

**LREC 2016 Workshop**

**Translation Evaluation:  
From Fragmented Tools and Data Sets  
to an Integrated Ecosystem**

**PROCEEDINGS**

Edited by

Georg Rehm, Aljoscha Burchardt, Ondřej Bojar, Christian Dugast,  
Marcello Federico, Josef van Genabith, Barry Haddow, Jan Hajič,  
Kim Harris, Philipp Koehn, Matteo Negri, Martin Popel,  
Lucia Specia, Marco Turchi, Hans Uszkoreit

24 May 2016



Proceedings of the LREC 2016 Workshop  
“Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem”

24 May 2016 – Portorož, Slovenia

Edited by Georg Rehm, Aljoscha Burchardt, Ondřej Bojar, Christian Dugast, Marcello Federico, Josef van Genabith, Barry Haddow, Jan Hajič, Kim Harris, Philipp Koehn, Matteo Negri, Martin Popel, Lucia Specia, Marco Turchi, Hans Uszkoreit

<http://www.cracking-the-language-barrier.eu/mt-eval-workshop-2016/>

Acknowledgments: This work has received funding from the EU’s Horizon 2020 research and innovation programme through the contracts CRACKER (grant agreement no.: 645357) and QT21 (grant agreement no.: 645452).



# Organising Committee

- Ondřej Bojar, Charles University in Prague, Czech Republic
- Aljoscha Burchardt, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany\*
- Christian Dugast, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany
- Marcello Federico, Fondazione Bruno Kessler (FBK), Italy
- Josef van Genabith, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany
- Barry Haddow, University of Edinburgh, UK
- Jan Hajič, Charles University in Prague, Czech Republic
- Kim Harris, text&form, Germany
- Philipp Koehn, Johns Hopkins University, USA, and University of Edinburgh, UK
- Matteo Negri, Fondazione Bruno Kessler (FBK), Italy
- Martin Popel, Charles University in Prague, Czech Republic
- Georg Rehm, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany\*
- Lucia Specia, University of Sheffield, UK
- Marco Turchi, Fondazione Bruno Kessler (FBK), Italy
- Hans Uszkoreit, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany

\* main editors and chairs of the Organising Committee

## Programme Committee

- Nora Aranberri, University of the Basque Country, Spain
- Ondřej Bojar, Charles University in Prague, Czech Republic
- Aljoscha Burchardt, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany
- Christian Dugast, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany
- Marcello Federico, Fondazione Bruno Kessler (FBK), Italy
- Christian Federmann, Microsoft, USA
- Rosa Gaudio, Higher Functions, Portugal
- Josef van Genabith, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany
- Barry Haddow, University of Edinburgh, UK
- Jan Hajič, Charles University in Prague, Czech Republic
- Kim Harris, text&form, Germany
- Matthias Heyn, SDL, Belgium
- Philipp Koehn, Johns Hopkins University, USA, and University of Edinburgh, UK
- Christian Lieske, SAP, Germany
- Lena Marg, Welocalize, UK
- Katrin Marheinecke, text&form, Germany
- Matteo Negri, Fondazione Bruno Kessler (FBK), Italy
- Martin Popel, Charles University in Prague, Czech Republic
- Jörg Porsiel, Volkswagen AG, Germany
- Georg Rehm, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany
- Rubén Rodríguez de la Fuente, PayPal, Spain
- Lucia Specia, University of Sheffield, UK
- Marco Turchi, Fondazione Bruno Kessler (FBK), Italy
- Hans Uszkoreit, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany

## Preface

Current approaches to Machine Translation (MT) or professional translation evaluation, both automatic and manual, are characterised by a high degree of fragmentation, heterogeneity and a lack of interoperability between methods, tools and data sets. As a consequence, it is difficult to reproduce, interpret, and compare evaluation results. In an attempt to address this issue, the main objective of this workshop is to bring together researchers working on translation evaluation and practitioners (translators, users of MT, language service providers etc.).

This workshop takes an in-depth look at an area of ever-increasing importance. Two clear trends have emerged over the past several years. The first trend involves standardising evaluations in research through large shared tasks in which actual translations are compared to reference translations using automatic metrics or human ranking. The second trend focuses on achieving high quality (HQ) translations with the help of increasingly complex data sets that contain many levels of annotation based on sophisticated quality metrics – often organised in the context of smaller shared tasks. In industry, we also observe an increased interest in workflows for HQ outbound translation that combine Translation Memories, MT, and post-editing. In stark contrast to this trend to quality translation and its inherent overall approach and complexity, the data and tooling landscapes remain rather heterogeneous, uncoordinated and not interoperable.

The event brings together researchers, users and providers of tools, and users and providers of manual and automatic translation evaluation methodologies. We want to initiate a dialogue and discuss whether the current approach involving a diverse and heterogeneous and distributed set of data, tools, scripts, and evaluation methodologies is appropriate enough or if the community should, instead, collaborate towards building an integrated ecosystem that provides better and more sustainable access to data sets, evaluation workflows, tools, approaches, and metrics that support processes such as annotations, quality comparisons and post-editing.

The workshop is meant to stimulate a dialogue about the commonalities, similarities and differences of the existing solutions in the three areas (1) tools, (2) methodologies, (3) data sets. A key question concerns the high level of flexibility and lack of interoperability of heterogeneous approaches, while a homogeneous approach would provide less flexibility but higher interoperability and thus allow, e. g., integrated research by means of an MT app store (cf. the *Translingual Cloud* anticipated in the META-NET Strategic Research Agenda). How much flexibility and interoperability does the translation community need? How much does it want? How can communication and collaboration between industry and research be intensified?

We hope that the papers presented and discussed at the workshop provide at least partial answers on these, and other, crucial questions around the complex and interdisciplinary topic of evaluating translations, either produced by machines or by human experts.

G. Rehm, A. Burchardt, O. Bojar, C. Dugast, M. Federico, J. van Genabith, B. Haddow, J. Hajič,  
K. Harris, P. Koehn, M. Negri, M. Popel, L. Specia, M. Turchi, H. Uszkoreit May 2016

# Programme

09.00 – 09.10 Welcome – introduction – context

---

## **Session 1: Tools, Methods and Resources for Research**

09.10 – 09.30 Julia Ive, Aurélien Max, François Yvon, Philippe Ravaud:  
Diagnosing High-Quality Statistical Machine Translation Using  
Traces of Post-Editon Operations

09.30 – 09.50 Ondřej Bojar, Filip Děchtěrenko, Maria Zelenina:  
A Pilot Eye-Tracking Study of WMT-Style Ranking Evaluation

09.50 – 10.00 Anabela Barreiro, Francisco Raposo, Tiago Luís:  
CLUE-Aligner: An Alignment Tool to Annotate Pairs of  
Paraphrastic and Translation Units (short presentation)

10.00 – 10.10 Zijian Győző Yang, László János Laki, Borbála Siklósi:  
HuQ: An English-Hungarian Corpus for Quality Estimation  
(short presentation)

10.10 – 10.30 Discussion of the papers presented in Session 1

---

10.30 – 11.00 *Coffee break*

---

## **Session 2: Shared Tasks**

11.00 – 11.20 Ondřej Bojar, Christian Federmann, Barry Haddow,  
Philipp Koehn, Matt Post, Lucia Specia:  
Ten Years of WMT Evaluation Campaigns: Lessons Learnt

11.20 – 11.40 Luisa Bentivogli, Marcello Federico, Sebastian Stüker, Mauro Cettolo, Jan Niehues:  
The IWSLT Evaluation Campaign: Challenges, Achievements, Future Directions

11.40 – 12.00 Discussion of the papers presented in Session 2

---

---

**Session 3: Evaluation Tools and Metrics (part A)**

- 12.00 – 12.20 Katrin Marheinecke:  
Can Quality Metrics Become the Drivers for Machine Translation Uptake?  
An Industry Perspective
- 12.20 – 12.40 Kim Harris, Aljoscha Burchardt, Georg Rehm, Lucia Specia:  
Technology Landscape for Quality Evaluation:  
Combining the Needs of Research and Industry
- 12.40 – 13.00 Eleftherios Avramidis:  
Interoperability in MT Quality Estimation or wrapping useful stuff in various ways

---

13.00 – 14.00 *Lunch break*

---

**Session 3: Evaluation Tools and Metrics (part B)**

- 14.00 – 14.20 Arle Lommel:  
Blues for BLEU: Reconsidering the Validity of Reference-Based MT Evaluation
- 14.20 – 14.40 Roman Sudarikov, Martin Popel, Ondřej Bojar, Aljoscha Burchardt, Ondřej Klejch:  
Using MT-ComparEval
- 14.40 – 15.00 Michal Tyszkowski, Dorota Szaszko:  
CMT: Predictive Machine Translation Quality Evaluation Metric
- 15.00 – 15.20 Aljoscha Burchardt, Kim Harris, Georg Rehm, Hans Uszkoreit:  
Towards a Systematic and Human-Informed Paradigm  
for High-Quality Machine Translation
- 15.20 – 16.00 Discussion of the papers presented in Session 3 (part A, part B)

---

16.00 – 16.30 *Coffee break*

---

16.30 – 17.30 Summary – final discussion – next steps: towards an integrated ecosystem?

---

17.30 *End of workshop*

# Table of Contents

<i>Interoperability in MT Quality Estimation or wrapping useful stuff in various ways</i> Eleftherios Avramidis .....	1
<i>CLUE-Aligner: An Alignment Tool to Annotate Pairs of Paraphrastic and Translation Units</i> Anabela Barreiro, Francisco Raposo, Tiago Luís .....	7
<i>The IWSLT Evaluation Campaign: Challenges, Achievements, Future Directions*</i> Luisa Bentivogli, Marcello Federico, Sebastian Stüker, Mauro Cettolo, Jan Niehues .....	14
<i>A Pilot Eye-Tracking Study of WMT-Style Ranking Evaluation</i> Ondřej Bojar, Filip Děchtěrenko, Maria Zelenina .....	20
<i>Ten Years of WMT Evaluation Campaigns: Lessons Learnt*</i> Ondřej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, Lucia Specia .....	27
<i>Towards a Systematic and Human-Informed Paradigm for High-Quality Machine Translation</i> Aljoscha Burchardt, Kim Harris, Georg Rehm, Hans Uszkoreit .....	35
<i>HuQ: An English-Hungarian Corpus for Quality Estimation</i> Zijian Győző Yang, László János Laki, Borbála Siklósi .....	43
<i>Technology Landscape for Quality Evaluation: Combining the Needs of Research and Industry*</i> Kim Harris, Georg Rehm, Aljoscha Burchardt, Lucia Specia .....	50
<i>Diagnosing High-Quality Statistical Machine Translation Using Traces of Post-Editon Operations</i> Julia Ive, Aurélien Max, François Yvon, Philippe Ravaud .....	55
<i>Blues for BLEU: Reconsidering the Validity of Reference-Based MT Evaluation</i> Arle Lommel .....	63
<i>Can Quality Metrics Become the Drivers of Machine Translation Uptake? An Industry Perspective</i> Katrin Marheinecke .....	71

<i>Using MT-ComparEval</i>	
Roman Sudarikov, Martin Popel, Ondřej Bojar, Aljoscha Burchardt, Ondřej Klejch .....	76
<i>CMT: Predictive Machine Translation Quality Evaluation Metric</i>	
Michal Tyszkowski, Dorota Szaszko .....	83

\* paper was submitted and reviewed as abstract.

# Interoperability in MT Quality Estimation or wrapping useful stuff in various ways

Eleftherios Avramidis

German Research Center for Artificial Intelligence (DFKI)  
Language Technology Group  
Alt Moabit 91c, 10559 Berlin  
eleftherios.avramidis@dfki.de

## Abstract

The situation on the interoperability of Natural Language Processing software is outlined through a use-case on Quality Estimation of Machine Translation output. The focus is on the development efforts for the `QUALITATIVE` tool, so that it integrates a multitude of state-of-the-art external tools into one single Python program, through an interoperable framework. The presentation includes 9 approaches taken to connect 25 external components, developed in various programming languages. The conclusion is that the current landscape lacks important interoperability principles and that developers should be encouraged to equip their programs with some of the standard interaction interfaces.

**Keywords:** Machine Translation, Quality Estimation, interoperability

## 1. Introduction

Software development in Computational Linguistics and Natural Language Processing (NLP) has been for many years a means for scientific experimentation, primarily confined on an academic environment. This often had as a result that the software written and the data collected for this purpose lack most software engineering principles, as they mainly address the need for performing research experiments that can get results in a timely manner. During the last decade, Language Technology (LT) has jumped from the research labs to the industry and many open-source tools and data that had originally been written for research purposes have ended being used in a wide scale, albeit leading to a very diverse and multilateral landscape.

For both industry and academic research, re-using software and data seems the obvious solution. It means saving effort and time in development to focus on the innovation, but also easily reproducing (and therefore confirming) state-of-the-art methods. Luckily, most of the tools and data are available with open or reusable licenses, but combining many of them into one LT application remains a challenge: every tool may be written in a different programming language, utilizing a different file format or providing a different (or no) external interface.

Our current contribution outlines the situation through a use-case on Quality Estimation of Machine Translation (MT) output. We look on the `QUALITATIVE` tool (Avramidis et al., 2014), that combines a multitude of state-of-the-art tools into one single application, which can be used both for research and real-time use. We describe the approaches taken to achieve an interoperable framework that bridges the communication between the several components in order to achieve the desired processing of data in a functional way. Although we do not provide a unified wide-scale solution, by sharing the experience of our own development from the perspective of MT Evaluation, we aim to highlight a part of the ecosystem and raise the awareness of how difficult and challenging it is to get everything together without re-writing from scratch.

## 2. Previous work

The idea of complex interoperable pipelines is not new in NLP or MT specifically. Most of such tasks consist of many tools in order to process the data and extract all sorts of linguistic knowledge and analyses. Among the most popular tools are the pipelines for training and evaluating Statistical Machine Translation systems. Consequently, we are reviewing some of the most prominent relevant frameworks.

**EMS** (Koehn, 2010) is the pipeline for training models for Statistical MT through Moses. Originally written as a single very long Perl script, it has been extended and adapted through the years to include more than 30 components. EMS wraps each external tool in a `bash` script launched through the shell, whereas data transfer between the components is done through temporary files on the disk or shell pipes. The majority of the launched programs are written in Perl and C++, whereas there are also some in Python and Java. The advantages of EMS are that it is very modular and it can wrap any program that operates on the Linux commandline. Additionally, it can run as an injection script for the Sun Grid in order to distribute and parallelize tasks across many computational servers. EMS could in principle be adapted to function for other type of experiments apart for MT, though not many examples have been reported.

**TREX** (Popel and Žabokrtský, 2010) is a similar pipeline offering sentence-level processing, a Perl API and a socket server, whereas its main use is focused on Statistical MT with deep transfer methods. **LOONYBIN** (Clark and Lavie, 2010) follows a similar approach. Coded in Jython, it allows the user to use modules in Python for wrapping tools in hyper-workflows. Similar approaches are followed by tools such as **EMAN** (Bojar and Tamchyna, 2013). One disadvantage of such tools is that `bash` wrappers, temporary files and pipes can mostly operate efficiently for batches of data and not for single sentences, as the case is for many user-oriented applications. Additionally, such a pipeline itself cannot be very efficiently incorporated in another pipeline, unless it is written in the same tool. There are other experiment pipeline tools such as DagMan, Dryad,

DuctTape, PCL, Pegasus, SoapLab, Taverna and UIMA, but we won't focus on them, since they have not been widely used in MT research.

An interesting contribution towards interoperability was the **Open Machine Translation Core** (Johnson, 2013) which attempted to define an abstract interface that standardizes functionality common to all MT systems. Despite the fact that a Java prototype was presented, to our knowledge no progress has been shown in the direction of this standardization by now.

**QUEST** (Specia et al., 2013) was the first software that appeared to wrap many tools for Quality Estimation. Written in Java, it incorporates directly a few other Java applications as libraries, whereas other tools are also wrapped by using the `bash` shell and intermediate temporary batch files. The machine learning part is held by a separate Python script. For this reasons, running a unified pipeline with realtime user requests (e.g. server mode) is non-trivial.

### 3. Basic architecture

In our approach, the main core of the program is written in Python. Python has been chosen because it offers the flexibility of dynamic programming, which allows for quick and relatively easy experimentation in many NLP tasks. Functionality from several powerful scientific and machine learning toolkits is available through imported libraries. Additionally, a Python script can be connected to real user applications, either through a web server (e.g. Django), or by offering its functionality via a socket service. This choice offers a flexible framework for both experimentation and practice, although it has got its own limitations (e.g. processing cannot be distributed in many computational machines without additional engineering).

As explained in Avramidis et al. (2014), the program is internally organized in several modules:

- **data reading** receives a file and loads the data in the memory via the respective data structures
- **preprocessing** sends a sentence for the required preprocessing task (e.g. tokenization, compound splitting, truecasing etc.)
- **machine translation** sends source sentences to MT engines and receives their translation
- **feature generation** sends the source sentences and their translations to feature generator classes and tools and receives the respective vectors of numerical features
- **machine learning** serves for the communication with machine learning toolkits for two functions: training and testing. During training, it sends a batch of vectors, each one with a golden label and it receives a model. During testing, the model is loaded and given a vector, the predicted label is returned.

For each module, the commands are organized so that they form a specific interface as a principle of internal *modularity*. This way, the same functionality can be implemented by different classes. For example, every feature generator

class has to implement at least one function that receives a source sentence and its translations and returns a vector of numerical features.

## 4. Connecting external components

We present two main categories of communicating with external software components, based on whether the execution of the external software is controlled by the our Python script, which we will call the “host”, or whether it is run as a remote service.

### 4.1. Inherent integration

In these functions, the execution of the external software is encapsulated into the host. The goal is to keep the external tool running in the background so that it can receive requests from the host. It gets automatically unloaded when the host program is finished. The part of the host program or the code which handles the specificities of the communication is referred to as at “connector”.

#### 4.1.1. Native Python libraries

Many pieces of Python open-source software already offer their functionality in openly available libraries. This is the easiest and most efficient type of integration, as all of the public functions of the included software can be directly called from within our host Python code. The software served by this method includes:

- BLEU (Papineni et al., 2001), Levenshtein Distance (Levenshtein, 1966) and RgbF (Popović, 2012) for MT evaluation scores
- **HIERSON** (Popović, 2011) for automatic detection of MT errors
- **KENLM** (Heafield, 2011) for language modelling
- **MLPYTHON**<sup>1</sup>, **ORANGE** (Demšar et al., 2004) and **SCIKIT-LEARN** (Pedregosa et al., 2011) for machine learning functions.
- **NLTK** (Loper and Bird, 2002) for several simple NLP tasks
- **NUMPY** (Van Der Walt et al., 2011) for memory-efficient handling of numerical arrays and **SCIPY** (Oliphant, 2007) for scientific (e.g. complex mathematical or statistical) functions.

#### 4.1.2. Java programs

Py4j<sup>2</sup> was chosen as a solution to integrate functionality from open-source Java programs into Python. The Java Virtual Machine (JVM) starts in the background including the required Java Packages (jar) in the classpath. Then, a Py4j gateway connects with the JVM via a socket and makes all public classes and functions loadable and callable from within Python. Python types are automatically converted to Java types and vice versa. If the processes are thread-safe

<sup>1</sup><http://www.dmi.usherb.ca/~larocheh/mlPython/>

<sup>2</sup><http://www.py4j.org>

on the Java side, they can be also parallelized in several Python threads.

This method is used to connect with:

- BERKELEY PARSER (Petrov et al., 2006) for parsing with Probabilistic Context Free Grammars (PCFG),
- LANGUAGE TOOL (Naber, 2003; Miłkowski, 2012) for rule-based language checking and
- METEOR (Lavie and Agarwal, 2007; Denkowski and Lavie, 2014) for MT evaluation scoring.

This method is efficient and allows wide access and parametrization to the functionality of the external Java program. Nevertheless, it also requires good knowledge to its internal structure, e.g. via a Java API documentation or by reading the Java source code. This is needed because the imported objects, functions and variables have to be treated in Python the same way they would do in Java. Additionally, the host needs to know or maintain a knowledge of the system socket where the JVM operates, which makes it complicated to run many hosts on the same JVM. In a few cases, parts of the source code had to be modified and be re-build, since not all required functions were declared as public, which is a major requirement.

#### 4.1.3. SWIG

Simplified Wrapper and Interface Generator (SWIG) allows wrapping C++ code as a Python library. Creating such a connector allows to parse C/C++ interfaces and generate the 'glue code' for Python to call into the C/C++ code. In our program we have not developed such a connector, but we have experimented with SWIG-SRILM (Madnani, 2009), an existing wrapper around SRILM (Stolcke, 2002).

#### 4.1.4. Pipes

An external commandline-based software is launched by the host as a sub-process in the background. The standard input, the standard output and the error output can be captured within a Python object (a *pipe*). Therefore, a program-specific connector needs to be written. It should be aware of the commandline behaviour of the software and simulate that through the pipe. The sub-process is treated as a black-box, i.e. no access to particular internal functions is possible.

For example, a standard tokenizer from the MOSES scripts would read from the standard input all characters, waiting for and "end of line". Once the "end of line" is received, the tokenization takes place and the tokenized string is returned through the standard output.

This approach is mainly used for Perl scripts and C++ programs but can be adapted for any commandline application. Such software includes:

- MOSES scripts for pre-processing and post-processing, such as punctuation normalizer, tokenizer, compound splitter (Koehn and Knight, 2003), true-caser (Och et al., 2003), de-truecaser, de-tokenizer etc. Although re-implementations for most of these exist in Python and therefore could be directly included in our code, one may still require to stick to the

original MOSES Perl scripts, if they want to re-use pre-trained MOSES translation models or acquire results comparable with other scientific works that use these state-of-the-art Perl scripts.

- TREETAGGER for POS tagging (Schmid, 1994) integrated via the TreeTaggerWrapper (Pointal, 2015).

The advantage of this method is that it can be adapted for many programs without requiring knowledge of their internal coding or functioning, while it still allows loading a tool into memory once and sending individual requests when the host program needs it. The disadvantage is that the only way of interaction is through the standard input and output, which offer no flexibility for parametrization or passing more complex types. Additionally, reading standard output often requires excessive use of regular expressions to understand some complex output, which would otherwise be intended for the visual understanding of the user. Unexpected errors and exceptions are hard to capture, too. We should also mention that some tools only work with input and output files (batch mode) and do not support per-request communication with standard input and output. Finally, serious deficiencies have been noted concerning the buffering support of the pipes, which may cause prevent data to be transferred through the standard input/output.

#### 4.1.5. Shell with external files

The data to process is written by the host on a temporary file. The external program is launched once, asked to process the given temporary file as an input and write its output in another temporary file, which consequently gets read by the host. This is the last resort for having the host communicate with external tools, since loading the entire program per request and writing external files is not efficient for single sentences and is useful only for processing batches of requests. We also noticed that some programs of this kind do not allow many instances to be run in parallel (e.g. because they require an exclusive lock on some internal files, whose location is often non-parametrizable).

We used this method for aligning sentences with GIZA++ (Och and Ney, 2003), acquiring baseline features from QUEST and doing PCFG parsing with BITPAR (Schmid, 2004) with the help of a wrapper (van Cranenburgh, 2010). This method was useful only for experiments that did not require parallelization and single requests.

## 4.2. Integrating functionality as a remote service

An additional possibility of integrating an external tool is by sending requests to it as a remote service. In this case, the external tool must provide a server which initially loads the program and implements a network protocol of requests and responses. It waits until a request is received from the host, in order to run the required functions. The result of the functions is then sent with a corresponding response. Four such protocols and the respective tools we have used are:

- JSON: with MT-MONKEY (Tamchyna et al., 2013), which acts as a hub and a load balancer for fetching translations from several MT engines

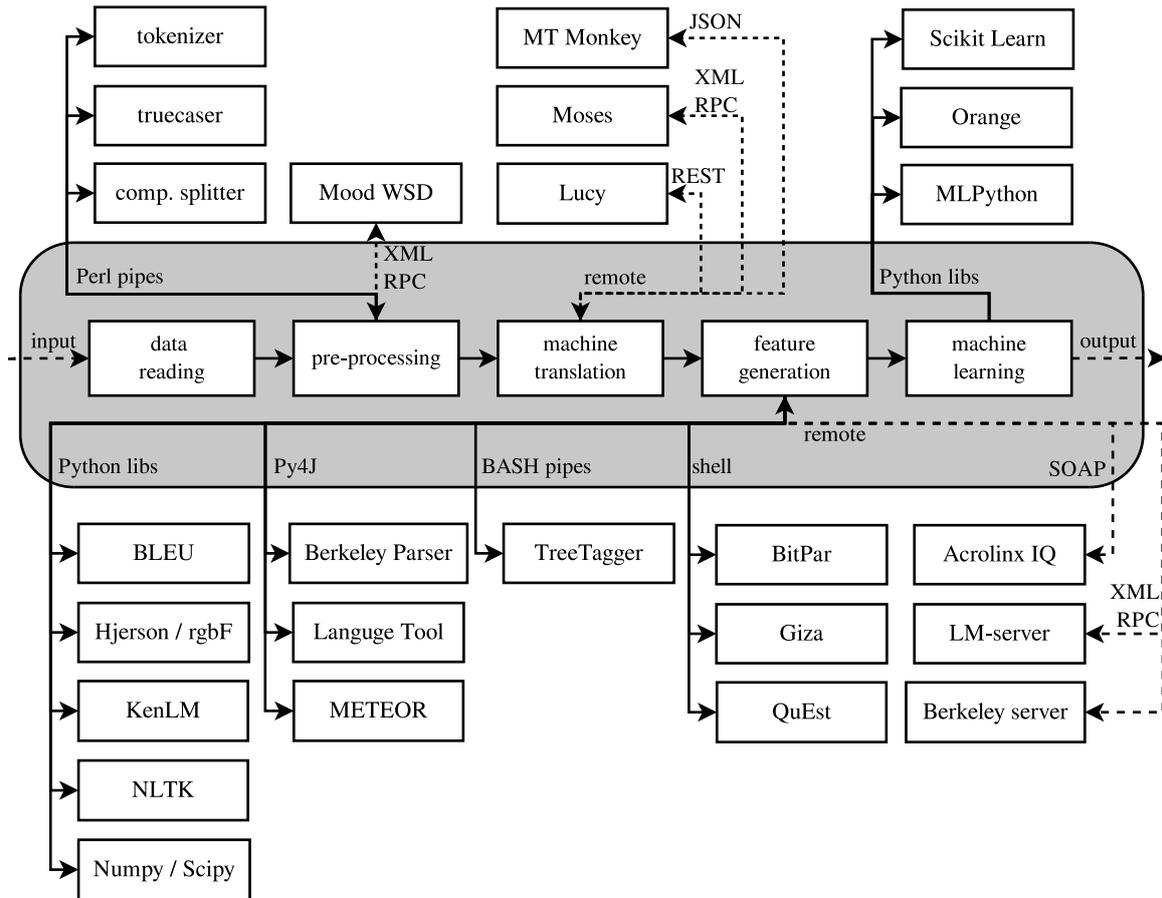


Figure 1: Full diagram of the components that have been integrated into the application

- **SOAP:** with ACROLINX IQ (Siegel, 2011) for language checking
- **REST:** with the LUCY rule-based MT system (Alonso and Thurmair, 2003).
- **XML-RPC:** with MOSES (Koehn et al., 2006) for Statistical MT, with LM-SERVER (Madnani, 2009) for language model scoring, with our own XML-RPC wrapper of BERKELEY PARSER and with MOOD, (Weissenborn et al., 2015) a Word Sense Disambiguation analyzer.

Such an integration is straightforward if the tool already provides such a protocol interface, since the protocols allow for easy mapping of function and variable types across many different programming languages. This solution is based on a network connection, so it is also desirable when one needs to distribute different computationally or memory intensive modules to many computational servers. Nevertheless, such a network communication may be considerably slower due to the network overhead. Additionally, starting and stopping remote services cannot be easily controlled by the host, unlike to the encapsulation described in the previous section.

## 5. Discussion

The ecosystem is indeed complicated. Integrating existing software saves time from re-implementing it and can confirm replicability of scientific experiments. Nevertheless, as we outlined in our use-case, the different types of software may require different kind of integration. Such an integration often requires low-level or even backwards engineering, which means a lot of non-creative effort.

An obvious conclusion through our experience is that reusability and efficient interoperability mostly depends on the will of the original developer. Adding support for a network service or exporting a Python library is straightforward for the original developers of a software, in contrast to the huge effort required for a third-party developer to understand the functionality and wrap it one way or another. It suffices to mention that out of the 25 external tools and libraries that we integrated, only 5 provided original support (remote service or library) for being integrated with a programming language other than the one they were originally developed in.

In that direction, the specification of a unified way to communicate across different code and platforms would be precious. Whatsoever, even encouraging developers to con-

sider serious solutions for the interoperability of their software would be a major first step. Among the most obvious solutions, we would consider wrapper libraries in the most popular scripting languages (e.g. Python, Perl) and exposing full functionality through a ReSTful service (Richardson and Ruby, 2008), possibly along with “autodiscovery” API functions via the WSDL (Christensen et al., 2001).

### Acknowledgment

This work has received support by the EC’s FP7 (FP7/2007-2013) under grant agreement number 610516: “QTLeap: Quality Translation by Deep Language Engineering Approaches”. Early stages have been developed with the support of the projects TaraXÚ and QT-Launchpad.

### References

- Alonso, J. A. and Thurmair, G. (2003). The Compendium Translator system. In *Proceedings of the Ninth Machine Translation Summit*. International Association for Machine Translation (IAMT).
- Avramidis, E., Poustka, L., and Schmeier, S. (2014). Qualitative: Open source Python tool for Quality Estimation over multiple Machine Translation outputs. *The Prague Bulletin of Mathematical Linguistics*, 102:5–16.
- Bojar, O. and Tamchyna, A. (2013). The Design of Eman, an Experiment Manager. *Prague Bull. Math. Linguistics*, 99:39–58.
- Christensen, E., Curbera, F., Meredith, G., and Sanjiva Weerawarana. (2001). Web Services Description Language (WSDL) 1.1. Technical report, World Wide Web Consortium.
- Clark, J. H. and Lavie, A. (2010). LoonyBin: Keeping Language Technologists Sane through Automated Management of Experimental (Hyper)Workflows. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 1301–1308.
- Demšar, J., Zupan, B., Leban, G., and Curk, T. (2004). Orange: From Experimental Machine Learning to Interactive Data Mining. In *Principles of Data Mining and Knowledge Discovery*, pages 537–539.
- Denkowski, M. and Lavie, A. (2014). Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA, jun. Association for Computational Linguistics.
- Heafield, K. (2011). KenLM : Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, number 2009, pages 187–197, Edinburgh, Scotland, jul. Association for Computational Linguistics.
- Johnson, I. (2013). Open Machine Translation Core: An Open API for Machine Translation Systems. *The Prague Bulletin of Mathematical Linguistics*, 100:91–100.
- Koehn, P. and Knight, K. (2003). Empirical Methods for Compound Splitting. In *Proceedings of the 10th Annual Conference of the European Association for Machine Translation*, volume 1, page 8, Budapest, Hungary.
- Koehn, P., Shen, W., Federico, M., Bertoldi, N., Callison-Burch, C., Cowan, B., Dyer, C., Hoang, H., Bojar, O., Zens, R., Constantin, A., Herbst, E., and Moran, C. (2006). Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 177–180, Prague, Czech Republic, jun.
- Koehn, P. (2010). An Experimental Management System Philipp Koehn. *The Prague Bulletin of Mathematical Linguistics*, (94):87–96.
- Lavie, A. and Agarwal, A. (2007). METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic, jun. Association for Computational Linguistics.
- Levenshtein, V. (1966). Binary Codes Capable of Correcting Deletions and Insertions and Reversals. *Soviet Physics Doklady*, 10(8):707–710.
- Loper, E. and Bird, S. (2002). NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP ’02, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Madnani, N. (2009). Source Code: Querying and Serving N-gram Language Models with Python. *The Python Papers Source Codes*.
- Milkowski, M., (2012). *Translation Quality Checking in LanguageTool*, pages 213–223. Corpus Data across Languages and Disciplines. Peter Lang, Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Wien.
- Naber, D. (2003). A rule-based style and grammar checker. Technical report, Bielefeld University, Bielefeld, Germany.
- Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. In *Computational Linguistics*.
- Och, F., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., and Others. (2003). Syntax for statistical machine translation. In *Johns Hopkins University 2003 Summer Workshop on Language Engineering, Center for Language and Speech Processing, Baltimore, MD, Tech. Rep. Cite-seer*.
- Oliphant, T. E. (2007). Python for Scientific Computing. *Computing in Science & Engineering*, 9(3):10–20.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). Bleu: a Method for Automatic Evaluation of Machine Translation. IBM Research Report RC22176(W0109-022), IBM.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, 12:2825–2830.
- Petrov, S., Barrett, L., Thibaux, R., and Klein, D. (2006). Learning Accurate, Compact, and Interpretable Tree An-

- notation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440, Sydney, Australia, jul. Association for Computational Linguistics.
- Pointal, L. (2015). TreeTagger Wrapper.
- Popel, M. and Žabokrtský, Z. (2010). TectoMT: Modular NLP Framework. In *Proceedings of the 7th International Conference on Advances in Natural Language Processing, IceTAL'10*, pages 293–304, Berlin, Heidelberg. Springer-Verlag.
- Popović, M. (2011). Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, 96(-1):59–68.
- Popović, M. (2012). rgbF: An Open Source Tool for n-gram Based Automatic Evaluation of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, (98):99–108.
- Richardson, L. and Ruby, S. (2008). *RESTful web services*. ” O’Reilly Media, Inc.”.
- Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.
- Schmid, H. (2004). Efficient Parsing of Highly Ambiguous Context-free Grammars with Bit Vectors. In *Proceedings of the 20th International Conference on Computational Linguistics, COLING ’04*, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Siegel, M. (2011). Autorenunterstützung für die Maschinelle Übersetzung. In *Multilingual Resources and Multilingual Applications: Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*, Hamburg.
- Specia, L., Shah, K., de Souza, J. G. C., and Cohn, T. (2013). QuEst - A translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84, Sofia, Bulgaria, aug. Association for Computational Linguistics.
- Stolcke, A. (2002). SRILM — an Extensible Language Modeling Toolkit. In *System*, volume 2, pages 901–904. ISCA, sep.
- Tamchyna, A., Dušek, O., Rosa, R., and Pecina, P. (2013). MTMonkey: A Scalable Infrastructure for a Machine Translation Web Service. *The Prague Bulletin of Mathematical Linguistics*, 100:31–40, oct.
- van Cranenburgh, A. (2010). Enriching Data-Oriented Parsing by blending morphology and syntax. Technical report, University of Amsterdam, Amsterdam.
- Van Der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science and Engineering*, 13(2):22–30.
- Weissenborn, D., Hennig, L., Xu, F., and Uszkoreit, H. (2015). Multi-Objective Optimization for the Joint Disambiguation of Nouns and Named Entities. In *Proceedings of the 53rd Annual Meeting of the Association*
- for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 596–605, Beijing, China. Association for Computer Linguistics.

