PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 100 OCTOBER 2013 83–89

# COSTA MT Evaluation Tool: An Open Toolkit for Human Machine Translation Evaluation

Konstantinos Chatzitheodorou[a], Stamatis Chatzistamatis[b]

[a] Aristotle University of Thessaloniki
[b] Hellenic Open University

**Abstract**

A hotly debated topic in machine translation is human evaluation. On the one hand, it is extremely costly and time consuming; on the other, it is an important and unfortunately inevitable part of any system. This paper describes COSTA MT Evaluation Tool, an open stand-alone tool for human machine translation evaluation. It is a Java program that can be used to manually evaluate the quality of the machine translation output. It is simple in use, designed to allow machine translation potential users and developers to analyze their systems using a friendly environment. It enables the ranking of the quality of machine translation output segment-by-segment for a particular language pair. The benefits of this tool are multiple. Firstly, it is a rich repository of commonly used industry criteria (fluency, adequacy and translation error classification). Secondly, it is freely available to anyone and provides results that can be further analyzed. Thirdly, it estimates the time needed for each evaluated sentence. Finally, it gives suggestions about the fuzzy matching of the candidate translations.

## 1. Introduction

Machine translation (MT) refers to the use of a machine for performing translation tasks which convert a text from a source language into a target language. Given that there may exist more than one correct translation of any given sentence manual evaluation of MT output is difficult and persistent problem. On the one hand, it is "holy grail" in MT community; on the other, it is becoming impractical because it is a time-consuming, costly and, sometimes, a subjective process. Answering questions about the accuracy and fluency, and categorizing translation errors are just as important as

Corresponding author: chatzik@itl.auth.gr

Cite as: Konstantinos Chatzitheodorou, Stamatis Chatzistamatis. COSTA MT Evaluation Tool: An Open Toolkit for Human Machine Translation Evaluation. The Prague Bulletin of Mathematical Linguistics No. 100, 2013, pp. 83–89. doi: 10.2478/pralin-2013-0014.

the MT itself. Moreover, human evaluation results give the opportunity to compare system performance and rate its progress. At the same time, researchers suffer from the lack of suitable, consistent, and easy-to-use evaluation tools.

During the DARPA GALE evaluations (Olive et al., 2011), a similar tool was designed but it was only made available to participants in the GALE program. Appraise is an other open-source tool for manual evaluation of MT output. It allows to collect human judgments on translation output, implementing annotation tasks such as translation quality checking, ranking of translations, error classification, and manual post-editing. It is used in the ACL WMT evaluation campaign (Federmann, 2012). Last but not least, PET is a stand-alone tool that has two main purposes: facilitate the post-editing of translations from any MT system so that they reach publishable quality and collect sentence-level information from the post-editing process, e.g.: post-editing time and detailed keystroke statistics (Aziz et al., 2012).

We implemented a simple stand-alone tool which facilitate MT evaluation as much as possible and to give easy access to collected evaluation data for further analysis. The typical requirements of such a tool in the framework of machine translation (MT) research are discussed in this section. Section 2 discusses usage and the corresponding graphical user interface of the tool as well the analysis of the results. Section 3 describes the evaluation criteria used and, finally, Section 4 concludes and gives an outlook on future work.

## 2. The Tool

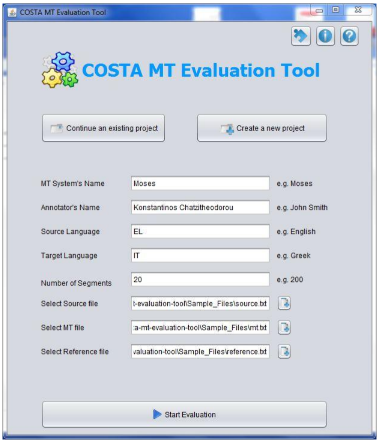
COSTA MT Evaluation Tool helps users to manually evaluate the quality of the MT output. The tool uses standard Java libraries; hence it works on any platform running a Java Virtual Machine. There is no special installation; the tool runs by just double clicking the file into any target directory.

