Error Detection for Post-editing Rule-based Machine Translation

Justina Valotkaite
Research Group in Computational Linguistics
University of Wolverhampton
justina.valotkaite@gmail.com

Munshi Asadullah
Research Group in Computational Linguistics
University of Wolverhampton
asad.anto@gmail.com

Abstract

The increasing role of post-editing as a way of improving machine translation output and a faster alternative to translating from scratch has lately attracted researchers' attention and various attempts have been proposed to facilitate the task. We experiment with a method to provide support for the post-editing task through error detection. A deep linguistic error analysis was done of a sample of English sentences translated from Portuguese by two Rule-based Machine Translation systems. We designed a set of rules to deal with various systematic translation errors and implemented a subset of these rules covering the errors of tense and number. The evaluation of these rules showed a satisfactory performance. In addition, we performed an experiment with human translators which confirmed that highlighting translation errors during the post-editing can help the translators perform the post-editing task up to 12 seconds per error faster and improve their efficiency by minimizing the number of missed errors.

1. Introduction

Since its introduction Machine Translation (MT) has improved considerably and recently it has started gaining recognition in the translation industry. However, translations of MT systems have not yet reached the level of human quality. One of the ways of improving MT outputs is by performing the task of post-editing (PE), which nowadays, is becoming a common practice. According to Suzuki (2011), "to make the best of machine translation humans are urged to perform post-editing efficiently and effectively". As a starting point, in this study, we focus on Rule-based Machine Translation (RBMT), since we believe these systems produce errors in a more systematic manner, which makes capturing these errors more feasible.

In their outputs RBMT systems tend to repeat the same mistakes. Therefore, while post-editing, humans are forced to correct the same mistakes repeatedly and this makes the post-editing task draining and monotonous. In this study we aim at investigating a way of providing support for the post-editors by designing linguistically motivated rules for error detection that could be integrated into a post-editing tool. Our hypothesis is that these rules could help post-editors by indicating problems in the output which need to be fixed, and as a consequence help minimise post-editing time.

Recent work has addressed error detection and its visualization following shallow, statistic approaches for the error detection and focusing mostly on SMT. Koehn and Haddow (2009) introduced a tool for the assistance of human translators with functionalities such as prediction of sentence completion, options from the translation table and post-editing. Experiments with a Statistical Machine Translation (SMT) system and ten translators revealed that the translators were fastest when post-editing with the tool.

Xiong et al. (2010) proposed the integration of two groups of linguistic features, i.e. lexical and syntactic features, into error detection for Chinese–English SMT. These features were put together with word posterior probability features using a classifier to predict if a word was correct or incorrect. Various experiments were carried out and the results revealed that the integration of linguistic features was very useful for the process of error detection because the linguistic features outperformed word posterior probability in terms of confidence estimation in error detection.

Bach et al. (2011) proposed a framework to predict SMT errors at word and sentence levels for Arabic–English translation. They used a large dataset with words and phrases that had been previously post-edited as features for training the error detection model. As part of their experiments they also introduced a visualization prototype for errors in order to improve the productivity of post-editors by helping them quickly identify sentences that might have been translated incorrectly and need correction. Their method was based on confidence scores, i.e. predictions at phrase and word level for good, bad and medium quality translations. The results showed that the MT error prediction accuracy has increased from 69.1 to 72.2 in F-score.

While our goal is very similar to that of these papers, we address the error detection of the MT output from a more linguistically motivated perspective. We derive linguistic rules from an error analysis of Portuguese–English sentences from two text domains and two variants of Portuguese (European and Brazilian) translated by two RBMT systems – Systran[1] and PROMT.[2] We consider detection rules to be a practical and potentially helpful solution for RBMT systems, since these are known for making repetitive mistakes. In other words, if a system cannot cope with some specific language phenomenon, for example, recognizing a certain word or a construction in a source language (SL), it is likely to keep making the same mistake whenever that phenomenon is encountered. In addition to understanding whether we can successfully detect the errors with these rules and whether highlighting them can help human translators, we are interested in assessing how general the rules (and the errors made by the systems) are across MT systems of the same type (rule-based) and across significantly different text domains.

