PET: a Tool for Post-editing and Assessing Machine Translation

Wilker Aziz[†], Sheila C. M. de Sousa[†] and Lucia Specia[§]

† Research Institute in Information and Language Processing
University of Wolverhampton
Stafford Street, Wolverhampton, WV1 1SB, UK
{w.aziz, sheila.castilhomonteirodesousa}@wlv.ac.uk

§ Department of Computer Science
University of Sheffield
Regent Court, 211 Portobello, Sheffield, S1 4DP, UK
L.Specia@sheffield.ac.uk

Abstract

Given the significant improvements in Machine Translation (MT) quality and the increasing demand for translations, post-editing of automatic translations is becoming a popular practice in the translation industry. It has been shown to allow for larger volumes of translations to be produced, saving time and costs. In addition, the post-editing of automatic translations can help understand problems in such translations and this can be used as feedback for researchers and developers to improve MT systems. Finally, post-editing can be used as a way of evaluating the quality of translations in terms of how much effort these translations require in order to be fixed. We describe a standalone 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.

Keywords: Machine Translation, Post-editing of Machine Translation, Evaluation of Machine Translation

1. Introduction

Post-editing Machine Translation (MT) output is now seen as a potentially successful way of incorporating MT into human translation workflows in order to minimize time and costs in the translation industry. The editing of semi-automatic translations is a common practice among users of Translation Memory (TM) tools, which provide user-friendly and functional environments for translators, for example: SDL Trados[1], Wordfast[2] and Déjà Vu X2[3]. Many of these tools now incorporate MT systems. For example, SDL Trados provides the same post-editing interface for TMs and a few MT systems. Although less common, some MT systems also enable the use of TMs besides providing their own translation output. These include systems such as Google Translate[4] and Systran[5].

However, existing post-editing environments have three main limitations: restricted availability and flexibility, and lack of detailed statistics from post-editing jobs. Most of them are proprietary tools only available as part of a major (and more expensive) product distribution. Apart from a few options, mostly regarding their interface, they cannot be modified in any way. Furthermore, these tools generally only allow the post-editing of one or a very small number of specific specific MT systems, which restricts their application. As such, they do not allow, for example, the comparison of translations produced by different MT systems in terms of post-editing effort.

An important use of post-editing is the collection of information that can be used for measuring translation quality and diagnosing translation problems, but this broadly neglected in existing tools. An exception is Translog[6], a tool developed specifically for the purpose of logging very detailed information about operations performed on a text. Translog keeps track of each and every move of the translator/writer, presenting the possibility of replaying the writing process in full, as if it were a video. The post-editing of machine translations is one of the applications of this tool, however it does not provide specific facilities for translation post-editing, for example, access to external resources such as dictionaries, the possibility of assessing translations or defining restrictions on the post-editing process.

Another example of tool that allows collecting some information from the post-editing process is Caitra[7]. Caitra (Koehn, 2009) is aimed at interactive translation, where the translator can choose to use the assistance provided by the tool, such as (i) predictions - the tool proposes suggestions for sentence completion; (ii) options - the tool displays multiple alternatives available to translate the sentence; and (iii) MT post-editing - the tool provides the best complete translation from the MT system, which the user can accept or edit if necessary. Caitra collects information related to post-editing time, such as the time spent on different types of edits and pauses, and statistics such as types of edits and

[1] http://www.trados.com/en/
[2] http://www.wordfast.net/
[3] http://www.atril.com/en/software/deja-vu-x-professional
[4] http://translate.google.com/
[5] http://www.systran.co.uk/
[6] http://www.translog.dk/
[7] http://www.caitra.org/

keystrokes. However, it uses a specific MT system, Moses (Koehn et al., 2007), and therefore comparisons of multiple MT systems are not possible. In addition, it has limited additional facilities for post-editing and does not allow subjective assessments of translations.

A similar standalone tool for the post-editing of multiple MT system was developed and used during the DARPA GALE evaluations (Olive et al., 2011). However, this tool was only made available to participants in the GALE program. To the best of our knowledge, it does not include common facilities for post-editors, such as dictionaries, etc.

