Hybrid Machine Translation for German in taraXÜ:

Can translation costs be decreased without degrading quality?

Aljoscha Burchardt, Christian Federmann, Hans Uszkoreit
DFKI Language Technology Lab
Saarbrücken & Berlin, Germany
E-mail: {burchardt,cfedermann,uszkoreit}@dfki.de

Abstract

A breakthrough in Machine Translation is only possible if human translators are taken into the loop. While mechanisms for automatic evaluation and scoring such as BLEU have enabled fast development of systems, these systems have to be used in practice to get feedback for improvement and fine-tuning. However, it is not clear if and how systems can meet quality requirements in real-world, industrial translation scenarios. taraXÜ paves the way for wide usage of hybrid machine translation for German. In a joint consortium of research and industry partners, taraXÜ integrates human translators into the development process from the very beginning in a post-editing scenario collecting feedback for improvement of its core translation engines and selection mechanism. taraXÜ also performs pioneering work by integrating languages like Czech, Chinese, or Russian, that are not well studied to-date.

Keywords: Hybrid Machine Translation, Human Evaluation, Post-Editing

1. Introduction

Machine Translation (MT) is a prime application of Language Technology. Research on Rule-Based MT (RBMT) goes back the early days of Artificial Intelligence in the 1960s and some systems have reached a high level of sophistication (e.g., Schwall & Thurmair, 1997; Alonso & Thurmair, 2003). Since the mid 1990, Statistical MT (SMT) has become the prevalent paradigm in the research community (e.g. Koehn et al., 2007; Li et al., 2010). In the translation and localization industry, Translation Memory Systems (TMS) are used to support human translators by making informed suggestions for recurrent material that has to be translated.

As human translators can no longer satisfy the constantly raising translation need, important questions that need to be investigated are:

1) How good is MT quality today, especially for translation from and to German?
2) Which paradigm is the most promising one?
3) Can MT aid human translators and can it help to reduce translation costs without sacrificing quality?

These questions are not easy to answer and it is clear that research on the matter is needed. The quality of MT output cannot be objectively assessed in a once-and-for-all measure (see e.g. Callison-Burch et al., 2006) and it also strongly depends on the nature of the input material. Various MT paradigms have different strengths and shortcomings, not only regarding quality. For example, RBMT allows for a good control of the overall translation process, but setting up and maintaining such a system is very costly as it requires trained specialists. SMT is cheap, but it requires huge amounts of compute power and training data, which can make it difficult to include new languages and domains. TMS can produce human quality, but are limited in coverage due to their underlying design. Finally, the question of how human translators can optimally be supported in their translation workflow has largely been untouched.

Machine Translation for German The number of available mono- and bi-lingual resources for German is quite high. In the "EuroMatrix"[1] which collects resources, corpora, and systems for a large number of language pairs, German ranges on the third place behind English and French. Still, only little research has been focused on MT for language pairs including German, especially for translation tasks to and from languages other than English.