# CLUE-Aligner: An Alignment Tool to Annotate Pairs of Paraphrastic and Translation Units

Anabela Barreiro<sup>1</sup>, Francisco Raposo<sup>2</sup>, Tiago Luís<sup>3</sup>

<sup>1</sup>INESC-ID, <sup>2</sup>IST, <sup>3</sup>VoiceInteraction

<sup>1</sup>abarreiro@inesc-id.pt,

<sup>2</sup>francisco.afonso.raposo@ist.utl.pt,

<sup>3</sup>tiago.luis@voiceinteraction.pt

## Abstract

Currently available alignment tools and procedures for marking-up alignments overlook non-contiguous multiword units for being too complex within the bounds of the proposed alignment methodologies. This paper presents the CLUE-Aligner (Cross-Language Unit Elicitation Aligner), a web alignment tool designed for manual annotation of pairs of paraphrastic and translation units, representing both contiguous and non-contiguous multiwords and phrasal expressions found in monolingual or bilingual parallel sentences. Non-contiguous block alignments are necessary to express alignments between multiwords or phrases, which contain insertions, i.e., words that are not part of the multiword unit or phrase. CLUE-Aligner also allows the alignment of smaller individual or multiword units inside non-contiguous multiword units. The interactive web application was developed under the scope of the eSPERTo project, which aims to build a linguistically enhanced paraphrasing system. However, a tool for manual annotation of alignment and for visualization of automatic phrase alignment can prove useful in human and machine translation evaluation.

**Keywords:** Alignment tool, phrase alignment, paraphrastic and translation units, non-contiguous multiword units

## 1. Introduction

Alignment is an NLP task consisting of the identification of translation or paraphrastic relationships among words, multiwords or phrasal expressions in bilingual or monolingual parallel sentences, i.e., sentence pairs that have been identified as translations or paraphrases of each other. Alignments are important data for statistical machine translation (SMT), where most models are built on the grounds of word and phrase-level alignment pairs acquired mostly automatically via statistical techniques. Unsupervised learning approaches used in SMT systems are trained to produce probabilistic alignments within each sentence.

Even though MT systems can learn alignments from data using unsupervised algorithms, recent work has focused on supervised methods that use manually aligned sentences (Ambati et al., 2010). Supervised learning normally takes context, syntax and other grammatical and semantic information into consideration. Thus, manually annotated alignments represent an important asset in the development and evaluation of SMT systems (Gao and Vogel, 2010), but they are also useful for other applications, such as translation and paraphrase lexicon induction, crosslingual contrastive studies, terminology extraction, and word sense disambiguation (Vickrey et al., 2005), among others.

The current state of the art in supervised learning and manual alignment is characterised by three main problems: (i) lack of multilingual datasets, i.e., the number of publicly available manual alignments is restricted to a few pairs of languages (cf. section 2.), (ii) lack of linguistically motivated alignment guidelines that take into account non-contiguous multiword units and other non-adjacent linguistic phenomena or syntactic discontinuity, such as extraposition or topicalization, among others, and (iii) lack of tools that enable the correct alignment of non-contiguous multiwords and phrasal expressions. These shortcomings in manual alignment standard practices drove us to the development of a new web alignment tool – CLUE-Aligner

– Cross-Language Unit Elicitation Aligner<sup>1</sup> – that allows representing semantically equivalent non-adjacent structures, such as non-contiguous multiword units in both translation and paraphrasing. This paper presents the current CLUE-Aligner interface, which does not implement automatic aligning yet, even though it can be used for correcting automatic alignments, and describes an experimental research on the use of CLUE-Aligner interactive tool that we have used in the alignment of a reference multilingual sub-corpus of the common test version of the Europarl Proceedings (Koehn et al., 2003; Koehn, 2005), comprising all possible combinations between English, French, Portuguese, and Spanish parallel texts. The alignment task resulted in a set of guidelines for the alignment of pairs of translation units – the CLUE4Translation Alignment Guidelines<sup>2</sup> – and a gold collection with translation pairs – the Gold CLUE4Translation. These resources have been developed as an experimental research within the eSPERTo project<sup>3</sup>, whose main objective was the development of a context-sensitive and linguistically enhanced paraphrase system that can be used in a large variety of applications, including adaptation between variants of the same language or integration of paraphrases in the translation workflow. CLUE-Aligner is also a by-product of the eSPERTo project.

## 2. Related Work

The concept of an alignment as the representation of a translation relation between words was introduced and ap-

<sup>1</sup>CLUE-Aligner’s characteristics were briefly introduced in Barreiro (2016). The often referred to as the “clue alignment approach” (Tiedemann, 2003; Tiedemann, 2011) is based on mainly word-level alignment clues. Our approach is based on manual alignments of cross-language multiwords and phrasal expressions.

<sup>2</sup>[www.l2f.inesc-id.pt/~abarreiro/clue/translation-alignment-guidelines.pdf](http://www.l2f.inesc-id.pt/~abarreiro/clue/translation-alignment-guidelines.pdf)

<sup>3</sup>eSPERTo – System for Paraphrasing in Editing and Revision of Text – available at: <https://esperto.l2f.inesc-id.pt/esperto/esperto/demo.pl>

plied to statistical machine translation in the early nineties (Brown et al., 1990), and it was used as a primary resource for phrase-based (Och and Ney, 2000) and syntax-based machine translation (Galley et al., 2004). Most of the literature focuses on automatically building alignments using unsupervised algorithms, where the machine itself decides which segments of a sentence in a source language to align with which segments of a sentence in a target language and learns alignments from data (Och and Ney, 2000). Research on alignment quality improvement through supervised training is also available, where supervised methods use high quality alignments, which are often hand-made by linguists (Blunsom and Cohn, 2006). Several guidelines for manual alignment can be found in the literature for English–French, in the context of the Blinker project (Melamed, 1998), Czech–English (Kruijff-Korbayová et al., 2006; Bojar and Prokopová, 2006), and Spanish–English texts (Lambert et al., 2005), among others. Some basic annotation guidelines for the alignment of paraphrases have also been proposed.<sup>4</sup>

Despite the aforementioned efforts, manual alignments scarcity remains a problem, and the available alignments are mostly bilingual, with the exception of 6 multilingual sets annotated by Graça et al. (2008). In addition, previously proposed alignment guidelines cover cross-linguistic phenomena superficially, excluding the important alignment challenges (and challenges to machine translation) presented by non-contiguous support verb constructions and other multiwords and phrasal expressions covered in the CLUE4Translation Alignment Guidelines and reproduced graphically in the CLUE-Aligner.

There are several tools for manual annotation of word alignment that support non-contiguous multiwords, e.g. Alpaco (Rassier and Pedersen, 2003), based on the Blinker project, Yawat (Germann, 2008), SWIFT (Gilmanov et al., 2014), among others. These tools use distinct visualization schemes. The most common alignment visualization types are links (or lines), matrix, and dynamic markup. Source or target-ordered bitext and coloring schemes have also been used. A thorough description of the different visualization types can be found in Germann (2008). The CLUE-Aligner was inspired in a matrix type visualization alignment tool, Linear-B (Callison-Burch and Bannard, 2004), (Callison-Burch, 2007). CLUE-Aligner’s design aims to bring more flexibility to the alignment task, allowing the block-alignment of contiguous and non-contiguous multiwords and phrasal expressions found in monolingual or bilingual pairs of parallel sentences, depending on whether these sentences are paraphrases or translations of each other. The use of a matrix visualization scheme was complemented with a coloring scheme that helps distinguish between sure and possible contiguous and non-contiguous units, and internal units (or internal blocks), as described in section 3.. The matrix has the advantage of facilitating the visualization of the segments that have already been annotated making it easier to focus on the annotation of the remaining ones with a clear view of the intersection between

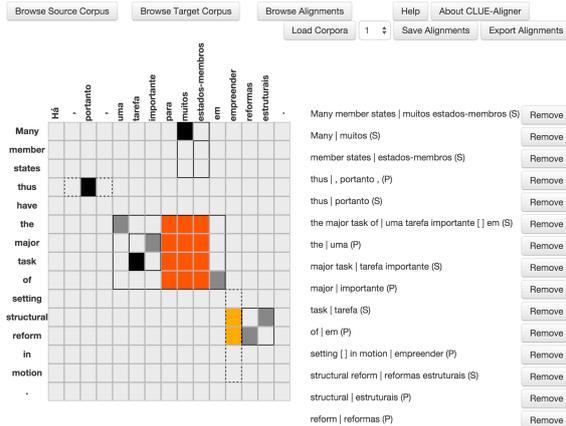


Figure 1: CLUE-Aligner interface

source and target segments. However, the strongest feature of CLUE-Aligner is the alignment and storage of pairs of paraphrastic units with indication of the place of insertions, which are represented by "[ ]" (e.g., *I urge [ ] to | Exorto [ ] a*, as in Figure 2). This feature is valuable for the use of these paraphrastic pairs in the construction of translation grammars and syntactic parsers where precision is important, in machine learning to help learning constituents, among other purposes and applications.

### 3. CLUE-Aligner

In CLUE-Aligner, each pair of sentences is represented by a grid, and all translational equivalences are graphically represented in that grid by the intersection of single segments corresponding to individual words or blocks corresponding to multiword units and phrasal expressions. Each word of the parallel sentences is displayed individually in the horizontal or vertical direction. The intersection between the words of the source and the target language forms a cell, which represents a segment.<sup>5</sup> CLUE-Aligner allows the annotation of non-contiguous multiword units. Non-contiguous block alignments are necessary to express alignments between multiwords or phrases, which contain words that are not part of the multiword or phrase. It is also possible to annotate smaller individual or multiword units inside non-contiguous multiword units. Figure 1 illustrates the current version of the CLUE-Aligner interface, displaying an English–Portuguese pair of sentences with single word, contiguous multiword unit, and non-contiguous multiword unit alignments.

#### 3.1. Types of Alignment

Based on a cross-linguistic analysis of particular cases of multiwords and other phrasal units, in our research work we annotated pairs of translation units as sure or possible alignments (Och and Ney, 2000; Graça et al., 2008) (cf. 3.1.1.), some of them as single word alignments (cf. 3.1.2.), others as block alignments (cf. 3.1.3.).

<sup>4</sup>[http://www.cs.jhu.edu/~ccb/publications/paraphrase\\_guidelines.pdf](http://www.cs.jhu.edu/~ccb/publications/paraphrase_guidelines.pdf)

<sup>5</sup>In our research work, a source and a target language can correspond to the same language, as in the case of paraphrases.

### 3.1.1. Sure and Possible Alignments

Sure alignments (S-alignments) establish relationships between semantically equivalent expressions. They correspond to expressions/translations that satisfy the criteria for optimum equivalence.<sup>6</sup> S-alignments have the property of reciprocity with regards to source-target direction. For example, the English domain term *venture capital markets* S-aligns<sup>7</sup> with the equivalent term *mercados de capital de risco* in Portuguese.

In contrast with sure alignments, possible alignments (P-alignments) correspond to expressions/translations that meet some, but not all of the requirements for absolute equivalence. P-alignments do not have the property of reciprocity with regards to source-target direction. For example, the English verb *began* P-aligns<sup>8</sup> with the idiomatic verbal expression *a vu le jour*, literally, *has seen the day*, in French.

### 3.1.2. Single Word Alignments

In the alignment task, when the annotator clicks on a cell in the grid once, the cell becomes black, which means it is an individual sure or S-alignment (non-ambiguous and optimal). When the annotator clicks on a cell in the grid again, the cell becomes grey, which means it is an individual possible or P-alignment (ambiguous or non-optimal). Both black and grey cells in the grid represent individual word alignments. Black cells represent full semantic correspondence. Grey cells represent approximate semantic correspondence. Individual word S and P-alignments are one-to-one alignments.<sup>9</sup> In Figure 1, the single word alignments *reform* – *reformas* and *structural* – *estruturais* are gray and not black because of the singular-plural disagreement.

### 3.1.3. Block Alignments

In addition to one-to-one alignments, there are one-to-many, or many-to-many word alignments, which correspond to the alignment of multiword units, phrasal expressions, or larger segments of a sentence, such as full paragraphs.<sup>10</sup> One-to-many and many-to-many word alignments are visually represented by block alignments. In

<sup>6</sup>Optimum equivalence refers to the highest level of translation equivalence on both linguistic and extra-linguistic levels (Bayar, 2007).

<sup>7</sup>Henceforth, “S-aligns” will be used to designate the establishment of a sure alignment between one or more elements of a source and a target language pair.

<sup>8</sup>Henceforth, “P-aligns” will be used to designate the establishment of a possible alignment between one or more elements of a source and a target language pair.

<sup>9</sup>A one-to-one word alignment represents the mapping of a single word in the sentence of the source language to a single word in the sentence of the target language, i.e., two words in semantic correspondence between the two languages of the language-pair (or in two paraphrases of the same language).

<sup>10</sup>A one-to-many word alignment represents the mapping of a single word in the sentence of the source language to a sequence of two or more words in the sentence of the target language; a many-to-many segment alignment represents the mapping of a sequence of words in the sentence of the source language to a sequence of words in the sentence of the target language.

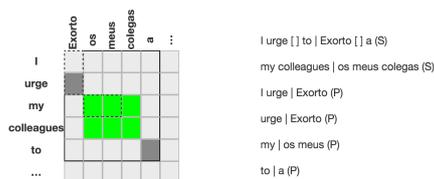


Figure 2: Phrase alignments in non-contiguous block

CLUE-Aligner, block alignments can be created by clicking and dragging the mouse cursor. A dashed line border in the grid represents a P-block-alignment, and a plain line border in the grid represents an S-block-alignment.

CLUE-Aligner also allows the annotation of non-contiguous multiword units. Non-contiguous block alignments are necessary to express alignments between multiwords or phrases, which contain words that are not part of the multiword or phrase. Non-contiguous multiword units are blocks which contain insertions - words or groups of words which are not part of the block alignment. Insertions are represented by orange cell groups (inside the block) corresponding to each insertion’s line/column (depending on whether the insertion belongs to the source/target sentence). The annotator can individually indicate which words are insertions by creating these orange cell groups. Light and dark orange cell groups represent P-insertions and S-insertions, respectively. This sureness property of insertions is determined according to the sureness of the block they belong into. Insertions (in the source and target sentences) can also be aligned. When two insertions (both within the same non-contiguous block) are aligned, their grid intersection (the cells that form this intersection) is green (cf. Figure 2). These green cell groups can also be created by the annotator. Light and dark green cell groups represent P-block and S-block (insertion) alignments, respectively, and the annotator may change this sureness property independently of the “parent” block’s sureness.

## 3.2. CLUE-Aligner Interface

CLUE-Aligner is implemented by a Perl script, which handles the server side processing of all files (such as parsing user uploaded files and updating alignments information) and generates an HTML page, displaying all the interface gadgets, with Javascript (making use of the jQuery library) for handling the logic of the application.

CLUE-Aligner Web interface divided in 3 main sections. The first section is a row of buttons and a drop-down menu displayed on top of the web page. These buttons are: “Browse Source Corpus”, “Browse Target Corpus”, “Browse Alignments”, “Load Corpus”, “Save Alignments”, and “Export Alignments”. The Browse Corpus buttons are used to upload the corpora files. CLUE-Aligner’s goal is to align words and multiwords within sentences, which means the corpora must be sentence aligned (parallel). Each half of the corpus corresponds to the source or target corpus. This means that there is a file for each half of the corpus (2 files per parallel corpus). These are text files whose format consists of one sentence per line surrounded by <s> and </s> tags. These markup tags can store arbitrary in-

formation (e.g., original corpus line numbers) as attributes, which may be important to the annotator, but unnecessary to, and therefore ignored by, the CLUE-Aligner. Sentences on the same line (in each corpus file) are considered to be aligned by CLUE-Aligner. The "Browse Alignments" button is used for uploading the alignments file where the alignments information is stored. This file is only uploaded if the annotator is resuming previous work - if the annotator is starting from scratch, there is no need to upload any file because there are no alignments saved yet. Previous work can also correspond to automatic aligning. Automatic alignments can be saved into that file (which must conform to specified format) for further manual annotator refinement in CLUE-Aligner. Note that when using a previously saved alignments file, the source and target corpus must be the same as they were when the file was saved. After selecting these files, the user must click the "Load Corpus" button so that CLUE-Aligner can load that information (corpora and alignments) and display it as grid and list of alignments (in the second and third sections). The "Save Alignments" button is used for saving the annotator's progress, by downloading an alignments file. This file contains the necessary information to resume the work performed by the annotator in the next session. It is a non-readable text file containing only indices indicating which cells in the alignment tables should be aligned according to what has been recorded in the previous session. These files may eventually be generated by an automatic aligner in the future to speed up the alignment task. The "Export Alignments" button is used for exporting all the alignments, i.e., pairs of paraphrastic (or identical) units in text format, where each line "<SOURCE TEXT> | <TARGET TEXT> (<SURENESS>)" corresponds to an alignment. This file is in text format containing one pair of alignments per line, separated by '|'. The alignments containing pairs of paraphrastic units can be used to train machine learning systems. Figure 3 illustrates the format of the alignments data and alignments text files corresponding to Figure 4. Finally, the drop-down menu in this first section lets the annotator switch between parallel sentences in the corpus.

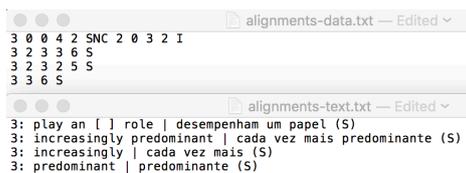


Figure 3: Alignments data and alignments text files

The second section consists of the grid itself, where the annotator can align words or multiwords. This grid is composed of cells that represent the intersection of two words, each from one of the two parallel sentences. Single word alignments are created by changing a single cell to black (sure) or grey (possible) by left-clicking the corresponding cell. Multiwords are aligned by left-clicking and dragging the cursor to include a cell group that corresponds to the desired units' intersection. This action will create an S-block alignment. Every alignment is displayed in text format in

the third section of the interface. It is also possible to annotate single/block alignments inside other blocks. It is not possible, however, to annotate blocks that cross over each other - a block can only be completely inside or completely outside another block. Non-contiguous block alignments can be created in CLUE-Aligner by annotating insertions. This can be done by first creating another block (spanning all rows/columns belonging to the block for target/source insertions, respectively) representing the corresponding insertions of the block and then switching that block's type in the third section of the interface (explained in the next paragraph). This is visualized in the third section where the missing words are represented by "[ ]". Besides being able to exclude words from a non-contiguous block alignment, the annotator can also align some (or all) of those excluded words by annotating insertion alignments. This can be done by first creating a block spanning the cell group intersection of the corresponding insertions and then switching that block's type in the third section of the interface (explained in the next paragraph). Switching a block's sureness is also done in the third section of the interface. These insertion alignments are also listed in the third section.

The third section is a listing of all alignments for the current pair of sentences. In this listing, word and multiword alignments can be visualized, changed and removed. Each alignment is displayed in the format "<SOURCE TEXT> | <TARGET TEXT> (<SURENESS>)". Each one of these lines is followed by a "Remove" button that can be clicked if the annotator wishes to remove the alignment (from both the list and grid). The annotator can left click an alignment to switch between possible and sure alignments and right click to switch block types. Block types switches can be between "regular"- "insertion" blocks or "regular"- "insertion alignment" blocks. What determines which one of these switches is done depends on the context of each specific block. If the block spans all rows/columns of an outer block, then CLUE-Aligner knows this can only be transformed into an "insertion" block and interprets the switching accordingly. If not, then CLUE-Aligner knows this can only be transformed into an "insertion alignment" block and switches the type accordingly. These type switches can be undone by left clicking any cell belonging to the corresponding green/orange group. Every change is also reflected in the grid.

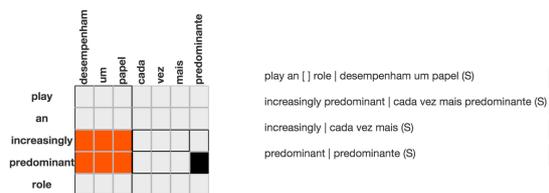


Figure 4: Insertion of an adjective modifier

Figure 4 exemplifies both "insertion" and "insertion alignment" blocks in the grid, with the corresponding alignments in text format. As can be seen, the adjective modifier insertions *increasingly predominant* in the English sentence is excluded from the alignment *play [ ] a role – desempenham*

*um papel*. To achieve this, the annotator must first create a regular block spanning those cells (for both "insertion" and "insertion alignment" blocks), yielding the temporary state pictured in Figure 5 (grid and list of pairs). The annotator must then right click the alignment *increasingly predominant – desempenham um papel* in that (temporary) list to achieve the intended result.<sup>11</sup>

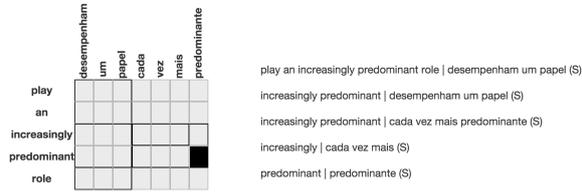


Figure 5: Insertions example temporary grid

Until now, we have used CLUE-Aligner to perform manual alignments only, but CLUE-Aligner can start by loading previously (and, possibly, automatically) generated alignments (segments) for the corpora parallel sentences. During the annotation task, the annotator manually corrects any inaccurate automatic alignments (if provided), and/or defines the new alignments for multiwords and phrasal expressions, which represent translation units. Multiwords or phrasal units in the translation unit display window can be used in future automatic monolingual and bilingual alignments. In our research, we have used them to create the Gold CLUE collection.

#### 4. Guidelines and Gold Collection

Our alignment guidelines contain statements by which to determine courses of action regarding the alignment of multiwords and phrasal expressions that correspond to pairs of translation units. The alignment task of the translation pairs of units resulted in the gold collection, achievable due to the CLUE-Aligner.

##### 4.1. CLUE4Translation Alignment Guidelines

The CLUE4Translation Alignment Guidelines are a work in progress set of guidelines that summarize the most important decisions for the alignment of pairs of multiword and other phrasal units in parallel sentences of the multilingual Europarl training corpus (Koehn, 2005). In addition to the alignment of one-to-one, one-to-many and many-to-one words, they take into special account the annotation of "phrase alignment", which includes multiwords and phrasal expressions that correspond to translation units. In the Guidelines, alignments are linguistically-informed and motivated. Although our research has led to the identification of four main classes of challenges to the alignment (lexical and semantico-syntactic, morphological, morpho-syntactic, and semantico-discursive), the focus of the CLUE4Translation Alignment Guidelines is on the lexical and semantico-syntactic phenomena, namely the align-

<sup>11</sup>In this example, the fully induced block *increasingly predominant – cada vez mais predominante* alignment is explicitly marked following the Logos Model approach, where adverbs are attached to the verbal adjectives they modify.

ment of cross-linguistic multiwords and phrasal expressions as representation objects in the alignment between the parallel sentences. These guidelines are presented with examples of support verb constructions, compound verbs, and prepositional predicates, among others, used to justify and motivate the annotation decisions. The guidelines also discuss problems, such as variability, insertion of external elements inside multiword units, and preposition selection by each language. Based on a cross-linguistic analysis of particular cases of multiwords and phrasal expressions, we have annotated valid pairs of translation units as S or P-alignments.

The CLUE4Translation Alignments Guidelines were inspired by the Logos Model (Scott, 2003; Barreiro et al., 2011), which relies on deep semantico-syntactic analysis to translate not only contiguous multiword units, such as the support verb construction *to draw a distinction between*, but also non-contiguous multiword units, such as the support verb construction *to bring* [INSERTION] *to a conclusion*, as in the sentence *I would urge the European Commission to bring the process of adopting the directive on additional pensions to a conclusion*, often mistranslated by machine translation systems.

##### 4.2. Gold CLUE4Translation

The Gold CLUE4Translation is the gold collection made of 6 sets of aligned corpora resultant from our alignment task. We used the common test version of the European Parliament Proceedings taken from Q4/2000 portion of the data, 2000-10 to 2000-12 (Koehn, 2005). The bilingual texts are available at the European Parliament Proceedings Parallel Corpus website.<sup>12</sup> The reference sub-corpus is aligned at the sentence level, ranging from sentence number 101 to sentence number 500.

The gold collection was achieved using the following methodology. A polyglot linguist, with knowledge of all languages contemplated in this study, annotated manually a total of 2,400 sentence alignments (400 x 6 language pairs) and built the CLUE4Translation Alignment Guidelines based on linguistic knowledge inspired in the Logos Model, paying special attention to multiwords and phrasal expressions that correspond to translation units. The annotator has performed two tasks simultaneously, multiword identification and alignment. The annotation criterium used by the annotator to decide whether an alignment is an S-alignment rather than a P-alignment was the quality of being optimal or non-optimal and the quality of reciprocity with regards to source and target language. Statistics on the type (S or P) and the total number of word alignments (WA) and multiword unit alignments (MWU) for each language pair are presented in Table 1.

Figure 6 exemplifies the adverbial insertions *still* and *ainda*, excluded from the alignment pair *There is [ ] a need to – é [ ] necessário que*. Since these insertions are direct translations of each other, they are, therefore, aligned with each other. The same can be said of the noun phrase insertions *entrepreneurs* and *os empresários*, excluded from the alignment pair *give [ ] easier access to – seja fácil [ ] recorrer a*, which are S-block-aligned as a pair of translation

<sup>12</sup>[www.statmt.org/europarl/archives.html#v1](http://www.statmt.org/europarl/archives.html#v1)

Pair	WA			MWU		
	S	P	Total	S	P	Total
EN-ES	7581	430	8011	2962	785	3747
EN-FR	6995	586	7581	3401	740	4141
EN-PT	6443	607	7050	1547	1399	2946
ES-FR	8488	589	9077	2806	403	3209
PT-ES	7945	566	8511	3139	387	3526
PT-FR	7022	787	7809	3220	700	3920

Table 1: Number of WA and MWU per language pair

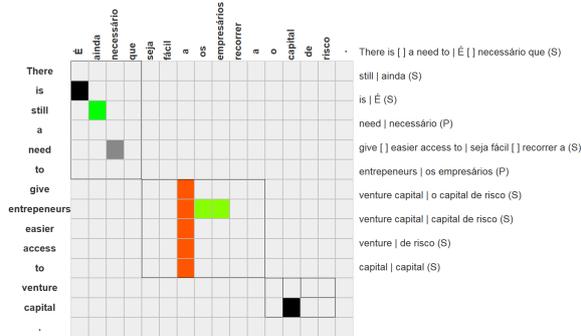


Figure 6: Insertions example grid

units. Note that even though these insertions are inside an S-aligned block, they can also be P-aligned (as is the case with *entrepreneurs* and *os empresários*). An outer block alignment's "sureness" does not impose the same "sureness" for inner insertion alignments. Individual alignments outside the insertions' intersection can also be either S or P-aligned, as for example *is* – *é* and *need* – *necessário*, respectively.

## 5. Conclusions and Future Directions

Linguistic-based alignments extracted from good quality translation corpora can contribute to increased precision and recall in SMT systems, with the subsequent improvement of translation quality. They are also a valuable asset for applications that require monolingual paraphrases.

In this paper, we have presented CLUE-Aligner, a new web alignment tool that allows the alignment of non-contiguous multiwords and phrasal expressions to improve translation applications. Even though we move forward by creating an alignment tool that handles non-adjacent structures, there is still room for improvement, such as being able to feed the CLUE-Aligner with existing translation or paraphrastic knowledge previously aligned or generated with a linguistic processing tool. Future work aims the enhancement of CLUE-Aligner to align and extract automatically large amounts of alignment pairs to be applied to paraphrasing and machine translation case studies.

## Acknowledgements

This research work was supported by Fundação para a Ciência e Tecnologia (FCT), under project eSPERTO – EXPL/MHC-LIN/2260/2013, UID/CEC/50021/2013, and post-doctoral grant SFRH/BPD/91446/2012. The authors

of this paper would like to thank the three anonymous reviewers for their helpful comments.

## 6. Bibliographical References

- Ambati, V., Vogel, S., and Carbonell, J. (2010). Active semi-supervised learning for improving word alignment. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*.
- Barreiro, A., Scott, B., Kasper, W., and Kiefer, B. (2011). OpenLogos Rule-Based Machine Translation: Philosophy, Model, Resources and Customization. *Machine Translation*, 25(2):107–126.
- Barreiro, A. (2016). Contributos para o Aumento de Qualidade na Língua Digital. In José Teixeira, editor, *O Português como Língua num Mundo Global: problemas e potencialidades*. Edições Húmus.
- Bayar, M. (2007). *To Mean or Not to Mean*. Ph.D. thesis, Kadmous cultural foundation.
- Blunsom, P. and Cohn, T. (2006). Discriminative word alignment with conditional random fields. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, ACL-44*, pages 65–72. Association for Computational Linguistics.
- Bojar, O. and Prokopová, M. (2006). Czech-English Word Alignment. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1236–1239. ELRA, May.
- Brown, P. F., Cocke, J., Pietra, S. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85.
- Callison-Burch, C. and Bannard, C. (2004). Improving statistical translation through editing. European Association for Machine Translation (EAMT-04) workshop. In *European Association for Machine Translation*.
- Callison-Burch, C. (2007). *Paraphrasing and Translation*. Ph.D. thesis, University of Edinburgh, Edinburgh, Scotland.
- Galley, M., Hopkins, M., Knight, K., and Marcu, D. (2004). What's in a translation rule? In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT-NAACL '04)*, Boston, USA.
- Gao, Q. and Vogel, S. (2010). Consensus versus expertise: a case study of word alignment with mechanical turk. In *Association for Computational Linguistics*, pages 30–34.
- Germann, U. (2008). Yawat: Yet Another Word Alignment Tool. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session, HLT-Demonstrations '08*, pages 20–23.
- Gilmanov, T., Scrivner, O., and Kübler, S. (2014). SWIFT Aligner, A Multifunctional Tool for Parallel Corpora: Visualization, Word Alignment, and (Morpho)-Syntactic Cross-Language Transfer. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.
- Graça, J., Pardal, J. P., Coheur, L., and Caseiro, D. (2008). Building a Golden Collection of Parallel Multi-Language

- Word Alignment. In Nicoletta Calzolari, et al., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.
- Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical Phrase-Based Translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL '03)*, pages 48–54, Morristown, NJ, USA. Association for Computational Linguistics.
- Koehn, P. (2005). EuroParl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86. AAMT.
- Kruijff-Korbayová, I., Chvátalová, K., and Postolache, O. (2006). Annotation Guidelines for Czech-English Word Alignment. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1256–1261.
- Lambert, P., Gispert, A. D., Banchs, R., and Mariño, J. B. (2005). Guidelines for Word Alignment Evaluation and Manual Alignment. *Language Resources and Evaluation*, 39(4):267–285.
- Melamed, I. D. (1998). Annotation Style Guide for the Blinker Project. Technical report, IRCS.
- Och, F. J. and Ney, H. (2000). Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Meeting on Association For Computational Linguistics*, Hong Kong.
- Rassier, B. and Pedersen, T. (2003). Alpaco: Aligner for parallel corpora.
- Scott, B. E. (2003). The Logos Model: An Historical Perspective. *Machine Translation*, 18(1):1–72.
- Tiedemann, J. (2003). Combining Clues for Word Alignment. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 12–17, Budapest, Hungary.
- Tiedemann, J. (2011). *Bitext Alignment*. Morgan and Claypool.
- Vickrey, D., Biewald, L., Teyssier, M., and Koller, D. (2005). Word-Sense Disambiguation for Machine Translation. In *EMNLP*, pages 771–778.

## The IWSLT Evaluation Campaign: Challenges, Achievements, Future Directions

Luisa Bentivogli<sup>1</sup>, Marcello Federico<sup>1</sup>, Sebastian Stüker<sup>2</sup>, Mauro Cettolo<sup>1</sup>, Jan Niehues<sup>2</sup>

<sup>1</sup>FBK - Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy

<sup>2</sup>KIT - Karlsruhe Institut of Technology, Adenauerring 2, 76131 Karlsruhe, Germany

### Abstract

Evaluation campaigns are the most successful modality for promoting the assessment of the state of the art of a field on a specific task. Within the field of Machine Translation (MT), the International Workshop on Spoken Language Translation (IWSLT) is a yearly scientific workshop, associated with an open evaluation campaign on spoken language translation. The IWSLT campaign, which is the only one addressing speech translation, started in 2004 and will feature its 13th installment in 2016. Since its beginning, the campaign attracted around 70 different participating teams from all over the world. In this paper we present the main characteristics of the tasks offered within IWSLT, as well as the evaluation framework adopted and the data made available to the research community. We also analyse and discuss the progress made by the systems along the years for the most addressed and long-standing tasks and we share ideas about new challenging data and interesting application scenarios to test the utility of MT systems in real tasks.