In this paper we present PET, a simple standalone tool that allows the post-editing of any MT system and records (by default) the following information at sentence-level, among others: time, customizable quality scores, timestamped edits, keystrokes, and edit distance from the original MT and its post-edited version. PET can also be used to set constraints on a post-editing task, such as a maximum post-editing time on a per-sentence basis. As an MT system-independent tool, PET makes it possible to collect post-editing/revision information in a controlled way for multiple MT systems. The tool, along with some of its customization options, is presented in Section 2. We also describe two uses of the tool: (i) the collection of post-editings to compare different MT systems and to compare post-editing against translating from scratch (Section 3.1.), and (ii) the collection of post-editings to serve as training data to build and compare translation quality prediction models (Section 3.2.).

2. PET: a Post-Editing Tool

The Post-Editing Tool (PET) was developed mainly to serve the purpose of collecting implicit and explicit effort indicators. While the main use of PET is the revision or post-editing of draft translations, such a tool can also collect information regarding translation from scratch. Among several possibly interesting effort indicators one might seek, the most appealing to us is the time spent on performing a task.

PET was developed in an object-oriented fashion using standard Java-6 libraries, hence it works on any platform running a Java Virtual Machine. For post-editing, the interface displays source and target language texts in two columns. Figure 1 shows these annotation window, where the left hand side column is for the visualization of the source text and the right hand side column enables the editing of its translation. For translation, the right hand side column is empty. The unit of text to translate or edit is defined in the input file: it can be a sentence, like in Figure 1, a paragraph, phrase, or a text of any length. Units are seen in context, that is, surrounded by some preceding and forthcoming units.

Each unit is translated/edited at a time and navigation is achieved using the navigation bar on the right hand side. For the active unit, an extra box at the top of the window can display additional information, such as the original translation, an alternative translation (from other MT system, for example), or a reference (human) translation.

Once a unit is completed, an assessment window can be displayed to collect additional information about that unit, for example, overall translation quality or post-editing effort scores, as shown in Figure 2. The type of assessment to be collected, as well as its scale, is set in a configuration file. Any number of assessment questions can be used, and multiple windows will be created if necessary. Optional comments regarding the assessment are also possible.

2.1. Translation/post-editing jobs

Units to be translated/edited are grouped in "jobs". A job is therefore a sequence of units assigned to a human annotator. It may contain units to translate or edit, or a mixture of both, where each unit is identified by a unique index. For any unit, the only mandatory information to be provided as part of the input file is the source text. In the case of post-editing/revision, a job must contain a source text and a draft translation (machine or human). In addition, it may contain the reference translation.

Units may also contain a number of attributes that can be made visible to the translator or kept in the input/output files only, such as the "producer" of the translation, to indicate the MT system or human translator who produced a given draft translation. Other default attributes include the maximum length allowed for the translation or the maximum time allowed for its editing. Any extra attributes that do not require changing the tool's behavior can be added to the input file.

In addition, PET's API provides an interface to add new attributes in the form of constraints and events allowing to further customize a job. This is done by stating that a specific class adds a behavior to the job and it is controlled by a set of attributes. For instance, for the definition of the maximum post-editing time for a unit, a "maxtime" attribute enables the constraint "Deadline" which triggers the event "EndTaskByForce". As a consequence, PET forces a unit to end once a specific amount of time has passed since the