[1] http://www.euromatrixplus.net/matrix/

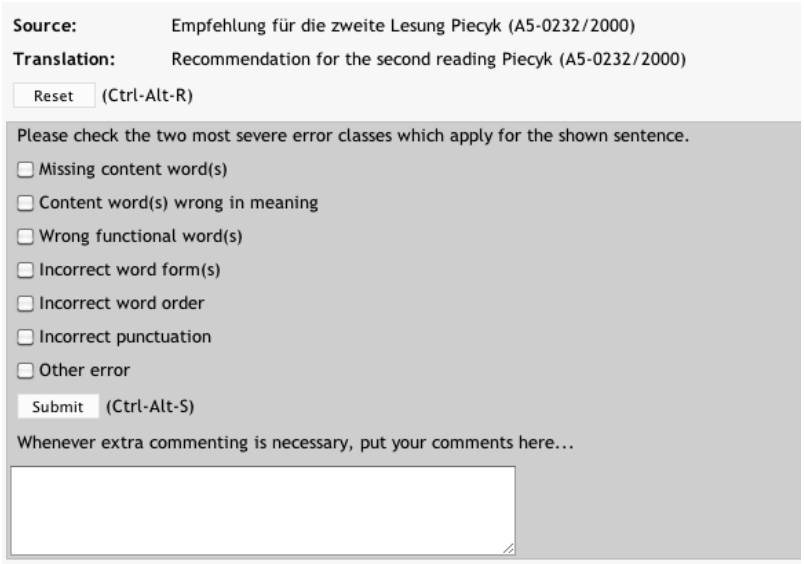

Figure 0: Error classification interface used within taraXÜ .

This paper reports on taraXÜ[2], which aims to address the aforementioned questions in a consortium consisting of partners from both research and industry. taraXÜ takes the selection from hybrid MT results including RBMT, TMS, and SMT as the first part of its analytic process. Then a self-calibration[3] component applies, extended by controlled language technology and human post-processing to match real-world translation concerns. A novelty in this project is that human translators are integrated into the development process from the very beginning: Within several human evaluation rounds, the automatic selection and calibration mechanisms will be refined and iteratively improved. This paper focuses on hybrid translation (Section 2) and the large-scale human evaluation rounds in taraXÜ (Section 3). In the conclusion and outlook (Section 4), ongoing and future research is sketched.

2. Hybrid Machine Translation

Hybrid MT is a recent trend (e.g. Federmann et al., 2009; Chen et al., 2009) for leveraging the quality of MT. Based on the observation that different MT systems often have complementary strengths and weaknesses, different methods for hybridization are investigated that aim to "fuse" an improved translation out of the good parts of several translation candidates.

Complementary Errors Typical difficulties for SMT are morphology, sentence structure, long-range re-ordering, and missing words, while strengths are disambiguation and lexical choice.

RBMT systems are typically strong in morphology, sentence structure, have the ability to handle long-range phenomena, and also ensure completeness of the resulting translation. Weaknesses arise from parsing errors and wrong lexical choice. The following examples illustrate the complementary nature of such systems' errors.

1) **Source**: Then, in the afternoon, the visit will culminate in a grand ceremony, at which Obama will receive the prestigious award.
2) **RBMT**[4]: Dann wird der Besuch am Nachmittag in einer großartigen Zeremonie gipfeln, an der Obama die berühmte Belohnung bekommen wird.
3) **SMT**[5]: Dann am Nachmittag des Besuchs in eine beeindruckende Zeremonie mu ndet , wo Obama den angesehenen Preis erhalten werden.

As you can see in the translation of Example 1), the RBMT system generated a complete sentence, yet with a wrong lexical choice for award. The SMT system on the other hand generated the right reading, but made morphological errors and did not generate a complete German sentence. In the translation of Example 4), a parsing error in the analysis phase of the RBMT system led to an almost unreadable result while the SMT decoder

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

[3] Due to limited space, this won't be discussed herein.

[4] System used: Lucy MT (Alonso and Thurmair, 2003)

[5] System used: phrase-based Moses (Koehn et al., 2007)

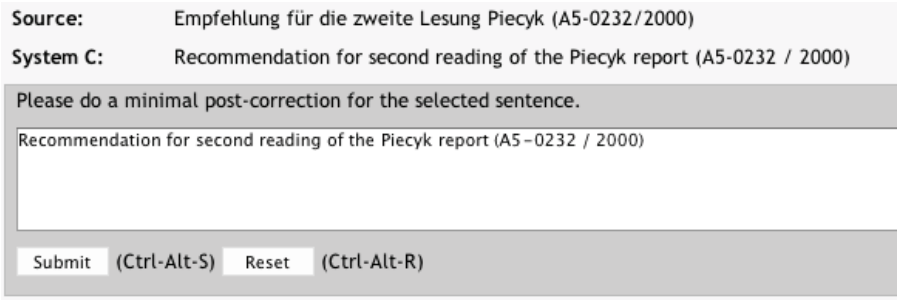

Figure 0: Post-editing interface used within taraXÜ.

generated a generally intelligible translation, yet with stylistic and formal deficits.