In the remainder of this paper, we first describe our linguistic analysis (Section 2), to then describe the implementation of the rules (Section 3**)** and present a post-editing experiment with human translators (Section 4).

2. Linguistic Analysis

For the linguistic analysis we randomly selected 300 Portuguese sentences from two corpora: 150 sentences from Europarl (Koehn 2005), which is a collection of parliamentary speeches representing European Portuguese, and 150 from Fapesp (Aziz and Specia 2011), which is a collection of scientific news for Brazilian Portuguese. The minimum length of Europarl sentences was 3 words, maximum – 115, average – 27; whereas the minimum length of Fapesp sentences was 3 words, maximum – 88, and average 31. We then translated these sentences into English using Systran and PROMT, totalling 8,455 words.

Un-translated words	Inserted article*
Inflectional error*	Incorrect preposition
Incorrect voice *	Inserted preposition*
Mistranslated pronoun*	Inserted pronoun*
Missing pronoun*	Incorrect adjective translation*
Incorrect subject-verb order*	Incorrect order of nouns and their adjectival modifiers*
Missing article*	Incorrect date translation format/ numbering system*
Incorrect other word order*	Incorrect article / an article replaced by another POS*
Incorrect lexical choices	Incorrect translation of Portuguese reflexive verbs
Repeated words	Incorrect translation of Portuguese weekdays
Added words	Translated Portuguese surnames*
Missing words	Translated Portuguese abbreviation*
Main message is different	Missing subjects/ predicates
Capitalization problems*	POS error: a verb instead of an adjective etc (the same root)*
Missing if-clause*	Missing preposition*

Table 1. Error classification for the English-Portuguese language pair. Categories marked with * were later modelled by rules (Section 3)

The linguistic analysis was carried out as follows. First, we manually analyzed each sentence, identified various translation errors and

[1] http://www.systran.co.uk/
[2] http://www.promt.com/

assigned them to different error categories. We identified errors by correcting the sentences until they were of acceptable quality but at the same time trying to keep them as close as possible to their machine translated versions. In cases when errors co-occurred, i.e. two categories could have been applied for the same issue, these were counted twice. The error classification introduced in this paper was inspired by the classification schemes introduced by Flanagan (1994), Farrús (2010), and Specia et al. (2011).

Table 1 presents the error classification. The current analysis showed that the Portuguese–English translation outputs contained the most frequent and typical language-independent MT errors, such as "Incorrect lexical choice", "Inflectional errors", "Untranslated words". On the other hand, we also identified some language-specific errors, typical to the Portuguese-English language pair. Table 2 illustrates a subset of these: the most frequent error categories identified during the linguistic analysis.

Error category	Percentage of total errors			
	Systran		PROMT	
	Europarl	Fapesp	Europarl	Fapesp
Incorrect lexical choices	31.29	31.73	34.41	34.56
Inflectional error	9.84	5.61	7.76	5.87
Mistranslated pronoun	9.52	6.41	9.22	5.87
Untranslated words	4.19	4.33	2.20	6.21
Incorrect other word order	8.39	5.93	4.83	3.02

Table 2. The most frequent error categories in the corpora and systems analysed

An interesting example of language-specific category is the "Incorrect translation of Portuguese weekdays". In Portuguese, names of weekdays except *sábado* (*Saturday*) and *domingo* (*Sunday*) are compounds made of two individual words: a numeral and a noun. For instance, *segunda-feira* (*Monday*), *quinta-feira* (*Thursday*), etc. However, both systems failed to produce correct translations for these compounds. Instead, Systran produced a literal translation for both individual words (*quinta-feira* was translated as *fifth-fair*); whereas, PROMT translated them as equivalent weekdays in English but also added a verb phrase which was not present in the source text and did not make sense in the given context (*quinta-feira* was translated as *Thursday-sells at a fair*).

Based on the frequency of the errors in each category and on an analysis on the feasibility of creating rules for them, we selected 20 categories (marked * in Table 1) for which we designed the rules that later could be implemented and used to support the post-editing task.

