Involving Language Professionals in the Evaluation of Machine Translation

Eleftherios Avramidis[1], Aljoscha Burchardt[1], Christian Federmann[1], Maja Popović[1], Cindy Tscherwinka[2], David Vilar[1]

[1]DFKI – Language Technology Lab, Berlin & Saarbrücken, Germany
[2]euroscript Deutschland, Berlin, Germany

[1]Firstname.Lastname@dfki.de
[2]Cindy.Tscherwinka@euroscript.de

Abstract

Significant breakthroughs in machine translation only seem possible if human translators are taken into the loop. While automatic evaluation and scoring mechanisms such as BLEU have enabled the fast development of systems, it is not clear how systems can meet real-world (quality) requirements in industrial translation scenarios today. The TARAXŰ project paves the way for wide usage of hybrid machine translation outputs through various feedback loops in system development. In a consortium of research and industry partners, the project integrates human translators into the development process for rating and post-editing of machine translation outputs thus collecting feedback for possible improvements.

Keywords: machine translation, human evaluation, error analysis

1. Introduction

Translation is a difficult task – even for humans. Machine translation (MT) quality has improved greatly over the last years, nevertheless the evaluation of machine translation output is an intrinsically difficult task as well. While ranking different translation systems (often using automatic scores such as BLEU (Papineni et al., 2001)) is an important first step towards their improvement, it does not provide enough scientific insights.

This paper describes the results of a large-scale human evaluation round carried out in the framework of the TARAXŰ[1] project. The approach rises from the need to detach MT evaluation from a pure research-oriented development scenario and to bring it closer to the end users. Therefore, we will present an evaluation round performed in close cooperation with translation industry. The evaluation process has been designed in order to answer particular questions closely related to the applicability of MT within a real-time professional translation environment. The whole evaluation task has been performed by qualified professional translators.

2. Human evaluation design

Several large-scale human evaluation rounds are foreseen within the duration of TARAXŰ project. The first round has already been completed and the results are presented in this paper. The involved languages were German, English and Spanish. Later evaluation rounds will include more languages that are not well studied to-date, such as Czech, Chinese and Russian. The evaluation tasks are performed by external Language Service Providers, as they offer human translation services and act as experts.

[1]http://taraxu.dfki.de/

Evaluation round. The translation outputs evaluated during the round which will be presented in this work are produced by German-to-English, English-to-German and Spanish-to-German machine translation systems. The test corpora consist of two domains: News taken from previous WMT tasks (1,030 sentences from the WMT 2010 News test set (Callison-Burch et al., 2010)) and technical documentation extracted from the freely available OpenOffice project (Tiedemann, 2009). Four different translation systems were considered:

Moses (Koehn et al., 2007): a phrase-based statistical machine translation (SMT) system, trained on the standard Europarl and News corpora of WMT 2010.

Google Translate: a web-based machine translation engine also based on statistical approach. Since this system is known as one of the best MT engines, it has been included in order to allow us to assess the performance level of our SMT system and also to compare it directly with other MT approaches.

Lucy MT (Alonso and Thurmair, 2003): a commercial rule-based machine translation system with sophisticated hand-written transfer and generation rules, which has shown good performance on previous shared tasks.

Trados: a professional Translation Memory System (TMS) whose translation memory has been enriched with the same parallel data that our SMT system was trained on.

The obtained outputs are then given to the professional human annotators in order to perform the following three sentence-level evaluation tasks:

Ranking: *rank the outputs* of four different MT systems according to how well these preserve the meaning of the source sentence.

Error classification: select the best ranked translation output and *define the two main types of errors* (if any) in it. A following subset of the error types suggested by (Vilar et al., 2006) is used: missing content word(s), wrong content word(s), wrong functional word(s), incorrect word form(s), incorrect word order, incorrect punctuation and other error.

Post-editing: select the translation output *which is easiest to post-edit* (which is not necessarily the best ranked) and perform the editing. The translators were asked to perform only the minimal post-editing necessary to achieve acceptable translation quality.

The browser-based evaluation tool Appraise (Federmann, 2010) to collect human judgments and post-editings. It should be noted that the Google Translate system was not considered as an option for error classification and post-editing. We took this decision in order to avoid futile efforts because we have no way to influence on improving this system. In case Google was the best ranked system, the translators were offered the second ranked system for the classification task, whereas for the post-editing task they could choose among ranks 2, 3 and 4.

3. Results

In this section, we will present results for each evaluation task. It should be noted that the Trados memory was filled automatically with the WMT News texts which are also a part of the test data – therefore Trados provides ecxeptionally good translation results for those sentences that are very similar to those of the WMT News texts.

