The TaraXŰ Corpus of Human-Annotated Machine Translations

Eleftherios Avramidis[1], Aljoscha Burchardt[1], Sabine Hunsicker[2], Maja Popović[1], Cindy Tscherwinka[2], David Vilar[3], Hans Uszkoreit[1]

[1] DFKI, Germany
[2] euroscript Germany
[3] Pixformance, Germany

firstname.lastname@dfki.de[1]
firstname.lastname@euroscript.de[2]

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. This paper describes the corpus developed as a result of a detailed large scale human evaluation consisting of three tightly connected tasks: ranking, error classification and post-editing.

Keywords: corpus, machine translation, evaluation

1. Introduction

This paper describes the corpus created in the framework of the TARAXŰ[1] project. The approach rises from the need to detach Machine Translation (MT) evaluation from a pure research-oriented development scenario and to bring it closer to the end users. Therefore, three evaluation rounds were performed in close co-operation with translation industry. The evaluation process has been designed in order to answer particular questions closely related with the applicability of MT within a real-time professional translation environment. All evaluation tasks have been performed by qualified professional translators.

The evaluation rounds, resulting in the corpus discussed in this paper, built on one another in a logical procession: the first round created baseline results, whereas each further round was concerned with more elaborated measuring methods and more specific factors impacting translation quality. Findings of evaluating the results from these rounds have been published in (Avramidis et al., 2012) and (Popović et al., 2013). Parts of the corpus have more recently been used in the QTLaunchPad project[2] where they served as the basis for a more detailed error analysis. The corpus is openly available through META-SHARE.

2. The corpus

The corpus contains machine translations created by several MT engines for the same source text as well as the output of different evaluation tasks. It covers the following language pairs:

- German ↔ English
- German ↔ Spanish
- German ↔ French
- Czech ↔ English

[1] http://taraxu.dfki.de
[2] http://www.qt21.eu/launchpad/

translation direction	source sentences	source words
German→English	3731	87939
English→German	5497	130003
German→French	1218	25912
French→German	3618	90228
German→Spanish	1636	33347
Spanish→German	3760	84823
Czech→English	731	12009
English→Czech	622	13722

Table 1: Total number of source sentences and running words used in evaluation tasks for each translation direction.

As each evaluation round dealt with different scenarios and language combinations, the amount of data available per evaluation task, translation direction and translation system differs. Table 1 gives an overview of the total amount.

Two different domains were used: news text from the WMT10 and WMT11 shared tasks (Callison-Burch et al., 2011) and technical documentation from OpenOffice and KDE (Tiedemann, 2009). Except for KDE, all corpora were used in all rounds. From round two on, some input from earlier rounds was reused to test the reliability of evaluators. Translations were produced using several different translation engines:

Moses (Koehn et al., 2007): a phrase-based statistical machine translation (SMT) system trained on news texts and technical documentation.

Jane (Vilar et al., 2010): a hierarchical phrase-based SMT system trained on news texts and technical documentation.

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.[3]

Google Translate[4]: a web-based machine translation engine also based on statistical approach.

Trados[5]: 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.

Moses and Jane were trained on news text and technical data following the WMT11 baseline. The Trados translation memory used the same parallel data as the SMT systems to provide appropriate matches where possible. The rule-based systems were adapted to the source texts by importing terminology. Google Translate is known as one of the best general purpose MT engines. As such, it has been included in order to allow us to assess the performance level of our SMT systems and also to compare it directly with other MT approaches.

3. Evaluation tasks

The evaluation itself was performed by professional translators working for several language service providers. Choosing properly educated, fully bilingual translators ensured that the focus was on the usability of the output given the evaluation task. Translators' feedback was collected using the graphical interface of Appraise[6] (Federmann, 2010). The released corpus contains the output from four different tasks:

3.1. Ranking

This task was part of all evaluation rounds and its definition is as follows:

- for each source sentence, rank the outputs of different MT systems according to how well these preserve the meaning of the source sentence.

Ties were not allowed in the first evaluation round, but in consecutive rounds. This is the most basic evaluation task as it is only concerned with comparing the understandability of different translation outputs. Translation outputs of all source sentences in Table 1 were compared.
It should be noted that the Google Translate system was considered only for this task. We took this decision in order to avoid futile efforts because we have no way to influence on improving this system – we included it in the evaluation only for the sake of comparison with the other MT engines which we could improve.

[3]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.

[4]http://translate.google.com/

[5]http://www.trados.com/en/

[6]Appraise modifications branch for the aims of our project is available at https://github.com/lefterav/Appraise

3.2. 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. There were two different annotation schemes for this task, a shallow and a more fine-grained variation.

- shallow classification:
  classify the two main types of errors (if any) in the best ranked translation output. We use a following subset of the error types: missing content word(s), wrong content word(s), wrong functional word(s), incorrect word form(s), incorrect word order, incorrect punctuation and other error.

- more fine-grained classification:
  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.

Due to the complexity of this task, only two evaluation rounds included error classification. The amount of source sentences used in shallow evaluation, and the number of source sentences together with number of translations produced by different translation systems used for fine-grained error classification are given in Table 2.

3.3. Post-editing

Evaluating machine translation for a professional translation environment means evaluating the requirements of post-editing machine translation output, i.e. editing the output to create a fully fluent and adequate translation which is of the same quality level as a "normal" human translation. We distinguish between two subtasks: "select and post-edit" and "post-edit all". The first subtask was performed in all evaluation rounds, the second one only in one. The subtasks were defined as follows:

Select and post edit: for each source sentence, select the translation output which is easiest to post-edit and perform the editing.