### 2.1. Usage

Each evaluation task in COSTA MT Evaluation Tool is called a “project”. Each project requires the user to provide three parallel text files (UTF-8). Every line of these files should contain one sentence.

1. Source file contains the source sentences.
2. MT file contains the candidate translations.
3. Reference file contains the reference translations.

COSTA MT Evaluation Tool gives the opportunity to the user to choose the number of sentences and interrupt or restart the project at any time. Moreover, users can have many projects on hold. The main window of the tool is divided into 4 parts: i) the part of the source text, ii) the part of the machine translation, iii) the part of the reference translation, and iv) the part of the translation error classification as shown in Figure 1.



MT System's Name: Moses (e.g. Moses)
Annotator's Name: Konstantinos Chatzitheodorou (e.g. John Smith)
Source Language: EL (e.g. English)
Target Language: IT (e.g. Greek)
Number of Segments: 20 (e.g. 200)
Select Source file: t-evaluation-tool\Sample_Files\source.txt
Select MT file: :a-mt-evaluation-tool\Sample_Files\mt.txt
Select Reference file: valuation-tool\Sample_Files\reference.txt

*Figure 1. Main screen*

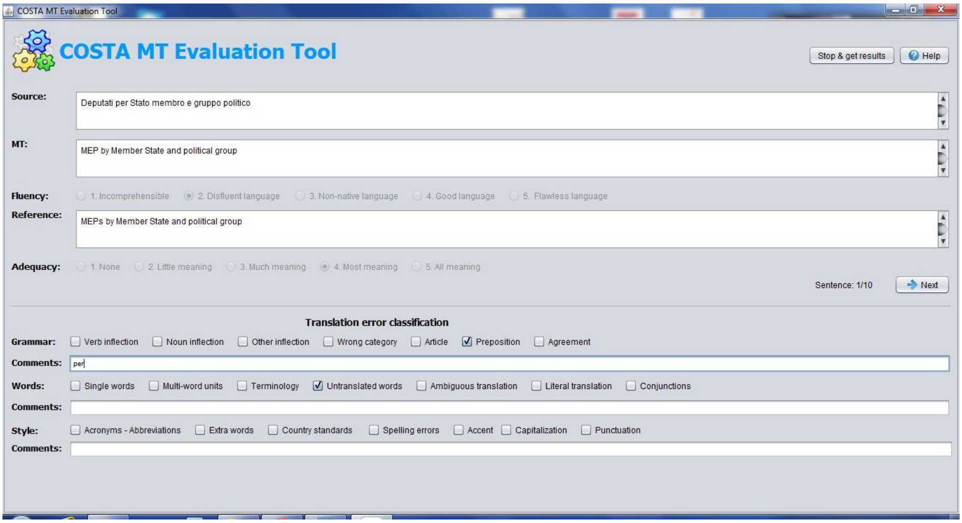
By pressing NEXT, a new sentence comes for evaluation if the current sentence is already evaluated. Annotators can stop the evaluation process by pressing Stop & Get at any time. In that case, the results for all the already evaluated sentences will be counted.

### 2.2. Getting Results

COSTA MT Evaluation Tool presents the users with automated reports on the results of evaluations as it is shown in Figure 3. Once evaluation is completed, the Tool will create a text file (UTF-8) in the target directory. The base filename consists of <the system's name> + <the annotator's name> + <the source> and <target> languages, followed by _results.txt. For instance, a typical name for a Moses English into Greek MT system with annotator Mr. Smith will be:

Moses_Smith_EN_GR_results.txt

This file can easily be imported into Excel, SPSS, MATLAB, and most other statistical software suites for further analysis and significance testing. In addition, it can be read by other tools or machine learning algorithms in order to estimate the quality of future MT outputs. The header of the file contains all the information for the system as well as the average fluency and adequacy scores and the count of the errors. Moreover, each line of the rest of this file contains the analytical results for each evaluated sentence.