3.1. Ranking

The results for the ranking task are shown in Table 1. The first row presents the overall average ranks for the four listed systems, **bold face** indicating the best system. Furthermore, the results are presented separately for each translation direction, namely German-to-English, Spanish-to-German and English-to-German, as well as for each domain, namely WMT News and OpenOffice technical documentation.

human ranking	Lucy	Moses	Google	Trados
Overall	2.00	2.38	**1.86**	3.74
de-en	2.01	2.46	**1.73**	3.80
es-de	**1.85**	2.42	1.99	3.72
en-de	2.12	2.28	**1.89**	3.71
News	**2.52**	2.59	2.69	2.21
OpenOffice	1.72	2.77	**1.56**	3.95

Table 1: Human ranking results as the average position of each system in each task.

It can be observed that the ranks of the machine translation systems are comparably close. A noticeable result is that Google performs worst on the WMT corpus although the data should – in principle – have been available online for training. This might, however, explain the good perfomance of this web-based system on the OpenOffice corpus. On the other hand, for the OpenOffice task Moses showed the worst performance – the reason is that it has been trained only on the out-of-domain WMT data.

BLEU (%)	Lucy	Moses	Google	Trados
Overall	15.6	15.0	**20.2**	2.9
de-en	22.7	18.8	**29.8**	5.0
es-de	12.3	12.9	**15.9**	1.8
en-de	13.6	14.6	**17.2**	2.5
News	14.1	16.3	**17.7**	2.1
OpenOffice	19.5	12.0	**26.6**	4.9

Table 2: Average BLEU scores (%) for each system in each task.

Table 2 shows the average BLEU scores for illustration. The main difference is that the Google Translate system is the best one for each language pair and each task. This can be expected, since in several WMT evaluation tasks it is shown that the correlation between the BLEU score and the human rankings is not particularly high, mainly because the BLEU score is biased towards statistical systems thus underestimating rule-based systems (such as Lucy).

3.2. Error classification

The results of the error classification are presented in Table 3. It can be seen that the most frequent errors in all systems are *wrong lexical choices* (wrong content and functional words), and the next frequent error type is *incorrect word order*. This indicates the need for improvement of reordering and lexical choice techniques for all translation approaches. Another interesting observation is the very low number of missing content words for the Lucy system.

	Lucy	Moses	Trados
Missing content word(s)	3.2	16.8	12.6
Wrong content word(s)	34.6	24.6	33.2
Wrong functional word(s)	18.6	11.8	11.0
Incorrect word form(s)	13.1	14.6	9.1
Incorrect word order	16.1	22.0	13.4
Incorrect punctuation	3.7	3.4	2.1
Other error	10.8	6.7	18.6

Table 3: Human error classification: overall percentage of errors for each translation system.

3.3. Post-editing

A central question that is to be answered by the evaluation round is whether there is a difference between those sentences that are ranked best (i.e. that are the best MT result) and those sentences that are chosen by human professionals as the easiest for post-editing. It has been shown that 74% of those hypotheses selected for post-editing were ranked as the best or the second best in the ranking task. 20% were ranked third, and 6% had the worst rank. An example of discrepance between the "best ranked" and "easiest to post-

edit" sentence is presented in Table 4. The chosen sentence contains untranslated words (Warenhäusern) and therefore got a bad ranking – on the other hand, such a lexical error is very easy to post-edit. Another example is a missing or extra negation particle ("not") – this is a very severe error in terms of translation quality, i.e. conveying the meaning of the source sentence, but very easy to post-edit.

Rank	Translation output
1	Our experience shows that the majority of the customers in the three department stores views not more at all on the prices.
2	Our experience shows that the majority of the customers doesn't look on the prices in the three department stores any more.
3	Our experience shows that the majority of the customers does not look at the prices anymore at all in the three department stores.
4	Our experience shows that the majority of customers in the three Warenhäusern do not look more on prices.
Edited	Our experience shows that the majority of customers in the three department stores no longer look at the prices.

Table 4: Example of discrepance between ranking and post-editing: the worst ranked sentence is chosen for post-editing.

3.3.1. Automatic classification of edits

In order to obtain more insight into the nature of errors corrected by post-editing thus learning more about differences between the systems and possibilities for improvement, automatic error analysis is performed using the post-edited translations as references. The original translation hypotheses are compared with the post-edited ones in order to estimate which type of editing are most frequent for each of the systems. The following five types of edits (Popović and Ney, 2011) are taken into account: correcting word form (morphology), correcting word order, adding missing word, deleting extra word and correcting lexical choice. Table 5 presents overall percentage for each of the five correction types for the three systems.

The main observation from the overall results is that the most frequent correction for all systems is the lexical choice and the next frequent correction is the word order, which suggests the same as the human error classification: the main weak points of all systems are incorrect lexical choice and incorrect word order. Furthermore, it can be seen that the rule-based Lucy system better handles morphology and word ordering, whereas the statistical-based Moses system produces less lexical errors.

