Using a New Analytic Measure for the Annotation and Analysis of MT Errors on Real Data

Arle Lommel[1], **Aljoscha Burchardt**[1], **Maja Popović**[1],
Kim Harris[2], **Eleftherios Avramidis**[1], **Hans Uszkoreit**[1]
[1]DFKI / Berlin, Germany
[2] text&form / Berlin, Germany
name.surname@dfki.de[1]
kim_harris@textform.com[2]

Abstract

This work presents the new flexible Multidimensional Quality Metrics (MQM) framework and uses it to analyze the performance of state-of-the-art machine translation systems, focusing on "nearly acceptable" translated sentences. A selection of WMT news data and "customer" data provided by language service providers (LSPs) in four language pairs was annotated using MQM issue types and examined in terms of the types of errors found in it.

Despite criticisms of WMT data by the LSPs, an examination of the resulting errors and patterns for both types of data shows that they are strikingly consistent, with more variation between language pairs and system types than between text types. These results validate the use of WMT data in an analytic approach to assessing quality and show that analytic approaches represent a useful addition to more traditional assessment methodologies such as BLEU or METEOR.

1 Introduction

For a number of years, the Machine Translation (MT) community has used "black-box" measures of translation performance like BLEU (Papineni et al., 2002) or METEOR (Denkowski and Lavie, 2011). These methods have a number of advantages in that they can provide automatic scores for MT output in cases where there are existing reference translations by calculating similarity between the MT output and the references. However, such metrics do not provide insight into the specific nature of problems encountered in the translation output and scores are tied to the particularities of the reference translations.

As a result of these limitations, there has been a recent shift towards the use of more explicit error classification and analysis (see, e.g., Vilar et al. (2006)) in addition to automatic metrics. The error profiles used, however are typically ad hoc categorizations and specific to individual MT research projects, thus limiting their general usability for research or comparability with human translation (HT) results. In this paper, we will report on annotation experiments that use a new, flexible error metric and that showcase a new type of MT research involving collaboration between MT researchers, human translators, and Language Service Providers (LSPs).

When we started to prepare our annotation experiments, we teamed up with LSPs and designed a custom error metric based on the "Multidimensional Quality Metric" MQM designed by the QTLaunchPad project (http://www.qt21.eu/launchpad). The metric was designed to facilitate annotation of MT output by human translators while containing analytic error classes we considered relevant to MT research (see **Section 2**, below). This paper represents the first publication of results from use of MQM for MT quality analysis.

Previous research in this area has used error categories to describe error types. For instance, Farrús et al. (2010) divide errors into five broad classes (orthographic, morphological, lexical, semantic, and syntactic). By contrast, Flana-

gan (1994) uses 18 more fine-grained error categories with additional language-pair specific features, while Stymne and Ahrenberg (2012) use ten error types of somewhat more intermediate granularity (and specifically addresses combinations of multiple error types). All of these categorization schemes are ad hoc creations that serve a particular analytic goal. MQM, however, provides a *general* mechanism for describing a family of related metrics that share a common vocabulary. This metric was based upon a rigorous examination of major human and machine translation assessment metrics (e.g., LISA QA Model, SAE J2450, TAUS DQF, ATA assessment, and various tool-specific metrics) that served as the basis for a descriptive framework for declaring what a particular metric addresses. While the metric described in this paper is still very much a purpose-driven metric, it is declared in this general framework, which we propose for use to declare specific metrics for general quality assessment and error annotation tasks.

For data, we chose WMT data (Bojar et al., 2013) to represent the state of the art output for MT in research. However, LSPs frequently reported to us that the mostly journalistic WMT data does not represent their business data (mostly technical documentation) or typical applications of MT in business situations. In addition, it turned out that journalistic style often contains literary flourishes, idiosyncratic or mixed styles, and deep embedding (e.g., nested quotations) that sometimes make it very difficult to judge the output.

