IDIAP RESEARCH REPORT

TRANSLATION ERROR SPOTTING FROM A USER'S POINT OF VIEW

Thomas Meyer[a]

Idiap-RR-31-2012

NOVEMBER 2012

[a]Idiap Research Institute, P.O. Box 592, CH-1920 Martigny

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Translation Error Spotting from a User's Point of View

Thomas Meyer
Idiap Research Institute / Martigny, Switzerland
EPFL-EDEE doctoral school / Lausanne, Switzerland
Thomas.Meyer@idiap.ch

Abstract

The evaluation of errors made by Machine Translation (MT) systems still needs human effort despite the fact that there are automated MT evaluation tools, such as the BLEU metric. Moreover, assuming that there would be tools that support humans in this translation quality checking task, for example by automatically marking some errors found in the MT system output, there is no guarantee that this actually helps to achieve a more correct or faster human evaluation. The paper presents a user study which found statistically significant interaction effects for the task of finding MT errors under the conditions of non-annotated and automatically pre-annotated errors, in terms of the time needed to complete the task and the number of correctly found errors.

1 Introduction

In current research in Machine Translation (MT) and especially Statistical Machine Translation (SMT), apart from the development of actual systems and algorithms, there has also been a considerable effort on how to evaluate the translation quality of the output by such systems. This lead for instance to automated evaluation metrics which are able to measure the quality of MT output on a large scale by comparing it, using various matching techniques, against one or more human reference translations. These metrics however, may not be accurate enough to account for all MT errors and even in the most recent MT system competition the participants themselves are still asked to manually evaluate the output of the other participants' systems[1].

Moreover, if one not only considers the assessment of the overall MT output quality but also post-editing, i.e. the conversion of an automatically generated translation into an acceptable and readable text, the human effort becomes even more important and detailed error analysis is necessary. If one hypothesizes that there are tools to support the evaluators in the task of translation error analysis and/or post-editing, it would be worth knowing to what extent such a tool would lessen the human effort to spot MT errors. The task of human evaluators to check for translation quality and to identify errors made by MT systems is referred to in the following as 'translation error spotting'. In this article, we study the use and helpfulness of a translation error spotting tool for human evaluators.

The paper starts with a short description of the state-of-the-art and the current problems in evaluating MT system output (Section 2). We then present a user study performed by different groups of users under different conditions which are described in Section 3. The methods used to analyze the performance of the subjects on the translation error spotting task and the results of the data analyses are shown in Section 4. The paper is concluded in Section 5, outlining possible future work.

2 Evaluating Machine Translation Output

Translation quality is hard to evaluate because translators almost never completely agree on a unique translation, even for short sentences. Several translations might be perfectly correct even though they use different expressions and gram-

[1] statmt.org/wmt12/translation-task.html

matical constructs. The effort needed to post-edit and check human translated text is considerable and is even higher for output by MT systems, as they still cannot reach the correctness provided by a human translator.

The overall quality of MT output (without specific error analysis) can be scored over large sets of texts with automated metrics that use various matching techniques to compare to one or more reference translations such as n-gram matching (BLEU) (Papineni et al., 2002), synonym and paraphrase matching (METEOR) (Denkowski and Lavie, 2011) or the computation of an edit distance (TER) (Snover et al., 2006). Often, however, only one human reference translation is at hand and the scoring suffers from the fact that the MT system output might still be correct but is not close to the reference. Humans on the other hand, can compare and rank, for instance, the *fluency* and the *adequacy* of a translation against a reference. However, when these criteria have to be judged in large amounts of texts, the evaluation can be costly and time-consuming.

Besides judging the overall quality of MT output, one might also want to perform a detailed error analysis, either for post-editing the automatically obtained translation or in order to have a more diagnostically oriented MT evaluation that enables to analyse specific weaknesses of a system for certain types of errors. Error classification allows to compare translations produced by different MT systems and to formalize error counting and relationships between different types of errors (Flanagan, 1994). A framework (named FEMTI) of MT error analysis and a quality model interrelated with a classification of contexts where MT software is used has been provided by Hovy et al. (2002) and more recently, a classification of five linguistic error types (similar to the ones we use below, see Section 3) has been used to compare different MT system outputs, their strengths and weaknesses and the well given correlation to human judgement (Popovic and Ney, 2011).

