Sisyphos-II
MT-Evaluation tools

D. Kaumans, Gr. Thurmair, Linguatec
2012-04-27

1 Introduction

This is a set of tools to evaluation MT output interactively[1]. It supports the main non-automatic evaluation metrics used today, which is:

- Determination of the quality of an MT output, in terms of adequacy and fluency (called 'absolute evaluation'). This answers the question 'How good is the MT output'.
- Determination of the quality of an MT output in comparison to another MT output (called 'comparative evaluation'). It answers the question 'Which output (of two systems) is better?'. Note that it does not answer the question on the real output quality.
- Determination of the distance of an MT output to a correct human translation (called 'postediting evaluation'). It answers the question on the effort needed to create a good translation from a raw MT output, both in terms of edit distance and of required postediting time.

Three little standalone tools have been created to support these evaluations; they can be given to external evaluators (freelancers etc.), together with a pack of evaluation data, so evaluators can process them offline, and return the results. This workflow can be seen as an alternative to online-access tools as used in WMT.

2 Installation

Installation requirement is a Java runtime (1.7 and higher).

The tools are displayed in a zip file. They must be extracted into a directory of users' choice; this directory will contain both the programme and the files used for processing. Below this working directory there is a directory 'lib' containing an auxiliary jar-file (for XML code handling).

The programmes are called:

- AbsoluteEvaluation.jar
- ComparativeEvaluation.jar
- PostEditingEvaluation.jar

The installation package also contains three example files, for easier startup, and the DTDs for the evaluation files.

It also contains this documentation.

[1] The first version of Sisyphus was created by the Belgian METAL team in 1987, in pre-Windows times, to speed up system development. The kind of tools is still needed…

3 Functionality

The main functionality of the tools is:

- Import of a new evaluation 'package'
- Interactive support of the evaluation procedure
- Creation of result files containing statistics.

The **data flow** is depicted in Fig. 1. The main files are the translation and evaluation xml files. Each tool works with two XML files, called 'translation-{abs|comp|post}.xml' (created by the import function from the source and target language files produced by the MT systems), storing the data to be evaluated, and 'evaluation-{abs|comp|post}.xml', created during interactive evaluation, storing the evaluation result. The file names are fixed. The result of the evaluation is stored in the evaluation xml files; an overview file can be created containing basic statistics.

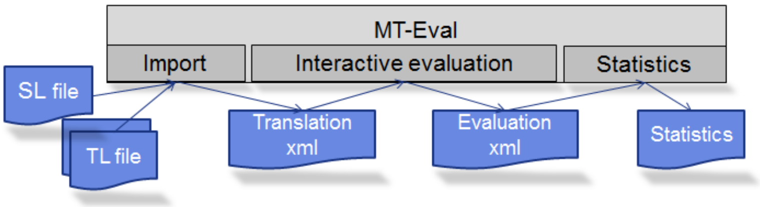


Fig. 1: Data flow

3.1 Import of evaluation data

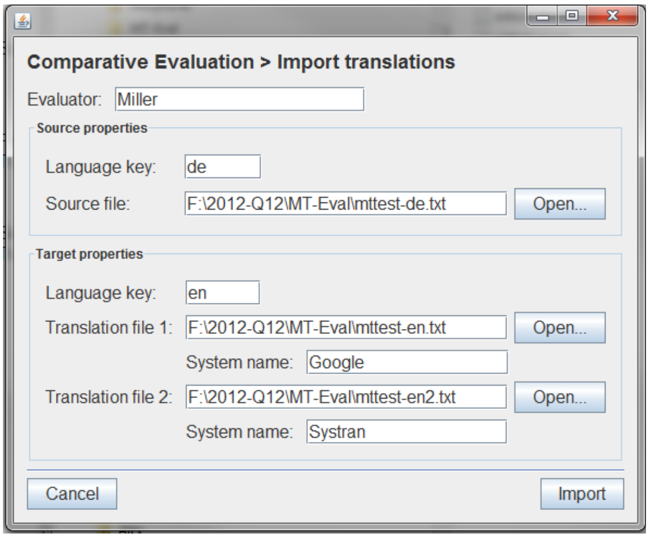
The tool expects the evaluation data in the following format:

- UTF8 character code
- one line per sentence
- one file per language
- parallel numbering of sentences.