As a result, we decided to use both WMT data and customer MT data that LSPs contributed from their daily business to see if the text types generate different error profiles. This paper accordingly presents and compares the results we obtained for both types of sources. For practical purposes, we decided to analyze only "near miss" translations, translations which require only a small effort to be converted into acceptable translations. We excluded "perfect" translations and those translations that human evaluators judged to have too many errors to be fixed easily (because these would be too difficult to annotate). We therefore had human evaluators select segments representing this especially business-relevant class of translations prior to annotation.

A total of nine LSPs participated in this task, with each LSP analyzing from one to three language pairs. Participating LSPs were paid up to €1000 per language pair. The following LSPs participated: Beo, Hermes, iDisc, Linguaserve, Logrus, Lucy, Rheinschrift, text&form, and Welocalize.

2 Error classification scheme

The Multidimensional Quality Framework (MQM) system[1] provides a flexible system for declaring translation quality assessment methods, with a focus on analytic quality, i.e., quality assessment that focuses on identifying specific issues/errors in the translated text and categorizing them.[2] MQM defines over 80 issue/error types (the expectation is that any one assessment task will use only a fraction of these), and for this task, we chose a subset of these issues, as defined below.

- **Accuracy**. Issues related to whether the information content of the target is equivalent to the source.
  - **Terminology**. Issues related to the use of domain-specific terms.
  - **Mistranslation**. Issues related to the improper translation of content.
  - **Omission**. Content present in the source is missing in the target.
  - **Addition**. Content not present in the source has been added to the target.
  - **Untranslated**. Text inappropriately appears in the source language.
- **Fluency**. Issues related to the linguistic properties of the target without relation to its status as a translation.
  - **Grammar**. Issues related to the grammatical properties of the text.
    - **Morphology (word form)**. The text uses improper word forms.
    - **Part of speech**. The text uses the wrong part of speech

[1] http://www.qt21.eu/mqm-definition/

[2] This approach stands in contrast to "holistic" methods that look at the text in its entirety and provide a score for the *as a whole* in terms of one or more dimensions, such as overall readability, usefulness, style, or accuracy. BLEU, METEOR, and similar automatic MT evaluation metrics used for research can be considered holistic metrics that evaluate texts on the dimension of similarity to reference translations since they do not identify specific, concrete issues in the translation. In addition, most of the options in the TAUS Dynamic Quality Framework (DQF) (https://evaluation.taus.net/about) are holistic measures.

        - **Agreement**. Items within the text do not agree for number, person, or gender.
        - **Word order**. Words appear in the incorrect order.
        - **Function words**. The text uses function words (such as articles, prepositions, "helper"/auxiliary verbs, or particles) incorrectly.
    - **Style**. The text shows stylistic problems.
    - **Spelling**. The text is spelled incorrectly
        - **Capitalization**. Words are capitalized that should not be or vice versa.
    - **Typography**. Problems related to typographical conventions.
        - **Punctuation**. Problems related to the use of punctuation.
    - **Unintelligible**. Text is garbled or otherwise unintelligible. Indicates a major breakdown in fluency.

Note that these items exist in a hierarchy. Annotators were asked to choose the most specific issue possible and to use higher-level categories only when it was not possible to use one deeper in the hierarchy. For example, if an issue could be categorized as *Word order* it could also be categorized as *Grammar*, but annotators were instructed to use *Word order* as it was more specific. Higher-level categories were to be used for cases where more specific ones did not apply (e.g., the sentence *He slept the baby* features a "valency" error, which is not a specific type in this hierarchy, so *Grammar* would be chosen instead).

3 Corpora

The corpus contains Spanish→English, German→English, English→Spanish, and English→German translations. To prepare the corpus, for each translation direction a set of translations were evaluated by expert human evaluators (primarily professional translators) and assigned to one of three classes:

1. **perfect (class 1)**. no apparent errors.
2. **almost perfect or "near miss" (class 2)**. easy to correct, containing up to three errors.
3. **bad (class 3)**. more than three errors.

