Learning from human judgments of machine translation output

Maja Popović, Eleftherios Avramidis, Aljoscha Burchardt,
Sabine Hunsicker*, Sven Schmeier, Cindy Tscherwinka*, David Vilar, Hans Uszkoreit
DFKI / Berlin, Germany
name.surname@dfki.de

* euroscript Deutschland / Berlin, Germany
name.surname@euroscript.de

Abstract

Human translators are the key to evaluating machine translation (MT) quality and also to addressing the so far unanswered question when and how to use MT in professional translation workflows. Usually, human judgments come in the form of ranking outputs of different translation systems and recently, post-edits of MT output have come into focus. This paper describes the results of a detailed large scale human evaluation consisting of three tightly connected tasks: ranking, error classification and post-editing. Translation outputs from three domains and six translation directions generated by five distinct translation systems have been analysed with the goal of getting relevant insights for further improvement of MT quality and applicability.

1 Introduction and related work

A widely used practice for MT evaluation is ranking outputs of different machine translation systems by human annotators, e.g. in WMT shared tasks (Callison-Burch et al., 2012). While this is an important step towards an understanding of their quality, it does not provide enough scientific insights. In the last years, human error analysis is often carried out in order to better understand some phenomena (Vilar et al., 2006), and recently more and more attention is paid to various aspects of post-editing effort (Specia, 2011; Koponen, 2012). However, to the best of our knowledge, no study has been carried out yet which puts all these aspects together.

This paper describes the results of detailed human evaluation covering all three aspects: ranking, error classification and post-editing. The approach arises from the need to detach MT evaluation from a pure research-oriented development scenario and to bring it closer to the end users. Therefore, evaluation has been performed in close co-operation with translation industry. All evaluation tasks have been performed by qualified professional translators. 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, such as: Is the best ranked translation output also the best for post-editing? What criteria guide the selection process? Which error types are occurring in different translation systems? What types of post-edit operations are carried out in particular translation outputs? What are the differences between different language pairs and domains?

2 Human evaluation design

Human evaluation has been performed focussing on three different aspects: ranking, analysis of translation errors and post-editing effort. The involved languages were German, English, Spanish and French. The evaluation tasks were performed by external language service providers. We asked them to treat this job as any other job, i.e. apply the usual standards. They usually assign one translator per test set, and the raters are normally native speakers of the target language. We did not influence the number of translators working on it.

2.1 Translation systems used

The evaluated translation outputs presented in this work are produced by German-English, German-French and German-Spanish machine translation

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 231–238.

systems in both directions. The test sets consist of three domains: news texts taken from WMT tasks (Callison-Burch et al., 2010), technical documentation extracted from the freely available OpenOffice project (Tiedemann, 2009) and client data owned by project partners.

The following translation systems were considered:

Moses (Koehn et al., 2007): a phrase-based statistical machine translation (SMT) system trained on news texts and technical documentation (no client data were available for training).

Jane (Vilar et al., 2010): a hierarchical phrase-based SMT system trained on news texts and technical documentation (no client data were available for training).

Lucy MT (Alonso and Thurmair, 2003): a commercial rule-based machine translation (RBMT) system with sophisticated hand-written transfer and generation rules adapted to domains by importing domain-specific terminology.

RBMT: Another widely used commercial rule-based machine translation system whose name is not mentioned here.[1]

Google Translate[2]: a web-based machine translation engine also based on statistical approach. Since this system is known as one of the best general purpose 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.

Trados[3]: a professional Translation Memory System (TMS) whose translation memory has been enriched with the same News parallel data that our SMT systems were trained on.

The translation outputs were generated by the described systems prepared for the German-English, German-French and German-Spanish language pairs in both directions, and then given to the professional human annotators in order to perform the defined sentence-level evaluation tasks using the browser-based evaluation tool Appraise (Federmann, 2010). The reference translation was not shown in any task, only the source sentence and the translation output.

[1]We have been asked to anonymise this system; for this reason, we refer to Lucy and this other system as RBMT1 and RBMT2 without revealing which is which.

[2]`http://translate.google.com/`

[3]`http://www.trados.com/en/`

2.2 Evaluation tasks

The evaluation tasks were defined as follows:

Ranking: for each source sentence, *rank the outputs* of five different MT systems (Trados is excluded, explanation below) according to how well these preserve the meaning of the source sentence. Ties were allowed.