**Keywords:** Evaluation Campaign, Spoken Language Translation, Machine Translation

### 1. Introduction

Evaluation based on measurable and shared criteria has always been an essential component of scientific research, and constitutes the hallmark of any well established research field. Shared evaluation criteria and accepted evaluation practices help in promoting the most promising scientific approaches, and thus foster the quick production of technological advancements. They also contribute to strengthen the scientific relationships and the self-awareness within a research community, and they can encourage the involvement of newcomers in the field, by providing clearly defined scientific and technological objectives, and benchmarks for evaluating them. Evaluation campaigns are the most successful modality for promoting the assessment of the state of the art of a field on a specific task.

Within the field of Machine Translation (MT), the *International Workshop on Spoken Language Translation* (IWSLT) is a yearly scientific workshop, associated with an open evaluation campaign on spoken language translation. The IWSLT campaign, which is the only one addressing speech translation, started in 2004 and will feature its 13th installment in 2016. IWSLT's evaluations are not competition-oriented, since their goal is to favor cooperative work and scientific exchange. In this respect, IWSLT proposes challenging research tasks and an open experimental infrastructure for the scientific community working on spoken language translation.

In the following, after introducing the evaluation campaign, we present the peculiarities and challenges of spoken language translation (Section 2). We then describe the main characteristics of the offered tasks, as well as the data sets and the evaluation infrastructure made available to the community (Section 3). We also present how human evaluation evolved from adequacy/fluency assessment to relative ranking, and finally to post-editing performed by professional translators, pursuing the objective of maximising the benefit to the research community, both in terms of information about MT systems and data and resources to be reused

(Section 4). To complete the overview on the evolution of the evaluation campaign, we analyse the progresses made by the systems along the years for the most addressed and long-standing tasks (Section 5). Finally, we conclude presenting ideas about new challenging data and interesting application scenario to test the utility of MT systems in real tasks (Section 6).

### 2. The Evaluation Campaign

The IWSLT workshop was started in 2004 with the purpose of enabling the exchange of knowledge among researchers working on speech-to-speech translation and creating an opportunity to enhance the MT systems by comparing technologies on a common test bed. The campaign built on one of the outcomes of the C-STAR (Consortium for Speech Translation Advanced Research) project, namely the BTEC (Basic Travel Expression Corpus) multilingual spoken language corpus (Takezawa et al., 2002), which served as a primary source of evaluation. Since its beginning, increasingly challenging translation tasks were offered and new data sets covering a huge number of language pairs were shared with the research community.

In the twelve editions organized from 2004 to 2015, the campaign attracted around 70 different participating teams from all over the world. Figure 1 presents the number of different teams participating in each round of the campaign.

The task of speech translation is particularly challenging for a number of reasons. On one side, MT systems are required to deal with the specific features of spoken language. With respect to written language, speech is structurally less complex, formal and fluent. It is also characterized by shorter sentences with a lower amount of rephrasing but a higher pronoun density (Ruiz and Federico, 2014). On the other side, speech translation (Casacuberta et al., 2008) requires the integration of MT with automatic speech recognition, which brings with it the additional difficulty of translating content that may have been corrupted by speech recognition errors.

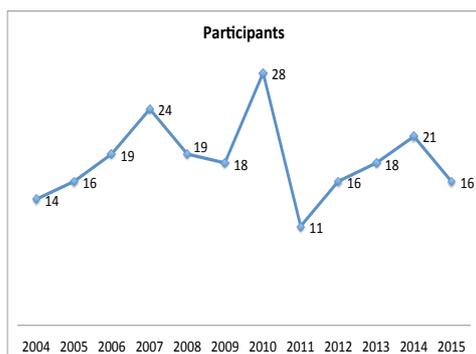


Figure 1: Number of teams that participated in the IWSLT evaluation campaigns.

Along the years, three main evaluation tracks were progressively introduced, addressing all the core technologies involved in the spoken language translation task, namely:

- Automatic speech recognition (ASR), *i.e.* the conversion of a speech signal into a transcript
- Machine translation (MT), *i.e.* the translation of a polished transcript into another language
- Spoken language translation (SLT), *i.e.* the conversion and translation of a speech signal into a transcript in another language

In the first IWSLT campaign in 2004 only the MT track was offered. Since correct human transcriptions were given as input to the MT systems, the task allowed to focus on the specific challenges related to the translation of spoken language.

Starting from 2005, also the SLT track was proposed, in order to include the additional challenge of dealing with automatic transcriptions of the audio signal, and thus investigating the impact of recognition errors on the MT performance. Participants in the SLT track could either use their own ASR systems or the ASR outputs provided by the organizers to facilitate participation. Depending on the year, different types of ASR outputs were released, such as first best output, n-best lists, lattices, ROVER combination of the outputs submitted to the ASR tracks.

The ASR track, which was offered starting from IWSLT 2011, is out of the scope of this paper since it is specifically devoted to the evaluation of speech recognition systems and does not address MT evaluation.

### 3. Tasks and Challenges

The first IWSLT task (Akiba et al., 2004) addressed the translation of *read-speech* transcripts in the travel domain. It was based on the BTEC corpus, which is a collection of sentences that bilingual travel experts consider useful for people going to or coming from another country and cover utterances for every potential subject in travel situations. The BTEC task was replicated in the second round of IWSLT (Eck and Hori, 2005) and was offered as “Classical” task until 2010 so to give continuity with previous

editions and allow new and old participants to test their systems against a standard setting.

Starting from 2006, new and progressively more challenging tasks were added to the BTEC task, aiming at keeping the interest of the research community high by introducing more realistic scenarios. The new focus was the translation of *spontaneous speech*, while the tourism domain was maintained.

For these so-called “Challenge” tasks, different types of speech data – recorded in realistic settings – were collected, namely *answers* to travel-related questions (Paul, 2006), *monolingual dialogues* from travel agent and client interactions via telephone (Fordyce, 2007), *machine-mediated dialogues* where foreign travelers were asked to use a state-of-the-art speech-to-speech translation device to communicate with local staff (Paul, 2008), *cross-lingual human-mediated dialogues* in travel situations, where the uttered sentences were simultaneously translated by interpreters (Paul, 2009; Paul et al., 2010), and finally human dialogs in travel situations closely related to the Beijing 2008 Olympic Games (Federico et al., 2012b).

In 2010, the seventh round of IWSLT presented a mixture of innovation and continuity with the previous campaigns. Besides the Classical BTEC and Challenge Dialog tasks, a completely new task was piloted, which marked a major change with respect to previous tasks.

The new pilot task was based on TED Talks,<sup>1</sup> a collection of recordings of public speeches covering a wide variety of topics. Each talk is delivered in a brilliant and original style by a very skilled speaker and, while addressing a wide audience, it pursues the goal of both entertaining and persuading the listeners on a specific idea. For each talk, transcriptions and translations into several languages are provided by volunteers worldwide.

The proposed new challenge departed from and completed the application scenarios proposed till then in the IWSLT evaluations. On one side, the communication modality changed from dialogue to monologue and the language style passed from spontaneous to planned. On the other side, TED Talks data are recordings of really occurring open-domain speeches vs. speeches recorded in realistic situations within a restricted domain. Furthermore, from an application perspective, the TED Talks task is a captioning scenario, which suggests translation tasks ranging from off-line translation of written captions, up to on-line speech translation, requiring a tight integration of MT with ASR possibly handling stream-based processing.

The TED Talks task embeds interesting research challenges which are unique among the available speech recognition and machine translation benchmarks, such as coping with (i) background noise—e.g., applause and laughter from the audience—, (ii) different speakers—e.g., accents including non native speakers, varying speaking rates, prosodic aspects—, and (iii) limited in-domain training data and variability of topics and styles.

The TED Talks task became the main IWSLT task in 2011, and was offered to participants up to the last IWSLT edition in 2015 (Federico et al., 2011; Federico et al., 2012b; Cet-

<sup>1</sup>www.ted.com

tolo et al., 2013; Cettolo et al., 2014; Cettolo et al., 2015). A major benefit to the community with respect to previous tasks lies in the public availability of TED Talks. While the BTEC corpus and all the other datasets used in the “Challenge” tasks were licenced only to IWSLT participants, TED talks video recordings, transcripts, and translations are distributed from the TED website under a Creative Commons license. Aiming at maximizing the sharing of resources, starting from 2012, the TED datasets used in the IWSLT evaluations were distributed through the WIT<sup>3</sup> web repository (Cettolo et al., 2012).<sup>2</sup> The purpose of this repository is to make the collection of TED talks effectively usable by the NLP community. Besides offering ready-to-use parallel corpora, the WIT<sup>3</sup> repository also offers MT benchmarks and text-processing tools designed for the TED talks collection.

The various IWSLT tasks described above were offered for a remarkable number of language pairs which changed along the years. Both distant language pairs—typically involving English and Japanese, Chinese, Korean, Arabic, Turkish—and languages belonging to the same family, such as German, French, Italian, English, and many others, were addressed. All details can be found in the IWSLT overview papers.

Finally, official evaluation specifications were defined for the IWSLT tasks and require MT output to be (i) case sensitive and (ii) with punctuation marks. These specifications were chosen to serve the double purpose of delivering usable translations and making IWSLT evaluation results comparable to outcomes of other MT evaluation initiatives. In addition, automatic evaluation scores have always been calculated also for the case-insensitive (lower-case only) and no-punctuation setting.

In line with other major evaluation campaigns in the MT field, both automatic metrics and human assessments are used to evaluate submissions to IWSLT. As for automatic metrics, BLEU has always been the primary metric used to rank the participating systems; furthermore, along the years additional standard metrics have been calculated, such as METEOR, WER, PER, TER, GTM, and NIST.

An important novelty introduced in IWSLT 2015 is the availability of an evaluation server, developed with the purpose of allowing participants to assess their progresses automatically and in identical conditions. Participants could submit the translation of any development set to the evaluation server, receiving scores calculated with BLEU, NIST, and TER. The evaluation server was used by the organizers for the automatic evaluation of the official submissions and, after the evaluation period, the evaluation on test sets was enabled to all participants as well. The evaluation server is maintained active and new datasets will be added for evaluations in the next campaigns.

#### 4. Human Evaluation

Although automatic evaluation plays a very important role in fostering MT research, human evaluation is crucial aspect for an evaluation campaign. On the one hand, it provides the most direct and reliable assessment of translation

quality; on the other, it is used to validate and improve automatic metrics by measuring their correlation with human judgments.

A distinguishing characteristic of IWSLT is the attention paid to the quality of human evaluation. For this reason, human evaluation was not done on a voluntary basis but was typically carried out by paid evaluators. However, it is well-known that collecting human judgments of MT outputs is time consuming and expensive, especially on the scale of an evaluation campaign. In order to find a trade off between human evaluation quality and costs, evaluation was limited to a subset of submitted runs and test data.

In the first IWSLT campaign, the standard methodology followed in other MT evaluations was adopted, where systems were judged on the basis of *fluency* and *adequacy* (White et al., 1994). Fluency refers to the degree to which the translation is well-formed according to the grammar of the target language, while adequacy refers to the degree to which the translation contains the information present in the source. This methodology was used for the first three evaluation campaigns, while in IWSLT 2007 a new methodology was introduced. In fact, studies on the reliability of human evaluation demonstrated that ranking judgments, in which annotators rank MT systems with respect to each other, are shown to have higher inter-annotator and intra-annotator agreement than adequacy and fluency judgments (Callison-Burch et al., 2007). For this reason, in IWSLT 2007 the *Ranking* task was introduced. In this task, for each source sentence five MT outputs (randomly sampled from those submitted) are presented to the evaluator, who must rank them from best to worst using a five-point scale. The collected judgments are used to obtain the ranking scores, which are calculated as the average number of times that a system was judged better than any other system. In addition to the ranking task, the evaluation based on fluency and adequacy was also carried out until IWSLT 2010 for comparison purposes.

IWSLT 2011 represents a major change in the evolution of human evaluation, since it focused solely on the ranking task and introduced a number of novelties with respect to the traditional ranking evaluation carried out in previous campaigns.

The major change was that the evaluation was not carried out by hired expert graders but relying on crowdsourced data. This choice was motivated by the results of a previous experiment on IWSLT data (Bentivogli et al., 2011), which demonstrated the feasibility of using crowdsourcing methodologies as an effective way to reduce the costs of MT evaluation without sacrificing quality.

The cost reduction obtained by using crowdsourcing allowed the modification of the ranking methodology in various respects, with the aim of maximizing the overall evaluation reliability. First, the traditional five-fold ranking task involving the evaluation of five translated sentences at a time was abandoned in favor of a direct comparison between two translated sentences only, which is a more reliable task due to the lower cognitive load required to perform it. Second—and differently from previous campaigns—to ensure system ranking reliability, full coverage of pairwise comparisons was achieved following

<sup>2</sup><http://wit3.fbk.eu>

a *round-robin* tournament, in which each system competes against every other system.

Following the practice consolidated in the previous campaign, the IWSLT 2012 evaluation was also carried out with ranking judgments collected through crowdsourcing. However, the goal for 2012 was to find a tournament structure comparable with round robin in terms of reliability, but requiring less comparisons in favor of cost effectiveness. The most suitable structure, given its ability of ranking all competitors and the relatively few comparisons required, turned out to be the *Double Seeded Knockout with Consolation* tournament, which was thus adopted for the evaluation.

IWSLT 2013 saw the introduction of the last major novelty in human evaluation. The Ranking task was substituted by a *Post-Editing* task and, accordingly, HTER (Human-mediated Translation Edit Rate) was adopted as the official evaluation metric to rank the systems.

Post-Editing, i.e. the manual correction of machine translation output, has long been investigated by the translation industry as a form of machine assistance to reduce the costs of human translation. Nowadays, Computer-aided translation (CAT) tools incorporate post-editing functionalities, and a number of studies (Federico et al., 2012a; Green et al., 2013) demonstrate the usefulness of MT to increase professional translators' productivity. The MT TED task offered in IWSLT can be seen as an interesting application scenario to test the utility of MT systems in a real subtitling task.

From the point of view of the evaluation campaign, the goal was to adopt a human evaluation framework able to maximize the benefit to the research community, both in terms of information about MT systems and data and resources to be reused. With respect to previously adopted evaluation methodologies (i.e. adequacy/fluency and ranking tasks), the post-editing task has the double advantage of producing (i) a set of edits pointing to specific translation errors, and (ii) a set of additional reference translations. Both these byproducts are very useful for MT system development and evaluation. Human evaluation based on post-editing was adopted also in IWSLT 2014 and 2015.

## 5. Trends in System Performance

Our analysis focuses on the MT tracks organised over the period 2012-2015, which considered the translation of TED talks from English into language X, as well as the translation of TEDx talks given in language X into English. Tracking the progress on this task is not straightforward, as every year new evaluation sets and new training data were released. In fact, machine translation performance varies from evaluation set to evaluation set, independently from the relative improvements of the systems over the years. These random variations can be so large that they may hide the progress of the systems. Another factor that influences the absolute performance of a system is the amount of available training data. Exploiting more data, especially in-domain data, generally leads to better performance.

In order to neutralize the random effects introduced by the different test sets and the different in-domain training sets, we do compare performance of systems relative to standardised baseline systems. In particular, each baseline system is trained in exactly the same way, over the years, with

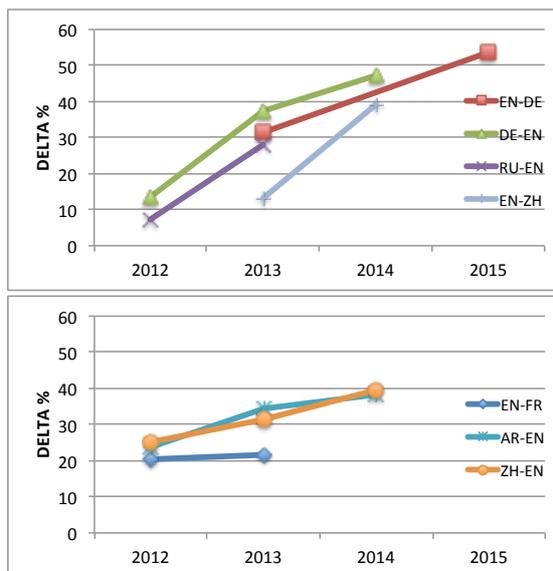


Figure 2: Performance trends over popular language pairs in terms of relative (%) improvement over the standardised baseline system.

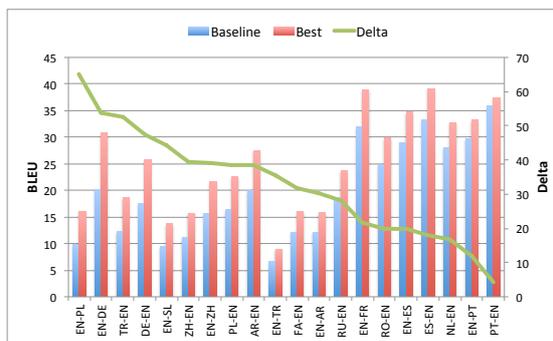


Figure 3: Best relative improvements (Delta) measured for each language pairs covered by at least two IWSLT editions, during the period 2012-2015. BLEU scores of the corresponding systems and baselines are reported, too.

the in-domain training data for each year and tested on the corresponding evaluation set.

Figure 2 plots the relative differences between the BLEU scores of the best system and its corresponding baseline, for a range of language pair and years. The figures are provided for the most popular translation directions and to the years where a positive trend was observed. More precisely, this excludes the cases in which the best system in one year overtakes the baseline by less than the systems of the previous years did. Our underlying assumption is that performance of MT systems developed for this task should not get worse over time. The fact that this *monotonic progress* behavior is not observed in the IWSLT evaluation is mainly due to the participant turnover, i.e. the top system of one year does not show up in the following years.

Figure 3 shows instead the overall best results in terms of BLEU score improvement - i.e. best system vs. base-

Table 1: Example of a sentence pair from the QED data

Language	Transcript
English	So in this video I'm just going to do a ton of examples.
German	Daher werde ich in diesem Video viele Beispiele durchrechnen.

line - for all translation directions that were proposed at least twice during the period 2012-2015. Language pairs are sorted by the observed improvement (Delta) with respect to the reference baseline system. This plot also shows the actual BLEU scores of the systems and baselines. Although BLEU scores on different language pairs are not directly comparable among each other, they can give a rough idea of the level of performance achieved by the baseline systems and consequently of the level of difficulty of each translation direction.

By considering the language pairs in Figure 3, the average relative BLEU score improvement over the considered period is about 33%. In particular, remarkable performance gains were achieved for English-German (53.64%), German-English (47.26%), Chinese-English (39.48%), English-Chinese (38.57%) and Arabic-English (38.43%). In fact, these improvements are the result of significant progressions in performance over time (see Figure 2). On the other hand, less progress (21.67%) has been observed on a very popular language pair such as English-French (Figure 3). A probable explanation could be that this translation pair is hard to improve because its performance is already high (BLEU score is over 35). In fact, Figure 2 confirms that lower improvements (Delta values) are in general observed for languages having better performing baselines.

## 6. Future Directions for Spoken Language Translation Evaluation

The TED translation task of IWSLT has become a seasoned task by now. Its introduction was motivated by its higher complexity with respect to the previous travel tasks, and by the availability of high quality data. In order to keep the tasks interesting and to follow current trends in research and industry, we are going to expand and develop the IWSLT tasks further, starting with the evaluation campaign of 2016. We will augment the TED Talk task by including more challenging data from the QCRI Educational Domain (QED) Corpus<sup>3</sup> (Abdelali et al., 2014). Further, we will introduce a new task on Skype conversations. Unlike in previous years we will limit the scope of the evaluation to few languages: English, German, French, and one low resourced European language. The main reason for this is to avoid dispersion of participants in too many tasks.

### 6.1. Extended Lecture Task

TED talks are challenging due to their variety in topics, which can be considered unlimited for all practical purposes. With respect to the type of language, TED talks are,

<sup>3</sup><http://alt.qcri.org/resources/qedcorpus/>

Table 2: Example of a sentence pair from the Skype data

Language	Transcript
German	ähm wir haben grade über Platten geredet, und über, über Musik, Musik Stream, was mich halt irgendwie nervt ist das bei so vielen Platten vorn so krass viel Werbung dazwischen geschaltet wird, und das find ich äh sehr störend, ja.
English	We just talked about albums and about streaming music, which just bugs me somehow, that for so many albums, so much advertising is placed before and in between them. And I find that very disruptive, yes.

however, very well behaved. Before being delivered, TED talks are rehearsed rigorously. Therefore, the talks tend not to show spontaneous speech phenomena, but are rather well formed. However, the majority of talks held in the world are not that well formed and well rehearsed, but rather more spontaneous and of lower quality. A prominent example of such type of talk is given by academic lectures. In order to address more lifelike talks, we are going to include data from the QED corpus (Abdelali et al., 2014) into our lecture task. This data is obtained from subtitles created on the Amara platform of videos from Khan Academy, Coursera, Udacity, etc. Table 1 gives an example of a transcription and translation from the corpus.

### 6.2. Skype Translation Task

Recently Microsoft has introduced its Skype Translator.<sup>4</sup> Translating Skype or video conference conversations is a challenging task due to the nature of the language used in conversations, which is often not planned, informal in nature, ungrammatical, using special idioms etc. Therefore, while maybe not as broad in domain as talks and lectures, this task represents a challenge that goes beyond the translation of TED talks.

The test data that will be made available from Microsoft Research consist of bilingual conversations, where each speaker was speaking in his own language but was able to understand the other dialog partner's language. In this way natural conversations could be recorded. Audio was then manually processed to produce transcripts, transformed transcripts (cleaned of disfluencies), and translations (in or out of English). Table 2 shows an example from such a dialogue in English and German.

### 6.3. Evaluation

We expect to evaluate the extended lecture task under the post-editing perspective, exactly as we have done for the TED talk task. For the Skype Translator task, instead, we plan to opt for an adequacy-oriented evaluation, given that the focus of this communication scenario is the exchange of content between two parties. For the incoming campaign, we plan to apply human evaluation only for the extended

<sup>4</sup>(<http://research.microsoft.com/en-us/about/speech-to-speech-milestones.aspx>)

lecture task and to ground it again on the post-edition of MT outputs by professional translators. For the Skype Translator task, on the basis of the performance and output variability that we will observe, we will decide if to apply in the future (starting from 2017) human evaluations based on ranking or Likert scales.

## 7. Acknowledgments

Work by FBK's authors was partially funded by the EU H2020 project CRACKER, grant agreement no. 645357.

## 8. Bibliographical References

- Abdelali, A., Guzman, F., Sajjad, H., and Vogel, S. (2014). The amara corpus: Building parallel language resources for the educational domain. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).
- Akiba, Y., Federico, M., Kando, N., Nakaiwa, H., Paul, M., and Tsujii, J. (2004). Overview of the IWSLT04 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–12, Kyoto, Japan.
- Bentivogli, L., Federico, M., Moretti, G., and Paul, M. (2011). Getting Expert Quality from the Crowd for Machine Translation Evaluation. In *Proceedings of the MT Summit XIII*, pages 521–528, Xiamen, China.
- Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). (Meta-) Evaluation of Machine Translation. In *Proceedings of the 2nd Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic.
- Casacuberta, F., Federico, M., Ney, H., and Vidal, E. (2008). Recent efforts in spoken language processing. *IEEE Signal Processing Magazine*, 25(3):80–88, May.
- Cettolo, M., Girardi, C., and Federico, M. (2012). WIT<sup>3</sup>: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the Annual Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May.
- Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., and Federico, M. (2013). Report on the 10th IWSLT Evaluation Campaign. In *Proceedings of the Tenth International Workshop on Spoken Language Translation (IWSLT 2013)*, Heidelberg, Germany.
- Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., and Federico, M. (2014). Report on the 11th IWSLT Evaluation Campaign, IWSLT 2014. In *Proceedings of the Eleventh International Workshop on Spoken Language Translation (IWSLT 2014)*, Lake Tahoe, USA.
- Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., Cattoni, R., and Federico, M. (2015). The IWSLT 2015 Evaluation Campaign. In *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT 2015)*, Da Nang, Vietnam.
- Eck, M. and Hori, C. (2005). Overview of the IWSLT 2005 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–22, Pittsburgh, PA.
- Federico, M., Bentivogli, L., Paul, M., and Stüker, S. (2011). Overview of the IWSLT 2011 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 11–27, San Francisco, USA.
- Federico, M., Cattelan, A., and Trombetti, M. (2012a). Measuring user productivity in machine translation enhanced computer assisted translation. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)*.
- Federico, M., Cettolo, M., Bentivogli, L., Paul, M., and Stüker, S. (2012b). Overview of the IWSLT 2012 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 11–27, Hong Kong, HK.
- Fordyce, C. S. (2007). Overview of the IWSLT 2007 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–12, Trento, Italy.
- Green, S., Heer, J., and Manning, C. D. (2013). The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 439–448. ACM.
- Paul, M., Federico, M., and Stüker, S. (2010). Overview of the IWSLT 2010 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 3–27, Paris, France.
- Paul, M. (2006). Overview of the IWSLT 2006 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–15, Kyoto, Japan.
- Paul, M. (2008). Overview of the IWSLT 2008 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–17, Waikiki, Hawaii.
- Paul, M. (2009). Overview of the IWSLT 2009 Evaluation Campaign. In *Proceedings of the sixth International Workshop on Spoken Language Translation*, pages 1–18, Tokyo, Japan.
- Ruiz, N. and Federico, M. (2014). Complexity of spoken versus written language for machine translation. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 173–180, Dubrovnik, Croatia.
- Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H., and Yamamoto, S. (2002). Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 147–152.
- White, J. S., OConnell, T., and OMara, F. (1994). The arpa mt evaluation methodologies: evolution, lessons, and future approaches. In *Proceedings of AMTA*, pages 193–205.

## A Pilot Eye-Tracking Study of WMT-Style Ranking Evaluation

Ondřej Bojar,<sup>\*</sup> Filip Děchtěrenko,<sup>‡</sup> Maria Zelenina<sup>#</sup>

<sup>\*</sup>Charles University in Prague, Faculty of Mathematics and Physics, ÚFAL

<sup>‡</sup>Institute of Psychology, The Czech Academy of Sciences

<sup>#</sup>ITMO University, St. Petersburg, Russia

bojar@ufal.mff.cuni.cz, filip.dechterenko@gmail.com, marie.zelenina@gmail.com

### Abstract

The shared translation task of the Workshop of Statistical Machine Translation (WMT) is one of the key annual events of the field. Participating machine translation systems in WMT translation task are manually evaluated by relatively ranking five candidate translations of a given sentence. This style of evaluation has been used since 2007 with some discussion on interpreting the collected judgements but virtually no insight into what the annotators are actually doing. The scoring task is relatively cognitively demanding and many scoring strategies are possible, influencing the reliability of the final judgements. In this paper, we describe our first steps towards explaining the scoring task: we run the scoring under an eye-tracker and monitor what the annotators do. At the current stage, our results are more of a proof-of-concept, testing the feasibility of eye tracking for the analysis of such a complex MT evaluation setup.

**Keywords:** MT Evaluation, WMT, Eye Tracking

### 1. Introduction

Despite benefits and popularity of automatic Machine Translation (MT) evaluation metrics, human annotation stays the gold standard for evaluating and comparing MT outputs. Many styles and purposes of manual MT evaluation have been studied in the past, from judging *fluency* and *adequacy* (Koehn and Monz, 2006) over a single absolute score scale (Graham, 2015) or relative ranking of complete MT hypotheses or their parts (Callison-Burch et al., 2007) to correcting outputs (Specia and Farzindar, 2010), marking of errors (Fishel et al., 2012; Lommel et al., 2014b), cloze tests (Ageeva et al., 2015) or approximations of task-based evaluation (Berka et al., 2011); see (Koehn, 2007) for a small survey.

Moreover, automatic MT evaluation metrics usually heavily rely on human judgements. One part of this input are naturally the reference translations or some clever variation of them (Dreyer and Marcu, 2012; Bojar et al., 2013a). Another very important part are the manual quality judgements, because more and more metrics are trained to approximate best the manual scores (Stanojević et al., 2015). Thus, the quality of automatic metrics depends on human evaluations that are used as training data and benchmarks. To sum up, human evaluation of MT to date is crucial for setting an adequate standard of quality for MT systems and steering development loop in the right direction (including the area of automatic MT evaluation). Therefore, to create better MT systems, we need to deeply understand the process of human evaluation and analyse its possible shortcomings.

In this study, we focus on the human annotation procedure used to evaluate submissions for the Shared Translation Task in Workshop of Machine Translation (WMT<sup>1</sup>). WMT is an annual flagship event for the MT community, and the yearly overview paper sets the state of the art in the field. WMT relies on manual MT evaluation, and considering the importance and the scope of influence of WMT, the quality of the manual evaluation is of critical importance.

Concerns about reliability of the manual evaluation used for WMT Shared Task have been raised frequently by the research community. Arguably, the most concerning issue is the relatively low rate of inter-annotator agreement (IAA). There have been many attempts to understand the reasons for annotator disagreement. For example, Bojar et al. (2011) noted some problems of the technique in the way judgements were interpreted, the calculation of scores due to ties or a relation between the number of times a system is judged and the evaluation score. Also, they observed that sentence length played a role: for longer sentences, IAA was lower, possibly due to limited attention span of annotators or due to larger portions of text that can be incomparably wrong.

Vela and van Genabith (2015) attempted to explain the low IAA scores by the lack of special knowledge and skills translators acquire during their study process, in contrast to WMT annotators who are recruited from MT researchers or lay people on Amazon Mechanical Turk. Vela and van Genabith (2015) reproduced WMT13 evaluation procedure, but employed professional translation students as annotators. They generally managed to reproduce the results of WMT evaluation, but IAA scores were indeed higher. This can be attributed to both special knowledge, such as a strong background in linguistics, better understanding of hierarchy and categorisation of errors, general homogeneity of the group in the level of education and background, and specially trained skills.

Lommel et al. (2014a) performed a general in-depth analysis of IAA on errors in MT, without using the WMT interface. They identified several points of disagreement between annotators. Those include scope of span-level annotations, defined as precise scope of errors; error categorisation; hierarchy (how crucial is the error). While this sheds the light on issues which lead to confusion between annotators, an important observation is that evaluators disagree on specific, fine grained description of errors, but generally agree on binary decision whether MT output is erroneous or not. The same observation was made by Styrmne and Ahrenberg (2012), who reported higher IAA scores when a

<sup>1</sup><http://www.statmt.org/wmt16>

simplified error taxonomy was used. This, however, contradicts to some extent the low IAA scores of WMT evaluation where no error labels are used at all.

To sum up, low IAA scores are an important issue that impacts MT research. In order to reach higher IAA, we need to understand the process of manual evaluation deeper. Despite various attempts, at the moment there is no clearly defined and exhaustive explanation for this phenomenon.

In this work, we make preparatory steps towards analysing the problem of low IAA by paying special attention to cognitive processing of the sentences by the annotators. To indirectly observe how annotators process sentences and make decisions on their quality, we use eye-tracking techniques.

## 2. Goal of this Pilot

In this pilot study, we run a small sample of the WMT manual ranking for MT systems that took part in the WMT13 shared task (Bojar et al., 2013b) in English-to-Czech translation<sup>2</sup>. Our primary objectives were:

- to implement and test the technical means for eye tracking of WMT rankings,
- to observe the accuracy of eye tracking and the range of possible recording problems before running an experiment at a larger scale,
- to collect preliminary observations about time spent on various parts of the screen, attention span and strategies of judging,
- to formulate specific questions that should be answered in a larger study.

## 3. Related Work

Eye tracking has been successfully applied to different areas of MT research, including translation memory matches (O’Brien, 2006), post-editing (Green et al., 2013; Vieira, 2014) and others. It allows to look at the behaviour of human translators, editors and annotators from a white box perspective, indirectly observing the cognitive effort they need for the task and the areas of text that attracts more attention. Eye tracking has the advantage of being relatively cheap and effortless since participants do not require any additional training. This, together with the reliability of results, makes the use of eye tracking very attractive for research on human processing of MT data.

One of the important applications of eye tracking for MT research is analysis of MT errors. Intuitively, errors differ in severity. The more serious the error, the more it complicates the understanding of the text. If a chunk of a text is difficult to understand, the reader would show certain gaze patterns when reading this text. Strymne et al. (2012) showed that different types of errors had different levels of deviation in those measurements from the error-free areas. This is explained by different level of difficulty caused by different types of errors. (Hill and Keller, 2014) analysed in

<sup>2</sup><http://www.statmt.org/wmt13/translation-task.html>

depth cognitive basis of error analysis, taking into account error classifications by type and severity.

Another direction that has been well researched is automatic MT evaluation using eye-tracking data. This direction of research goes back to 1990s, with Fuji (1999) who measured “informativeness”, “comprehensiveness” and “fluency” of MT texts via eye-tracking data and proposed metrics based on gaze patterns as an alternative to traditional evaluation metrics. More recently, Doherty et al. (2010) explored correlations between the quality of MT output sentences and eye-tracking data collected from reading those sentences. The findings show that gaze time and fixation counts have medium to strong correlation with the manually evaluated quality of translated texts. Average fixation times and pupil dilations did not show correlation. The results suggest that further research on the use of eye tracking for semi-automatic evaluation of the quality of MT has a high potential.