This is the basic format as produced by systems like MOSES.

By clicking on [Import] in one of the tools, the import screen is displayed, asking for

- The name / id of the evaluator
- Source and target language involved
- File name of the source and the target language(s) file
- Source of translation (which system did the translation)


Comparative Evaluation > Import translations

Evaluator: Miller

Source properties

Language key: de

Source file: F:\2012-Q12\MT-Eval\mttest-de.txt

Target properties

Language key: en

Translation file 1: F:\2012-Q12\MT-Eval\mttest-en.txt

System name: Google

Translation file 2: F:\2012-Q12\MT-Eval\mttest-en2.txt

System name: Systran

Fig. 2: Data import

With this information, an XML file is created which is used during the evaluation process. Its name is 'translations-{comp|abs|post}.xml' (depending on the tool). This file is used as input by the interactive evaluation process.

3.2 Interactive Evaluation

The evaluation interaction differs depending on the tool. It displays sentences with their translations, in random order. Each tool has a section where the source and translations are displayed, and below that a section with the evaluation options. At the bottom of the screen, buttons for the different system possibilities are located:

Navigation in the evaluation data is done with [Next] and [Previous]; [End Session] terminates the current session, [Import] creates a new evaluation file, [Review] accesses evaluation results of a previous session, and [Statistics] displays a table with evaluation results.

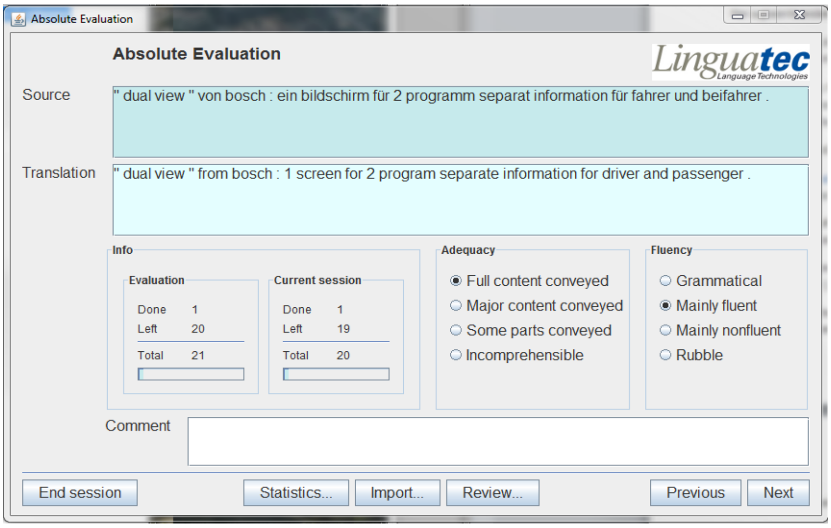
3.2.1 Absolute Evaluation

For a given translation, its quality is determined.

The translation is displayed, and users can evaluate the adequacy and the fluency of the translation. Each time a 4-point scale is presented, users select one of the options in both areas.

- For adequacy, the options are: { *full content conveyed | major content conveyed | some parts conveyed | incomprehensible* }
- For fluency, the options are:

  { *grammatical | mainly fluent | mainly nonfluent | rubble* }



Absolute Evaluation

Source: " dual view " von bosch : ein bildschirm für 2 programm separat information für fahrer und beifahrer .

Translation: " dual view " from bosch : 1 screen for 2 program separate information for driver and passenger .

Info

Evaluation		Current session	
Done	1	Done	1
Left	20	Left	19
Total	21	Total	20

Adequacy
● Full content conveyed
○ Major content conveyed
○ Some parts conveyed
○ Incomprehensible

Fluency
○ Grammatical
● Mainly fluent
○ Mainly nonfluent
○ Rubble

Comment

Fig. 3: Absolute evaluation

By clicking on [Next] the result is stored, and the next sentence is presented, [Previous] displays previous evaluation data, for corrections.

