пока не будут заполнены все выборные должности .

Figure 1: Error annotation interface configured for two MT outputs and with the default error typology.

2.3 Translation Quality Rating Interface

As shown in Figure 2(a), the quality rating interface displays the source sentence, a reference translation (optional) and the MT output(s) to be rated. When the assessor clicks on a point in the scale, the annotation is automatically saved and the point is highlighted in red. By clicking on the button "Done", the assessor confirms that the evaluation item has been completed. This layout is suitable for adequacy/fluency evaluation, ranking outputs relatively to each other, and in general all those assessment tasks that require rating MT outputs.



SOURCE: The Committee shall elect its own Chairperson from among its Members.
REFERENCE: 委员会应从其成员中选举自己的主席。
OUTPUT 1: 委員會應選舉自己的主席，其成員中。
OUTPUT 2: 委员会应选举自己的主席从其成员中选出。
OUTPUT 3: 委员会将选举自己从其成员中的主席。

(a) Quality Rating

SOURCE: Grants and sponsors were sought .
OUTPUT 1: assegnato e sponsor erano cercato .

	assegnato	e	sponsor	erano	cercato	.	
Grants	□	□	□	□	□	□	Grants
and	□	■	□	□	□	□	and
sponsors	□	□	■	□	□	□	sponsors
were	□	□	□	▨	□	□	were
sought	□	□	□	□	■	□	sought
.	□	□	□	□	□	■	.

(b) Word Alignment

Figure 2: (a) Quality Rating interface with 3 systems and a 5-point scale (b) Word Alignment interface.

2.4 Word Alignment Annotation Interface

The word alignment interface displays a traditional alignment matrix, where the rows correspond to the words of the sentence in one language and the columns to the words of its translation. Word alignments can be edited by clicking the respective matrix cells to add or remove links between words. The interface is designed to allow the alignment of discontinuous text segments. Figure 2(b) shows an alignment example where light grey, dark grey, and black cells respectively represent unlinked words, possible and sure alignments.

3 Applications of MT-EQuAl and Forthcoming Extensions

MT-EQuAl is currently being used by professional translators on English to Italian data to assess the performance of the alignment models and annotate translation errors of the MT systems developed within the MateCat project.[3] MT-EQuAl was also extensively used within an industrial project for the evaluation of commercial MT systems. To this purpose, professional translators performed error annotation and quality rating on data for three different language pairs (English to Arabic/Chinese/Russian). MT-EQuAl is being actively developed on the basis of the feedback and requirements collected from its users. We are also currently implementing the automatic computation of Inter-Annotator Agreement scores, as an additional feature to further improve the toolkit.

Acknowledgments

This work has been partially supported by the EC-funded project MateCat (ICT-2011.4.2-287688).

References

Wilker Aziz, Sheila Castilho Monteiro de Sousa, and Lucia Specia. 2012. PET: a tool for post-editing and assessing machine translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Konstantinos Chatzitheodorou and Stamatis Chatzistamatis. 2013. COSTA MT Evaluation Tool: An Open Toolkit for Human Machine Translation Evaluation. *Prague Bulletin of Mathematical Linguistics*, pages 83–89.

Christian Federmann. 2012. Appraise: an open-source toolkit for manual evaluation of mt output. *Prague Bulletin of Mathematical Linguistics*, pages 25–35.

Sara Stymne. 2011. Blast: A tool for error analysis of machine translation output. In *Proceedings of the ACL-HLT 2011 System Demonstrations*, pages 56–61, Portland, Oregon, June. Association for Computational Linguistics.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error Analysis of Machine Translation Output. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 697–702, Genoa, Italy, May. European Language Resources Association (ELRA).

[3] www.matecat.com