Both WMT and "customer" data[3] were rated in this manner and pseudo-random selections (selections were constrained to prevent annotation of multiple translations for the same source segment within a given data set in order to maximize the diversity of content from the data sources) taken from the class 2 sentences, as follows:

- **Calibration set**. For each language pair we selected a set of 150 "near miss" (see below) translations from WMT 2013 data (Bojar et al., 2013).
    - For English → German and English → Spanish, we selected 40 sentences from the top-ranked SMT, RbMT, and hybrid systems, plus 30 of the human-generated reference translations.
    - For German → English and Spanish → English, we selected 60 sentences from the top-ranked SMT and RbMT systems (no hybrid systems were available for those language pairs), plus 30 of the human-generated reference translations.
- **Customer data**. Each annotator was provided with 200 segments of "customer" data, i.e., data taken from real production systems.[4] This data was translated by a variety of systems, generally SMT (some of the German data was translated using an RbMT system).
- **Additional WMT data**. Each annotator was also asked to annotate 100 segments of previously unannotated WMT data. In some cases the source segments for this selection overlapped with those of the calibration set, although the specific MT outputs chosen did not (e.g., if the SMT output for a given segment appeared in the calibration set, it would not reappear in this set, although the RbMT, hybrid, or human translation might). Note that the additional WMT data provided was different for each LSP in order to maximize coverage of annotations in line with other research goals; as such, this additional data does not factor into inter-annotator agreement calculations (discussed below).

[3] WMT data was from the top-rated statistical, rule-based, and hybrid systems for 2013; customer data was taken from a variety of in-house systems (both statistical and rule-based) used in production environments.

[4] In all but one case the data was taken from actual projects; in the one exception the LSP was unable to obtain permission to use project data and instead took text from a project that would normally not have been translated via MT and ran it through a domain-trained system.

It should be noted that in all cases we selected only translations for which the source was originally authored in the source language. The WMT shared task used human translations of some segments as source for MT input: for example, a sentence authored in Czech might be translated into English by humans and then used as the source for a translation task into Spanish, a practice known as "relay" or "pivot" translation. As we wished to eliminate any variables introduced by this practice, we eliminated any data translated in this fashion from our task and instead focused only on those with "native" sources.

3.1 Annotation

The annotators were provided the data described above and given access to the open-source translate5[5] annotation environment. Translate5 provides the ability to mark arbitrary spans in segments with issue types and to make other annotations. All annotators were invited to attend an online training session or to view a recording of it and were given written annotation guidelines. They were also encouraged to submit questions concerning the annotation task.

The number of annotators varied for individual segments, depending on whether they were included in the calibration sets or not. The numbers of annotators varied by segment and language pair:

- **German→English**: Calibration: 3; Customer + additional WMT: 1
- **English→German**: Calibration: 5; Customer + additional WMT: 1–3
- **Spanish→English**: Calibration: 4; Customer + additional WMT: 2–4
- **English→Spanish**: Calibration: 4; Customer + additional WMT: 1–3

After annotation was complete some post-processing steps simplified the markup and extracted the issue types found by the annotators to permit comparison.

3.2 Notes on the data

The annotators commented on a number of aspects of the data presented to them. In particular, they noted some issues with the WMT data. WMT is widely used in MT evaluation tasks, and so enjoys some status as the universal data set for tasks such as the one described in this paper. The available translations represent the absolute latest and most state-of-the-art systems available in the industry and are well established in the MT research community.

However, feedback from our evaluators indicated that WMT data has some drawbacks that must be considered when using it. Specifically, the text type (news data) is rather different from the sorts of technical text typically translated in production MT environments. News does not represent a coherent domain (it is, instead, a genre), but rather has more in common with general language. In addition, an examination of the human-generated reference segments revealed that the human translations often exhibited a good deal of "artistry" in their response to difficult passages, opting for fairly "loose" translations that preserved the broad sense, but not the precise details.

The customer data used in this task does not all come from a single domain. Much of the data came from the automotive and IT (software UI) domains, but tourism and financial data were also included. Because we relied on the systems available to LSPs (and provided data in a few cases where they were not able to gain permission to use customer data), we were not able to compare different types of systems in the customer data and instead have grouped all results together.