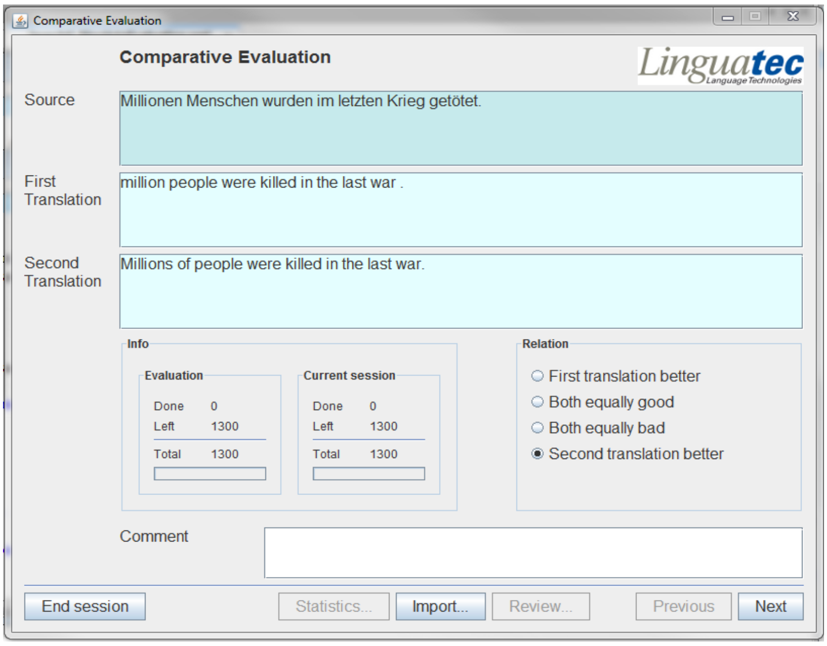
3.2.2 Comparative evaluation

The tool compares the quality of two translations against each other.

Two translations of a given sentence are displayed, for comparison. Users can decide which one is better, on a 4-point scale.

Comparison options are: { *first translation better | both equally good | both equally bad | second translation better* }.

The sequence of translation1 and translation2 is randomized to avoid biased evaluation (i.e. translation 1 is sometimes displayed first, sometimes second).



Comparative Evaluation

Source: Millionen Menschen wurden im letzten Krieg getötet.

First Translation: million people were killed in the last war .

Second Translation: Millions of people were killed in the last war.

Info

Evaluation		Current session	
Done	0	Done	0
Left	1300	Left	1300
Total	1300	Total	1300

Relation
○ First translation better
○ Both equally good
○ Both equally bad
● Second translation better

Comment

Fig. 4: Comparative Evaluation

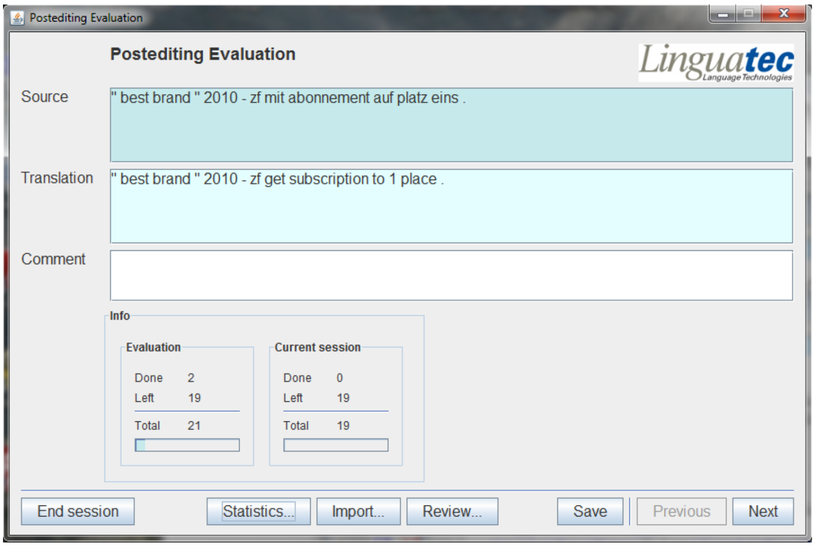
By clicking on [Next] the result is stored, and the next sentence is presented, [Previous] displays previous evaluation data, for corrections.

