Automatic and Manual Metrics for Operational Translation Evaluation

Workshop Programme

08:45 – 09:30 Welcome and Introduction by Workshop Organizers

09:30 – 10:30 Talks Session 1

Joke Daems, Lieve Macken and Sonia Vandepitte, *Two Sides of the Same Coin: Assessing Translation Quality in Two Steps Through Adequacy and Acceptability Error Analysis*

Leonid Glazychev, *How to Reliably Measure Something That's Not Completely Objective: A Clear, Working and Universal Approach to Measuring Language Quality*

Mihaela Vela, Anne-Kathrin Schumann and Andrea Wurm, *Translation Evaluation and Coverage by Automatic Scores*

Arle Lommel, Maja Popović and Aljoscha Burchardt, *Assessing Inter-Annotator Agreement for Translation Error Annotation*

10:30 – 11:00 Coffee break

11:00 – 13:00 Talks Session 2

Marianne Starlander, *TURKOISE: A Mechanical Turk-based Tailor-made Metric for Spoken Language Translation Systems in the Medical Domain*

Caitlin Christianson, Bonnie Dorr and Joseph Olive, *MADCAT Evaluation Approach: Operational Accuracy of MT applied to OCR*

Ekaterina Stambolieva, *Continuous Operational Evaluation of Evolving Proprietary MT Solution's Translation Adequacy*

Lars Ahrenberg, *Chunk Accuracy: A Simple, Flexible Metric for Translation Quality*

Michael Carl and Moritz Schaeffer, *Word Transition Entropy as an Indicator for Expected Machine Translation Quality*

Douglas Jones, Paul Gatewood, Martha Herzog and Tamas Marius, *A New Multiple Choice Comprehension Test for MT* and *Standardized ILR-Based* and *Task-Based Speech-to-Speech MT Evaluation*

Lena Marg, *Rating Evaluation Methods through Correlation*

Federico Gaspari, Antonio Toral, Arle Lommel, Stephen Doherty, Josef van Genabith and Andy Way, *Relating Translation Quality Barriers to Source-Text Properties*

13:00 – 14:00 Lunch break

14:00 – 15:00 Hands-On Session 1

15:00 – 16:00 Hands-On Session 2

16:00 – 16:30 Coffee break

16:30 – 17:30 Hands-On Session 3

17:30 – 18:00 Discussion, Potential for Future Collaboration, Next Steps, and Conclusion

Organizing Committee and Editors

Keith J. Miller	The MITRE Corporation
Lucia Specia	University of Sheffield
Kim Harris	text&form GmbH, Germany
Stacey Bailey	The MITRE Corporation

Table of Contents

Extended Abstracts 1

Two Sides of the Same Coin: Assessing Translation Quality in Two Steps through Adequacy and Acceptability Error Analysis 2

How to Reliably Measure Something That's Not Completely Objective: A Clear, Working and Universal Approach to Measuring Language Quality 3

Human Translation Evaluation and its Coverage by Automatic Scores 4

Assessing Inter-Annotator Agreement for Translation Error Annotation 5

TURKOISE: A Mechanical Turk-based Tailor-made Metric for Spoken Language Translation Systems in the Medical Domain 6

MADCAT Evaluation: Operational Accuracy of MT Applied to OCR 7

Continuous Operational Evaluation of Evolving Proprietary MT Solution's Translation Adequacy 8

Chunk Accuracy: A Simple, Flexible Metric for Translation Quality 10

Word Transition Entropy as an Indicator for Expected Machine Translation Quality 11

A New Multiple Choice Comprehension Test for MT 13

Standardized ILR-Based and Task-Based Speech-to-Speech MT Evaluation 14

Rating Evaluation Methods through Correlation 15

Relating Translation Quality Barriers to Source-Text Properties 16

Understanding Stakeholder Requirements for Determining Translation Quality 17

Automated and Task-Based Evaluation of the Effects of Machine Translation Domain Tuning on MT Quality, Post-editing, and Human Translation 18

Full Papers 19

Human Translation Evaluation and its Coverage by Automatic Scores 20

Assessing Inter-Annotator Agreement for Translation Error Annotation 31

TURKOISE: a Mechanical Turk-based Tailor-made Metric for Spoken Language Translation Systems in the Medical Domain 38

Word Transition Entropy as an Indicator for Expected Machine Translation Quality 45

Standardized ILR-Based and Task-Based Speech-to-Speech MT Evaluation 51

Relating Translation Quality Barriers to Source-Text Properties 61

Author Index

Ahrenberg, Lars 10
Bailey, Stacey 18
Burchardt, Aljoscha 5, 31
Carl, Michael 11, 45
Christianson, Caitlin 7
Daems, Joke 2
Doherty, Stephen 16, 61
Dorr, Bonnie 7
Gaspari, Federico 16, 61
Gatewood, Paul 13, 14, 51
van Genabith, Josef 16, 61
Glazychev, Leonid 3
Herzog, Martha 13, 14, 51
Jones, Douglas 13, 14, 51
Lommel, Arle 5, 16, 31, 61
Macken, Lieve 2
Marg, Lena 15
Marius, Tamas 13, 14, 51
Melby, Alan 17
Miller, Keith J. 18
Olive, Joseph 7
Popović, Maja 5, 31
Schaeffer, Moritz 11, 45
Schumann, Anne-Kathrin 4, 20
Snow, Tyler 17
Stambolieva, Ekaterina 8
Starlander, Marianne 6, 38
Toral, Antonio 16, 61
Vandepitte, Sonia 2
Vela, Mihaela 4, 20
Way, Andy 16, 61
Wurm, Andrea 4, 20

Preface

While a significant body of work has been done by the machine translation (MT) research community towards the development and meta-evaluation of automatic metrics to assess overall MT quality, less attention has been dedicated to more operational evaluation metrics aimed at testing whether translations are adequate within a specific context: purpose, end-user, task, etc., and why the MT system fails in some cases. Both of these can benefit from some form of manual analysis. Most work in this area is limited to productivity tests (e.g. contrasting time for human translation and MT post-editing). A few initiatives consider more detailed metrics for the problem, which can also be used to understand and diagnose errors in MT systems. These include the Multidimensional Quality Metrics (MQM) recently proposed by the EU F7 project QTLaunchPad, the TAUS Dynamic Quality Framework, and past projects such as the Framework for Evaluation of MT in ISLE (FEMTI[1]), developed out of work in the EAGLES and the joint EU and NSF (US) International Standards for Language Engineering (ISLE) programs. Some of these metrics are also applicable to human translation evaluation. A number of task-based metrics have also been proposed for applications such as topic ID / triage, and reading comprehension.

The purpose of the Workshop on Automatic and Manual Metrics for Operational Translation Evaluation (MTE) was to bring together representatives from academia, industry and government institutions to discuss and assess metrics for manual and automatic quality evaluation, as well as how these might be leveraged or further developed into task-based metrics for more objective "fitness for purpose" assessment, and to compare them with well-established metrics for automatic evaluation such as BLEU, METEOR and others, including reference-less metrics for quality prediction. The workshop used datasets already collected and manually annotated for translation errors by the QTLaunchPad project (http://www.qt21.eu/launchpad/) and covers concepts from many of the metrics proposed by participants through a half-day of hands-on tasks.

We received 29 papers/abstract proposals, of which 12 were selected for presentation slots at the workshop. The workshop papers and abstracts cover metrics for machine (and/or human) translation quality evaluation and quality estimation, including metrics that are automatic, semi-automatic and manual. Papers also address comparisons between these metrics as well as correlations between the metrics and the task suitability of MT output.

The full-day workshop consisted of two parts: 1) half day for the presentation and discussion of recent work on the topics of interest; 2) half day for hands-on activities during which participants were asked to perform task-based quality evaluation on machine translation data, including MQM-based annotation as well as annotation and other tasks suggested by selected workshop submissions.

As a follow-on to the hands-on activities and general discussions during the workshop, the organizers performed a post-workshop analysis of the human evaluation data collected along with the automated metrics. The results and annotated data will be available to any interested parties for further investigation on the workshop's external website at http://mte2014.github.io.

The Organizing Committee

[1] http://www.issco.unige.ch:8080/cocoon/femti/

Extended Abstracts

Two Sides of the Same Coin: Assessing Translation Quality in Two Steps through Adequacy and Acceptability Error Analysis

Joke Daems, Lieve Macken, Sonia Vandepitte
Department of Translation, Interpreting and Communication, Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
E-mail: joke.daems@ugent.be, lieve.macken@ugent.be, sonia.vandepitte@ugent.be

A translator has to find the balance between adhering to the norms of the source text (adequacy) and respecting the norms of the target text (acceptability) (Toury, 1995). The quality of a translation can then be judged on its (non-)adherence to these norms. This is a common quality judgment for machine translation, where evaluators give translated segments an adequacy and acceptability (sometimes 'fluency') score on a scale from one to five (White, 1995).

When looking at translation quality assessment through error analysis, however, the dichotomy between acceptability and adequacy is not always as distinct. Existing metrics do provide error categories relating to both types of issues. For example, QTLaunchPad's MQM has a category for fluency and one for accuracy; TAUS suggests a category for accuracy, one for terminology, and two categories that could relate to acceptability (language and style); FEMTI proposes suitability, accuracy and wellformedness; and MeLLANGE offers categories for language and content transfer. Yet these categories are all part of one and the same evaluation step: evaluators have to identify issues and assign the correct category to these issues. Research has shown that deciding whether an error belongs to adequacy or acceptability is one of the most difficult aspects of error analysis for human annotators, together with having to assign an error weight to each error instance (Stymne & Ahrenberg, 2012).

We therefore propose facilitating the error annotation task by introducing an annotation process which consists of two separate steps that are similar to the ones required in the European Standard for translation companies EN 15038: an error analysis for errors relating to acceptability (where the target text as a whole is taken into account, as well as the target text in context), and one for errors relating to adequacy (where source segments are compared to target segments). We present a fine-grained error taxonomy suitable for a diagnostic and comparative analysis of machine translated-texts, post-edited texts and human translations. Categories missing in existing metrics have been added, such as lexical issues, coherence issues, and text type-specific issues. Annotator subjectivity is reduced by assigning error weights to each error category beforehand, which can be tailored to suit different evaluation goals, and by introducing a consolidation step, where annotators discuss each other's annotations.

The approach has been tested during two pilot studies with student translators who both post-edited and translated different texts. Inter-annotator agreement shows that the proposed categorization is clear and that it is necessary to include a consolidation phase. Annotations after consolidation were used to analyze the most common errors for each method of translation and to provide an average error score per word for each text. In a next phase, the annotations were manually grouped into source text-related error sets: a source text passage and the translations for that passage that contain errors. Error sets allow for a diagnostic evaluation: which source text-segments that were problematic for machine translation are still problematic after post-editing and how? How many and which post-editing errors originate from the machine translation output?

Though the approach in its current form requires much time and human effort (the annotation process in itself costs around 45 minutes for 150 words of a new MT text, with acceptability annotations requiring the most time: 30 minutes), it does provide rich data needed to improve translation quality. Familiarity with a text can seriously decrease annotation time, and the time for HT or PE if also lower than for MT. We are currently optimizing the annotation process to increase the speed and reduce manual effort, and we believe that the processing of the annotations and the creation of the error sets can, at least in part, be automated.

How to Reliably Measure Something That's Not Completely Objective:
A Clear, Working and Universal Approach to Measuring Language Quality

Leonid Glazychev
CEO, Logrus International Corporation (www.logrus.net)
2600 Philmont Avenue, Suite 305, Huntingdon Valley, PA 19006 USA
E-mail: leonidg@logrus.net, lglazychev@outlook.com

While everybody needs to measure language quality, and numerous practical models as well as theoretical approaches have been developed over decades, all these models and approaches were concentrating on particular factors to measure and their relative weights, i.e. what is important and what is not. At the same time practical, real-world solutions targeted at human reviewers and applicable beyond the MT domain, when we have to deal with new translations of unknown origin (either human or [post-edited] MT) with no reference translations available, are scarce. Creating one requires providing answers to the following questions:

- How exactly can we reliably measure something that is not completely objective by design?
- How trustworthy the results of each particular human review are, and what is the best way to analyse and interpret them?
- How to develop the quality measurement approach/metric that is not simply justified, flexible and reliable enough, but can also be utilized in real life as part of the production process?

The presentation outlines the approach developed by the author at Logrus International Corporation and based on years of research and practical work on clients' projects. The suggested model is flexible (can be easily adapted to particular type of content or requirements), universal (can be applied to both human and machine translation) and practical (can be implemented as part of the production process).

The concept is based on selecting primary factors influencing the perception and priorities of the target audience, separating global and local issues and dividing all quality-related factors into three basic categories: objective, semi-objective and subjective. Each category is described in detail, including its nature, limitations and specifics. This classification is paired with a multidimensional approach to quality built around four "cornerstones": Adequacy, Readability, Technical Quality, and Major Errors.

The presentation provides concrete recommendations on the process, which includes:

- Applying threshold-based (pass/fail) criteria for most important global, semi-objective factors, such as content adequacy to the original and overall readability (fluency).
- Considering major errors (grossly distorting the meaning, creative offensive statements, etc.)
- Counting and classifying technical errors in materials that passed the first two tests and applying error-weighting templates appropriate for the occasion.

The presentation gives a deeper insight into interpreting quality review results and explains why this hybrid approach combining threshold-based (rubric) and regular quantitative criteria is optimal, discusses grading scales, etc. Practical details include the following:

- Why one should not over-rely on particular figures
- Why it is not recommended to wrap everything into a single quality evaluation grade, and why we need to substitute it with four different values defining the "quality square".
- Why is the scale used for grading so important

Both the "full" and "light" quality evaluation models are presented. The latter is significantly less effort-consuming and consequently less precise, but is invaluable for public initiatives involving unpaid community-sourced effort or for cases of quality evaluation on a shoestring budget.

All recommendations are illustrated using actual statistical results obtained through a community review of the localized version of a popular website by 18 professional translators.

Human Translation Evaluation and its Coverage by Automatic Scores

Mihaela Vela, Anne-Kathrin Schumann, Andrea Wurm
Department of Applied Linguistics, Translation and Interpreting, Saarland University
Campus A2 2, 66123 Saarbrücken, Germany
E-mail: m.vela@mx.uni-saarland.de, anne.schumann@mx.uni-saarland.de,
a.wurm@mx.uni-saarland.de

Approaches to the evaluation of machine translation output are numerous and range from fully automatic quality scoring to efforts aimed at the development of "human" evaluation scores. The goals for which such evaluations are performed are manifold, covering system optimisation and benchmarking as well as the integration of MT engines into industrially deployable translation workflows. The discipline of translation studies, on the other hand, can look back onto a long line of thought on the quality of translations. While the discipline has traditionally been centered on the human translator and her individual competence, the notion of "translation quality", in translation studies, has in the last decades assumed a multi-faceted shape, embracing aspects that go beyond an individual's competence of optimising the relation between linguistic naturalness and semantic fidelity or her ability to use rule sets specific to a given language pair.

This paper presents a study on human and automatic evaluations of translations in the French-German translation learner corpus KOPTE (Wurm, 2013). The aim of the paper is to shed light on the differences between MT evaluation scores and approaches to translation evaluation rooted translation studies. We illustrate the factors contributing to the human evaluation of translations, opposing these factors to the results of automatic evaluation metrics, by applying two of the most popular automatic evaluation metrics, namely BLEU (Papineni et al., 2002) and Meteor (Denkowski and Lavie, 2011), to a sample of human translations available from KOPTE. The goal of these experiments is threefold. Firstly, we want to study whether the automatic scores can mimic the fine-grained distinctions of the human translations expert who evaluated the translations available from KOPTE or, at least, make meaningful distinctions when applied to human translations. Secondly, we are interested in investigating how automatic evaluation scores evolve if the number of chosen references is increased. Finally, we are also interested in examining whether a higher number of references influence the correlation of the automatic scores with the human expert grades for the same translation. Our experiments suggest that both BLEU and Meteor systematically underestimate the quality of the translations tested.

By means of a qualitative analysis of human translations we then highlight the concept of legitimate variation and attempt to reveal weaknesses of automatic evaluation metrics. More specifically, our qualitative analysis suggests that lexical similarity scores are neither to cope satisfactorily with standard lexical variation (paraphrases, synonymy) nor with dissimilarities that can be traced back to the source text or the nature of the translation process itself.

References

Michael Denkowski and Alon Lavie. 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*.

Kishore Papineni, Salim Roukos, ToddWard, andWei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, pages 311–318.

Andrea Wurm. 2013. Eigennamen und Realia in einem Korpus studentischer Übersetzungen (KOPTE). *transkom*, 6:381–419, 2

Assessing Inter-Annotator Agreement for Translation Error Annotation

Arle Lommel, Maja Popović, Aljoscha Burchardt
DFKI
Alt-Moabit 91c, 10559 Berlin, Germany
E-mail: arle.lommel@dfki.de, maja.popovic@dfki.de, aljoscha.burchardt@dfki.de

One of the key requirements for demonstrating the validity and reliability of an assessment method is that annotators be able to apply it consistently. Automatic measures such as BLEU traditionally used to assess the quality of machine translation gain reliability by using human-generated reference translations under the assumption that mechanical similar to references is a valid measure of translation quality. Our experience with using detailed, in-line human-generated quality annotations as part of the QTLaunchPad project, however, shows that inter-annotator agreement (IAA) is relatively low, in part because humans differ in their understanding of quality problems, their causes, and the ways to fix them. This paper explores some of the facts that contribute to low IAA and suggests that these problems, rather than being a product of the specific annotation task, are likely to be endemic (although covert) in quality evaluation for both machine and human translation. Thus disagreement between annotators can help provide insight into how quality is understood.

Our examination found a number of factors that impact human identification and classification of errors. Particularly salient among these issues were: (1) disagreement as to the precise spans that contain an error; (2) errors whose categorization is unclear or ambiguous (i.e., ones where more than one issue type may apply), including those that can be described at different levels in the taxonomy of error classes used; (3) differences of opinion about whether something is or is not an error or how severe it is. These problems have helped us gain insight into how humans approach the error annotation process and have now resulted in changes to the instructions for annotators and the inclusion of improved decision-making tools with those instructions. Despite these improvements, however, we anticipate that issues resulting in disagreement between annotators will remain and are inherent in the quality assessment task.

TURKOISE: A Mechanical Turk-based Tailor-made Metric for Spoken Language Translation Systems in the Medical Domain

Marianne Starlander
University of Geneva, FTI-TIM, 40 Bd du Pont d'Arve, CH-1211 Genève 4
E-mail: Marianne.Starlander@unige.ch

In this paper, we will focus on the evaluation of MedSLT, a medium-vocabulary hybrid speech translation system intended to support medical diagnosis dialogues between a physician and a patient who do not share a common language. How can the developers ensure a good quality to their users, in a domain where reliability is of the highest importance?

MedSLT was designed with a strong focus on reliability in the correct transmission of the message. One of the characteristics of MedSLT is its rule-based architecture that uses an interlingua approach to produce highly reliable output. This approach avoids surface divergences in order to keep only the meaning of the sentences. Consequently, sentences are translated more freely and as a consequence of our speech input, the sentences are particularly short. These characteristics entail quite low BLEU scores as well as little correlation with human judgment. Besides these automatic metrics, we also completed several human evaluations; using different scales (including fluency and adequacy as well as ranking). None of our experimented metrics gave us satisfactory results in the search of an operational metric for speech translation systems in a safety-critical domain such as the medical diagnosis domain. We have thus decided to experiment with manual metrics in order to find an evaluation that could be implemented without producing human references and at reasonable cost, within a minimum time span.

In the following paper we will describe the path that led us to using Amazon Mechanical Turk[1] (AMT) as an alternative to more classical automatic or human evaluation. We started using adequacy and fluency metrics but soon decided to experiment with a tailor-made and task-specific human metric, adapted to our domain but that could be used by a wider group of evaluators thanks to the AMT while guaranteeing certain coherence between the evaluators. The proposed metric is called **TURK**OISE, designed to be used by unskilled AMT evaluators while guaranteeing reasonable level of coherence between evaluators.

Our study focuses on inter-rater agreement comparing this aspect for our in-house small group of translator-evaluators compared to a wider group of AMT workers. We would also like to quantify the effort in running the AMT evaluation in order to compare the resources needed. Developers and researchers tend to minimize the effort related with the creation of reference translations in order to use BLEU or other reference-based metrics. Hence, we assume that if AMT workers are found to be reliable, this type of evaluation would be, at least, as cost and time effective as the classical automatic metrics but providing the advantage of reflecting the end-user's quality level request.

Our main results of this experiment are that AMT workers are found to be reaching comparable levels of inter-rater agreement when using the classic fluency and adequacy metrics, but also TURKoise, being our tailor-made evaluation scale.

[1] www.mturk.com

MADCAT Evaluation: Operational Accuracy of MT Applied to OCR

Caitlin Christianson, Bonnie Dorr*, Joseph Olive
*Defense Advanced Research Projects Agency and *Florida IHMC & University of Maryland*
E-mail: {caitlin.christianson.ctr,joseph.olive.ctr}@darpa.mil, bdorr@ihmc.us

Evaluating progress is an important aspect of NLP research that can help identify the most effective techniques and systems. Prior evaluation techniques have been valuable in monitoring progress and comparing different MT systems, but they have failed to answer an important question relevant to research sponsors with an interest in operational MT use, namely what accuracy is necessary for any given application. To answer this question, we devised an experiment to solicit input from experienced users of translated material by providing them documents with varying levels of translation accuracy and asking them which of these documents would be useful for a given task. We were interested mainly in three tasks: editing, gisting, and triage. Documents deemed editable would be publishable with human editing, documents deemed gistable would be suitable for human readers to determine the basic meaning, and documents deemed triageable would be suitable for determining mission relevance.

Current MT algorithms were used to translate Arabic and Spanish documents, and accurate human translations were obtained. The MT outputs were then edited to reflect the meaning of the human translated documents. Errors were counted (insertions, deletions, substitutions and moves of any number of adjacent words) and divided by the number of words in the source document. Both machine-translated documents had, on average, 45% errors. The MT output was then corrected by randomly choosing the edited corrections in steps of 5% to generate 10 documents. The original MT output and human translations were added to these for a total of 12 documents. The results showed that triageable, gistable and editable documents required accuracy of 55%, 70%, and 85%, respectively. This work led to a new evaluation paradigm, Human-mediated Translation Error Rate (HTER; Olive and Christianson, 2011; Dorr et al., 2011). This meaning-based metric compares machine-translated text to a "gold standard" translation of the same text created by a team of human translators. The MT output is edited to obtain text that conveys the same meaning as the "gold standard" text; the number of edits are counted and divided by the number of words.

Our research in applying HTER has focused on the use of various evaluation methods to determine MT accuracy in relation to technology applications. An OCR-based example of this relation was investigated for the Multilingual Automatic Document Classification, Analysis, and Translation (MADCAT) program. An Arabic data set was generated to determine for measuring translation accuracy. However, to ascertain the applicability of MADCAT systems on operational data, the program acquired some hand-written documents collected in the field in Iraq. Consistent improvements in the ability of MADCAT to translate program-generated documents were obtained during the first 4 years of the project. However, for the field-collected documents, the starting point was much lower – not even triageable. By year 5, improvements were still significant – above gistable.

We have made the case for evaluation in context, using HTER on the final MT output, rather than standard transcription error rates used in OCR. We have also argued for determining performance levels on data with operational characteristics and relating accuracy judgments to utility levels of relevance to the end user.

References

Dorr, B.; Olive, J; McCary, J.; Christianson, C. (2011) "Chapter 5: Machine Translation Evaluation and Optimization," in Olive, J; Christianson, C; McCary, J. (Eds.), *Handbook of NLP and MT: DARPA Global Autonomous Language Exploitation*, pp. 745—843.

Olive, J; Christianson, C. (2011) "The GALE Program," in Olive, J; Christianson, C; McCary, J. (Eds.), *Handbook of NLP and MT: DARPA Global Autonomous Language Exploitation*, pp. vii—xiv.

Approved for Public Release, Distribution Unlimited

Continuous Operational Evaluation of Evolving Proprietary MT Solution's Translation Adequacy

Ekaterina Stambolieva
euroscript Luxembourg S.à. r.l.
Bertrange, Luxembourg
E-mail: Ekaterina.stambolieva@euroscript.lu

Little attention is given to the focus on continuous diagnostic monitoring of the adequacy of translations (Koehn, 2010) dependent on specific business scenarios. Numerous organizations, including ours, post-edit Machine Translation (MT) to accelerate translation time-to-delivery and reduce translation costs. Unfortunately, in many cases, MT quality is not good enough for the task of post-editing and MT systems struggle to deliver native-fluency translations (Allen, 2003). Many researchers (Krings 2001, He et al. 2010, Denkowski and Lavie 2012, Moorkens and O'Brien 2013) agree that human end-user (translators, project coordinators with solid linguistic and translation knowledge, among othres) evaluation input contributes to MT quality improvement. Armed with translators' feedback, benchmarks and metrics such as QTLaunchPad[1]'s MQM (Doherty et al., 2013) along with taraXÜ2[2]'s confidence score (Avramidis et al., 2011) tackle the MT quality problem in search of effective quality evaluation. Nevertheless, all of these do not solve the impending industry problem – evaluating and comparing over-time MT solution modifications. This paper contributes to the development of a Continuous Operational MT Evaluation (COMTE) approach, which concentrates on repeated evaluation of MT system improvements based on human end-user feedback.

It is crucial to secure high MT adequacy on each stage of the MT system modifications that reflect the human translators' assessment and expectation of the output. COMTE contributes to quality improvement in modified MT solutions based on end-users' feedback, and helps to increase post-editing task suitability. COMTE does not directly evaluate translation quality, like scores such as BLEU (Papineni et al., 2002) and METEOR (Banerjee et al., 2005), or metrics such as MQM. Instead COMTE assesses the over-time MT system improvement. It focuses on measuring translation adequacy and fluency based on developments, which solve MT solution issues and are suggested by the system's end-users. We propose to measure continuously translation adequacy in the task of post-editing by employing two well-known evaluation metrics. We explore the correlation between the metrics and the scheduled human-evaluation-driven MT system modifications.

The two founding scores of the approach are: Edit Distance (ED) (Przybocki et al., 2006) and Fuzzy Match (FM) (Bowker, 2002). ED measures in how many edits machine translation output transforms into a human translated segment. ED is employed as a simple metric that measures post-editing effort. On the other hand, FM is a metric inherited by computer-assisted translation (CAT) tools. It shows what percentage of the current text for translation can be fluently translated by selecting an existing translation from business-dependent Translation Memories (TM). In the language business, a FM threshold is set, on which many pricing strategies depend. All text that has a FM lower than the fixed threshold is machine translated. A TM match is retrieved for the rest. Importantly, this approach requires zero additional annotation effort – all the information is derived from collected structured translators' feedback of the MT output. We also show how ED and FM correlate depending on the business scenario and MT quality improvements.

Our approach suggests a reliable strategy for performing operational translation evaluation of evolving in-house MT systems with wide applicability in the language industry. The evolution of the systems is based on human end-user feedback collected in a systematic way, following a 4-category error typology. We present empirical evidence that ED and FM correlate with successful system improvements, and conclude that they can thus be used to automatically assess system development.

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

References

Jeffrey Allen. 2003. Post-editing. In Harold Somers, editor, *Computers and Translation: A Translator's Guide*, pages 297-317, John Benjamins B.V.

Eleftherios Avramidis, Maja Popovic, David Vilar Torres and Aljoscha Burchardt. 2011. Evaluate with Confidence Estimation: Machine ranking of translation outputs using grammatical features. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT-11)*, Edinburgh, United Kingdom.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Matric for MT Evaluation with Improved Correlation with Human Judgements. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguists (ACL-2005)*, Ann Arbor, Michigan.

Lynne Bowker. 2002. *Computer-Aided Translation Technology: A practical Introduction*, pages 98-101, University of Ottawa Press.

Michael Denkowski and Alon Lavie. 2012. Challenges in Predicting Machine Translation Utility for Human Post-Editors. In *Proceedings of ATMA 2012*, San Diego.

Stephen Doherty, Federico Gaspari, Declan Groves, Josef van Genabith, Lucia Specia, Aljoscha Burchardt, Arle Lommel and Hans Uszkoreit. 2013. *Mapping the Industry I: Findings on Translation Technogogies and Quality Assessment*. European Commission Report.

Yifan He, Yanjun Ma, Johann Roturier, Andy Way and Josef van Genabith 2010. Improving the post-editing experience using translation recommendation: A user study. In *Proceedings of the 9th Annual AMTA Conference*, pages 247-256, Denver, CO.

Phillip Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, New York, pages 217-218.

Chunk Accuracy: A Simple, Flexible Metric for Translation Quality

Lars Ahrenberg
Department of Computer and Information Science
Linköping University
E-mail: lars.ahrenberg@liu.se

Many approaches to assessment of translations are based on error counts. These are usually supported by detailed taxonomies that highlight different quality aspects. Even if provided with guidelines the categorization and localization of errors can be quite difficult and time-consuming for a human annotator. Efforts may be wasted on details that are not really relevant for the purpose at hand. For example, a post-editor may be more helped by getting information on the locations of errors than a detailed classification of them.

A framework such as MQM: Multidimensional Quality Metrics (Uszkoreit&Lommel, 2013) is very helpful as a guide to what may be relevant for a given evaluation purpose. There is still a problem of applying criteria, however, once you have a taxonomy. Even if your selection is small, it is still often multi-dimensional, and ambiguities are likely to arise. For example, the distinction between error categories such as Wrong Translation and Missing Word may be clear in principle, but can be hard to make in a concrete case. Also, the question remains how a multi-dimensional selection is used to compare systems. As Williams (2001: 329) puts it: "The problem is this: assuming you can make a fair assessment of each parameter, how do you then generate an overall quality rating for the translation?"

I suggest that these two problems can be at least partly remedied by the following measures: (1) use the simplest possible taxonomies and give priority to counts before types; (2) use chunks as the loci of problems; a chunk can be read as a single unit by a human and eases the task of assigning a problem to a particular word, as for instance in the case of agreement errors. Still, it is more informative than counting errors for the whole text or complete sentences. For example, a post-editor may be shown not just that there are errors in a sentence, but in which part of the sentence the errors are located.

In the simplest case chunks need only be categorized into problematic (P) and correct (C). The metric then becomes C/(C+P) (or a percentage). To increase granularity, we can use a n-ary scale (for example good, bad, and ugly as is currently popular) and report a distribution over these categories. To get more informative we can categorize problems as those pertaining to adequacy (relation to corresponding source chunks), fluency (target language problems) and others. And then climb further down a taxonomy such as the MQM as motivated by the evaluation purpose.

Chunk accuracy can be applicable whenever a micro-level analysis is called for, e.g., in assessment of student translations, in post-editing settings, or even for MT development. It can be automated to a some extent, thus reducing the human effort. While aggregating observations at the micro-level, and reporting quality characteristics, it is not a final assessment, however. It reports values, not thresholds.

In my presentation, I will further develop my arguments and make comparisons of chunk accuracy to other known frameworks for error analysis.

References

H. Uszkoreit and A. Lommel (2013). Multidimensional Quality Metrics: A New Unified Paradigm for Human and Machine Translation Quality Assessment.(http://www.qt21.eu/launchpad/sites/default/files/MQM.pdf)

M. Williams (2001). The Application of Argumentation Theory to Translation Quality Assessment. Meta 46(2): 326-344.

Word Transition Entropy as an Indicator for Expected Machine Translation Quality

Michael Carl and Moritz Schaeffer
Copenhagen Business School, Denmark
Email: mc.ibc@cbs.dk, moritzschaeffer@gmail.com

While most machine translation evaluation techniques (BLEU, NIST, TER, METEOR) assess translation quality based on a single (or a set of) reference translations, we suggest to evaluate the literality of a set of (human or machine generated) translations to infer their potential quality. We provide evidence which suggests that more literal translations are produced more easily, by humans and machine, and are also less error prone. Literal translations may not be appropriate or even possible for all languages, types of texts, and translation purposes. However, in this paper we show that an assessment of the literality of translations allows us to (1) evaluate human and machine translations in a similar fashion and (2) may be instrumental to predict machine translation quality scores.

While "translators tend to proceed from more literal versions to less literal ones" (Chesterman, 2011) it is controversial what it actually means for a translation to be literal. In this paper, we follow a strict definition which defines literal translations to "consist of the same number of lexical words, representing equivalent grammatical categories, arranged in the same literal order and underlying semantically equivalent sentences" (Krzeszowski, 1990). This definition is operationalized by the following criteria:

1. Word order is identical in the source and target languages
2. Source and target text items correspond one-to-one
3. Each source word has only one possible translated form in a given context.

In this talk we focus on point 3: Nine English source texts were machine-translated into Spanish and post-edited by nine different post-editors. The post-edited texts were subsequently corrected by independent reviewers. The translations were semi-automatically aligned on a sentence and a word level. Keystroke and gaze data was collected during a part of the post-editing sessions. See Carl et al, (2014) for a more detailed description of the experimental data.

We computed the edit distance between the MT output and the post-edited text (*MT-PE*) and the edit distance between the post-edited translation and its reviewed version (*PE-RE*). We take the *MT-PE* distance as a quality indicator for the MT output: the more a post-editor modifies the MT output the worse can be expected the MT quality to be and the bigger will be the *MT-PE* distance.