Doherty and O’Brien (2014) and Klerke et al. (2015) used the rationale that a translation is good if it is easy to read and (at the same time) it successfully serves its purpose (e.g. a translated user manual is helpful to solve a software problem). The studies conclude that, while being comprehensible enough to be useful in a real-world scenario, outputs of modern MT systems require manual post-editing to reach native level of comprehensiveness.

Overall, eye-tracking data has been successfully used to shed light on complex questions in MT from the cognitive perspective. We therefore hope it can also help in understanding the cognitive processes behind the traditional WMT style of evaluation.

## 4. Experimental Setup

For this pilot study, we asked our colleagues and students. In total, we got 8 volunteers (4 male, 4 female), five of which have taken part in WMT manual ranking before, so they were familiar with the task and screen layout. Six of the annotators were linguists or had some background in linguistics. All but two annotators had normal or corrected-to-normal vision (wearing glasses or contact lenses). The two problematic ones agreed to run the task without glasses they normally wear, because we were not able to calibrate the eye tracker with the glasses on. They confirmed they can read the text well, the exercise was only more strenuous for them.

### 4.1. Screen Layout

The WMT ranking is traditionally performed in a web-based user interface. Since 2013, WMT relies on an updated version of Appraise (Federmann, 2012) to present the annotation screens and collect judgements.

We simplified the Appraise HTML+CSS layout by removing light gray labels for the various versions of the sentence and by extending the vertical and horizontal space to increase the reliability of the recorded gaze positions. We also added a simple JavaScript code that reported coordinates of all HTML elements of interest. We collected the bounding boxes of all letters and punctuation symbols in the sentences, and of all the buttons of the user interface



Figure 1: Our simplified Appraise evaluation screen with bounding boxes of letters, buttons and also some larger areas highlighted. There is the source and reference at the top, followed by five candidates. The “Rank 1” till “Rank 5” orange buttons are operational, the green and red “Best” and “Worst” boxes are labels only. Surprisingly, some of the reported coordinates are systematically misplaced, e.g. the buttons and esp. the blockquote element around each candidate translation. For the analysis, we relied primarily on the letters, wrapping all letters in a text snippet into a rectangular area of interest.

(rank buttons and the submit button; we did not allow the annotators to skip sentences), see Figure 1.

For the purposes of eye tracking, screen presentation had to be tightly coupled with the control of the eye tracker. Instead of running a web browser during the session, we rendered the annotation screens in FireFox beforehand, extracted the relevant area from standard screenshots and emulated the function of radio buttons by overlaying pictures in Matlab using Psychtoolbox extension (Brainard, 1997).

## 4.2. Eye-Tracker Setup

Eye movements were recorded using EyeLink II eyetracker with frame rate 250 Hz. Each annotator viewed the monitor (17” CRT) 50 cm from the screen<sup>3</sup> and their heads were restrained using chinrest.

For the sake of reliable measurements, we calibrated and validated the eye-tracker after every screen. This proved useful also because it allowed us to provide annotators with a break every few screens as needed.

All the annotators were presented with the exact same set of annotation screens in the same order. Depending on time availability and tiredness, they were allowed to exit the session after any screen. One annotator managed to annotate as few as 8 screens in the hour allocated for him (due to severe problems with calibration and the fact that the task was new to him), others completed 16, 27 or even 32 screens in less than one hour.

<sup>3</sup>For one of the annotators who was not wearing his glasses, we reduced the viewing distance to 42 cm.

## 4.3. Interpreting the Recordings

The work of one annotator on one screen constitutes a trial. Because the trials were long from the eye-tracking point of view (75.32 s per trial on average), it was common that annotators blinked. Blinks were projected into the eye data as rapid decrease in pupil size measured by eye tracker. We removed samples from the data where pupil size decreased below 0.7 of average pupil size in the trial. We also removed 30 ms of eye tracking data before and after the samples identified as blinks.

When we projected the recorded gaze trajectories onto the annotation screens, it was apparent that the calibration was slightly distorted in many cases. The distortions were different for each screen and person. Sometimes, the distortions were non-linear (not a simple translation or skew), with a large portion of the area calibrated well but e.g. a corner running away. At the same time, it was usually very clear from the trajectories which text the annotator was reading at a particular time. We therefore abandoned the idea of interpreting the recordings at the level of words and resorted to larger areas of interest.<sup>4</sup> To improve the reliability of the results, we decided to manually adjust the areas of interest for each annotator and screen individually. Original and adjusted areas are illustrated in Figure 4.

## 5. Observations

This section summarizes our first observations of the data.

### 5.1. Incomparable Types of Errors

First of all, it is reasonable to suggest that incomparable candidate sentences would cause annotator disagreement. In the long term, we would like to separate disagreement due to incomparable candidates from disagreement due to lack of attention, but the current recordings seem too scarce for such an analysis. Identifying incomparable candidates (be it from the final rankings or from eye-tracking data) could be useful beyond simple system evaluation. For instance, when deciding which MT system to use for a particular purpose, it may be useful to measure the extrinsic performance on incomparable sentences, because that would help to identify the qualities needed for the task.

We propose the following simple approximation of sentence pair incomparability given multiple annotations. We list all pairs of candidate translations and note how often the pair was annotated  $<$  (indicating that the first system was preferred),  $=$  or  $>$ . Then we consider the distribution of the three options  $<$ ,  $=$ ,  $>$  and calculate the entropy. When a pair is unanimously ranked as e.g.  $<$  then the entropy will be zero and when the annotators are not sure, the entropy will be high.

Looking at sentence pairs with a high entropy, we observed multiple times from the data that this happens in candidate pairs where fluency is high but the adequacy suffers and vice versa. If the first candidate is ungrammatical but it would become perfect when fixing some word-level errors (such as some morphological agreement) and the second

<sup>4</sup>We would like to come back to the analysis at the level of words and characters, because at least a dozen of screens seems to be recorded accurately enough.

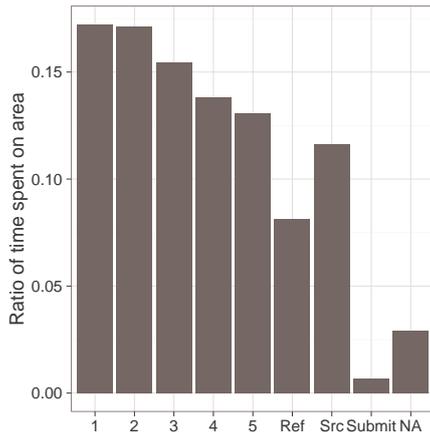


Figure 2: The ratio of time, which annotators spent on different areas. NA represents samples where eye gaze was out of the areas.

candidate is perfectly grammatical but the meaning is distorted, then the agreement on ranking would be typically low.

One option for future research is to collect eye tracking data large and precise enough to see if each annotator has a stable capacity of noticing various error types and a stable error hierarchy when comparing such candidates.

## 5.2. Less Attention Lower on the Screen

We quantified the allocation of attention between candidates as number of eye gaze samples falling inside each individual box. As can be seen on Figure 2, time spent on each candidate decreased as sentences were positioned lower at the screen. Participants looked most of the time at one of the boxes, only 2.3% of eye gaze data fell outside our defined boxes. This observation is verified by expressing the relationship using Pearson correlation coefficient ( $r = -0.32$ ) and this relationship was stable for all-but-one annotator ( $r = -0.51 - -0.25$ ).

With increasing order of trials, time needed per one screen decreased ( $r = -0.39$ ) and again, with exception of the same participant as above, this pattern was stable ( $r = -0.79 - -0.02$ ).

One possible explanation for a lower agreement (e.g. as approximated by our entropy of labelling) is that candidate appearing lower on the screen are ranked less reliably due to decrease in attention.

We observed from our data that this is not true. Candidate pairs when one of the candidates is at the bottom of the screen do not have any lower entropy. More generally, we took the Pearson correlation coefficient between the entropy of labelling described above and the sum of candidate positions (so if the two compared sentences are the fourth and fifth on the screen, this would be 9), and the entropy described above. The correlation is as low as 0.02.

It is also not true that candidates further apart from each other on the screen (i.e. taking the difference of positions) would have lower entropy.

Pattern	% Trials	Pattern	% Trials
S	81.2	SR123	34.1
R	6.8	SRSR1	7.4
1	5.7	SR121	6.2
2	2.3	SRS12	5.1
5	2.3	SRSRS	4.0
SR	73.9	SR1R1	2.3
S1	7.4	SR1S1	2.3
R1	5.7	S1234	2.3
1S	2.3	SR12S	2.3
SR1	53.4	SR1234	25.0
SRS	18.2	SRSR12	5.7
S12	4.5	SRS123	5.1
R12	4.0	SR1212	4.0
SR12	44.9	SR1232	4.0
SRSR	11.4	SRSRS1	2.8
SRS1	6.2	S12345	2.3
SR1R	4.0	SR123R	2.3
S123	2.8	SR1S12	2.3
SR1S	2.8	SR1213	2.3
		SR12345	19.9
		SRSR123	4.0
		SR12123	4.0
		SRS1234	3.4
		SR1S123	2.3
		SRSRS12	2.3

Table 1: A summary of frequent beginnings of gaze trajectories. Patterns over 2% listed.

We suggest to make a similar analysis on order in which sentences are being read.

## 5.3. Ranking Strategies

Given the detailed eye-tracking data we have, it is interesting to learn if people follow some common annotation strategy or a pattern on the screen. Many possible ways of extracting patterns from the gaze trajectory are possible, and it would be also interesting to relate the eye tracker data to the choices made and adjusted by clicking the rank buttons. For now, we limit our observations to common *beginnings* of the trajectories (removing short noise).

Table 1 summarizes the percentage of trials that the annotator started by looking into a particular area (S for Source, R for Reference, digits for individual candidate translations; disregarding too short observations). The majority of paths start in the source, but about 7% start in the reference and 6% in the first candidate.

Extending the observation to the first two areas, reading first the source and then the reference is the most common pattern (74%), followed by considering the source and then immediately the first candidate (7%) or the reference and the first candidate (6%).

Further in the table, we see that a quick comparison of the source and reference is (SRSR) is also frequent (11%).

Finally, about 20% of trials follow the most natural sequence SR12345. Other patterns are also common, e.g. comparing pairs of candidates (...121... etc.). It would be very interesting to look at such pairwise comparisons and sentence similarity or agreement in pairwise ranking.

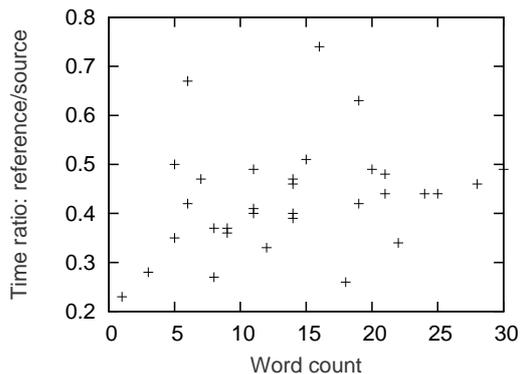


Figure 3: The correlation of source word count and the ratio of time spent in the Reference vs. the Source (Pearson 0.272).

#### 5.4. Source or Reference?

All our annotators were native Czechs with good English skills, so evaluating English-to-Czech MT was an appropriate task for them. Still, it was clear that reading the English source is more demanding for them. On the other hand, the references are not always very literal and the annotators were aware of the fact that the systems would be better judged based on the source than on the reference. (It is even possible, that the reference would contain some additional information, either from the context, or because Czech was the source text in some cases). A natural question to ask is whether the annotators put more attention to the English source or to the Czech reference.

To quantify this, we consider the time spent gazing in the area of the reference vs. the time spent on the source. We take a simple ratio of these figures for every screen and annotator and average the ratio over screens. Annotators do not differ much in this respect, their Reference/Source ratio is  $0.41 \pm 0.06$  which means they spend more time (59%) in the source.

Figure 3 puts the Reference/Source ratio in context with source sentence length (word count before tokenization). We see a slight preference for relying on the Reference as the sentences get longer.

This result confirms our expectation: longer sentences are more demanding to process, so the annotators resort to the reference more often.

#### 5.5. Parts of Sentences Skipped

One concern we may have is whether the annotators read the whole sentences. An example of a portion of sentence skipped can be seen in Figure 4: the fifth candidate is read only to about one half and then the judge realizes that the beginning was already bad enough to make the judgement. The candidate gets rank 4, the worst value assigned by the annotator in this screen.

We checked the whole set of 163 screens (i.e.  $5 \cdot 163 = 815$  candidates to read) and skipping such a large portion was a very rare exception. One more annotator skipped the tail of the very same sentence and we have seen one more sim-



Figure 4: An example gaze trajectory. We see rather good accuracy at the five candidate translations, a skew in the area of the source sentence (top left) and a drift by about one line in the area of the reference (top right). The second half of the last candidate was not read by the judge at all. The blue dashed rectangles indicate areas of interest derived from the exact rendering of the page, the black rectangles are our manually adjusted areas to compensate for calibration errors.

ilarly bad sentence evaluated while skipping its last three words. In about 19 cases (2.3%), the annotator skipped the last word or two, but the exact same sentence end appeared higher on the screen, so the annotator has probably compared the text visually. Once, the two words skipped at the end of the sentence were not exactly identical but swapped. And once, one word at a sentence beginning was skipped, but it was visually identical to the beginning of another sentence. In sum, our annotators were very careful in reading all the sentences so no unreliability can be attributed to skipping the text.

## 6. Discussion

In this pilot eye tracking of WMT manual ranking, we learned that the accuracy of the technology is very limited. In the majority of trials, so far we couldn't get the gaze position precisely enough to identify words focused, let alone their parts (like morphemes indicating declination). The difficulties with calibration depend on many factors, primarily related to the subject (glasses, tiredness, acceptable tightness of the head band) but also to surrounding light conditions. Even with the exceptionally well recorded trials, the gaze usually stops before the middle of the last word, so character-level diagnostics is not plausible unless the sentences are constructed in some special way.

On the other hand, the areas of attention can be easily identified (although with a necessary manual correction) and with detailed timing information, we believe very interesting insights can be obtained.

To overcome the limited accuracy, we consider replacing the eye-tracker with a special presentation mode where the whole screen except a small area around the mouse pointer would be blurred.

## 7. Future work

The small dataset we created can be analyzed in many further ways, for instance, it would be interesting to check whether WMT annotators exhibit the same reactions to different error types as observed by Stymne et al. (2012). This need not be the case, since WMT annotators *expect* errors and search for them in a way.

We suggest to separately investigate the influence of attention span on agreement. It is known that human attention span is as limited as 8 seconds (Watson, 2015), which is not sufficient to carefully read and compare all 5 target sentences, as offered by WMT evaluation interface. Certain observations have been made by the research community that support this hypothesis, see e.g. Vela and van Genabith (2015) mentioned above.

We also suggest to explore the issue of accumulating tiredness during the exercise. The degradation of people's attention during the hourlong recording could influence the results.

## 8. Conclusion

We conducted a pilot study on using eye-tracking technologies to look into the process of WMT manual evaluation. Our approach allows to implicitly show the strategies used by annotators, and issues they struggled with. Analysis of such data can help to understand the issue of annotator disagreement, which is an important problem in manual annotation.

Our preliminary findings show that annotators follow different strategies when working with the data. It remains for future analysis to check if each particular person prefers a small subset of the strategies, or if the strategies vary during the session. We plan to relate the observed strategies to the choices made and adjusted by the annotator, and also to the inter-annotator agreement.

We are very hopeful that further explorations will shed light on the phenomenon of annotator disagreement and help improve the reliability of manual MT evaluation.

Finally, we make our data publicly available at:

<http://hdl.handle.net/11234/1-1679>

## Acknowledgement

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 645452 (QT21). This work was partially financially supported by the Government of Russian Federation, Grant 074-U01.

This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

## 9. Bibliographical References

Ageeva, E., Forcada, M. L., Tyers, F. M., and Pérez-Ortiz, J. A. (2015). Evaluating machine translation for assimilation via a gap-filling task. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 137–144, Antalya, Turkey, May.

- Berka, J., Černý, M., and Bojar, O. (2011). Quiz-Based Evaluation of Machine Translation. *Prague Bulletin of Mathematical Linguistics*, 95:77–86, March.
- Bojar, O., Ercegovčević, M., Popel, M., and Zaidan, O. F. (2011). A grain of salt for the wmt manual evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11. Association for Computational Linguistics.
- Bojar, O., Macháček, M., Tamchyna, A., and Zeman, D. (2013a). Scratching the Surface of Possible Translations. In *Proc. of TSD 2013*, Lecture Notes in Artificial Intelligence, Berlin / Heidelberg. Západočeská univerzita v Plzni, Springer Verlag.
- Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013b). Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria, August. Association for Computational Linguistics.
- Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, pages 433–436.
- Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, June. Association for Computational Linguistics.
- Doherty, S. and O'Brien, S. (2014). Assessing the usability of raw machine translated output: A user-centered study using eye tracking. *International Journal of Human-Computer Interaction*, 30(1):40–51.
- Doherty, S., O'Brien, S., and Carl, M. (2010). Eye tracking as an mt evaluation technique. *Machine translation*, 24(1):1–13.
- Dreyer, M. and Marcu, D. (2012). HyTER: Meaning-Equivalent Semantics for Translation Evaluation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 162–171, Montréal, Canada, June. Association for Computational Linguistics.
- Federmann, C. (2012). Appraise: an Open-Source Toolkit for Manual Evaluation of MT Output. *The Prague Bulletin of Mathematical Linguistics*, 98:25–35.
- Fishel, M., Bojar, O., and Popović, M. (2012). Terra: a Collection of Translation Error-Annotated Corpora. In *Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC'12)*, pages 7–14, Istanbul, Turkey, May. ELRA, European Language Resources Association.
- Fuji, M. (1999). Evaluation experiment for reading comprehension of machine translation outputs. In *Proceedings of MT Summit VII*, pages 285–289.
- Graham, Y. (2015). Improving evaluation of machine translation quality estimation.
- Green, S., Heer, J., and Manning, C. D. (2013). The efficacy of human post-editing for language translation. In

- Proceedings of the SIGCHI conference on human factors in computing systems*, pages 439–448. ACM.
- Hill, R. L. and Keller, F. (2014). Human detection of translation errors in text: unwrapping the dynamic process through eye-tracking. In *Translation in Transition: between cognition, computing and technology*.
- Klerke, S., Castilho, S., Barrett, M., and Sogaard, A. (2015). Reading metrics for estimating task efficiency with MT output. In *Proc. of EMNLP*, page 6.
- Koehn, P. and Monz, C. (2006). Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City, June. Association for Computational Linguistics.
- Koehn, P. (2007). Euromatrix–machine translation for all european languages. *Invited Talk at MT Summit XI*, pages 10–14.
- Lommel, A., Popovic, M., and Burchardt, A. (2014a). Assessing inter-annotator agreement for translation error annotation. In *MTE: Workshop on Automatic and Manual Metrics for Operational Translation Evaluation*.
- Lommel, A. R., Burchardt, A., and Uszkoreit, H. (2014b). Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumatica: tecnologies de la traduccio*, 0(12):455–463, 12.
- O’Brien, S. (2006). Eye-tracking and translation memory matches. *Perspectives: Studies in translatology*, 14(3):185–205.
- Specia, L. and Farzindar, A. (2010). Estimating machine translation post-editing effort with hter. In *Proceedings of the Second Joint EM+/CNGL Workshop Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEC 10)*, pages 33–41.
- Stanojević, M., Kamran, A., Koehn, P., and Bojar, O. (2015). Results of the WMT15 Metrics Shared Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisboa, Portugal, September. Association for Computational Linguistics.
- Stymne, S. and Ahrenberg, L. (2012). On the practice of error analysis for machine translation evaluation. In *LREC*, pages 1785–1790.
- Stymne, S., Danielsson, H., Bremin, S., Hu, H., Karlsson, J., Lillkull, A. P., and Wester, M. (2012). Eye tracking as a tool for machine translation error analysis. In *LREC*, pages 1121–1126.
- Vela, M. and van Genabith, J. (2015). Reassessing the wmt2013 human evaluation with professional translators trainees. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 161–168.
- Vieira, L. N. (2014). Indices of cognitive effort in machine translation post-editing. *Machine Translation*, 28(3-4):187–216.
- Watson, L. (2015). Humans have shorter attention span than goldfish, thanks to smartphones. *The Telegraph*, 15.

## Ten Years of WMT Evaluation Campaigns: Lessons Learnt

Ondřej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, Lucia Specia

Charles University in Prague, Microsoft Research, University of Edinburgh, University of Edinburgh/ JHU,  
JHU, University of Sheffield

bojar@ufal.mff.cuni.cz, chrife@microsoft.com, bhaddow@inf.ed.ac.uk, phi@jhu.edu  
post@cs.jhu.edu, l.specia@sheffield.ac.uk

### Abstract

The WMT evaluation campaign (<http://www.statmt.org/wmt16>) has been run annually since 2006. It is a collection of shared tasks related to machine translation, in which researchers compare their techniques against those of others in the field. The longest running task in the campaign is the translation task, where participants translate a common test set with their MT systems. In addition to the translation task, we have also included shared tasks on evaluation: both on automatic metrics (since 2008), which compare the reference to the MT system output, and on quality estimation (since 2012), where system output is evaluated without a reference. An important component of WMT has always been the manual evaluation, wherein human annotators are used to produce the official ranking of the systems in each translation task. This reflects the belief of the WMT organizers that human judgement should be the ultimate arbiter of MT quality. Over the years, we have experimented with different methods of improving the reliability, efficiency and discriminatory power of these judgements. In this paper we report on our experiences in running this evaluation campaign, the current state of the art in MT evaluation (both human and automatic), and our plans for future editions of WMT.

**Keywords:** Machine Translation, Evaluation, Shared Tasks

### 1. Introduction

The First Workshop in Statistical Machine Translation was held in 2006, and it has been held annually since then, becoming the First WMT Conference in Machine Translation (WMT 2016) this year. In the first year of WMT there was a shared translation task which attracted 12 task description papers. In 2015 there were 5 different tasks and 46 task description papers, whilst in 2016 there will be 10 different tasks, covering translation of text and images, handling of pronouns in translation, MT evaluation, system tuning, automatic post-editing and document alignment.

The core component of WMT has been the main translation task (which in most years is the only translation task). The first translation task used Europarl (Koehn, 2005) for the test set; since then, we have constructed the test set from news text, with the complex structure and broad topic coverage providing a significant challenge to MT systems. Since 2009 the news test sets have been created specifically for the shared task, by crawling news articles in various languages and translating to the other task languages, providing the MT research community with valuable resources for future research. We have also varied the language pairs from year to year to present different challenges to researchers, although there has always been an emphasis on European languages. The language pairs included in each year's evaluation are shown in Table 1.

A central theme in the WMT shared tasks has been the evaluation of MT. We have explored this extensively, focusing on both human and automatic evaluation. The main translation task has always employed large-scale human evaluation to determine the quality and ranking of the systems; how precisely this is done has varied over the years (Section 2.). The human ranking has enabled the development of automatic metrics by providing a gold standard against which metrics can be compared. Since 2008, the metrics task has asked participants to develop tools to evaluate MT output against one or more references (Section 3.). In 2012, we introduced the quality estimation task, which takes met-

rics a step further, attempting to evaluate the quality of MT output without use of a reference (Section 4.).

### 2. Manual Evaluation

Since the very beginning, WMT organizers have taken the position that machine translation performance should be evaluated from time to time against human opinion:

*While automatic measures are an invaluable tool for the day-to-day development of machine translation systems, they are only a imperfect substitute for human assessment of translation quality*  
... (Koehn and Monz, 2006)

This is not to disparage automatic metrics, which have played a crucial role in the progress of the field and the improvement of MT quality over time. It is only to say that they are at best a proxy for what we really care about, and must be regularly anchored to human opinion. The WMT therefore produces an annual *human ranking of systems* for each task, from best to worst. In addition to helping direct researchers to the systems whose features they might wish to copy, this gold-standard system ranking is used to evaluate automatic metrics (a metric metric).

Of course, the question of which system is the best or worst is a fraught one. There are any number of answers, and subsequent questions. The first is best *for what purpose?* For a person trying to understand a foreign-language news article, an MT system that can convey the gist of an article is necessary, but quality might need to be sacrificed for speed. On the other hand, a student trying to learn how to translate an article may require a system that can also correctly generate grammatical and natural-sounding sentences. Evaluations are often broken down along these concepts of *adequacy* and *fluency*.

In fact, in the first two editions of the WMT shared translation task we used adequacy/fluency judgements on a 5-point scale as our main evaluation measure. Not satisfied with the results though, we started experimenting with

Language Pair	'06	'07	'08	'09	'10	'11	'12	'13	'14	'15	'16
Czech ↔ English		•	•	•	•	•	•	•	•	•	•
Finnish ↔ English										•	•
French ↔ English	•	•	•	•	•	•	•	•	•	•	
German ↔ English	•	•	•	•	•	•	•	•	•	•	•
German ↔ Spanish			•								
Haitian Creole → English						•					
Hindi ↔ English									•		
Hungarian ↔ English			•	•							
Romanian ↔ English											•
Russian ↔ English								•	•	•	•
Spanish ↔ English	•	•	•	•	•	•	•	•			
Turkish ↔ English											•

Table 1: Language pairs in the main translation task.

Metric	'06	'07	'08	'09	'10	'11	'12	'13	'14	'15	'16
Adequacy / Fluency	•	•									
Sentence Ranking		•	•	•	•	•	•	•	•	•	•
Constituent Ranking		•	•								
Constituent Judgement (Y/N)			•								
Sentence Comprehension				•	•			○			
Direct Assessment											•
Used MTurk					•		•	•			•

Table 2: Metrics used in the human evaluation over the years for all languages pair (•) or only English → Czech (○).

other methods and over the years, WMT has tried several different ones, encoded in different evaluations, summarized in Table 2. Brief explanations of the approaches follow:

- *Fluency / Adequacy*. Annotators were presented with a sentence, and were asked to rank it separately for both fluency and adequacy, on five-point scales.
- *Sentence Ranking*. Annotators are presented with the outputs of multiple systems, along with the source and reference sentence, and asked to rank them, from best to worst.
- *Constituent Ranking*. Annotators were asked to rank the quality of the translations of automatically-identified constituents, instead of the complete sentences.
- *Constituent Judgement (Y/N)*. Annotators were asked to provide a binary judgement on the suitability of the translation of a constituent.
- *Sentence Comprehension*. Annotators were asked to edit MT output for fluency (without providing the reference), and then (separately) to determine via binary judgement whether those edits resulted in good translations.
- *Direct Assessment (DA)*. Annotators are asked to provide a direct assessment of the quality of a single MT output compared to a single reference, using an analog scale.

The adequacy/fluency judgements were abandoned as the 5-point measurements proved to be quite inconsistent and

hard to normalize, and they were not popular with the annotators. Viewing the distributions of scores provided by individual annotators showed them to be very different in shape, often skewed in different directions, so there was no clear way to combine judgements from multiple annotators. There was also complaints from annotators about the extreme difficulty in annotating long sentences of, frequently scrambled, MT output.

Two early measures of quality focused only on noun phrase constituents that were automatically identified in the reference and then extracted from system outputs via projections across automatic alignments. Constituent ranking (2007–2008) asked annotators to compare and rank these constituents, while binary constituent judgements (2008) asked them only whether a constituent (provided in context and approximately highlighted) were “acceptable” compared to the reference. An advantage of these binary judgements was very high annotator agreement rates; this is likely due in part to their relatively short length.

Another means of directly assessing output quality (and thereby inferring a system ranking) is Sentence Comprehension, used in 2009 and 2010. In this task, one set of judges was asked to edit a sentence’s fluency (without access to the source or reference); these edited sentences were then later evaluated to see whether they “represent[ed] fully fluent and meaning-equivalent alternatives to the reference sentence”. This mode of evaluation did not correlate well with relative ranking, however, and was abandoned in 2011 in order to focus annotators’ efforts on that method.

In an effort to find a better evaluation method, we introduced Sentence Ranking in 2007. One big advantage of Sentence Ranking is that it is conceptually very simple: of-

fer the annotator two samples of MT output (and a reference) and ask them which they prefer. In practice, in order to gather judgements more efficiently, we present the annotator with 5 different MT outputs at a time, which then yields ten pairwise comparisons. We have experimented with presenting more or fewer sentences at a time, but 5 seems to be a good compromise between efficiency and reliability. We have also experimented with collecting judgements on Amazon’s Mechanical Turk (2012 and 2013), in an effort to reduce the effort required from researchers. While relatively effective, the effort required to ensure that the work was completed faithfully, and the even lower annotator agreement rates, caused us to abandon it.

Since 2011, Sentence Ranking has been the only method of human evaluation we have used, but during that time the details have evolved in response to criticism. In particular, Bojar et al. (2011) pointed out various problems with the way the comparisons were collected and interpreted which led to changes in the procedure. A particular problem with Sentence Ranking is that the method involves collecting *relative* judgements of MT performance, but attempts to combine these to give an *absolute* measure of translation performance. Unless a sufficient number of carefully chosen comparisons are made, then systems can be treated unfairly by being compared too often to a very bad, or very good system (or the reference, which may be in there for control). Furthermore, systems were getting credit for ties, so systems which were very similar to others were doing better than they should. Finally, Bojar et al. (2011) showed that the agreement on the Sentence Ranking task falls off rapidly as sentence length increased.

Further analysis of the Sentence Ranking approach was provided by Lopez (2012) who pointed out the difficulties in obtaining a reliable total ordering of systems from the pairwise judgements. Further work (Koehn, 2012) suggested that we really needed to collect more judgements in order to display significant differences between the systems, and also established a means of clustering systems into equivalence classes of mutually indistinguishable systems, based on bootstrap resampling. Thus, since 2013, the system rankings have been presented as a partial ordering over systems, instead of a total ordering, where systems in the same group are considered to be tied. (However, the total ordering is still used when evaluating metrics).

One important point has not been addressed. Over the years, WMT has experimented with many different means of producing a system ranking. These rankings are then used as a gold standard for metrics tasks, and are also published as an official ranking, which researchers make use of in determining which system description papers to plumb for ideas to improve their own systems. Each year, different methods have been evaluated and then kept or discarded according to a number of criteria, such as annotator agreement numbers, or time spent. However, how can we really know which of these is the best? This point was raised by Hopkins and May (2013), who then provided a Bayesian model formulation of the human ranking problem, which allowed them to use perplexity to compare different system rankings. Influenced by this idea, in 2014, we compared the ability of three different models trained on a large set of

pairwise rankings, using accuracy on held-out comparisons instead of perplexity. The method that won was a new approach that based on the TrueSkill algorithm (Sakaguchi et al., 2014). This has been in use since.

To conclude, the WMT manual evaluation has engaged in a deep and extensive experimentation over the years. The Sentence Ranking task has formed the core of our evaluation approach, and has seen many variations from year to year. We have made progress on many of the problems with evaluation. However, many problems remain: the relatively low annotator agreement rates, the immense amount of annotator time required, and the difficulty of scaling the sentence ranking task to many systems. In 2016, we plan to run a pilot investigation based on Direct Assessment of machine translation quality, which we hope will further alleviate some of these issues.

### 3. Automatic Evaluation

Since the second year of the WMT campaigns, targeted effort was also devoted to evaluation of automatic metrics<sup>1</sup> of MT quality, or **metrics task** for short. This meta-evaluation is an important complement to the shared translation task, because automatic metrics are used throughout the development of MT systems and also in automatic system optimization (Neubig and Watanabe, 2016). The utility of some of the metrics in system optimization has been tested in the sister **tuning task** in 2011 and 2015 and also planned for 2016.

Metrics of MT quality are evaluated at two levels:

**System-level** evaluation tests, how well a metric can replicate the human judgement about the overall quality of MT systems on the given complete set of test set sentences.

**Segment-level** evaluation tests how well a metric can predict the human judgement for each input sentence.

In both cases, participants of the metrics task are given input sentences, outputs of MT systems and one reference translation. Note that the reliance on a single reference is not ideal. It is well known that the reliability of automatic MT evaluation methods is limited if only one reference is available (see the WMT 2013 overview paper for an empirical evaluation of BLEU with up to 12 references for translation into Czech). The quality estimation task (Section 4.) focuses on the setup where no reference is available at all. Table 3 summarizes the participation and methods used to evaluate the system-level and segment-level parts of the task. The task had always received a good number of participating teams. The number of evaluated metrics varies considerably across the years, because in some years, multiple variations of some metrics were evaluated. Starting from 2013, we distinguish “baseline metrics”. These metrics are run by the organizer in addition to the submitted ones. Baseline metrics include the `mteval` scoring script and all the metrics available in Moses. We report the exact configuration flags for them, so they should be reliably reproducible.

Throughout the years, the metrics task has always relied on the manual evaluation (Section 2.), so the gold standard

<sup>1</sup>Despite the term “metrics”, none of the measures or methods is a metric in the mathematical sense.

	'07	'08	'09	'10	'11	'12	'13	'14	'15	'16
Participating Teams	-	6	8	14	9	8	12	12	11	
Evaluated Metrics	11	16	38	26	21	12	16	23	46	
Baseline Metrics							5	6	7	
System-level evaluation methods										
Spearman Rank Correlation	•	•	•	•	•	•	•	◦		
Pearson Correlation Coefficient							◦	•	•	•
Segment-level evaluation methods										
Ratio of Concordant Pairs		•	•							
Kendall's $\tau$				•	•	•	*	*	*	*
Tuning Task					•				•	•

• main and ◦ secondary score reported for the system-level evaluation.

•, \* and \* are slightly different variants regarding ties.

Table 3: Summary of metrics tasks over the years.

human judgements do come from different styles of evaluation. A major move from Sentence Ranking to Direct Assessment is considered in 2016, which would particularly affect the segment-level metric evaluation. In Direct Assessment, the judgements have to be sampled differently from the system-level and segment-level evaluation, and there is a concern whether we will be able to find enough distinct speakers for each of the language pairs. Preliminary experiments are now under way.