3.2.3 Postediting evaluation

The tool measures the time needed to postedit a translation output into a correct format (HTER). It can afterwards also be used to compute the edit distance.

The translation of the source sentence is displayed. The translation field is editable, so users can edit the MT output.

The time from the first display of the sentence until the pressing of the [Save] button is stored (in seconds). There is also a 'comment' field which can be used to give comments on the translation / postediting. Navigation is done with the [Next] and [Previous] buttons.



Postediting Evaluation

Source: " best brand " 2010 - zf mit abonnement auf platz eins .

Translation: " best brand " 2010 - zf get subscription to 1 place .

Comment:

Info

Evaluation		Current session	
Done	2	Done	0
Left	19	Left	19
Total	21	Total	19

Fig. 5: Postediting evaluation

3.2.4 Common features

All tools have common features; this relates mainly to the concepts of sessions. Usually people cannot do the complete evaluation in one go, but do it in several sessions.

Within a session, users can move back and forth in the evaluated sentences, and also go back and correct an evaluation, by clicking on [Previous]. Also, a statistics on the progress of the current session is displayed, as well as of the whole task. This is for motivation reasons. If users want to stop they click on [End session].

If a session is closed, another XML file containing the evaluation results is written / updated. This file is called evaluation-{abs|comp|post}.xml.

Users can also access the evaluations of a previous session by clicking on [Review]. This allows them to change evaluation results from previous sessions (i.e. modify the evaluation-xml file). The system displays the evaluated sentence pairs, users can click on the one they want to change, and click on [edit] to edit it. This is relevant as sometimes the evaluation criteria change after having seen the first couple of data.

3.3 Evaluation

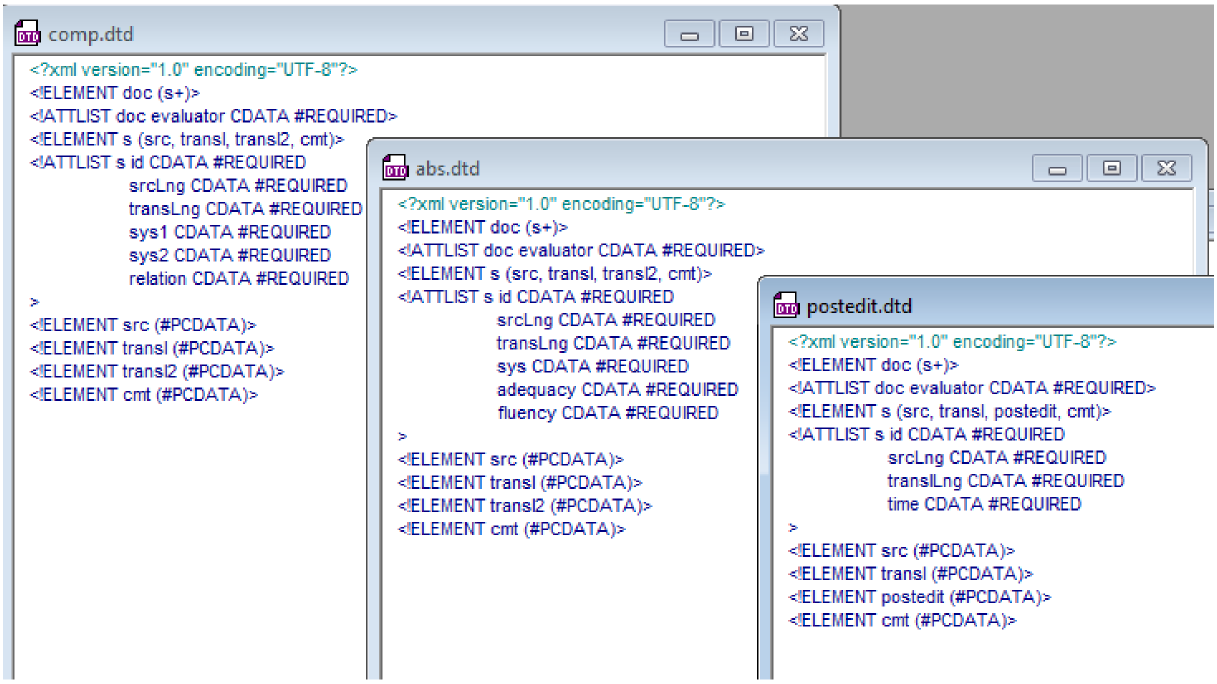
Users have the option to see an overview of the evaluation at any time of their work. They can click on [Statistics], and then a first statistics on the number of sentences, and how they were evaluated, is shown. Users can print this into a file.