We computed the word translation entropy of the human post-edited texts *HH(e):* the word translation probabilities *p(e→si)* of an English word e into the Spanish word *si* were computed as the ratio of the number of alignments e-si in the post-edited texts. Subsequently, the entropy of an English source word *e* was computed as:

$$HH(e) = -1 * \Sigma i\ p(e \rightarrow si) * log2(p(e \rightarrow si)). \quad (1)$$

We also computed the word transition entropy in the machine translation search graph MH(e) based on the transition probabilities *p(e→si)* from *s-1* to *si* as provided in the Moses search graphs. Up to 10% of the most unlikely transitions were discarded, and transition weights of the remaining translation options were mapped into probabilities.

Given these metrics, we can make a number of assumptions: If the MT output of a source word e was not modified by any post-editor, then *HH(e)=0*. Conversely, *HH(e)* would reach its maximum value if the MT output for e was modified by every post-editor in a different way. If a segment was not modified at all, *MT-PE* would be 0 and hence we expect a positive correlation between *HH(e)* and *MT-PE*.

Further, we expect that translations become worse as the entropy *MH(e)* increases, since it might be more difficult for an MT system to decide which translation to choose if several word transition probabilities are similarly likely, and thus the likelihood may increase for a sub-optimal translation choice. Subsequently, we expect to see more post-editing activities on translations with higher *MH(e)* values, and thus a positive correlation between *HH(e)* and *MH(e)*. Finally, we expect that textual changes of the MT output require more gaze and translation time, so that we also expect a positive correlation between post-editing activities and *MH(e).* In our analysis we show that:

1. *HH(e)* correlates positively with the edit distance *MT-PE*. That is, the edit distance increases if different post-editors translate a source word *e* in different ways.
2. There is a negative correlation between *MT-PE* and *PE-RE*: the more a text was edited in the post-edited phase, the less it was modified during revision, and vice versa.
3. *HH(e)* correlates with the gaze duration on (and translation production time of) *e*. That is, it is more time consuming for a translator to translate a source language word which can be translated in many different ways, than a source word which translates only into few target words, with high probability.
4. *HH(e)* correlates with *MH(e).* That is, if human translators translate a source word in many different ways, also the SMT system has many translation options for that word.

In this paper we pinpoint a correlation between the entropy of human translation realizations and the entropy of machine translation representations. As such, this is not surprising, since statistical machine translation systems are trained on, and thus imitate, the variety of human produced translations. Entropy is tightly linked to translation literality, and as translations become less literal (be it for structural reasons or for translator's choices) state-of-the-art statistical machine translation systems fail, while human translators seem to deploy as of now non-formalized translation strategies, to select amongst the many possible the/a good translation. This turning point may serve as an indicator for translation confidence, beyond which the quality of MT output becomes less reliable, and thus MT post-editing may become less effective.

References

Michael Carl, Mercedes Martínez García, Bartolomé Mesa-Lao, Nancy Underwood (2014) "CFT13: A new resource for research into the post-editing process". In *Proceedings of LREC*.

Chestermann, Andrew. (2011) Reflections on the literal translation hypothesis. In *Methods and Strategies of Process Research*, edited by Cecilia Alvstan, Adelina Held and Elisabeth Tisselius, Benjamins Translation Library, pp. 13-23, 2011.

Krzeszowski, Tomasz P. (1990) Contrasting Languages: The Scope of Contrastive Linguistics. Trends in Linguistics, Studies and Monographs, Mouton de Gruyter.

A New Multiple Choice Comprehension Test for MT

Douglas Jones, Paul Gatewood, Martha Herzog, Tamas Marius*

MIT Lincoln Laboratory *DLI Foreign Language Center**

244 Wood Street, Lexington, MA *542 Rifle Range Road, Monterey, CA*

E-mail: daj@ll.mit.edu, paul.gatewood@ll.mit.edu, MHerzog2005@comcast.net, tamas.marius@dliflc.edu

We present results from a new machine translation comprehension test, similar to those developed in previous work (Jones et al., 2007). This test has documents in four conditions: (1) original English documents; (2) human translations of the documents into Arabic; conditions (3) and (4) are machine translations of the Arabic documents into English from two different MT systems. We created two forms of the test: Form A has the original English documents and output from the two Arabic-to-English MT systems. Form B has English, Arabic, and one of the MT system outputs. We administered the comprehension test to three subject types recruited in the greater Boston area: (1) native English speakers with no Arabic skills, (2) Arabic language learners, and (3) Native Arabic speakers who also have English language skills. There were 36 native English speakers, 13 Arabic learners, and 11 native Arabic speakers with English skills. Subjects needed an average of 3.8 hours to complete the test, which had 191 questions and 59 documents. Native English speakers with no Arabic skills saw Form A. Arabic learners and native Arabic speakers saw form B.

The overall comprehension results for English natives reading Form A were 88% for the original English texts, 53% for MT1, and 76% for MT2. System level BLEU scores were 0.0624 and 0.2158 respectively. Comprehension scores were not strongly correlated with BLEU scores. For the Arabic language learners who saw Form B, the comprehension results were 91% for English, 46% for Arabic, and 76% for MT2. For the Arabic native speakers who saw Form B, comprehension results were 82% for English, 80% for Arabic, and 72% for MT2. The Arabic learners, who had an average of 2.5 semesters of Arabic instruction at the college level, demonstrated comprehension at a level between that of MT1 and MT2 as read by native English speakers. No MT results were as good as native speakers reading their native language.

We used the standardized language skill descriptions defined by the Interagency Language Roundtable (ILR); see (ILR, 2014). To measure machine translation capabilities, as opposed to human foreign language capabilities, we constructed a variant of the Defense Language Proficiency Test, following the general DLPT design principles, but modified to measure the quality of machine translation. This test is a multiple-choice format, ILR-based machine translation test format described in the paper "ILR-Based MT Comprehension Test with Multi-Level Questions" by Jones et al. in the proceedings of HLT 2007. Test documents were rated for ILR reading skills and were split between Levels 2, 2+ and 3. Questions were also rated for ILR level: Level 1, 2, 2+, and 3; comprehension results generally reflected the difficulty levels.

References

ILR. (2014). Interagency Language Roundtable (website). http://www.govtilr.org/.

Jones, D., et al (2007). ILR-Based MT Comprehension Test with Multi-Level Questions. In Human Language Technologies 2007: NAACL, Short Papers, pages 77–80.

This work is sponsored by the Defense Language Institute under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

Standardized ILR-Based and Task-Based Speech-to-Speech MT Evaluation

Douglas Jones, Paul Gatewood, Martha Herzog, Tamas Marius*

MIT Lincoln Laboratory *DLI Foreign Language Center**

244 Wood Street, Lexington, MA *542 Rifle Range Road, Monterey, CA*

E-mail: daj@ll.mit.edu, paul.gatewood@ll.mit.edu, MHerzog2005@comcast.net, tamas.marius@dliflc.edu

This paper describes a new method for task-based speech-to-speech machine translation evaluation, in which tasks are defined and assessed according to independent published standards, both for the military tasks performed and for the foreign language skill levels used. We analyze task success rates and automatic MT evaluation scores for 220 role-play dialogs. Each role-play team consisted of one native English-speaking soldier role player, one native Pashto-speaking local national role player, and one Pashto/English interpreter. The overall PASS score, averaged over all of the MT dialogs, was 44%. The average PASS rate for HT was 95%.

Scenarios were of two general types: a basic definition without any complications, and a contrasting definition with some type of obstacle, perhaps minor, that needed to be overcome in the communication. For example, in a basic Base Security scenario, a Local National may seek permission to pass a checkpoint with valid identification. In a contrast scenario, he may lack the identification, but seek an alternative goal that does not involve passing the checkpoint. Overall PASS/FAIL results for the HT condition were 95% for basic scenarios and 94% for contrasting scenarios with obstacles. For MT we observed 67% PASS for basic and 35% for contrast scenarios. The performance gap between HT at 94~95% and MT with basic scenarios at 67% was 27% on average, whereas the difference between MT in basic scenarios and MT in contrasting scenarios was 32%.

The dialogs were also assessed for language complexity. Scenarios with language complexity at the ILR Levels 1, 1+ and 2 had PASS scores of 94%, 100% and 92% respectively in the HT condition. For MT the overall results were 47%, 48% and 31%. In other words, MT does not work as well when the language is fundamentally more complex. The average BLEU score for English-to-Pashto MT was 0.1011; for Pashto-to-English it was 0.1505. BLEU scores varied widely across the dialogs. Scenario PASS/FAIL performance was also not uniform within each domain. Base Security scenarios did perform relatively well overall. Some of the scenarios in other domains were performed well with MT but performance was uneven.

Role players performed 20 tasks in 4 domains. The domain-level PASS scores ranged from 89% to 100% in the HT condition. For MT we observed 83% PASS rate in one domain, Base Security, with the remaining three domains ranging from 26% to 50%. The dialogs were human-scored in two main ways: (a) aggregate PASS/FAIL outcomes, and (b) a secondary assessment of specific communication initiatives. Inter-coder agreement for task PASS/FAIL scoring, which required an assessment of several performance measures per task, averaged 83%. Agreement for the specific communication initiatives was 98%.

We learned that success rates depended as much on task simplicity as it did upon the translation condition: 67% of simple, base-case scenarios were successfully completed using MT, whereas only 35% of contrasting scenarios with even minor obstacles received passing scores. We observed that MT had the greatest chance of success when the task was simple and the language complexity needs were low.

Rating Evaluation Methods through Correlation

Lena Marg
Welocalize, Inc.
Frederick, MD, United States
E-mail: lena.marg@welocalize.com

While less attention may have been dedicated to operational or task-based evaluation metrics by the MT research community, in our case (i.e. Language Service Provider), every evaluation is by definition task-based as it is carried out at the request of or tailored to a specific end-client and therefore with a defined purpose, scope and end-user in mind. This being the case, there still seems to be a need for more appropriate and easy-to-use metrics, both for evaluating the quality of raw and post-edited MT versus human translation.

In 2013, we put together a database of all evaluations (automatic scorings, human evaluations including error categorization and productivity tests including final quality assessments) carried out that year in Welocalize, in order to establish correlations between the various evaluation approaches, draw conclusions on predicting productivity gains and also to identify shortcomings in evaluation approaches. The database was limited to evaluations of that year for consistency in approach with regard to human evaluations and productivity tests compared to previous years.

Among the findings we observed were that the Human Evaluations of raw MT (especially the "Accuracy" score) seemed to be stronger predictors for potential productivity gains than automatic scores; Human Evaluation error categorizations provided initial glimpses of (cognitive effort) trends, but the markings seemed to be unreliable to some extent; further analysis, adjustment and fine-tuning of the (final) QA process are needed.

As part of the workshop, I would like to share findings from our data correlation analysis, which metrics turned out to be most valid and where we identified shortcomings. I will also be able to share first steps taken to improve our evaluation protocols in ongoing tests.

Description of metrics used in correlation database

The automatic score used for the data correlation is BLEU. When produced by an MT system, it would be based on MT versus human reference from a TM. In the case of productivity tests they can also be generated from MT versus post-edited version of the given content.

Human Evaluations of raw MT output are scored on a scale from 1-5, with 5 indicating "very good" quality and 1 indicating "very low" quality. They are divided into three parts: Accuracy score, Fluency score and Error Categorization. Human Evaluations are typically carried out on a set of manually selected strings representative of the content to be evaluated (i.e.: string length; typical "pitfalls" such as handling of software options, measurements and conversions, "To"-structures, gerunds, marketing speak, enumerations, elliptical structures etc.).

Productivity Tests are carried out in iOmegaT, an instrumented version of the open-source CAT tool co-developed by John Moran and Welocalize, which captures the translation time and number of edits for each segment. Test kits contain a mix of segments to translate from scratch and segments to post-edit, and linguists are usually asked to carry out 8 hours of translation/post-editing work.

Similar to the LISA QA Model and SAE J2450, the current QA metrics are a quantitative-based method of translation quality assessment which measures the number, severity and type of errors found in a text and calculates a score, which is indicative of the quality of a given translation.

Relating Translation Quality Barriers to Source-Text Properties

Federico Gaspari*, Antonio Toral*, Arle Lommel^,
Stephen Doherty°, Josef van Genabith§, Andy Way*

** School of Computing*	*^ DFKI GmbH*	*° School of Humanities &*	*§ DFKI GmbH*
Dublin City University	*Language Technology*	*Languages*	*Language Technology*
Glasnevin	*Alt Moabit 91c*	*University of New South Wales*	*Campus D3 2*
Dublin 9	*D-10559 Berlin*	*Sydney 2052*	*D-66123 Saarbrücken*
Ireland	*Germany*	*Australia*	*Germany*

E-mail: {fgaspari, atoral, away}@computing.dcu.ie, arle.lommel@dfki.de,
s.doherty@unsw.edu.au, josef.van_genabith@dfki.de

This study presents work on the identification of translation quality barriers. Given the widely perceived need to enhance MT quality and the reliability of MT evaluation for real-life applications, this study is of potential interest to a variety of MT users and developers. Our study focuses on identifying the source-side linguistic properties that pose MT quality barriers for specific types of MT systems (statistical, rule-based and hybrid) and for output representative of different quality levels (poor-, medium- and high-quality) in four translation combinations, considering English to and from Spanish and German. Using the diagnostic MT evaluation toolkit DELiC4MT and a set of human reference translations, we relate translation quality barriers to a selection of 9 source-side PoS-based linguistic checkpoints (adjectives, adverbs, determiners, common nouns, nouns, proper nouns, particles, pronouns and verbs).

DELiC4MT is an open-source toolkit for diagnostic MT evaluation. Its diagnostic dimension derives from its ability to focus on user-defined linguistic checkpoints, i.e. phenomena of the source language that the user decides to analyse when evaluating the quality of MT output. Linguistic checkpoints can correspond to interesting or difficult lexical items and/or grammatical constructions for which a specific translation quality assessment is required. They can be defined at any level of granularity desired by the user, considering lexical, morphological, syntactic and/or semantic information.

DELiC4MT has so far been used to evaluate the overall quality of MT systems with respect to their performance on user-defined source-side linguistic phenomena. The novelty of this work lies in the application of this toolkit to the investigation of translation quality barriers. These are investigated according to two main variables. Firstly, we consider different MT system types: this variable enables us to compare the performance of statistical, rule-based and hybrid MT software on a selection of source-language linguistic checkpoints. Secondly, we look at human quality rankings of the MT output: this variable concerns the quality band assigned by human evaluators to the output of each MT system, whereby each sentence was rated as either good (rank 1), near-miss (rank 2) or poor (rank 3). We are thus able to evaluate the performance of the MT systems on each checkpoint separately for those sentences that fall into each of these rating bands.

We show that the combination of manual quality ranking and automatic diagnostic evaluation on a set of PoS-based linguistic checkpoints is able to identify the specific quality barriers of different MT system types across the four translation directions under consideration. On the basis of this evaluation, we have analysed the correlation between the scores obtained for each of these source-side linguistic phenomena and the human quality ratings, thus assessing the extent to which these phenomena can be used to predict human quality evaluation. Considering all the MT system types evaluated together, it turns out that the best predictors are verbs (r=0.795), proper nouns (r=0.658) and pronouns (r=0.604), while the worst one is by far adverbs (r=0.02).

Keywords: MT quality barriers, diagnostic evaluation, statistical/rule-based/hybrid MT, linguistic features

Understanding Stakeholder Requirements for Determining Translation Quality

Tyler Snow and Alan Melby
Brigham Young University
4064 JFSB, Provo, Utah 84602
E-mail: tylerasnow@gmail.com, alan.melby@gmail.com

This paper presents the results of a large-scale study on the translation-related language quality assessment practices of language service providers, content creators who purchase translation, and free-lance translators. Conducted by the Globalization and Localization Association (GALA) as part of the EU-funded QTLaunchPad Project, this study is intended to provide concrete feedback to influence the development of the Multidimensional Quality Metrics (MQM) system for analyzing translation quality. By specifying what the "real-world" requirements for a translation quality assessment system are, it will help ensure that MQM is aligned with industry best practice and is flexible enough to meet the requirements of the full range of potential users in industry and research.

The study began with a survey sent out to thousands of individuals in the above-mentioned stakeholder segments around the world concerning quality management as applied to their translation activities. Approximately 300 persons participated in the survey, and approximately 60 percent of those indicated they would be interested in follow-up interviews. Key findings include:

(1) There is no industry consensus on appropriate quality processes, and assessment processes are highly diverse, ranging from informal, subjective readings to highly rigorous, analytic approaches. There are currently no widely accepted best practices.

(2) The most common method involves "spot checks" conducted on small samples of translated data to determine whether texts are "good enough" or need additional remediation.

(3) Most of those surveyed use "analytic" quality assessment methods that evaluate the translated text closely to identify and quantify specific errors in the text. Less common alternatives include a "holistic" approach that involves rating the overall translation on one or more dimensions.

(4) The most common specific metrics today are either in-house adaptations of the LISA QA Model or ones built into tools such as CAT tools or purpose-built translation quality-checking tools.

(5) Many quality assessment processes use a scorecard to aid in evaluation. Evaluators go through the text to mark and categorize errors, information about which is entered into the scorecard to calculate a quality score (usually expressed as a percentage value).

(6) The most frequently checked issues are: technical issues related to internationalization/ localization engineering, accuracy (e.g., mistranslation), fluency (e.g., linguistic features and grammar), terminology compliance, typography, compliance with legal requirements, and consistency. But many stakeholders wish that metrics would address other features such as offensiveness, readability, functionality of code (to ensure that localization has not "broken" it), productivity, and adherence to specifications.

Understanding these quality processes and the requirements that various stakeholder groups have within the translation process is crucial for improving the quality assessment of translation and providing results that accurate reflect the "quality" of texts in real-world situations. This presentation provides an overview of the findings of the survey and qualitative interviews, specifically as they relate to the MQM system for defining quality metrics. It will identify the most common error categories found and discuss how they are used in industry settings and the practical issues that the various stakeholder segments experience in their efforts to define, determine, and assure quality.

Automated and Task-Based Evaluation of the Effects of Machine Translation Domain Tuning on MT Quality, Post-editing, and Human Translation

Stacey Bailey and Keith J. Miller
The MITRE Corporation
McLean, Virginia, USA
E-mail: sbailey@mitre.org, keith@mitre.org

Domain tuning (DT) is the process of tailoring a machine translation (MT) system to better handle data relating to a particular topic area, either by training the MT system with data that is representative of the topic's subject matter (e.g., scientific and technical literature) or by adding terminology that is relevant to that subject matter. While DT can improve the quality of MT output, knowing how, when, and to what extent users should invest in developing corpus and lexical resources for DT is unclear. This research begins to address these questions by investigating the effects of domain-tuning on the quality of the output of two commercial MT systems.

This research evaluates two approaches to machine translation domain tuning (MTDT): (1) training a custom engine using parallel, domain data and (2) lightweight tuning using domain-specific glossaries. The research combined automatic evaluation and in-depth task-based evaluation of Chinese-to-English translation in the cyber domain. This study provided a 3-way comparison between 1) post-editing MT output from two commercial MT systems, 2) human translation of texts with no MT, and 3) human translation without MT but with domain term translations provided.

The three working hypotheses were that 1) DT improves the quality of machine translation output over baseline capabilities, as measured by automatic evaluation metrics, 2) Human translation time can be reduced by requiring human translators to post-edit the output of domain-tuned MT systems and 3) The linguistic quality of the target language document can be improved by requiring human translators to post-edit the output of domain-tuned MT systems as opposed to starting with source text only. It was hypothesized the post-editing DT would improve speed and quality of translation as compared to both post-editing of baseline MT and human translation without MT.

For each MT engine, there were four engine variations compared, yielding a total of eight MT test conditions: Post-editing using (1) the MT engine without any DT. (2) the MT engine plus lightweight DT with a found domain-specific lexicon. (3) the MT engine plus a statistically retrained engine based on the training data, and (4) the MT engine plus both a statistically retrained engine and a found lexicon. There were two additional conditions compared to these MT conditions: (5) Manual translation that does not use MT but does use a domain-specific lexicon for highlighting found terms with glosses provided for the translator. (6) Manual translation with no MT or term highlighting.

16 participants were given abstracts that included just the source Chinese text or the source text plus either the output of the MT (one of the MT test conditions) or the manually highlighted terms. They were asked to correct the MT output or produce a final translation from scratch. Translation times were recorded, and after each translation, the participants were given a survey about the utility of the resources provided and their opinions of the translation quality.

The results suggest that, generally speaking, DT can improve performance on automatic MT metrics, but it is not straightforward to predict whether a particular type of DT will definitely improve performance on a given domain. For some conditions and metrics, the performance dropped with DT. With respect to translation rates, results were also mixed. Rates were faster for some MT conditions and slower for others. Most notably, it was slowest on output from the two MT conditions based on the most involved DT.

Finally, six quality control (QC) translators were given the Chinese segments with the collected English translations. The QC-ers reviewed the source segment, rated each translation, and counted errors. Follow-on work will correlate these data with the automatic metrics and time data.

Full Papers

Human Translation Evaluation and its Coverage by Automatic Scores

Mihaela Vela, Anne-Kathrin Schumann, Andrea Wurm
Department of Applied Linguistics, Translation and Interpreting, Saarland University
Campus A2 2, 66123 Saarbrücken, Germany
E-mail: m.vela@mx.uni-saarland.de, anne.schumann@mx.uni-saarland.de, a.wurm@mx.uni-saarland.de

Abstract
This paper presents a study on human and automatic evaluations of translations in a French-German translation learner corpus. The aim of the paper is to shed light on the differences between MT evaluation scores and approaches to translation evaluation rooted in a closely related discipline, namely translation studies. We illustrate the factors contributing to the human evaluation of translations, opposing these factors to the results of automatic evaluation metrics, such as BLEU and Meteor. By means of a qualitative analysis of human translations we highlight the concept of legitimate variation and attempt to reveal weaknesses of automatic evaluation metrics. We also aim at showing that translation studies provide sophisticated concepts for translation quality estimation and error annotation which the automatic evaluation scores do not yet cover.

Keywords: translation evaluations, translation quality, translation learner corpus

1. Translation evaluation

Approaches to the evaluation of machine translation output are numerous and range from fully automatic quality scoring to efforts aimed at the development of "human" evaluation scores. The goals for which such evaluations are performed are manifold, covering system optimisation and benchmarking as well as the integration of MT engines into industrially deployable translation workflows. Despite all differences, however, most evaluation approaches that are described in the MT literature, conceptualise translation quality as a compromise between *adequacy*, the degree of meaning preservation, and *fluency*, target language correctness (Callison-Burch et al., 2007).

The discipline of translation studies, on the other hand, can look back onto a long line of thought on the quality of translations. While the discipline has traditionally been centred on the human translator and her individual competence, the notion of "translation quality", in translation studies, has in the last decades assumed a multi-faceted shape, embracing aspects that go beyond an individual's competence of optimising the relation between linguistic naturalness and semantic fidelity or her ability to use rule sets specific to a given language pair. These aspects include functional, stylistic and pragmatic factors and are supposed to be taught and evaluated in a systematic fashion. In this section, we investigate commonalities and differences between approaches to evaluation developed both in MT and translation studies. Due to the amount of available literature, our overview is necessarily incomplete, but still insightful with respect to the factors that influence and the underlying theoretical concepts that guide translation evaluation in the two disciplines.

1.1. Automatic evaluation metrics for MT

For the MT developer, the ideal environment to test an MT system is to call a program/script which calculates how well the system performs. Based on this scenario, several automatic language-independent evaluation metrics have been developed. The big advantage of automatic evaluation metrics is that they can be applied to large amounts of data in a language-independent, fast and cheap fashion, especially if compared to human evaluation. It is also due to the automatic evaluation metrics that MT research progressed so much in the last years.

Automatic evaluation metrics try to estimate the closeness between a "hypothesis" translation and one or more "reference" translations. In the last years, the most frequently used evaluation metric has been IBM BLEU (Papineni et al., 2002). BLEU accounts for adequacy and fluency by calculating word precision. The overgeneration of common words is handled by clipping precision, meaning that a reference word is exhausted after it is matched. Usually BLEU takes into account the modified n-gram precision for N=4, combining the result into the geometric mean. In order to penalise hypothesis translations which are shorter than the reference translations, the computed modified precision is scaled by the brevity penalty (BP).

There are several other metrics used for tuning and evaluating MT systems. Another often used metric is NIST (Doddington, 2002). NIST is derived from BLEU and computes the modified n-gram precision for N=5 into the arithmetic mean. But NIST also takes into consideration the information gain of each n-gram, giving more weight to more informative (less frequent) n-grams and less weight to less informative (more frequent) n-grams. Another often occurring evaluation metric is Meteor (Denkowski and Lavie, 2011). Meteor evaluates a candidate translation by calculating precision and recall on the unigram level and combining them into a parametrised harmonic mean. The result from the harmonic mean is then scaled by a fragmentation penalty which penalizes gaps and differences in word order. Other widely used evaluation metrics in MT research are WER (Levenshtein, 1966) as well as PER (Tillmann et al., 1997). WER (word error-rate) computes the normalised Levensthein distance (Levenshtein, 1966) between a hypothesis translation and a reference translation. PER (position-independent error rate) is based on WER, but ignores the ordering of the words in a sentence by just

counting the number of deletions, insertions, and substitutions that are necessary to transform the candidate sentence into the reference sentence.

1.2. Human translation evaluation

Human MT evaluation approaches employ the (often tacit) knowledge of human annotators to assess the quality of automatically produced translations along the two axes of target language correctness and semantic fidelity. The simplest evaluation method seems to be a ranking of a set of hypothesis translations according to their quality. According to Birch et al. (2013), this form of evaluation was used, among others, during the last STATMT workshops and can thus be considered rather popular. Federmann (2012) presents a software that integrates facilities for such a ranking task.

Another evaluation method that measures semantic fidelity by determining the degree of parallelism of verb frames and semantic roles between hypothesis and reference translations is HMEANT (Birch et al., 2013), based on MEANT (Lo and Wu, 2011). Unfortunately, Birch et al. (2013) report difficulty in producing coherent role alignments between hypotheses and reference translations, a problem that affects the final HMEANT score calculation.

An indirect human evaluation method that is also employed for error analysis are reading comprehension tests (e.g. Maney et al. (2012), Weiss and Ahrenberg (2012)). Other evaluation metrics try to measure the effort that is necessary for "repairing" MT output, that is, for transforming it into a linguistically correct and faithful translation. One such metric is HTER (Snover et al., 2006) which uses human annotators to generate "targeted" reference translations by means of post-editing, the rationale being that by this the shortest path between a hypothesis and its correct version can be found. Snover et al. (2006) report a high correlation between evaluation with HTER and traditional human adequacy and fluency judgements. Last but not least, Somers (2011) mentions other repair-oriented measures such as post-editing effort measured by the amount of key-strokes or time spent on producing a "correct" translation on the basis of MT output.

1.3. Translation evaluation in translation studies

In translation studies, "good" translation, for a long time, was viewed as an optimal compromise between meaning preservation and target language correctness. Thus, the notion of "translation quality" matched the dichotomy between *adequacy* and *fluency* as put forward by today's MT researchers.

However, in recent years "translation quality" has assumed a more complicated conceptual outline. Mainstream translation studies, by now, postulate that, depending on the communicative context within and for which a translation is produced, the relation between source and target text can vary greatly. That is, the degree of linguistic or semantic "fidelity" of a good translation towards the source text depends on functional criteria. Consequently, translation strategies as well as translation evaluation procedures become dependent on functional criteria, a view that is most prominently advocated by the so-called *skopos the- ory* (cf. Dizdar (2003)). From this it follows that translation errors are not simply linguistically incorrect structures or "mistranslated segments", but functional defects that can occur on all levels of text production (Nord, 2003), including errors in the use of phraseology, idioms, syntactic structures, grammatical, modal, temporal, stylistic, cohesive and other features. Moreover, the nature of the translation process itself, the transfer of a text into a new semiotic system, can result in *translation-specific errors* that occur when the translation does not fulfill its function because of pragmatic (e. g. text-type specific forms of address), cultural (e. g. text conventions, proper names, or other conventions) or formal (e. g. layout) defects (Nord, 2003). Depending on the appropriate translation strategy for a given translation task, these error types may be weighted differently. Consequently, the concept of "equivalence" which in its oldest form used to echo today's MT concept of "adequacy", in modern translation studies, depends not only on semantic equality, but also on aesthetic, connotational, textual, communicative, situational, functional and cognitive aspects (for a detailed discussion see Horn-Helf (1999)). In MT evaluation, most of these aspects have not yet or only in part been considered.

For large-scale evaluation purposes, the translation industry has developed normative standards and proofreading schemes. For example, the DIN EN 15038:200608 (Deutsches Institut für Normung, 2006) discusses translation quality aspects, quality management and qualificational requirements for translators and proofreaders, while the SAE J2450 standard (Society of Automotive Engineers, 2005) presents a weighted "translation quality metric". An application perspective is given by Mertin (2006) who discusses translation quality management procedures from an industry point of view and, among other things, develops a weighted translation error scheme for proofreading.

1.4. Discussion

The above discussion of approaches to translation evaluation put forward by machine translation researchers and researchers in the field of translation studies reveals that both the practical evaluation methods and the underlying theoretical concepts vary greatly between the two disciplines. The most important differences are the following:

- In translation studies, translation evaluation is considered an expert task, for which translation-specific expert knowledge is required on top of a specific *Multilingual* (source and target language) competence. According to normative standards, proofreaders must be experienced professional translators.
- Evaluation, in translation studies, is normally not carried out on the sentence level, since sentences can contain more than one "translation problem". Con-

sequently, the popular MT practice of ranking whole sentences according to some automatic score, by anonymous evaluators or even users of Amazon Turk (e. g. in the introduction to Bojar et al. (2013)), from a translation studies point of view, is unlikely to provide reasonable evaluations.

- The view that translation quality can be defined along the two axes of adequacy and fluency does not fit the complicated source/target text relations that have been acknowledged by translation studies. Even more importantly, evaluation methods based on simple measures of linguistic equality fail to provide straightforward criteria for distinguishing between *legitimate* and *illegitimate* variation. Moreover, semantic and pragmatic criteria as well as the notion of "reference translation" remain unclear.

However, the realisation that evaluation methods need to be improved is not new to the MT community: Birch et al. (2013) state that ranking judgments are difficult to generalise, while Callison-Burch et al. (2007) discuss the reliability of BLEU. Moreover, the depth and degree of sophistication of evaluation methods are clearly dependent on the goal for which translations are produced. Therefore the questions whether and, if yes, how MT evaluation research can benefit from the more fine-grained distinctions commonly used in translation studies is still an open research topic.

2. The KOPTE corpus

2.1. Corpus design

KOPTE (Wurm, 2013) is aFrench-German corpus of translations produced in class by translation students at the Department of Applied Linguistics, Translation and Interpreting at Saarland University. The aim of the corpus is to enable research on translation evaluation in a university training course (master's degrees) for translators and to enlighten student's translation problems as well as their problem solving strategies. The corpus covers 985 translations of 77 newspaper texts comprising a total of 318 467 tokens. The source texts are French newspaper texts that had to be translated into corresponding German press articles, that is, maintaining the dominant textual function of the original texts. Each of the translations was graded according to the German grade system on a scale ranging from 1 (=very good) to 6 (=very bad) with in-between intervals at the levels of 0.3 and 0.7.[1]

2.2. Translation evaluation in KOPTE

The evaluation of the student translations was carried out by an experienced translation teacher. Grading was based on an evaluation of both good solutions and translation errors which are weighted on a scale ranging from plus/minus 1 (minor) to plus/minus 8 (major). The final grade is calculated by summing up positive and negative scores before subtracting the negative score from the positive one. A score of around zero corresponds to the grade "good" (=2), to achieve "very good" (=1) the student needs a surplus of positive evaluations.

The evaluation scheme based on which student translations were graded comprises both external and internal factors. *External* characteristics describe the communicative context in which the source text functions and the translation brief (author, recipient, medium, location, time).

Internal factors, on the other hand, include eight categories: form, structure, cohesion, stylistics/register, grammar, lexis/semantics, translation-specific problems, function. Some internal subcriteria of these categories are summarised in Table 1. A quantitative analysis of error types in KOPTE shows that semantic/lexical errors are by far the most common error type in the student translations (Wurm, 2013).

Evaluations in KOPTE rely on the expertise of just one evaluator for the reason that, in a classroom setting, multiple evaluations are not feasible. Although multiple evaluations would have been considered valuable, KOPTE evaluations were provided by an experienced translation scholar with long-standing experience in teaching translation. Moreover, the evaluation scheme is much more detailed than error annotation schemes that are normally described in the MT literature and it is theoretically well-motivated. An analysis of the median grades in our data sample shows that grading varies only slightly between different texts, considering the maximum variation potential ranging from 1 to 6, and thus can be considered consistent.

[1] More information about KOPTE is available from http://fr46.uni-saarland.de/index.php?id=3702&L=%2524L

Criteria	Examples of subcriteria
author, recipients, medium, topic, location, time	
form	paragraphs, formatting
structure	thematic progression, macrostructure, illustrations
cohesion	reference, connections
stylistics	style, genre
grammar	determiners, modality, syntax
semantics	textual semantics, idioms, numbers, terminology
translation problems	erroneous source text, proper names, culture-specific items, ideology, weights, measurements, pragmatics, allusions
function	goal dependence

Table 1: Internal evaluation criteria in the KOPTE annotation scheme.