### 3.1. How Metrics are Evaluated

As indicated in Table 3, the metrics task has seen a few changes of the exact evaluation method.

**Evaluating System-Level Evaluation** System-level methods were first evaluated using Spearman rank correlation, comparing the list of systems for a particular language pair as ordered by the metric (given the test set of sentences are reference translations) and as ordered by humans (on the sample of sentences from the test set that actually receive some human judgements). Spearman rank correlation was selected in the first year, because it is applicable also to the ordinal scales of adequacy and fluency which were used in 2006 and 2007. Since 2007, Pearson correlation coefficient could have been also used (as the system scores were on continuous scales), but the switch happened only in 2013. The benefit of Pearson over Spearman is that it considers the distances between the systems, so it should be more stable for systems of similar quality.

**Evaluating Segment-Level Evaluation** Segment-level evaluation has so far relied on pairwise judgements of translation quality. Given two candidate translations of an input sentence, the segment-level metric gets a credit if it agrees with the human judgement, i.e. the two pairwise judgements are “concordant”. The exact calculation of the final score changed throughout the years: in 2008 and 2009, a simple ratio ranging from 0 to 1 was used: the number of concordant pairs out of the total number of pairs evaluated. Starting from 2010, the score was modified to penalize discordant pairs, falling under the general definition of Kendall rank correlation coefficient, or Kendall's  $\tau$  for short, with  $[-1, 1]$  as the range of possible values:

$$\tau = \frac{|Concordant| - |Discordant|}{|Concordant| + |Discordant|} \quad (1)$$

There has always been a question of how to handle tied comparisons, either the humans or the metric (or both) assigning the same rank/score to the two candidates. Each type of tied pairs can be included in the denominator and if it is, it may be also included in the numerator (bonified or penalized). After the discussion available in Macháček and Bojar (2013) and Macháček and Bojar (2014), the current method:

- ignores pairs where humans tied altogether,
- does not give any credit or bonus to pairs where the metric predicted a tie,
- but includes these metric-tied pairs in the denominator.

Moving to the Direct Assessment or some other absolute scale in the human evaluation would allow use to use Pearson correlation coefficient instead of Kendall's  $\tau$ .

**Significance** From the beginning, it was not quite clear how to establish significance of the observed differences in metric evaluation, especially at the system level where the number of participating systems is less than 20, providing a low sample size.

Starting from 2013, system-level scores for each given language pair were reported with empirical confidence bounds constructed by resampling the “golden truth”: given the complete set of human judgements, 1000 variations are constructed by resampling with repetition, leading to 1000 different scorings of the systems.<sup>2</sup> Each participating metric provides a single scoring of the systems and this scoring is correlated with the 1000 golden truths, giving us 1000 results reflecting the variance due to the set of sentences and annotators included in the golden truth.

As noticed by Graham and Liu (2016), confidence intervals obtained from this sampling cannot be used to infer whether one metric significantly outperforms another one, because the number of “significant” pairs would be overestimated. Instead, Graham and Liu (2016) proposes a novel method, artificially generating a large number of MT systems (by

<sup>2</sup>Many of these scorings share the same order of the systems. Unlike Spearman rank correlation, the Pearson correlation coefficient used since 2013 however appreciates also differences in the scores.

mixing the outputs of the real MT systems participating in the translation task) and asking metrics task participants to score e.g. not 5 but 10000 MT systems on the given test set. We will try to adopt this approach in 2016, testing in practice, how many metrics task participants can cope with these enlarged sets of MT systems.

### 3.2. Observations in Metrics Task

While metrics tasks across the years cannot be directly compared because a whole range of conditions keeps changing, the overall setting remains stable and some general observations can be made:

- BLEU has been surpassed by far by many diverse metrics. On the other hand, we acknowledge that it remains the most widely used and also scores on average well among the baseline metrics, with CDER (Leusch and Ney, 2008) being a competitor.
- The level of 0.9 of system-level correlation into English was reached by the best metrics in 2009, rising up to 0.98 in 2011. These levels were achieved by **aggregate** or **combination metrics** that include many features and standard metrics; sometimes the combination is **trained** on a past dataset. IQmt-ULCh, SVMrank (2010) and MTeRater-Plus (2011) are the early examples, followed by a row of other combination metrics in recent years (e.g. BEER, DPMFcomb, RATATOUILLE in 2014 or 2015). MTeRater is an interesting outlier in that its main component is based on many features from automatic essay scoring (preposition choice, collocations typical for native use, inflection errors, article errors).
- Benefits were confirmed many times from **including paraphrases or synonyms** incl. Wordnet (e.g. Meteor, Tesla in 2010 and 2011), refining the metric to consider the coverage of individual **parts of speech** (e.g. PosBLEU 2008, SemPOS 2009, 2012), focusing on **content words** (Tesla, SemPOS), **dependency relations** (already 2008) or **semantic roles** (already 2007), evaluating at the level of **character sequences** (i-letter-BLEU 2010, chrF 2015, BEER).
- In 2012, we saw a drop in into-English evaluation mainly due to a different set of participating metrics. Such a **“loss of wisdom”** is unfortunate and the baseline metrics run by the organizers are one of possible means to avoiding it. In an ideal world, the authors of the top performing metrics every year would incorporate their metrics to Moses, to ensure that the metric gets evaluated in the coming years. Achieving this state is obviously complicated by the reliance of some of the metrics on diverse language-dependent resources which are not always publicly available. Meteor remains the only such maintained metric throughout the years. Hopefully, some of the trivial but well-performing metrics based on characters (chrF, i-letter-BLEU) will join the baselines soon.

## 4. Quality Estimation

Quality Estimation (QE) offers an alternative way of assessing translation quality. QE metrics are fully automated and, unlike common evaluation metrics (Section 3.), do not rely

on comparisons against human translations. QE metrics aim to provide predictions on translation quality for MT systems in use, for any number of unseen translations. They are trained metrics, built using supervised machine learning algorithms with examples of translations labelled for quality (ideally, by humans). Predictions can be provided at different granularity levels: word, phrase, sentence, paragraph or document. Different levels require different features, label types and algorithms to build prediction models.

While work on QE started back in the early 2000’s (Blatz et al., 2004), the use of MT was substantially less widespread back then, and thus the need for this type of metric was less evident. A new surge of interest appeared later (Specia et al., 2009; Soricut and Echihiabi, 2010), particularly motivated by the popularisation of MT in commercial settings.

QE was first organised as a shared task (and a track at WMT) in 2012 (Callison-Burch et al., 2012). The main goals were to provide a baseline approach, devise evaluation metrics, benchmark existing approaches (features and algorithms), and establish the state-of-the-art performance in the area. The task focused on quality prediction at sentence level. Only one dataset was provided, for a single language pair (English-Spanish), on the News domain, translated by one MT system. For training and evaluation, translations were manually annotated by professional translators for quality in terms of “perceived” post-editing effort (1-5 scores). A system to extract baseline QE features and resources to extract additional features were also provided. The baseline system used a Support Vector Machine regression algorithm trained on the features provided. This was found to be a strong baseline (both features and algorithm) and has been used in all subsequent editions of the task.

As we continued running the task in subsequent years (Bojar et al., 2013; Bojar et al., 2014; Bojar et al., 2015), our main goals have been to provide, each year, new subtasks (while keeping the popular ones), additional language pairs, and larger and more reliably labelled datasets. For most subtasks, the evaluation metrics have also been redefined over the years. Table 4 summarises the main components of the shared task over the years.

More specifically, we introduced variants of post-editing effort prediction – edit distance (a.k.a. HTER) and post-editing time – for sentence level (2013), and other subtasks at new granularity levels: (i) a system selection subtask to learn how to rank alternative MTs for the same source sentence, precisely the same goal as the metrics task (Section 3.), but without reference translations (2013); (ii) a word-level subtask concerned with predicting a binary (good/bad) or 3-way (keep, delete, replace) tag for each word in a target sentence (2013), as well as more fine-grained error categories annotated by humans (omission, word order, word form, etc., in 2014); (iii) a paragraph-level subtask to predict a Meteor score for an entire paragraph (2015); (iv) a document-level subtask to predict a task-based human-targeted score for the entire document (2016); and (v) a phrase-level subtask, where binary labels (good/bad) are to be predicted for entire “phrases”, as segmented by the MT system (2016). Baseline systems and resources were provided for all these subtasks.

The main language pair has remained English-Spanish

(en→es), the only constant language over all editions for the sentence and word-level subtasks. This was mostly due to the availability of (labelled) data for this pair. However, other language pairs have been explored over the years for most subtasks. English-German (en→de) was used on various occasions, including all subtasks in 2014 and the paragraph-level subtask in 2015. German-English (de→en) was also used in the latter subtask, in all subtask in 2014, and in the MT system selection task in 2013.

The sizes of the datasets varies over the years. A good indicator is the sentence-level subtask. The figures in the last row of Table 4 refer to the largest number of sentences for any score prediction subtask in a given year.

The number of participating teams has remained considerably stable over the years (10–14), but teams tend to submit systems for various subtasks, as well as for the same subtask when multiple languages are available. The submission figures in Table 4 include only submissions for different subtasks and language pairs.

The evaluation of participating systems varies across subtasks. For sentence, paragraph and document levels, systems can be submitted for two variants of each task: scoring (for various labels, e.g. 1-5, 1-3, HTER, time, Meteor) and ranking, where only a relative ranking of test instances is required. Scoring is evaluated using standard error metrics (e.g. Mean Absolute Error) against the true scores and, since 2015, using Pearson’s correlation. Ranking is evaluated using Spearman’s correlation, as well as a ranking metric proposed for the task in 2012: DeltaAvg, which compares the ranking of instances given by the system against the human ranking for different quality quantiles of the test set. For the word and phrase-level tasks, per-class precision, recall and F-measure metrics are computed, with F-measure for the “bad” class used as main metric in the binary variant.

Overall, the shared tasks have led to many findings and highlighted various open problems in the field of QE. Here we summarise the most important ones:

- **Training data:** The size of the training data is important for all prediction levels, but is even more critical for word and phrase levels. For sentence level, it does not seem to be the case that having more than 2K sentences makes a significant difference in performance. The quality of the data has proved a more important concern. The dataset used for the sentence and word level subtasks in 2015, for example, although large, was of questionable quality (spurious or missing post-editings) and had a very skewed label distribution, which made model learning harder.
- **Algorithms:** There is no consensus on the best algorithm for each subtask. Various popular regression algorithms have ranked best for sentence (and paragraph) level in different years, including SVM, Multilayer Perceptron, and Gaussian Process. For word (and phrase) level, sequence labelling algorithms such as Conditional Random Fields perform best.
- **Tuning:** Feature selection and hyperparameter optimisation proved essential. The winning submissions

in most years performed careful (or even exhaustive) search for both features and hyperparameter values.

- **Features:** While a range of features has been used over the years, shallow, often language-independent features, tend to contribute the most. The majority of submissions built on the set of baseline features provided. Recently, word embeddings and other neural inspired features have been successfully explored. While features for sentence and word/phrase-level prediction are clearly very distinct from one another, for paragraph level, most systems used virtually sentence level features. We hope that more interesting discourse features will be exploited in 2016 given the much longer documents provided as instances. A critically important feature for all levels is the *pseudo-reference* score, i.e., comparisons between the MT system output and a translation produced by another MT system for the same input sentence.
- **Labels:** Prediction of objective scores, such as post-editing distance and time, has led to better models (in terms of improvements over the baseline system and correlation with human scores) than prediction of subjective scores such as 1-5 labels. Post-editing time seems to be the most effective label. However, given the natural variance across post-editors, this is only the case when data is collected by and a model is built for a single post-editor.
- **Granularity:** The word-level subtask has proved much more challenging than the sentence-level one, often obtaining very marginal improvements over naive baselines. In the tasks we have run so far, this could have been due to: little training data, limited number of examples of words with errors (class unbalance), and potentially noisy automatic word labelling. We attempted to solve some of these limitations by providing data annotated manually for errors (2014), but for cost reasons the largest dataset we could collect has just over 2K segments. A larger dataset (14K segments) was collected based on post-editions in 2015, but the post-editing, and hence the labelling generated from it, are of questionable quality. In 2016, we are providing an even larger dataset (15K segments) post-edited by professional translators. The new phrase-level subtask in 2016 should also help overcome some of the limitations of the word-level one, by providing more natural ways in which to segment the text for errors. The paragraph-level subtask in 2015 did not attract much attention, perhaps due to the use of an automatic metric as quality label (Meteor). In 2016 we provide actual (much longer) documents labelled by humans.
- **Progress over time:** As with any other shared task, measuring progress over time is a challenge since we have new datasets (and often new training sets) every year. Progress in the QE task can however be speculated in relative terms, more specifically, with respect to the improvement of submitted systems over

	'12	'13	'14	'15	'16
Participating Teams	11	14	10	10	-
Evaluated QE Systems	20	55	57	34	-
Subtasks	1	4			
Sentence Level	•	•	•	•	•
Word Level		•	•	•	•
Paragraph Level				•	
Document Level					•
Phrase Level					•
Language Pairs	en→es	en→es, de→en	en↔de, en↔es	en→es, en↔de	en→es
Largest Dataset (snt)	2,254	2,754	4,416	14,088	15,000

Table 4: Details on different editions of the QE task over the years.

the baseline system. This is possible for the sentence-level subtask, since the language pair and baseline system have remained constant over the years. We have observed, year after year, that more systems are able to beat the baseline, and by a larger margin.

## 5. Plans for Future Editions

In recent years, we have used Sentence Ranking as the sole method of automatic evaluation (refining it according to certain criticisms (Bojar et al., 2011; Lopez, 2012; Koehn, 2012)), but ongoing problems with reliability, interpretability and poor scalability with increasing numbers of systems have driven the search for alternatives. In 2016, we will pilot a new technique for manual evaluation of MT output. This is based on recent work demonstrating an effective means for collecting adequacy and fluency judgements using crowd-sourcing (Graham et al., 2016). This *Direct Assessment* of machine translation quality is similar to our early attempts to judge quality with adequacy and fluency judgements (Koehn and Monz, 2006; Callison-Burch et al., 2007), but improves upon it in critical ways. Crucially, an analog scale is presented to the user in the form of a slider bar, which underneath maps to a 100-point scale, instead of the 5-point Lickert scale we used in the past, which gave us inconsistent results that were difficult to interpret. Annotators are required to do large batches of assessments in a single sitting, which allows their scores to be normalized more reliably. By embedding deformed outputs and comparing their scores to those of their uncorrupted counterpart, inconsistent, unreliable, and untrustworthy annotators can be identified, and their outputs discarded.

The potential advantages of Direct Assessment are:

- It offers good reliability, as measured by inter-annotator agreement;
- the cost of assessment scales linearly in the number of systems assessed (instead of quadratically, as with Sentence Ranking);
- it provides absolute measures which can be compared year-over-year; and
- the concepts of adequacy and fluency are readily interpretable, in a way that the scores derived from Sentence Ranking are not.

Sentence Ranking will remain our primary evaluation for this year, but the results of this evaluation will be compared to those of the DA evaluation in order to help assess its

suitability for future evaluations.

One of the big issues we face in MT evaluation is the question of *for what purpose?* In other words, the way we evaluate our MT system may depend quite strongly on what we want to use it for, whether for gisting, post-editing, direct publication, language learning, automated information extraction, or something else. The Sentence Ranking method is particularly weak in this regard, since we do not give the raters any guidance as to how they should judge the translations. In some sense, we have punted on the difficult question of purpose, allowing each annotator to be guided by his or her own intuitions. This likely explains some of the low annotator agreement rates. Using adequacy and fluency separately is an improvement as the terms have meaningful interpretation, although they are still intrinsic rather than extrinsic measures. In the end, we believe that the work of the WMT manual evaluation has improved our knowledge for how to assess human quality of MT, providing a rich well from which to draw for those wishing to focus on more targeted and specific applications.

For QE, after the 2016 edition we will have covered all possible granularity levels. The plan is to keep the most popular and the most challenging ones, with a particular emphasis on word and phrase-level prediction. Instead of more language pairs, we will prioritise larger and better datasets for fewer language pairs. Another direction we aim to pursue is better integration with other WMT evaluation tasks, e.g. using the test sets and system translations from the translation task, and reusing the manual evaluations as training data. In the past this has proved difficult logistically because of the tasks' timeframe or unsuccessful because the manual evaluations (esp. rankings) were not adequate for QE. The planned changes in the manual evaluation procedure should make this integration possible.

## 6. Conclusions

The WMT shared tasks have given us a platform to explore all forms of Machine Translation (MT) evaluation; human evaluation, automatic evaluation with a reference, and quality estimation. Not only that, but WMT has helped to drive research in MT evaluation, firstly by having high profile shared tasks to engage the community; and secondly by the extensive data sets that we provide. Each year, we prepare new translation test sets, and annotated data sets for quality estimation. During the tasks, we collect and release all

translation system submissions, all the human judgements, all the submissions to metrics, and all the quality estimation data. These are made available from the WMT website (for this year it is [www.statmt.org/wmt16](http://www.statmt.org/wmt16)) and are used frequently in subsequent research.

MT evaluation is a hard problem, and is capable of generating significant controversy in the MT community, as we have observed when evaluation results were presented. This difficulty is indicated by the number of changes, experiments, and refinements we have introduced over the years. This year, with the piloting of Direct Assessment, we return to a direct measure of the quality of a system output that we abandoned a number of years ago, and are hopeful that the reformulation of the problem will make DA more successful than our earlier experiments. If so, one option for the QE task in subsequent years is for it to model the prediction of DA scores.

### Acknowledgements

This work received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 645442 (QT21) and 645357 (Cracker).

- Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. (2004). Confidence Estimation for Machine Translation. In *Proc. of the 20th International Conference on Computational Linguistics*, pages 315–321, Geneva, Switzerland.
- Bojar, O., Ercegović, M., Popel, M., and Zaidan, O. (2011). A Grain of Salt for the WMT Manual Evaluation. In *Proc. of the Sixth Workshop on Statistical Machine Translation*, pages 1–11, Edinburgh, Scotland.
- Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In *Proc. of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria.
- Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 Workshop on Statistical Machine Translation. In *Proc. of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA.
- Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. (2015). Findings of the 2015 Workshop on Statistical Machine Translation. In *Proc. of the Tenth Workshop on Statistical Machine Translation*, pages 1–46.
- Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). (Meta-) Evaluation of Machine Translation. In *Proc. of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic.
- Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2012). Findings of the 2012 Workshop on Statistical Machine Translation. In *Proc. of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada.
- Graham, Y. and Liu, Q. (2016). Achieving Accurate Conclusions in Evaluation of Automatic Machine Translation Metrics. In *Proc. of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (to appear)*, San Diego, CA.
- Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2016). Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, FirstView:1–28, 1.
- Hopkins, M. and May, J. (2013). Models of Translation Competitions. In *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1416–1424, Sofia, Bulgaria.
- Koehn, P. and Monz, C. (2006). Manual and Automatic Evaluation of Machine Translation between European Languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City.
- Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proc. of MT Summit*.
- Koehn, P. (2012). Simulating Human Judgment in Machine Translation Evaluation Campaigns. In *Proc. of IWSLT*, pages 179–184.
- Leusch, G. and Ney, H. (2008). BLEUSP, INVWER, CDER: Three improved MT evaluation measures. In *NIST Metrics for Machine Translation Challenge*, Waikiki, Honolulu, Hawaii, October.
- Lopez, A. (2012). Putting Human Assessments of Machine Translation Systems in Order. In *Proc. of the Seventh Workshop on Statistical Machine Translation*, pages 1–9, Montréal, Canada.
- Macháček, M. and Bojar, O. (2014). Results of the WMT14 Metrics Shared Task. In *Proc. of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, Baltimore, MD, USA.
- Macháček, M. and Bojar, O. (2013). Results of the WMT13 Metrics Shared Task. In *Proc. of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria.
- Neubig, G. and Watanabe, T. (2016). Optimization for Statistical Machine Translation: A Survey. *Computational Linguistics*, To appear.
- Sakaguchi, K., Post, M., and Van Durme, B. (2014). Efficient elicitation of annotations for human evaluation of machine translation. In *Proc. of the Ninth Workshop on Statistical Machine Translation*, pages 1–11, Baltimore, Maryland, USA.
- Soricut, R. and Echihiabi, A. (2010). TrustRank: Inducing Trust in Automatic Translations via Ranking. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 612–621, Uppsala, Sweden.
- Specia, L., Turchi, M., Cancedda, N., Dymetman, M., and Cristianini, N. (2009). Estimating the Sentence-Level Quality of Machine Translation Systems. In *Proc. of the 13th Conference of the European Association for Machine Translation*, pages 28–37, Barcelona, Spain.

# Towards a Systematic and Human-Informed Paradigm for High-Quality Machine Translation

Aljoscha Burchardt, Kim Harris, Georg Rehm, Hans Uszkoreit

DFKI GmbH

Alt-Moabit 91c, Berlin, Germany

{firstname.lastname}@dfki.de

## Abstract

Since the advent of modern statistical machine translation (SMT), much progress in system performance has been achieved that went hand-in-hand with ever more sophisticated mathematical models and methods. Numerous small improvements have been reported whose lasting effects are hard to judge, especially when they are combined with other newly proposed modifications of the basic models. Often the measured enhancements are hardly visible with the naked eye and two performance advances of the same measured magnitude are difficult to compare in their qualitative effects. We sense a strong need for a paradigm in MT research and development (R&D), that pays more attention to the subject matter, i.e., translation, and that analytically concentrates on the many different challenges for quality translation. The approach we propose utilizes the knowledge and experience of professional translators throughout the entire R&D cycle. It focuses on empirically confirmed quality barriers with the help of standardised error metrics that are supported by a system of interoperable methods and tools and are shared by research and translation business.

**Keywords:** Machine Translation, Platforms, Human Evaluation

## 1. Introduction

Since the advent of modern statistical machine translation (SMT), much progress in system performance has been achieved that went hand-in-hand with ever more sophisticated mathematical models and methods. Numerous small improvements have been reported whose lasting effects are hard to judge, especially when they are combined with other newly proposed modifications of the basic models. Often the measured enhancements are hardly visible with the naked eye and two performance advances of the same measured magnitude are difficult to compare in their qualitative effects. On the other hand, most of the fundamental known barriers to MT quality have not yet overcome.

We sense a strong need for a paradigm in MT research and development, that pays more attention to the subject matter, i.e., translation, and that analytically concentrates on the many different challenges for quality translation. The approach we propose utilizes the knowledge and experience of professional translators throughout the entire R&D cycle. It focuses on empirically confirmed quality barriers with the help of a standardised parameterisable error metric. The metric is supported by a system of methods and tools and shared by research and translation business. These components, which have already been created and tested, are seen as core components of an envisaged cloud-based platform, which will be sketched out in the last part of the paper.

The remainder of this paper explains these ideas in more detail.

## 2. Human-Informed MT Development Cycle

The prevalent (S)MT development cycle consists of a number of experiments in which system parameters, feature sets, preprocessing steps, etc. are more or less systematically varied followed by a testing phase (“generate-and-test”).

The power of SMT lies in its massive utilization of human translation expertise. In rule-based systems only those parts

of human knowledge are used that could be encoded in the dictionaries and rule sets of the system usually a mix of intellectually compiled explicitly stated linguistic regularities and exceptions.

Statistical methods acquire implicit human knowledge about translation and linguistic well-formedness by learning huge numbers of patterns from texts, especially from translated texts in connection with their source texts. In this way they can model semantic and stylistic preferences and constraints that could not be encoded in any of the hand-crafted rule systems.

Within the testing phase human knowledge is again being used in a rather indirect and implicit way, i.e., by comparing the output of the MT engine with one or more human reference translations using simple mathematical measures such as BLEU.

If one has the goal of working towards High-Quality Machine Translation (HQMT), this approach is scientifically questionable at best, for a number of reasons including:

- It is widely known that simple automatic measures such as BLEU correlate only mildly with translation quality. If we rely on them, only optimising our systems towards BLEU scores exclusively, we run the risk of reporting spurious improvements, under- or overestimating system variants, oscillating on plateaus, etc.
- It has been shown that the highest BLEU improvements are often made on segments that are unintelligible anyway, i.e., completely unintelligible translations get a little less unintelligible, but, nevertheless, they remain unintelligible (but the BLEU score is improved). This approach does not contribute to the goal of working towards high-quality translation.
- BLEU relies critically on one or more reference translations used for the comparison. We have performed an internal study using a Chinese → English reference corpus comprising 11 documents (1000 sentences),

each translated by a different human expert to evaluate an online MT system. The results are startling: the choice of *one* reference document led to a variation in BLEU scores of up to 7.64 points, depending on which reference was chosen (average BLEU: 18.11). Using all 11 reference translations together led to a BLEU score of 53.42.

- While higher BLEU scores *indicate* improved translation quality, they cannot be taken as *scientific evidence*.
- Single scores do not provide many (scientific) insights. Tuning and optimisation steps are usually epiphenomenal. They are not suitable to generalise results, to apply them to new translation tasks, or to make predictions.
- BLEU scores do not detect errors, nor do they provide any information on the type or source of errors.

All researchers should be eager to analyse the results of their experiments as thoroughly as possible in order to compare them to the work of others and to be in the best possible position to generate hypotheses for improving the approach and to drive future experiments. There are two classes of quality indicators: (i) Translation errors and (ii) cases where a generated correct translation is better or worse than another possible correct translation. Whereas (ii) is rather important for human translation in the highest quality segment, for today’s MT only errors matter.<sup>1</sup> So far we do not have any reliable ways for automatically detecting errors and their error types. Thus we are convinced that any serious attempt to improve translation quality must include feedback by human experts such as translators and linguists early in the development process. This feedback can be given, e. g., in the form of post-edited (i.e. corrected) translations or explicit annotation of errors using standardised markup. In the language industry, both approaches are established best practice for assessing (machine) translation quality.

If the MT research community wants to produce research results that are supposed to be meaningful to the language industry, we have to extend our approaches, systems and paradigms in such a way as to be able to assess and report translation quality in the required way (in addition to what is needed in the SMT R&D cycle, i.e., automatic scores like BLEU). Figure 1 shows how the current SMT development cycle can be extended to include human feedback. The blue box represents the typical existing SMT development cycle. In addition, we propose to include at certain intervals or checkpoints a language expert who inspects the translation results, annotates and classifies errors and provides feedback to the MT developer who then starts another development cycle based on the insights gained. At some point, the language expert will most probably compare newly generated translations to previous output to see if the intended improvements have materialised, to check if

<sup>1</sup>BLEU score measurement also punishes correct translations if they differ from the reference translations, may they be better or worse.

there have been any unintended side-effects, and to spot the most pressing quality barriers.

This proposed approach makes it necessary to reserve a budget for human language professionals in MT projects (and to make sure that the human analysis and annotation process is optimally supported by tools), but we are convinced that this investment will pay off. The data gathered from human analysis and annotation should be used to build linguistically informed methods for quality estimation and error detection to eventually support (semi-)automatic analytic workflows.

### 3. High-Quality Translation Paradigm

The use of MT is increasingly popular for ‘gisting’ (information-only translation) through free online systems such as Google Translate or Bing Translator. These services have created huge new markets for translation. Already back in 2012, Google alone automatically translated as much content in a single day as all professional translators combined in an entire year, and was used by more than 200 million people every month<sup>2</sup> – by now the usage figures are surely much higher.

Still, all popular, freely available online translation services follow the “one size fits all” approach, i.e., they are not customisable. This approach is inherently incompatible with Europe’s pressing demand for being able to produce large volumes of high-quality outbound translations either fully automatically or through human translators supported by machines. In this high-quality scenario, MT has to behave much more like Translation Memories (TMs) that are widely used in translation industry, especially when dealing with repetitive material such as technical documentation. TMs support translators by suggesting perfect or almost perfect translations based on previous translations. The translator can then accept and edit the suggestion or translate from scratch.

Already in past publications such as, for example, (Rehm and Uszkoreit, 2013; Burchardt et al., 2014; Popović et al., 2014), we have made the point of breaking out of the dead-end the MT research landscape is currently trapped in by advocating a paradigm shift. Instead of only adjusting known SMT algorithms and features to produce marginally better results, we call for a different approach of carrying out MT research in Europe, an approach that addresses the goal of producing quality translations and that takes into account very thoroughly the needs and priorities of European MT and Language Service Provider (LSP) companies, thus initiating a close collaboration for creating new breakthroughs in research and business opportunities at the same time.

A trivial, yet far-reaching insight is that not all translations are equally useful for human translators. For simplicity, they are often divided into the three discrete classes of (i) error-free translations, (ii) translations that can efficiently be post-edited and (iii) translations that are so bad that they would not help a human translator. While class (iii) might

<sup>2</sup><http://googleblog.blogspot.de/2012/04/breaking-down-language-barriersix-years.html>

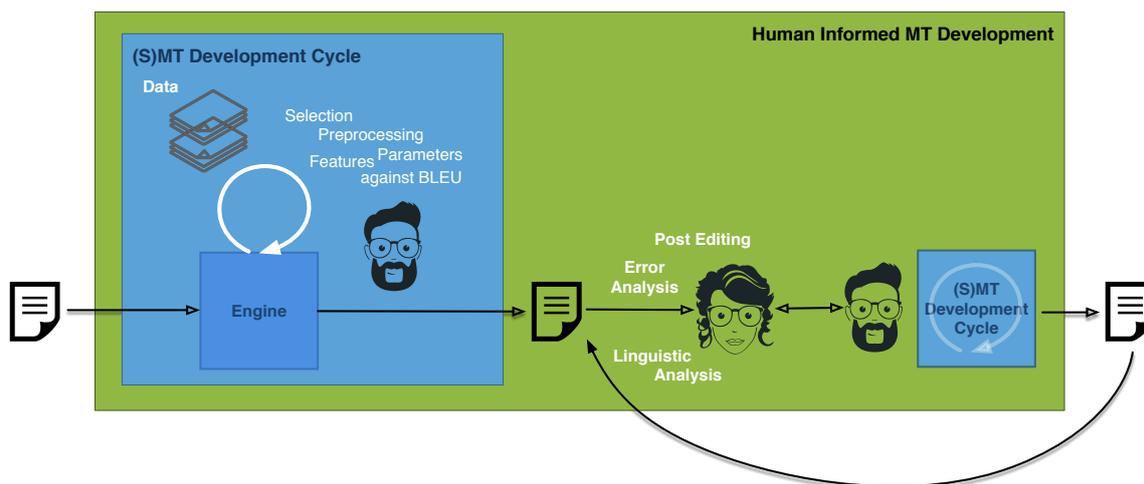


Figure 1: Human-informed MT development cycle

still provide some guidance to people in an information search scenario (gisting), they do not play a role in the quality translation scenario.

As a consequence, *improvement* in this paradigm, say from system variant  $A$  to variant  $A'$ , requires that the proportion of the quality classes changes so that we have more translations of the “better” classes in the end. Figure 2 provides one such example where there has been an increase in the number of perfect, error-free translations. The precise criteria like error types, severity classes, scoring model, etc. need to be worked out between research and industry taking into account task-specific factors such as the language pair, document type, domain, target audience, etc.

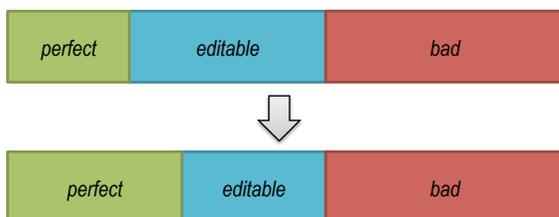


Figure 2: Example of improvement in high-quality translation paradigm

This focus on HQMT requires new diagnostic methods. We provide some suggestions in the next section.

#### 4. Standardised Error Metrics and Benchmarks

HQMT relies on improved translation models that must be based on novel, reliable and informative quality measures. Simplistic common measures such as BLEU or edit-distance based measures such as TER may even incorrectly punish perfectly adequate translations which differ from a given reference (or references), e. g., in completely legitimate word order and morphological realisation. Currently, the only way of assessing translation quality with an adequate level of reliability and granularity (word/phrase level)

necessarily involves intellectual work such as post-editing or explicit error annotation.

In the new type of MT development, annotations can be added on the following three levels as needed:

**Phenomenological level** Annotation of issues in the translated output (target side) with translation errors such as, e. g., Omission, Terminology, or Grammar.

**Linguistic level** Annotation of the translation source or target side with information like part of speech, phrase boundaries or more specific phenomena under consideration such as long-distance dependencies or multi-word expressions.

**Explanatory level** Annotation of the source (also referencing the target) with (typically speculative) reasons for translation failure such as model class,  $n$ -gram size, data sparseness, etc.

The annotation on the phenomenological level usually involves language professionals like human translators while the other two levels require linguistic skills and expertise on the MT system level that researchers from linguistics, language technology, and related areas typically have.

**Standardised error markup with MQM** While the notions “error markup” and “issue markup” are often used interchangeably, there is an important difference that we only briefly sketch in this article. There is no transcendent, absolute notion of translation quality. Thus, an issue such as an inconsistency in terminology, for example, referring to an object as “PC” in one sentence and as “computer” in another, might be counted as an error, e. g. in a reference manual, but it can be perfectly acceptable, maybe even preferred, in a newspaper article. Translation quality is always relative to the intended communicative purpose and context that can best be captured in a formal specification. The Multidimensional Quality Metrics (MQM) (Lommel et al., 2014) is based on this principle of flexibility to translate different purposes into dimensions and selective subsets of

issues to be checked (and weighted). MQM was designed around a master vocabulary comprising 100+ issue types for describing task-specific metrics in a highly customizable way. It provides a unified approach for (diagnostic) evaluation of MT with approaches used for human translation quality checking in industry. It was designed as a non-strict superset of prominent metrics (LISA QA Model, SAE J2450, ATA certification, etc.). An early version was standardised in the W3C recommendation ITS 2.0.