These developments have also lead to actual software tools that support human MT evaluators in their task. A framework where the fluency and adequacy can be scored in a multiple choice way has been proposed by Koehn (2010, p. 219). Stymne (2011) introduced a tool (called BLAST) that learns from manually annotated data and reference translations to automatically mark errors in sentences output by MT systems.

As these are very recent ideas, there has not yet been a published evaluation of the usefulness of such tools. In the user study described in the following, we test whether an MT error spotting tool can help to increase the performance for MT error analysis.

3 Translation Error Spotting: A User Study

In the following, we used the BLAST tool to annotate errors in sentences that were output by a trained English to French SMT system built with the Moses toolkit (Koehn et al., 2007) and trained on the Europarl corpus (Koehn, 2005). 14 translated test sentences were randomly chosen from the UN EN/FR corpus[2]. We ignore the fact that the tool itself makes errors in annotating the mistakes in an MT system output – which might occur, as the error annotation tool does the mark-up automatically and therefore cannot reach perfect performance. The study does therefore not provide an evaluation of the tool itself but of its usefulness as is under real user conditions.

By analyzing the errors made by our SMT system we defined four types of errors: Error type A: *missing words* (words that were deleted by the system), type B: *grammatical mistakes*, type C: *untranslated words* (words that are still in English in the French output) and type D: *the use of wrong expressions* (wrongly translated words and constructs that a native French speaker would not use).

We randomly split the 14 sentences into two sets of 7 sentences each; Set 1 contained 33 errors in total (6 of error type A; 14 of type B; 3 of type C and 10 of type D). In Set 2 were 25 errors (4 of type A; 10 of type B; 5 of type C; 6 of type D). In both sets therefore, grammar errors were most frequent, followed by wrong expressions, deleted words and untranslated words. For Set 2, we pre-processed the 7 sentences with the MT error spotting tool in order to annotate errors for testing.

In principle, there are two groups of users for an MT error spotting tool: On the one hand there are the experienced evaluators of (machine) translation such as post-editors, translators and linguists. On the other hand there might be less experienced users that only occasionally have to check for translation quality and exactly for this reason

[2]Freely available at the same location as given in footnote 1.

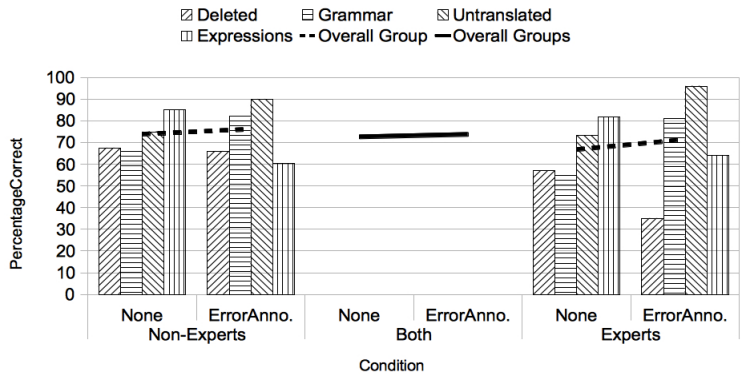

Figure 1: Percentages of correctly spotted MT errors for each error type, under the conditions of non-annotated errors and annotated errors for the Non-Expert and Expert group. The average performance over all errors in each user group is indicated with dashed lines, the continuous line shows the slight performance increase with error annotation when averaged over both user groups.

could profit of the tool that guides them to find the mistakes.

In the user study we performed, we distributed Sets 1 and 2 of 7+7 sentences to such two groups of users of which we have known that the one consisted mostly of linguists, some professional translators and a few people working on Machine Translation. The other group consisted of computer-literate users not working on translation or language specifically. In both groups, the subjects had knowledge of both languages, English and French. The user responses were handled anonymously. The subjects were only informed that they had to look for MT errors, but they did not know that an error spotting tool made the annotations.