3. Experiments

In order to test the reproducibility of the evaluation performed by the human expert through automatic evaluation scores, that is, to test whether they measure a similar concept of quality, we applied two of the most popular automatic evaluation metrics, namely BLEU (Papineni et al., 2002) and Meteor (Denkowski and Lavie, 2011), to a sample of human translations available from KOPTE. The goal of these experiments was threefold. Firstly, we wanted to study whether the automatic scores can mimic the finegrained distinctions of the human expert or, at least, make meaningful distinctions when applied to human translations. Secondly, we were interested in investigating how automatic evaluation scores evolve if the number of chosen references is increased. Finally, we were also interested in examining whether a higher number of references influence the correlation of the automatic scores with the human expert grades for the same translation.

For clarifying these three questions, we conducted two sets of experiments. Automatic scores were calculated on the text level. The correlation between the human judgement and the BLEU and Meteor scores assigned to each translation was calculated using Kendall's rank correlation coefficient as described in Sachs and Hedderich (2009). The following correlations were calculated: the correlation between the human expert grades and BLEU, the correlation between human expert grades and Meteor and between BLEU and Meteor.

3.1. Setup and results

In the *first set of experiments* we conducted three experiments. In the first experiment, we applied the automatic evaluation metrics to the source texts listed in table 2. For each text, we chose the translation with the best human grade as reference translation, the remaining translations were considered "hypothesis" translations. The number of evaluated translations, the resulting median human grades, the median BLEU and Meteor scores and the correlation scores (all excluding the reference translation) obtained for each text are listed in Table 2.

In the second and third experiment we repeated the same procedure, but this time with three, respectively five, reference translations. In these experiments source texts with less than four hypotheses were excluded from the data set. The results are listed, analogously to the first experiment, in Table 3 and Table 4.

In the *second series of experiments*, we conducted the same experiments as in the first series, considering, however, this time the worst-graded translations as reference translations. The goal of this was to study whether the quality of the reference translations used for evaluation purposes changes the result of the evaluation. In the first experiment, we selected three reference translations, respectively the three lowest-graded translations. Table 5 lists the corresponding values for this setup. In the last experiment, we chose the five lowest-graded translations as reference translations. The results of this experiment can be found in Table 6.

We tested whether the BLEU and Meteor scores obtained in the two experimental series were significantly different from each other. To this end, we calculated the Wilcoxon rank sum test as described in Sachs and Hedderich (2009) for the automatic scores obtained upon evaluation against three references in the two series. For BLEU, the measured p-value was 0.9004, indicating lack of significance. For Meteor, we got a slightly significant p-value of 0.0467. Figure 1 depicts the differences between the lowest-graded and best-graded BLEU scores, as well as for the lowest-graded and best-graded Meteor scores. The differences between the Meteor scores becomes clearer when looking separately at the distribution of the Meteor scores in Figure 2.

3.2. Interpretation

The above tables show that the amount of source texts and their corresponding human translations used for the experiment decreases as the number of references is increasing. If in the first experiment 152 translations were evaluated, in the second and third experiment we dealt with 108, respectively 68, translations. Analysing the BLEU and Meteor scores one can notice that for the experiments with the bestgraded human references the highest mean per source text for BLEU is 0.26 (source text AT008 in

Table 4) whereas the highest mean for Meteor is 0.45 (source text AT008 in Table 3 and Table 4). The median of the human-graded translations for the same source text is 2.85, respectively 2.5denoting not very good, but good, readable and understandable translations. For the set of experiments with the worst-graded human translations as references, the situation does not change much, the best mean for BLEU remains at 0.26 and the best mean for Meteor increases slightly up to 0.47 (source text AT008 in Table 6). Although the BLEU and Meteor scores did not increase significantly, the median of the human assigned grades is 1.5 for this setting and source text, showing that the remaining hypothesis translations were indeed good to very good translations - a difference that the automatic evaluation scores did not capture. Overall, both the BLEU and Meteor scores obtained on the human translations of our KOPTE sample seem too low.

With respect to the relation between human and automatic evaluation, we observe that neither BLEU nor Meteor (except in a few exceptional cases with mainly few "hypotheses") correlate with the human quality judgements, however, they show a tendency to correlate with each other. Moreover, the increase of reference translations does not improve the BLEU and Meteor scores. Furthermore, the fact that the scores obtained in the two experimental series are not strongly different from each other gives reason to ask whether this kind of evaluation is actually meaningful. One reason for this similarity could be a somewhat "equal distance" between references and translations in the two series, however, this explanation seems somewhat devious. Even more disturbing is the observation that in four out of five experiments the highest BLEU and Meteor scores were obtained for AT008 although the human translations available for this source text, as indicated by the median of the human grades, are not the best in the data set. We believe that one reason for this result is the fact that the French source text contains many numbers (4.17%) and person names (6.63%) which are not changed upon translation, but allow for easy matching. Overall, these findings raise doubts concerning the concept of "reference translation": What is it actually that translations are evaluated against in practical evaluation settings and how much do the quality of the reference itself and its properties influence evaluation results?

Source text	**Human trans./ source text**	**Median grades**	**Mean BLEU**	**Mean Meteor**	**Correlation Human-BLEU**	**Correlation Human-Meteor**	**Correlation BLEU-Meteor**
AT001	7	2. 7	0.15	0.33	-0. 39	-0. 73	0.24
AT002	12	2. 3	0.15	0. 35	-0. 20	-0. 43	0.49
AT004	12	2. 7	0.19	0. 37	0.14	0.11	0.63
AT005	12	2. 3	0. 20	0. 36	0. 32	0. 45	0.45
AT008	10	2.15	0. 23	0. 38	-0. 43	-0. 29	0.78
AT010	11	2. 7	0. 25	0. 41	0. 06	-0.10	0.56
AT012	9	2. 0	0. 22	0. 40	-0. 30	-0. 36	0.50
AT015	5	2. 0	0.11	0. 28	0. 36	0.12	0.60
AT017	7	2. 3	0. 22	0. 38	-0. 20	0. 06	0.71
AT021	4	3. 0	0.18	0. 39	-0. 55	-0. 55	1.00
AT023	6	2. 3	0. 22	0. 38	0. 50	-0. 07	-0.20
AT025	4	2.15	0.13	0. 36	0. 33	0. 0	0.00
AT026	21	3. 0	0.12	0. 26	-0.19	-0. 35	0.67
AT039	13	3. 0	0.10	0. 29	-0. 08	0. 03	0.49
AT052	7	2. 0	0.17	0. 31	-0. 32	0. 05	0.00
AT053	7	2. 3	0.18	0. 32	0. 62	0. 39	0.33
AT059	5	2. 0	0. 24	0. 36	0. 00	0. 22	0.80

Table 2: Source texts, number of human translations per source text, median of the obtained grade per source text, mean of the BLEU and Meteor scores per source text and one best-graded reference and Kendall's rank correlation coefficients for the first experiment.

Source text	Human trans./ source text	Median grades	Mean BLEU	Mean Meteor	Correlation Human-BLEU	Correlation Human-Meteor	Correlation BLEU-Meteor
AT001	5	3. 0	0. 17	0. 36	-0.12	0. 36	0.60
AT002	10	2. 3	0. 17	0. 36	-0.14	0. 05	0.38
AT004	10	2. 85	0. 20	0. 37	0. 39	0.16	0.51
AT005	10	2. 3	0. 20	0. 40	-0.10	0. 05	0.47
AT008	8	2. 5	0. 25	0. 45	-0. 67	-0.15	0.00
AT010	9	2. 7	0. 23	0. 41	-0.10	-0. 50	0.28
AT012	7	2. 3	0. 23	0. 43	0. 00	0.11	0.52
AT017	5	2. 3	0. 21	0. 43	0.12	0. 36	0.60
AT023	4	2. 5	0. 21	0. 38	0. 41	0. 81	0.67
AT026	19	3. 3	0.10	0. 26	-0. 31	-0. 41	0.77
AT039	11	3. 0	0.11	0. 34	0. 06	0.14	0.74
AT052	5	2. 0	0.18	0. 40	0.12	0. 36	0.20
AT053	5	2. 3	0.17	0.35	0. 36	-0.12	0.40

Table 3: Source texts, number of human translations per source text, median of the obtained grade per source text, mean of the BLEU and Meteor scores per source text and three best-graded references and Kendall's rank correlation coefficients for the second experiment.

3.3. Legitimate and illegitimate variation in human translation

Any attempt to come up with explanations for the inability of the two automatic metrics to simulate the evaluation behaviour of the human expert leaves much room for exploration, however, we believe that one reason at least is the large amount of *legitimate Variation* (in addition to illegitimate variation, that is, translation errors) that can be found in human translations. Since we did not have the resources for an exhaustive study, we selected three source texts, namely AT008, AT023 and AT053 and performed a qualitative analysis of translation variants found in the German versions of these texts. The phenomena we found can partly be described as well-known translation problems (e. g. proper nouns, colloquial and figurative speech, culture-specific elements), others can be circumscribed as the use of simple synonyms and paraphrases. We will now discuss some examples in more detail.

The first phenomenon to deal with is synonymy. In Example 1, the verb *arroser*, having in this context the meaning *payment of bribe*, is translated into German by using the verbs *schmieren* and *bestechen.* The translation *mit Spendengeldern überschüttet (overwhelmed by moneyfrom donations)* can be viewed as a translation error. Note also the use of different tenses in the different German translations, which is legitimate.

(1) arros6
schmierten
bestochen
mit Spendengeldern überschüttet
geschmiert

Source text	Human trans./ source text	Median grades	Mean BLEU	Mean Meteor	Correlation Human-BLEU	Correlation Human-Meteor	Correlation BLEU-Meteor
AT002	8	2. 5	0. 17	0. 36	-0.08	0. 00	0.43
AT004	8	3. 0	0. 20	0. 36	0. 00	0. 23	0.71
AT005	8	2. 3	0. 20	0. 42	0. 00	0. 08	0.43
AT008	6	2. 85	0. 26	0. 45	-0. 55	-0.14	0.33
AT010	7	2. 7	0. 23	0. 41	0. 00	-0.12	0.05
AT012	5	2. 3	0. 23	0. 43	0. 22	0. 22	0.40
AT026	17	3. 3	0.11	0. 31	-0. 24	-0. 34	0.62
AT039	9	3. 0	0.10	0. 37	0. 22	0. 55	0.22

Table 4: Source texts, number of human translations per source text, median of the obtained grade per source text, mean of the BLEU and Meteor scores per source text and five best-graded references and Kendall's rank correlation coefficients for the third experiment.

Source text	Human trans./ source text	Median grades	Mean BLEU	Mean Meteor	Correlation Human-BLEU	Correlation Human-Meteor	Correlation BLEU-Meteor
AT001	5	2. 3	0.17	0. 36	0.11	0.12	0.6
AT002	10	2.15	0.18	0. 40	-0.12	-0. 12	0.6
AT004	10	2. 5	0. 20	0. 39	-0. 07	0. 21	0.51
AT005	10	2. 0	0. 20	0. 42	-0. 05	0. 20	0.69
AT008	8	1.7	0. 24	0. 46	-0. 07	-0.15	0.79
AT010	9	2. 7	0. 22	0. 43	0. 1	-0. 37	0.5
AT012	7	2. 0	0. 22	0. 40	0. 26	0. 80	0.33
AT017	5	2. 3	0. 20	0. 46	N. A.	N. A.	N. A.
AT023	4	0.19	0. 20	0. 38	-0. 91	-0. 91	1.0
AT026	19	2. 3	0.10	0. 34	0. 23	-0. 22	0.41
AT039	11	2. 7	0.11	0. 37	0. 24	0. 20	0.53
AT052	5	1.7	0.19	0. 37	0. 22	0. 90	0.4
AT053	5	2. 0	0.18	0.35	0. 51	0. 51	0.2

Table 5: Source texts, number of human translations per source text, median of the obtained grade per source text, mean of the BLEU and Meteor scores per source text and Kendall's rank correlation coefficients for evaluation with three "bad" references.

Example 2 is also concerned with synonymy. In this example, the adjective *ambivalent (ambivalent)* is correctly translated by *ambivalent* as well as by its synonyms *gegensätzlich* and *widersprüchlich.* Even the phrase *von großer Ambivalenz geprägt* is a valid synonym of *ambivalent*, the translator choosing here to nominalise the adjective.

(2) ambivalentes
von großer Ambivalenz geprägt
gegensätzlich
ambivalent
widersprüchlich

Another good illustration for synonymy is Example 3. Here, the verb *affirmer (to state)* is translated with *betonte, hat versichert* and *bestätigte*, all valid translations for the French verb. The translation *Aussage (statement)* is not incorrect, the translator having decided to use a nominalisation instead of a predicative rendering of the French structure. Nominalisations are indeed typical for German newspaper texts, so this solution can be considered valid. Another source of variation is the ambigous French collocation *justes paroles (true/right words).*

(3) Nicolas Sarkozy, en affirmant a Libreville que ...
a prononc6 de justes paroles
Nicolas Sarkozy hat die Wahrheit gesagt, als er während seiner Rede in Libreville betonte
Nicolas Sarkozy hat (der Regierung) in Libreville versichert
Nicolas Sarkozy hat die richtigen Worte gefunden, als er in Libreville bestätigte
Nicolas Sarkozys Aussage in Libreville

The next phenomenon to look at are bigger linguistic units, such as phrases which can also be used to exemplify the application of different translation strategies. In Example 4 we remark that three translators have maintained the original construction featuring a relative clause, whereas the other two translators decided to avoid this less typical construction in their German versions of the text. Of special interest is the last translation, in which the translator has changed the text's perspective on the situation: Instead of maintaining the French perspective, that is, instead of talking of Germany as a foreign country, the translator chose to adapt the text to the German perspective by introducing the adverb *hierzulande (in our country)* while omitting the country name. By doing so, the translator realises a fundamental quality aspect of translation, namely the power of cross-cultural conceptual transfer. Both foreignisation and domestication are important translation strategies which, depending on

Source text	Human trans./ source text	Median grades	Mean BLEU	Mean Meteor	Correlation Human-BLEU	Correlation Human-Meteor	Correlation BLEU-Meteor
AT002	8	2. 0	0.17	0. 39	-0.29	-0. 29	0.43
AT004	8	2. 3	0. 20	0. 40	0. 23	0. 23	0.64
AT005	8	1.85	0.19	0. 43	-0. 08	0.15	0.43
AT008	6	1.5	0. 26	0. 47	0. 21	-0. 07	0.6
AT010	7	2. 3	0. 22	0. 44	0. 26	-0.16	0.52
AT012	5	1.7	0. 22	0. 40	0. 45	0. 89	0.4
AT026	17	2. 3	0.11	0. 35	0.12	-0.15	0.61
AT039	9	2. 3	0.11	0. 36	0.18	0. 30	0.55

Table 6: Source texts, number of human translations per source text, median of the obtained grade per source text, mean of the BLEU and Meteor scores per source text and Kendall's rank correlation coefficients for evaluation with five "bad" references.

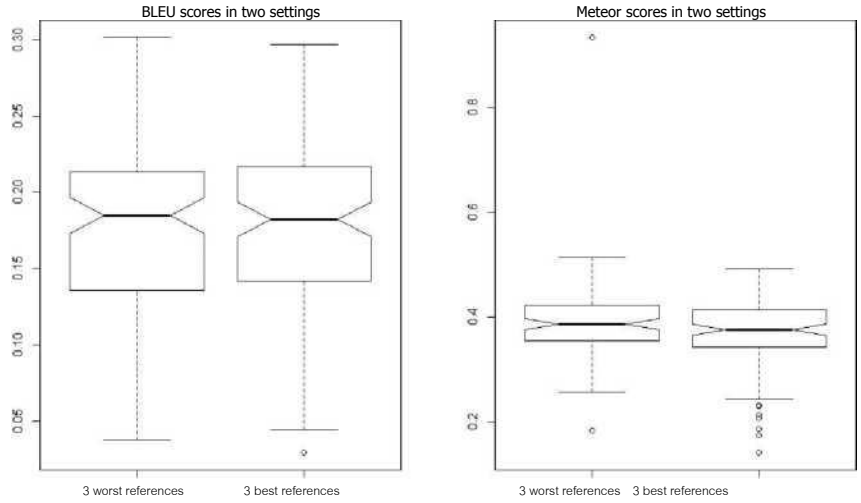

Figure 1: The difference between the BLEU and Meteor scores for the three lowest-graded and the three best-graded translations as references.

the intended use of the translation, can be applied with equal legitimacy. Also note that with respect to the foreign-domestic distinction the second translation remains neutral which, in situations in which the function of the translation is underspecified, can be an equally good solution. Other sources of (legitimate) variation in this example are the competing use of *Gesetzgeber vs. Gesetz* and the German rendering of *vie privee.* In translations number 3 and 4, the latter is actually omitted in favour of phrases which simply express the notion of importance. Only translation number 1 is marked as erroneous in KOPTE due to lexical and registerial inconsistencies.

(4) en Allemagne, ou la loi est particulierement protectrice pour la vie priv6e des citoyens
in Deutschland, wo die Rechtsprechung ganz besonders das Privatleben der Bürger schützt
In Deutschland gewährleisten Gesetze den Schutz der Privatsphäre eines jeden Bürgers
In Deutschland, wo das Datenschutzgesetz eine große Rolle spielt
Deutschland, wo Datenschutz großgeschrieben wird da hierzulande der Gesetzgeber besonders auf den Schutz der Privatsphäre seiner Bürger achtet

Example 5 illustrates translation Variation resulting from different strategies in dealing with source text elements that are untypical in the target language. Here, the open enumeration in parentheses can be considered at least unusual for German newspaper texts. However, some translators have decided to stick to the linguistic structure of the source text by adding the proper nouns in parentheses, while translator 4 has chosen a relative clause and the third translator just ignored the proper nouns in her translation, assuming that *PPR* includes the brands *Fnac* and *Gucci* (this omission is marked as an error in KOPTE). Note also the use of different variants for indicating the openness of the enumeration *(..., unter anderem, etc.)* and of different generic head nouns, which can be attached to the name *PPR* in different ways, for example, by means of compounding. All of these variants are correct.

(5) groupe PPR (Fnac, Gucci...)
Unternehmen PPR (Fnac, Gucci,...)
PPR-Konzern (Fnac, Gucci...)
Unternehmen PPR
Gruppe PPR, zu der unter anderem Fnac und Gucci gehören
Firma PPR (Fnac, Gucci, ...)
Konzern PPR (Fnac, Gucci etc.)

Example 6 illustrates the translation of proper nouns, more precisely, of a book title. While almost all translators decided to use the original title, just one considered it necessary to include also its translation. Translation number 2 featuring only a slightly diverging German translation of the book title has an error marker in KOPTE. The first translation includes a spelling error. The examples also illustrate strongly diverging variants for the translation of the year and overall different sentence structures, all of which are correct.

(6) La Terre vue du ciel, son best-seller de 1999
in seinem 1999 erschienen Bestseller "La terre vu du ciel"
Bestsellers aus dem Jahre 1999 "Die Erde aus Sicht des Himmels "
Bestsellers von 1999 „La Terre vue du ciel"
"La Terre vue du ciel", Y.A.B.'s Bestseller aus dem Jahr 1999
"La Terre vue du ciel" (Die Erde vom Himmel aus gesehen), dem Bestseller Arthus-Bertrands von 1999
„La Terre vue du ciel" beigetragen, Arthus-Bertrands Bestseller von 1999

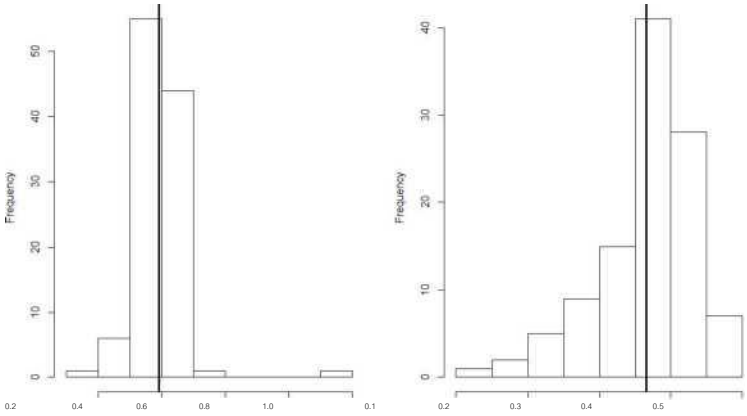

Figure 2: The distribution of the Meteor scores for the three lowest-graded and the the best-graded translations as references.

In order to solve translation problems arising from colloquial or figurative speech as given in Example 7 a translator requires very good knowledge not only of the target language, but also of the source language, including culturespecific knowledge, and a certain amount of creativity. A literal translation of the French example phrase would be *the resultsfrom the little brothers,* while the intended meaning is *earnings from merchandising products.* From that perspective, the solutions given by the human translators are, with two exceptions, all good. The first translation *Einnahmen der Vorgänger (earnings of the predecessor)* and *Verdienste zusätzlicher kleiner Artikel (income from additional small products)* are marked as translation errors in KOPTE, while translation number 4 has both a good solution *(Merchandising)* and a mistake (an incorrect preposition). As can be seen from the example, this kind of difficulty triggers heavy lexical variation in the translations.

(7) resultats des petits freres
Einnahmen der Vorgänger
Verdienste zusätzlicher kleiner Artikel
Einnahmen durch andere Produkte
Erlöse von Merchandising
Einnahmen aus dem Merchandising
Nebeneinkünfte

In Example 8, *je vis mal qu'on parle de fric* is a colloquial expression meaning *I don't like to talk about money* which, again, requires from the translator more than just proficiency in the source and target language, but in fact a creative solution which is not straightforward to come up with. The translation variants given below show that almost all translators understood the original phrase and tried to find the most appropriate solution such as *ich spreche nicht gern über Geld (I don't like talking about money).* Still, one translator (translator number 4) misunderstood the French phrase and rendered it as *ich verstehe nicht viel vom Geld (I don't know much about money).* This translation cannot be considered a valid translation variant. Again, a lot of lexical variation can be observed in this example.

(8) je vis mal qu'on parle de fric
ich spreche nicht gern über Geld
ich mag es nicht, wenn man vom Geld redet
für mich geht es nicht nur um Geld
ich verstehe nicht viel vom Geld
mir wird schlecht, wenn man von Geld spricht
das gefällt mir nicht, dass man von Kohle spricht

Units of measurement are also a source of translation variation. In Example 9 we observe that in the French source text the number appears as *100 000 euros,* whereas in the translations we have different variants ranging from *100 000 Euro* to *100 000€,* all being accepted variants. An additional phenomenon occurring in the example is synonymy in the translation of the French noun *amende (fine).* Here we can notice that all three German translations (*Geldstrafe*, *Strafe*, *Bußgeld*) of *amende* are good translations.

(9) une amende de 100 000 euros
Geldstrafe in Höhe von 100 000 Euro
Strafe von 100 000€
Geldstrafe von 100.000,- EUR
Geldstrafe in Höhe von 100.000 Euro
Bußgeld in Höhe von 100 000€

Example 10 also deals with numerals, this time related to age. The presented translation variants show that some translators adopted the original construction *Fotografen Yann Arthus-Bertrand, 63,* whereas other translators decided to change the structure of the phrase by putting the numeral in front of the proper name as *63jährigen Fotografen Yann Arthus-Betrand (63 years old photographer Yann Arthus-Betrand).* Other sources of variation can be found in the use of parentheses, various spelling variants for the German equivalent of *photographe* or of the generic noun *Jahre (years),* all of which are appropriate.

(10) photographe Yann Arthus-Bertrand, 63 ans
63jährigen Fotografen Yann Arthus-Betrand
Fotographen Yann Arthus-Bertrand (63 Jahre)
Fotografen Yann Arthus-Bertrand (63)
63-jährigen Fotografen Y.A.B.
Fotografen Yann Arthus-Bertrand, 63

In our analysis, source text elements that cannot be translated literally, but instead call for a creative solution, because of the lack of a direct German translation were classified as translation problems. Example 11 is such a phrase. The noun *pivot* meaning in this context *central figure* is combined with *l'influence frangaise (French influence)* into a phrase which, again, requires more than language proficiency from the translator. The presented translation variants are, with one exception, valid, one of them being, in fact, a very good solution (*Schlüsselfigur für den Einfluss Frankreichs (key figure ofthe French influence)).* The translation which cannot be considered valid is *Stützpunkt des Einflusses Frankreichs (the base for the influence of France)*. Again we see that this kind of difficulty triggers strong variation on the lexical level. We also observe some spelling variants.

(11) pivot de l'influence frangaise
Stützpunkt des Einflusses Frankreichs
zentralen Figur des französischen Einfluss
Stütze für den Einfluss Frankreichs
Schlüsselfigur für den Einfluss Frankreichs
Garant für den französischen Einfluß

Also a very difficult phrase to be translated is the French phrase in Example 12. Finding the best translation for this phrase is not straightforward and, in addition, requires culture-specific knowledge. From the five translations listed in Example 12 only two are valid, namely *Ältesten von Afrika* and *"Alten Herrn von Afrika"*, both of which, are, in fact, almost as opaque as the French originals. The fact that only two of five translators found a good solution for this phrase is an indicator of the difficulty of this kind of source text element.

(12) "doyen de l'Afrique"
obersten Würdenträgers Afrikas
"Alten Herrn von Afrika"
"Abtes von Afrika"
"Ältesten von Afrika"
"doyen de l'Afrique"

The examples in this section show that variation in translation can be caused by various source text elements. For some of these phenomena, some translators chose to add explanations, additional information, to adapt the perspective to the German target audience or to adapt the formatting of the enumeration, whereas other translators chose to translate literally. The examples also show that by far the bigger part of the observed variation is indeed legitimate or, in other words, variation in translation is not a sign of low quality, but a direct expression of the creative powers of natural language.

However, it is also obvious that - while many variants in our examples are correct and legitimate - not all are equally good. Best solutions for given problems are distributed unequally across the translations, but it is impossible to combine them into artificial "optimal" translations due to syntactic, grammatical, stylistic etc. constraints - language is not random. Moreover, extensive variation can also be found on the syntactic, but also the grammatical levels. For example, some translators chose to break the rather complicated syntax of the French original into simpler, easily readable sentences, producing, in some cases, considerable shifts in the information structure of the text - often a legitimate strategy. Considering the performance of the automatic scores, our study - that still calls for larger-scale and in-depth verification - suggests that neither BLEU nor Meteor are able to distinguish between *legitimate* and *illegitimate* variation. Thus, they overrate surface differences and thus assign very low scores to many translations that were found to be at least acceptable by a human expert. Furthermore, both scores failed to mimic the fine-grained quality distinctions made by the human expert - whether they can grasp more coarse-grained differences is still an open research question.

4. Conclusion

This paper presents a study on two different views on translation evaluation, one from the MT perspective and one from the perspective of translation studies. The goal of this paper was to bring these two disciplines together by investigating the behaviour of automatic evaluation metrics on a set of human translations from the KOPTE corpus. In Section 1 we concentrated on discussing various approaches to translation quality assessment. We discussed both the automatic evaluation metrics as well as the human evaluation of MT output as used in MT research. We also outlined fundamental notions of translation quality from the perspective of translation studies. In Section 2 we introduced the KOPTE corpus, more specifically the corpus design and the annotation of translation features and their evaluation. In Section 3 we described the experiments performed with BLEU and Meteor on our KOPTE sample as well as the results obtained from Kendall's rank correlation coefficient. The experiments suggest that both BLEU and Meteor systematically underestimate the quality of the translations tested and fail to provide meaningful evaluations in the sense understood by translation studies. A qualitative analysis of some of the evaluated translations supports our finding that lexical similarity scores are neither able to cope satisfactorily with standard lexical variation (paraphrases, synonymy) nor with dissimilarities that can be traced back to the source text or the nature of the translation process itself. Moreover, our results shed doubt on the concept of "reference translation", showing that automatic evaluation results tend to be dependent on some properties of the source text itself, e. g. the amount of constant elements (e. g. numbers or person names) that do not change upon translation.

5. Acknowledgement

This work has partially been supported by the CLARIN-D (Common Language Resources and Technology Infrastructure) project (http://de.clarin.eu).

6. References

Alexandra Birch, Barry Haddow, Ulrich Germann, Maria Nadejde, Christian Buck, and Philipp Koehn. 2013. The feasibility of HMEANT as a human MT evaluation metric. In *Proceedings of the 8th Workshop on SMT*, pages 52-61.

Ondrej Bojar, Christian Buck, Chris Callison-Burch, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Herv6 Saint-Amand, Radu Soricut, and Lucia Specia, editors. 2013. *Proceedings ofthe 8th Workshop on SMT*. ACL.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proceedings ofthe 2nd Workshop on SMT*, pages 136-158.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the 6th Workshop on SMT*, pages 85-91.

Deutsches Institut für Normung. 2006. *DIN EN 15038:2006-08: Übersetzungsdienstleistungen Dienstleistungsanforderungen.* Beuth.

DilekDizdar. 2003. Skopostheorie. In *Handbuch Translation*, pages 104-107. Stauffenburg.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings ofthe 2nd International Conference on HLT*, pages 138-145.

Christian Federmann. 2012. Appraise: An open-source toolkit for manual evaluation of machine translation output. *PBML*, 98:25-35, 9.

Brigitte Horn-Helf. 1999. *Technisches Übersetzen in Theorie und Praxis.* Franke.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707-710.

Chi-Kiu Lo and Dekai Wu. 2011. MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *Proceedings of the 49th Annual Meeting of the ACL*, pages 220-229.

Tucker Maney, Linda Sibert, Dennis Perzanowski, Kalyan Gupta, and Astrid Schmidt-Nielsen. 2012. Toward determining the comprehensibility of machine translations. In *Proceedings ofthe 1st PITR*, pages 1-7.

Elvira Mertin. 2006. *Prozessorientiertes Qualitätsmanagement im Dienstleistungsbereich Übersetzen. Peter* Lang.

Christiane Nord. 2003. Transparenz der Korrektur. In *Handbuch Translation*, pages 384-387. Stauffenburg.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting ofthe ACL*, pages 311-318.

Lothar Sachs and Jürgen Hedderich. 2009. *Angewandte Statistik. Methodensammlung mit R*. Springer.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA*, pages 223-231.

Society of Automotive Engineers. 2005. *SAE J2450:2005-08: Translation Quality Metric*. SAE.

Harold Somers. 2011. Machine translation: History, development, and limitations. In *The Oxford Handbook of Translation Studies*, pages 427-440. Oxford University Press.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Alexander Zubiaga, and Hassan Sawaf. 1997. Accelerated DP based search for statistical translation. In *Proceedings of the EUROSPEECH*, pages 2667-2670.

Sandra Weiss and Lars Ahrenberg. 2012. Error profiling for evaluation of machine-translated text: a polish-english case study. In *Proceedings ofthe Eighth LREC*, pages 1764-1770.

Andrea Wurm. 2013. Eigennamen und Realia in einem Korpus studentischer Übersetzungen (KOPTE). *transkom*, 6(2):381-419.

Assessing Inter-Annotator Agreement for Translation Error Annotation

Arle Lommel, Maja Popović, Aljoscha Burchardt
DFKI
Alt-Moabit 91c, 10559 Berlin, Germany
E-mail: arle.lommel@dfki.de, maja.popovic@dfki.de, aljoscha.burchardt@dfki.de

Abstract

One of the key requirements for demonstrating the validity and reliability of an assessment method is that annotators be able to apply it consistently. Automatic measures such as BLEU traditionally used to assess the quality of machine translation gain reliability by using human-generated reference translations under the assumption that mechanical similar to references is a valid measure of translation quality. Our experience with using detailed, in-line human-generated quality annotations as part of the QTLaunchPad project, however, shows that inter-annotator agreement (IAA) is relatively low, in part because humans differ in their understanding of quality problems, their causes, and the ways to fix them. This paper explores some of the facts that contribute to low IAA and suggests that these problems, rather than being a product of the specific annotation task, are likely to be endemic (although covert) in quality evaluation for both machine and human translation. Thus disagreement between annotators can help provide insight into how quality is understood.

Our examination found a number of factors that impact human identification and classification of errors. Particularly salient among these issues were: (1) disagreement as to the precise spans that contain an error; (2) errors whose categorization is unclear or ambiguous (i.e., ones where more than one issue type may apply), including those that can be described at different levels in the taxonomy of error classes used; (3) differences of opinion about whether something is or is not an error or how severe it is. These problems have helped us gain insight into how humans approach the error annotation process and have now resulted in changes to the instructions for annotators and the inclusion of improved decision-making tools with those instructions. Despite these improvements, however, we anticipate that issues resulting in disagreement between annotators will remain and are inherent in the quality assessment task.

Keywords: translation, quality, inter-annotator agreement

1. Introduction

The development and improvement of Machine Translation (MT) systems today makes heavy use of human knowledge and judgments about translation quality. Human insight is typically provided in one of four ways:

1. human-generated reference translations
2. rating of MT output based on perceived quality
3. post-edits of MT output (implicit error markup)
4. explicit error markup of MT output.

However, it is well known that human judgments of translation show a high degree of variance: in WMT testing, the inter-annotator agreement (IAA), i.e., agreement between two or more annotators, in a rating task did not exceed 0.40 ($\varkappa$, described in section 3) and *intra*-annotator agreement (i.e., the agreement of raters *with themselves* when faced with the same assessment task multiple times) did not exceed 0.65 (Bojar et al., 2013:6–8). By contrast, for most IAA tasks, agreement of at least 0.85 is required for a measure to be considered reliable.