Post-edit all: For each source sentence, post-edit all produced translation outputs.

For the "select and post-edit" subtask, all source sentences from Table 1 were taken into account. However, for the "post-edit all" subtask, similarly to error classification, only a subset of source sentences was taken into account due to complexity of post-editing large amounts of low quality translations. This subtask was partly motivated by the need to compare the performed edit operations in selected sentences with the rest of translation outputs.

(a) shallow error classification task

translation direction	source sentences
German→English	1492
English→German	1655
Spanish→German	1798

(b) fine-grained error classification task

translation direction	source	Moses	Jane	RBMT1	RBMT2
German→English	696	160	196	160	160
English→German	1356	280	589	247	240
German→French	550	152	195	120	83
French→German	689	191	200	170	128
German→Spanish	859	113	535	113	98
Spanish→German	2092	523	523	523	523

Table 2: Number of source sentences and translation outputs of each system in the error classification tasks.

Translators were asked to perform only the minimal post-editing necessary to achieve an acceptable translation quality. An option "Translate from scratch" was available as well and the translators were instructed to use it when they thought that creating a completely new translation was faster than post-editing, e.g. in the case where all translation outputs were of bad quality. This mimics the workflow in a professional translation environment where translators may use fuzzy matches from a translation memory, but also discard them when editing would require too much time and effort. The amount of evaluated sentences available from these data is given in Table 3.

3.4. Quality scoring

This task has been performed only in the last evaluation round. Instead of ranking translation outputs in relation to each other, translators were asked to provide a judgement on the translation quality of each individual output following the instruction:

- classify each translation output into one of three categories: acceptable, easy to correct, not easy to correct

For each source sentence, three translation outputs were scored – one produced by statistical system (Moses), one produced by a rule-based system, and a Trados translation. The amount of evaluated source sentences available for this task is given in Table 4.

4. Ongoing work

Findings of evaluating the results from these rounds have been published in (Burchardt et al., 2011), (Avramidis et al., 2012), (Popović et al., 2013), (Burchardt et al., 2013). Additionally, parts of the corpus described have been used in the QTLaunchPad project, where the human-centric approach to MT evaluation has been further developed. In one line of research, selected corpora have been filtered in order to represent those translations that are either perfect or can easily be fixed by humans, in order to derive a multidimensional quality metric (MQM). On the basis of the resulting error corpus, we studied the error distribution within this translation quality band. In another line of research, the TARAXŰ data has been used to train quality estimation models, having as a goal to automatically predict aspects of MT quality without reference translations or human annotators. This work will be published elsewhere.

5. Summary

We presented a corpus consisting of machine translations annotated and evaluated by. professional human translators. We believe that it will be of value to the machine translation community as professionally annotated data has often been brought forward as central desideratum. After our positive experience with the cooperation between MT research and language professionals in the project, we want to advocate this human-centric approach to MT research and development and we hope that the community will derive valuable knowledge from the corpus we presented.

6. Acknowledgments

Many thanks to Christian Federmann for the technical support with Appraise. The work described in this paper has been partly developed within the TARAXŰ project financed by TSB Technologiestiftung Berlin – Zukunftsfonds Berlin, co-financed by the European Union – European fund for regional development and partly within the QTLaunchPad project, which receives funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 296347.

7. References

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

Eleftherios Avramidis, Aljoscha Burchardt, Christian Federmann, Maja Popović, Cindy Tscherwinka, and David Vilar. 2012. Involving language professionals in the evaluation of machine translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 12)*, pages 1127–1130, Istanbul, Turkey, May.

Translation direction	source	Moses	Jane	RBMT1	RBMT2	Trados
German→English	516	516	512	498	474	467
English→German	467	423	432	394	394	427
German→French	240	238	240	240	229	43
French→German	163	122	122	116	77	117
German→Spanish	113	113	113	92	73	113
Spanish→German	519	517	516	518	519	519

Table 3: Number of source sentences and translation outputs of each system that were processed by professional translators in the "post-edit all" task.

Translation direction	source sentences
German→English	775
English→German	1450
French→German	1450
German→Spanish	775
Spanish→German	1450

Table 4: Number of source sentences used in the quality scoring task.

Aljoscha Burchardt, Christian Federmann, and Hans Uszkoreit. 2011. Hybrid Machine Translation for German in taraXÜ: Can translation costs be decreased without degrading quality? In *Conference of the German Society for Computational Linguistics and Language Technology (GSCL-11)*, Hamburg, Germany, September.

Aljoscha Burchardt, Cindy Tscherwinka, Eleftherios Avramidis, and Hans Uszkoreit, 2013. *Machine Translation at Work*, volume 458 of *Studies in Computational Intelligence*, pages 241–261. Springer, 1.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 Workshop on Statistical Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT 2011)*, pages 22–64, Edinburgh, Scotland, July.

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, Chris Zens, Richard a nd 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.

Maja Popović, Eleftherios Avramidis, Aljoscha Burchardt, Sabine Hunsicker, Sven Schmeier, Cindy Tscherwinka, David Vilar, and Hans Uszkoreit. 2013. Learning from human judgments of machine translation output. In *Proceedings of the MT Summit XIV*, pages 231–238, Nice, France, September.

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