An additional factor is that the sentences in the calibration sets were much longer (19.4 words, with a mode of 14, a median of 17, and a range of 3 to 77 words) than the customer data (average 14.1 words, with a mode of 11, a median of 13, and a range of 1 to 50 words). We believe that the difference in length may account for some difference between the calibration and customer sets described below.

4 Error analysis

In examining the aggregate results for all language pairs and translation methods, we found that four of the 21 error types constitute the majority (59%) of all issues found:

- Mistranslation: 21%
- Function words: 15%
- Word order: 12%
- Terminology: 11%

None of the remaining issues comprise more than 10% of annotations and some were found so

[5] http://www.translate5.net

infrequently as to offer little insight. We also found that some of the hierarchical distinctions were of little benefit, which led us to revise the list of issues for future research (see **Section 4.2** for more details).

4.1 Inter-Annotator Agreement

Because we had multiple annotators for most of the data, we were able to assess inter-annotator agreement (IAA) for the MQM annotation of the calibration sets. IAA was calculated using Cohen's kappa coefficient. At the word level (i.e., seeing if annotators agreed for each word, we found that the results lie between 0.2 and 0.4 (considered "fair"), with an average of pairwise comparisons of 0.29 (de-en), 0.25 (es-en), 0.32 (en-de), and 0.34 (en-es), with an overall average of 0.30

4.2 Modifications

This section addresses some of the lessons learned from an examination of the MQM annotations described in **Section 4.1**, with a special emphasis on ways to improve inter-annotator agreement (IAA). Although IAA does not appear to be a barrier to the present analytic task, we found a number of areas where the annotation could be improved and superfluous distinctions eliminated. For example, "plain" *Typography* appeared so few times that it offered no value separate from its daughter category *Punctuation*. Other categories appeared to be easily confusible, despite the instructions given to the annotators (e.g., the distinction between "Terminology" and "Mistranslation" seemed to be driven largely by the length of the annotated issue: the average length of spans tagged for "Mistranslation" was 2.13 words (with a standard deviation of 2.43), versus 1.42 (with a standard deviation of 0.82) for "Terminology'.' (Although we had expected the two categories to exhibit a difference in the lengths of spans to which they were applied, a close examination showed that the distinctions were not systematic with respect to whether actual terms were marked or not, indicating that the two categories were likely not clear or relevant to the annotators. In addition, "Terminology" as a category is problematic with respect to the general-domain texts in the WMT data sets since no terminology resources are provided.)

Based on these issues, we have undertaken the following actions to improve the consistency of future annotations and to simplify analysis of the present data.

- The distinction between *Mistranslation* and *Terminology* was eliminated. (For calculation purposes *Terminology* became a daughter of *Mistranslation*.)
- The *Style/Register* category was eliminated since stylistic and register expectations were unclear and simply counted as general *Fluency* for calculation purposes.
- The *Morphology (word form)* category was renamed *Word form* and *Part of Speech*, *Agreement*, and *Tense/mood/aspect* were moved to become its children.
- *Punctuation* was removed, leaving only *Typography,* and all issues contained in either category were counted as *Typography*
- *Capitalization*, which was infrequently encountered, was merged into its parent *Spelling*.

In addition, to address a systematic problem with the *Function words* category, we added additional custom children to this category: *Extraneous* (for function words that should not appear), *Missing* (for function words that are missing from the translation), and *Incorrect* (for cases in which the incorrect function word is used). These were added to provide better insight into the specific problems and to address a tendency for annotators to categorize problems with function words as Accuracy issues when the function words were either missing or added. This revised issue type hierarchy is shown in Figure 1.

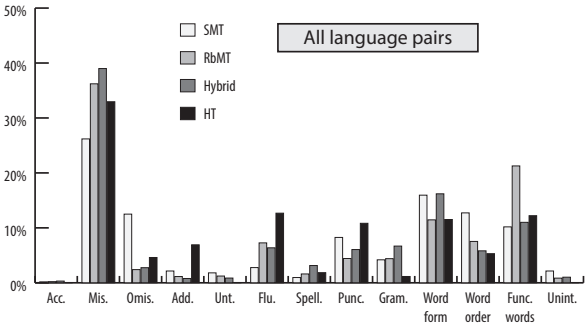

Figure 3: Average Sentence-level error rates [%] for all language pairs.