**Post-Editing** Apart from direct error annotation, one can also make use of human feedback on the basis of automatically translated output that was corrected by human translators through post-editing. This output can be used to update, reinforce, and correct systems' translation hypotheses and together with explicit error markup will help to overcome real quality barriers and also fix relatively minor issues such as punctuation or agreement errors that seem to have been over-looked in the development of MT engines for gisting, yet render most output improper for outbound translation.

**Evaluation workflow** Projects such as QTLaunchPad and QT21 have developed valuable experience for what we hope becomes best practice in future evaluation scenarios. Given some MT translated corpus and initial hypotheses of what issues may be encountered, the following steps are included in an example evaluation workflow:

1. Definition of a concrete metric for the given purpose starting from an existing metric ("benchmark") or from scratch.
2. Filtering the translation corpus to be evaluated in a triage:
  - a. Perfect translations.
  - b. Almost good translations that need further analysis.
  - c. Bad translations that do not qualify for further inspection.

These steps can be performed manually, in a semi-automatic or even in an automatic way using sampling and filtering strategies or with the help of a quality estimation toolkit such as QuEst ((Shah et al., 2013)), depending on the size of the corpus, available human resources, and required precision and recall. What follows is:

3. Annotation/Post-editing of the segments of type b.
4. Inspection of the errors/edits to:
  - Confirm if the system output supports the hypotheses
  - Get a quantitative basis to decide on MT development priorities
  - Get a qualitative idea of remaining quality barriers

Figure 3 illustrates this proposed workflow. It is advisable to perform error annotation and post-editing in parallel so that analysis and correction are handled in a consistent manner as there are usually multiple ways of analysing and fixing a translation error.

**Test Suites** Test suites are a best practice instrument in areas such as grammar checking, to ensure that a parser is able to analyse certain sentences correctly or test the parser after changes to see if it still behaves in the expected way. In the context of HQMT, we use the term "test suite" to refer to a selected set of input-output pairs that reflects interesting or difficult, error-prone cases. Test suites have not generally been used in MT research. Reasons for this might include the theoretical issue that there is no eternal notion of "good translation" and the more practical issue that there are usually many different good translations for a given input. Even if one could assume the existence some gold-standard translation, there would be no simple notion of deviation that could be used. In line with what we have argued for above, human analysis will be needed for evaluating MT performance on test suites.

Nevertheless, we think that testing system performance on empirically grounded error classes will lead to insights that can guide future research and improvements of systems. By using suitable test suites, MT developers will be able to see how their systems perform compared to scenarios that are likely to lead to failure and can take corrective action, e. g., by creating targeted training corpora focussing on certain error types. Test suites can also be the basis for new types of benchmarks and shared tasks that are based on empirically attested quality barriers; at the time of writing, we are working on a test suite for the language pair German-English, which will be published in 2016.

## 5. Integrated MT Development Platform

MT research has contributed to the development of a large set of tools required to build MT systems, training data, and evaluating corpora. These resources exist in various locations and often require substantial IT and system development skills to put them together and to make them work in an operating environment. As a result, most resources are not being reused to the extent they should be, some are not being reused at all after the end of the project in which they were created. Some of the tools that could have been useful outside the R&D community, especially to language service providers (LSPs), have appealed primarily to researchers and computer scientists rather than to language professionals and, thus, their use in and impact on the language industry has remained limited at best. Even the large volumes of valuable data accumulated over the years by the Workshop on Statistical Machine Translation (WMT) community (primarily in the projects EuroMatrix, EuroMatrix-Plus, MosesCore and CRACKER) have mostly been stored in hundreds of unconnected text files that are hard to search and to combine.

The field of leading-edge MT R&D has reached a level of complexity with its workflows, networks of people and communities, as well as resources and components involved that it is about time to discuss the pros and cons of an integrated development platform. An integrated

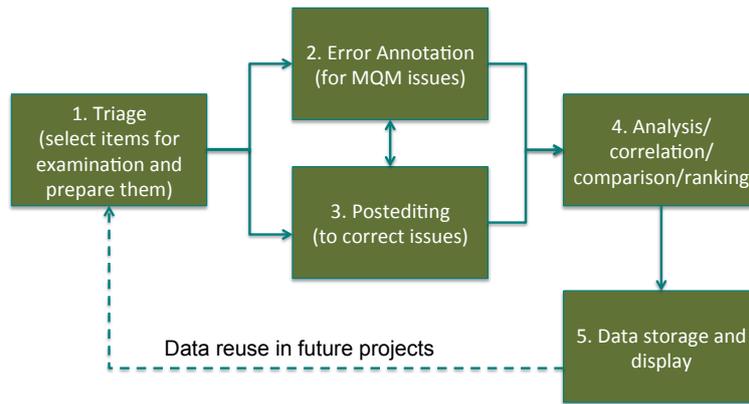


Figure 3: Post-Editing and Error Annotation Workflow Example

development environment will be even more necessary when we look at the *additional* ingredients needed for the human-informed and quality-driven MT paradigm we have sketched above, which will add to the already complexity no doubt. If designed the right way, the integrated development platform we have in mind will have positive effects on MT technology evolution by speeding up the search and evaluation cycles clustered around shared tasks and collectively approached challenges. A technology infrastructure for a major research effort in HQMT needs to serve several major purposes:

- It should help research groups to develop, test and demonstrate their new methods in realistic workflows with existing state-of-the-art components, realistic tasks, and benchmarks. It should also help them to compare the performance of their components with the best existing technology.
- It should support shared research and shared evaluations by providing the data, environments, workflows, and evaluation tools for collective system building and collective comparative assessment campaigns.
- It should preserve, document, administer and provide data, technologies and evaluations for new research groups, potential users, research planners, funding agencies and the media.
- It should provide access to state-of-the-art tools for functions and processes that may lie outside the core competency of any specific group.

To this end, the integrated platform needs to go beyond existing resources such as the open resource exchange infrastructure META-SHARE (Piperidis et al., 2014), the open source tools and core components for building SMT systems like Moses or Jane (Koehn et al., 2007; Vilar et al., 2010); tools for quality estimation and error analysis such as Qualitative (Avramidis et al., 2014), MTCompareEval (Klejch et al., 2015) or Hjerson (Popović, 2011); web platforms for selecting or training MT systems like iTrans-

late4EU<sup>3</sup> or Let’sMT<sup>4</sup>; CAT tools and workbenches like MateCat and Casmacat (Federico et al., 2014; Alabau et al., 2014); Post-Editing and error-annotation tools like PET (Aziz et al., 2012) or MT-Equal (Girardi et al., 2014); the web-based service collection of PANACEA (Poch et al., 2012), or the accumulated data, tools, and scripts of the WMT shared-task repositories. The targeted infrastructure should incorporate as many existing useful resources as possible instead of rebuilding components, data, tools and service platforms. It must enable truly collaborative research and make resources more accessible to all.

Figure 4 provides an overview of the kind of platform we envisage. In the center there is a core research infrastructure that consists of a data repository that is connected through a back-end to a front-end that takes over the function of the overall cockpit. This core infrastructure should be linked to the different services, tools, data sources, workflows, and stakeholders included in the figure in coloured boxes. Note that the content of these boxes is not fully exclusive. For now, we have left it underspecified what services can and will actually be hosted by the backend and what should be accessed via APIs.

An important part of the cockpit is data management, i. e., a data model together with data collections, DB user interfaces and data maintenance tools for the management of MT-related data collections. For now, we propose a simple data model and suggest to base a first iteration of the cockpit on the the existing open source tool translate5.

### 5.1. Data Model

For the envisaged platform, we suggest to define a very general relational model for MT-relevant data including training data, reference data, benchmark data, evaluation results and test suites. This model can be employed for designing databases for existing and new data sets of various types. It also supports user interfaces for typical viewing and editing tasks.

The proposed data model is simple and versatile. It picks up an original idea of Harris (Harris, 1988) who proposed to store bi-texts in databases whose two dimensions are the

<sup>3</sup><http://itranslate4.eu/en/>

<sup>4</sup><http://www.letsmt.eu>

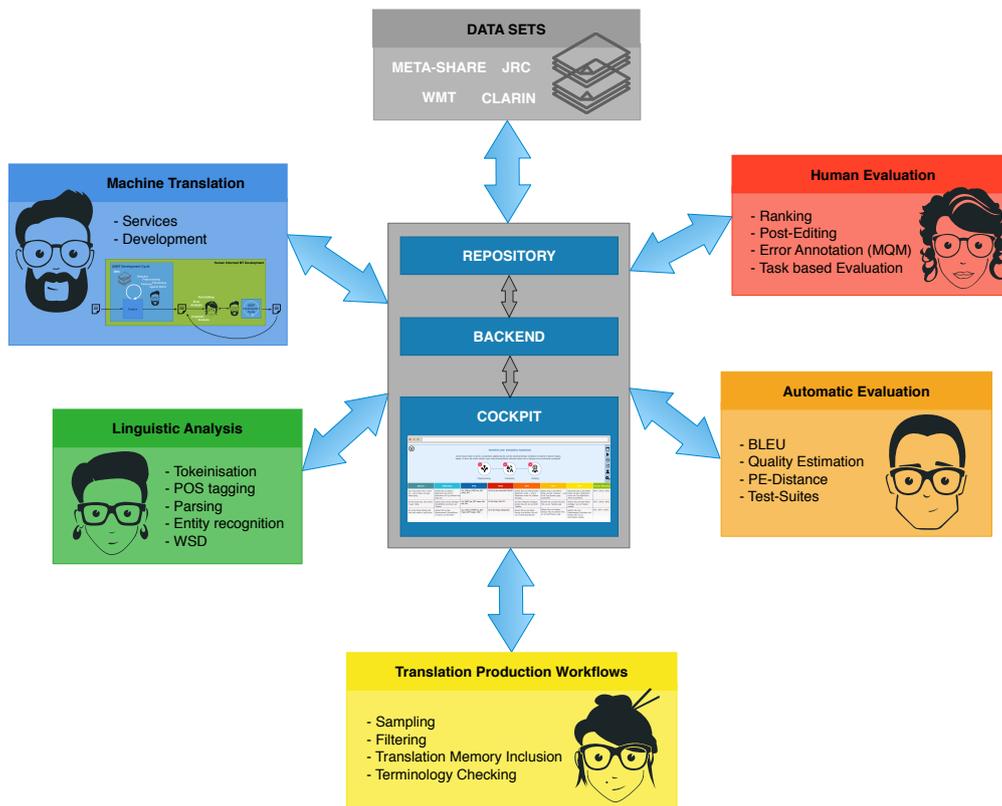


Figure 4: Integrated MT Development Platform

segments and the languages. Assume a tabular database in which every row contains all the translations and annotations of a segment. The columns are dedicated to different translations, versions and annotations of the segments. Closest to the original idea of bi-texts are columns for a translation of a text into another language. However, columns could also be used for storing edited versions of segments or texts. They can hold alternative translations by human experts or machines into the same target language. Columns can also hold comments such as assessment scores or annotations referring to the segments in another column, such as POS tagging, syntactic structures, or marked errors. They can accommodate stand-off annotation consisting of lists of mark-up tags with their respective scopes given in offset notation. However, they can also contain in-line markup applied to a copy of the raw text in another column. Finally, they can also be used to annotate relational information pertaining to two or more columns such as alignments or comparative quality rankings. A relational database management system would be needed to prevent overly complex databases. Assigned primary keys will facilitate linking and joining of the tables.

Authoring, translation, assessment and annotation can be conceptualised and realised as entering data into a new column. Pre-and post-editing could either be realised as editing data in an existing column or in a new, copied column, depending on the interest in documenting the editing

step. In keeping with the user- and human-centred research paradigm, such a user interface would be suited for translation professionals who are used to working with similar (albeit less powerful) interfaces in many computer-assisted translation (CAT) tools. Figure 5 provides a mockup of an example workflow including preprocessing, translation, and human ranking results.

Modelled workflows in translation management can be easily supported by assigning and removing read/write privileges and by appropriate reporting functions. The same is true for workflows in collective research such as multi-site system testing and in shared evaluation tasks such as the application of alternative systems to the same texts or competitive quality assessments. Such workflows can include the evaluation tasks and the realisation and testing of second-order translation systems such as combos. In addition, full versioning of data sets will ensure that users will be able to trace the complete provenance of all data; in current workflows, multiple different versions of resources may be in circulation leading to situations in which it is not clear after the fact which version of a resource was used to generate another resource.

Just as several general architectures for text analytics use layers of annotation as the output interfaces between the modules, a general architecture for MT built on our data model would use new columns to display results.

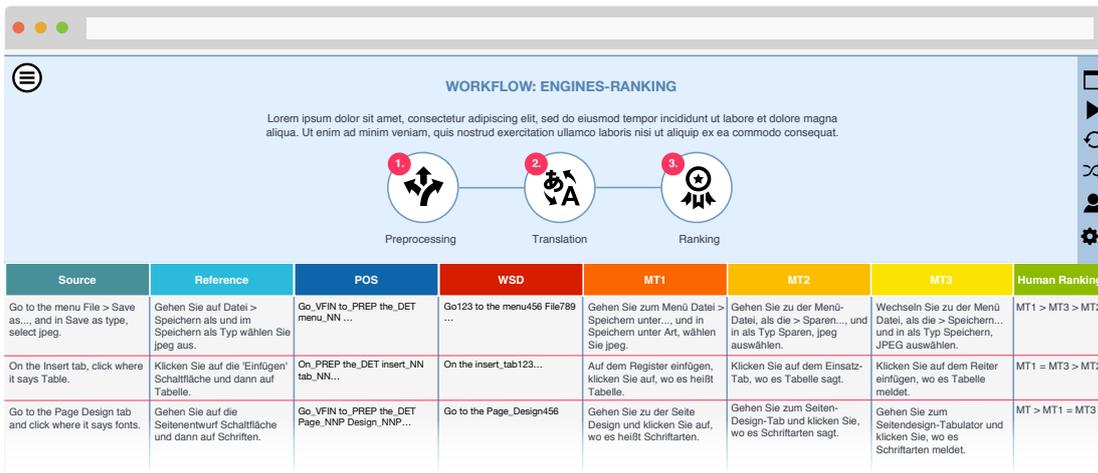


Figure 5: Mockup of the infrastructure’s cockpit

5.2. Translate5

Translate5, implemented by MittagQI is a database-driven tool with a graphical user interface that implements the column-based data model sketched above (see Figure 6 for a screenshot).

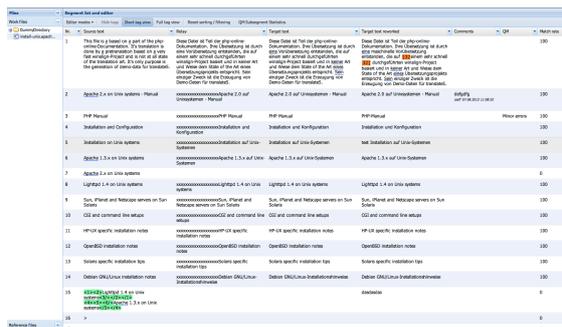


Figure 6: translate5 column view

It was originally implemented as a proofreading and post-editing environment for the translation industry. It uses a MySQL database, in which source texts, translations and annotations are stored. In the QTLaunchPad project we extended the tool to support MQM error annotation. In order to stress-test the system with large amounts of data, we imported all WMT data from 2008 to 2014 into translate5 without encountering any performance issues. Translate5 can be used for manual translation, pre- and post-editing and for quality assessment. For automatic processing steps, as well as for data import and export, a first set of APIs has been implemented as well as queuing, management and load-balancing of external and internal processes including dependency management. Simple reporting functions are already in place, others will follow soon. Translate5 features user administration, as well as simple workflow specification facilities and client management functions. Translate5 currently supports post-editing, MQM error tagging, and simple ranking. Further improve-

ments of the feature set will be provided with support of the project CRACKER.

6. Conclusion

In this paper, we have argued that research and development in Machine Translation has to make a more direct use of the knowledge of language experts such as translators and linguists. To this end, we suggest a human-informed development cycle that works on empirically confirmed quality barriers with the help of standardised error metrics and benchmarks.

As the technical foundation for a new kind of intensified collaboration between MT developers and language professionals, we outline a platform that assembles a system of methods and tools that are shared by research and the translation industry in MT R&D activities. One open source tool that could serve as the nucleus for this envisaged paradigm is translate5 that has been extended to support MQM error markup in the QTLaunchPad project and is currently further developed with support of the CRACKER project (Rehm, 2015).

In fact, some of the currently running European projects like QT21 and QTLeap are already implementing certain aspects of the emerging paradigm by including human annotation and evaluation into the MT development methodology, supported by CRACKER. Yet, we are convinced that implementing the vision put forward in this paper requires substantial support, both in terms of willingness on the side of the research community and in terms of support on the side of funding agencies and policy makers. The support of this quality-driven and analytical approach to MT development we see in industry is a step in the right direction.

Acknowledgements

This article has received support from the EC’s Horizon 2020 research and innovation programme under grant agreements no. 645452 (QT21) and no. 645357 (CRACKER). We thank the three anonymous reviewers for their valuable comments.

## 7. Bibliographical References

- Alabau, V., Buck, C., Carl, M., Casacuberta, F., García-Martínez, M., Germann, U., González-Rubio, J., Hill, R. L., Koehn, P., Leiva, L. A., Mesa-Lao, B., Ortiz-Martínez, D., Saint-Amand, H., Sanchis-Trilles, G., and Tsoukala, C. (2014). CASMACAT: A computer-assisted translation workbench. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 25–28.
- Avramidis, E., Poustka, L., and Schmeier, S. (2014). Qualitative: Open source python tool for quality estimation over multiple machine translation outputs. *The Prague Bulletin of Mathematical Linguistics*, 102:5–16, 10.
- Aziz, W., Castilho, S., and Specia, L. (2012). PET: a Tool for Post-editing and Assessing Machine Translation. In *Eighth International Conference on Language Resources and Evaluation*, pages 3982–3987, Istanbul, Turkey.
- Burchardt, A., Lommel, A., Rehm, G., Sasaki, F., van Genabith, J., and Uszkoreit, H. (2014). Language Technology Drives Quality Translation. *MultiLingual*, (143):33–39, April. Issue April/May.
- Federico, M., Bertoldi, N., Cettolo, M., Negri, M., Turchi, M., Trombetti, M., Cattelan, A., Farina, A., Lupinetti, D., Martines, A., Massidda, A., Schwenk, H., Barrault, L., Blain, F., Koehn, P., Buck, C., and Germann, U. (2014). The matecat tool. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference System Demonstrations, August 23-29, 2014, Dublin, Ireland*, pages 129–132.
- Girardi, C., Bentivogli, L., Farajian, M. A., and Federico, M. (2014). MT-EQuAl: a Toolkit for Human Assessment of Machine Translation Output. In *COLING 2014*, pages 120–123, Dublin, Ireland, August. Dublin City University and ACL.
- Harris, B. (1988). Bi-text, a new concept in translation theory. *Language Monthly*, 54:8–10.
- Klejch, O., Avramidis, E., Burchardt, A., and Popel, M. (2015). MT-ComparEval: Graphical evaluation interface for machine translation development. *The Prague Bulletin of Mathematical Linguistics*, 104:63–74.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, Richard and Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, ACL '07*, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Lommel, A. R., Burchardt, A., and Uszkoreit, H. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0(12):455–463, 12.
- Piperidis, S., Papageorgiou, H., Spurk, C., Rehm, G., Choukri, K., Hamon, O., Calzolari, N., del Gratta, R., Magnini, B., and Girardi, C. (2014). META-SHARE: One year after. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, Iceland, May.
- Poch, M., Toral, A., Hamon, O., Quochi, V., and Bel, N. (2012). Towards a user-friendly platform for building language resources based on web services. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).
- Popović, M., Avramidis, E., Burchardt, A., Hunsicker, S., Schmeier, S., Tschewinka, C., Vilar, D., and Uszkoreit, H. (2014). Involving language professionals in the evaluation of machine translation. *Journal on Language Resources and Evaluation*, 48(4):541–559.
- Popović, M. (2011). Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, 96:59–68.
- Georg Rehm et al., editors. (2013). *The META-NET Strategic Research Agenda for Multilingual Europe 2020*. Springer, Heidelberg, New York, Dordrecht, London. Buy this book at [springer.com](http://springer.com) or [amazon.de](http://amazon.de).
- Rehm, G. (2015). CRACKER: Cracking the Language Barrier. In Ilknur Durgar El-Kahlout, et al., editors, *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT 2015)*, page 223, Antalya, Turkey, May.
- Shah, K., Avramidis, E., Biçici, E., and Specia, L. (2013). QuEst – design, implementation and extensions of a framework for machine translation quality estimation. *The Prague Bulletin of Mathematical Linguistics*, 100:19–30, 9.
- Vilar, D., Stein, D., Huck, M., and Ney, H. (2010). Jane: Open source hierarchical translation, extended with re-ordering and lexicon models. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR (WMT 2010)*, pages 262–270, Uppsala, Sweden, July.

## HuQ: An English-Hungarian Corpus for Quality Estimation

Zijian Győző Yang\*, László János Laki†, Borbála Siklósi\*

\*Pázmány Péter Catholic University, Faculty of Information Technology and Bionics

†MTA-PPKE Hungarian Language Technology Research Group

Práter str. 50/A, 1083 Budapest, Hungary

{yang.zijian.gyozo, laki.laszlo, siklosi.borbala}@itk.ppke.hu

### Abstract

Quality estimation for machine translation is an important task. The standard automatic evaluation methods that use reference translations cannot perform the evaluation task well enough. These methods produce low correlation with human evaluation for English-Hungarian. Quality estimation is a new approach to solve this problem. This method is a prediction task estimating the quality of translations for which features are extracted from only the source and translated sentences. Quality estimation systems have not been implemented for Hungarian before, thus there is no such training corpus either. In this study, we created a dataset to build quality estimation models for English-Hungarian. We also did experiments to optimize the quality estimation system for Hungarian. In the optimization task we did research in the field of feature engineering and feature selection. We created optimized feature sets, which produced better results than the baseline feature set.

**Keywords:** quality estimation, machine translation

### 1. Introduction

The measurement of the quality of translation output has become necessary especially in the field of machine translation (MT). A reliable quality score for MT could save a lot of time and money for translators, companies, researchers and ordinary users. Knowing the quality of machine translated segments can accelerate the translators' work, or can help human annotators in their post-edit tasks, or can filter out and inform about unreliable translations. Last but not least, quality indicators can help MT systems to combine the translations to produce better output. There are two kinds of evaluation methods for MT. The first type uses reference translations, i.e. it compares machine translated sentences to human translated reference sentences, and measures the similarities or differences between them. To evaluate the quality of MT, after an automatic translation, we also have to create a human translated sentence (for the sentences of the test set) to compare it to the machine translated output. Creating human translations is expensive and time-consuming, thus these methods, such as BLEU (Papineni et al., 2002) and other methods based on BLEU, TER (Snover et al., 2006), HTER (Snover et al., 2006) etc., cannot evaluate in run-time, and the correlation between the results of these methods and that of human evaluation is very low in the case of translations from English to Hungarian. A completely new approach is needed to solve these problems, i.e. a method which can predict translation quality in real-time and does not need reference translations.

The other type of evaluation methods is called Quality Estimation (QE). This is a supervised approach that does not use reference translations. This method addresses the problem by evaluating the quality of machine translated segments as a prediction task. Using QE we can save considerable time and money for translators, human annotators, researchers, companies and ordinary users.

In this study, we use the QuEst framework (Specia et al., 2013), developed by Specia et al., to train and apply QE

models for Hungarian, which to our knowledge has not been done before. Hence, first, we needed to create a QE corpus for Hungarian. Then, using this corpus we built different kinds of optimized English-Hungarian QE models. For optimizing we developed new semantic features using WordNet and word embedding models.

Hungarian is an agglutinating and compounding language. There are significant differences between English and Hungarian, regarding their morphology, syntax and word order or number. Furthermore, the free order of grammatical constituents, and different word orders in noun phrases (NPs) and prepositional phrases (PPs) are also characteristics of Hungarian. Thus, features used in a QE task for English-Spanish or English-German, which produced good results, perform much worse for English-Hungarian. Hence, if we would like to use linguistic features in QuEst, we need to integrate the available Hungarian linguistic tools into it.

The structure of this paper is as follows: First we will shortly introduce the QE approach. Then, we will present the corpus we created for English-Hungarian QE. Finally, our experiments, optimizations and results in the task of QE are described.

### 2. Related Work

In the last couple of years there have been several WMT workshops with quality estimation shared tasks,<sup>1</sup> which provided datasets for QE researches. The datasets are evaluated with HTER, METEOR, ranking or post-edit effort scores. But, unfortunately, there is no dataset for Hungarian. In this research we created a QE dataset for English-Hungarian. For human judgement we used the Likert scoring scale.

QE is a prediction task, where different quality indicators are extracted from the source and the machine translated segments. The QE model is built with machine learning algorithms based on these quality indicators. Then the QE

<sup>1</sup><http://www.statmt.org/wmt15/quality-estimation-task.html>

model is used to predict the quality of unseen translations. The aim is that the scores, predicted with the QE model highly correlate with human judgments, thus the QE model is trained on human evaluations.

Recently, in the field of QE, research has focused on feature selection (Biçici, 2013) using a variety of machine learning algorithms and feature engineering (Camargo de Souza et al., 2013). In feature selection task, Beck et al. tried more than 160 features in an experiment for English-Spanish to predict HTER (Beck et al., 2013). Other key aspects in field of QE are: a.) providing larger datasets; b.) feature selection using a variety of machine learning algorithms and feature engineering for word-level, sentence-level and document level QE; c.) exploring differences between sentence-level and document-level prediction; d.) analyzing training data size and quality (Bojar et al., 2015).

According to the WMT15 (Bojar et al., 2015) shared task results, for English-Spanish the LORIA/17+LSI+MT+FILTRE (Langlois, 2015) system gained the best performance with 0.36 Spearman’s  $\rho$  and with 0.39 Pearson’s correlation. The LORIA uses the baseline features, two Latent Semantic Indexing features and 31 features based on pseudo-references. For English-German the RTM-DCU/RTM-SVR (Bicici et al., 2015) system achieved the best result with -0.62 Spearman’s  $\rho$  and with 0.59 Pearson’s correlation. The RTM-DCU is based on referential translation machines using support vector regression for document and sentence-level prediction.

In our research we did experiments for Hungarian QE in providing a dataset, word-level feature engineering and feature selection.

### 3. Quality Estimation

In the QE task, we extract different kinds of features as quality indicators from the source and translated sentences. Following the research of Specia et al., we can separate the features in different kinds of category (Specia et al., 2013). From the source sentences, complexity features can be extracted (e.g. number of tokens in the source segment). From the translated sentences, we extract fluency features (e.g. percentage of verbs in the target sentences). From the comparison between the source and the translated sentences, adequacy features are extracted (e.g. ratio of percentage of nouns in the source and target). We can also extract indicators from the MT system, these are the confidence features (e.g. features and global score of the SMT system). From another point of view, we can also divide the features into two main categories: “black-box” features (independent from the MT system) and “glass-box” features (MT system-dependent). Since in our experiments we have translations from different MT systems, we did use only the “black-box” features. After feature extraction, using these quality indicators, we can build QE models with machine learning methods. The aim is that the predictions of the QE models are highly correlated with human evaluations. Thus, the extracted quality indicators need to be trained on human judgments.

Adequacy	Fluency
1: none	1: incomprehensible
2: little meaning	2: disfluent Hungarian
3: much meaning	3: non-native Hungarian
4: most meaning	4: good Hungarian
5: all meaning	5: flawless Hungarian
0: I do not understand this English sentence	

Table 1: Adequacy and fluency scales for human evaluation

### 4. HuQ Corpus

To build the English-Hungarian QE system, we needed a training corpus. In our experiments, we created a corpus called Hungarian QE (HuQ). The HuQ corpus contains 1500 English-Hungarian sentence pairs. To build the HuQ corpus, we used 300 English sentences of mixed topics from the Hunglish corpus (Halácsy et al., 2005). We translated these 300 sentences into Hungarian with different MT systems. After the translation, to create human judgements, we evaluated these translated segments with human annotators. For creating human scores, we developed a website<sup>2</sup> with a form for human annotators to evaluate the translations. In this website we can see an English source sentence and its Hungarian translation, originating from one of the translation sources. However, the evaluators were not aware of the origin of the translation. The annotators could give quality scores from 1 to 5, from two points of view (Koehn, 2010): adequacy and fluency (see Table 1). We added a 0 score (*I do not understand the English sentence*) to filter out wrong evaluations. All the 1500 sentences were evaluated by 3 human annotators: L, M and T. All the annotators were native Hungarian speakers who have minimum B2 level English language skill. The 3 annotators have different evaluation attitudes:

- L: linguist,
- M: MT specialist,
- T: language technology expert.

To follow and control the annotators effectively, or to discuss the annotation aspects with the annotators personally, to avoid misunderstandings, we did not use crowdsourcing for the evaluation. In order to ensure a consistence annotation scheme, the 3 annotators evaluated a set of 50 translations in a personal meeting. These translations are not included in the training set.

There are 3 topics in the HuQ corpus: subtitles, literature and law. The subtitles are simple daily used sentences containing a high ratio of slang words. The language of literature has more complex grammatical constructions with many rare words used. The segments from law are official texts with complex grammar.

We used 5 different translations for each of the 300 sentences. One of them is human translation from the Hunglish corpus, the remaining translations are from 4 different MT systems:

<sup>2</sup><http://nlp.itk.ppke.hu/node/65>

1. MetaMorpho (Novák et al., 2008) rule based MT system,
2. Google Translate,
3. Bing Translator,
4. MOSES statistical MT toolkit (Koehn et al., 2007).

The Google Translate and the Bing Translator are statistical MT systems. The main advantage of these two systems is that these are trained on huge corpora. Thus, the commonly used phrases will be translated in high quality, but in the case of unseen or rare segments or word forms, the quality will be low. In contrast, the MetaMorpho rule based MT system can handle numerous grammatical forms. Thus, it can gain high quality both in adequacy and fluency. The MOSES MT toolkit was trained on the Hunglish corpus, which contains  $\sim 1.1$  million English-Hungarian sentence pairs, which is not big enough to produce high quality translations. There is a typical difference between statistical MT systems and rule based MT systems for English-Hungarian. In Table 2 we can see an example: *Smith turned the question over in his mind*. The main difficulty for automatic MT systems in this sentence is that not Smith turned over, but the question turned over (by Smith). The MetaMorpho system, using the grammatical analyzer could handle this problem correctly, but the statistical systems could not, because the probability of “Smith turning over” is higher than a “question turning over”. This problem appears in the human evaluation scores as well. We can see in Table 2, in the case of Google Translation, that the 3 annotators gave 3 different scores. One reason for the difference is that the 3 annotators had different attitudes, another reason is the ambiguity. If we translate the Hungarian sentence back, it means: *Smith turned around in his mind, above the question*. Thus, L gave 1 because this translation is totally different from the source sentence. But T gave 5, because these phrases: “in mind”, “turn question”, together definitely have the main meaning that Smith analyzed the question, which has the same meaning as the source sentence. M agrees with both L and T, he is halfway between them. For building the QE models, we used the arithmetic mean of the scores of the 3 annotators:

- AD: arithmetic mean of the adequacy scores,
- FL: arithmetic mean of the fluency scores,
- AF: arithmetic mean of the AD and FL scores.

We also created classification scores, because there are many cases, when we do not need 5 grades. For instance, the companies and translators need only 2 or 3 classes: need post-edit – do not need post edit; correct – need correction, etc. We created 3 classes from the AD, FL and AF scores:

- BAD:  $1 \leq x \leq 2$ ,
- MEDIUM:  $2 < x < 4$ ,
- GOOD:  $4 \leq x \leq 5$ ,

where:  $x = AD, FL$  or  $AF$ . The classification scores are:

- CLAD: classification scores from AD,
- CLFL: classification scores from FL,
- CLAF: classification scores from AF.

## 5. Methods, experiments and optimization

Using the HuQ corpus with AD, FL, AF, CLAD, CLFL and CLAF, we built the QE models. For building the QE model, features as quality indicators are needed to be extracted from the corpus. Then, with a machine learning method, human or automatic evaluation scores are used to build the QE model. To create the quality indicators from features, we used the QuEst framework. In this study, 103 features (103F) were extracted from the corpus. The set of 103 features contains 76 features implemented by Specia et al. and 27 additional features developed by us. In the 103F, there are adequacy features (e.g. ratio of percentage of nouns in the source and target, ratio of number of tokens in source and target, etc.), fluency features (e.g. perplexity of the target, percentage of verbs in the target, etc.) and complexity features (e.g. average source token length, source sentence log probability, etc.). The 27F contains 3 dictionary features and 24 features using WordNet and word embedding models.

The first task was doing evaluations with differently-sized portions of the HuQ corpus. Secondly, we evaluated the HuQ corpus with standard automatic metrics. Thereafter, we built different QE models for English-Hungarian. First, we tried the 17 baseline features (17F) (Specia et al., 2013) for Hungarian. The 17F is language and language tool independent. Then we performed experiments with the 103F (17F is subset of 103F). The problem was that the 103F contains features that use language dependent linguistic tools (e.g. Stanford parser (De Marneffe et al., 2006), Berkeley Parser (Petrov et al., 2006) etc.). Most of these tools, however, are not applicable to Hungarian. Thus, we integrated the available Hungarian linguistic tools into QuEst: For Part-of-Speech (POS) tagging and lemmatization, we used PurePos 2.0 (Orosz and Novák, 2013), which is an open source, HMM-based morphological disambiguation tool. Purepos2 has the state-of-the-art performance for Hungarian. It has the possibility to integrate a morphological analyzer. Thus, to get the best performance, we used Humor (Prószéky, 1994), a Hungarian morphological analyzer. For NP-chunking, we used HunTag (Recski and Varga, 2009) that was trained on the Szeged Treebank (Csendes et al., 2005). HunTag is a maximum entropy Markov-model based sequential tagger. There are many language specific features that could not be extracted, because there are no Hungarian language tools for them.