It must be put forth as a fundamental assumption that there is no single, objectively "correct" translation for a given text, but rather a range of possible translations that range from perfectly acceptable to totally unacceptable. Moreover, Translation quality is always relative to given specifications or the given job. Factors like resource availability, production environment, target audience, etc. can determine whether a certain translation is considered correct or not. For example, in an on-demand instant MT system, quality may be determined by whether or not the text enables the reader to accomplish a task. In such cases texts may show low levels of Accuracy and Fluency and yet still be considered to meet quality expectations.[1] Although we will not discuss this issue in depth in this paper, it should be kept in mind.

The realization that there is a spectrum of acceptable translations rather than a single optimal output and that raters will often disagree in their opinions are reasons why automatic measures of MT quality like BLEU have been designed to be able compare MT output with multiple human translation references from the very beginning (Papimeni et al. 2002).

Considering the four types of human insight listed at the start of this paper, the question of inter-annotator agreement boils down, in part to questions such as: How similar are two or more human reference translations? How similar are ratings? How similar are post-edits? How similar are explicit error markups? In all of these cases, any subsequent experiments using performance measures like BLEU or METEOR or analysis tools like Hjerson (Popović 2011) rely on the assumption that the human input provides a reliable basis.

To the best of our knowledge, the question of how many reference translations, ratings, or post-edits are needed

[1] As a result of this realization, there has been a recent shift towards the use of explicit specifications that guide translation, assessment, and postediting (Melby, Fields, & Housley, 2014).

per sentence to substantiate reliable and replicable quality judgments about MT performance has not yet found a widely accepted answer. In this paper, we will report first steps in evaluating inter-annotator agreement for the case of explicit error markup.

As MT errors can overlap or interact in many ways, we will focus on machine translations that show only few errors to minimize the problem of overlapping errors.

One reason for human disagreement in the case of analysis based on post-edits or manual error annotation is the simple fact that errors can often be analyzed (or explained) in multiple ways. For example, a seemingly missing plural *-s* in an English noun phrase might constitute an *agreement* error (*Fluency*) or indicate a *mistranslation* of a noun, which was meant to be singular (*Accuracy*). When translating from Chinese, for example, such factors may lead to different opinions of human translators since Chinese does not mark number; such confusion is likely inherent in the task since there are multiple valid ways to understand an error.

The remainder of this paper will focus on some of the issues that complicate the determination of IAA with examples from a human annotation campaign undertaken by the QTLaunchPad project.

2. Experimental setup

In the annotations described in this paper multiple professional translators from commercial language service providers (LSPs) were asked to evaluate a set of 150 sentences in one of four language pairs (EN>ES, ES>EN, EN>DE, and DE>EN) using the open-source translate5 (http://www.translate5.net) tool.

The sentences were selected from the WMT 2012 shared task data produced by state-of-the-art MT systems. The sentences were selected so that only those with a "native" source were used (i.e., only those sentences where the source segment had been written in the source language rather than translated from another language).[2] To select the sentences for annotation, human evaluators reviewed the MT output for the 500 translations of each of the systems—SMT, RbMT, and (for English source only) hybrid—plus the 500 reference human translations. These reviewers ranked each translated segment according to the following scale:

- Rank 1: Perfect output (no edits needed)
- Rank 2: "Near misses" (1–3 edits needed to be acceptable)
- Rank 3: "Bad" (>3 edits needed)

From the Rank 2 sentences, we pseudo-randomly selected a corpus of 150 sentences, to create the "calibration set." The calibration set consisted of the following breakdown of segments by production type:

- EN>ES and EN>DE: 40 segments each SMT, RbMT, and hybrid, plus 40 human translations.
- ES>EN and DE>EN: 60 segments each SMT and RbMT, plus 40 human translations.

These corpora were uploaded into the translate5 system and the annotators were all provided with a set of written guidelines[3] and invited to attend or view a recording of a webinar[4] introducing them to the tool and task.

The segments were annotated by three (DE>EN), four (EN>ES, ES>EN), or five (EN>DE) annotators. Annotators were encouraged to interact with our team and to ask questions. The annotators used translate5 to associate issues with specific spans in target sentences. The list of issues used was the following:

- Accuracy
  - Terminology
  - Mistranslation
  - Omission
  - Addition
  - Untranslated
- Fluency
  - Register/Style
  - Spelling
    - Capitalization
  - Typography
    - Punctuation
  - Grammar
    - Morphology (word form)
    - Part of speech
    - Agreement
    - Word order
    - Function words
    - Tense/mood/aspect
  - Unintelligible

The definitions for each of these issues are provided in the downloadable guidelines previously mentioned. Annotators were instructed to select "minimal" spans (i.e., the shortest span that contains the issue) and to add comments to explain their choices, where relevant.

[2] WMT data includes both sentences written in the source language and those translated into the source language from another language.

[3] http://www.qt21.eu/downloads/shared-task-webinar.mov

[4] http://www.qt21.eu/downloads/webinar-on-shared-task.pdf

Annotators found the numbers of issues given in Table 1.

	ES>EN	EN>ES	DE>EN	EN>DE	All
Annot. 1	157	387	219	216	—
Annot. 2	229	281	266	278	—
Annot. 3	98	289	327	277	—
Annot. 4	255	235	—	315	—
Annot. 5	—	—	—	278	—
TOTAL	739	1192	812	1364	4107
AVG	185	298	271	273	257
AVG/Seg	1.23	1.99	1.80	1.82	1.71

Table 1: Number of issues found in corpus per annotator and language pair.

The distribution of identified issues in this corpus is described in Burchardt et al. (2014) and is not covered here, as the analysis of specific issue types and their distribution is beyond the scope of this paper.

3. Assessing Inter-Annotator Agreement

As part of the evaluation of the results of the annotation task, we wished to determine inter-annotator agreement (IAA), sometimes known as inter-rater reliability. Demonstrating a high degree of IAA is a necessary step to showing that an assessment metric is reliable. In addition, demonstrating reliability helps, but is not sufficient, to demonstrate that a metric is fair.

There are a number of different approaches to demonstrating IAA. One approach is to look at absolute agreement between raters. This approach typically overstates agreement, however, because it does not take into account the probability of agreement by chance. For example, if items are assessed on a 1 to 5 scale with an equal distribution between each of the points on the scale, an assessment that *randomly* assigns scores to each item would achieve an absolute agreement approaching 0.2 (i.e., 20% of numbers would agree) as the sample size approaches infinity. As a result, for many tasks a different measure, Cohen's kappa ($\varkappa$)[5] is preferable because it attempts to take the probability of random agreement into account, although the assumption that annotators will make random choices in the absence of a clear option is debatable, a point we will return to, so $\varkappa$ scores may *understate* agreement (Uebersax, 1987). Nevertheless, in order to provide comparison with assessments of IAA given in WMT results, this study uses $\varkappa$ scores.

To calculate scores, we examined the positional tagging for issues in pairwise comparisons between each annotator, averaging the results within each language pair. Figure 1 shows an example in which one annotator tagged two issues and the second tagged one. In the last row the lighter cells show the area of disagreement.

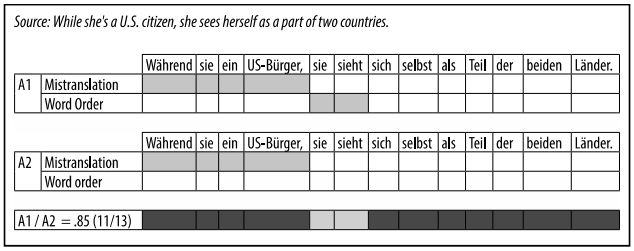

Figure 1: IAA for an English>German translation (absolute agreement average = .85, Kappa IAA = .72)

Because $\varkappa$ is appropriate only for pair-wise comparisons, we evaluated the similarity between each pair of annotators separately and took the average score, as shown in Figure 2. In this example three different IAA figures are assessed, one for each of the three possible pair-wise comparisons. In this example, Rater 1 and Rater 3 are quite similar with $\varkappa$ = 0.89 while Rater 2 differs from both of them with $\varkappa$ = 0.57 with Rater 1 and $\varkappa$ = 0.53 with Rater 3. Although both Rater 2 and Rater 3 identified the same types of errors (and were alike in not identifying the Agreement error identified by Rater 1), they disagreed on the precise spans for those errors, leading to lower $\varkappa$ scores.

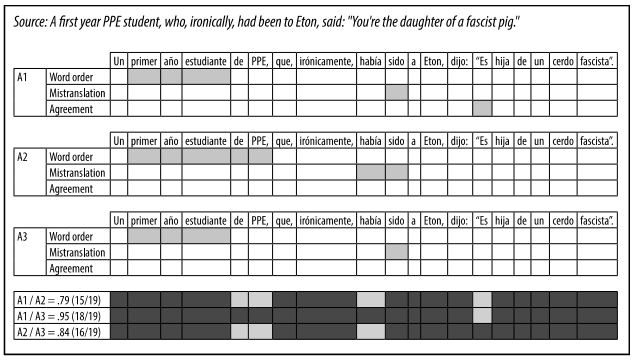

Figure 2: IAA for an English>Spanish translation (absolute agreement average = .86, Kappa IAA = .66)

Note that if a simpler segment-level measure that counts only whether the same issue classes were identified for each segment were used instead, the results would be rather different. In that case the example in Figure 1 would yield an agreement figure of 0.5 (there would be a total of two issues for the segment and the annotators would agree on one). For the example in Figure 2, by contrast, Rater 1 would show the same agreement with Raters 2 and 3 (.67) while Rater 2 and Rater 3 would show perfect agreement (1.0) since they identified the same issues, even though they disagreed on the scope.

[5] Kappa is calculated as follows:

$$\varkappa = \frac{P(a)-P(e)}{1-P(e)}$$

where P(a) is probability of actual agreement, i.e., $\sum_k p(a1 = k, a2 = k)$ and P(e) is probability of agreement by chance, i.e., $\sum_k p(a1 = k) * p(a2 = k)$, where k denotes class (in this case the error tag) and a1, a2 refer to the two annotators.

Using kappa allowed us to calculate $\varkappa$ scores for the test data sets, as shown in Table 2. (The EN>DE pair was annotated by five reviewers, but one was received after this analysis was completed.) The results of this analysis lie between 0.18 and 0.36 and are considered to be "fair" (see Bojar et al., 2013:6–8, for discussion of $\varkappa$ levels). The overall average is 0.30.

	ES>EN	EN>ES	DE>EN	EN>DE
a1-a2	0.30	0.35	0.23	0.36
a1-a3	0.18	0.36	0.36	0.28
a2-a3	0.19	0.28	0.29	0.33
a1-a4	0.25	0.33		0.30
a2-a4	0.26	0.36		0.34
a3-a4	0.34	0.35		0.30
Average	0.25	0.34	0.29	0.32

Table 2: Kappa coefficients measuring inter-annotator agreement for MQM error annotation

By comparison, the WMT organizers evaluated $\varkappa$ scores for their rating tasks in which raters were asked to assign quality rates from 1–4 (2011/2012) or 1–5 (2013). The $\varkappa$ scores for this task are presented in Table 3.

	ES>EN	EN>ES	DE>EN	EN>DE
WMT 2011	0.38	0.37	0.32	0.49
WMT 2012	0.36	0.25	0.38	0.30
WMT 2013	0.46	0.33	0.44	0.42
Average	0.40	0.32	0.38	0.40

Table 3: $\varkappa$ scores for WMT ranking tasks.

As can be seen, the IAA scores for the human annotation task are lower than those for the rating task, with the highest scores in the annotation task roughly on par with the lowest scores in the rating task. While such a result might seem discouraging, we believe there are a number of reasons for this result and that our seemingly low results may reveal problems hidden in many translation quality assessment tasks/methods. The remainder of this paper will address some of these results.

4. Scope of span-level annotation

One fundamental issue that the QTLaunchPad annotation encountered was disagreement about the precise scope of errors. In the example shown in Figure 2, for instance, Annotator 1 marked the following issue spans:

> Un **primer año estudiante** de PPE, que, irónicamente, había **sido** a Eton, dijo: "**Es** hija de un cerdo fascista".

while Annotator 2 marked the following spans:

> Un **primer año estudiante de PPE**, que, irónicamente, **había sido** a Eton, dijo: "Es hija de un cerdo fascista".

Here they fundamentally agree on two issues (a *Word order* and a *Mistranslation*) and disagree on the third (an *Agreement*). However, for the two issues they agree on, they disagree on the span that they cover. Annotators were asked to mark *minimal* spans, i.e., spans that covered *only* the issue in question, but they frequently disagreed as to what the scope of these issues was.

In the case of *primer año estudiante* vs. *primer año estudiante de PPE*, two word orders are equally acceptable: *estudiante de primer año de PPE* and *estudiante de PPE de primer año.* Thus it seems that the reviewers agreed that the phrase *(Un) primer año estudiante de PPE* was problematic, but disagreed as to the solution and whether *de PPE* needed to be moved or not.

In the case of *sido* vs. *había sido*, the correct rendering would be *había ido* ('had gone') instead of *había sido* 'had been'. Annotator 1 thus correctly annotated the minimal span, while Annotator 2 annotated a longer span. However, it may be that the two reviewers perceived the issue differently and that the cognitively relevant span for Annotator 1 was the word *sido* while for Annotator 2 it was the entire verbal unit, *había sido*.

In these two cases we see that reviewers can agree on the nature (and categorization) of issues and yet still disagree on their precise span-level location. In some instances this disagreement may reflect differing ideas about optimal solutions, as in the case of whether to include *de PPE* in the *Word order* error. In others the problem may have more to do with perceptual units in the text.

In such cases we are uncertain how best to assess IAA. Using the model presented in the previous section these are marked as agreement for some words and disagreement for others. The net effect is that, at the sentence level, they have partial agreement.

	ES>EN	EN>ES	DE>EN	EN>DE
Accuracy	0.0%	0.1%	0.0%	0.4%
Addition	0.5%	1.3%	0.4%	2.2%
Agreement	0.4%	2.8%	0.3%	1.4%
Capitalization	0.0%	0.6%	0.3%	0.3%
Fluency	0.0%	0.0%	0.0%	0.0%
Function words	9.2%	10.1%	4.1%	1.9%
Grammar	3.0%	0.3%	0.1%	9.5%
Mistranslation	6.4%	6.9%	4.4%	8.0%
Morphology	0.0%	0.1%	1.0%	0.1%
POS	1.1%	0.5%	1.2%	0.0%
Punctuation	2.0%	0.7%	1.2%	1.5%
Spelling	0.4%	0.6%	0.1%	0.2%
Style/Register	7.1%	7.4%	3.8%	6.3%
Tense/Aspect/Mood	1.6%	4.4%	0.5%	2.3%
Terminology	6.3%	14.2%	8.9%	2.8%
Typography	0.0%	0.4%	0.0%	0.0%
Unintelligible	0.1%	0.0%	1.7%	1.2%
Untranslated	0.3%	0.0%	0.3%	0.5%
Word order	8.0%	10.1%	24.2%	6.1%

Table 4: Percentage of instances for each issue class in which annotators disagreed on precise spans.

Quantitatively, the impact of different assessments of spans can be see in Table 4, which shows, based on a pairwise comparison of annotators, the percentage of cases in which annotators differ in their assessment of the location (but not the nature) of spans. Note that this analysis does not distinguish between cases where spans are actually related and where they are independent instances of the same category (e.g., two annotators annotate totally different *Mistranslations* in a segment), so it may overstate the numbers slightly.

In this case it can be seen that *Word order* has the highest overall rate of instances in which annotators disagreed on the precise location of spans. From other analysis done in the QTLaunchPad project we know that word order is particularly problematic for German>English translations, and here we see a high confusion rate for this issue type. It is not surprising that *Word order* ranks so highly in terms of confusion because often there are different ways to interpret ordering errors. So even though annotators largely agree on the existence of the problem, they often disagree on the location.

5. Unclear error categorization

In the example discussed in the last section, one item was tagged by Annotator 1 and missed by other reviewers. Annotator 1 tagged it as *Agreement*, but a close examination of the issue leaves it unclear why *Agreement* was chosen. The use of *Es* is clearly a mistake since it cannot generally mean "You're". After consulting with a Spanish native speaker, it appears that the error should definitely have been tagged (so two of three reviewers missed the problem) but that there are multiple possible categorizations depending on how *You're* should be rendered in Spanish. Possible options include the following:

- **Mistranslation**. *Es* 'is' clearly not the intended meaning. *Es* can thus be treated as a mistranslation for *Tu eres* or *Eres* 'You (informal) are'.
- **Omission**. If a formal register is intended (an unlikely choice for a human translator in this case, but possible since MT systems might be optimized to usw the formal), then *Usted es* would be the appropriate text, and there would be an *Omission* of *Usted*.
- **Agreement**. Since Spanish can, in most circumstances, drop subject pronouns (although, generally, *Usted* should not be tropped), *Es* could exhibit an *Agreement* problem with the implicit subject *tu.*

Since the exact nature of the problem is not clear from the text (source or target), the rules used in the annotation task would specify that the first possible one in the list of issues be taken, unless a more specific type also applies. In this case, then, *Mistranslation* would be the appropriate issue type. However, if the annotator did not perceive the phrase as having the wrong meaning, but rather an awkward phrasing, then the annotator might never arrive at this option.

	ES>EN	EN>ES	DE>EN	EN>DE
Accuracy	0.2%	0.2%	0%	1.0%
Addition	2.1%	4.8%	4.0%	3.5%
Agreement	6.2%	7.3%	3.6%	4.7%
Capitalization	0.3%	2.9%	1.1%	1.2%
Fluency	0%	3.0%	0%	0.2%
Function words	30.4%	21.9%	18.9%	7.6%
Grammar	6.3%	1.0%	0.7%	16.8%
Mistranslation	23.6%	22.8%	27.1%	24.1%
Morphology	0.2%	0.3%	3.8%	5.4%
Omission	5.3%	6.6%	7.6%	5.2%
POS	2.9%	2.2%	2.4%	1.0%
Punctuation	4.0%	4.5%	9.1%	9.3%
Spelling	0.8%	2.2%	1.1%	0.9%
Style/Register	16.3%	9.1%	3.3%	11.0%
Tense/Aspect/Mood	3.9%	11.3%	3.1%	7.1%
Terminology	12.9%	24.5%	19.1%	13.1%
Typography	0.2%	0.8%	0.2%	0.0%
Unintelligible	0.2%	0.0%	1.3%	1.2%
Untranslated	0.9%	0.9%	0.9%	0.5%
Word order	7.2%	5.9%	8.9%	4.4%
No error	15.4%	10.4%	7.6%	8.7%

Table 5: Percentage of instances at the sentence level in which one annotator noted an issue and another annotator did not, by language pair.

The problem of differing assessments of the nature of problems is pervasive in our corpus, as shown in Table 5, which provides the percentage of times in which one annotator marked a sentence as having a specific issue and another annotator did not mark that same issue type as occurring within the sentence.

The figures in Table 5 were derived in pair-wise comparisons between annotators. For each case if one annotator noted a specific class of issue, regardless of location within the segment, and another annotator also annotated the same issue class as occurring in the segment, the annotators were deemed to be in agreement. If one annotator noted an issue class and another did not then they were deemed to be in disagreement. This provides a rough measure for the frequency with which the issues might be annotated in different ways. Examining the totals reveals the following notable points:

- *Mistranslation* and *Terminology* show high levels of confusion. (Burchardt et al. 2014 discusses the confusion between these two and the correlation with the length of the problematic span, with *Mistranslation* being used for longer spans in general while *Terminology* is used primarily for single-word spans.)
- The *Function word* category also shows very high confusion, with very different profiles between the language paits. Overall, this category was one of the most frequently occurring and problematic in the entire corpus.
- *Word order* shows high levels of agreement between annotators, although the span-level agreement is significantly lower.
- There is a relatively high percentage of sentences where some annotators say that there are no errors and other annotators say there are some errors.

5.1. Confusion within the hierarchy

It is important to note that the MQM issues exist in a hierarchy and the annotators were instructed that if no issue applied precisely at one level in the hierarchy, they should select the next highest level. As a result, annotators may be confused about which class applies to a specific error and find the issue types confusing. When we ran the annotation campaign a number of annotators came back to us with cases where they were unsure as to which level was appropriate for a given issue.

For example, if the annotator encountered a grammatical error but none of the children of *Grammar* applied (e.g., a sentence has a phrase like "he slept the baby" in which an intransitive very is used transitively, but there is no precise category for this error, which is known as a *Valency* error), then the parent (*Grammar* should be used). As a result, many issues could be annotated at multiple levels in the hierarchy, especially if the precise nature of the error is not entirely clear, as with the example given above, where it *could* be *Agreement*, but is not clearly so. In such cases some annotators may pick a specific category, especially if they feel comfortable with the category, while others may take the more general category in order to be safe in a situation where they are not certain.

5.2. Lack of clear decision tools

One of the problems annotators faced was the mismatch in knowledge between the team that created the MQM metric and themselves. Many of the training materials we created assumed a certain degree of background knowledge in linguistics that it turned out we could not assume. A simple list of issue types and definitions along with some general guidelines were insufficient to guide annotators when faced with unfamiliar issues which they intuitively know how to fix but which they are not use to classifying in an analytic scheme.

As a result, we discovered that annotators need better decision-making tools to guide them in selecting issues, especially when they are easily confusable, as is the case with issues in the hierarchy, or when there are multiple, equally plausible explanations for an error. By formalizing and proceduralizing the decision-making process, confusion could be reduced.

6. Annotators' personal opinions

Finally, we cannot discount the possibility that different translators may simply disagree as to whether something constitutes an error or not, based on dialect, ideolect, education, or even personal opinion. Such cases, where one speaker of a language sees a sentence as acceptable and another does not, have long been the bane of linguistics professors who want to have a clear-cut case for putting a star (*) on unacceptable sentences. In addition, although we provided the annotators with detailed guidelines for the issue types, they may have disagreed as to whether something was serious enough to annotate. Thus the individual annotators' opinions are likely to have a substantial impact on overall IAA, albeit one hard to quantify without an extensive qualitative consultation with annotators in a lab setting.

7. Lessons learned and conclusions

Analytic measures like MQM offer the potential to gain insights into the causes of translation problems and how to resolve them. Although IAA in our first studies reported here is lower than ideal, we believe that our findings point out a covert problem in most annotation and quality evaluation tasks,. As we discovered, the human annotators' meta-understanding of language is quite variable, even when working with professional translators. Even with an analytic framework and guidelines there is significant, and perhaps unavoidable, disagreement between annotators. To a large extent this disagreement reflects the variability of human language.

Evaluation methods that rely on reference translations such as METEOR or BLEU must assume that the range of available translations provides a "good enough" approximation of the range of language variation that can be expected in translations. This assumption may be valid for limited cases in which the training data used for MT is substantially similar to the reference translations, but in cases with heterogeneous training data and references, it is entirely possible that reference-based methods may penalize acceptable translations because they differ from references and reward less optimal answers because they are mechanically similar to references.

In order to improve future MQM assessment and improve IAA rates, we have revised the issue hierarchy to reduce certain distinctions (e.g., between *Typography* and *Punctuation*) that offered little discriminatory power, while we have added more details to other categories such as splitting the *Function words* category to allow more detailed analysis of specific problems. We have also created a formal decision tree and improved guidelines[6] to assist with future annotation work and to help annotators distinguish between problematic categories.

While improved IAA is an important goal where possible, the exact nature of disagreement where clarification of issue types and procedures does not result in agreement can also provide insight into how humans conceive of translation quality. If one of the goals of MT research is to deliver translation closer to human quality, a better understanding of the variables that impact quality judgments is vital, as is understanding the extent of variation that comprises "acceptable" translation. This study and the issues it raises will help provide better understanding of these factors. We intend to continue this analysis using our improved issue hierarchy and decision tools in an annotation campaign planned for March 2014.

8. Acknowledgements

This study was conducted as part of the European Commission-funded project QTLaunchPad (Seventh Framework Programme (FP7), grant number: 296347). The authors thank Ergun Bicici (Dublin City University) for assistance with calculating IAA scores. Thanks also to Nieves Sande (DFKI) for her kind assistance with Spanish-language issues.

9. 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, pp 1–44. (http://www.statmt.org/wmt13/pdf/WMT01.pdf.)

Burchardt, A.; Gaspari, F.; Lommel, A.; Popović, M. and Toral, A. (2014). Barriers for High-Quality Machine Translation (QTLaunchPad Deliverable 1.3.1). http://www.qt21.eu/launchpad/system/files/deliverables/QTLP-Deliverable-1_3_1.pdf.

Melby, A.; Fields, P.J. and Housley, J. (forthcoming). Assessment of Post-Editing via Structured Translation Specifications To appear in M. Carl, S. O'Brien, M. Simard, L. Specia, & L. Winther-Balling (Eds.), *Post-editing of Machine Translation: Processes and Applications*. Cambridge: Cambridge Scholars Publishing.

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 on Association for Computational Linguistics.* Stroudsburg, PA: Association for Computational Linguistics, pp. 311–18.

Popović, M. (2011). Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *Prague Bulletin of Mathematical Linguistics* 96, pp. 59–68.

Uebersax, J.S. (1987). Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin* 101, 140–46.

[6] http://www.qt21.eu/downloads/annotatorsGuidelinesNew.pdf

TURKOISE: a Mechanical Turk-based Tailor-made Metric for Spoken Language Translation Systems in the Medical Domain

Marianne Starlander

University of Geneva, FTI-TIM, 40 Bd du Pont d'Arve, CH-1211 Genève 4

E-mail: Marianne.Starlander@unige.ch

Abstract

In this paper, we will focus on the evaluation of MedSLT, a medium-vocabulary hybrid speech translation system intended to support medical diagnosis dialogues between a physician and a patient who do not share a common language (Bouillon et al, 2005). How can the developers be sure of delivering good translation quality to their users, in a domain where reliability is of the highest importance?

With MedSLT sentences are usually translated freely and, as a consequence of spoken input, they are often short. These characteristics entail low BLEU scores (Starlander and Estrella, 2011) as well as poor correlation when using human judgments.

In the present paper we will describe the path that led us to using Amazon Mechanical Turk (AMT) as an alternative to more classical automatic or human evaluation, and introduce task-specific human metric, TURKOISE, designed to be used by unskilled AMT evaluators while guaranteeing reasonable level of coherence between the evaluators.

Keywords: spoken language translation, evaluation, crowdsourcing

1. Introduction

Speech recognition and machine translation are now widely available on laptops and mobile devices: typical examples are the speech-enabled Google Translate mobile application and Jibbigo (recently acquired by Facebook). A more specialized system is the US military application, now also running on smart-phones (Weiss et al, 2011). These technologies can be of use in many situations, especially when fast, low-cost translations are required. In this paper, we will investigate the application of the above technologies for medical communication purposes. We will focus on the evaluation of MedSLT, a medium-vocabulary hybrid speech translation system intended to support medical diagnosis dialogues between a physician and a patient who do not share a common language (Bouillon et al, 2005). The central question faced by the developers is to deliver good translation quality to their users, in a domain where reliability is of the highest importance. Here, we will consider the question of determining if the translations provided by the system are suitable for the task: enabling communication between the physician and his patient without generating ambiguity or errors that could potentially endanger the patient.

In previous research we have evaluated the usability, translation quality and recognition quality of MedSLT (Starlander and Estrella, 2009). This research has revealed discrepancies between usability measures, automatic measures and human evaluation of translation and recognition quality. This made us further investigate the question of translation quality in (Starlander and Estrella, 2011) in order to develop an evaluation method equally suitable both for rule-based spoken language translation (SLT) systems such as MedSLT, and for SMT based systems. One of the findings of the study was that a classic metric like BLEU is not well-suited for the evaluation of MedSLT output, due to the architecture of the system. The problem is that MedSLT was designed with a strong focus on reliability in correct transmission of the message. One of the characteristics of MedSLT is its rule-based architecture, which uses an interlingua approach to produce highly reliable output. This approach discards surface form to keep only the meaning of the sentences. Consequently, sentences are translated freely; for example, "Do you have a sore throat" is translated as "Le duele la garganta" (closer to "Does your throat hurt") instead of the more literal "Tiene dolor de garganta". Due to these characteristics, and also to the fact that our sentences are short (10% of the corpus consists of sentences counting less than 4 words), the BLEU scores obtained in (Starlander and Estrella, 2011) were low, and did not correlate well with human judgment. This concurs with the common opinion in the MT literature (Callison-Burch et al, 2006, Popescu-Belis et al, 2004) that automatic metrics like BLEU are often not well suited for rule-based machine translation (RBMT) systems, given that they tend to reward translations that are literal and close to a given reference. The results of our experiments are presented in the following section. In section 4, we describe the results obtained when using AMT. Indeed, despite the increasing use of AMT – in a great variety of fields –, the question of the reliability of this type of evaluation compared with a small amount of expert evaluators still remains open, especially with the kind of complex task (requiring bilingual evaluators) we are submitting. On the other hand, AMT workers might arguably be more appropriate judges of translation quality, as they are closer to the real users of the system (on patient side) and are less likely to focus on the linguistic form than translators.

Our study focuses on inter-rater agreement, comparing this statistic for our small in-house group of translator-

evaluators against a wider group of AMT workers. We also quantify the effort involved in running the AMT evaluation in order to compare the resources needed. Developers and researchers tend to minimize the effort related to the creation of the reference translations needed in order to use BLEU or other reference-based metrics. If AMT workers are found to be reliable, we argue that this type of evaluation is at least as cost- and time-effective as classical automatic metrics, while also providing the advantage of reflecting the end-user's quality level request.

2. Experiment

This experiment retraces the road from the classic human metrics, to a tailor made human metric and finishes by testing our metric using non-expert evaluators in order to compare our proposed 4-point scale with the traditional 5-point scale fluency/adequacy implemented both by a small group of selected in-house-evaluators and by AMT workers.

We had envisaged experimenting with quality estimation (QE) methods such as described in Specia et al (2010) but our corpus being small, we decided to carry out a first trial with AMT before utterly pursuing in the QE direction.

2.1 Data

Our data set is composed of utterances collected during a test-phase in 2008, where we simulated medical diagnosis dialogues with English-speaking physicians and Spanish-speaking standardised patients at the Dallas Children's Hospital using our RBMT system MedSLT. The total corpus is made of approximately 1200 utterances, but the test corpus for the actual study consists of an excerpt of approximately 220 English to Spanish translations. We removed sentences that were too short to be of real interest for the task, hence ruling out all the one-word utterances (mainly *yes, no, un poco, mucho*...) and we only kept the types (removing all multiple occurrence of a sentence and translation.

2.2 Evaluation tasks

The current experiment was divided into the following three tasks:

Monolingual evaluation task:

- fluency of translation results from English to Spanish (Flawless, Good, Non-native, Disfluent, Incomprehensible Spanish)

Bilingual evaluation tasks:

- adequacy of English-Spanish translations using a 5-point scale (All information is present, Almost all information is present, A lot of information is present, Only a little information is present, No information is present)
- rating of English-Spanish translations using our custom 4-point scale (Starlander 2009) as described in section 4.

For these evaluation tasks we used the classical fluency and adequacy human metrics, as well as our own task-centred human metric for the medical domain (Starlander, 2009). This metric relies on human judgements and uses a 4-point scale:

CCOR (4): The translation is completely correct. All the meaning from the source is present in the target sentence.

MEAN (3): The translation is not completely correct. The meaning is slightly different but it represents no danger of miscommunication between doctor and patient.

NONS (2): This translation doesn't make any sense, it is gibberish. This translation is not correct in the target language.

DANG (1): This translation is incorrect and the meaning in the target and source are very different. It is a false sense, dangerous for communication between doctor and patient.

The two categories NONS and DANG might seem similar, but the un-negligible difference between these two categories is that a sentence from the category NONS is much easier identified by the end users (being gibberish or at least incorrect target language); while as a sentence from the category DANG would be a perfectly well built sentence but where the original meaning has been altered in such a way that the translation could represent a danger for the patient using the SLT system. In the table below we provide some evaluation examples using the above-defined scale.

Source	Target	E1	E2	E3
Are you having fever?	¿El dolor está aliviado cuando tiene fiebre?	1	1	1
Did you see a doctor this week?	¿Ha consultado un médico esta semana?	4	4	4
Do you have a headache?	¿Tiene tos ayer?	1	2	1

Table 1: Examples of the 4-point scale application.

Our tailor-made 4-point scale for SLT in safety-critical domain had so far been used only by very limited numbers of evaluators.

In section 3, we will describe the results obtained by both evaluation groups: our group of in-house translators and the AMT workers. We will analyse the resulting AMT evaluations under the following aspects. First, we aim to determine which of the different scales leads to the best inter-annotator agreement. Then, we will observe how evaluations by AMT workers relate to evaluations by in-house translators for this domain, and finally in section 3.4 we will study how the Kappa evolves according to the number of total evaluation assignments.

Finally, we hope to provide a better insight on how to use AMT on this type of evaluation task. We will conclude on the potential of TURKOISE as an alternative to automatic metric or classical human judgement limited to a few expert evaluators.

2.3 Participants

Classically, when launching a human evaluation, the problem is to find suitable evaluators. In our particular context, at the Faculty of Translation and Interpretation of the University of Geneva, it is not a too difficult task to find freshly graduated translators, final year students or fellow translators willing to participate for free at evaluation tasks like ours. This however is only true in our particular context, in "real life" finding enough evaluators can soon turn out to be a very time and money consuming task if not a total nightmare. We have always been able to recruit, indeed, relatively small groups of translators and non-translators, but answering fast (most of the time within 24h).