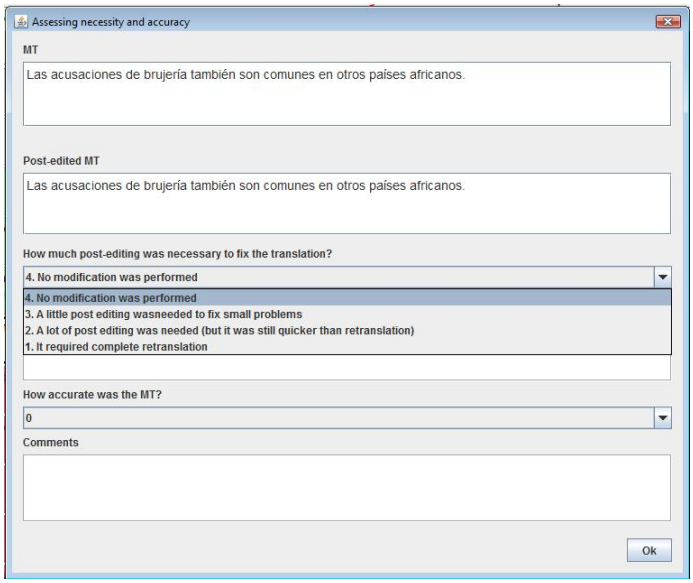

Figure 2: Assessment window

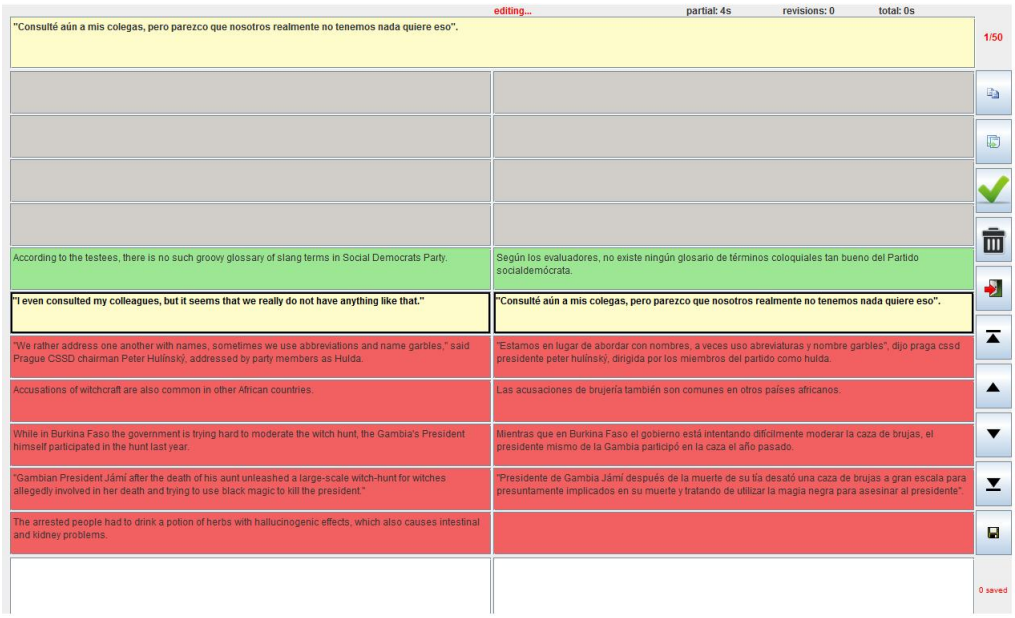

Figure 1: Annotation window

translation/editing of the unit started.

A job can be paused, interrupted and re-stared at any moment after a unit is completed. Intermediate files are generated and the overall job time is computed from summing over the editing time of individual units.

2.2. User interface

Using a simple configuration file, PET allows the customization of the following, among other features:

- how many units are displayed at a time
- what is displayed at the top box (e.g source text, reference text)
- which attributes are displayed
- whether explicit assessments, and which assessments, are requested in the assessment window
- whether a unit should be hidden before its editing time starts to be recorded
- whether a unit can be edited multiple times

Other customizable features are described in PET's user manual. In addition, the interface uses two customizable boxes (bottom of the window, Figure 1). These boxes may render additional information about words and/or phrases in the source (left box) and in the translation (right box) texts. Useful sub-sentential information includes paraphrases, alternative translations, definitions and links to external sources with relevant information. This additional information needs to be provided in an XML file for a given translation/post-editing job. Entries are key-value pairs where the keys are words and/or phrases that may occur in the text and values are the additional information for those words/phrases. By default the tool displays only the values that match the content of the active unit. The default rendering and selection criteria may be overridden using PET's API.

2.3. Input/output format and quality indicators

The input format for the tool is XML, which facilitates the use of new attributes. For example, a post-editing job can be defined by the following basic elements for each unit, as shown in Figure 3: type of job (type), identifier of the unit (id), source file with source and reference texts (S producer and R producer), and system that produced the translation (MT producer).

The outcome of a task is organized as an annotation object per unit. This object contains the final translation, the effort indicators (e.g., time) obtained during the translation/editing and any additional assessment. If PET is set to allow multiple edits of the same unit, for every unit there will be a list of annotation objects marked with revision stamps.