We created the rules by analysing the errors of the target language (TL) sentences and comparing these to their corresponding SL sentences. In total, we produced a set of 40 contrastive rules which covered various problematic linguistic issues. For instance, if the systems made mistakes while dealing with present, past and future verb tenses and chose incorrect tenses for the TL translations, rules were designed for these specific issues. It is important to emphasize that although the two variants of Portuguese are considerably different, we focused on creating the rules that apply for both.

Rules are of the if-then type: *if in Portuguese* <...>, *then in English* <...>. For example, "If the verb X is in the past simple tense in Portuguese, then in English the translation of the same verb X must be in the past simple tense".

We then selected a subset of these rules which could be implemented for a pilot study. We took the following rules dealing with tense and number errors for the inflectional category due to the availability of the necessary pre-processing tools for the two languages and because they represented one of the most frequent error types in the output sentences:

- If the noun X is in singular/plural in Portuguese, the translation of the noun X should also be in singular/plural in English.
- If the verb X is in the $1^{st}/2^{nd}/3^{rd}$ person singular/plural in Portuguese, the translation of the verb X should also be in the $1^{st}/2^{nd}/3^{rd}$ person singular/plural in English;

- If the verb X is in the infinitive/ present simple tense/past simple tense/future simple tense in Portuguese, the translation of X in English should also be in the infinitive/ present simple tense/past simple tense/future simple tense;
- If the Portuguese construction contains "ir (*to go*) + infinitive", then the English translation should be the future simple tense "(subjective pronoun) + will + simple verb/(subjective pronoun) / to be + going to + infinitive" (e.g. vou falar = I will speak; vamos verificar = we are going to check);
- If the Portuguese verb construction contains "não + V_{prs} + 3rd person sg", the English equivalent should be the construction "(subjective pronoun) + auxiliary verb + 3rd p. sg. + not + infinitive" (e.g. não fala = (subj. pronoun) does not speak);
- If a Portuguese verb phrase is of progressive aspect, i.e. "estar + a + infinitive", it should be in the present continuous tense in English "subjective pronoun + the form of to be + V_{ing}" (e.g. está a falar = (s)he is speaking).
- If the Portuguese verb construction is the following "Vprs +3rd p. sg. + a + infinitive, in English it should be the English construction V_{prs} +3rd p. sg. + infinitive" (e.g. ajuda a conter = helps to contain).

3. Evaluation of the Categories and Rules

In order to analyse how systematic the selected categories are and check the coverage of the rules created and the ones selected for the implementation, we performed two small scale experiments on two new datasets. For the first experiment we randomly selected 100 additional sentences from the original corpora: 50 from Europarl and 50 from Fapesp, and translated them using Systran and PROMT. The output of both datasets resulted in 5,676 words. The minimum length of Europarl sentences was 3 words, maximum – 68, average – 26; whereas the minimum length of Fapesp sentences was 3 words, maximum – 62, average – 30.

During the analysis, we fixed the translations to identify translation errors as before and assigned them to our error categories. The results revealed that out of our 30 categories, 26 were present in the new dataset, despite its smaller size. Only four categories - "Missing if-clause", "Incorrect adjective translation", "Missing subjects/predicates" and "Incorrect translation of Portuguese weekdays" - were not found in the new dataset. It is also important to emphasize that no new categories were identified in this dataset. From these results it can be concluded that the error classification for the Portuguese–English language pair in Table 1 is representative of these two text domains and RBMT systems.

Furthermore, we checked the frequency of all errors and in particular those of the inflectional category. The translations in the new dataset contained 393 errors, out of which 54 were attributed to the inflectional error category. To verify the coverage of the rules and in particular of those dealing with inflectional errors, we computed the percentage of errors in the new dataset that could be dealt with by the rules derived for the original dataset. The coverage was computed by dividing the number of errors for which there were no rules created by the total number of errors. The results showed that the coverage of the whole set of rules was 98.21%, while the coverage of the rules of the inflectional category was 92.59%.