This revised hierarchy will be used for ongoing annotation in our research tasks. We also realized that the guidelines to annotators did not provide sufficient decision-making tools to help them select the intended issues. To address this problem

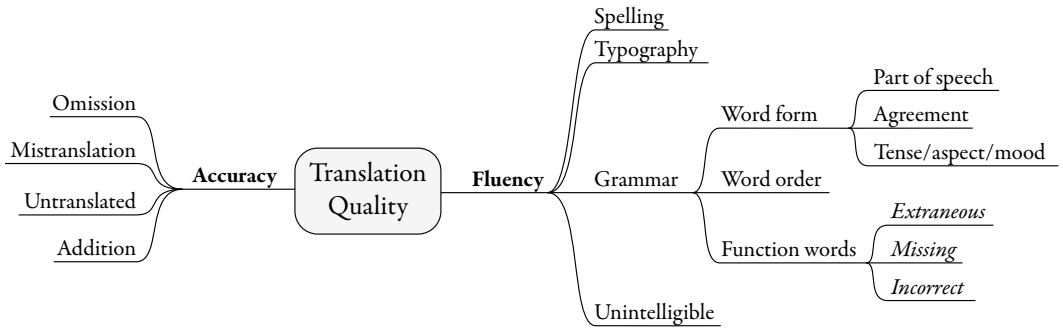

Figure 1: Revised issue-type hierarchy.

we created a decision-tree to guide their annotations. We did not recalculate IAA from the present data set with the change in categories since we have also changed the guidelines and both changes will together impact IAA. We are currently running additional annotation tasks using the updated error types that will result in new scores.

Refactoring the existing annotations according to the above description, gives the results for each translation direction and translation method in the calibration sets, as presented in Figure 2 (with averages across all language pairs as presented in Figure 3). Figure 4 presents the same results for each language pair in the customer data. As previously mentioned, we were not able to break out results for the customer data by system type.

4.3 Differences between MT methods

Despite considerable variation between language pairs, an examination of the annotation revealed a number of differences in the output of different system types. While many of the differences are not unexpected, the detailed analytic approach taken in this experiment has enabled us to provide greater insight into the precise differences rather than relying on isolated examples. The overall results for all language pairs are presented in Figure 3 (which includes the results for the human translated segments as a point of comparison).

The main observations for each translation method include:

- **statistical machine translation**
  - Performs the best in terms of *Mistranslation*
  - Most likely to drop content (*Omission*); otherwise it would be the most accurate translation method considered.
  - Had the lowest number of *Function Words* errors, indicating that SMT gets this aspect substantially better than alternative systems.
  - Weak in *Grammar*, largely due to significant problems in *Word Order*
- **rule-based machine translation**
  - Generated the worst results for *Mistranslation*
  - Was least likely to omit content (*Omission*)
  - Was weak for *Function Words*; statistical enhancements (moving in the direction of hybrid systems) would offer considerable potential for improvement
- **hybrid machine translation** (available only for English→Spanish and English→German)
  - Tends to perform in between SMT and RBMT in most respects
  - Most likely method to produce mistranslated texts (*Mistranslation*)

When compared to the results of human translation assessment, it is apparent that all of the near-miss machine translations are somewhat more accurate than near-miss human translation and significantly less grammatical. Humans are far more likely to make typographic errors, but otherwise are much more fluent. Note as well that humans are more likely to add information to translations than MT systems, perhaps in an effort to render texts more accessible. Thus, despite substantial differences, all of the MT systems are overall more similar to each other than they are to human translation. However, when one considers that a far greater proportion of human translation sentences were in the "perfect" category and a far lower proportion in the "bad" category, and that these comparisons focus only on the "near miss sentences," it