4) **Source**: Right after hearing about it, he described it as a "challenge to take action."
5) **RBMT**: Nachdem er richtig davon gehört hatte, bezeichnete er es als eine "Herausforderung, um Aktion auszuführen."
6) **SMT**: Gleich nach Anhörung darüber, beschrieb er es als eine "Herausforderung, Maßnahmen zu ergreifen."

Hybrid combination can hence lead to better overall translations.

A Human-centric Hybrid Approach In contrast to other hybrid approaches; taraXÜ is in the first place designed to support human post-editing, e.g., in a translation agency. Two different modes have to be handled by the project's selection mechanism:

- **Human post-editing**: Select the sentence that is easiest to post-edit and have the user edit it.
- **Standalone MT**: Select the overall best translation and present it to the user.

For the translation of 4), the best selection in Standalone MT mode would probably be 6), which is a useful translation, e.g., for information gisting. In Human post-editing mode, 5) would be a better selection as it can relatively quickly be transformed into 7), which is a human-quality translation.

7) **Human edit of 5)**: Gleich, nachdem er davon gehört hatte, bezeichnete er es als eine "Herausforderung, zu handeln."

One goal of taraXÜ is the design and implementation of such a novel selection mechanism; however this is still work in progress and will be described elsewhere. Apart from properties of the source sentence (domain, complexity, etc.) and the different translations (grammatical correctness, sentence length, etc.), the selection mechanism will also take into account "metadata" of the various systems involved such as runtime, number of out-of-vocabulary-warnings, number of different readings generated, etc.

One industry partner in the project consortium provides modules for language checking that will not only be used in the selection mechanism, but also in pre-processing of the input. Starting from the observation that many translation problems arise from problematic input, another goal of taraXÜ is to develop automatic methods for pre-processing input before it is sent to MT translation engines.

3. Large-Scale Human Evaluation

Several large-Scale human evaluation rounds are foreseen within the duration of taraXÜ, mainly for the calibration of both the selection mechanism as well as the pre-editing steps, but also for measuring the time needed for post-editing, and for getting a detailed error classification on the translation output from the various MT systems under investigation. The evaluation rounds are performed by external Language Service Providers that usually offer human translation services and hence are considered to act as non-biased experts.

Evaluation Procedure The language pairs that will be implemented and tested during the runtime of taraXÜ are listed in Table 1.

German	⇔	English French Japanese Russian Spanish
English	⇔	Chinese Czech

Table 1: Language pairs treated in taraXÜ.

We use an extended version of the browser-based evaluation tool Appraise (Federmann, 2010) to collect human judgments on the translation quality of the various systems under investigation in taraXÜ. A screen-shot of the error classification interface can be seen in Figure 1, the post-editing view is presented in Figure 2.

Pilot Evaluation Round The first (pilot) evaluation round of taraXÜ includes the language pairs EN→DE, DE→EN, and ES→DE. The corpus size per language pair is about 2,000 sentences, the data taken mainly from previous WMT shared tasks, but also extracted from freely available technical documentation. Two evaluation tasks will be performed by the human annotators, mirroring the two modi of our selection mechanism:

1) In the first task, the annotators have to rank the output of four different MT systems depending on their translation quality. In a subsequent step, they are asked to classify the two main types of errors (if any) of the chosen best translation. We use a subset of the error types suggested by (Vilar et al., 2006), as shown in Figure 1.
2) The second task for the human annotators in the first evaluation round is selecting the translation that is easiest to post-edit and to perform the editing. Only a minimal post-editing should be performed.