For the second evaluation experiment we randomly selected 100 sentences from two new corpora: 50 sentences from CETEMPublico[3] which covers news in European Portuguese, and 50 sentences from CETENFolha[4] which covers news in Brazilian Portuguese. The minimum length of CETEMPublico sentences was 6 words, maximum – 58, average – 27; whereas the minimum length of CETENFolha sentences was 11 words, maximum – 51, average – 24.

We translated the sentences using both RBMT systems, resulting in 5,039 words. Once again, we checked the coverage of the categories and rules. The results revealed that all 30 categories introduced in the error classification were present in this dataset and no new categories were identified. In total, the output sentences contained 513 errors, with 39 of them of the inflectional

[3] http://www.linguateca.pt/cetempublico/informacoes.html
[4] http://www.linguateca.pt/cetenfolha/

category. The coverage of the whole set of rules was 93.67%, while the coverage of the inflectional rules selected for implementation was 87.18%.

From these results we can conclude that it is possible to systematically categorise errors in RBMT systems and that linguistic rules with sufficient coverage can be created in new datasets for such categories.

4. Implementation and evaluation of rules

A few pre-processing tasks were performed in order to obtain certain linguistic information necessary for the implementation of the rules, such as a part-of-speech, lemma, morphological information (number and gender). First we performed word alignment between the Portuguese-English sentence pairs by using the aligner GIZA++ (Och and Ney, 2003). GIZA++ aligns sentences at the token level producing in this case four pairs of the alignment datasets. As some incorrect alignments were found in the sentences, we manually corrected them in order to obtain a "clean" dataset, that is, a dataset that allows evaluating the rules themselves, isolating any effect of low quality word alignments. Each aligned sentence pair was checked and the necessary corrections were performed, i.e. some incorrect aligned links were deleted, while other necessary links were inserted.

The sentences were also parsed in order to obtain their morphological information. For this purpose we used the parsers Palavras [5] for Portuguese and ENGCG [6] for English, both available online. The Palavras parser was reported to have 99.2% correct morphological tagging (Bick, 2000) and ENGCC was reported 99.8% recall in morphological tagging (Voutilainen and Heikkilä 1994). Any other parser producing morphological information could in principle be used. For this pilot study, seven rules dealing with errors of tense and number agreement were implemented using Python.

[5] http://beta.visl.sdu.dk/constraint_grammar.html
[6] http://www2.lingsoft.fi/cgi-bin/engcg

We then analysed the performance of the rules manually, i.e. each output sentence was checked individually to find out how many translation errors the system identified correctly, how many were missed and if it identified any false positives.

	Europarl-Systran	Europarl-PROMT	Fapesp-Systran	Fapesp-PROMT
Recall	0.46	0.70	0.66	0.63
Precision	0.62	0.58	0.42	0.37
F-Measure	0.53	0.63	0.51	0.47

Table 3. The evaluation of the system

Precision, Recall and F-measure were calculated. The results are shown in Table 3.

The system found a high number of false positives in the output sentences. The main reasons for this is parser errors and inconsistencies and the fact that often more than one rule can be applied and no rule precedence scheme was defined at first. An example of a case where multiple rules could be applied is the following Portuguese construction "*the present tense form of ir* (*to go*) + *infinitive*". It expresses a future action and usually is translated into English using the future simple tense i.e. *vou / vais / vai / vamos / vão + falar = I / you /(s)he / we / they will speak*. However, when the system encountered this construction, it identified the correct English translation as incorrect. For example:

(1-PT): *Vou verificar* se nada de isso foi efectivamente feito.
(1-EN): *I will check if nothing of that was effectively an act.*

This happened because the first rule processed by the system (by default) stated that "*if the verb X is in the present simple tense, the English translation of the same verb X should be in the present simple too*". Therefore, when the system encountered *vou verificar*, it did not recognize the pattern of *vou + verificar* as a future tense construction but rather only the verb in the present simple tense *vou* and in the English side it expected to find *I go + check*. However, further in the list there was a rule explaining this specific pattern and indicating the correct translation. Defining rule precedence schemes is not a trivial problem. While this was possible for our small set of rules, this issue will require further investigation as this set grows to incorporate other linguistic phenomena. .