PET provides a few built-in effort indicators and assessment types, but many others can be added via the PET's API. The default effort indicators and assessment types are:

- Editing time: time spent translating or editing a unit;
- Assessing time: time spent assessing a unit (quality, effort, etc.);
- Assessment tag: actual assessment tag amongst a predefined set;
- Keystrokes: number of keys pressed during the post-editing grouped by type of keys (deletion, alphanumeric, etc.);
- HTER: Human Translation Edit Rate (Snover et al., 2006): edit distance between the draft translation and its post-edited version.

Time, one of the most important indicators collected by the tool, is computed from the moment the target box of the unit is clicked to the moment the task is completed (either the job is closed or the navigation button "next" is pressed).

```
<task type="pe" id="3">
  <S producer="xfiles.en">Excuse me.</S>
  <R producer="xfiles.pt">- Com licença,</R>
  <MT producer="google">Desculpe-me.</MT>
</task>
```

Figure 3: Extract of input file

```
−<task id="3" status="FINISHED" type="pe">
   <S producer="xfiles.en">Excuse me.</S>
   <R producer="xfiles.pt">- Com licença,</R>
   <MT producer="google">Desculpe-me.</MT>
  −<annotations revisions="1">
    −<annotation r="1">
       <PE producer="pet">Desculpe-me.      </PE>
       <indicator id="editing">3s</indicator>
       <indicator id="assessing">0s</indicator>
       <comment/>
     </annotation>
   </annotations>
</task>
```

Figure 4: Extract of ouput file

The outcome of a job is also stored in an XML file, such as the one in Figure 4. In this case, only the post-edited version and time indicators were produced (for editing and assessing the unit) and the unit was edited only once ("revisions = 1").

The distribution of the tool includes scripts to create input files and parse output files.

3. Examples of use

3.1. Post-editing versus translation

Sousa et al. (2011) reports an objective way of measuring translation quality in terms of post-editing time using PET. The goals of the experiments were (i) to check whether post-editing sentences was quicker than translating them from scratch, and (ii) as a by product, since multiple MT systems were used, to compare these systems by ranking them according to the amount of time that was required for humans to post-edit their output.

Eleven human translators were asked to post-edit English-Portuguese sentences from TV series subtitles translated using four systems (Google Translate, Systran, SDL Trados and Moses) and also to translate such sentences from scratch. Translators received a random selection of sentences and translations produced by different MT systems to translate or post-edit using PET.

For the translation task, annotators were also asked to assess the sentence according to a scale of difficulty (1 = difficult; 2 = moderate; 3 = easy). For the post-editing task, annotators were asked to indicate the post-editing effort for each unit (1 = requires complete retranslation; 2 requires some retranslation, but post editing still quicker than retranslation; 3 = very little post editing needed; 4 = fit for purpose).

PET was used to collect the assessment scores given by the translators to every unit and also the time spent performing the translation/post-editing of that unit.

Because time was an important effort indicator in this experiment, PET's feature that hides a unit before its editing starts was used. This prevented annotators from reading the source and translation texts (and possibly thinking about a correct translation) before they started the post-editing/translation.

Using the information gathered by the tool, it was possible to rank the translation systems according to i) the scores assigned by the annotators and ii) the average time the annotators spent post-editing the output of each system. These annotations also allowed contrasting post-editing and translation tasks in terms of time. Table 1 summarizes the results for these two aspects using time as measure. It shows that post-editing the output of any system is faster than translating subtitles from scratch. Post-editing was faster than translation in 72%-94% of the cases, depending on the quality of the translation system. On average for the four MT systems, it was found that post-editing sentences is 40% faster than translating them from scratch.

System	Faster than HT
Google	94%
Moses	86.8%
Systran	81.2%
Trados	72.4%

Table 1: How often post-editing a translation system output was faster than translating the text from scratch.