For more detailed evaluation, the evaluation XML files used by the tools must be consulted, like for inter-annotator agreement, or for edit-distance computation. The format of the different tools differs slightly; the DTD of them is given in Fig. 6. Examples of the files are given in Fig. 7 (for easier processing, all XML markups are in separate lines).

From this XML file, the interesting data can be extracted, e.g.:

- for Kappa calculation: sentence IDs, evaluator, evaluation results
- for edit distance calculation: translated text and postedit text, etc.

Users should save away the evaluation XML files from the working directory of the MT-Eval tools, to protect them from being overwritten by the next evaluation task.

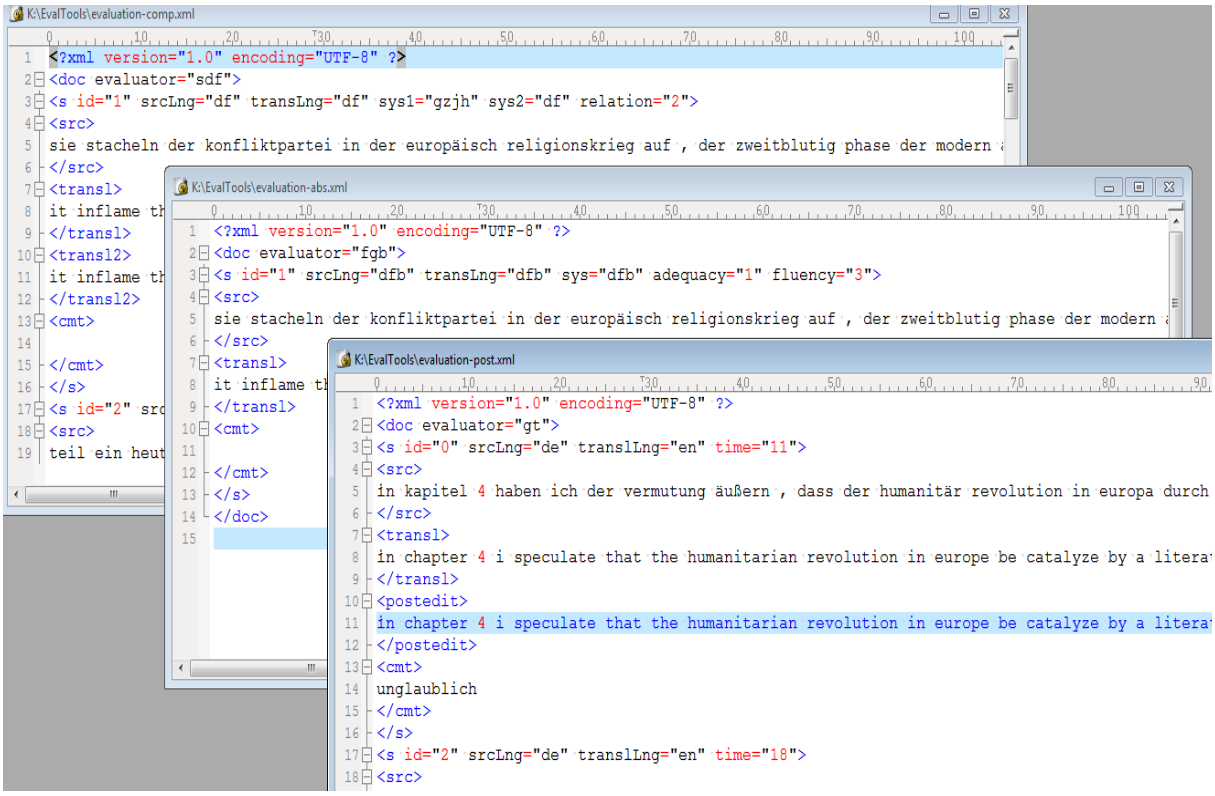


```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT doc (s+)>
<!ATTLIST doc evaluator CDATA #REQUIRED>
<!ELEMENT s (src, transl, transl2, cmt)>
<!ATTLIST s id CDATA #REQUIRED
            srcLng CDATA #REQUIRED
            transLng CDATA #REQUIRED
            sys1 CDATA #REQUIRED
            sys2 CDATA #REQUIRED
            relation CDATA #REQUIRED
>
<!ELEMENT src (#PCDATA)>
<!ELEMENT transl (#PCDATA)>
<!ELEMENT transl2 (#PCDATA)>
<!ELEMENT cmt (#PCDATA)>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT doc (s+)>
<!ATTLIST doc evaluator CDATA #REQUIRED>
<!ELEMENT s (src, transl, transl2, cmt)>
<!ATTLIST s id CDATA #REQUIRED
            srcLng CDATA #REQUIRED
            transLng CDATA #REQUIRED
            sys CDATA #REQUIRED
            adequacy CDATA #REQUIRED
            fluency CDATA #REQUIRED
>
<!ELEMENT src (#PCDATA)>
<!ELEMENT transl (#PCDATA)>
<!ELEMENT transl2 (#PCDATA)>
<!ELEMENT cmt (#PCDATA)>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT doc (s+)>
<!ATTLIST doc evaluator CDATA #REQUIRED>
<!ELEMENT s (src, transl, postedit, cmt)>
<!ATTLIST s id CDATA #REQUIRED
            srcLng CDATA #REQUIRED
            translLng CDATA #REQUIRED
            time CDATA #REQUIRED
>
<!ELEMENT src (#PCDATA)>
<!ELEMENT transl (#PCDATA)>
<!ELEMENT postedit (#PCDATA)>
<!ELEMENT cmt (#PCDATA)>
```