The largest number of false positives occurred due to the flaws of the parsers. Example (2) illustrates a false positive case due to errors of the parser. The system flags the English verb *comments* (3rd person singular) as an error although it was correctly translated. This happened because the parser identified *comments* as a plural noun, and not as a verb. Therefore, when the system found *comenta (comments)* as a verb in present tense (3rd person singular), it expected to find a verb in the English side.

(3-PT): *"Queremos aumentar o intercâmbio com instituições internacionais que são referência em a pesquisa em música e ciência", comenta Ferraz.*

(3-EN): *"We want to increase the exchange with international institutions that are a reference in the inquiry in music and science", comments Ferraz.*

Some examples when the rules correctly detect translation errors include the incorrect tense and number translation (4) and the incorrect translation (5), i.e. a noun instead of a verb:

(4-PT): *Todos os restantes discordavam.*

(4-EN): *All the remainder was disagreeing* (*disagreed*).

But in that moment we were covering (we covered) almost only projects with support of the Fapesp, which was not the case.

(5-PT): *Penso que porque, mesmo mantendo as posições marxistas dialéticas, o ensaio era uma desmontagem de o marxismo fechado.*

(5-EN): *Bandage (I think) that because, even maintaining the dialectic Marxist positions, the test was a desmontagem of the shut Marxism.*

The implementation and the evaluation of the system showed that it is possible to have a working rule-based system which can detect certain translation errors using linguistic rules. Although in the current version some errors still remain to be captured due to their complexity and the limitations of the approach (the rules cover a small range of translation problems and only a sentence boundaries), we believe that error detection based on linguistic information is a promising direction to improve MT quality. While having a large number of false positives can still be an issue, this is less problematic than missing true errors. Although further experiments are necessary in that direction, our preliminary analysis in Section 6 indicates that translators can miss certain errors if these are not highlighted.

5. Experiments with Human Translators

A post-editing experiment was carried out with human translators in order to determine the usefulness of having errors highlighted in the RBMT output. Here we aim to investigate whether it is possible to help the human translators perform the task of post-editing faster and more efficiently when the MT errors are detected and highlighted.

To proceed with the experiment, we used the post-editing system PET (Aziz et al., 2012).[7] The tool gathers various useful effort indicators while post-editing is performed. We measured the *time* translators spent post-editing sentences. The tool also renders HTML, so highlighting errors was trivial.

For the test set in this experiment we randomly selected 60 sentence pairs from both Europarl and Fapesp. These sentences were then manually annotated, i.e. translation errors in the English as well as their corresponding source segments in the Portuguese sentences were marked. We resorted to manual highlighting rather using the errors detected by our system due to its limited coverage (only certain inflectional errors) and its relatively low performance.

After the manual error annotation, the sentence pairs were given to six human translators. We divided the test set into two parts, i.e. 30 sentence pairs with no errors highlighted and 30 sentence pairs with errors highlighted. All translators post-edited sentences with and without highlights. As the sentences were randomly selected, they contained different numbers of mistakes. Therefore, we analysed time on a per error (and not per sentence) basis. The errors were highlighted using different colours, each colour representing an individual error type from the set of 20 categories.

We produced guidelines in order to help human translators perform their task by explaining in detail how to use the post-editing tool and how they were expected to perform the task. All

[7] http://pers-www.wlv.ac.uk/~in1676/pet/

participants were native speakers of Portuguese. They were fluent in English and had some experience with translation tasks. European Portuguese translators were given European Portuguese sentences, whereas Brazilian translators post-edited Brazilian Portuguese sentences. We asked translators to post-edit machine translated sentences by making as few changes as possible and in such a way that the sentences would be grammatically correct and understandable. For sentences with errors highlighted, translators were also asked to evaluate the usefulness of having errors highlighted by choosing one of three possible options:

- Very useful.
- Some of them were useful.
- Not useful at all.

Translators were given one week to perform the task. Results were computed for three main aspects: the number of correctly identified and missed errors, the time taken to post-edit sentences in both datasets and the translators' evaluation of the usefulness of the highlights.