**Source:** Deputati per Stato membro e gruppo politico

**MT:** MEP by Member State and political group

**Fluency:** ○ 1. Incomprehensible ● 2. Disfluent language ○ 3. Non-native language ○ 4. Good language ○ 5. Flawless language

**Reference:** MEPs by Member State and political group

**Adequacy:** ○ 1. None ○ 2. Little meaning ○ 3. Much meaning ● 4. Most meaning ○ 5. All meaning

Sentence: 1/10

**Translation error classification**

**Grammar:** □ Verb inflection □ Noun inflection □ Other inflection □ Wrong category □ Article ☑ Preposition □ Agreement

**Comments:** per

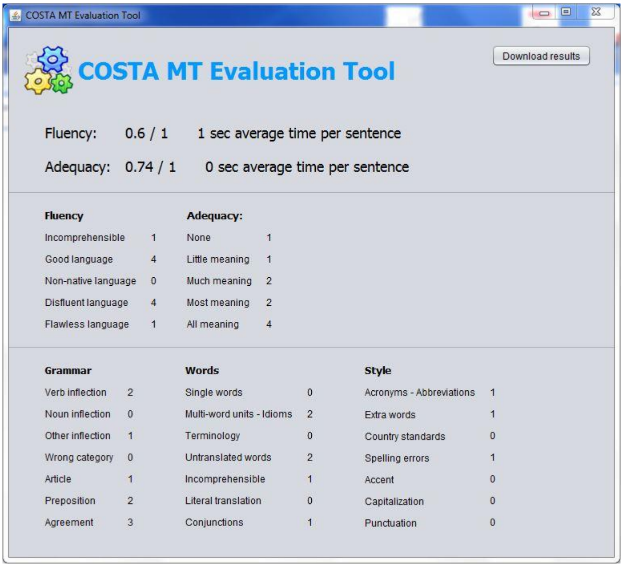
**Words:** □ Single words □ Multi-word units □ Terminology ☑ Untranslated words □ Ambiguous translation □ Literal translation □ Conjunctions

**Comments:**

**Style:** □ Acronyms - Abbreviations □ Extra words □ Country standards □ Spelling errors □ Accent □ Capitalization □ Punctuation

**Comments:**

*Figure 2. Evaluation environment*



Fluency: 0.6 / 1 1 sec average time per sentence

Adequacy: 0.74 / 1 0 sec average time per sentence

| **Fluency** | | **Adequacy:** | |
|---|---|---|---|
| Incomprehensible | 1 | None | 1 |
| Good language | 4 | Little meaning | 1 |
| Non-native language | 0 | Much meaning | 2 |
| Disfluent language | 4 | Most meaning | 2 |
| Flawless language | 1 | All meaning | 4 |

| **Grammar** | | **Words** | | **Style** | |
|---|---|---|---|---|---|
| Verb inflection | 2 | Single words | 0 | Acronyms - Abbreviations | 1 |
| Noun inflection | 0 | Multi-word units - Idioms | 2 | Extra words | 1 |
| Other inflection | 1 | Terminology | 0 | Country standards | 0 |
| Wrong category | 0 | Untranslated words | 2 | Spelling errors | 1 |
| Article | 1 | Incomprehensible | 1 | Accent | 0 |
| Preposition | 2 | Literal translation | 0 | Capitalization | 0 |
| Agreement | 3 | Conjunctions | 1 | Punctuation | 0 |

*Figure 3. Evaluation results*

## 3. Evaluation Metrics

COSTA MT Evaluation Tool enables users to evaluate the MT performance using the two main criteria:

1. Fluency and adequacy
2. Translation error classification

### 3.1. Fluency and Adequacy

The objective of the fluency evaluation is to determine how "fluent" a translation appears to be, without taking into account the correctness of the information. The evaluation does this segment-by-segment on a 1–5 scale without referring to any reference text. The objective of the adequacy evaluation is to determine the extent to which all of the content of a text is conveyed, regardless of the quality of the language in the candidate translation. The evaluation does this segment-by-segment on a 1–5 scale. The annotator is given the following definitions of adequacy and fluency (Koehn, 2007):

| **Fluency** | **Adequacy** |
|---|---|
| 5. Flawless language | 5. All meaning |
| 4. Good language | 4. Most meaning |
| 3. Non-native language | 3. Much meaning |
| 2. Disfluent language | 2. Little meaning |
| 1. Incomprehensible | 1. None |

Since, recent evaluation campaigns have shown that judgments of fluency and adequacy are closely related, COSTA MT Evaluation Tool firstly asks annotators to evaluate the fluency without referring to any reference text and secondly the adequacy with reference to the reference text (White, 1995). The evaluation of translation error classification is optional.