The results for Trados should be interpreted as follows. A large portion of the evaluation data did not reach a high degree of similarity for the content of the Trados Memory. Therefore many sentences remained untranslated which accounts for the high number of lexical errors. The low number of morphological and reordering errors is easily explained by the fact that the content of the memory stems from human translations in the first place. The fact that morphological and reordering errors occur at all indicates that the training material that has been used to enrich the memory already contained impure translations.

More detailed results can be seen in Table 6. The percentage of edits is presented for each language pair and each domain. However, these detailed results are not reported for Trados for the reasons explained above, but only for Lucy and Moses system. The following can be observed:

- Word forms: Lucy performs significantly better than Moses for the English-to-German task. For the other tasks, results are comparable. The reason is the rich morphology of the German language which can be better dealt with a rule-based system. Nevertheless, Spanish-to-German is "easier" for statistical systems than English-to-German in terms of word forms since the Spanish morphology is richer than English. These results indicate possibilities for improving English-to-German Moses system.
- Word order: Lucy performs better than Moses for all language pairs and domains. Possible improvements could consist of improving reorderings for the Moses systems.
- Missing words: again significantly lower numbers for Lucy outputs. One of the reason both for reorderings and for missing words could be the special positions of German verbs which are hard to deal with by statistical translation systems. This indicates a possibility for improvement as well.
- Extra words: for this error type, Moses performs better than Lucy. This can be attributed to word and phrase penalties in statistical translation systems.
- Lexical choice: for German-to-English and for the News domain, both systems have similar performance. However, for translation into German, Moses performs significantly better than Lucy. The probable reason is that whereas rule-based systems deal better with linguistic characteristics, statistical ones better handle lexical variations if trained in-domain. Further illustration of this can be seen from the results of the OpenOffice domain: Lucy performs significantly better, since Moses was trained on the out-of-domain WMT data. Possible directions for improvements that are currently being studied are including appropriate terminologies into Lucy, as well as in-domain training or domain adaptation for Moses.

4. Summary and Outlook

In this paper, we have shown evidence that a human-centric hybrid approach to machine translation is a promising way of further improvement of this technology. into industrial translation workflows. Even in this early stage, the TARAXŰ project has generated positive feedback and raised interest, especially from the side of the industrial partners. By the time of writing, the first (pilot) evaluation round of the TARAXŰ project including the language pairs German→English, English→German,

	correcting word form	correcting word order	adding missing word	deleting extra word	correcting lexical choice overall
Lucy	4.3	7.0	4.4	6.2	23.7
Moses	4.9	9.0	7.5	4.9	21.8
Trados	2.6	4.9	8.1	6.5	47.7

Table 5: Five types of edits for three translation systems: values are normalised over the total number of words generated by the corresponding system.

Lucy/Moses	correcting word form	correcting word order	adding missing word	deleting extra word	correcting lexical choice
de-en	2.4/2.6	7.8/9.7	4.3/7.3	6.3/5.3	20.6/21.6
en-de	5.8/6.4	7.4/8.8	5.8/6.8	5.0/3.8	26.3/20.8
es-de	5.9/5.9	5.9/7.3	3.2/8.3	7.2/5.4	26.3/22.6
News	4.3/5.7	6.8/8.6	3.7/6.6	5.3/4.7	19.2/20.7
OpenOffice	2.9/4.1	6.8/11.2	2.8/7.0	6.3/8.0	16.6/26.9

Table 6: Five types of edits separately for each language pair and each domain: values are normalised over the total number of words generated by the corresponding system. Trados is not taken into account (see the main text).

and Spanish→German has finished and further evaluation rounds are being planned that will iteratively extend the numbers of languages covered and include questions related to topics such as error types, post-editing efforts for each system, effects of pre-editing, etc.

Acknowledgments

This work has been developed within the TARAXŰ project financed by TSB Technologiestiftung Berlin – Zukunftsfonds Berlin, co-financed by the European Union – European fund for regional development.

5. References

Juan A. Alonso and Gregor Thurmair. 2003. The comprendium translator system. In *Proceedings of the Ninth Machine Translation Summit*.

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, July. Association for Computational Linguistics. Revised August 2010.

Christian Federmann. 2010. Appraise: An Open-Source Toolkit for Manual Phrase-Based Evaluation of Translations. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, May.

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 ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. IBM Research Report RC22176(W0109-022), IBM.

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

Jorg Tiedemann. 2009. News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Advances in Natural Language Processing*, volume V, chapter V, pages 237–248. Borovets, Bulgaria.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error Analysis of Machine Translation Output. In *International Conference on Language Resources and Evaluation*, pages 697–702, Genoa, Italy, may.