To investigate another evaluation approach not specific to our context, we have chosen to submit the same tasks to workers recruited on Amazon Mechanical Turk (AMT). The idea behind extending this human evaluation to AMT workers is of course to offer human evaluation but in a faster and cheaper way (Callison-Burch et al., 2009) as has been done at the IWSLT campaigns since 2010 (Paul et al, 2010 and Federico et al; 2012), but also to extend the already wide usage of AMT to an even larger variety of tasks. AMT workers have been involved in tasks ranging from labelling to transcription and have recently moved to spoken dialogue systems evaluation (Jurcícek et al, 2011) with success. Another incentive to experiment with AMT is our interest in comparing the results obtained with evaluators of different backgrounds. Indeed, we have observed that translators tend to focus on the form rather than on the meaning of the provided sentences, which is not the most relevant aspect in our context. Arguably, the AMT are a closer population to the real end-users of such systems as MedSLT. As a consequence of the wider usage of AMT workers for an increasing amount of tasks and resource creation, criticism has risen concerning these practices (Fort et al, 2011). The question of low payment and discrepancies in the quality of the tasks fulfilled are surely topics that deserve discussion. On the first topic, I would like to point out that most of our in-house evaluations have been done in a benevolent manner, in the spirit of helping out fellow translators or researchers. In future work, once a minimum of reliability from the AMT workers could be established for the proposed task in this paper, we would certainly like to investigate more about the ethical and sociological impact of AMT on the research. But at present, we will present the characteristics of the participants to our current experiment.

2.3.1 In-house translators

For the fluency and adequacy evaluation task, we recruited five in-house trained English to Spanish translators. They completed the evaluation in Microsoft Excel spreadsheets. They first graded the fluency by reading only the Spanish target and second, they rated the adequacy of the translations using the classical 5-point scale.

The evaluation using the above described 4-point scale (CCOR, MEAN, NONS, DANG) had been done in previous research. In the first stage, we asked the in-house interpreters of the Dallas Children's Hospital that participated to the data collection in 2008 to evaluate by email a series of sentences (190 per evaluator). This could somehow be compared to an AMT setting, since we did not know these translators, and also because they did not evaluate all the sentences but only a subset. We had divided the data in such a way that in the end we should have obtained five human evaluations for each sentence. Atthe same time, we asked these participants to provide a reference translation on a subset of the evaluated translations. We succeeded in gathering three reference translations for each sentence but managing the translators for the evaluation and the production of reference translation was a very time consuming task.

In a second stage, we asked known in-house translators and finally we asked three non-translators to evaluate the entire set of sentences (222 sentences). The time cost for the evaluation in the second phase is comparable to the AMT response time: ranging from 1 hour to 24 hours, according to the respective work-load of our evaluators. Collecting reference translations in order to calculate BLEU and other classical automatic metrics was a much more time consuming task that is however difficult to evaluate.

2.3.2 Crowdsource evaluation

On AMT, translations to evaluate are presented to AMT workers grouped in human intelligence tasks HITs (HITs) of 20 sentences each. These HITs are set up by combining our data with html templates adapted to the different evaluation scales. AMT workers are paid for the evaluation task. We allowed a maximum of 10 assignments per HIT, meaning that 10 different workers would evaluate each sentence. We proceeded in two steps: in the first phase of the experiment we restricted the access to 5 assignments, i.e. a total of five evaluations. The idea is to observe the effect on inter-rater agreement when multiplying the amount of evaluators substantially. We therefore opened the task in a second phase to yet another five assignments in order to observe the behaviour of inter-rater agreement.

In the first phase we studied the reliability of the AMT workers. As we are mainly interested in inter-annotator agreement for the different scales, we have to be particularly careful on how to design the AMT experiment in order to enable detection and exclusion of suspicious responses. Several methods already exist, such as filtering out extremely short task duration (Kittur, 2008), but this alone is not sufficient in our case. Our main difficulty is making sure that the workers who accept our HITs do in fact have the language skills required, namely good knowledge of both languages involved. While workers are generally careful not to work on HITs that they cannot complete satisfactorily, as this might lead to rejection of their work, we cannot exclude that some might try despite insufficient language skills.

In order to recruit AMT workers that seem to present the required skills (fluency in English and Spanish), we first posted a selection HIT of 20 gold-standard sentences. This gold standard set was composed of sentences where all five of our in-house evaluators had reached unanimous judgement. The AMT workers obtaining a minimum of 75% of agreement on the selection HIT were attributed a qualification, and given access to the three real tasks divided into a total of 29 HITS. Very rapidly (within 24 hours) we had assigned the qualification to 20 "suitable" workers, and within three days, our five assignments for all the proposed hits where completed. As mentioned above, we had limited the number of assignments to five for the first stage of the experiment with AMT in order to first verify the feasibility of the task, since it involved English-Spanish bilingual AMT workers. We were surprised to find sufficient English-Spanish participants on the AMT to reach our assignment of five within such a limited time. We were able to give the aforementioned qualification to a total of 11 workers, within 24 hours. However, on the second phase, only 3 out of 13 invited participants participated within an acceptable responding time ranging (within a week). We could have opened the task to more AMT workers, but this meant reopening a qualification round, which we decided not to do and provide the results for a maximum of eight assignments instead.

For the second phase of the experiment, we first identified which qualified AMT workers had not yet achieved the task in the first stage, and quite simply, by changing the level of qualification, we were able to invite them to participate to the new stage of the task by opened 5 assignments for each task (fluency, adequacy and TURKoise). The aim of this new phase was to study the impact of multiplying the number of assignments, through doubling the number of evaluations for each translation, and finally trying to identify what the critical threshold would be: when does the inter-rater agreement reach a peak or degrade. We intended to observe how the inter-rater agreement evolves when multiplying the number of evaluations. The idea behind this second experiment was to find out the impact of having more evaluators working on a task. We hypothesized that if we obtain the same type of inter-rater agreement (i.e. evaluation quality) with 3 assignments, 5 assignments and 8 assignments, it would mean that a small crowd is sufficient to achieve our goal of proposing a reliable Turk-based evaluation.

The next section describes the results obtained for the three tasks and two groups of evaluators. In the section Kappa Evolution we will present the results obtained in the second phase of the experiment.

3. Results

Handling human metrics always implies checking their coherence and inter-rater agreement. We start with the human metrics: fluency and adequacy, using the classic 5-point-scale, then we briefly present the results obtained with our tailor-made human metric. In the first column, we present the results achieved by our in-house evaluators and in the second column contains the results for the AMT workers. We calculated the percentage of unanimous and majority agreement, but we also present Fleiss' Kappa. In our previous study (Starlander & Estrella 2011) we discussed the difficulty of interpretation of Kappa, and decided to follow (Hamon et al, 2009)'s example and calculate the percentage of total agreement between judges, that is the number of times all judges agreed on a category. We apply this on all three evaluation task (fluency, adequacy and TURKoise).

3.1 Fluency

In table two below, the difficulty and subjectivity of the fluency evaluation tasks appears clearly. The percentage of the majority agreement only amounts to 84% for the in-house evaluators and 79% for the AMT workers. For this task, the Fleiss Kappa is slightly better for the in-house translators.

Fluency	In-house	AMT
unanimous	21%	22%
4 agree	32%	29%
3 agree	31%	28%
majority	84%	79%
Fleiss Kappa	0.174	0.164

Table 2: Fluency

As expected, the best agreement is obtained for the highest categories, when the sentences are fluent (obtaining 5 out of 5). This observation can be done for both categories of participants. In this table we present the result for five assignments. As we will further comment in section 3.4, the fluency task clearly comes out as the mostly subjective one, were the variety of answers is rich and the agreement poor. Using AMT for this type of task could be an advantage as the crowd would compensate for the agreement.

3.2 Adequacy

Regarding the adequacy of the translations obtained by MedSLT, a divergence between the percentage of majority agreement and the Fleiss Kappa appears. For the first, the in-house evaluators reach 93% compared to 86% for the AMT workers, while as for the latter, the AMT receives a clearly higher Fleiss Kappa (0.236) than the in-house translators (0.121).

Adequacy	In-house	AMT
unanimous	26%	36%
4 agree	40%	27%
3 agree	27%	24%
majority	93%	86%
Fleiss Kappa	0.121	0.236

Table 3: Adequacy

This result, is quite encouraging, and tends to indicate that the path of using AMT for evaluation of SLT output could

be followed with success. We are keen on comparing this relatively good result with those coming from the tailor-made 4-point scale metric designed for AMT: TURKoise.

3.3 TURKoise

As expected from the results obtained for the adequacy evaluation, the results for TURKoise by the AMT are absolutely comparable to the results obtained by our in-house evaluators. Again, the Fleiss Kappa for the AMT workers is around 0.232, which, although this figure is in the lower range of the interpretation grid for Kappa (Landis & Koch, 1977), the trend is out-classing the equivalent in-house evaluation.

TURKoise	In-house	AMT
unanimous	15%	*32%*
4 agree	35%	26%
3 agree	42%	37%
majority	92%	95%
Fleiss Kappa	0.199	*0.232*

Table 4: TURKoise.

In most cases, the graders disagree on only one point of the scale, but agreeing on the general quality of the rated sentence.

%	CCOR	MEAN	NONS	DANG	all
total number of eval.	735	301	28	44	1108
unanimous	31%	1%	0.0%	0.0%	32%
4 agree	20%	5%	0.5%	1.4%	26%
3 agree	22%	14%	0.0%	1.8%	37%

Table 5: Agreement by categories for TURKoise.

The agreement is far higher for the CCOR category than for the NON or DANG categories, this could be observed in general for all evaluations (fluency, adequacy and TURKoise).

Following these observations, we would like to make some experiments by collapsing similar categories: regrouping on one hand CCOR and MEAN and on the other hand NONS and DANG together, as it is well known that the more point on a scale the less the agreement. In previous research we had observed much higher Kappa on binary evaluation task (ranking task). It would be interesting to test this with the AMT in further research.

In a preliminary study about the participating population, our evaluators recruited so far were always translators and we decided to add non-translators to our evaluators' population, in order to verify the impression we had about our evaluators being particularly severe. As you can see from the table below, it is almost impossible to reach full agreement of 6 evaluators, it only happens in 18.5% of the tested sentences.

In-house evaluators	% of full agreement
All 6 evaluators	18.5%
Translators (3)	33.8%
Non-translators (3)	41.9%

Table 6: Inter-rater agreement in % of full agreement.

However, in the table above, we still observe low inter-annotator agreement when using our 4-point scale, but a more "positive attitude" from the non-translators and a higher inter-annotator agreement, which suggests that translators might not be the best judges as they find it difficult not to focus on the form. Of course, this type of categorisation for our evaluators is impossible to apply to crowdsourcing, since we don't know the profiles of our AMT workers. The only information we gathered is that they are fluent enough to achieve our task. But this is also where the idea of using crowdsourcing occurred to us, since the amount of evaluators would have two effects: smoothing the inter-rater differences and provide a significant amount of bilingual evaluators without being translators.

3.4 Kappa evolution

When using crowdsourcing, the temptation of multiplying the number of evaluators to compensate for the inherent incoherence of human evaluation is great, but the question we wanted to investigate is: Does an increasing number of evaluators really make a difference? The fact that AMT workers are quite easily available and cost-effective does not automatically mean that "more is better".

In the figure below we present the results obtained in our second AMT phase, when we added three supplementary evaluations for each sentence. We compare Fleiss' Kappa obtained for respectively three, five and eight assignments compared to the Kappa achieved by our 5-inhouse evaluators, who each evaluated the entire corpus (222 sentences).

Number of eval.	Fluency	Adequacy	TURKoise
3-times AMT	-0.052	0.135	0.181
5-times AMT	0.164	0.236	0.232
8-times AMT	0.134	0.226	0.227
5-inhouse eval.	0.174	0.121	0.199

Table 7: Kappa evolution according to the number of evaluations

It is interesting to observe that the task obtaining the worse Kappa is fluency. It appears that the more "subjective" the task the more evaluations are needed to compensate and reach a slightly better Kappa. Generally, it appears that the AMT does well, and achieves comparable Kappa levels with the human in-house evaluators. Clearly, for the fluency task, three assignments are not enough, since the achieved Kappa doesn't show any agreement and even show a slight disagreement between the evaluators (Landis & Koch, 1977). Asking

AMT workers to achieve a total of five evaluations in a crowd effort seems to be the best alternative regarding the above table.

Another interesting result from this experiment is the fact that TURKoise achieves a higher Kappa with five assignments than with eight, and confirms that the Kappa is higher than for the in-house evaluators as soon as more than three assignments are accomplished by AMT workers.

4. Conclusion and further work

These preliminary results using AMT workers are encouraging. Overall, it seems that crowdsourcing can be considered to be an effective method to replace or enhance small group-evaluations. The results obtained by both groups are quite comparable on the level of inter-rater agreement. The results obtained tend to show that our selected AMT workers are reliable and surely are cost- and time- effective. Our experiment could also determine which of the three evaluation scales lead to the inter-rater agreement. Our test shows that there is no need to multiply the number of assignments in order to gather more than 5 evaluations for each sentence. In our experiment we needed a total of 23 AMT workers to achieve this goal. The total cost can thus even more be restricted than for our experiment, since five assignments gave better results than eight. Hence, we could identify the "ideal" size of the effectively used crowd to be 5 assignments ventilated on as many AMT workers as necessary, knowing that the total cost for phase one was of 55$, we can also conclude that this method is cost-effective (compared to 50$ for non-benevolent each evaluator): expanding the number of assignments while still remaining competitive in terms of cost and quality. Our general conclusion is thus that there is a clear potential of using TURKoise (i.e. with AMT workers) as an alternative to classic fluency and adequacy human evaluation but also to classic reference-based automatic evaluation.

Then, as mentioned in the introduction, a supplementary aim of our study is to provide the research community with an evaluation method that would not be biased in favour of SMT or RBMT, but that would equally suit to compare spoken language translation output being produced by SMT, RBMT or hybrid systems. We would therefore further investigate the suitability of TURKoise using it this time on both MedSLT output and on the translation provided by an SMT system (probably Google translate) to build on the results from (Starlander & Estrella, 2011). For this purpose we will thus also add a human binary ranking task of MedSLT translations vs. translations obtained with *Google Translate*, and maybe also experiment with AMT workers using a simplified 2-point scale resulting from the above mentioned scale (Meaning, Danger).

In the present paper we focused on a reduced version of our Corpus made of English to Spanish translations, being originally questions asked by the physicians and translated into Spanish by MedSLT to enable the patients to interact with the doctor. The whole corpus So far we have not yet tackled the trickier data of Spanish to English patient responses. These utterances are often even shorter and rely on ellipsis resolution (Bouillon & al., 2005), which represent extra difficulties to be handled by TURKoise. Hence, the next step would be to extend the present study to the remaining real data collection.

Further work, once these experiments achieved, would be to exploit the collected ratings through our "controlled" crowdsourcing to implement quality estimation specifically chosen to suite SLT systems in a safety critical domain, with the "handicap" of having only a very small amount of specific data to add to the general models used (Specia & al. 2010). A viable perspective could be to enhance the corpus through targeted web-search, as there is a massive amount of medical diagnosis type data available out on the net.

5. Acknowledgments

We would like to acknowledge our benevolent in-house evaluators for participating freely to our experiment. Special thanks to my colleague Johanna Gerlach that shared her precious experience and knowledge of the AMT platform.

6. References

Bouillon, P., Rayner, M., Chatzichrisafis, N., Hockey, B.A., Santaholma, M., Starlander, M., Isahara, H., Kanzaki, K., Nakao, Y. (2005). A generic Multi-Lingual Open Source Platform for Limited-Domain Medical Speech Translation. *Tenth Conference of the European Association of Machine Translation,* Budapest, Hungary, pp. 50--58.

Bouillon, P., *et al*. Les ellipses dans un système de Traduction Automatique de la Parole. *Proceedings of TALN/RECITAL*. p. 53—62.

Callison-Burch, C., Osborne. (2006). M. Re-evaluating the role of BLEU in machine translation research. *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy pp. 249--256.

Callison-Burch, C. (2009). Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. *Proceedings of the 2009 Empirical Methods in Natural Language Processing (EMNLP)*, Singapore, pp. 286--295.

Federico, M., Cettolo, M., Bentivogli, L., Paul, M., Stüker, S. (2012). Overview of the IWSLT 2012 evaluation campaign. *IWSLT-2012: 9th International Workshop on Spoken Language Translation Proceedings*, Hong Kong, December 6th-7th, pp.1--33.

Fort, K., Adda, G., & Cohen, K. B. (2011). Amazon Mechanical Turk: Gold mine or coal mine? Computational Linguistics, 37(2), 413--420.

Google translate site, http://translate.google.com/

Jibbigo information site : http://www.cmu.edu/silicon-valley/news-events/news/2009/jibbigo.html

Jurcícek, F.; Keizer, S.; Gasic, M.; Mairesse, F.; Thomson, B.; Yu, K. & Young, S. (2011), Real User Evaluation of Spoken Dialogue Systems Using Amazon Mechanical Turk., *in* 'INTERSPEECH' , ISCA, , pp. 3061--3064.

Hamon, O, Fügen, C., Mostefa, D., Arranz, V., Kolss, M., Waibel, A. et Choukri, K: (2009) End-to-end evaluation in simultaneous translation. Proceedings of the 12th Conference of the European Chapter of the ACL, Athens, Greece, 30 March – 3 April.

Kittur, A., Chi, E. H. and Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. *Proc. of CHI '08*, ACM, Florence, Italy, pp. 453--456.

Landis, J.R., Koch, G.G. (1977). The Measurement of Observer Agreement for Categorical Data, *Biometrics*, **33**, pp. 159--174.

Paul, M., Federico, M., Stüker, S. (2010). Overview of the IWSLT 2010 evaluation campaign. *Proceedings of the 7th International Workshop on Spoken Language Translation*, 2-3 December 2010, Paris, France, pp.3--27.

Specia, L., Dhwaj, R. and Turchi, M. (2010). Machine translation evaluation versus quality estimation. *Machine Translation (24)*, pp. 39--50.

Starlander, M., Estrella, P. (2009). Relating recognition and translation quality with usability of two different versions of MedSLT. *Machine Translation Summit XII*. Ottawa, Ontario, Canada, pp.324--331.

Starlander, M., Estrella, P. (2011). Looking for the best evaluation method for interlingua-based spoken language translation in the medical domain. *Proceedings of the 8th international NLPSC workshop. Special theme: Human-machine interaction in translation*, Copenhagen, Denmark, pp.81--92.

Weiss, B., Schlenoff, C. (2011). Performace Assessments of Two-Way, Free-Form, Speech-to-Speech Translation Systems for Tactical Use. *ITEA Journal 2011*(32), pp. 69--75.

Word Transition Entropy as an Indicator for Expected Machine Translation Quality

Michael Carl and Moritz Schaeffer
Copenhagen Business School
Denmark
E-mail: mc.ibc@cbs.dk, ms@cbs.dk

Abstract

While most machine translation evaluation techniques (BLEU, NIST, TER, METEOR) assess translation quality based on a set of reference translations, we suggest to evaluate the *literality* of a set of (human or machine generated) translations to infer their potential quality. We provide evidence which suggests that more literal translations are produced more easily, by humans and machine, and are also less error prone. Literal translations may not be appropriate or even possible for all languages, types of texts, and translation purposes. However, in this paper we show that an assessment of the literality of translations allows us to (1) evaluate human and machine translations in a similar fashion and (2) may be instrumental to predict machine translation quality scores

Keywords: word translation entropy, edit distance, literality threshold

1. Introduction

While it has been stated that "translators tend to proceed from more literal versions to less literal ones" (Chesterman, 2011: 28) it is controversial what it actually means for a translation to be literal. For some, literal translations are syntactically correct "word-for-word" translations which lack sophistication, and may be wrong if it comes to non-technical types of texts. For humans, as well as for machine translation (MT) systems it is possible to produce more or less literal translations, but it is more difficult to produce and post-edit non-literal translations. As Schaeffer and Carl (2014) show, extra effort is required to 1. generate less literal translation alternatives and 2. cross-check whether the produced non-literal translation still fulfils its function in the given context and translation purpose.

In this paper, we show that the quality of machine generated translations deteriorates as they become less literal. We follow a strict definition which defines literal translations as consisting "of the same number of lexical words, representing equivalent grammatical categories, arranged in the same literal order" (Krzeszowski, 1990: 135). This definition is operationalized by the following criteria:

1. Word order is identical in the source and target languages
2. Source and target text items correspond one-to-one
3. Each source word has only one possible translated form in a given context.

Schaeffer and Carl (2014) assess this literality definition and compare translation from scratch and post-editing. In this paper we trace the production of (non-) literal translations in both, humans and machine, by focussing on point 3 of the above definition. Using a subset of the CasMaCat data, as described in (Carl et al, 2014), we investigate post-editing output and post-editing behavior of nine different linguists working on the same MT output of nine different source texts. We compute the entropy of word translation encodings in the MT search graph and the entropy of word translation realizations in the post-edited texts, and correlate this with the cognitive post-editing effort (gaze behavior, number of keystrokes, post-editing time, and edit distance). We discover a correlation of entropy in human translations and the SMT decoder, and show that much of the post-editing effort is related to difficulties in lexical choice (i.e. increased entropy values), which makes the machine fail more easily, and is more effortful for a human translator to post-edit.

We first describe the experimental setup and our metrics in section 2. In section 3 we discuss possible implications for quality metrics in MT.

2. Experimental Setup and Methods

Within the framework of the CasMaCat project (http://www.casmacat.eu/), nine English source texts were machine-translated into Spanish and post-edited by nine different post-editors. The post-editors worked with three different GUI settings, **PE**, **ITP** an **AITP**, for 'traditional' post-editing, interactive post-editing and advanced interactive post-editing, respectively (Martinez-Gomez et al, 2012). A key-logging device recorded all keyboard activities with a time stamp, cursor offset and segment information. The post-edited texts were subsequently corrected by four independent reviewers. Also during revision, keyboard activities were logged. The translations were semi-automatically aligned on a segment and a word level. Gaze data was collected during a part of the post-editing sessions. A more complete description of the data set is available in (Carl et al, 2014).

In section 2.1 we show that post-editing effort correlates with MT quality, and that more post-editing leads to less revision and vice versa. In section 2.2 we look at the entropy of post-edited texts and see that the more translation choices lead to more cognitive effort. In section 2.3 we assess translation choices in the MT search graph.

2.1 Edit Distance and post-editing effort

From the CasMaCat data, we computed the edit distance between the MT output and the post-edited translations (*MT-PE*) and the edit distance between the post-edited translation and their reviewed version (*PE-RE*). We take *MT-PE* as a quality indicator for the MT output: the more a post-editor modifies the MT output the worse the quality and the bigger the *MT-PE* edit distance. Similarly, we take the *PE-RE* distance to indicate the quality of the post-edited output, as we expect reviewers to modify only flawed passages in the post-edited text.

In this section we first investigate the correlation of *MT-PE* distance and the effort that a post-editor spends on the post-edited segments in terms of gaze time and duration of coherent typing. We then investigate the distribution of workload between post-editing and revision.

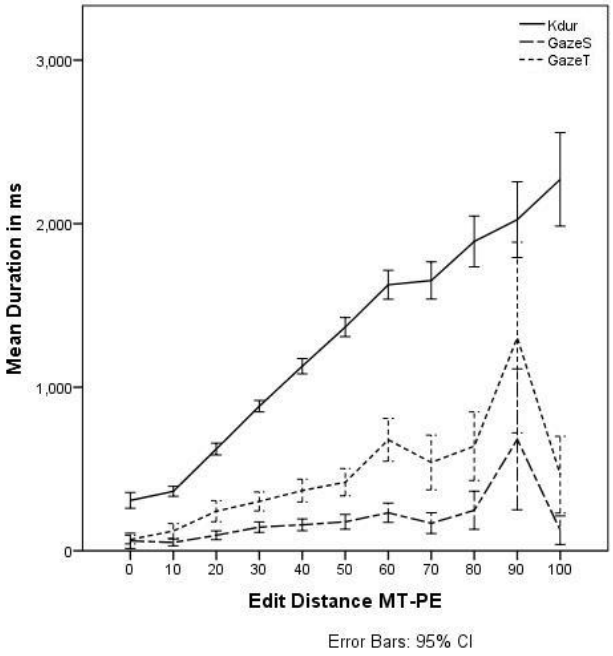

Figure 1: Correlation between edit distance (MT-PE) and mean duration Kdur, GazeS and GazeT per segment.

Post-editing effort

For the post-editing phase, we observe a strong correlation between the *MT-PE* edit distance and the duration of the coherent keyboard activities (*Kdur*). *Kdur* is defined as the total duration of typing "bursts" with no inter keystroke pause longer than 5 seconds.

There is approximately three times more gaze data on the target text than on the source text. We observed a relatively strong correlation between *MT-PE* edit distance and the average gaze duration of the post-editor on target words (*GazeT*), but - unlike in from-scratch translation - the average gaze duration on the source words (*GazeS*) is largely independent from the post-editing duration.

Given that the distribution of gaze duration also indicates allocation of cognitive resources, post-editors (as well as translators) struggle relatively more with the target text than with the source text. Figure 1 plots all three correlations in one Graph.

A strong positive correlation was found between edit distance *MT-PE* and *Kdur* (r = .99) and the regression model predicted 99% of the variance. The model was a good fit for the data ($F(10) = 780$, $p < .000$). A strong positive correlation was also found between edit distance *MT-PE* and *GazeT* (r = .77) and the regression model predicted 60% of the variance. The model was a good fit for the data ($F(10) = 13$, $p < .005$). There was also a relatively strong positive correlation between edit distance *MT-PE* and *GazeS* (r = .60) but the regression model only predicted 36% of the variance. However, the model was a good fit for the data ($F(10) = 5$, $p < .05$).

Post-editing and revision effort

The cumulative time it takes to post-edit and review a translation can be seen as an indicator for the amount of cognitive effort involved in carrying out these tasks, while the edit distance between the MT output and the final (reviewed) text reflects the quality of the MT output. In this section we show that, on average, the cumulative effort is constant, that is, if quick post-editing leads to longer revision times and shorter revision implies more post-editing.

We correlated the edit distance *PE-RE* with the post-editing and reviewing time per word. We observe a positive correlation between *PE-RE* and reviewing time, but a negative correlation between *PE-RE* and post-editing time.

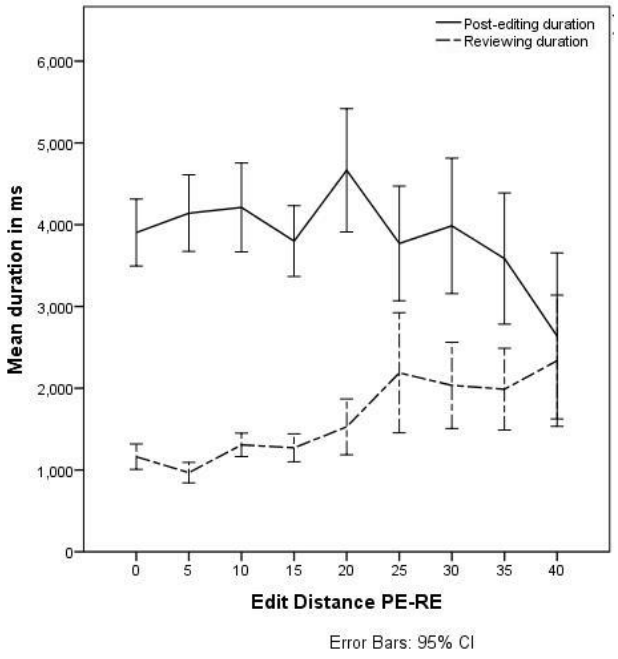

Figure 2: Correlation between edit distance (PE-RE) and reviewing duration (dotted line) and post-editing duration (black line), in ms per word.

Figure 2 depicts a strong positive correlation between edit distance *PE-RE* and reviewing duration (r = .92). The regression model predicted 85% of the variance and the model was a good fit for the data ($F(8) = 41$, $p < .001$).

The correlation between *PE-RE* and post-editing duration was negative and relatively strong (r = .71). This correlation turned out to be only reliable if *PE-RE* values of zero were excluded. The regression model predicted 50% of the variance and the model was a good fit for the data ($F(7) = 6$, $p < .05$).

The edit distance *PE-RE* is generally smaller than *MT-PE*, because reviewers tend to change the post-edited text less than post-editors change the MT output. The edit distance *PE-RE* in Figure 2 ranges therefore only from 0 to 40 as compared to the edit distance *MT-PE* in Figure 1.

The results from Figure 1 suggest that post-editors require more effort when they produce more modifications to remedy the (faulty) MT output. Figure 2 suggests the same for reviewers: the more a reviewer changes the post-edited text, the more time is required. Figure 2 also suggests that the less time a post-editor spends on the MT output, the more time will the reviewer spend on it. In other words, those chunks of text which receive little attention from the post-editor receive more time from the reviewer and vice versa.

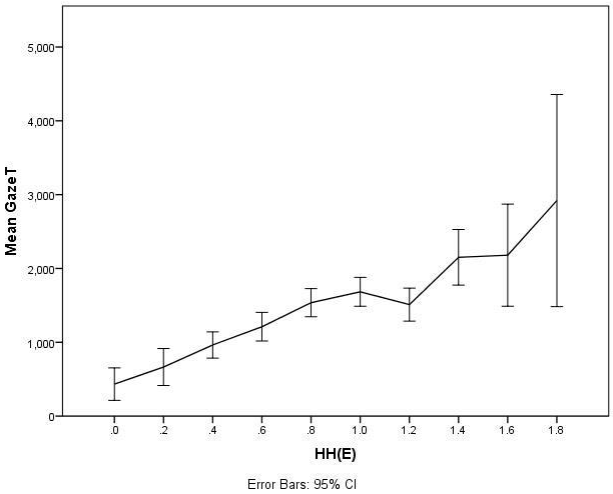

Figure 3: Correlation between HH(e) and GazeT.

2.2 Human Word Translation Entropy

We also computed the word translation entropy of the post-edited texts. The word translation probabilities $p(e \rightarrow s_i)$ of an English word *e* into the Spanish word s_i were computed as the ratio of the number of alignments e-s_i in the post-edited English-Spanish sentence pair *E-S*. Schaeffer and Carl (2014) discuss an example in detail in which 31 translations of the same text are analysed. Subsequently, the entropy of a post-edited source word *e* was computed based on the probability distribution of the different translation realizations s_i:

$$(1) \quad HH(e) = -\sum_i p(e \rightarrow s_i) * \log_2 (p(e \rightarrow s_i))$$

Note that *HH(e)* may be different for different contexts in which *e* occurs. If the machine translation output of an English source word *e* was not modified by any post-editor or all post-editors edited the MT output in the same way, then *HH(e)=0*. Conversely, *HH(e)* would reach its maximum value if the MT output for *e* was modified by every post-editor in a different way. Thus, as post-editors modify the MT output, they are likely to transform the more literal MT output into a less literal version.

Accordingly, we expect a positive correlation between *HH(E)* and *GazeT*, the gaze duration on the target words, since it is cognitively more challenging to choose between more than one translation alternative than to accept whatever the MT system has produced.

Figure 3 plots this correlation between *HH(e)* and gaze time on the translations of *e*. We found a strong positive correlation between *HH(e)* and *GazeT* (r = .97) and the regression model predicted 94% of the variance. The model was a good fit for the data ($F(9) = 130$, $p < .001$).

2.3 Machine Translation Word Entropy

In addition to the entropy *HH(e)* of the post-edited translations, we also computed the entropy *MH(e)* for all the possible Spanish translations $s_{1..n}$ as coded in the machine translation search graph of the Moses system. *MH(e)* is computed based on the transition probabilities from a Spanish word *s-1* to s_i. As an illustration, Figure 4 shows a part of a search graph for the translation German: "*was hast du gesagt?*" into English.

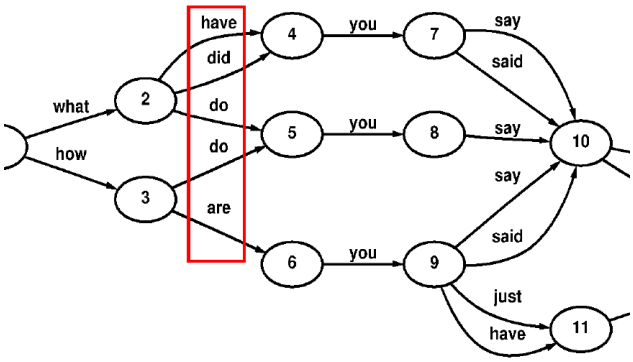

Figure 4: Search graph (adapted from Och et al 2003)

The framed target words (*have, did, do, are*) represent four different translation realizations for the source word "*hast*". As there are several different translation realizations from the preceding English word (*what*, *how*), the entropy of *MH(hast)* will be higher than for the successive German word *MH(du)* for which the search graph encodes "*you*" as the only possible translation. As the search graph encodes in some cases several dozen different translations, up to 10% of the most unlikely transitions were discarded, and transition weights of the remaining translation options were mapped into probabilities. We then compute *MH(e)* the same way as *HH(e)* in equation (1).

We expected that translations become worse as the entropy *MH(e)* increases, since it might be more difficult for the MT decoder to decide which translation to choose if several word transition probabilities are similarly likely, the word translation entropy increases, and thus with this, the likelihood for the MT system to make a sub-optimal choice also increases. To assess this assumption, we correlated *MH(e)* and post-editing duration. As discussed earlier, post-editing duration can be seen as an indicator for the MT quality, so that we also expect a positive correlation between post-editing duration and *MH(e)*.