Fig. 6: DTDs of evaluation files



K:\EvalTools\evaluation-comp.xml

```
<?xml version="1.0" encoding="UTF-8" ?>
<doc evaluator="sdf">
<s id="1" srcLng="df" transLng="df" sys1="gzjh" sys2="df" relation="2">
<src>
sie stacheln der konfliktpartei in der europäisch religionskrieg auf , der zweitblutig phase der modern
</src>
<transl>
it inflame th
</transl>
<transl2>
it inflame th
</transl2>
<cmt>

</cmt>
</s>
<s id="2" src
<src>
teil ein heut
```

K:\EvalTools\evaluation-abs.xml

```
<?xml version="1.0" encoding="UTF-8" ?>
<doc evaluator="fgb">
<s id="1" srcLng="dfb" transLng="dfb" sys="dfb" adequacy="1" fluency="3">
<src>
sie stacheln der konfliktpartei in der europäisch religionskrieg auf , der zweitblutig phase der modern
</src>
<transl>
it inflame tl
</transl>
<cmt>

</cmt>
</s>
</doc>
```

K:\EvalTools\evaluation-post.xml

```
<?xml version="1.0" encoding="UTF-8" ?>
<doc evaluator="gt">
<s id="0" srcLng="de" translLng="en" time="11">
<src>
in kapitel 4 haben ich der vermutung äußern , dass der humanitär revolution in europa durch
</src>
<transl>
in chapter 4 i speculate that the humanitarian revolution in europe be catalyze by a litera
</transl>
<postedit>
in chapter 4 i speculate that the humanitarian revolution in europe be catalyze by a litera
</postedit>
<cmt>
unglaublich
</cmt>
</s>
<s id="2" srcLng="de" translLng="en" time="18">
<src>
```

Fig. 7: Examples of evaluation files