We manually analysed each translator's work by comparing their post-edited sentences with the previously annotated sentences. The results revealed that in the test set of the non-highlighted (**NH**) sentences the range of correctly identified errors by translators varies from 90% to 95.56%. The results for the sentences with the highlights (**WH**) showed a noticeable improvement, i.e. from 95.24% to 100%. It can thus be concluded that the performance of the translators improved when post-editing sentences with errors highlighted as the number of errors missed was significantly reduced. The reasons for translators missing errors in the non-highlighted sentences could be various, including the fact that perhaps the translators got used to having errors highlighted and the fact that some errors were not very significant for adequacy purposes, for example an incorrect extra article. However, the experiment showed that highlighting errors can be very helpful in attracting the attention of translators.

In order to find out if there was any significant difference between performing the two tasks in terms of time, we counted each translator's average time per error for both datasets. To get these estimates we divided the total time spent for post-editing each dataset (WH and NH) for each translator by the total number of errors in that dataset. The results are shown in Figure 1.

As it can be seen from Figure 1, translators' time per error NH and WH varies from 33 to 23 seconds (Translator 2), from 43 to 31 seconds (Translator 5) and from 27 to 17 seconds (Translator 6). These translators post-edited the WH sentences 10-12 seconds faster than the NH sentences.

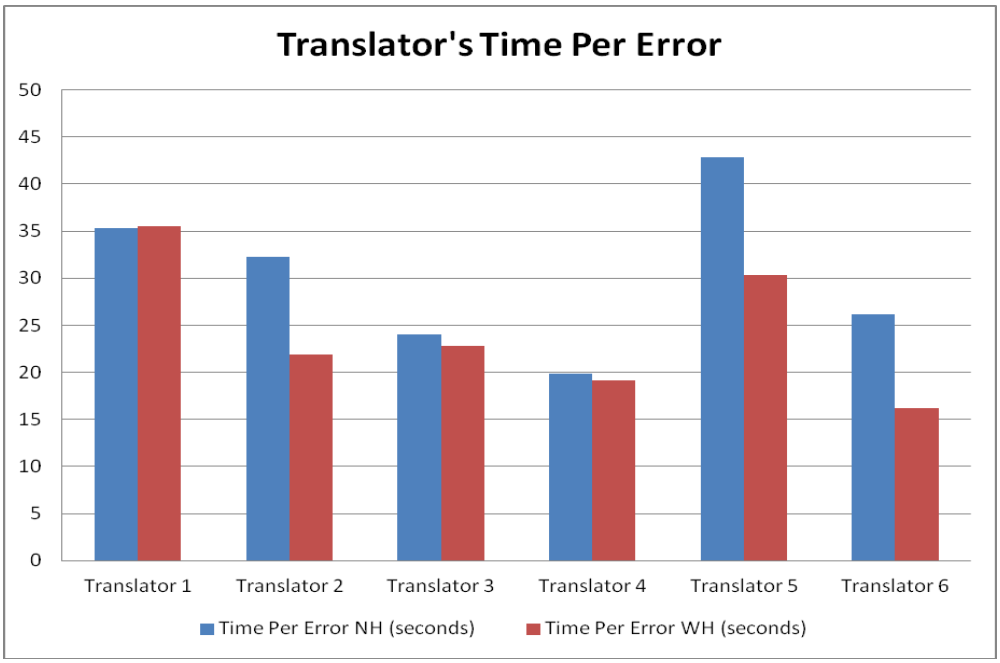

Figure 1. Time per error taken to post-edit the sentences

For the rest of the translators, the results were less significant. The time of Translators 3 and 4 for the WH sentences was slightly better than for the NH sentences, varying from 22 to 24 seconds, and from 17 to 19 seconds respectively. On the other

hand, Translator's 1 time was 36 seconds for the NH sentences and 35 seconds for the WH sentences. This could be explained by the fact that these translators missed a considerable number of errors, thus it is not surprising that there was no improvement in their results in terms of time. Although these improvements seem to be modest, we believe that when one extrapolates them to thousands of sentences with potentially dozens of errors, having errors highlighted can make a considerable difference in productivity.