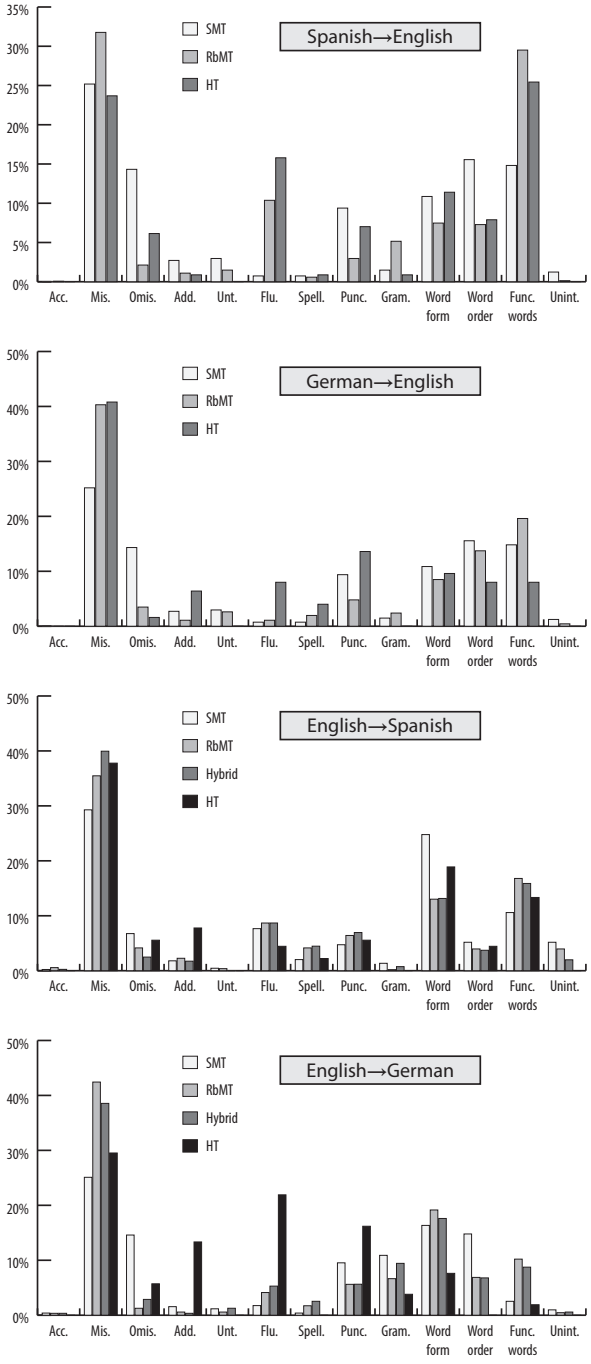

Figure 2: Sentence-level error rates [%] for each translation direction and each translation method for WMT data.

is apparent that outside of the context of this comparison, human translation still maintains a much higher level of Accuracy and Fluency.

In addition, a number of the annotators commented on the poor level of translation evident in the WMT human translations. Despite being professional translations, there were numerous instances of basic mistakes and interpretive translations that resulted in translations that would generally be considered poor references for MT evaluation (since MT cannot make interpretive translations). However, at least in part, these problems with translation may be attributed to the uncontrolled nature of the source texts, which tended to be more literary than is typical for industry uses of MT. In may cases the WMT source sentences presented translation difficulties for the human translators and the meaning of the source texts was not always clear out of context. As a result the WMT texts provide difficulties for both human and machine translators.

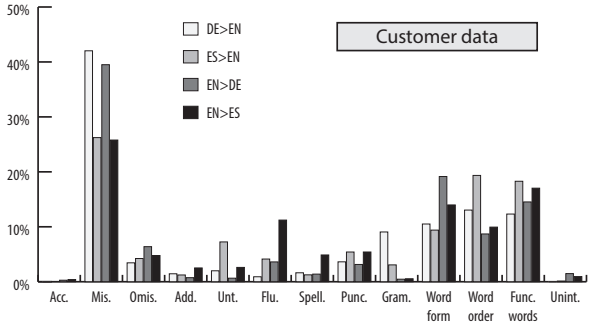

Figure 4: Sentence-level error rates [%] for each translation direction for customer data.