For the machine learning task, we used the Weka system (Hall et al., 2009). We created 7 classifiers with 10 fold cross-validation: Gaussian Processes with RBF kernel, Support Vector Machine for regression with Normalized-PolyKernel (SMOreg), Bagging (with M5P classifier), Linear regression, M5Rules, M5P Tree and for classification we used Support Vector Machine with NormalizedPolyKernel (SMO). We only show the results of the SMOreg and SMO, because these gained the best results. For evaluating

MT system	Example	Adequacy			Fluency		
		L	M	T	L	M	T
Source	Smith turned the question over in his mind.						
Reference	Smith megvizsgálta a kérdést.						
MetaMorpho	Smith a kérdést forgatta a fejében. (Smith turned the question in his mind.)	2	5	5	4	5	5
Google	Smith megfordult a kérdés felett a fejében. (Smith has turned in his mind above the question.)	1	3	5	5	3	4
Bing	Smith megfordult a kérdés a fejében. (Smith the question turned in his mind.)	4	5	4	4	4	4
MOSES	Cyrus smith a kérdést. (Cyrus smith the question.)	1	1	1	1	1	4

Table 2: Example of translation difference

the performance of our methods, we used the statistical correlation, the MAE (Mean absolute error), the RMSE (Root mean-squared error) and the Correctly Classified Instances (CCI) evaluation metrics. The correlation ranges from -1 to +1, and the closer the correlation to -1 or +1, the better it is. In the case of MAE and RMSE the closer the value to 0, the better.

We developed 27 new word-level semantic features. Our aim was to quantify the similarity and relatedness of the topic or meaning of the source and the target sentences. We created bag of words (BOW) from the source and the target segments.

We used 3 features extracted from an English-Hungarian dictionary used by the MetaMorpho system, which contains 365000 entries. We created noun, verb, adjective BOW from the source and the target sentences, then we counted the source-target word pairs from the BOW, which are included in the dictionary. After all, we divided the matches by the length of the source sentence, the length of the target sentence and we counted the F1 score of them.

We developed an additional 24 features using WordNet and word embedding models. We used the Princeton WordNet 3.0 (Fellbaum, 1998) and the Hungarian WordNet (Miháلتz et al., 2008). We collected the synsets of the nouns in the source and the target segments. Then, we collected the hypernyms of the synsets up to two levels. Using the collected synsets and hypernym synsets we counted the weighted intersection of synsets of the source and the target words. Features are extracted from the result synsets. We counted the instances of the result synset and divided the sum with the length of the source sentence, the length of the target sentence, the number of nouns in the source sentence, the number of nouns in the target sentence and we counted the F1 score of them. Using these counts, we also created features with the verbs, the adverbs and the adjectives.

However, if looking up words in WordNet did not provide any results, which is quite often the case because of the small coverage of the Hungarian WordNet, we used word embedding models to substitute synset results (Mikolov et al., 2013b; Mikolov et al., 2013a). Thus, first we trained a CBOW model with 300 dimensions on a 3-billion-word lemmatized Hungarian corpus. The reason for using the lemmatized version was to have set of semantically related words, rather than syntactically related ones. Due to the ag-

glutinating behaviour of Hungarian, building an embedding model from the raw text would have provided syntactically similar groups of words, and only a second key of similarity would have been their semantic relatedness (Siklósi and Novák, 2016). However, in the lemmatized model, this problem was eliminated. Thus, if there was no result for a word from WordNet, its top 10 nearest neighbours were retrieved from this embedding model, resulting in a list of quasi synonyms, and used the same way as WordNet synsets. However, as these lists did not necessarily correspond to exact synonyms of the original word, the weight of this feature was lower (set to 0.1).

We carried out experiments for five different settings:

1. task (T1): we did statistical and inter-annotator agreement measurements on the HuQ corpus.
2. task (T2): we compared and evaluated the quality of the MT systems.
3. task (T3): the HuQ corpus is evaluated using automatic evaluation methods: TER, BLEU and NIST (Lin and Och, 2004)
4. task (T4): using the HuQ corpus and the 103F, we built QE models with different portions of the HuQ corpus trained on AF: 100, 500, 1000 and 1500 sentence pairs.
5. task (T5): using the HuQ corpus, the 17F and the 103F, we built QE models trained on the automatic evaluation metrics, the AD, the FL, the AF, the CLAD, the CLFL and the CLAF scores.
6. task (T6): using the HuQ corpus and the optimized feature sets, we built the QE models trained on the AD, the FL, the AF, the CLAD, the CLFL and the CLAF scores.

The experiment with human scores needed to be optimized for English-Hungarian. For optimizing, we used the forward selection method. First, we extracted and evaluated each feature separately. Then we chose the feature that produced the best result. Thereafter, we combined the chosen feature with each remaining feature, and we added the feature that produced the best combined result in each round.

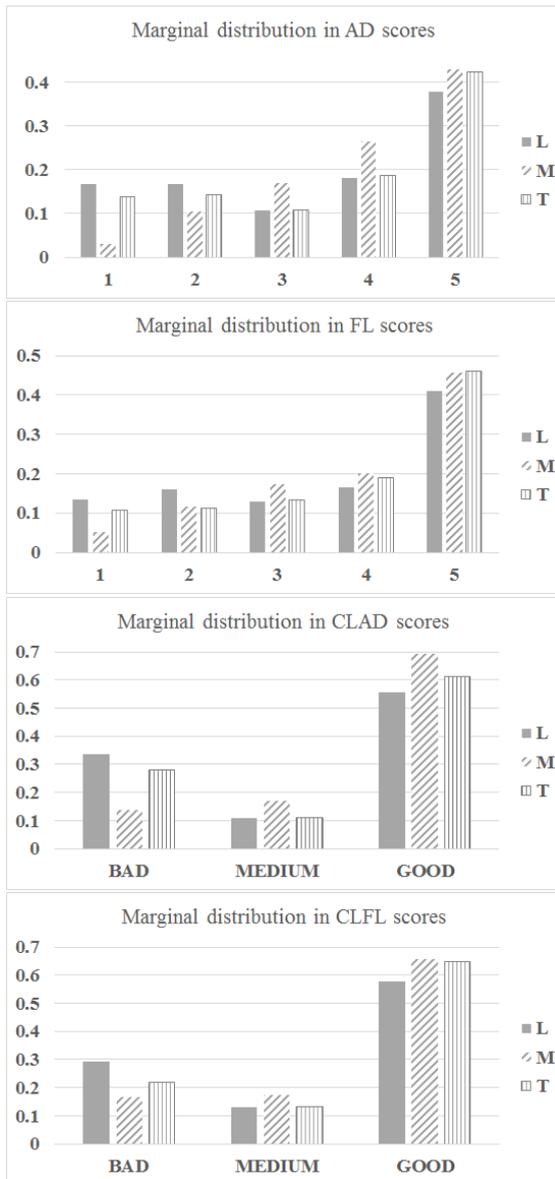


Figure 1: Marginal distributions

Then, we continued adding features until the combined result did not improve any further.

## 6. Results and Evaluation

In Table 3 we can see the inter-rater agreement found in T1 and in Figure 1 we can see the marginal distributions. Because of the ambiguities described in Section 4, the Fleiss Kappa values of inter-annotator agreement between the 3 annotators is moderate.

In T2 (see Table 4), as we expected, MOSES achieved the lowest result and MetaMorpho performed best.

The results of T3 describe the quality of the HuQ corpus. The system-level results of the T3 evaluation: TER: 0.6107; BLEU: 0.3038, NIST: 5.1359. According to the TER and

the BLEU scores, ~30% of the HuQ corpus contains correct translations.

According to CLAF scores, we also counted the "GOOD" classes. There are 780 instances of GOOD, which means 52% of the HuQ corpus contains correct or close to correct translations. According to AF scores, we counted the perfect translations (score value = 5), there are 387 instances of perfect translation, which means 25.8% of HuQ are correct translations.

In T4, as we can see in Table 5, increasing the size of HuQ, we got better results:

- the AF-500 could gain ~24% higher correlation than the AF-100,
- the AF-1000 could gain ~3% higher correlation than the AF-500,
- the AF-1500 could gain ~1.5% higher correlation than the AF-1000.

The results of T5 experiments show the performance of building the QE models to predict the standard automatic evaluations (see in Table 6) and the human judgements. As we can see in Table 7, the AD-103F could gain ~10% higher correlation than the 17F baseline set, the FL-103F could gain ~6% higher correlation than the 17F baseline set, the AF-103F could gain ~7% higher correlation than the 17F baseline set. It means that, the baseline features are not the optimized features for Hungarian. Thus, we need to find the most relevant features for Hungarian QE.

In T6, first, we used the 103F to build QE models trained on AD, FL, AF, CLAD, CLFL and CLAF human scores. Then, we optimized the models to Hungarian. After optimizing, as we can see the results in Table 7 and in Table 8, the optimized features for Hungarian could gain ~14% (optimized AD - 29 features), ~10% (optimized FL - 32 features), ~12% (optimized AF - 26 features), ~6% (optimized CLAD - 21 features), ~5% (optimized CLFL - 10 features) and ~4% (optimized CLAF - 12 features) higher correlation than the baseline features. The optimized AD set contains 5, the optimized FL set contains 8, the optimized AF contains 5, the optimized CLAD contains 2, the optimized CLFL contains 1 and the optimized CLAF contains 2 semantic features developed in this research. As we can see in the results of the optimization, each optimized feature set contains semantic features developed in this research, which means that the semantic features are important features for QE.

## 7. Conclusion

We created the HuQ corpus for quality estimation of English-Hungarian machine translation. The corpus contains 1500 quality scores of translations, which are given by human annotators. Then, using the HuQ corpus, we built different QE models for English-Hungarian translations. In our experiments, we used automatic metrics and human judgements as well. In the experiments we tried 103 features including 27 newly developed semantic features using WordNet and word embedding models. Then, we optimized the quality models to English-Hungarian. In

	AD	FL	CLAD	CLFL	CLAF
Fleiss Kappa	0.357	0.463	0.44	0.521	0.493
Krippendorff Alpha	0.357	0.463	0.44	0.521	0.493
Average Pairwise Cohen’s Kappa	0.360	0.464	0.444	0.522	0.494
Average pairwise percent	52.467%	61.222%	70.022%	74.444%	70.6%

Table 3: Evaluation of annotator-rater agreement

	AD mean	FL mean	AF mean
MetaMorpho	3.8707	3.8651	3.8679
MOSES	3.0175	3.1872	3.1024
Google	3.6395	3.5729	3.6062
Bing	3.2166	3.2256	3.2211

Table 4: Quality of MT systems

	Correlation	MAE	RMSE
AF-100	0.2700	0.8159	1.0613
AF-500	0.5155	0.8478	1.0603
AF-1000	0.5480	0.8147	1.0481
AF-1500	<b>0.5618</b>	<b>0.7962</b>	<b>1.0252</b>

Table 5: Evaluation of T4

	Correlation	MAE	RMSE
TER	0.3550	0.3275	0.4357
BLEU	0.4404	0.2201	0.3474
NIST	0.3669	2.6695	3.4777

Table 6: Quality of MT systems

	Correlation	MAE	RMSE
AD-17F	0.3832	0.9429	1.1990
AD-103F	0.4847	0.8805	1.1199
Optimized AD	<b>0.5245</b>	<b>0.8397</b>	<b>1.0869</b>
FL-17F	0.5400	0.8229	0.8345
FL-103F	0.6070	0.7723	1.0297
Optimized FL	<b>0.6413</b>	<b>0.7440</b>	<b>0.9878</b>
AF-17F	0.4931	0.8345	1.0848
AF-103F	0.5618	0.7962	1.0252
Optimized AF	<b>0.6100</b>	<b>0.7459</b>	<b>0.9775</b>

Table 7: Evaluation QE using the human judgements

	CCI	MAE	RMSE
CLAD-17F	0.5493	0.3590	0.4591
CLAD-103F	0.5766	0.3492	0.4483
Optimized CLAD	<b>0.6093</b>	<b>0.3370</b>	<b>0.4346</b>
CLFL-17F	0.5887	0.3434	0.4419
CLFL-103F	0.6246	0.3310	0.4275
Optimized CLFL	<b>0.6407</b>	<b>0.3299</b>	<b>0.4262</b>
CLAF-17F	0.5780	0.3433	0.4417
CLAF-103F	0.6033	0.3347	0.4318
Optimized CLAF	<b>0.6180</b>	<b>0.3299</b>	<b>0.4263</b>

Table 8: Evaluation of QE using the classification metrics

the optimization task, we used forward selection to find the best features. We could produce optimized sorted feature sets, which produced more than 10% better correlation than the baseline set. In our experiments, our HuQ corpus and QE models can be used for predicting the quality of machine translation outputs for English-Hungarian.

In the future, we would like to enlarge the corpus. We also would like to examine the effect of utilizing crowdsourcing to increase the size of HuQ. Last, but not least, we will do experiments and evaluations in a ranking task.

### Acknowledgements

The authors would like to thank Andrea Dömötör for her generous help in the annotation work.

### References

- Beck, D., Shah, K., Cohn, T., and Specia, L. (2013). Shelfite: When less is more for translation quality estimation. In *Proceedings of the Workshop on Machine Translation (WMT)*, August.
- Biçici, E. (2013). Feature decay algorithms for fast deployment of accurate statistical machine translation systems. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria, August. Association for Computational Linguistics.
- Bicici, E., Liu, Q., and Way, A. (2015). Referential translation machines for predicting translation quality and related statistics. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 304–308, Lisbon, Portugal, September. Association for Computational Linguistics.
- Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. (2015). Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal, September. Association for Computational Linguistics.
- Camargo de Souza, J. G., Buck, C., Turchi, M., and Negri, M. (2013). FBK-UEdin participation to the WMT13 quality estimation shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 352–358, Sofia, Bulgaria, August. Association for Computational Linguistics.
- Csendes, D., Csirik, J., Gyimóthy, T., and Kocsor, A. (2005). The Szeged Treebank. In *Lecture Notes in Computer Science: Text, Speech and Dialogue*, pages 123–131. Springer.

- De Marneffe, M.-C., MacCartney, B., Manning, C. D., et al. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454.
- Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.
- Halácsy, P., Kornai, A., Németh, L., Sas, B., Varga, D., Várad, T., and Vonyó, A. (2005). A Hunglish korpusz és szótár. In *III. Magyar Számítógépes Nyelvészeti Konferencia*. Szegedi Egyetem.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.
- Langlois, D. (2015). Loria system for the wmt15 quality estimation shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 323–329, Lisbon, Portugal, September. Association for Computational Linguistics.
- Lin, C.-Y. and Och, F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Miháltz, M., Hatvani, C., Kuti, J., Szarvas, G., Csirik, J., Prószéky, G., and Várad, T. (2008). Methods and results of the hungarian wordnet project. In *Proceedings of the Fourth Global WordNet Conference GWC 2008*, pages 310–320.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.
- Novák, A., Tihanyi, L., and Prószéky, G. (2008). The metamorpho translation system. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 111–114, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Orosz, G. and Novák, A. (2013). Purepos 2.0: a hybrid tool for morphological disambiguation. In *RANLP'13*, pages 539–545.
- 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*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Petrov, S., Barrett, L., Thibaux, R., and Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 433–440. Association for Computational Linguistics.
- Prószéky, G. (1994). Industrial applications of unification morphology. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, pages 213–214, Stuttgart, Germany, October. Association for Computational Linguistics.
- Recski, G. and Varga, D. (2009). A Hungarian NP Chunker. *The Odd Yearbook. ELTE SEAS Undergraduate Papers in Linguistics*, pages 87–93.
- Siklósi, B. and Novák, A. (2016). Beágyazási modellek alkalmazása lexikai kategorizációs feladatokra. *XII. Magyar Számítógépes Nyelvészeti Konferencia*, pages 3–14.
- Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.
- Specia, L., Shah, K., de Souza, J. G., and Cohn, T. (2013). Quest - a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84, Sofia, Bulgaria, August. Association for Computational Linguistics.

## Technology Landscape for Quality Evaluation: Combining the Needs of Research and Industry

Kim Harris<sup>1,2</sup>, Aljoscha Burchardt<sup>2</sup>, Georg Rehm<sup>2</sup>, Lucia Specia<sup>3</sup>

<sup>1</sup>text&form GmbH (Berlin), <sup>2</sup>Deutsches Forschungszentrum für Künstliche Intelligenz (Berlin), <sup>3</sup>University of Sheffield

### Abstract

Translation quality evaluation (QE) has gained significant uptake in recent years, in particular in light of increased demand for automated translation workflows and machine translation. Despite the need for innovative and forward-looking quality evaluation solutions, the technology landscape remains highly fragmented and the two major constituencies in need of collaborative and ground-breaking technology are still very divided. This paper will demonstrate that closer cooperation between users of QE technology in research and industry to create a holistic but highly adaptable environment for all aspects of the translation improvement process, most significantly quality evaluation, can lead the way to novel and ground-breaking achievements in accelerated improvement in machine translation results.

**Keywords:** Machine Translation, Evaluation, Human Translation

### 1. Introduction

Currently, the approaches and tools applied by research and industry to evaluate the quality of translation differ widely from each other, in terms of both methodology and implementation. Yet, the needs of both constituencies are largely identical: Both want to determine overall translation quality for various purposes, both want to understand the underlying issues – or errors – and fix them, and, most importantly, both want to improve translation output, i. e., prevent those issues from recurring in the future.

While most language service providers primarily perform QE on translations carried out by professional translators, there is a positive trend towards the integration of machine translation (MT) solutions in “traditional” translation workflows (Autodesk, 2011). Consequently, the demand for efficient QE processes to improve the content as it moves through the typical translation cycle has increased. In a 2013 survey performed by the QTLaunchpad Consortium (Doherty et al., 2013) two-thirds of all language industry respondents said they were currently using or planned to use machine translation in their translation business, and almost 70% said they use human evaluation methods to assess the quality of MT output, with only 22% using automatic evaluation metrics such as BLEU and TER.

Language service providers are often bound by the (human) translation technology dictated by their customers or that offers features that make the translation process more efficient and thus more widely accepted by the translator community. A number of LSPs have incorporated MT generated content into these translation environments and succeeded in integrating quality *estimation* tools in their workflows to filter out the automatic translations that are not worth editing. However, these processes neither fully integrate research approaches nor do they directly support the **improvement** of the generated content for future use. MT is still widely seen as a black box, and very few have the resources to invest in closer ties to the research community, in the rare cases where this is actively pursued.

This approach is the complete reverse of that applied by the research community to evaluate MT output. Historically, research has largely relied on automatic evaluation metrics such as BLEU (Papineni et al., 2002), TER (Snover et al., 2006) and METEOR (Banerjee and Lavie, 2005) to assess

the quality of the MT system output for a specific language pair based on one or more reference (human) translations. While the generated score gives an indication of the *overall* quality, it does not provide information on the *reason* for the outcome, nor does it reveal how to improve the translation in the future.

Some research has seen a shift towards human evaluation in the form of MT translation ranking and other primarily non-linguistic evaluations performed to a significant extent by untested and unqualified crowdsourced resources (Graham et al., 2016) or researchers with no translation background (Bojar et al., 2015). The integration of professional translators in the evaluation process is still lagging, largely due to the lack of collaboration with the language industry on a broader scale.

This gap between these two constituent drivers of machine translation has become somewhat of a conundrum: Commercial LSPs are unable – even unwilling – to invest in their own systems because they have no access to the necessary expertise, no financial resources and see relative stagnation in MT innovation and therefore no business case for the investment. The research community has been sufficiently successful in proving its own results for its own purposes with automatic scoring and minimal human ranking efforts, and therefore sees little reason to invest financially and otherwise in the integration of professional translators into the research loop to find more novel and less automatic ways of looking deeper into the crystal ball.

### 2. Fragmentation in the Translation Industry

As a result, there is little overlap in the methods and tools currently used by these two groups for quality evaluation and even less interaction between or influence of one over the other in a move towards more interconstituent standardization. This, however, does not only lie in the lack of cooperation between the research community and the language industry, but also in the inherent fragmentation of the processes and tools implemented by either constituency respectively.

## 2.1. The Question of Quality

The greatest challenge and root of much debate and discord relates to defining quality. What is it exactly? According to (Koby et al., 2014) “*quality translation demonstrates accuracy and fluency required for the audience and purpose and complies with all other specifications negotiated between the requester and provider, taking into account end-user needs.*” While there are many other scientific definitions similar to this one, in reality quality is whatever the customer wants it to be. This in itself demonstrates just how diverse and heterogenous quality standards and all aspects of translation quality must be and have always been. As a result, the evaluation of this quality poses a significant challenge if the number of factors affecting quality is multiplied by the number of criteria used to evaluate it.

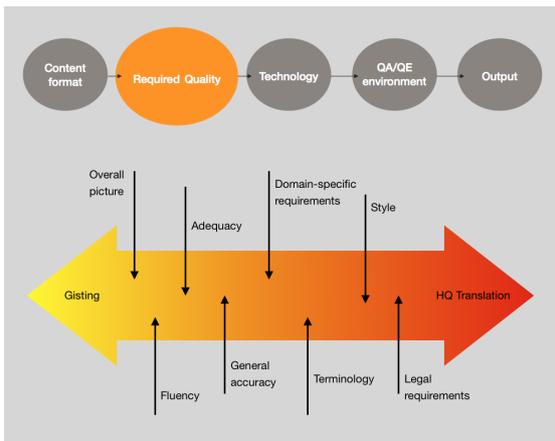


Figure 1: Quality scale in translation workflow

As shown in Figure 1, there are nuances in quality expectations that vary from case to case, and these will depend on a range of factors that influence the expected quality, including the purpose of the content, its format, domain, time constraints, financial issues and other customer and content-related factors such as tools and publication (Zaret, 2016). All of these factors impact not only the type of quality evaluation performed on the translation but the environment in which evaluation can, or even must, be performed. We can demonstrate the diversity of fit-for-purpose quality expectations by comparing two vastly different scenarios. Customer A requires the translation of a legally binding financial document for broader publication and application. Customer B has general e-mail correspondence between two subsidiaries for internal use only. Not only is the domain different, but so is the purpose. Whereas gisting and some light post-editing might be feasible for Customer B, an absolutely flawless and highly accurate translation will be required by Customer A. Quality for Customer B is proper conveyance of the overall meaning, which is insufficient for legally binding documents.

## 2.2. Translation Technology Landscape

Given the sheer size, diversity and unabating growth of the language industry, and the lack of standardization in key areas such as format and quality, it is hardly surprising that

the industry is enormously fragmented. Translation has become somewhat ubiquitous with the rise of free online translation services such as Google and Bing. Yet, there are over 25,000 registered language service providers worldwide using hundreds of different technologies to perform translation and quality assessment. Fragmentation appears to meet the needs of those who have a demand.

The drive to reach global markets in a competitive landscape has been quintessential in the positive impetus experienced by the language industry, but it has also played a major role in the development of highly specialized and often customized technologies and environments specific to both customer and content. Repositories for open source tools and language resources such as META-SHARE<sup>1</sup> and language technology associations such as LT-Innovate<sup>2</sup> reference hundreds of language tools and resources and demonstrate clearly how significant and how fragmented the language industry is, from both an industry and a research perspective.

The user-driven sophistication of standard technology used by language service providers is striking when compared to that of many open source solutions, particularly those used by the research community. The most successful translation environments are those that offer efficient workflows and features that are profitable to the supplier and provide the level of quality, speed and price required by the buyer of language services. Tools that are too cumbersome or do not support the most common file formats and markup will find little uptake in the industry. SDL Trados Studio<sup>TM</sup>, shown in Figure 2, is currently the most widely used environment for professional translation and MT integration, however, other applications such as MemSource and MemoQ and hundreds of smaller, specialized applications, all of which offer optimized translation features, multiformat support and MT integration and services are on the rise. Needless to say that most tools used by the language industry are neither interoperable nor compatible except in their most basic text form.

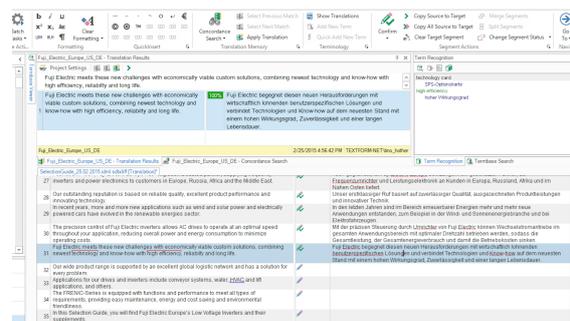


Figure 2: SDL Trados Studio<sup>TM</sup> user interface

## 2.3. Quality Evaluation in the Language Industry

While translation technology has experienced a relative boom during the past few years, not least due to the dawn

<sup>1</sup><http://www.meta-share.eu>

<sup>2</sup><http://www.lt-innovate.org>

of accessible machine translation and the *need for speed* on global markets, standardized, integratable tools to help assess and improve the quality of translated content have not. The evaluation of translation quality represents an area where the absence of reliable and meaningful standardisation and evaluation methods for buyers, suppliers, MT adopters, among others, is particularly serious (Doherty et al., 2013). As shown in Figure 3, language service providers use a vast number of different evaluation methods and standards to assess the quality of their translation output, with proprietary tools and those integrated in other tools making up two-thirds. This is a clear indication that currently none of the aforementioned translation technologies offer suitable or satisfactory integrated QE features at the level required by the user, particularly in light of the fact that well over two-thirds of all respondents still use human quality evaluation *only*.

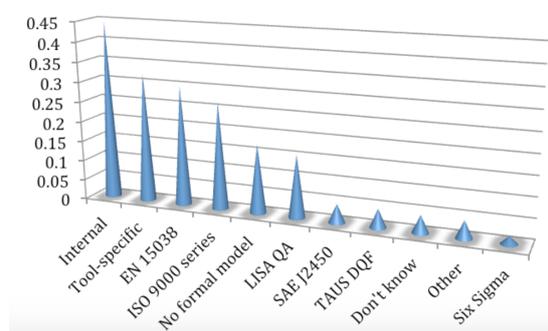


Figure 3: QE tools used by language industry (Doherty et al., 2013)

MT adopters in the language industry do use some of the metrics made available through the research community, such as BLEU and (H)TER, to evaluate the quality of the output, but this is primarily performed in order to filter out what these metrics would consider bad translations based on their scores so that post-editors do not need to do this themselves. It is still considered the most efficient way to perform an overall assessment, but there are no other efficient ways to evaluate translation quality in detail than to do this manually.

MQM, the quality metric developed by the QTLaunchpad Consortium<sup>3</sup> addresses some of these standardization issues with respect to error categorization and the flexible creation of error typologies. It can be integrated into the methods and standards shown in Figure 3, and adapted to fulfill all quality specifications of any given translation task flexibly and easily. This methodology has received positive feedback from a number of research and industry users and has been harmonized with the TAUS DQF<sup>4</sup> to promote industry-wide uptake and push consolidation in the area of quality evaluation.

<sup>3</sup><http://www.qt21.eu/launchpad/>

<sup>4</sup><http://www.taus.net>

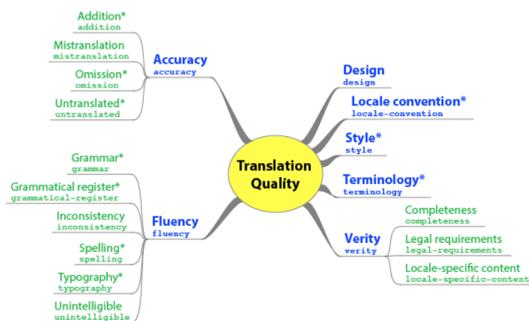


Figure 4: Example of MQM error typology

### 3. Fragmentation in the Research Community

The research community has been pivotal in the continued development and success of machine translation technologies; however, it is a community that spends much of its effort working in silos, developing tools and solutions for highly specific problems or challenges in a particular aspect of their own research. Rarely does one see a collaborative **interoperable** platform of complementary tools that have the potential to address a larger complex of problems, and even more seldom is an ongoing collaborative effort with the user community of these tools to find applications for them in real-world scenarios. The META-SHARE repository alone boasts 2,725 language resources at the time of writing, 102 of which contain the keyword *quality*.

Automatic evaluation metrics such as BLEU and TER are two of the most popular and inexpensive automated metrics and have been known to demonstrate a relatively high correlation with human judgements. The resulting quality scores are based on comparisons with sets of HT references, which can be useful for certain estimation tasks; however, they do not provide the ability to assess why scores improve or worsen, and they focus almost exclusively on the score, offering insufficient insight into real error analysis and improvement.

The number of automatic evaluation metrics alone is a clear indication of just how granular an evaluation metric is to a particular subtask of a specific task. Much like the fragmentation found in industry, many of these metrics will have some degree of overlap, yet there seems to be little interest in adapting or combining existing tools instead of developing new ones.

A number of quality estimation and evaluation tools developed by the research community have attempted to combine various aspects of the actual translation quality with the use of automatic metrics, such as QuEst<sup>5</sup>, Asiya (see Figure 5) and Appraise<sup>6</sup>, the latter of which also integrates human error annotation in its quality metrics.

<sup>5</sup><http://www.quest.dcs.shef.ac.uk>

<sup>6</sup><https://github.com/cfedermann/Appraise>

The screenshot shows the Asiya-Online interface. At the top, there are three sections for uploading files: 'Source file', 'Reference files', and 'System translation files'. Each section has a text input field and a 'Upload File' button. Below this is the 'Evaluation Options' section, which includes a 'Metric selection' dropdown menu with a list of metrics like BLEU, NIST, TER, etc. There are also 'Clear Files' and 'Run Asiya!' buttons. At the bottom, there is an 'Asiya Report:' section with a table of evaluation results.

Systems	Document	Segment	BLEU	GTM-3	NIST	WER	-PER	OI	TERbase	METEOR	ROUGE	BL
sys.txt	no name	1	0.1134	0.2535	1.3333	-0.75	-0.625	0.3077	-0.75	0.1931	0.4	0.4

Figure 5: Asiya-Online with evaluation metrics

#### 4. The Motivational Divide Between Research and Industry

As discussed earlier in the paper, the objectives of both communities are identical: to determine overall translation quality for various purposes, to understand the underlying issues – or errors – and fix them, and, most importantly, to improve machine translation output, i. e., prevent occurring issues from recurring in the future. Why, then, have we not seen more cooperation towards these common goals?

Although the objectives are seemingly similar, the motivation that drives them is completely different. Industry, on the one hand, needs reliable, faster solutions that are scalable and financially viable. Quality is no longer a unique selling point. It is a requirement, regardless of how the customer defines it. Settings up machine translation systems and automated quality metrics can be expensive, complex, complicated and embody the proverbial black box for many language service providers. The systems either rely too heavily on large amounts of data and experienced resources with the right background in computer science, or on intrinsic linguistic programming that is time-consuming and only applicable to a handful of language pairs. Neither scenario has proven promising to the majority of LSPs. Real progress is slow, innovative technology drives are few and far between, and the cost of ramping up an MT workflow for a customer often brings with it a significant financial risk.

What is lacking in the language industry is the motivation to participate in a collaborative paradigm shift towards human-informed MT development. There is little interest in collaboration, which stems largely from its cottage-industry heritage, as well as a fear of promoting their own professional demise. Diversity of language is a welcome excuse to remain as fragmented as possible. It is the Darwinian survival of the fittest.

This concept of survival is not unknown to the research community either, and it is the force that drives the lone-ranger mentality in many aspects of its work. Most institutions

are not interested in finding industry applications for their research but choose to focus on proving the point of their research in order to find and receive funding.

As with language service providers, financial considerations are the key factor when deciding how to spend a budget. Working with industry partners is understandably more expensive than hiring primarily unqualified Mechanical Turkers or finding colleagues or crowd-sourced resources to perform some of the manual tasks involved in some research. It is little wonder that the results are far from ideal, although research would be hard-pressed to agree that lack of skills and qualification may be the cause, but the investment in much more promising collaboration with professionals is seen as too time-consuming and too costly.

##### 4.1. Closing the Gap

Bringing the language industry into the research fold and vice-versa is a win-win situation for both. The development of language technology in a multi-billion dollar language industry with an annual growth rate of almost 5%<sup>7</sup> is extremely lucrative for those whose business is language, and if the research community can demonstrate visible, profitable and concrete technological innovation and breakthroughs in application scenarios, they will make a good case for significantly more funded research in the field.

Quality evaluation development that incorporates the needs of both communities can provide the necessary impetus for more collaborative efforts and promote a greater level of understanding of the work each group does. Some open-source tools such as translate5<sup>8</sup> are now beginning to understand these parallels and are developing environments that combine the business features required by industry with the scientific features required by research. The goal is to turn translate5 into a flexible repository and data curation tool for MT research going beyond the functionality that can be provided by open resource exchange and sharing facilities such as META-SHARE (Burchardt et al., 2016).

##### 4.2. Single Environment for Multiple Objectives

A holistic environment that combines quality evaluation requirements for professional translation and machine translation output in both business and research applications and offers flexible tool integration for different evaluation scenarios will provide the foundation for novel and groundbreaking research in improving machine translation quality. Incorporating the linguistic and language-related knowledge of industry experts into machine translation research can uncover previously unattainable information that is vital to the improvement process.