Error classification: Error classification is a rather complex and time-consuming task, therefore only translation outputs generated by a subset of source sentences was processed. The following error categories on the word level were taken into account: incorrect lexical choice, terminology error, morphological error, syntax error, misspelling, insertion, punctuation error and other error. For each category, two grades were defined: severe and minor. In addition, the category of missing words was defined on the sentence level: the evaluators should only decide if omissions are present in the sentence or not. For the translation outputs of particular low quality, a special category "too many errors" was offered.

Post-editing: This task was divided into two subtasks:

- **Select and post edit**: for each source sentence, select the translation output *which is easiest to post-edit* (which is not necessarily the best ranked) and perform the editing.
- **Post-edit all**: For this task again a subset of source sentences was taken into account due to complexity of post-editing translations of a low quality. For each source sentence in the selected subset, post-edit all produced translation outputs.

For both post-editing sub-tasks, the translators were asked to perform only the minimal post-editing necessary to achieve acceptable translation quality. An option "Translate from scratch" was available as well: the translators were instructed to use it when they think that a

	News	OpenOffice	Client	Total
de-en	1788	418	500	2706
de-es	514	414	548	1476
de-fr	912	412	382	1706
en-de	1744	414	0	2158
es-de	101	413	1028	1542
fr-de	1852	412	0	2264
Total	6911	2483	2458	11852

Table 1: Test sets – number of source sentences per language pair and domain.

completely new translation is faster than post-editing.

The sizes of test sets for each language pair and domain can be seen in Table 1.
The results are presented in the following sections. It should be noted that the Google Translate system was not considered as an option for error classification and post-editing, and the Trados memory was not used for ranking. The reasoning behind these decisions is:

- Google: This system was taken into account only for the sake of comparison – we have no way to influence on potential improvements of this system.

- Trados: Translation memories are the most widely used by human translators, therefore they should be part of the evaluation. Their performance depends on their content, which is usually extended and maintained over years. However, within the scope of this work, it was not possible to design a fully fair comparison between this technology and MT systems in general. Translation Memories (TM) like Trados are usually filled over time by human translators in production workflows. As we did not have such resources for the TARAXŰ data, we automatically filled one TM with part of the data. For this reason, we cannot expect it to perform like a ”normal” TM. We integrated it into our tool chain on a more explorative basis.

3 Results

3.1 Ranking

The results for the ranking task are shown in Table 2. The rank for each system is calculated as percentage of sentences where the particular system has better or equal rank than the other systems. The first row presents the overall ranks for the five systems, then separate results are presented for each domain as well as for each translation direction. **Bold face** indicates the best system and *italic* the second best system.

Google performs best most often. However, for the German-to-Spanish translation, RBMT1 performs almost equally. In general, it can be observed that the two rule-based systems perform comparably well except for the language pair German-to-Spanish where the RBMT1 system performs significantly better than RBMT2. This strengthens the observation that rule-based systems heavily rely on the amount of effort that is put into the development of certain language pairs. Apart from that, all systems perform comparably close.

3.2 Error classification

The error classification results are presented in Table 3 – for each translation system, raw error counts in its output were normalised over total number of sentences generated by this system. Thus, the percentage “12.9” in the column ”Jane” and row ”lexical choice (minor)” can be interpreted as follows: in 100 sentences translated by Jane there is a total of 12.9 minor lexical choice errors. The exact error distribution among the sentences is not indicated by these numbers. It can be seen that the most frequent errors in all systems are *wrong lexical choices*, *terminology errors* and *syntax errors*. One observation is that the rule-based systems have more problems with terminology errors than statistical systems. On the other hand, they produce less severe morphological errors and significantly less omissions.

Another interesting observation is that the number of severe errors for all error types is higher than the number of minor errors. Exceptions include incorrect lexical choice for the rule-based systems, where both are nearly the same.