The output of each system was also assessed in terms of TER/HTER using different types of reference translations: (i) the original single reference subtitle in Portuguese (R_0); (ii) the targeted reference (P_i), that is, a single post-edited version of the machine translation; and (iii) all the translations and post-edited versions of the machine translation collected as part of the task: R_0, as above, plus R_{1-5} = five reference translations collected from the translation job, and R_{6-17} = twelve reference translations obtained via the post-editing of draft translations. The aim was to measure how close to any manually obtained translation the MT and TM outputs were and what percentage of the draft translations was reused in the PE task. Table 2 shows the performance of the four systems in terms of TER/HTER. Note that apart from the first row (R_0), the reference sets contain targeted references.

The quality of the post-edited translations was also a concern in this experiment. Although the translators were asked to perform the minimum necessary operations while post-editing, they were instructed to produce translations that were "ready for publishing". An automatic evaluation was conducted in order to compare each of the 12 sets of post-edited translations to the 5 sets of translations

References	Google	Moses	Systran	Trados
R_0	0.79	0.75	0.88	1.01
P_i	0.06	0.21	0.22	0.66
R_{0-17}	0.06	0.19	0.21	0.62

Table 2: TER/HTER scores with different types of references (P_i and R_{0-17}).

Post-editing time vs	HTER	Assessments
Spearman's ρ	0.72 ± 0.1	-0.76 ± 0.1
Pearson's	0.46 ± 0.1	-0.53 ± 0.1

Table 3: Correlation between the post-editing time and HTER or human assessment. The values represent the average of all participants in the task. All individual scores were significant with p-value < 0.01.

produced from scratch and the reference translations created independently from the machine translations (R_{0-5}). The comparison resulted in an average BLEU score (Papineni et al., 2002) of 69.92 ± 4.86 and average TER scores of 0.26 ± 0.04, suggesting that post-editing does not imply loss of translation quality compared to translation from scratch. We note that in this comparison post-edited translations were evaluated against translations produced from scratch, as opposed to our main setup in which machine translations are compared to their post-edited versions (targeted evaluation).

Additionally, for each system (i), its post-edited machine translations (P_i) were compared to all the other references, that is, the post-edited machine translations of all the other systems ($\forall_{j \neq i} P_j$) and the translations produced from scratch (R_{0-5}). The resulting TER score of 0.18 ± 0.042 confirms our assumption that post-editing does not harm translation quality, since this level of difference is expected to result from the use of equally valid paraphrases in the translations.

To validate the use of post-editing time as a valuable effort indicator, segment-level correlation coefficients between post-editing time and HTER, and between post-editing time and the explicit assessment scores given by the annotators were computed. Table 3 shows a strong Spearman's rank correlation ρ between the two pairs of variables. This shows that more edits require more post-editing time and that less post-editing time indicates higher assessment scores. Pearson's coefficient shows that this correlation is not always linear.

3.2. Post-editing for quality estimation

It is generally agreed that the post-editing of MT can be more productive than translation from scratch provided that the automatic translations have a satisfactory level of quality. However, it is very common for an MT system to produce a mixture of good and bad quality translations. Therefore, the post-editing of certain segments will require much more effort than that of other segments, sometimes even more than translating those segments from scratch. Identifying such segments and filtering them out from the post-editing task is a problem addressed in the field of Quality Estimation (QE) for MT.

QE metrics are usually prediction models induced from data using standard machine learning algorithms fed with examples of source and translation features, as well as some form of annotation on the quality of the translations. Recent work on the topic focuses on having humans implicitly or explicitly assigning absolute quality scores to translations, which has shown more promising results. In particular, (Specia, 2011) describes experiments comparing the prediction of three types of quality scores: absolute scores reflecting post-editing effort, post-editing time (seconds per word) and edit distance from a good translation.

In that work, PET was used to facilitate the process of obtaining training data with explicit and implicit human annotations for translation quality. Two datasets were collected using *news* source sentences from development and test sets provided by WMT[8] (Callison-Burch et al., 2010), translated using a phrase-based SMT system built using Moses[9] (Koehn et al., 2007):

- news-test2009: $2,525$ French-English sentences.
- news-test2010: first $1,000$ English-Spanish sentences.

Translators received initial training on the tool and task and were instructed to perform the minimum number of editions necessary to make the translation ready for publishing. They were aware of the time measurement and its general purpose. Using PET, translators were asked to post-edit each sentence and to score the original translation according to its post-editing effort using the 1-4 scale described in Section 3.1.