As shown in Figure 5 this assumption is verified by our data. We found a strong positive correlation between *MH(e)* and *Dur* (r = .94) and the regression model predicted 89% of the variance. The model was a good fit for the data (F (6) = 39, p < .002). This regression model only takes into account *MH(e)* between 1 and 7.
Note that considerable higher durations for *MH(e)=0* can be observed than for many other *MH(e)* values. This is most likely due to the fact that OOV words appear in the MT search graph only in their source form (hence *MH(e) =0*), whereas as post-editors correct them, they produce a variety of different realizations. The MT search graph has actually very low transition weights for OOV, which are turned into *p=1.0* in the normalization step.

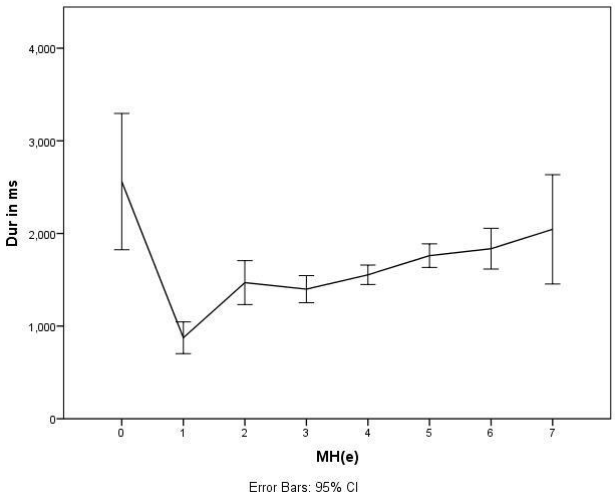

Figure 5: correlation between MH(e) and post-editing duration.

According to Figure 5, post-editors need, on average, longer to tackle OOV words than any other words, which may indicate that OOV words are also difficult to translate for humans. Another interesting observation in Figure 5 is that the variance in post-editing duration increases with the entropy *MH(e)*. Thus, for *MH(e)=7* we observe bigger translation duration variance than for *MH(e)=6.* An explanation may be that more ambiguous translations (i.e. higher entropy values) are also more difficult and thus more time consuming to post-edit.

Thus, apart from the exceptional OOV case for *MH(e)=0*, post-editing durations correlate with higher *MH(e)* values. Accordingly we also expect that MT output with higher *MH(e)* values will more likely be modified, and thus we anticipate a positive correlation between *HH(e)* and *MH(e)*. That is, we expected that the word translation entropy of the post-edited texts correlates with that of the MT search graphs.

The analysis of the two entropy measures confirms also this assumption. Figure 6 depicts the correlation of the entropy *MH* (horizontal) and *HH* (vertical): We found a strong positive correlation between *MH(e)* and *HH(e)* (r = .91) and the regression model predicted 83% of the variance. The model was a good fit for the data (F (6) = 25, p < .004). As before, also this regression model only takes into account *MH(e)* values between 1 and 7 and ignores *MH(e)* values of zero.

For *MH(e)=0*, the *HH(e)* values are higher than for *MH(e)=1*, which can, again, be explained by the special situation of OOV words in the SMT search graph.
It is important to recognise that while *MH(e)* and *HH(e)* correlate well, the ideal scenario would be that *HH(e)* and *MH(e)* correlate in a completely linear fashion. Our data suggests that this is not the case. Our data suggests that when the SMT system is relatively uncertain (*MH(e)* values of 7), the human translator considers a much smaller number of options (*HH(e)* values of around 1.5). This drop might indicate a threshold by which MT output is no more suitable for post-editing.

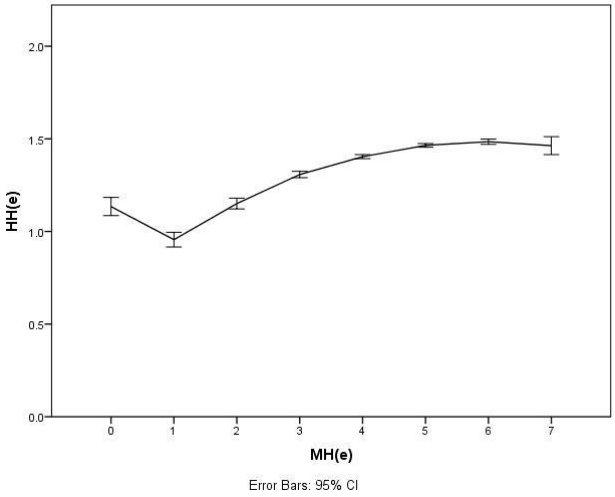

Figure 6: Correlation between HH(e) and MH(e).

3. Discussion

In this paper we pinpoint a correlation between the entropy of human translation realizations and the entropy of machine translation representations. Where humans produce many different translations from a word or a sentence, also MT systems are more ambiguous. This finding is not surprising, since statistical machine translation systems are trained on, and thus imitate, the variety of translations produced by human. It seems to attest as John M. Smar (2003) suspects that "the human conversation phase space, in all languages, and all digital forms (web publishing, email, audio, video, chat, and other searchable conversations) while still growing slowly in novelty, is becoming *effectively, approximately, or statistically* computationally closed relative to the rapid mapping of this space by technological intelligence."

According to our definitions, word-based translation entropy is tightly linked to translation literality, and as translations become less literal (be it for structural or cultural reasons or for translator's choices) state-of-the-art statistical machine translation systems fail, while human translators seem to deploy as of now non-formalized translation strategies, to select amongst the many possible translations a good one.

It seems that those parts of the "human conversation space" that become "ergodic"[1] are suitable for MT and for post-editing, provided the entropy is below a certain threshold. An entropy threshold may be an indicator for this turning point and may serve as an indicator for translation confidence, beyond which the quality of MT output becomes less reliable, and thus MT post-editing may become less effective.

Our investigation shows that translations are simultaneously less literal in man and machine. However, while the quality of MT output decreases as the translations become less literal, for humans, non-literal translations are more effortful to produce - but presumably still appropriate and correct in the given context.

The more the source and the target languages are different from each other, the more difficult it may be to understand a literal translation. The source language may contain a word for a concept which has no direct equivalent in the target language, appropriate lexical selection may be difficult to determine in a given context and for a given purpose, or the syntactic structure may impose non-literal rendering.

To enhance the quality of MT output, one line of research asks to what extent it is *possible* to produce non-literal translations. How can we describe non-literal translations formally and statistically? How can we train computers to produce non-literal translations correctly?
Research answers to this set of questions seek to include more background knowledge into the MT systems, such as more sophisticated decoding algorithms, taking into account syntactic and semantics representations, deep learning, etc. Another (much common) approach is just to add more data in the hope that more parts of the "human conversation space" are covered by pure observation and more reliable probabilities can be inferred.

Another set of research questions is geared towards asking to what extent is it *necessary* to produce all kinds of translations? That is, how far can we come if we restrict the texts to be translated and train special MT systems for the purpose at hand?
Answers to this set of questions point to controlled language translation for different text types, domains, and text genres. The underlying assumption is that there might be (language-pair, or translation purpose dependent) literality thresholds which ensure MT systems to comply with pre-defined quality requirements. Controlled language translation (Mitamura, 1999; Muegge 2007) has been one attempt to reduce the conversation space, and thus to predict the expected quality and literality of the MT output. The Multidimensional Quality Metrics[2] is the most recent initiative to formulate translation expectations according to which machines can be trained and quality requirements formalised.
Domain adaptation either in a batch mode[3] or in a setting of online and adaptive learning (Martinez-Gomez et al, 2012, Ortiz-Martinez, 2010), where the translation system learns at runtime from the corrections of a post-editor, is still another attempt to decrease the entropy of human translations. Luckily, all methods are complementary.

4. Conclusion

In this paper we show that the entropy of the transitions in the MT search graph is related to the MT quality and subsequently to post-editing duration. This suggests that certain entropy values might be suitable for translation confidence indicators and the MT output beyond some entropy threshold might not be useful for post-editing.

We define literality of translation by the amount of source-target reordering and the diversity of translation realizations. We look more closely into the latter point by investigating the entropy of human and machine word translation choices. We compare human word translation entropy with machine word translation entropy, translation quality and MT post-editing effort. In our analysis we show that:

1. The effort of the post-editor correlates positively with the edit distance between the MT output and the post-edited version of the text. The more a post-editor modifies a segment, the less time is needed for reviewing.
2. Human word translation entropy correlates with gaze duration on (and translation production time of) the translation: it is more time consuming for a translator to translate a source language word which can be translated in various different ways, than a source word which can only be translated into one or small number of different target words, with high probability.
3. Human word translation entropy correlates with machine word translation entropy: if post-editors translate a source word in many different ways, the SMT system also has many translation options for that word.

We discuss whether these findings may be instrumental for the design of a novel translation confidence metrics which rely on the entropy of word transitions in the MT search graph. More work is needed to replicate our findings and to formulate in more detail possible thresholds for different translation purposes.

In future work we also intend to take into account the other parameters of the literality definition as provided in the introduction: i.e. to what extent one-to-one

[1] An intuitive explanation of the term is provided on:
http://news.softpedia.com/news/What-is-ergodicity-15686.shtml

[2] MQM: http://www.qt21.eu/launchpad/

[3] See, for instance:
http://www.clsp.jhu.edu/workshops/archive/ws-12/groups/dasmt/ (last accessed 22.3.2014)

correspondence between source and target language items and change in the word order may have an impact on a literality quality threshold.

5. Acknowledgements

This work has been supported by the CasMaCat project, co-funded by the European Union under the Seventh Framework Programme, project 287576 (ICT-2011.4.2). We are grateful to the nine post-editors and four reviewers who took part in the second field trial, whose UAD contributed to build the CFT13 study

6. References

Michael Carl, Mercedes Martínez García, Bartolomé Mesa-Lao, Nancy Underwood (2014) "CFT13: A new resource for research into the post-editing process". In Proceedings of LREC

Chesterman, Andrew. (2011) Reflections on the literal translation hypothesis. In Methods and Strategies of Process Research, edited by Cecilia Alvstan, Adelina Held and Elisabeth Tisselius, Benjamins Translation Library, pp. 13-23, 2011

Martinez-Gomez, P., Sanchis-Trilles, G.; Casacuberta, F. 2012. "Online adaptation strategies for statistical machine translation in post-editing scenarios". Pattern Recognition, 45:9, pp. 3193 - 3203.

Krzeszowski, Tomasz P. (1990) Contrasting Languages: The Scope of Contrastive Linguistics. Trends in Linguistics, Studies and Monographs, Mouton de Gruyter

Franz Josef Och, Richard Zens, Hermann Ney (2003) "*Efficient Search for Interactive Statistical Machine Translation"* In EACL '03: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics, pp. 387--393

Teruko Mitamura (1999) Controlled Language for Multilingual Machine Translation In Proceedings of Machine Translation Summit VII

Uwe Muegge. "Controlled language: The next big thing in translation?" *ClientSide News Magazine* 7.7 (2007): 21-24. Available at: http://works.bepress.com/uwe_muegge/4

Moritz Schaeffer and Michael Carl (2014) Measuring the Cognitive Effort of Literal Translation Processes, In Proceedings of the Workshop on Humans and Computer-assisted Translation (HaCaT) at 14 Conference of the European Chapter of the Association for Computational Linguistics

John M. Smart (2003) The Conversational Interface: Our Next Great Leap Forward, http://www.accelerationwatch.com/lui.html#phasechange (last accessed 19.3.2014)

Daniel Ortiz-Martinez, Ismael Garcıa-Varea, Francisco Casacuberta (2010) Online Learning for Interactive Statistical Machine Translation, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 546–554,

Standardized ILR-Based and Task-Based Speech-to-Speech MT Evaluation

Douglas Jones, Paul Gatewood, Martha Herzog, Tamas Marius*

MIT Lincoln Laboratory
244 Wood Street, Lexington, MA

DLI Foreign Language Center*
542 Rifle Range Road, Monterey, CA

E-mail: daj@ll.mit.edu, paul.gatewood@ll.mit.edu, MHerzog2005@comcast.net, tamas.marius@dliflc.edu

Abstract

This paper describes a new method for task-based speech-to-speech machine translation evaluation, in which tasks are defined and assessed according to independent published standards, both for the military tasks performed and for the foreign language skill levels used. We analyze task success rates and automatic MT evaluation scores (BLEU and METEOR) for 220 role-play dialogs. Each role-play team consisted of one native English-speaking soldier role player, one native Pashto-speaking local national role player, and one Pashto/English interpreter. The overall PASS score, averaged over all of the MT dialogs, was 44%. The average PASS rate for HT was 95%, which is important because a PASS requires that the role-players know the tasks. Without a high PASS rate in the HT condition, we could not be sure that the MT condition was not being unfairly penalized. We learned that success rates depended as much on task simplicity as it did upon the translation condition: 67% of simple, base-case scenarios were successfully completed using MT, whereas only 35% of contrasting scenarios with even minor obstacles received passing scores. We observed that MT had the greatest chance of success when the task was simple and the language complexity needs were low.

Keywords: Machine Translation Evaluation, Independent Standards, Interagency Language Roundtable

1. Introduction

This paper[1] presents a new method for task-based speech-to-speech machine translation evaluation, in which tasks are defined and assessed according to independent published standards, both for the military tasks performed[2] and for the foreign language skill levels used[3]. We analyze task success rates and automatic MT evaluation scores for 220 role-play dialogs. Each role-play team consisted of one native English-speaking soldier role player, one native Pashto-speaking local national role player, and one Pashto/English interpreter. Machine translation (MT) and human translation (HT) conditions were assigned in a Latin Square design. Dialogs were assessed for language difficulty according to the Interagency Language Roundtable (ILR) speaking and listening skills. The overall PASS score, averaged over all of the MT dialogs, was 44%. The average PASS rate for HT was 95%.

Scenarios were of two general types: a basic definition without any complications, and a contrasting definition with some type of obstacle, perhaps minor, that needed to be overcome in the communication. For example, in a basic Base Security scenario, a Local National may seek permission to pass a checkpoint with valid identification. In a contrast scenario, he may lack the identification, but seek an alternative goal that does not involve passing the checkpoint, such as passing a message to a relative who is working behind the checkpoint. Overall PASS/FAIL results for the HT condition were 95% for basic scenarios and 94% for contrasting scenarios with obstacles. For MT we observed 67% PASS for basic and 35% for contrast scenarios. The performance gap between HT at 94˜95% and MT with basic scenarios at 67% is 27% on average, whereas the difference between MT in basic scenarios and MT in contrasting scenarios is 32%, as shown in Figure 1.

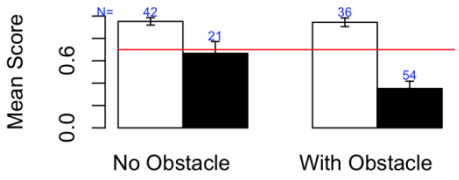

Figure 1: PASS/FAIL Results with Task Obstacles

The dialogs were also assessed for language complexity. Scenarios with language complexity at the ILR Levels 1, 1+ and 2 had PASS scores of 94%, 100% and 92% respectively in the HT condition, as shown in Figure 2. For MT the overall results were 47%, 48% and 31%. In other words, MT does not work as well when the language is fundamentally more complex. The average BLEU score for English-to-Pashto MT was 0.1011; for Pashto-to-English it was 0.1505. BLEU scores varied widely across the dialogs. Scenario PASS/FAIL performance was also not uniform within each domain. Base Security scenarios did perform relatively well overall. On the other hand, although Civil Affairs scenarios did not perform that well on average, some of the scenarios were performed well with MT.[4]

Role players performed 20 tasks in 4 domains. The domain-level PASS scores ranged from 89% to 100%

[1]This work is sponsored by the Defense Language Institute under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

[2] See (ILR, 2014) for a description of the Interagency Language Roundtable language skill levels.

[3]See (USArmy, 2014) for Training and Doctrine references.

[4]See (Jones et al., 2007) for an ILR-based evaluation of text MT with comprehension questions.

Figure 2: PASS/FAIL Scores by SME ILR Level for Speaking

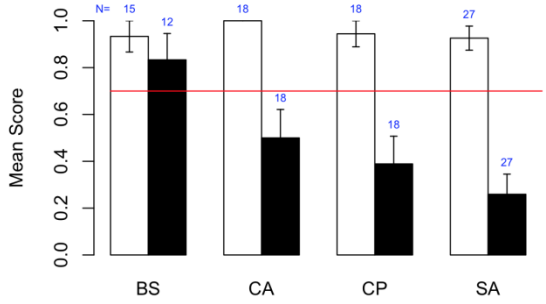

Figure 3: PASS/FAIL Results Across Domains

in the HT condition. For MT we observed 83% PASS rate in one of the four domains, Base Security (BS), with the remaining three domains ranging from 26% to 50% (Checkpoint Operations (CO), Civil Affairs (CA) and Situational Awareness (SA), an umbrella domain encompassing a variety of more complex tasks). The dialogs were human-scored in two main ways: (a) aggregate PASS/FAIL outcomes, and (b) a secondary assessment of specific communication initiatives. Inter-coder agreement for task PASS/FAIL scoring, which required an assessment of several performance measures per task, averaged 83%. Agreement for the specific communication initiatives was 98%. The PASS/FAIL scores for scenarios within the four domains are shown in Figure 3.

The overall impression is that MT can work for some scenarios, but language complexity and task obstacles may drastically reduce performance, and in fact these factors can be as important as the contrast between machine translation and human translation. In other words, the effect of the scenario complexity is stronger than the effect of using MT instead of an interpreter. MT may provide needed communication when the only barrier for accomplishing a simple task is the language barrier. When the task has unexpected obstacles, MT is more likely to fail.

2. Design

We constructed a framework for evaluating speech-to-speech machine translation technology in a way that isolates spoken language used in performing standardized military tasks. Role players performed their duties via a push-to-talk communication system that routed human and machine-generated spoken language to the relevant role players and to a system monitor. Whether any given scenario used HT or MT was determined by the randomized position in the Latin Square design. The role players were as follows: (1) an English-speaking Subject Matter Expert (SME), a person who has experience with the various military tasks; (2) a Foreign Language Expert (FLE), a Pashto-speaking playing the role of the Local National; (3) an Interpreter (INT) who is able to provide immediate Pashto/English translation for the FLE and the SME. The SME and the FLE cannot hear each other; they communicate only via the interpreter providing human translation (HT) or machine translation (MT). Role players interacted via MT interface in all conditions in order to maintain consistency in communication behavior. In the HT condition, the MT output was saved for further study, but not used in the role-play, the interpreter's live production of HT being substituted in its place.

The two human subjects engaged in a role-playing scenario. The SME took on the role of a US soldier charged to perform a certain duty based on relevant US Army training and doctrine. The SME spoke only in English. The FLE took on the role of an Afghan Local National. The FLE spoke only in Pashto. Both subjects were given common information about the scenario; however, each subject also received unique additional information. Each subject was given a goal for the scenario; the subjects had to try to attain their goals through voice communication. Translation during the scenario was either machine-based or human-based.

2.1. Configuration

Normally, the translation devices show three machine translation stages: first, the Automatic Speech Recognition output is shown in the native language. If there is an obvious mis-recognition of a word, the role-player will know not to expect a good translation. Second, the translation into the foreign language is shown. Since the English speakers did not understand Pashto, we were not concerned with masking the output. However, since the Pashto-speakers also speak and understand English, we configured their devices so that they did not show the English output, in order to avoid biasing the experiment in a way that would not reflect real-world use, where the Pashto speakers do not understand English. For this reason, FLEs were not allowed to monitor their English MT audio output. In this way the FLE role players, who were by necessity English speaking, could more accurately mimic a non-English speaking foreign national. The third machine translation stage, a back-translation into the native language for the role-player, was shown for all role-players.

The diagram in Figure 4 shows the general construct of the communication setup.

To review: the SME machine translation (MT) device was configured to provide textual feedback of both the automatic speech recognition (ASR) and back translation (English -> Pashto -> English). SME's were encouraged to

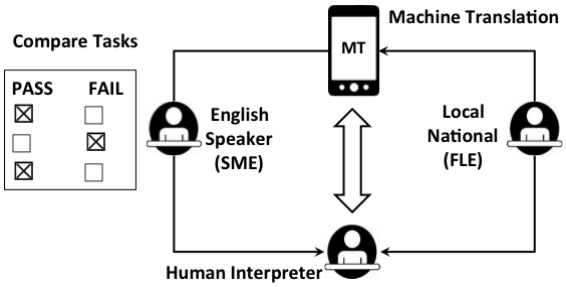

Figure 4: Evaluation Concept

make multiple attempts if necessary to achieve accurate ASR and if possible accurate back translation. Time constraints imposed a practical cap of 4-5 attempts per turn. On average, SMEs generated twice as many translations as were eventually used in communication, and FLEs generated 1.3 times as many.[5]

The transcript in Table 1 shows examples of how role players used back translations to decide whether to keep a translation or to try again. Recall that the role players are able to see a back-translation into their native languages before sending the foreign language output to their counterpart for communication. Although not a direct indication of whether the foreign language translation was good enough, if they observed obvious problems with the back-translation, they could reject the output and try again. For example, on Line 6, the FLE can observe that the presence of the name "Khan" (جان) does not make sense, and can try again. Lines 3-5 show how the SME rejected the first two translations based on the English back-translations (rejecting "Please help and take me to the explosion", trying again and also rejecting "Please help me gets to the explosion", and trying a third time, accepting "Please help me go to the explosion". In some cases, the role-players may have had unreasonably high expectations for fluency of a back-translation.

The overall intent was to provide the role-players with another chance to produce a translation when the back-translation looks too garbled, giving them an opportunity to try again.[6] The column on the right shows these rejected MT turns.

2.2. Terminology

By task, we mean US Army tasks as defined by TRADOC (Training and Doctrine) materials, such as Army Field Manuals, indexed by task identifier. The tasks used in our experiment are shown in Table 2. A *scenario* is a description of one of the these TRADOC tasks, specified in sufficient detail for the tasks to be performed consistently by role players. A *dialog* is the record of a specific role-playing event, a task performed by the role players. We refer to language expressed to initiate a communication goal as *communication initiatives*, regardless of whether it was understood or contributed to a passing score. For example, if the role player said, in English: "How many doctors are in the community?" the dialog would be scored positively for the initiative goal defined as "Does the role player ask about the local staff?" The overall *PASS/FAIL* goal for the dialog in this case was "Does the role player collect health information?" which required successful completion of several performance steps to receive a PASS. Inter-coder agreement for task PASS/FAIL scoring averaged 83%; agreement for the specific communication initiatives was 98%, as shown previously in Figure 3, the PASS/FAIL scores for scenarios.

2.3. Roles

The SME was expected to know the various training and doctrine principles necessary to perform the tasks within each scenario in a detailed and consistent fashion. They were expected to conduct these procedures as they would in a real world situation and maintain a respectful and courteous posture. In the case of scenarios in which the SME and FLE have opposing goals, the SME tried to find a compromise or other way to accommodate the FLE. SMEs used speech to communicate intent and instant messages (IM) to communicate physical context. The SME was shown how to operate the PC GUI including IM and the MT device during training.

The FLEs were expected to conduct themselves as they imagine they would under real world conditions while maintaining a respectful and courteous posture. If the FLE could not get a SME to agree to the goals as put forth in the scenario description the FLE would attempt to find a compromise or partial solution.

The interpreter provided real-time translation in both directions during HT scenarios and performed no interpretations during MT scenarios.

The monitors oversaw one set of role players (SME/FLE/INT). The monitors assigned, began, ended and monitored all scenarios. They were responsible for reverting a turn if procedural mistakes were made. They also broadcast IMs for any physical action that added to the shared context with the proper "SEE SME" or "SEE FLE" preface even when these were redundant to the player IMs. The monitors also took on the role of virtual SME and FLE by prompting the actual players via private IM any unaddressed communication goals before the end of the scenario. The monitors also scored the scenario. Runners oversaw individual role players and reported to the monitors. These were former SMEs or FLEs with experience in performing the tasks. An experiment manager was responsible for the overall experiment. The manager

[5] The rejected "practice" MT output created privately by each role-player was not used in constructing the reference translations and transcripts and hence was not used in calculating automatic scores, such as BLEU.

[6] For expository purposes in this paper, the English MT is shown in parenthesis for the FLE side, in addition to the Pashto back-translation seen by the role player. Recall that the FLE MT device was configured to only display Pashto, so for the FLE side, only the back-translations into Pashto were used.

Line	Role Player	Back-translation from MT Output Used by Role Players for Communication	"Practice" Back-translation MT Output not used by Role Players for Communication
1	SME:	I need your help	
2	*FLE:*	کولای شمنهزه *(I can not)*	
3	SME:		Please help and take me to the explosion
4	SME:		Please help me gets to the explosion
5	SME:	Please help me go to the explosion	
6	*FLE:*		دی خان وو ځای محفوظ ته ما *(I was khan is safe)*
7	*FLE:*	وپرپرم زه *(I am afraid)*	
8	SME:		You will be okay come with me
9	SME:		You will not gephardt come with me
10	SME:	You will not be hurt come with me	
11	*FLE:*		ده صحیح *(Okay)*
12	*FLE:*	مننه ده صحیح *(Okay thank you)*	
13	SME:		Can you tell these people to stay far away
14	SME:		Tell all these people to stay back
15	SME:		Gets the people to back up
16	SME:		Tell the people to go away
17	SME:		Do you tell the people to go away
18	SME:	Tell the people to go away	
19	*FLE:*	شم ولاړ ته یې زه چې وکړئ مر سر ته سره ما *(Help me so that I can go)*	
20	SME:		Gets the people to back up
21	SME:		Move the people away
22	SME:	Move the people	
23	*FLE:*	شومنهپوه *(I did not understand)*	
24	SME:		Help me move people away
25	SME:	Help me move people back	
26	*FLE:*	شمنهمر سر ته سره تاسو وپرپرم زه *(I am afraid I can not help you)*	
27	SME:		Move away to safety
28	SME:		And then go to safety
29	SME:	Then move back so you do not get hurt	
30	*FLE:*	کول شمنهکومکزه ایاوخت ځای معلوم ته ما *(I know where do I have time I did not help)*	

Table 1: Use of Back Translations Select MT Turns

Task ID	TRADOC Task		ILR Speaking Skill Level
		Civil Affairs	
CA2	331-38B-2020	Conduct a Local Medical Health Assessment	L2
CA4	331-38B-3015	Coordinate Handling of Supplies	L1+/2
CA5	331-38B-3033	Conduct Support to Civil Administration Operations	L2+
		Situational Awareness	
SA1	301-35M-1200	Implement Approach Strategies	L3
SA2	301-35M-1250	Assess Source for Truthfulness and Accuracy	L2
SA3	191-376-5126	Conduct Interviews	high L1
		Checkpoint Operations	
CP2	171-137-0001	Search Vehicles in a Tactical Environment	L1
CP6	191-376-5151	Control Access to a Military Installation	high L1
		Base Security	
BS1	191-376-4130	Operate a Roadblock as a Member of a Team	L1
BS4	191-376-5154	Respond to a Crisis Incident	L0+

Table 2: Task Inventory

typically monitored one of the role play teams in addition to guiding a second monitor when we had two parallel role playing teams.

2.4. Scenarios

The following lists show the information from the sample scenario that was provided to the role players. Both the SME and the FLE saw the material designated as "Shared Context".

- **SHARED_CONTEXT:** A US soldier is talking to a doctor. This Area of Operations (AO) is safe and secure.

The role players did not see information about each other's knowledge and goals. The SME saw these descriptions:

- **SME_KNOWS:** You are a Civil Affairs Soldier assigned to a civil-military operations center. You know how to conduct a Local Medical Health Assessment. Today you are concentrating on collecting human health information only.
- **SME_GOAL:** Follow the procedures for Conducting a Local Human Health Assessment. Collect health information including local facility names, the number of health workers and the nearest pharmacy. Determine the size of the population. Identify any endemic diseases and the leading causes of death.

Likewise, the FLE saw only this relevant part of the scenario description:

- **FLE_KNOWS:** You are the only doctor here with responsibility for the two villages and the surrounding area. Your office is the only clinic and pharmacy. You have a meager stock of only the most basic medications. There are about 200 extended families. Twenty percent of the population is over 65 years old, or about 250 people. Children under the age of 12 number about 400. Cholera is the leading cause of death. There are scattered cases of hepatitis B. You have also seen a rise in cases of brucellosis.
- **FLE_GOAL:** Describe the medical situation of the population in your area.

Neither the SME nor the FLE see the scoring criteria during the role-play in order to avoid over-scripting the dialogs. A full day of training was provided to the SME to cover the requirements according to the standard definitions of task, conditions, standards, performance steps and performance measures. After the role-play, the SME was assigned PASS/FAIL scores for the specific goals shown in Figure 5. We defined goals for the FLE to be scored in a similar fashion, although these are obviously not part of the training materials for soldiers. The FLE goals for this sample scenario are also shown in Figure 5.
Table 3 shows the list of scenarios in the experiment.

Sample Soldier SME Goals

1 Collect health information.
2 Ask about the local facilities.
3 Ask about the local staff.
4 Ask about the nearest pharmacy.
5 Ask about the size of the population.
6 Ask about any endemic diseases.
7 Ask about the leading cause of death.

Sample FLE Goals

1 Describe the local facilities.
2 Describe the local staff.
3 Describe the nearest pharmacy.
4 Convey data on the population.
5 Describe any endemic diseases.
6 Describe the leading cause of death.
7 Convey health information.

Figure 5: SME Goals for Sample Scenario

2.5. Sample Transcripts

Table 4 shows a sample transcript of the human interpreter for a Base Security scenario. Table 5 shows two MT transcripts. The first one is a partially fluent machine translation in a Base Security scenario. This is the same dialog shown in Table 1 which showed how some MT output is rejected using back translations. The second shows a lower quality MT interaction which causes some difficulty for the role players in the Civil Affairs scenario.

2.6. Scoring

Dialogs were scored in three different ways for each side of the conversation. *Communication Initiatives* are scored with one side of the conversation; they are scored as PASS if the Role Player attempts to communicate a particular objective, regardless of whether the FLE understands it or does anything in response. For example: "Does the SME ask about the local facilities?".
World Goals are scored with respect to both sides of the conversation for one role player; they scored as a PASS if the role player's real world outcome was met. The world goals may require action or communication on the part of the both role players in order to be assigned a PASS. They require successful two-way communication to meet the goal. For example, in response to the question "Does the SME collect health information?" the *PASS/FAIL Outcome* would be an aggregate score of all successful interactions, and that would be scored as PASS only if all of the role player's world goals were met. Otherwise, it is scored as FAIL. It requires successful two-way communication over the dialog as a whole.

The abbreviations for the score types are as follows: **SC**: SME Communication Initiative; **SW**: SME World Goal; **SO**: SME PASS/FAIL Outcome.
Scores for Soldier Communication Initiatives and PASS/FAIL outcomes for all of the tasks are shown

ScenID	Scenario Title and Domain
	Base Security
BS11	Slow US Convoy Blocking Road
BS45	Respond to a Crisis Incident
BS46	Local National wants to help in crisis
	Civil Affairs
CA21	Human Health Assessment
CA22	Animal Health Assessment
CA51	Return Displaced Civilians
CA54	Discuss Situation regarding Rule of Law
	Checkpoint Operations
CP21	Family Vehicle at Checkpoint
CP22	National with Borrowed Vehicle
CP61	Local National Wants to See Commander
CP66	Local National Wants Job
	Situational Awareness
SA11	Approach Tea Shop Owner for Information
SA12	Speak with Villager for Information
SA21	Assess Villager as Source for Information
SA22	Assess Insurgent Cell Member as Source
SA12	Interview Escaped Detainee
SA13	Local National Police Training

Table 3: Scenario Inventory

in Figure 6 and Figure 7 respectively. The key observations are: (1) the soldier role players generally succeed in their Communication Initiatives, both for MT and HT. However, the overall outcomes vary greatly on a scenario by scenario basis. The clearest distinction is between the Base Security (BS) scenarios at the more successful end, compared with the more complex Situational Awareness (SA) tasks with lower success rates.

2.7. Data Triage

There were three steps of data triage. First, we dropped two of the original twelve SME teams whose overall PASS average in the HT condition was lower than 70%. In order to assess the relative effect of machine translation, we required that the role players be able to perform the tasks required. In this triage step, we dropped all role plays for those two teams, both MT and HT. Second, we also dropped the subgoals for which the average success rate was less than 70%: of the 209 goals, 20 were dropped. Third, the second step of subgoal triage meant that 3 of the 20 scenarios lacked an overall PASS/FAIL score, so these were also excluded. Future experiments would repair the subgoals and associated overall PASS/FAIL scoring, and should only need the first triage step to exclude role-players who could not complete the tasks in the HT condition.

Human Interpreter
Civil Affairs Scenario CA12, Team 6

SME:	how many men and trucks can you provide ?
FLE:	*two cars four people four guns .*
SME:	do you know the roads very well ?
FLE:	*yes I know the roads very well .*
SME:	can you make sure nobody steals the supplies ?
FLE:	*okay . I'm trying my best .*
SME:	do you know the safest way to get there ?
FLE:	*yes I know a safe way .*
SME:	how many men can you provide ?
FLE:	*I can provide you with four men .*
SME:	how many trucks ?
FLE:	*two trucks .*
SME:	sounds good . I like the plan .
FLE:	*no problem . I'm thankful to you .*

Table 4: Sample HT Transcript

2.8. ILR Assessment of TRADOC Tasks

The tasks were assessed according to ILR skill level likely needed for successful task completion, as shown in previously in Table 2. The Speaking and Listening skill estimates were generally very close. The actual Speaking ILR levels observed for both HT and MT are shown in Figure 8. In general, the language used for MT was slightly less complex than for HT, and the levels were somewhat lower than what was estimated in advance of the experiment.