From this table, one can generate recommendations for the most effective system improvements. It seems that better syntactic modelling and improvement of terminology use should have the best effect. Lexical choice in general is also an issue, but the solution would probably require modelling of meaning, context and world knowledge, which may be even more demanding.

rank$\geq$	Google	Jane	Moses	RBMT1	RBMT2
Overall	**74.4**	47.6	57.6	*69.3*	67.4
News	**72.8**	42.6	55.3	*68.0*	66.2
OpenOffice	**66.5**	45.7	55.1	57.0	*58.9*
Client	**82.6**	58.2	64.0	*79.6*	75.2
de-en	**76.8**	43.8	51.8	63.2	*63.4*
de-es	**77.6**	53.6	55.8	*77.0*	44.5
de-fr	*69.6*	50.3	60.2	**69.9**	69.3
en-de	**76.3**	42.2	54.0	65.7	*67.7*
es-de	*77.1*	62.6	69.4	75.1	**78.4**
fr-de	68.2	40.3	55.6	*68.5*	**79.4**

Table 2: Ranking results: *rank*$\geq$ is defined as percentage of sentences where the particular system is ranked better or equal than the other systems.

N_{err}/N_{sent} (%)		Jane	Moses	RBMT1	RBMT2
lexical choice	minor	12.9	15.1	23.1	18.4
	severe	22.1	21.8	23.0	18.7
terminology	minor	5.9	6.6	11.6	10.4
	severe	27.0	29.3	44.2	37.2
morphology	minor	17.8	8.0	9.2	7.8
	severe	14.3	16.2	11.7	11.9
syntax	minor	7.8	7.6	9.8	9.7
	severe	27.5	36.6	28.3	28.4
misspelling	minor	5.8	2.5	3.4	3.8
	severe	5.6	1.7	1.8	1.3
insertion	minor	3.6	3.1	6.7	4.1
	severe	7.0	6.7	7.1	7.1
missing words		19.7	20.5	10.1	8.7
punctuation	minor	5.4	5.4	9.8	8.1
	severe	2.6	2.9	1.7	4.0
other	minor	1.3	1.5	1.6	2.8
	severe	3.4	2.3	2.7	3.3
too many errors		41.2	42.4	33.3	41.2

Table 3: Error classification results: for each system, raw error counts are normalised over the total number of evaluated sentences generated by this system.

3.3 Post-editing

Human post-edits were used to study the difference between selecting sentences for easier post-editing and for ranking (based on meaning), to compare different translation systems, as well as to compare sentences selected for post-editing with the rest of the sentences.

3.3.1 Select and post-edit

Table 4 presents the percentage of sentences selected for post-editing for four machine translation systems and Trados translation memory, **bold face** indicating the most selected system. The RBMT1 system is generally selected most often. A possible explanation is the fact that this system mostly produces well structured sentences even if they are only partially correct translations of the input sentence. In contrast, ungrammatical outputs (often generated by statistical systems) might contain better "material" but e.g. in the wrong word order therefore being dispreferred. Future work is needed to shed more light on these decisions. For the News task, it is comparable with the RBMT2 system, for the OpenOffice tasks with both statistical systems, and for the client data it is outperformed by Trados – this could be expected, since translation memories were filled with various types of client data. As for different translation directions, no significant differences can be observed except for the French-to-German where Trados was selected very seldom: nevertheless, for this translation direction no client data were available.

3.3.2 Selection for post-editing vs. ranking

Table 5 shows percentage of sentences selected as the best for post-editing for each of four ranks from the ranking task (1 being the best, 4 the worst). Intuitively one could expect that the tasks are (almost) the same; however, only 70% of selected sentences were ranked as best. About 20% of selected sentences were ranked as second best, and 10% had one of the two lowest ranks.

Table 6 shows an example of a third ranked translation selected for post-editing extracted from German-to-English client data: one word remained untranslated thus significantly degrading the quality. On the other hand, the correction of this sentence is easy – it requires only one edit operation, namely replacing this (German) word with the correct (English) one.

	% of selected sentences				
rank	Overall	Jane	Moses	RBMT1	RBMT2
1	71.7	65.1	63.9	78.8	73.6
2	19.1	21.4	22.9	15.3	18.7
3	6.5	9.3	9.2	4.2	5.5
4	2.7	4.2	4.0	1.7	2.1

Table 5: Percentage of sentences with a given rank selected as the best for post-editing; no Google for selection, no Trados for ranking.

source	Dazu ist ein Schraubendreher erforderlich.
Rank	Translation output
1	For this purpose a screwdriver is necessary.
2	In addition a screwdriver is necessary.
3	This requires a Schraubendreher.
4	This would require an Schraubendreher required.
edit(3)	This requires a screwdriver.