The final factor which we analysed in this experiment was the opinion of the translators about the usefulness of the highlighted errors. As mentioned before, after post-editing each sentence with highlights, translators were asked how useful the highlights were. The results show that in 68% of the cases the highlights were found to be very useful, in 27% of the cases - some of them were found to be useful, and in only 5% they were found not to be useful at all.

The results of the experiments with human translators showed that the having errors highlighted can help human translators perform the task of post-editing faster and more efficiently. The highlights were also positively evaluated by translators.

6. Conclusions and Future Work

We have shown that one can systematically categorize translation errors which RBMT systems make and create linguistic rules for the error categories identified. We also showed that the rules apply across MT systems and text domains, and that one can implement a system detecting certain translation errors on the basis of those rules. Having a linguistically motivated approach for the error detection has also been shown to be helpful for the post-editing task. The results of a post-editing experiment with human translators revealed that the highlighted errors in the RBMT output helped to perform the PE task faster up to 10-12 seconds per error and improve translators' efficiency in identifying errors by reducing the number of errors missed. In addition, the highlighted errors were positively evaluated by the translators. Thus it can be concluded that the approach for post-editing based on the error analysis and the automatic error detection is promising and should be elaborated further.

The major challenge for future work is to scale up the approach. In order to implement the remaining rules, more levels of linguistic pre-processing will be necessary, such as named entity recognition. More robust ways of dealing with flaws in linguistic processors (such as the current parser issues) are also necessary.

Acknowledgment

This project was supported by the European Commission, Education & Training, Erasmus Mundus: EMMC 2008-0083, Erasmus Mundus Masters in NLP & HLT program.

References

Aziz, W. and Specia, L. (2011) *Fully automatic compilation of Portuguese-English and Portuguese-Spanish parallel corpora*. In: 8th Brazilian Symposium in Information and Human Language Technology (STIL-2011), Cuiaba, Brazil

Aziz, W., Sousa, S. And Specia, L (2012*) PET: a tool for post-editing and assessing machine translation.* In: The Eighth International Conference on Language Resources and Evaluation, LREC '12, Instambul, Turkey

Bach, N., Huang, F., and Al-Onaizan, Y. (2011). *Goodness: A method for measuring machine translation confidence*. In: 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, 211–219.

Bick, E. (2000). *The parsing system "Palavras" - automatic grammatical analysis of Portuguese in a constraint grammar framework*. In: Proceedings of the 2nd International Conference on Language Resources and Evaluation, TELRI, Athens

Farrús, M. Costa-Jussà, M., Mariño, J., and Fonollosa, J. (2010) *Linguistic-based evaluation criteria to identify statistical machine translation errors*. In: Proceedings of European Association for Machine Translation (EAMT), Saint Raphael, France, 52–57.

Flanagan, M. (1994) *Error classification for MT evaluation*. In: Technology Partnerships for Crossing the Language Barrier. Proceedings of the First Conference of the Association for Machine Translation in the Americas. Columbia, Maryland, US, 65-72.

Koehn, P. (2005) *Europarl: a parallel corpus for statistical machine translation*. In: Proceedings of the 10th Machine Translation Summit, AAMT, Phuket, Thailand

Koehn, P. and Haddow, B. (2009) *Interactive Assistance to Human Translators using Statistical*

Machine Translation Methods. In: MT Summit XII, 73-80

Och, F. and Ney, H. (2003) *A Systematic Comparison of Various Statistical Alignment Models*. Computational Linguistics, vol. 29, no 1, 19-51

Suzuki, H. (2011) *Automatic Post-Editing based on SMT and its selective application by Sentence-Level Automatic Quality Evaluation*. In: Thirteenth Machine Translation Summit (AAMT), 2011, Xiamen, China, 156-163.

Voutilainen, A. and Heikkilä, J. (1994) An English constraint grammar (EngCG): a surface-syntactic parser of English. In Fries, Tottie and Schneider (eds.), Creating and using English language corpora, Rodopi, 189-199

Xiong, D., Zhang, M., Li, H. (2010) *Error detection for statistical machine translation using linguistic features.* In: ACL: the 48th Annual Meeting of the Association for Computational Linguistics, *Uppsala, Sweden*, 604-611.