As might be expected, the number of words used in communication correlates somewhat with the ILR Speaking levels, as shown in Figure 11, with $R^2 = 24\%$. In other words, the role player speaks more when the required language skill is more challenging.

2.9. Dialog Length

In Figure 10, a histogram of the number of turns needed to complete the dialog is shown for four cases. First, the number of turns used in dialogs receiving a PASS score. These are usually completed in under 15 turns. Failing dialogs never finished early; they required as many as 20 or more turns until they reached the time limit for the role-play, at which point they moved on. Passing dialogs in the HT condition had about as many turns as the passing dialogs in the MT condition. Very few HT dialogs failed; these were due to missed subgoals on the part of the role-players, rather than a general failure to communicate.

2.10. Subject Variation

We observed a noticeable amount of variation by subject, as shown in Figure 9. The highest performing team was S10 at 78% PASS rate for MT. A binomial test, using the overall PASS rate of 44% for MT, shows that this result would be achieved by chance just under 5% of the time, not quite enough to expect that this team is doing something special.

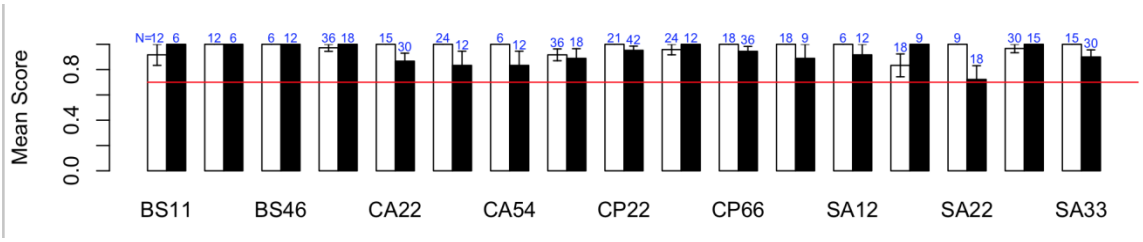

Figure 6: Communication Initiative Scores for Scenario

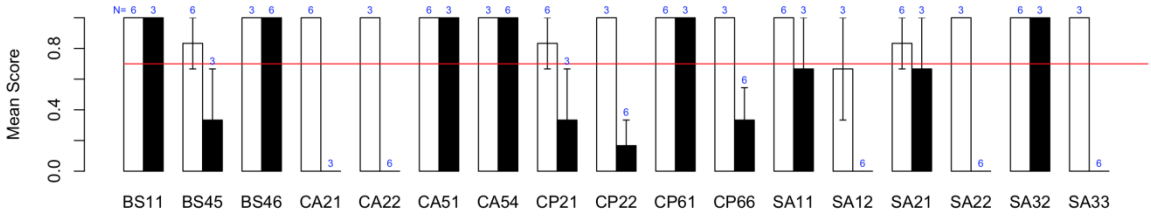

Figure 7: PASS/FAIL Scores for Scenarios

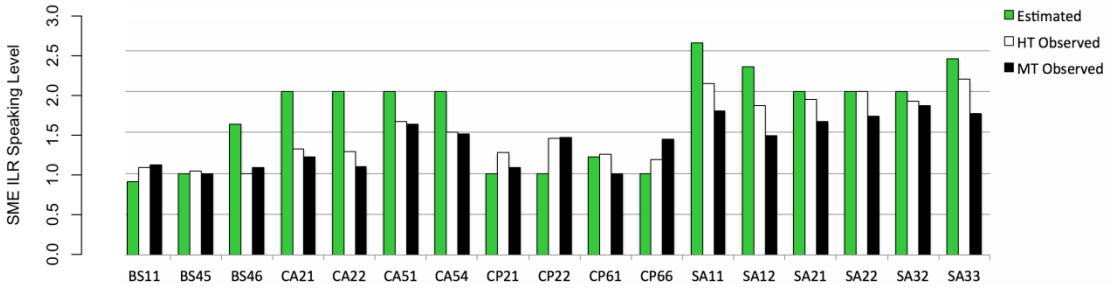

Figure 8: ILR Speaking Skills Estimated and Observed for Scenarios

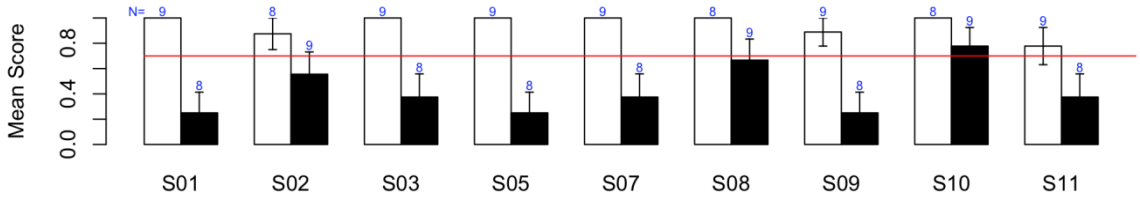

Figure 9: PASS/FAIL Scores by Subject Team

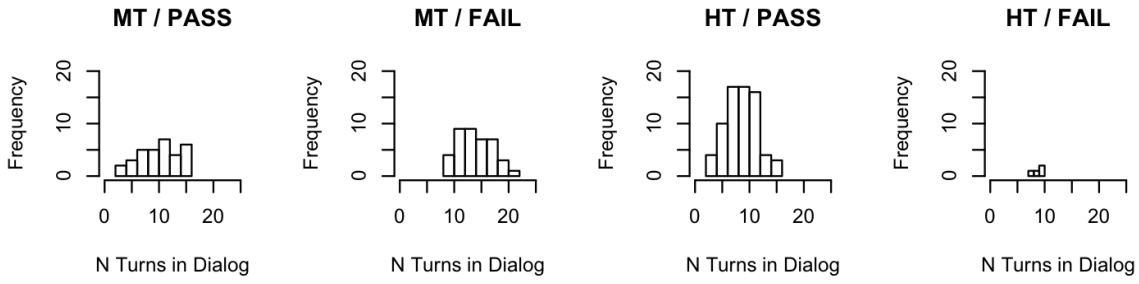

Figure 10: Length of Dialogs and PASS/FAIL Scores

Machine Translation
Base Security Scenario BS45, Team 6

SME:	I need your help .
FLE:	*I can not*
SME:	please help me go to the explosion .
FLE:	*I am afraid*
SME:	you will not be hurt . come with me .
FLE:	*Okay thank you*
SME:	tell the people to go away .
FLE:	*Help me so that I can go*
SME:	move the people .
FLE:	*I did not understand*
SME:	help me move people back .
FLE:	*I am afraid I can not help you*
SME:	then move back so you do not get hurt .
FLE:	*I know where do I have time I did not help*

Machine Translation
Civil Affairs Scenario CA42, Team 1

SME:	I send medical equipment and food .
FLE:	*You have old system*
SME:	do you have any suggestions ?
FLE:	*And I have done*
SME:	are you familiar with roads in this area ?
FLE:	*And this area is all dirt paved roads we know*
SME:	do you know a good road ?
FLE:	*And I know all this garbage paved roads*
SME:	where are the criminal groups near these roads ?
FLE:	*The roads fees criminal vehicles there*
SME:	what can you give me for security ?
FLE:	*I have four guys about guns trucks for you*
SME:	I will use these gun trucks .
FLE:	*That's very good*
SME:	how many trucks and how many men do you have ?
FLE:	*How many times the trucks and four guns person*
SME:	how many trucks ?
FLE:	*Four men and two ways*
SME:	how many trucks only ?
FLE:	*There are four two trucks*
SME:	only two vehicles ?
FLE:	*There are four guard with them*
SME:	I understand your plan and will take your supplies .
FLE:	*That's very good*

Table 5: Sample MT Transcripts

2.11. Automatic Scores

Machine translation is most typically evaluated with automatic methods such as BLEU and METEOR. They are typically used because of convenience and interpretability within the research and development community, and we have performed those measurements as well, as shown in

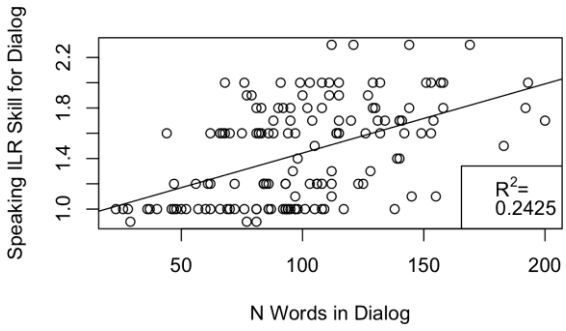

Figure 11: Words in SME Transcript and SME ILR Level for Speaking

Figure 12.[7]
The automatic scores were not correlated with the overall PASS/FAIL scores, either for BLEU or for METEOR. However, BLEU and METEOR scores correlate with each other, more strongly when compared at the dialog level and less strongly when the dialog scores are averaged over the role-player teams per scenario. One fundamental issue is one of granularity. The PASS/FAIL scores are assigned for each dialog, a relatively large unit. These are averaged over the performance by several role players, half of whom perform the scenario in the MT condition. Automatic scores such as BLEU and METEOR can be averaged over many different levels of granularity. However, collapsing the automatic scores to the scenario level throws away information. As might be expected, the BLEU scores correlate with METEOR better at the dialog transcript level ($R^2 = 66\%$) than at the scenario level, which averages the dialog performance over the role players for that scenario ($R^2 = 36\%$). Fitting a linear regression line onto the BLEU and METEOR data on this data set would show a negative correlation with both BLEU and METEOR. However, due to the small number of scenarios (there are only 20 scenarios in the role play), it would be a mistake to read too much into a negative correlation. The appropriate observation is to note that the automatic scores are not a reliable predictor of PASS/FAIL scores.
It is plausible, however, that a negative correlation could exist. One other fundamental issue is the difference between interactive machine translation and batch processing. In these interactive scenarios, a role player has the opportunity to try again after recognizing a bad translation. In fact, the more the role-player works to repair the communication, the longer trail that might be left of failed translations. These failures contribute to the BLEU score and may outweigh the successful translations in their quantity. For example, in the transcript in Table 6, the SME is very persistent in working around communication failures. Ultimately the SME was able to achieve this goal, despite failed interchanges within the dialog. Each of those failed interchanges would have contributed to a worse BLEU score despite an overall PASS for that dialog.

[7] We did not try to normalize the transcripts in producing these scores.

	Civil Affairs Scenario CA54, Team 9		
SME:	thank you for meeting with me .	SME:	by having your public trust your court
FLE:	*Not coming on us*		your town will be more functional .
SME:	what did you say ?	*FLE:*	*We have*
FLE:	*All the last*	SME:	what do you need from us to help you ?
SME:	what's wrong with your courts ?	*FLE:*	*In our security is*
FLE:	*Every work is security*	SME:	is all you need security ?
SME:	why are trials not happening ?	*FLE:*	*We have*
FLE:	*They are hidden nervous*	SME:	what did you say ?
SME:	why are you nervous ?	*FLE:*	*The security is for the country*
FLE:	*Put under the work*	SME:	I'm concerned about your court system
SME:	your public needs to trust you and be safe .		not your country .
FLE:	*He can be*	*FLE:*	*If we have security*
SME:	what did you say ?	SME:	to confirm all you need is security from me .
FLE:	*Our security officials work*	*FLE:*	*We were*
SME:	will you let us help you ?	SME:	okay . we will send people over tomorrow .
FLE:	*It is*	*FLE:*	*Thank you very much it*

Table 6: Working Through Garbled Translation

2.12. Contrast with Other Manual Methods

Manual methods such as Concept Transfer Rate have been used for speech-to-speech machine translation evaluation, for example (Sanders et al., 2013). Concept transfer rates are based on counting the number of key concepts communicated within a fixed time period. The new method that we describe here departs from these conventions by allowing the role players greater freedom to accomplish their tasks. One of the primary motivations was to avoid the risk of over-scripting role player behavior. For example, if communication breaks down early in the scenario, we do not want the role players to work from a script to artificially repair the situation. The main risk to the experiment is if the role players fail to properly perform their tasks even in the HT condition. To mitigate this risk, we only employed role players who had performed the required tasks in a recent military deployment, and we provided a full day of training using 10 additional scenarios from the same domains as the evaluation scenarios.

3. Acknowledgements

This work builds on prior work on ILR-Based Machine Translation evaluation performed by MIT Lincoln Laboratory, the National Institute of Standards and Technology and the US Army Machine Foreign Language Translation System program. We used the "Walkie-Talkie Over IP" (WOIP) system developed by Wade Shen as the communication system. He was instrumental in the design of the previous experiments for MFLTS, for which we wish to thank him. The original scenarios were developed at a military training exercise in 2009 with assistance from Patrick O'Malley, a retired US Army officer who supports TRADOC at Fort Huachuca, and Neil Granoien, a retired Vice Chancellor from the DLI who was working with us at that time. The role plays were conducted in Professor Ted Gibson's laboratory at MIT. We also wish to thank the role players: the military veterans who participated as SMEs, and the native Pashto speakers who participated as FLEs.

4. Conclusion

The key lessons we learned in the experiment is that it is not enough for the role players to express themselves. What they say has to be understood by their conversational partner for the scenario to be completed successfully. Moreover, success depended as much on task simplicity as it did upon the translation condition, given that 67% of the base case scenarios were successfully completed using MT, but only 35% of the contrasting scenarios with even minor obstacles received passing scores. In other words, we observed that MT had the greatest chance of success when the task was simple and the language complexity needs were low.

We feel that our earlier work suggested that ILR proficiency Level 2 was the ideal level for text translation of existing texts. That may be because concrete facts are more readily transferable from one language to another. Similarly, in speech to speech translation, Level 1 is may be ideal because such language focuses on the everyday, the here and now, the immediate situation discernible to all parties. The task is to work out a solution to a simple issue or problem within this situation. (*e.g. The road is blocked; how do I get home?*)

The implication of the performance variation shown in these results is that the technology should be tested in advance for specific situations in which it might be used, and not to assume that it just "works" for a particular broad domain.

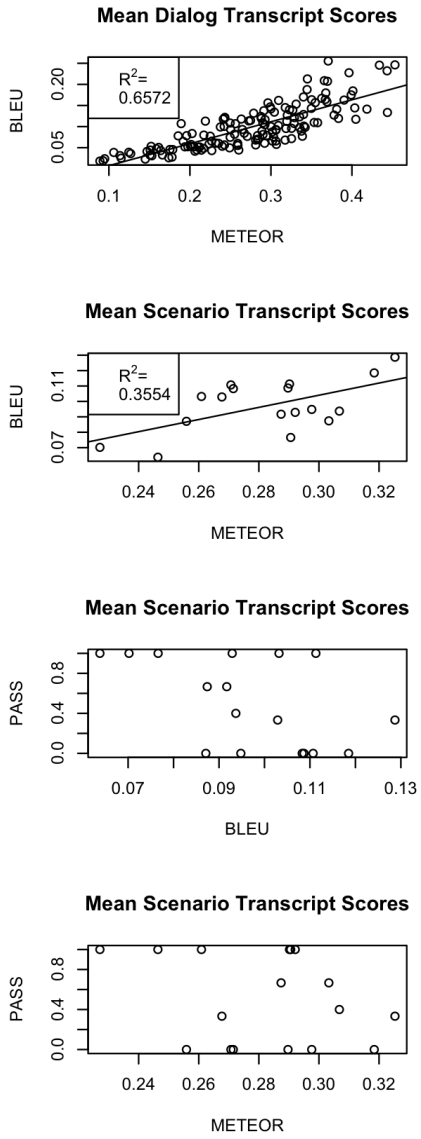

Figure 12: PASS/FAIL compared with BLEU and METEOR

5. References

ILR. (2014). Interagency language roundtable (website). http://www.govtilr.org/.

Jones, D., Herzog, M., Ibrahim, H., Jairam, A., Shen, W., Gibson, E., and Emonts, M. (2007). Ilr-based mt comprehension test with multi-level questions. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 77–80. Association for Computational Linguistics. Sanders, G. A., Weiss, B. A., Schlenoff, C., Steves, M. P., and Condon, S. (2013). Evaluation methodology and metrics employed to assess the transtac two-way, speech-to-speech translation systems. *Computer Speech & Language, 27(2):528–553.*

USArmy. (2014). Us army doctrine and training publications (website). http://armypubs.army.mil/doctrine/.

Relating Translation Quality Barriers to Source-Text Properties

Federico Gaspari[*], Antonio Toral[*], Arle Lommel[^], Stephen Doherty[°], Josef van Genabith[§], Andy Way[*]

[*] School of Computing
Dublin City University
Glasnevin
Dublin 9
Ireland

[^] DFKI GmbH
Language Technology
Alt Moabit 91c
D-10559 Berlin
Germany

[°] School of Humanities & Languages
University of New South Wales
Sydney 2052
Australia

[§] DFKI GmbH
Language Technology
Campus D3 2
D-66123 Saarbrücken
Germany

E-mail: {fgaspari, atoral, away}@computing.dcu.ie, arle.lommel@dfki.de, s.doherty@unsw.edu.au, josef.van_genabith@dfki.de

Abstract

This paper aims to automatically identify which linguistic phenomena represent barriers to better MT quality. We focus on the translation of news data for two bidirectional language pairs: EN↔ES and EN↔DE. Using the diagnostic MT evaluation toolkit DELiC4MT and a set of human reference translations, we relate translation quality barriers to a selection of 9 source-side PoS-based linguistic checkpoints. Using output from the winning SMT, RbMT, and hybrid systems of the WMT 2013 shared task, translation quality barriers are investigated (in relation to the selected linguistic checkpoints) according to two main variables: (i) the type of the MT approach, i.e. statistical, rule-based or hybrid, and (ii) the human evaluation of MT output, ranked into three quality groups corresponding to good, near miss and poor. We show that the combination of manual quality ranking and automatic diagnostic evaluation on a set of PoS-based linguistic checkpoints is able to identify the specific quality barriers of different MT system types across the four translation directions under consideration.

Keywords: MT quality barriers, diagnostic evaluation, statistical/rule-based/hybrid MT, linguistic features

1. Introduction

This study was conducted as part of the European Commission-funded project QTLaunchPad (Seventh Framework Programme (FP7), grant number: 296347), preparing groundwork for major developments in translation technology, with a special focus on identifying and overcoming barriers to translation quality.[1] Key goals of the project include providing test suites and tools for translation quality assessment, creating a shared quality metric for human and machine translation (MT), and improving automatic translation quality estimation. The project involves key research and industrial stakeholders interested in improving translation technology.

This paper presents work on the identification of translation quality barriers, one of the central objectives of QTLaunchPad. Given the widely perceived need to enhance MT quality and the reliability of MT evaluation for real-life applications, which has been confirmed further by QTLaunchPad surveys,[2] this study is of potential interest to a variety of MT users and developers. The main motivation behind the research is to systematically tackle quality barriers in MT, investigating closely the relationship between different types of MT systems, the overall quality of their output and the properties of the input. A key part of the work conducted in QTLaunchPad addresses this problem, with the goal of improving MT performance and extending its applicability.

Our study focuses on identifying the source-side linguistic properties that pose MT quality barriers for specific types of MT systems (statistical, rule-based and hybrid) and for output representative of different quality levels (poor-, medium- and high-quality) in four translation combinations, considering English to and from Spanish and German. Many commentators say that developers of SMT systems (in particular) are not able to predict which linguistic phenomena their systems are capable of handling. In this paper, on the contrary, we demonstrate the potential of combining manual MT quality ranking and DELiC4MT (an automatic diagnostic MT evaluation toolkit focusing on source-side linguistic phenomena that is described in more detail in Section 2.1) to identify translation quality barriers.

The remainder of the paper is organised as follows. After this introduction, Section 2 presents DELiC4MT, focusing on the novelty of its application to the discovery of translation quality barriers. Section 3 covers the evaluation, including the experimental set-up, the results for each of the four translation directions (paying special attention to the identified translation quality barriers in relation to the MT system types and to the quality rankings assigned to their output) and further correlation analysis. Finally, Section 4 summarises the main findings of the study and outlines possibilities for future work.

[1] www.qt21.eu/launchpad
[2] www.qt21.eu/launchpad/sites/default/files/QTLP_Survey2i.pdf

2. DELiC4MT for the Analysis of Translation Quality Barriers

2.1 DELiC4MT: an Open-Source Toolkit for Diagnostic MT Evaluation

DELiC4MT is an open-source toolkit for diagnostic MT evaluation (Toral et al., 2012).[3] Its diagnostic dimension derives from its ability to focus on user-defined linguistic checkpoints, i.e. phenomena of the source language that the user decides to analyse when evaluating the quality of MT output. Linguistic checkpoints can correspond to interesting or difficult lexical items and/or grammatical constructions for which a specific translation quality assessment is required. They can be defined at any level of granularity desired by the user, considering lexical, morphological, syntactic and/or semantic information.

Any of these layers of linguistic description can be combined to create checkpoints of variable composition, ranging from very basic and generic (e.g. focusing on any noun found in the input) to very complex and specific (e.g. all word sequences in the source text composed of a determiner, followed by any singular noun, followed by the literal word "of", followed by any plural noun, followed by a finite form of the verb 'go', etc.). The only constraint on the design of linguistic checkpoints for DELiC4MT is that they should consist of features supported by the language resources and processing tools previously used to annotate the data sets (most often a PoS tagger); clearly, some languages are better served than others in this respect. The data pre-processing steps that are required to use DELiC4MT are described in Section 3.1, while Section 3.3.2 discusses the way in which its output is presented to the user.

DELiC4MT produces a score, indicating how many of the relevant checkpoints detected on the source side were translated correctly by the MT system under investigation. This diagnostic feedback can then be incorporated into the further development, fine-tuning and customisation of the MT software to optimise its performance. One advantage of the toolkit over standard automatic MT evaluation metrics such as BLEU (Papineni et al., 2002) is that it supports more flexible, transparent and fine-grained evaluation: the scores of automatic MT evaluation metrics are often difficult to interpret and do not always help one to understand the actual linguistic strengths and weaknesses of an MT system.

Toral et al. (2012) describe the different modules that make up the DELiC4MT toolkit and present a step-by-step case study of how it can be applied to a specific language pair for an illustrative linguistic checkpoint defined by the user. A tutorial is also available, showing how the toolkit works, applying it to a specific language pair, test set and linguistic checkpoint.[4] DELiC4MT is also available via a web application and a web service, which are more convenient for users who wish to avoid the burden of installing, configuring and maintaining the software (Toral et al., 2013). The toolkit is language-independent and can be easily adapted to any language pair; it has, for example, been successfully applied to the diagnostic evaluation of MT quality for European language pairs (e.g. Naskar et al., 2011; Naskar et al., 2013), as well as for English in combination with Indian languages (Balyan et al., 2012; Balyan et al., 2013) on a range of checkpoints specific to the respective source languages.

2.2 DELiC4MT-Based Analysis of Translation Quality Barriers

DELiC4MT has so far been used to evaluate the overall quality of MT systems with respect to their performance on user-defined source-side linguistic phenomena. The novelty of the work presented in this paper lies in the application of this toolkit to the investigation of translation quality barriers. These are investigated with DELiC4MT according to two main variables. Firstly, we consider different MT system types: this variable enables us to compare the performance of statistical (SMT), rule-based (RbMT) and hybrid (HMT) MT software on a selection of source-language linguistic checkpoints, which are explained in more detail in Section 2.3. We thus have a clear view of those quality barriers encountered by the various types of MT software for each translation direction, broken down according to a range of checkpoints as salient linguistically-motivated morphosyntactic units of evaluation.

Secondly, we look at human quality rankings of the MT output: this variable concerns the quality band assigned by human evaluators to the output of each MT system, whereby each sentence was rated as either good (rank 1), near-miss (rank 2) or poor (rank 3). We are thus able to evaluate the performance of the MT systems on each checkpoint separately for those sentences that fall into each of these rating bands. Both variables under consideration lend themselves to comparative evaluations, which are investigated in Section 3 with a view to shedding light on translation quality barriers.

2.3 From Linguistic Checkpoints to Translation Quality Barriers

On the basis of some preliminary tests, we decided to focus our analysis on linguistic checkpoints consisting of individual PoS classes (rather than PoS sequences), which were deemed sufficiently fine-grained to obtain interesting and useful information on translation quality barriers. This decision mitigated the data sparseness problems that we would have run into using more elaborate and specific linguistic checkpoints, given the limited amount of data available (cf. Section 3.1).

Following some explorations of the possibilities, we eventually selected 9 linguistic checkpoints for our analysis, consisting of the following individual PoS classes: adjectives (ADJ), adverbs (ADV), determiners (DET), common nouns (NOC), nouns (NOU, combining NOC and NOP), proper nouns (NOP), particles (PAR),

[3] www.computing.dcu.ie/~atoral/delic4mt/

[4] http://github.com/antot/DELiC4MT/blob/master/doc/tutorial/delic4mt_tutorial.pdf

pronouns (PRO) and verbs (VER). These are grouping abstractions over the possibly different PoS sets used for the three languages under investigation, and we thought that they represented a reasonable balance between granularity and high-level description. The scores provided in the analysis for any of these checkpoints express the ratio between all the instances of the checkpoint detected on the source side and those that were translated correctly by the MT system in question. Thus, the lower the score for a checkpoint, the worse the quality of the translations in the MT output for words corresponding to that linguistic phenomenon (in this study, PoS class) in the input, which reveals a potential translation quality barrier when referenced against the human evaluation of the output.

3. Evaluation

3.1 Data, Pre-Processing and Experimental Set-Up

We conducted this analysis of translation quality barriers focusing on news data, relying on the 2013 WMT data sets for which human reference translations were available. Table 1 shows the data used for the evaluation, detailing the number of sentences and the types of MT systems available for each translation direction.

Translation Direction	Number of Sentences	MT Systems
EN→ES	500	SMT, RbMT, HMT
ES→EN	203	SMT, RbMT
EN→DE	500	SMT, RbMT, HMT
DE→EN	500	SMT, RbMT

Table 1: Datasets used for the evaluation.[5]

The sentences used were translated by the winning MT systems from the 2013 WMT shared task. In this case, the SMT system is a phrase-based system from one of the leading European academic teams in MT research, while both the RbMT and HMT systems are leading systems on the market nowadays. Since these systems were used in the shared task, they had training/reference data consisting of news articles and the translations were all of novel sentences from news articles. It should be noted that WMT uses paid human translators to generate source sentences in all language pairs, so, for example, a segment authored in Spanish would be translated by a human into German and then translated from German by the MT systems into the various WMT target languages (including back into Spanish). To control for the issue of "pivot" or "relay" translation, our corpus used only "native" source segments, i.e., those segments authored in the source language of each language pair we considered.

Two Language Service Providers (LSPs) plus an in-house team at DFKI (for ES→EN only) carried out human assessment of the quality of the MT output for these sentences in the various language pairs, ranking them into three quality categories: rank 1 (perfect output, not requiring any editing to be published); rank 2 (near-misses, i.e. sentences with fewer than 3 errors, thereby deemed to be easily post-editable); and, finally, rank 3 (poor-quality output, with 3 or more errors, requiring time-consuming and resource-intensive post-editing).

In order to pre-process the data so that it could be used with DELiC4MT, we PoS-tagged the source and target sides of the references. Freeling[6] (Padró and Stanilovsky, 2012) was used for English and Spanish, with TreeTagger[7] (Schmid, 1995) used for German. Subsequently, the source and target sides of the reference were word-aligned with GIZA++ (Och and Ney, 2003). As the reference datasets were rather small for word alignment, in order to obtain alignments of higher quality, they were appended to a bigger corpus of the same domain as the WMT data (news commentary),[8] before performing word alignment. Once the alignment was ready, we extracted the subset that corresponded to the sentences of the reference set and discarded the rest.

Before proceeding further, we need to provide clarification regarding the data sets used. For each of the four translation directions, the diagnostic evaluation presented here concerns the very same input when comparisons of MT systems take place on the whole input data. In contrast, this is not the case for the identification of the quality barriers considering the MT output categorised according to the three quality rankings. This is because DELiC4MT was run separately on a subset of the input, depending on how that subset was classified by the human judges, resulting in three different data sets divided according to their quality. This means, for example, that the subset of rank 1 sentences translated with the SMT system for EN→ES is different from the subset of the same rank and translation direction for the RbMT system, so no direct comparison is possible in such cases.

3.2 Results

This section presents in turn the results obtained with DELiC4MT (Y axis in the figures below) on the 9 chosen linguistic checkpoints (X axis in the figures) for each of the four translation directions. This enables us to directly relate the translation quality barriers identified for each MT system type as well as across the three quality rankings to specific source-text properties. The figures in this section presenting the data (1-16) would be better represented by scatter plots. However, some of the data points for individual PoS-based linguistic checkpoints for the different MT system types are very close, which makes it difficult to differentiate them. As a result, in the interest of clarity, all the figures 1-16 include the trend lines connecting the data points for the various PoS-based linguistic checkpoints.

[5] At the time of writing, the ES→EN data has only been partially rated, resulting in a smaller number of data points for this translation direction.

[6] http://nlp.lsi.upc.edu/freeling/

[7] www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[8] www.statmt.org/wmt13/translation-task.html#download

3.2.1. Results for EN→ES

One overall finding for the EN→ES language pair is that the SMT system is the best in general, followed by HMT and RbMT (in this order), even though SMT receives (virtually) the same scores as HMT for the PAR, PRO and VER linguistic checkpoints.

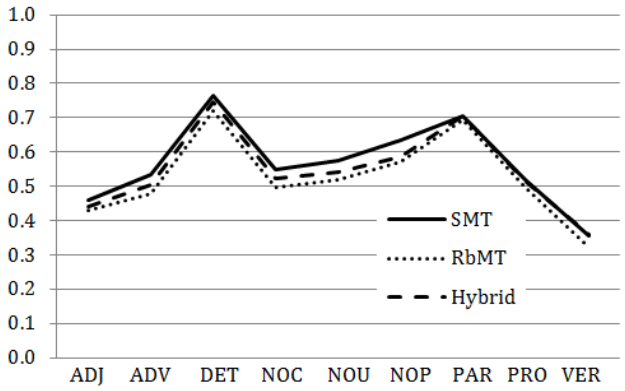

Figure 1: EN→ES results (overall).

Considering the top-ranking translations, SMT and HMT perform best for different linguistic checkpoints (except for NOC, where there is a tie). RbMT is on a par with SMT only for ADJ and NOC; otherwise it clearly lags behind the other two MT systems. It is particularly striking that HMT has a noticeably higher score than SMT for the VER checkpoint, corresponding roughly to a 10% improvement in relative terms; the difference is even more marked between HMT and RbMT, which has the worst performance on VER as far as high-quality translations are concerned.

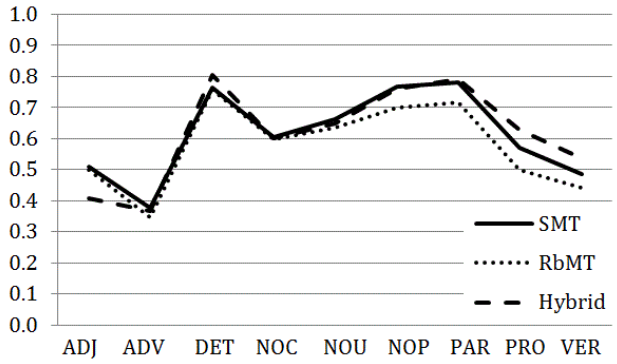

Figure 2: EN→ES results for rank 1.

As can be seen in Figure 3, rank 2 translations show similar results for all three systems, with RbMT lagging slightly behind, especially for ADV, NOC and VER. Equivalent trends can be observed in Figure 4 for rank 3 translations, with SMT obtaining an even bigger advantage on DET, NOU and NOP. For these three checkpoints HMT comes second, and RbMT last. However, we observe that the results by the three MT system types are very similar for the remaining checkpoints.

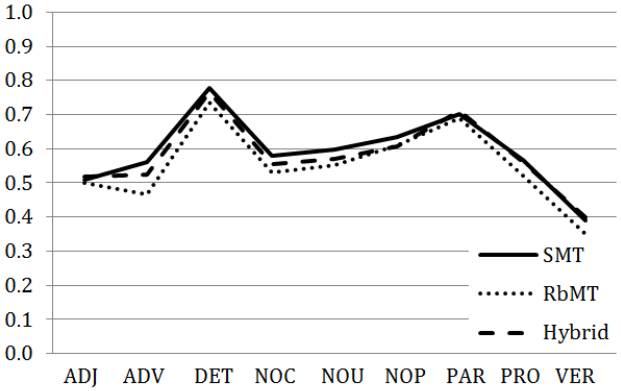

Figure 3: EN→ES results for rank 2.

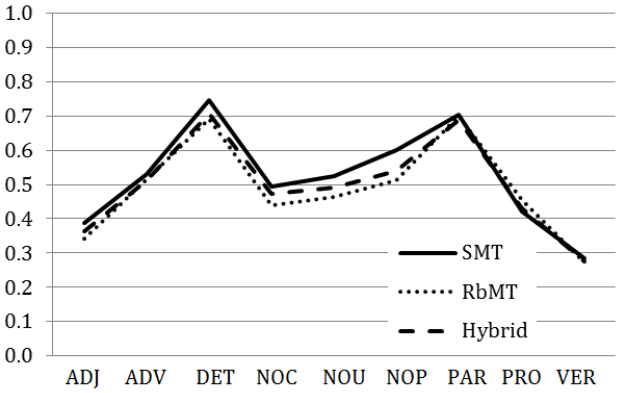

Figure 4: EN→ES results for rank 3.