All users were provided with the FR sentence output by the MT system and its original EN source sentence as a basis for comparison. The following examples illustrate the two conditions under which the two user groups had to find the errors (Example 1 without error annotation, Example 2 with annotated errors). The subjects had to indicate the words where they thought that there was something wrong or the words that were missing. After the study, we grouped the words indicated by the subjects according to the above-mentioned four error types:

EN-SRC: *While nearly every cell phone can play MP3 files, no MP3 player can make phone calls.*

1. FR-MT: **Tandis que presque chaque cellule téléphone peut jouer mp3 files, aucune mp3 acteur peut faire de téléphone.*

Errors: Type A: *des, ne*, Type B: *chaque*, Type C: *files*, Type D: *cellule téléphone, jouer, faire de téléphone*

EN-SRC: *As David Hume put it in discussing suicide, no man ever threw away life, while it was worth living.*

2. FR-MT-ANNO: **Comme David Hume [[mis en discuter, le suicide]] aucun homme jamais [[threw]] en vie, alors qu'[[il était bon de leur vie]].*

Errors: Type A: *à propos du, n'a*, Type B: *en*, Type C: *threw*, Type D: *mis en discuter, était bon de leur vie*

The survey was answered by 10 subjects in the Expert group and 12 subjects from the Non-Expert group. Apart from the percentage of correctly spotted errors, we also registered the time needed by each subject for each sentence.

4 Results

The performances on correctly spotted errors for both user groups of our study are given in Figure 1. Overall, under the condition of non-annotated errors, the Non-Expert group found more errors than the Experts. When adding error annotation to the sentences, both user groups performed slightly better, the Non-Experts still being better than the Experts, which indicates that spotting MT errors is a hard task even for Experts and that the performance might strongly be depending on each individual subject – whether Expert or not. Figure 1 also shows that spotting error type C (untranslated) was easiest, whereas type B (grammar) was the hardest to spot and here, indeed, the error annotation helped to increase the correctly found errors for both user groups. For error types A and D (deleted words and wrong expressions), the error annotation has an opposite effect: it lowers performance for both user groups.

We further analyzed our data for the second dependent variable in the study: the time needed by

the subjects to find the MT errors in the sentences. This is more difficult to measure, as some of the errors might be found almost at the same time and these findings might influence others or trigger a backtracking to parts of the sentence already read. We therefore measured the time needed by each subject to complete the error spotting per sentence, then normalized the time spent on one sentence by the number of errors that were in there and finally took the average amount of seconds over each set of sentences for both user groups, and individually for the Expert and Non-Expert group.

Figure 2 shows that the error annotation increases the average time needed when measured for both user groups. This changes when the time is measured for each user group individually: The Non-Experts become slower under the condition of error annotation, whereas the Expert group profits from the pre-annotated errors and can complete the task faster.

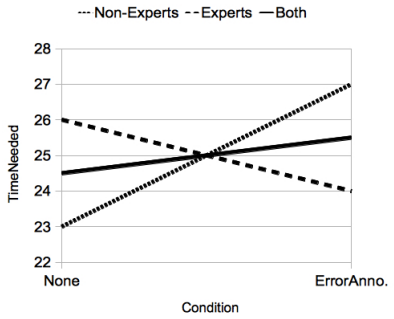

Figure 2: Average time needed in seconds to spot an MT error by Non-Experts, Experts and both groups with a significant interaction effect: The MT error annotation helps the Expert group to perform the task faster, while Non-Experts become slower.

In order to analyze the statistical significance of these performance differences, we used Multilevel Random Coefficient Modeling and more specifically Linear Mixed-Effects Models (LMEs) (Bliese, 2002; Pinhiero and Bates, 2000). LMEs allow to account for interaction effects, i.e. the fact that additional variance in ordinary statistical regression models can be dependent on group membership (*non-independence*), while normal contextual models 'pretend' that individual observations are made (*independence*).

We built two series of four LMEs: $m0$ – $m3$ for the first dependent variable in our study, $PercentageCorrect$; $tm0$ – $tm3$ for the second dependent variable $TimeNeeded$ (see Table 1 in Appendix A). The basic models ($m0$, $tm0$) to test against do not consider any condition, i.e. neither error annotation, nor user group membership. We can then factor in the two independent variables of our study: in $m1$ and $tm1$ we add $ErrorAnnotation$ as additional predictor on the percentage of spotted errors or the time needed, while in $m2$ and $tm2$, $Expertise$ is additionally modeled in to account for the possible covariance between the annotated errors and the group membership. The most complex models, $m3$ and $tm3$, finally model the interaction effects and test whether adding $ErrorAnnotation$ causes $PercentageCorrect$ or $TimeNeeded$ to vary *contrastively* depending on the group membership.