Table 6: Example of discrepance between ranking and post-editing: the third ranked sentence is chosen for post-editing.

3.3.3 Edit operations for different translation systems

In order to obtain better insight into the nature of post-edit corrections and to learn more about differences between the systems, automatic edit analysis was performed by using the Hjerson tool (Popović, 2011) using the post-edited translations as references. The following five types of edits were distinguished: correcting word form (morphology), correcting word order, adding missing word, deleting extra word and correcting lexical choice. The results are presented in the form of edit rates, i.e. the total number of edits normalised over the total number of words.

The overall edit rates for each of the five edit types for four machine translation systems and Trados memory are shown in Table 7. The most frequent correction for all systems is the lexical choice, followed by the word order. Furthermore, it can be seen that the rule-based systems better handle morphology and induce less missing words. The same tendencies were also observed in the human error classification results presented in Section 3.2, indicating that efforts should definitely be put on improving syntactical and lexical models

$N_{selected}/N_{total}$ (%)	Jane	Moses	RBMT1	RBMT2	Trados
Overall	12.4	18.6	**31.4**	23.7	13.8
News	9.2	20.8	31.2	**33.1**	5.7
OpenOffice	26.3	28.0	**29.9**	14.8	1.2
Client	8.0	8.4	32.8	14.4	**36.5**
de-en	14.6	20.5	**29.0**	25.1	10.8
de-es	13.3	17.0	**45.4**	4.3	20.3
de-fr	9.5	18.9	**29.8**	28.5	13.3
en-de	10.7	22.9	**28.2**	23.4	14.8
es-de	15.7	14.2	**36.4**	11.1	22.7
fr-de	11.4	18.5	20.9	**49.0**	0.3

Table 4: Percentage of sentences selected for post-editing for four machine translation systems and translation memory Trados.

	correcting word form	correcting word order	adding missing word	deleting extra word	correcting lexical choice
Jane	4.6	6.5	7.7	**21.2**	37.4
Moses	4.2	7.8	9.4	**21.0**	27.6
RBMT1	3.9	7.3	4.4	7.3	23.9
RBMT2	4.7	9.4	4.9	7.8	26.2
Trados	2.0	3.3	8.2	7.9	**61.7**

Table 7: Five types of edits for five translation outputs: values are normalised over the total number of words generated by the corresponding system.

for all machine translation systems. As for Trados, the majority of edits are lexical due to untranslated portions which are not contained in the memory. Apart from that, an unusally large amount of insertions can be observed in statistical translation outputs.

In order to understand better the described results, further analysis on different language pairs and domains was carried out. Different language pairs do not seem to have a big effect; on the other hand some edit rates significantly differ for distinct domains. The results for separated domains are presented in Table 8. First of all, it can be seen that a large number of insertions is produced by statistical systems only for the client data (see example in Table 9). The reason for this is the fact that the systems are not trained on this type of data, but on the News data having much longer sentences – the average sentence length for the News data is 22.6 words, for the technical documentation 14.6 words and for the client data 9.6 words. Another interesting observation is lower amount of lexical edits for Trados client data outputs – since the memory is actually filled with this data type, the translations of the client data are much better than of the other

source	<Frm id=”15”/>Gefahr durch elektrische Spannung
Jane	**The European Union** Frm id = ”15” /. risk posed by voltage
Moses	**This happening** Frm ID = ”” / 15-in danger from voltage
RBMT1	<Frm id=”15”/>Danger through electric tension
RBMT2	<Frm id=”15”/>Danger by electrical tension
Trados	<Frm id=”15”/>Gefahr durch elektrische Spannung
edit	<Frm id=15/>Danger by electrical tension

Table 9: Example of extra words produced by statistical systems on client data.

domains. Furthermore, more reordering edits were performed in the News data – another effect of the larger average sentence length.