3.2.2. Results for ES→EN

In overall terms, for the ES→EN translation direction the performance of SMT is consistently better than that of RbMT for all the 9 linguistic checkpoints, with approximately a 10% relative difference in the respective DELiC4MT scores. Particularly severe quality barriers for RbMT seem to be ADV, NOC, PRO and VER.

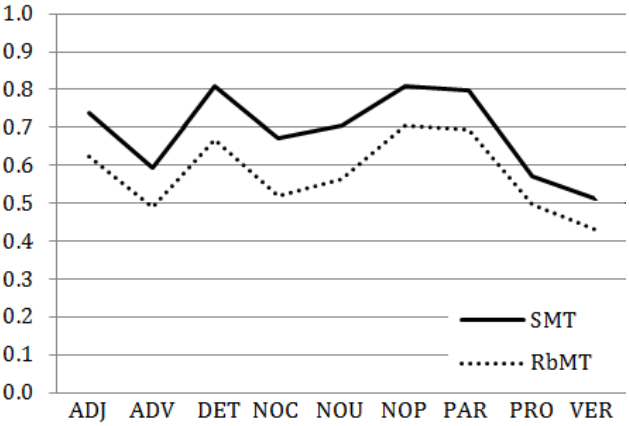

Figure 5: ES→EN results (overall).

More specifically, for rank 1 translations (bearing in mind the comparatively small numbers of checkpoint instances with respect to the other two quality bands, cf. Section 3.3.1 and in particular Tables 2 and 3), the performance of RbMT is particularly modest for ADJ, NOC, NOU, PAR, PRO and VER (Figure 6). On the other hand, SMT and

RbMT have very similar performances for ADV and NOP in rank 1 translations, showing that these two categories are not specifically affected by quality differences depending on the two MT system types.

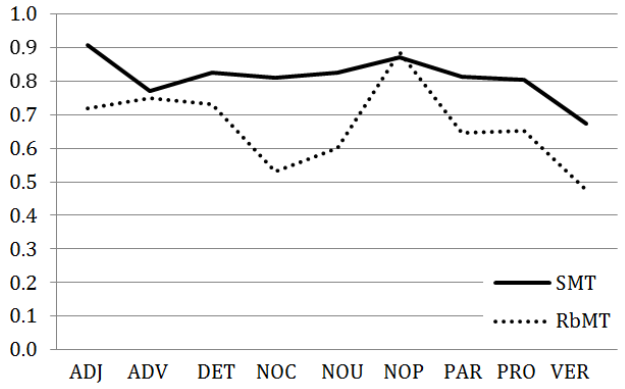

Figure 6: ES→EN results for rank 1.

The rank 2 translations show that SMT outperforms RbMT by a similar margin across all the linguistic checkpoints (Figure 7). As a result, in this case, the breakdown into the linguistic checkpoints does not allow us to gain particularly useful insights, showing that translation quality barriers are fairly consistent across the board for all the considered PoS-based linguistic checkpoints in near-miss translations, regardless of the type of MT system that generated them.

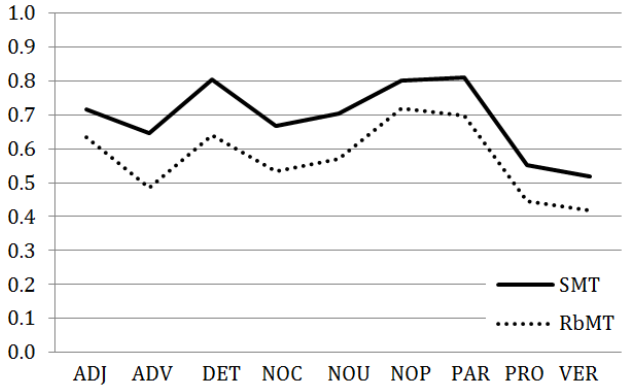

Figure 7: ES→EN results for rank 2.

The situation is more interesting for the rank 3 translations, where both SMT and RbMT show specific weaknesses in the translation of ADV, PRO and VER (Figure 8). Interestingly, although these three checkpoints show the lowest scores, they are also the ones where RbMT performs better than SMT, by a clear margin. For the remaining six checkpoints, the SMT output obtains higher scores, with a difference of approximately 10% in value at times.

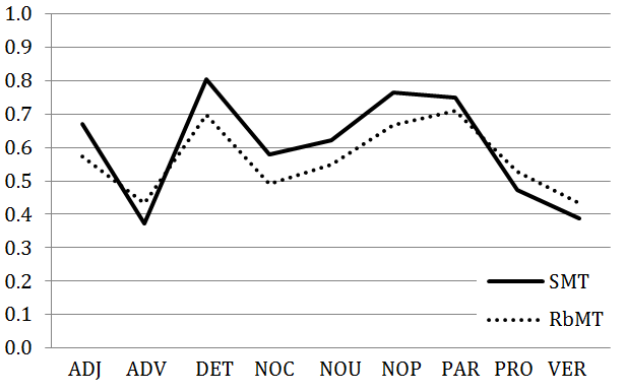

Figure 8: ES→EN results for rank 3.

3.2.3. Results for EN→DE

For the EN→DE translation direction, in overall terms, the performance of the three systems is very similar, with SMT giving slightly better scores than RbMT for all the checkpoints, while also beating HMT most of the time, except for PAR (where there is a tie), PRO and VER. As a result, it is difficult to identify prominent translation quality barriers from this analysis, except for a comparatively poor performance of RbMT, particularly for ADJ, NOC, NOU and PAR.

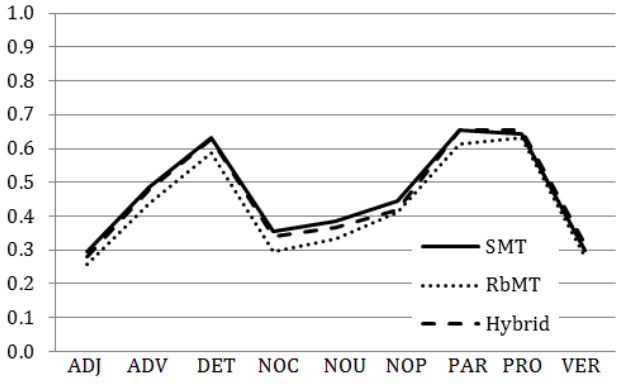

Figure 9: EN→DE results (overall).

Looking at the results by ranking, on the other hand, gives a more interesting picture. For rank 1 translations, SMT shows a particularly disappointing performance for NOC and NOU, while it is by far the top system for ADJ, NOP and PAR (Figure 10). RbMT receives the lowest score of the three systems for the ADJ checkpoint, where HMT also performs particularly badly. RbMT also showed the worst performance for VER, where HMT came out on top.

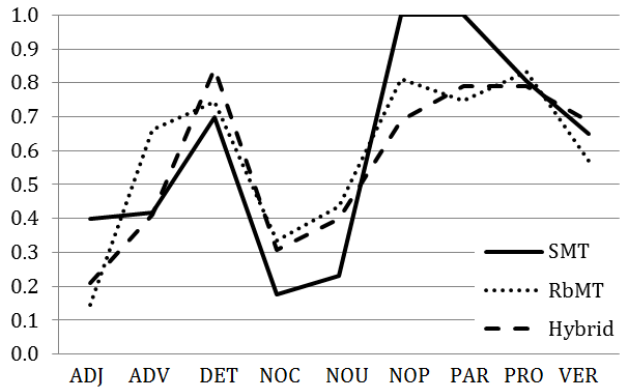

Figure 10: EN→DE results for rank 1.

The rank 2 translations (Figure 11) show a consistent trend, with SMT obtaining the best results for all the checkpoints (there is a tie with HMT for verbs), and RbMT lagging behind.

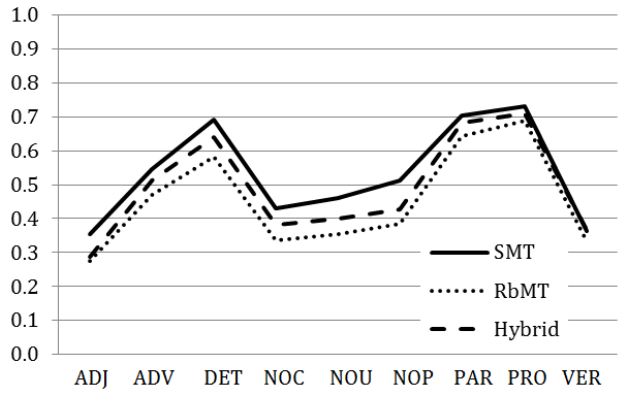

Figure 11: EN→DE results for rank 2.

Finally, looking at rank 3 translations (Figure 12), all three MT systems find ADJ and VER similarly problematic to translate (which was to be expected, due to a large extent to agreement problems), whereas RbMT runs into noticeably more difficulties with NOC. For the remaining checkpoints the scores of the three MT systems do not show clear differences, hence we cannot identify other particularly severe or interesting translation quality barriers for translations of modest quality in this particular translation direction.

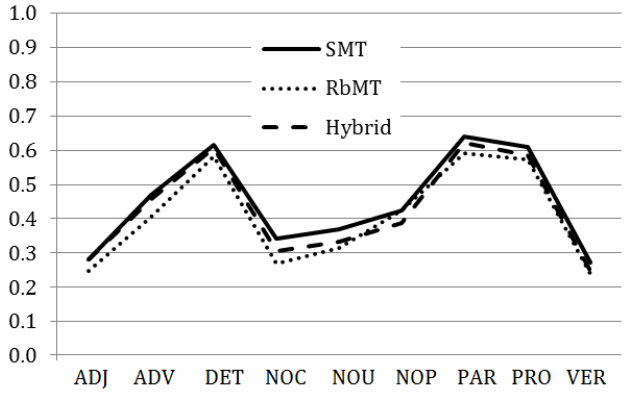

Figure 12: EN→DE results for rank 3.

3.2.4. Results for DE→EN

Finally, coming to the DE→EN translation direction, whose overall results are summarised in Figure 13, both SMT and RbMT encounter specific difficulties with the translation of ADJ, NOC and NOU checkpoints, with similarly low performance levels (the scores of RbMT are slightly lower in all these three cases). In contrast, DELiC4MT reveals that there are only relatively minor problems for the translation of DET, where both systems perform very well – determiners are much easier to translate from German into English, due to the much smaller set of non-inflected options available in the target. The other checkpoints show equivalent scores, but RbMT is comparatively weaker, especially for ADV and PRO.

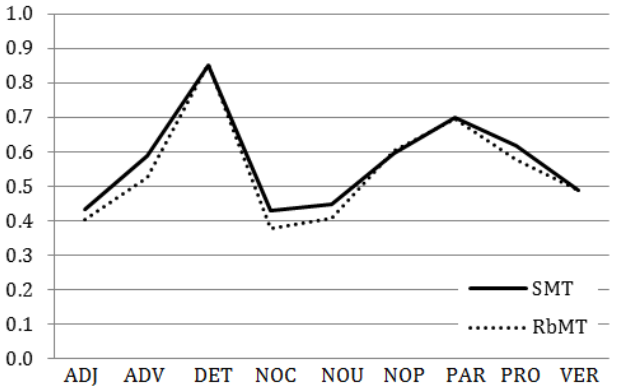

Figure 13: DE→EN results (overall).

With regard to the translation quality ranks, RbMT receives distinctly lower scores for ADV, NOP, PAR and PRO when considering the rank 1 translations (Figure 14). On the other hand, the performance of SMT is particularly bad for adjectives (20% lower than RbMT), thus pointing to a clear quality barrier within the better translations. RbMT also obtains a better evaluation than SMT for DET, where RbMT translates correctly 97.6% of them.

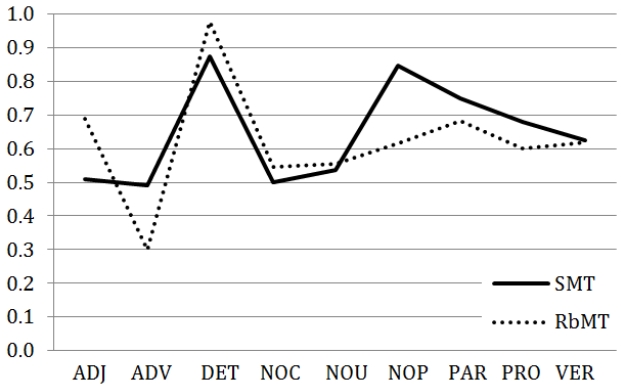

Figure 14: DE→EN results for rank 1.

As far as rank 2 translations are concerned (Figure 15), the performance of SMT and RbMT is very similar across all the checkpoints: some are handled slightly better by SMT (e.g. ADJ, NOU and PRO), while in particular for NOP the score of RbMT is higher.

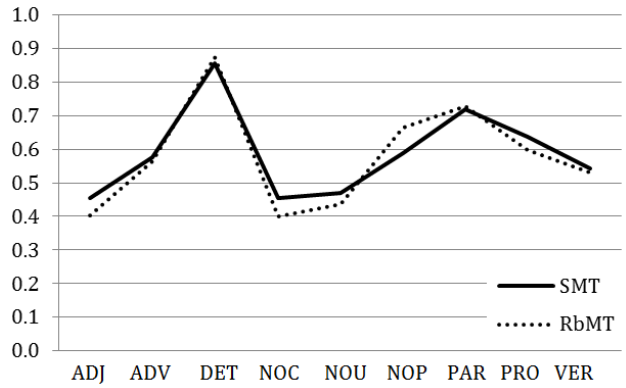

Figure 15: DE→EN results for rank 2.

Finally, for rank 3 translations (Figure 16) the performance tends to be equivalent again for most checkpoints, but RbMT struggles more with ADV, NOC and NOU. On the other hand, for these low-quality translations SMT seems to find more serious barriers in the translation of VER, for which RbMT receives a 5% higher score.

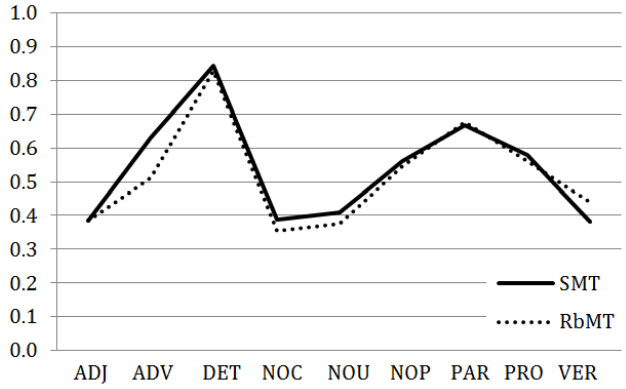

Figure 16: DE→EN results for rank 3.

3.3 Analysis

3.3.1. Correlations between Human Ratings and DELiC4MT

It should be noted that across all of the translation directions and MT system types, there tend to be comparatively few rank 1 translations, i.e. those rated as high-quality by the human judges. This considerably reduces the number of checkpoints detected in the input for those subsets of the data, thus making it particularly difficult to draw reliable generalisations in such circumstances, due to data sparseness problems. In Tables 2 and 3 we thus provide the number of instances of each checkpoint detected on the source/input side of the various data subsets (Table 2 shows the EN--ES language pair, and Table 3 presents the data for EN--DE), to help put in perspective the DELiC4MT scores and our findings in terms of translation quality barriers across the three ranks. For both language pairs there are much higher numbers of detected checkpoints for the rank 2 and rank 3 quality bands. When looking at SMT, RbMT and HMT alike, we can therefore be more confident in the analysis of our findings for near-miss and poor translations, whereas particular caution must be exercised when interpreting the results of rank 1 (i.e. good) translations.

		SMT		RbMT		HMT	
		ES>EN	EN>ES	ES>EN	EN>ES	ES>EN	EN>ES
RANK1	ADJ	53	59	27	44	-	71
	ADV	24	69	17	55	-	82
	DET	119	68	67	37	-	66
	NOC	193	197	100	97	-	183
	NOU	254	304	124	150	-	258
	NOP	61	107	24	53	-	75
	PAR	134	110	76	67	-	100
	PRO	41	56	29	48	-	70
	VER	177	193	157	102	-	198
RANK2	ADJ	196	492	196	496	-	448
	ADV	119	443	109	443	-	394
	DET	335	569	327	575	-	527
	NOC	639	1723	662	1772	-	1655
	NOU	853	2508	823	2499	-	2390
	NOP	214	785	161	727	-	735
	PAR	482	1156	459	1153	-	1073
	PRO	125	512	127	510	-	450
	VER	786	1449	687	1515	-	1373
RANK3	ADJ	70	380	93	412	-	433
	ADV	51	302	68	327	-	352
	DET	132	340	188	373	-	393
	NOC	286	1202	354	1275	-	1308
	NOU	371	1649	528	1840	-	1852
	NOP	85	447	174	565	-	544
	PAR	163	729	240	793	-	845
	PRO	72	274	81	289	-	329
	VER	273	1016	388	1069	-	1117

Table 2: Numbers of checkpoint instances detected on the source side for the EN--ES language pair.

		SMT		RbMT		HMT	
		DE>EN	EN>DE	DE>EN	EN>DE	DE>EN	EN>DE
RANK1	ADJ	173	5	45	20	-	38
	ADV	63	12	20	16	-	38
	DET	102	10	42	13	-	29
	NOC	356	40	152	44	-	140
	NOU	396	43	178	56	-	185
	NOP	39	3	26	12	-	45
	PAR	152	5	47	16	-	48
	PRO	87	16	35	18	-	30
	VER	179	40	89	42	-	75
RANK2	ADJ	591	180	587	360	-	419
	ADV	203	156	195	252	-	275
	DET	479	191	438	373	-	434
	NOC	2023	593	1655	1201	-	1312
	NOU	2294	905	1921	1878	-	1995
	NOP	270	312	264	677	-	683
	PAR	673	290	581	678	-	743
	PRO	298	167	278	317	-	336
	VER	680	493	640	896	-	1086
RANK3	ADJ	536	708	673	512	-	436
	ADV	203	488	254	388	-	343
	DET	403	749	504	560	-	487
	NOC	1737	2551	2299	1930	-	1732
	NOU	1995	3771	2574	2766	-	2539
	NOP	258	1220	275	836	-	807
	PAR	581	1435	778	1034	-	939
	PRO	251	509	323	355	-	326
	VER	578	1801	708	1396	-	1173

Table 3: Numbers of checkpoint instances detected on the source side for the EN--DE language pair.

To ascertain the correlation between the DELiC4MT scores and the human evaluations, we calculated Pearson's *r* values for the DELiC4MT scores and the human ratings. This correlation concerns individual PoS-based checkpoints and the human quality ranking of MT output of whole sentences for which the relevant checkpoint is detected by DELiC4MT on the source side. Noise introduced by the PoS tagger and the word aligner might have an impact on these results, however our previous work (Naskar et al., 2013) shows rather conclusively that the noise introduced by state-of-the-art PoS taggers and word aligners does not have a noticeable impact on DELiC4MT results. Normalising these results to positive values (since the rating is an inverse scale) gives the results shown in Figure 17.

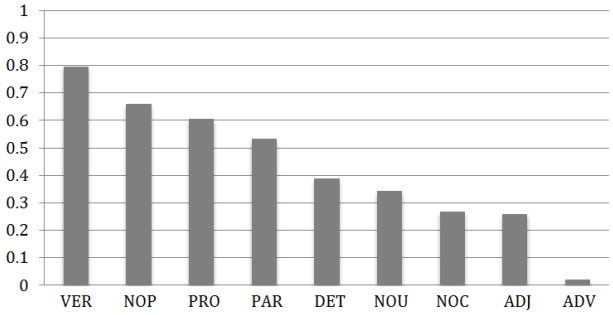

Figure 17: Pearson's *r* correlation for DELiC4MT scores and human quality ratings.

As seen in Figure 17, quality levels correlated with all of the PoS-based linguistic checkpoints, but many of the correlations are weak at best (and virtually nonexistent for ADV). Assuming significance over the approximately 4,400 translations examined, it seems that the VER, NOP, and PRO checkpoints stand out as the best predictors of human quality assessment. PAR and DET also exhibit reasonable correlation.

Figures 18 and 19 show the score clustering at each quality level along with the trend lines for each checkpoint. In an examination of human error annotation (Burchardt et al., 2014:39-4l) we found that PAR and DET were among the most problematic PoS classes for MT in general, and especially for RbMT. An examination of the respective scores and trend lines as shown in Figures 18 and 19 reveals that MT systems were generally quite reliable in producing high DELiC4MT scores for these items, with among the highest scores for these checkpoints across all quality bands. While they differentiate between the bands, due to the low standard deviation evident in each cluster, the differentiation is also quite small.

Furthermore, the use of particles and determiners is among the most variant of grammatical features across languages, and accurately transferring these items between languages is quite likely to be error prone. Accordingly, although the MT systems were consistently good in matching these two checkpoints, an examination of human error markup shows that a high DELiC4MT score for these two checkpoints is not necessarily a good predictor of overall quality (approximately 15% of the errors annotated in the corpus described in Burchardt et al. (2014) had to do with so-called "function words" such as particles and determiners), unlike VER, NOP, and PRO, where a high degree of correspondence between presence of these checkpoints in both source and target would generally be a good predictor of accuracy and quality. Thus a comparison of these results with the findings of the human annotation task described in Burchardt et al. (2014) shows that automatic analysis, such as this study carries out, can contribute to a better understanding of human annotation and vice-versa.

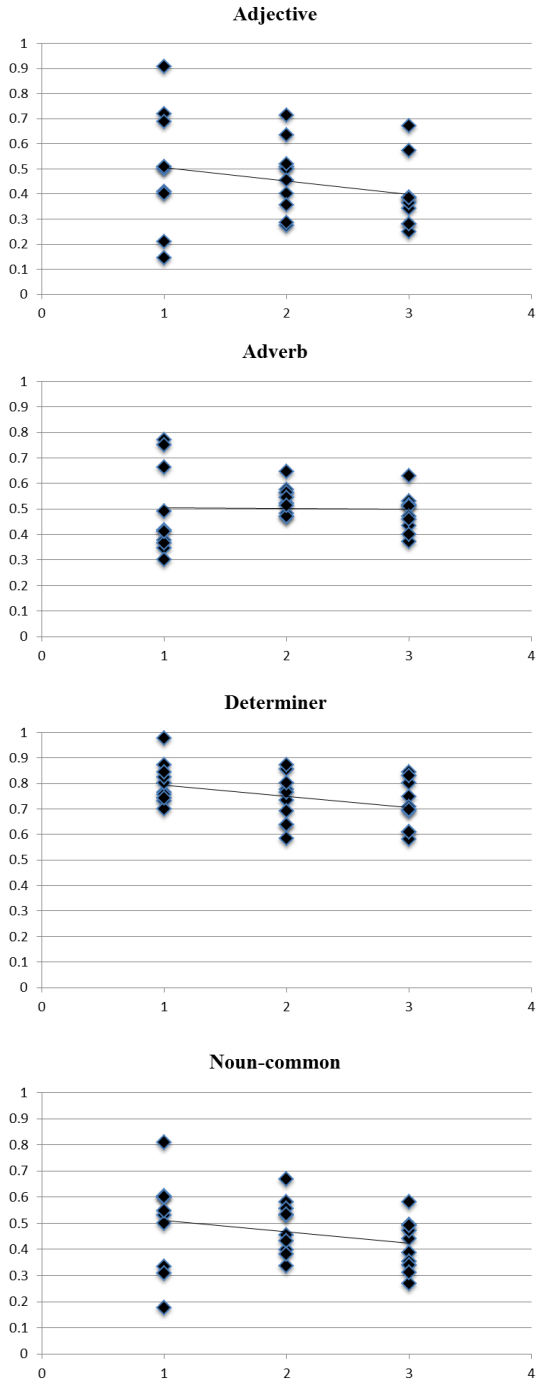

Figure 18: DELiC4MT scores by quality band with trend lines for ADJ, ADV, DET, NOC.

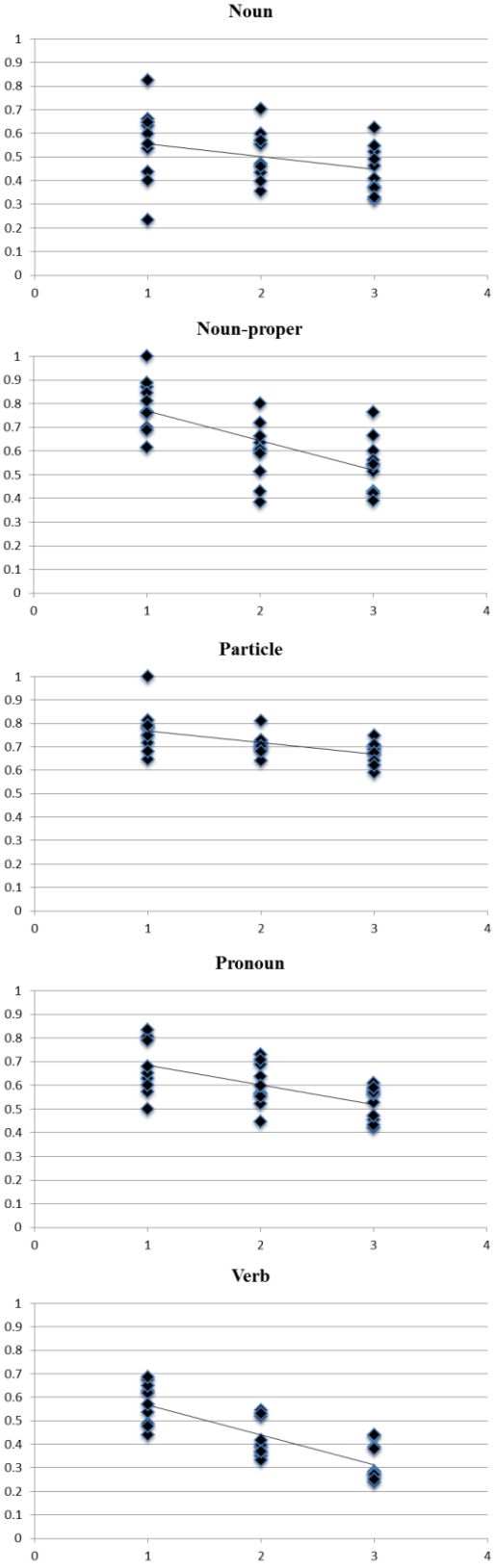

Figure 19: DELiC4MT scores by quality band with trend lines for NOU, NOP, PAR, PRO, VER.

3.3.2. DELiC4MT Sample Output

To clarify the mechanism of the analysis performed by DELiC4MT on the PoS-based linguistic checkpoints, here we show three sample outputs on the segment level (all of them for the language direction Spanish to English and for the VER checkpoint on output produced by the RbMT system). Given a checkpoint instance, these examples show the reference (source and target sides), the alignments (words between "<" and ">"), the MT output and the *n*-gram matches. As explained in more detail in Section 2.1, DELiC4MT detects the relevant PoS-based checkpoint on the source side, matches it with the aligned word in the MT output/hypothesis, and checks this against the corresponding word in the reference translation.

Source ref:	*Y aún así, <es> una estrella.*
Target ref:	*And yet, he <is> a star.*
MT output:	*And still like this, is a star.*
ngram matches:	*is (1/1)*

The first example shows a correct translation, scored successfully by DELiC4MT. The Spanish form "es" (3rd person of the present tense of the verb "ser", i.e. 'to be' in Spanish) is correctly translated to its equivalent in English, "is", matching the aligned reference translation.

Source ref:	*Fue un regalo que me <hizo> él*
Target ref:	*It was a gift he <gave> me*
MT output:	*It was a gift that did me he*
ngram matches:	*- (0/1)*

The second example shows a verb translated literally (and incorrectly): the source "hizo" (3rd person of the past tense of the verb "hacer", i.e. 'to make/to do' in Spanish) would normally correspond to "made/did" in English; however, in the expression "hacer un regalo" it corresponds to "give a present". The diagnostic evaluation tool correctly identifies this as a mistake, i.e. it detects a specific case contributing to translation quality barriers.

Source ref:	*Anto tiene asma, <respira> con dificultad*
Target ref:	*Anto has asthma, <he> <has> difficulty breathing*
MT output:	*Anto has asthma, it breathes with difficulty*
ngram matches:	*has (1/3)*

Finally, the third example shows a correct translation which DELiC4MT fails to assess positively: the verb "respira" (3rd person of the present tense of the verb "respirar", i.e. 'to breathe' in Spanish) is correctly translated as "breathing" in English; however, due to a wrong word alignment ("respira" is wrongly aligned to "he has", instead of to "breathing"), the score is not 1/1, but 1/3.

4. Conclusions and Future Work

This paper has explored the joint use of automatic diagnostic evaluation and human quality rankings to identify source-side linguistic phenomena that cause quality barriers in MT, looking at the two bidirectional language pairs EN↔ES and EN↔DE. We have evaluated output sentences produced by three types of MT systems (statistical, rule-based and hybrid) belonging to different

quality ranks (perfect, near-miss and poor translations), as classified by human annotators. The evaluation has been performed on a set of 9 PoS-based linguistic checkpoints with DELiC4MT, thus allowing us to draw conclusions on the quality barriers encountered by the different MT systems on a range of linguistic phenomena, for all three quality ranks across the four translation combinations.

On the basis of this evaluation, the paper has analysed the correlation between the scores obtained for each of these source-side linguistic phenomena and the human quality ratings, thus assessing the extent to which these phenomena can be used to predict human quality evaluation. Considering all the MT system types evaluated together, it turns out that the best predictors are VER (r=0.795), NOP (r=0.658) and PRO (r=0.604), while the worst one is by far ADV (r=0.02).

Regarding future work, taking into account the limitations of the current study (the small amount of data and somewhat limited translation combinations), we would like to confirm the findings reported here by performing experiments on larger data sets, including a more varied and extended set of language pairs for a wider collection of linguistic checkpoints. We are also planning to explore the use of diagnostic MT evaluation to analyse the errors identified by the Multidimensional Quality Metric (MQM) (Lommel and Uszkoreit, 2013). The MQM is a new paradigm for translation quality assessment, in which errors are categorised according to a hierarchy of issue types. By using DELiC4MT with a variety of suitable linguistic checkpoints to analyse translations annotated with the MQM, we intend to investigate which source-side linguistic phenomena cause the various MQM error types, as further indicators of translation quality barriers.

5. Acknowledgements

The work presented here has been conducted as part of the QTLaunchPad project, which has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 296347. Thanks are due to Aljoscha Burchardt, Maja Popović, Kim Harris and Lucia Specia for facilitating the annotation of the data used for this study and for interesting discussions that led to the work presented here, for which however the authors are solely responsible.

6. References

Balyan, R., Naskar, S.K., Toral, A. and Chatterjee, N. (2012). A Diagnostic Evaluation Approach Targeting MT Systems for Indian Languages. In *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012), COLING 2012*. Mumbai, India, December 2012, pp. 61--72.

Balyan, R., Naskar, S.K., Toral, A. and Chatterjee, N. (2013). A Diagnostic Evaluation Approach for English to Hindi MT Using Linguistic Checkpoints and Error Rates. In A. Gelbukh (Ed.), *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2013)*. Samos, Greece. 2013. LNCS 7817. Berlin: Springer, pp. 285--296.

Burchardt, A., Gaspari, F., Lommel, A., Popović, M., and Toral, A. (2014). *Barriers for High-Quality Machine Translation*. QTLaunchPad Deliverable 1.3.1. Available from www.qt21.eu/launchpad/system/files/deliverables/QTLP-Deliverable-1_3_1.pdf (accessed 10 February 2014).

Lommel, A. and Uszkoreit, H. (2013). Multidimensional Quality Metrics: A New Unified Paradigm for Human and Machine Translation Quality Assessment. Paper presented at *Localization World*, 12-14 June 2013, London, United Kingdom.

Naskar, S.K., Toral, A., Gaspari, F. and Way, A. (2011). A Framework for Diagnostic Evaluation of MT Based on Linguistic Checkpoints. In *Proceedings of Machine Translation Summit XIII*. Xiamen, China, 19-23 September 2011, pp. 529--536.

Naskar, S.K., Toral, A., Gaspari, F. and Groves, D. (2013). Meta-Evaluation of a Diagnostic Quality Metric for Machine Translation. In K. Sima'an, M.L. Forcada, D. Grasmick, H. Depraetere and A. Way (Eds.), *Proceedings of the XIV Machine Translation Summit*. Nice, France, 2-6 September 2013. Allschwil: The European Association for Machine Translation, pp. 135--142.

Och, F.J., and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, volume 29, number 1, pp. 19--51.

Padró, L. and Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference* . ELRA. Istanbul, Turkey. 21-27 May 2012, pp. 2473--2479.

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 on Association for Computational Linguistics*. Philadelphia, July 2002, pp. 311--318.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. P*roceedings of the ACL SIGDAT-Workshop.* Dublin, Ireland, pp. 47--50.

Toral, A., Naskar, S.K., Gaspari, F. and Groves, D. (2012). DELiC4MT: A Tool for Diagnostic MT Evaluation over User-defined Linguistic Phenomena. *The Prague Bulletin of Mathematical Linguistics*, 98(1), pp. 121--131.

Toral, A., Naskar, S.K., Vreeke, J., Gaspari, F. and Groves, D. (2013). A Web Application for the Diagnostic Evaluation of Machine Translation over Specific Linguistic Phenomena. In C. Dyer and D. Higgins (Eds.), *Proceedings of the 2013 NAACL HLTConference - Demonstration Session.* Atlanta, GA, USA. 10-12 June 2013. Stroudsburg, PA: Association for Computational Linguistics, pp. 20--23.