The HTER edit distance between the original automatic translation and its post-edited version was then computed. HTER computes the proportion of words or sequences of words that needed to be edited in order to change the MT output into a good translation. We set HTER options to tokenize the text, ignore case and use equal cost for all edits.

The annotation process resulted in three types of sentence-level annotation for each dataset: HTER, $[1-4]$ scores and post-editing time (average number of seconds to post-edit each word in the sentence).

Using a standard framework, three QE models were built for each language pair. The goal was then to assess these three models, which had been built using different annotation types, in a *task-based* evaluation. Unseen translations from other WMT datasets with the same genre and domain were selected:

- news-test2010: $2,489$ French-English translations.
- news-test2009: $2,525$ English-Spanish translations.

For each language-pair, four non-overlapping subsets of 600 translations were randomly selected from these WMT datasets. Quality predictions were generated for three of these subsets using each of the three variations of the QE models. The 600 translations in these three subsets were then ranked using the predicted score so that the (supposedly) best translations came first. Translations in the forth subset were not ranked.

Taking the four resulting datasets (three with translations sorted according to their estimated quality, and one with translations unsorted), the same two translators who had performed the annotation in the training sets (above) were

[8] www.statmt.org/wmt11
[9] www.statmt.org

then asked to use PET to post-edit as many sentences as possible following their order in four "tasks". Each task was performed on a different day, and translators were given **one-hour per task**. The order of the tasks was randomly defined, but each translator post-edited all four subsets:

- T1: 600 translations sorted - *HTER* model.
- T2: 600 translations sorted - *effort* model.
- T3: 600 translations sorted - *time* model.
- T4: 600 translations without any sorting.

The number of sentences post-edited in each subset varied from 33 to 97, but we note that sentences have different lengths, and thus looking at the counts of word in those sentences is more informative. The final ranking of the translation subsets was computed by counting the number of words that were post-edited in each test set. Based on these counts, Table 4 shows the average number of words post-edited per second (within the first hour of post-editing). These figures refer to the total number of words in the final post-edited sentences, including words which were kept as in the original MT.

Dataset		Words/second
fr-en	T1: *HTER*	0.96
	T2: *effort*	0.91
	T3: *time*	1.09
	T4: unsorted	0.75
en-es	T1: *HTER*	0.41
	T2: *effort*	0.43
	T3: *time*	0.57
	T4: unsorted	0.32

Table 4: Number of words that could be post-edited per second in sentences ranked according to different QE models in one hour.

For both language pairs, post-editing only the best machine translations according to any QE model allows more words to be post-edited in a fixed amount of time than post-editing randomly selected machine translations ("unsorted"). The best rate is obtained with *time* as response variable in both fr-en and en-es datasets. This shows that the implicit annotation of time using PET is a promising way of collecting training data for quality estimation. In this case, PET was used not only for data collection purposes, but also as a means to evaluate the usefulness of the QE models and compare different variations of such models.

4. Conclusions

We have presented a simple tool for post-editing and assessing automatic translations that is MT system independent and allows customization at various levels, including the types of assessments that can be collected and restrictions on the post-editing process (such as the length of post-edited units). We have also given a few examples of uses of such a tool: collecting information to train quality estimation models, comparing different translation tools and different quality estimation models, and contrasting post-editing and translation from scratch. The tool facilitates all these tasks, besides allowing for more controlled experiments, particularly with respect to time measurements.

The tool is available for download at: `http://pers-www.wlv.ac.uk/~in1676/pet/` and `http://www.dcs.shef.ac.uk/~lucia/resources/`.

5. References

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden.

Philipp Koehn, 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. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics: Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Philipp Koehn. 2009. A Process Study of Computed Aided Translation. *Machine Translation Journal*, 23(4):241–263.

Joseph Olive, Caitlin Christianson, and John McCary. 2011. *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Morristown.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231.

Sheila C. M. Sousa, Wilker Aziz, and Lucia Specia. 2011. Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles. In *Proceedings of the Recent Advances in Natural Language Processing Conference*, Hissar, Bulgaria.

Lucia Specia. 2011. Exploiting objective annotations for measuring translation post-editing effort. In *Proceedings of the 15th Conference of the European Association for Machine Translation*, pages 73–80, Leuven.