3.3.4 Selected sentences vs. the rest

In Section 3.3.2 it was shown that there is a difference between the selection mechanisms for

		correcting word form	correcting word order	adding missing word	deleting extra word	correcting lexical choice
News	Jane	5.4	10.8	6.3	7.4	25.2
	Moses	5.1	10.9	5.4	7.4	22.8
	RBMT1	4.0	9.5	4.5	7.3	23.2
	RBMT2	4.7	11.1	4.5	6.9	23.4
	Trados	1.6	3.5	8.7	6.8	**75.7**
OpenOffice	Jane	6.3	6.7	6.2	7.3	26.6
	Moses	6.2	7.5	6.6	6.8	25.0
	RBMT1	4.5	5.3	4.6	7.0	24.1
	RBMT2	5.4	7.5	4.3	8.7	27.0
	Trados	4.0	5.0	5.8	12.2	**54.9**
Client	Jane	4.8	6.1	8.5	**17.9**	29.9
	Moses	4.5	7.5	8.5	**18.0**	26.4
	RBMT1	3.3	4.9	4.0	7.7	24.9
	RBMT2	4.2	8.3	6.2	8.7	30.1
	Trados	0.8	1.4	9.2	5.8	**44.5**

Table 8: Five types of edits for five translation outputs separately for each domain: values are normalised over the total number of words generated by the corresponding system.

ranking translation outputs based on meaning and for choosing the output most suitable for post-editing. The first results indicated that a lower edit distance was an important factor for the latter. A further question is if only the total edit distance matters, or some edit types are more or less preferred than the others. In order to examine this, five type of edit rates described in Section 3.3.3 were calculated separately on the selected sentences and on the rest, and then the relative difference for each edit rate $(editRate(rest) - editRate(sel))/editRate(rest)$ was calculated.

The results are presented in Table 10. The first row shows the edit distances for selected sentences and for the rest, and the second row presents the relative differences. Overall, the relative difference between edit distances of two sentence sets is 36%, meaning that in the selected sentences there are 36% less edit operations than in the rest of the sentences. The differences are similar for all edit types being between 30% and 36%, only the missing words have 45% – adding missing words seems to be the least preferred edit operation. These results are offering interesting directions for future work, e.g. investigating differences separately for language pairs and domains, examining translations from scratch and comparing with other translation outputs, etc.

4 Conclusions and outlook

In this work, we have presented the results of a broad human evaluation where human translators have judged machine translation outputs of distinct systems via three different tasks: ranking, error classification and post-editing. We have systematically analysed the obtained results in order to better understand the selection mechanisms of human evaluators as well as differences between machine translation systems. The most severe problems that machine translation systems encounter are related to terminology/lexical choice and syntax. Human annotators seem to prefer well-formed sentences over unstructured outputs, even if the latter contain the "material" needed for creating a good translation. Further work is needed to study these hypotheses in more depth.

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. Thanks to our colleague Christian Federmann for helping with the Appraise system and to Lukas Poustka for the engineering work on preprocessing the test sets.

	total	form	order	missing	extra	lexical
edit rates (selected/rest)	39.0/61.0	2.9/4.5	5.3/7.8	3.6/6.7	6.0/9.0	21.2/33.0
relative difference (%)	36.0	36.2	31.9	**45.8**	34.2	35.8

Table 10: Total edit distance and five distinct types of edits for selected sentences and not selected sentences (first row) and their relative differences (second row).

References

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

Callison-Burch, Chris, 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.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, page 1051, Montral, Canada, June. Association for Computational Linguistics.

Federmann, Christian. 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.

Koehn, Philipp, 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.

Koponen, Maarit. 2012. Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, page 181190, Montral, Canada, June. Association for Computational Linguistics.

Popović, Maja. 2011. Hjerson: An open source tool for automatic error classification of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, pages 59–68.

Specia, Lucia. 2011. Exploiting Objective Annotations for Measuring Translation Post-editing Effort. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT 11)*, pages 73–80, Leuven, Belgium, May.

Tiedemann, Jorg. 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.

Vilar, David, 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.

Vilar, David, Daniel Stein, Matthias Huck, and Hermann Ney. 2010. Jane: Open source hierarchical translation, extended with reordering and lexicon models. In *ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, pages 262–270, Uppsala, Sweden, July.