For $PercentageCorrect$, when testing for statistically significant differences between the models $m0 - m3$ with ANOVA (Analysis of Variance), adding $ErrorAnnotation$ does not make a significant difference in terms of correctly spotted errors for the two types of users. For $TimeNeeded$ however, the testing for statistical significance results in a significant difference of model $tm3$ compared to model $tm2$ at a level of $p < 0.001$, which means that a clear interaction effect consists of the Expert group becoming faster, i.e. the error annotation leads them in finding the errors, whereas the Non-Expert group is apparently disturbed by the error annotation, i.e. it takes them longer to find the errors.

5 Conclusion and Future Work

The paper has presented the interaction effects that have to be considered for an MT error spotting tool that is developed to facilitate the post-editing and error analysis of MT output. Our data analyses show that such a tool is only useful for an Expert group consisting of experienced subjects, and only in terms of the time needed to find the errors but not for the correctness of errors found. For the Non-Expert group the tool clearly increases the time needed to find the mistakes.

In future work it would be interesting to extend the study to include more test sentences, finer-grained types of errors and more subjects or user groups. Also, in terms of time needed, eye or gaze tracking could give indications to the error types that are hardest to find and where in a sentence and how long exactly a user's attention is focused most.

Acknowledgments

Many thanks go to the participants to the study, for the time spent on performing 'translation error spotting'. We also would like to thank Patrick Jermann from EPFL-CRAFT whose availability and insights for the data analyses were most helpful.

References

Bliese, Paul D. 2002. Multilevel random coefficient modeling in organizational research: Examples using SAS and S-PLUS. In Drasgow, F and N. Schmitt, editors, *Modeling in Organizational Research: Measuring and Analyzing Behavior in Organizations*, pages 401–445. Jossey-Bass, Inc.

Denkowski, Michael and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*, Edinburgh, UK.

Flanagan, Mary A. 1994. Error classification for mt evaluation. In *Proceedings of 1st Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 65–72, Columbia, MD.

Hovy, Eduard, Margaret King, and Andrei Popescu-Belis. 2002. Principles of context-based machine translation evaluation. *Machine Translation*, 17:43–75.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbs. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics (ACL), Demonstration Session*, pages 177–180, Prague, Czech Republic.

Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of Machine Translation. In *Proceedings of 40th Annual meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA.

Pinhiero, José C. and Douglas M. Bates. 2000. *Mixed-effects models in S and S-PLUS*. Springer Verlag, New York, NY.

Popovic, Maia and Hermann Ney. 2011. Towards automatic error analysis of machine translation output. *Computational Linguistics*, 37(4):657–688.

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

Stymne, Sara. 2011. BLAST: A Tool for Error Analysis of Machine Translation Output. In *Proceedings of 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT), System Demonstrations*, pages 56–61, Portland, OR.

Appendix A: ANOVA comparison of LMEs

Model	**Log.Lik**	**df**	**L.ratio**	**p.value**
$m0$	-3922.60	3		
$m1$	-3921.87	4	1.443078	0.2296
$m2$	-3921.19	5	1.357636	0.2439
$m3$	-3921.11	6	0.168698	0.6813
$tm0$	-3895.39	3		
$tm1$	-3889.91	4	10.9538	0.0009
$tm2$	-3889.91	5	0.00092	0.9758
$tm3$	-3866.48	6	46.8621	$<$ **.0001**

Table 1: ANOVA of LMEs $m0$ – $m3$ for $PercentageCorrect$, where the error annotation does not significantly influence the performance on correctly spotted errors for both user groups ($m3$) and models $tm0$ – $tm3$ for $TimeNeeded$ which is significantly shorter for the Expert group under the condition of pre-annotated errors ($tm3$). The testing is done by a comparison of each model to its correspondent predecessor (i.e. $m1$ vs. $m0$, $m2$ vs. $m1$ and so on).