Some very first results of the ongoing examination of the first human evaluation round are shown in Table 2. The top of the table shows the over-all ranking among the four listed systems, bold face indicates the best system. Below are the results for translation from Spanish and English into German, respectively. On the bottom of the table, overall results on selected corpora are shown from the news domain (1,030 sentences from the WMT-2010 news test set of Callison-Burch et al. (2010), sub-sampled proportionally to each one of its documents) and from the technical documentation of the OpenOffice project.

One observation is that the systems' ranks are comparably close except for Trados, which is not a proper MT system. The very good result of Trados on the news corpora requires further investigation. 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 will also require some more detailed inspection. The latter might, however, explain the good perfomance of the web-based system on the OpenOffice corpus.

	Lucy	Moses	Trados	Google
Overall	2.00	2.38	3.74	1.86
DE-EN	2.01	2.46	3.80	1.73
ES-DE	1.85	2.42	3.72	1.99
EN-DE	2.12	2.28	3.71	1.89
WMT10	2.52	2.59	2.21	2.69
OpenOffice	1.72	2.77	3.95	1.56

Table 2: First human ranking results, as the average rank of each system in each task.

4. Conclusions and Outlook

In this paper, we have argued and shown evidence that a human-centric hybrid approach to Machine Translation is a promising way of integrating this technology into industrial translation workflows. Even in this early stage, taraXÜ has generated positive feedback and raised interest, especially on the side of the industry partners. We reported early results from the first (pilot) evaluation of taraXÜ, including language pairs EN→DE, DE→EN, and ES→DE. After analyzing the results of this pilot, further evaluation rounds will iteratively extend the numbers of languages covered and include questions related to topics such as controlled language, error types, and the effect of different subject domains. In the presentation of this paper, we will include a more detailed discussion of the first evaluation results.

5. Acknowledgements

This work has partly been developed within the taraXÜ project financed by TSB Technologiestiftung Berlin – Zukunftsfonds Berlin, co-financed by the European Union – European fund for regional development. This work was also supported by the EuroMatrixPlus project (IST-231720) that is funded by the European Community under the Seventh Framework Programme for Research and Technological Development.

6. References

Alonso, J. A. and Thurmair, G. (2003). The comprendium translator system. In Proceedings of the Ninth Machine Translation Summit.

Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., and Zaidan, O. (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. Association for Computational Linguistics. Revised August 2010.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the role of bleu in machine translation research. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, pages 249–256.

Chen, Y., Jellinghaus, M., Eisele, A., Zhang, Y., Hunsicker, S., Theison, S., Federmann, C., and Uszkoreit, H. (2009). Combining multi-engine translations with Moses. In Proceedings of the Fourth Workshop on Statistical Machine Translation, pages 42–46, Athens, Greece. Association for Computational Linguistics.

Federmann, C. (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. European Language Resources Association (ELRA).

Federmann, C., Theison, S., Eisele, A., Uszkoreit, H., Chen, Y., Jellinghaus, M., and Hunsicker, S. (2009). Translation combination using factored word substitution. In Proceedings of the Fourth Workshop on Statistical Machine Translation, pages 70–74, Athens, Greece. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C. J., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Li, Z., Callison-Burch, C., Dyer, C., Ganitkevitch, J., Irvine, A., Khudanpur, S., Schwartz, L., Thornton, W., Wang, Z., Weese, J., and Zaidan, O. (2010). Joshua 2.0: A toolkit for parsing-based machine translation with syntax, semir-ings, discriminative training and other good-ies. In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR, pages 133–137, Uppsala, Sweden. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). Bleu: a method for automatic evaluation of machine translation. IBM Research Report RC22176(W0109-022), IBM.

Schwall, U. and Thurmair, G. (1997). From metal to t1: systems and components for machine translation applications. In Proceedings of the Sixth Machine Translation Summit, pages 180– 190.

Vilar, D., Xu, J., D'Haro, L. F., and Ney, H. (2006). Error Analysis of Machine Translation Output. In International Conference on Language Resources and Evaluation, pages 697– 702, Genoa, Italy.