4.4 Comparison of WMT and customer data

By contrast, the customer data was more likely to consist of fragments (such as *Drive vibrates* or section headings) or split segments (i.e., one logical sentence was split with a carriage return, resulting in two fragments) that caused confusion for the MT systems. It also, in principle, should have had advantages over the WMT data because it was translated with domain-trained systems.

Despite these differences, however, the average profiles for all calibration data and all customer data across language pairs look startlingly similar, as seen in Figure 5. There is thus significantly more variation between language pairs and between system types than there is between the WMT data and customer data in terms of the error profiles. (Note, however, that this comparison addresses only the "near-miss" translations and cannot address profiles outside of this category; it also does not address the overall relative distribution into the different quality bands for the text types.)

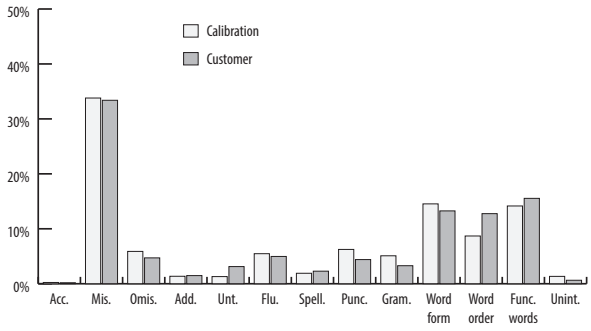

Figure 5: Sentence-level error rates [%] for calibration vs. customer data (average of all systems and language pairs).

5 Conclusions and outlook

The experiment here shows that analytic quality analysis can provide a valuable adjunct to automatic methods like BLEU and METEOR. While more labor-intensive to conduct, they provide insight into the causes of errors and suggest possible solutions. Our research treats the human annotation as the first phase in a two-step approach. In the first step, described in this paper, we use MQM-based human annotation to provide detailed description of the symptoms of MT failure. This annotation also enables us to detect the system type- and language-specific distribution of errors and to understand their relative importance.

In the second step, which is ongoing, linguists and MT experts will use the annotations from the first step to gain insight into the causes for MT failures on the source side or into MT system limitations. For example, our preliminary research into English source-language phenomena indicates that *-ing* verbal forms, certain types of embedding in English (such as relative sentences or quotations), and non-genitive uses of the preposition *of* are particularly contributory to MT failures. Further research into MQM human annotation will undoubtedly reveal additional source factors that can guide MT development or suggest solutions to systematic linguistic problems. Although many of these issues are known to be difficult, it is only with the identification of concrete examples that they can be addressed.

In this paper we have shown that the symptoms of MT failure are the same between WMT and customer data, but it is an open question as to whether the causes will prove to be the same. We therefore advocate for a continuing engagement with language service providers and translators using these different types of data. These approaches will help further the acceptance of MT in commercial settings by allowing them to be compared to HT output and will also help research to go forward in a more principled and requirements-driven fashion.

Acknowledgments

This work has been supported by the QTLaunchPad (EU FP7 CSA No. 296347) project.

References

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013). Findings of the 2013 workshop on statistical machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44.

Denkowski, M. and Lavie, A. (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*.

Farrús, M., Costa-jussà, M. R., Mariño, J. B., and Fonollosa, J. A. (2010). Linguistic-based evaluation criteria to identify statistical machine translation errors. In *Proceedings of the EAMT 2010*.

Flanagan, M. (1994). Error classification for MT evaluation. In *Proceedings of AMTA*, pages 65–72.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Stymne, S. and Ahrenberg, L. (2012). On the practice of error analysis of machine translation evaluation. In *Proceedings of LREC 2012*, pages 1785–1790.

Vilar, D., Xu, J., D'Haro, L. F., and Ney, H. (2006). Error analysis of statistical machine translation output. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 697–702.