### 3.2. Translation Error Classification

During the evaluation of fluency and adequacy, COSTA MT Evaluation Tool offers users the option to count and categorize errors. This type of evaluation can provide a descriptive framework that reveals relationships between errors. Furthermore, it can also help the evaluator to map the extent of the effect in chains of errors, allowing comparison among MT systems. At the same time, we propose these criteria as a new methodology of human translation error classification.

In total, there are three main categories each with seven subclasses. These categories were identified by observing the most frequent error types in MT outputs among Moses-based (Koehn et al., 2007) and free MT systems such as Google Translate and Bing Translator.

The first category concerns the grammatical and the linguistic accuracy of the machine translated texts. The second category concerns the use of the vocabulary and the third the format and style of the produced texts. Analytically, translation error classification works to the following criteria:

**Linguistic**

| | |
|---|---|
| Verb inflection | Incorrectly formed verb, or wrong tense. |
| Noun inflection | Incorrectly formed noun (e.g. as nominative nouns in apposition). |
| Other inflection | Incorrectly formed adjective or adverb. |
| Wrong category | Category error (e.g. noun vs. verb). |
| Article | Absent or unneeded article.<br>(e.g. The London vs. London) |
| Preposition | Incorrect, absent or unneeded preposition. |
| Agreement | Incorrect agreement between subject-verb, noun-adjective, past participle agreement with preceding direct object, etc. |

**Words**

| | |
|---|---|
| Single words | Sentence elements ordered incorrectly. |
| Multi-word units | Incorrect translation of multi-word expressions and idioms (e.g. to pay a visit). |
| Terminology | Incorrect terminology. |
| Untranslated words | Word not in dictionary. |
| Ambiguous translation | Ambiguous target language. |
| Literal translation | Word-for-word translation. |
| Conjunctions | Failure to reconstruct parallel constituents after conjunction, or failure to identify boundaries of conjoined units. |

**Style**

| | |
|---|---|
| Acronyms – Abbreviations | Incorrect abbreviations, acronyms and symbols. |
| Extra words | Extra words in target language. |
| Country standards | Incorrect format of dates, addresses, currency etc. |
| Spelling errors | Misspelled words. |
| Accent | Incorrect accents. |
| Capitalization | Incorrect upper or lower case. |
| Punctuation | Punctuation is incorrect, absent or unneeded. |

There are also three additional boxes to the bottom of the main screen, one for each translation error category, and where the evaluator could add comments.

## 4. Conclusion and Future Work

We have presented a simple tool for manual evaluation of MT. It is simple in use, designed to allow potential MT users and developers to analyze their systems using a friendly environment. It enables the ranking of the quality of MT output segment-by-segment for a particular language pair. At the same time, we propose these criteria as a new methodology of human translation error classification. Our future work includes:

1. Multiple MT systems evaluation
2. Multiple Reference evaluation
3. Extraction of feature that can be analyzed by machine learning algorithms for the estimation of the MT quality without reference translation.

The tool is available for download at:
`https://code.google.com/p/costa-mt-evaluation-tool/`

## Bibliography

Aziz, Wilker, Sheila Castilho Monteiro de Sousa, and Lucia Specia. 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 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Federmann, Christian. Appraise: An open-source toolkit for manual evaluation of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, 98:25–35, September 2012.

Koehn, Philipp. *Statistical Machine Translation*. Cambridge University Press, 2007.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Olive, Joseph, Caitlin Christianson, and John McCary. Handbook of natural language processing and machine translation: DARPA global autonomous language exploitation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Springer, 2011.

White, John S. Approaches to black box MT evaluation. In *MT Summit V Proceedings*, July 1995.

**Address for correspondence:**
Konstantinos Chatzitheodorou
`chatzik@itl.auth.gr`
Aristotle University of Thessaloniki
University Campus, GR-54124 Thessaloniki, Greece