Until now, the language industry has relied primarily on human resources to manually fix issues in the machine translation output to achieve a suitable level of quality. This process does not address, help understand, or permanently remedy underlying errors. It is not that the errors are not understood or that the user does not want to apply the information to improve the next translation. The system, tools

<sup>7</sup><http://www.pangeanic.com/knowledge-center/size-of-the-translation-industry/>

<sup>8</sup><http://www.translate5.net>

and workflow do not support the incorporation of this type of information, so the information is not collected despite its valuable potential. The heterogeneous translation environments and large number of quality standards complicate matters.

The research community, on the other hand, has focused much of its quality evaluation effort on improving the scores of automated metrics, sometimes based on reference translations completed by human resources, other times based on rankings and other forms of overall evaluations. Rarely does the feedback from linguistic experts find its way into ongoing research, and manual tasks such as annotation or error categorization are seen as too costly and ineffective. Without some of this information, it is difficult for the research community to see the benefits of applying it. Moreover, much of the research performed on its own is related to and can profit from research performed elsewhere.

A single, common environment that can connect all of these constituencies with each other, allow them to share information and results, experiment with data to which they would otherwise have no or little access can facilitate a level of communication that promotes cooperation and innovation. It can provide industry with a standardized platform that supports the import and export of files in any format, the definition of flexible quality metrics using MQM and other tools, the annotation and post-editing of machine translation for improvement cycles. It will make the efforts of the research community more accessible and comprehensible,

In turn, the research community will benefit from the work performed by industry users, making the quid pro quo collaboration on a unified platform affordable. It will have quick and easy access to data and results of other research users in an endless repository and the ability to plug-and-play almost any of the 2,725 language resources on META-SHARE.

## 5. Conclusions

The development of a holistic environment for translation quality evaluation that encompasses the requirements of both the research community and the language industry can have a significant positive impact on the future of language technology, in particular machine translation. It can provide the foundation for closer collaboration between the constituencies most interested in improving machine translation and secure the future of language technology and the translation industry.

## Acknowledgements

This article has received support from the EC's Horizon 2020 research and innovation programme under grant agreements no. 645452 (QT21) and no. 645357 (CRACKER). We thank the anonymous reviewers for their valuable comments.

## References

Autodesk. (2011). Translation and Post-Editing Productivity. In <http://translate.autodesk.com/productivity.html>.

- Banerjee, S. and Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 65–72, Ann Arbor, MI.
- Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. (2015). Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal, September. Association for Computational Linguistics.
- Burchardt, A., Harris, K., Rehm, G., and Uszkoreit, H. (2016). Towards a systematic and human-informed paradigm for high-quality machine translation. In *Translation evaluation – From fragmented tools and data sets to an integrated ecosystem, LREC 2016 workshop*.
- Doherty, S., Gaspari, F., Groves, D., van Genabith, J., Specia, L., Burchardt, A., Lommel, A., and Uszkoreit, H. (2013). Mapping the industry i: Findings on translation technologies and quality assessment. QTLaunchpad, FP7 funded by the European Union, Grant number 296347.
- Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2016). Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, FirstView:1–28, 1.
- Koby, G. S., Fields, P., Hague, D., Lommel, A., and Melby, A. (2014). Defining translation quality. *Revista Tradumàtica*, Traducció i qualitat (Número 12), December.
- Papineni, K., Roukos, S., Ward, T., and jing Zhu, W. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL 2002*, pages 311–318, Philadelphia, PA.
- Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.
- Zaret, A. (2016). A quality evaluation template for machine translation. *Translation Journal*, January.

# Diagnosing High-Quality Statistical Machine Translation Using Traces of Post-Editon Operations

Julia Ive<sup>1,2</sup>, Aurélien Max<sup>1</sup>, François Yvon<sup>1</sup>, Philippe Ravaud<sup>2</sup>

(1) LIMSI, CNRS, Univ Paris-Sud, Université Paris-Saclay, 91 403 Orsay, France,

(2) Cochrane France, INSERM U1153, 75181 Paris, France

{julia.ive, amax, yvon}@limsi.fr, philippe.ravaud@htd.aphp.fr

## Abstract

This paper proposes a fine-grained flexible analysis methodology to reveal the residual difficulties of a high-quality Statistical Machine Translation (SMT) system. This proposal is motivated by the fact that the traditional automated metrics are not enough informative to indicate the nature and reasons of those residual difficulties. Their resolution is however a key point towards improving the high-quality output. The novelty of our approach consists in diagnosing Machine Translation (MT) performance by making a connection between errors, the characteristics of source sentences and some internal parameters of the system, using traces of Post-Editon (PE) operations as well as Quality Estimation (QE) techniques. Our methodology is illustrated on a SMT system adapted to the medical domain, based on a high quality English-French parallel corpus of Cochrane systematic review abstracts. Our experimental results show that the main difficulties that the system faces are in the domains of term precision and source language syntactic and stylistic peculiarities. We furthermore provide general information regarding the corpus structure and its specificities, including internal stylistic varieties characteristic of this sub-genre.

**Keywords:** MT evaluation, high-quality SMT, post-editon

## 1. Introduction

Nowadays, narrowly-specialized MT systems are able to produce very high quality translations, as measured by automated metrics. In most cases, though, the final output still requires manual improvements to reach a publishable quality. However, standard automated metrics such as (H)BLEU (Papineni et al., 2002), (H)METEOR (Denkowski and Lavie, 2014) or (H)TER (Snover et al., 2006)<sup>1</sup> provide little clues regarding the remaining errors, and are of little help to suggest fixes or improvements.

The same can be said of automated error analysis techniques, which are often based on similar principles (Popovic and Ney, 2011; Bojar, 2011): In particular, they often consider the system as a black-box and tend to ignore the characteristics of the source text.

In this study, we propose an alternative fine-grained methodology that helps indicate translation difficulties in connection to the peculiarities of the source document, and also provide some hints as to the reasons of those difficulties in relation to the original corpus and the internal scoring procedures. Such a methodology proves especially useful in the context of high-quality MT, which requires more targeted and sophisticated solutions for further improvement. Our approach is illustrated using a medical SMT system built from a corpus of Cochrane medical systematic review abstracts. An English-French parallel corpus of such abstracts, including human and post-edited automatic translations, will be described.

The rest of this paper is organized as follows: in Section 2., we will present the main characteristics of the Cochrane corpus used. In Section 3., we will describe the chal-

lenges of the medical translation task in the context of the Cochrane Collaboration, before introducing our MT system analysis methodology in Section 4. We will finally present the results of the analysis applied to the Cochrane SMT system in Section 5., and conclude and discuss further prospects for MT evaluation and diagnosis in Section 6.

## 2. The Cochrane Bilingual Parallel Corpus

Cochrane France is part of the international non-profit Cochrane Collaboration<sup>2</sup> whose main mission is to globally spread high-quality evidence-based research in medicine. To this end, the Cochrane Collaboration publishes high-standard research reviews in English and selective translation of their abstracts into (as of now) 12 languages including French, Spanish, Japanese, and traditional Chinese. The review abstracts are publicly available online<sup>3</sup>. Full research reviews are openly accessible only for the low-income and middle-income countries.

Each Cochrane review abstract is made up of the following parts: (a) a plain language summary (PLS, 40% of the abstract, written in popular scientific style), focused on patient comprehension; (b) a scientific abstract (ABS, 60% of the open access abstract, written in scientific technical style), targeting medical experts.

The English-French Cochrane parallel corpus used in this study consists of the following:<sup>4</sup>

- **Cochrane Reference Corpus:** a high-quality corpus consisting of review abstracts translated by agencies and reviewed by domain professionals over a three-year period (2011-2013).

<sup>2</sup><http://www.cochrane.org>

<sup>3</sup><http://www.cochranelibrary.com>

<sup>4</sup>The corpus consisting of source text, machine translation output and PE output is available at <http://www.translatecochrane.fr/corpus>.

<sup>1</sup>Hereinafter, 'H' will be added to refer to the automated metrics applied to the references created by post-editing the evaluated MT output.

- **Cochrane Post-editing (PE) Corpus:** a lower quality corpus consisting of machine-translated review abstracts post-edited mainly by volunteer domain professionals over a 6-month period (Oct. 2013-May 2014). The MT was performed by different versions of the Cochrane SMT.
- **Cochrane Google Post-editing (PE) Corpus:** a lower quality corpus consisting of machine-translated review abstracts by the Google online system<sup>5</sup> post-edited by both professional translators and volunteer domain professionals over a 1-year period (Aug. 2014-Sep. 2015).

Table 1 provides statistics about each part of the corpus.

Corpus	# Lines	# Tokens, en (src)	# Tokens, fr (trg)
Cochrane Reference	≈ 130 K	≈ 2.9 M	≈ 3.6 M
Cochrane PE	≈ 21 K	≈ 500 K	≈ 600 K
Cochrane Google PE	≈ 31 K	≈ 740 K	≈ 890 K

Table 1: Corpora sizes

### 3. Automatic Translation of Cochrane Systematic Review Abstracts: Challenges and Solutions

The translation of English medical texts, in particular that of Cochrane systematic review abstracts, presents a series of challenges regarding:

1. the translation of the terminology and the professional jargon (e.g. abbreviations);
2. the translation of complex syntactic structures and compounds;
3. the adaptation to variations within the scientific style (this is particularly important in the Cochrane context, where different language styles are in use in the PLS and ABS sections).

We manually inspected the paraphrase tables extracted from PLS and ABS parts of the Cochrane Reference and PE Corpora to reveal the following stylistic differences between the registers (Denkowski and Lavie, 2014; Bannard and Callison-Burch, 2005):

1. terminology register (e.g., Source: "cycling", ABS: "cyclisme" 'cycling'<sup>6</sup>, PLS: "vélo" 'bicycle'; Source: "surgical fixation", ABS: "ostéosynthèse chirurgicale" 'surgical osteosynthesis', PLS: "fixation chirurgicale" 'surgical fixation');
2. professional jargon (e.g., Source: "once-daily", ABS: "une administration quotidienne" 'a daily administration', PLS: "une fois par jour" 'once a day'; Source: "viral", ABS: "viral" 'viral', PLS: "par des virus" 'by viruses');
3. selective translation of names (e.g., Source: "Cochrane Library", ABS: "Cochrane Library", PLS: "Bibliothèque Cochrane" 'Cochrane Library'; Source: "Cochrane Review", ABS: "Cochrane Review", PLS: "revue Cochrane" 'Cochrane review');

<sup>5</sup><https://translate.google.com>

<sup>6</sup>Hereinafter, literal translations are provided by the first author.

4. general language (e.g., Source: "to", ABS: "afin de" 'so that', PLS: "pour" 'to'; Source: "flexible", ABS: "flexible" 'flexible', PLS: "souple" 'soft').

The use of domain adaptation techniques, as well as more *ad-hoc* solutions, can help to obtain a better performance in medical MT (Costa-jussà et al., 2012; Wang et al., 2014; Boguraev et al., 2015). In any case, high-quality translation in specialized domains requires training data that closely match the test data.

The Cochrane SMT system for translating the systematic review abstracts is an example of such a narrowly-specialized system. In its current form, our system uses the Moses toolkit (Koehn et al., 2007). The Cochrane Reference corpus is used to train the main model (phrase table and reordering model `msd-bidirectional-fe`). Cochrane PE and additional corpora (WMT'14 medical task parallel data<sup>7</sup>) models (same components as for the main model) were used only for  $n$ -grams (up to  $n = 4$ ) when no translation is found by the first model. The monolingual parts of the corpora mentioned above, as well as general domain data (WMT'13 news data<sup>8</sup>) were used to train the corresponding language models.

The system was tuned using post-edited data, which is in line with the final quality requirements of producing comprehensible texts with minimum corrections to the MT output.

An automatic evaluation of this system was performed using a test set comprising 713 sentences for the PLS part and 949 sentences for the ABS part. Those sentences were extracted from the corresponding machine-translated and post-edited review abstracts.

Results, presented in Table 2, reveal a high level of translation performance according to the automatic metrics used, with a slightly better performance for the ABS section.

We also report a comparison with translations produced by the online Google system<sup>9</sup>, as well as with the translations of the target test set produced by a lower performance system trained only on the WMT'14 medical task parallel data (WMT'14 SMT). This system uses the language models built with the monolingual parts of the WMT'14 medical data and WMT'13 news data. It was tuned using the same post-edited Cochrane data as the Cochrane SMT.

The linear lattice BLEU oracle (LB-4g) was used to estimate the system potential (Sokolov et al., 2012). The atypically low oracle improvements in terms of the automatic metrics scores (+6 H-BLEU, +4 H-METEOR) suggest that the system produces translations that are close to the best translations it can produce given its training data.

Analysis of the HTERp traces confirmed the system performance differences for the PLS and ABS parts (see Table 3). For our experiments, we used the HTERpA configuration (Snover et al., 2009), optimized for human adequacy judgments, with the following components for processing French: the Snowball stemmer (Porter, 2001), and a paraphrase table extracted from the concatenation of

<sup>7</sup><http://statmt.org/wmt14/medical-task>

<sup>8</sup><http://statmt.org/wmt13/translation-task.html>

<sup>9</sup>the version publicly available in Sep. 2015

Metric	Cochrane SMT			WMT'14 SMT			Google SMT		
	ALL	PLS	ABS	ALL	PLS	ABS	ALL	PLS	ABS
H-BLEU↑	57	55	58	29	30	28	49	50	48
Oracle H-BLEU↑	63	62	64	40	41	39	NA	NA	NA
H-METEOR↑	73	72	74	56	55	56	67	67	66
Oracle H-METEOR↑	77	75	78	59	59	58	NA	NA	NA
H-TER↓	30	32	28	58	54	62	36	37	35
Oracle H-TER↓	30	32	28	55	50	59	NA	NA	NA

Table 2: Automatic evaluation results

the Cochrane Reference and PE corpora (Denkowski and Lavie, 2014; Bannard and Callison-Burch, 2005).

	PLS	ABS
HTERp Score ↓	25	25
# Hyp. Tokens	18534	31872
# Ref. Tokens	18502	32438
Operation	% Hyp. Tokens Edited	
Shift	4	5
Match	74	78
Stem match	3	3
Paraphrase	7	6
Substitution	8	7
Deletion	8	6
Edition	% Ref. Tokens Edited	
Insertion	7	7

Table 3: Number of hypothesis/reference tokens (words) aligned by an HTERp operation or a match

The post-edition operations performed to the output translation tend to be non-repetitive: only about 11% of edited tokens/pairs of tokens per operation are unique, but the most frequent post-edition operations (see Table 4) do not exceed 11% of all the changes per operation.

Operation	PLS		ABS	
	Tokens	%	Tokens	%
Stem Match	de → des	11	de → des	11
Paraphrase	les pansements → pansements à base	1	de la même fratrie → frères et sœurs	1
Substitution	les → des	2	, → ;	8
Deletion	de	6	les	5
Insertion	,	4	de	4

Table 4: Most frequent token changes per operation

As shown in Table 5, the most common Part-of-Speech (POS) substitution patterns reveal frequent modifications to nouns (NC) and to POS's that cooccur with them (DET, P, ADJ), potentially forming terms and terminological constructions, as well as grammatical changes to verbs (V (gram)) (Toutanova et al., 2003; Schmid, 1995).

PLS		ABS	
Pattern	%	Pattern	%
P → P	10	P → P	9
NC → NC	7	NC → NC	8
DET → DET	7	PUNC → PUNC	8
DET → P	5	DET → P	6
P → DET	4	DET → DET	4
V → V (gram)	3	ADJ → ADJ	4
ADJ → ADJ	3	P → DET	4
ADJ → NC	3	ADJ → NC	3
VPP → VPP	2	V → V (gram)	2
V → V	2	NC → P	2

Table 5: Most common POS substitution patterns

Such unusually high translation quality scores do not allow us, however, to dispense with a final post-edition step before publication. Also, improving the system to reduce

the post-editor burden remains an important goal. To this end, a fine-grained performance analysis is needed to detect the remaining translation difficulties and to guide future improvements to the system. Further, while analyzing the high-performance MT, we will talk about "residual" errors and difficulties.

#### 4. Diagnosing MT Performance

Since most human evaluation procedures are very costly, MT quality is traditionally measured using reference-based automatic metrics that compute a similarity score between the machine output and one or several human translations (or post-editions) (e.g., (H)BLEU (Papineni et al., 2002), (H)TER (Snover et al., 2006), (H)TER-plus (Snover et al., 2009), (H)METEOR (Denkowski and Lavie, 2014) etc.), which are based on an automatic alignment between words from the machine translation and words from the reference translation. Such alignments are often taken as the basis for an automated error analysis (e.g., (Popovic and Ney, 2011; Berka et al., 2012)). These methods, however, regard the system as a black-box and analyze only its output without any connection to the source text or to the system's specificities.

The trend to take more insight into system internals is observed for Quality Estimation (QE) of MT (Specia et al., 2010; Specia and Giménez, 2010), where most approaches based on Machine Learning techniques take into account both the output, its alignment to the source text, and additional systems scores (Wisniewski et al., 2014; Specia et al., 2015). Irvine *et al.* (2013) go one step further, trying to investigate the interconnection between the source, target and system-dependent characteristics in an attempt to detect domain adaptation errors. An approach of analyzing MT performance in a contrastive manner per linguistic phenomena (e.g., POS) is proposed by Max *et al.* (2010).

Inspired by these latter studies, we propose a new methodology for diagnosing MT performance that should help us to answer the following questions: Which kind of **translation difficulties does a system face?** Are those difficulties related to a greater extent to the **initial corpus quality or to the system scoring procedure?** To the best of our knowledge, this is the first attempt to analyze high-quality SMT by associating residual errors, detected during PE, with source characteristics and system parameters.

Taking into account the observations presented in Table 5, we decided to focus on the translation quality of certain syntactic constituents and POS, in particular noun phrases, as potential complex terminological structures, verbs and nouns (Klein and Manning, 2003).

We extracted the following groups of unique source *n*-grams (units): the ones corresponding to longest noun phrases (NP), then from the rest of the sentence we extracted units corresponding to the neighboring/single verbs (V) and nouns (N). The residual sentence spans of varying length, not covered so far, were put in a separate group (Rest). A sketch of our protocol is provided in Figure 1.

Further, we distinguished the following subordinate groups: the units that are present in the system's phrase table (PT) and also present in the 1-best hypothesis segmentation in

$\geq 80\%$  of their occurrences ( $k_{1-best}$ ); the ones that are present in PT but are absent from the 1-best segmentation in  $\geq 80\%$  of their occurrences ( $k_{pres}$ ); and the units that are absent from the PT ( $k_{abs}$ ).

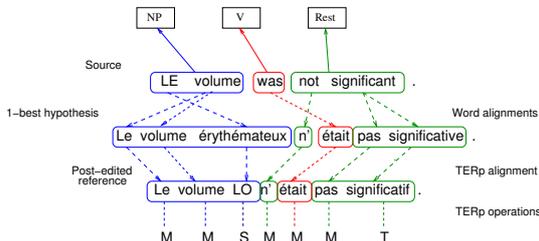


Figure 1: Illustration of our analysis strategy

Using the output word alignments, as well the hypothesis  $\rightarrow$  post-edited reference alignments produced by HTERp, we compute for each unit ( $k_i$ ) the averaged translation quality statistics for all its occurrences ( $t_j$ ), by comparing the aligned hypothesis segment ( $h_s$ ) to its aligned reference segment ( $r_m$ ). Hypothesis  $\rightarrow$  oracle hypothesis and oracle hypothesis  $\rightarrow$  post-edited reference HTERp alignments were used to calculate the averaged oracle translation quality statistics. More precisely, we estimate the following parameters:

1. unit frequency ( $fr$ );
2. unit length in words ( $\# w_k$ );
3. average per occurrence  $t_j$  percentage of the unit hypothesis segment words  $w_h$  aligned to reference segment words  $w_r$  with each TERp operation or a match (e.g., match ( $M$ ), substitution ( $S$ ), stem match ( $T$ ), paraphrase ( $P$ ) etc.), and correspondingly for the oracle hypothesis segment ( $M_O, S_O$  etc.):

$$M = \frac{\# M_{w_h}}{\# w_h} \quad (1)$$

To trace the connection between the system performance and source peculiarities, we calculate the unit *term rate*:

$$\text{term rate} = \frac{\# w_k^t}{\# w_k} \quad (2)$$

where  $w_k^t$  is the words of a unit marked as terms or parts of complex terms.

The term mapping was performed with the Metamap tool for medical texts (UMLS, 2009). Metamap searches were parametrized to avoid mapping to general concepts. A corpus statistics filter was used to further exclude highly frequent words.

Our methodology extends the approach described in (Irvine et al., 2013) and associates target errors with occurrences in the original training corpus. We do so by computing the prior translation entropy ( $H_{prior}$ ) of the distribution of the phrase translation probabilities  $p(\bar{t}|\bar{s})$  of all the possible

target bi-phrases  $\bar{t}$  with  $\bar{s}$  equal to the unit, taken from the PT with lemmatized  $\bar{t}$ :

$$H_{prior} = - \sum_{k=1}^n p_k(\bar{t}|\bar{s}) \log p_k(\bar{t}|\bar{s}) \quad (3)$$

We attempt to correlate the errors with the scoring procedure by measuring the presence of the reference translation in the oracle hypothesis. We extend the analysis of this correlation by computing the average posterior entropy ( $H_{post}$ ) of the normalized distribution of the 1-gram path posterior probabilities  $P(u|\varepsilon)$ , composing a unit.

$$H_{post} = - \sum_{k=1}^n P_k(u|\varepsilon) \log P_k(u|\varepsilon) \quad (4)$$

We calculate 1-gram posterior probabilities  $P(u|\varepsilon)$  from the estimation of path posterior probabilities as defined in (de Gispert et al., 2013):

$$P(u|\varepsilon) = \frac{\sum_{E \in \varepsilon_u} \exp(\alpha H(E, F))}{\sum_{E' \in \varepsilon} \exp(\alpha H(E', F))} \quad (5)$$

where  $\varepsilon$  is the space of translation hypotheses (a 10K-best list was chosen), and  $H(E, F)$  is the score assigned by the model to the sentence pair  $(E, F)$ .

The probabilities of the target bi-phrases  $\bar{t}$  and path posterior probabilities of 1-grams sharing the same lemma were added.

## 5. Evaluation Results

The proposed methodology was applied to the test set presented in Section 3. to analyze the functioning of the Cochrane SMT, as well as the functioning of the less competitive WMT'14 SMT. Examples of the test set sentences demonstrating the translation challenge are provided in Table 6.

During our analysis of residual translation difficulties of the Cochrane SMT, we attempted to find answers to the following questions:

### 1. What are the “worst” translated unit groups for the high-performance system?

We took the average percentage of matches per group  $M$  as an indicator of translation quality (see Figure 2a). We explored the group characteristics by analyzing their general statistics (see Table 7) and the *term rate* (see Figure 2c).

From Figure 2a we can see that the system faces difficulties translating the units of the V group (lowest average  $M \approx 53\%$ ), although the majority of those units are known to the model (97%, *1-best+Pres*, see Table 7).

For the NP group, Figure 2a shows the “worst” translation quality of the units that are absent from the PT ( $M=74\%$ , *Abs*), which need to be translated by composition.

Figure 2c detects the high term concentration for the N group units (average *term rate*=30%). Thus, the “worst” translated units of the N group ( $M=24\%$ , *Abs*) are mainly terms unknown to the model. The high rate of N units that are present in the 1-best segmentation (25%, *1-best*, see Table 7) suggests frequent term translation inconsistency due to lack of context information.

PLS	
Source	A lack of growth and poor nutrition are common in children with chronic diseases like cystic fibrosis and paediatric cancer.
Cochrane SMT	Un manque de la croissance et une mauvaise nutrition sont fréquents chez les enfants atteints de maladies chroniques comme la mucoviscidose et le cancer pédiatrique.
Oracle	'A lack of growth and bad nutrition are common in children suffering from chronic diseases like cystic fibrosis and paediatric cancer.'
PE	Un manque de la croissance et une mauvaise nutrition sont fréquents chez les enfants atteints de maladies chroniques comme la mucoviscidose et les cancers, chez les enfants 'A lack of growth and bad nutrition are common in children suffering from chronic diseases like cystic fibrosis and cancers, in children' Une croissance réduite et une mauvaise nutrition sont fréquents chez les enfants atteints de maladies chroniques comme la mucoviscidose et les cancers pédiatriques. 'A reduced growth and bad nutrition are common in children suffering from chronic diseases like cystic fibrosis and the paediatric cancers.'
ABS	
Source	Poor growth and nutritional status are common in children with chronic diseases.
Cochrane SMT	Une mauvaise croissance et le statut nutritionnel sont fréquents chez les enfants atteints de maladies chroniques.
Oracle	'A bad growth and the nutritional status are common in children suffering from chronic diseases.'
PE	Une mauvaise croissance et le statut nutritionnel sont fréquents chez l'enfant de 'A bad growth and the nutritional status are common in the child of' Une croissance réduite et un mauvais statut nutritionnel sont fréquents chez l'enfant atteint de maladie chronique. 'A reduced growth and a bad nutritional status are common in the child suffering from a chronic disease.'

Table 6: Examples of PLS and ABS test set sentences

The same difficulties are observed for the less competitive WMT'14 SMT: the V group units are the “worst” translated (lowest average  $M \approx 36\%$ ); translation of the NP group units absent from PT is of a low quality ( $M=61\%$ , *Abs*); translation of the term N units present in the 1-best segmentation is often inconsistent ( $M=44\%$ , *1-best*, *term rate=34%*, see Figure 3a, Figure 3c).

NP total : 3528						
	Cochrane SMT			WMT'14 SMT		
	1-best	Pres	Abs	1-best	Pres	Abs
%	10	27	63	9	10	81
# $w_k$	2	3	10	2	2	9
fr	1	1	1	1	1	1
N total : 336						
	Cochrane SMT			WMT'14 SMT		
	1-best	Pres	Abs	1-best	Pres	Abs
%	25	71	4	41	47	12
# $w_k$	1	1	1	1	1	1
fr	1	2	1	1	3	1
V total : 982						
	Cochrane SMT			WMT'14 SMT		
	1-best	Pres	Abs	1-best	Pres	Abs
%	18	79	3	32	62	6
# $w_k$	1	1	2	1	1	2
fr	1	3	1	1	4	1
Rest total : 931						
	Cochrane SMT			WMT'14 SMT		
	1-best	Pres	Abs	1-best	Pres	Abs
%	13	75	12	21	57	22
# $w_k$	2	2	2	1	1	2
fr	1	6	1	1	8	1

Table 7: General statistics per unit group

## 2. To which extent the high-performance system scoring procedure is responsible for the residual errors?

To answer this question we analyzed the per-group differences between system hypotheses and oracle hypotheses match percentage values  $\Delta M$  (see Figures 2a, 2b).

Additionally, to evaluate the scoring procedure we studied the correlation between the low/high match percentage zones (see Figure 2a) and the prior/posterior entropy values (see Figures 4a, 4b). E.g., we can see that the present in the PT (*1-best+Pres*) N group units with the high match percentage (average  $M \approx 73\%$ ) and the V group units with the low match percentage (average  $M \approx 57\%$ ) both correspond to the same average prior entropy value ( $H_{prior} \approx 2$ ), as well as to the absence of significant difference between the average posterior entropy values ( $H_{post} \approx 0.4$  and  $H_{post} = 0.3$  correspondingly).

With the average  $\Delta M$  of about 5%, we can conclude that in the majority of cases the system is unable to produce “correct” translations. The absence of correlation between the

match percentage and prior/posterior entropy values confirms that the scoring procedure is not responsible for most of the errors.

In comparison, the scoring procedure of the WMT'14 SMT can be improved more efficiently. The oracle changes to the WMT'14 SMT output ( $\Delta M$  of about 4%) are more significant since they are performed for more units. From Table 7 and Figures 3a, 3b, we see that the translation of 41% of the *1-best* N group units is improved with  $\Delta M=1\%$  (compare to 25% of N *1-best* units with  $\Delta M=1\%$  for the Cochrane SMT, see Figures 2a, 2b).

For the WMT'14 SMT we should also notice the presence of a more distinct correlation between the translation quality indicator and entropy values: e.g., the high posterior entropy value ( $H_{post} = 0.5$ ) for the *1-best* N units corresponds to the low match percentage ( $M=44\%$ , see Figures 4c, 3a).

## 3. What is the nature of the per-group residual errors?

The manual analysis of the “worst” ( $M \leq 20\%$ ) and “best” ( $M \geq 80\%$ ) translated unit occurrences for the Cochrane SMT within the target groups provides some insight as to the nature of the residual errors (see Table 8).

Confirming our previous observations, the remaining errors of the N and NP groups concern mainly terms unknown to the model (out-of-vocabulary (OOV)), as well as errors in term and professional jargon precision (e.g., Source: “*cardiotoxicity*”, MT: “*cardiotoxicité*” ‘cardiotoxicity’, PE: “*toxicité cardiaque*” ‘cardiac toxicity’, absent from the oracle hypothesis; Source: “*IDA*”, MT: “*une anémie ferriprive*” ‘iron deficiency anemia’, PE: “*l’IDA*” ‘IDA’, absent from the oracle hypothesis).

In the NP group we often face complex terminological constructions translated by composition (e.g., Source: “*people with functioning kidney transplants*”, MT: “*les personnes atteintes de fonctionnement de greffes de rein*” ‘people suffering from functioning of kidney transplants’, PE: “*des receveurs de greffe rénale fonctionnelle*” ‘functional renal transplant recipients’, absent from the oracle hypothesis).

The residual translation errors related to the V group are mostly caused by the specificities of the source language:

1. source syntactic/stylistic peculiarities (very often expletive constructions), requiring restructuring on the target language side (see Table 9);
2. tense and modality (e.g., Source: “*may reduce*”, MT: “*peut réduire*”, Oracle: “*peut réduire*” ‘can reduce’, PE: “*pourrait réduire*” ‘could reduce’).

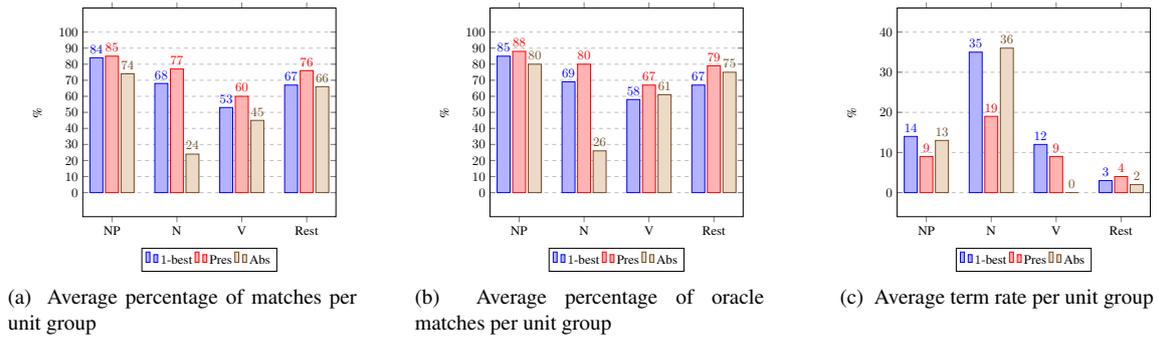


Figure 2: Translation quality statistics for Cochrane SMT

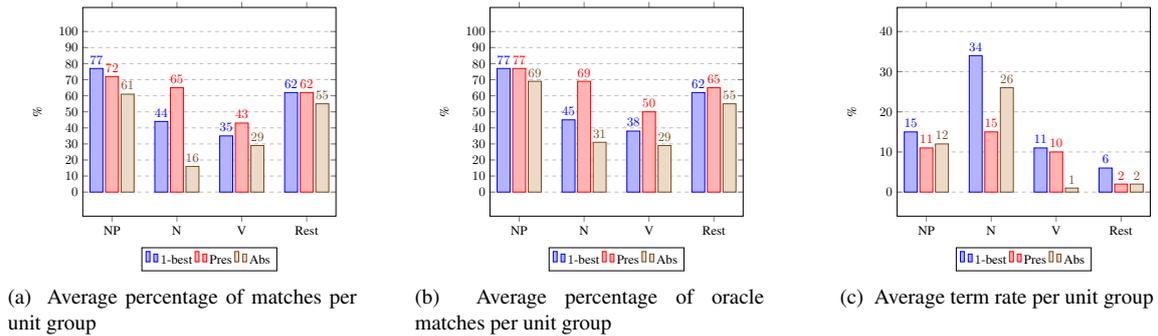


Figure 3: Translation quality statistics for WMT' 14 SMT

	NP				N				V			
	PLS		ABS		PLS		ABS		PLS		ABS	
	Worst	Best										
# total	1641		2495		304		365		1206		1604	
%	5	57	4	64	21	66	16	73	33	58	31	61
#w <sub>g</sub>	2	5	3	7	1	1	1	1	1	1	1	1
fr	3	2	2	2	2	4	3	7	15	23	26	30
M <sub>g</sub> %	0	100	0	100	0	100	0	100	0	100	0	100
M <sub>o</sub> %	34	100	30	100	27	100	25	100	18	100	16	100
term rate,%	11	10	15	10	44	18	28	20	6	6	14	10
H <sub>post</sub>	0.5	0.3	0.4	0.2	0.3	0.3	0.4	0.2	0.4	0.4	0.4	0.4

Table 8: Statistics about the “worst” and “best” translated unit occurrences

We should also notice an increased quantity of paraphrasing corrections performed to the V group (e.g., Source: “we searched all databases”, MT: “nous avons effectué des recherches dans toutes les bases de données” ‘we have performed searches in all the databases’, PE: “nous avons interrogé toutes les bases de données” ‘we have questioned all the databases’, oracle output corresponds to MT). Those rephrasings have a negative impact on the automatic evaluation metrics. The semantic and stylistic necessity of those changes need further investigations.

In comparison, stylistic changes within NP and N groups are quite rare (e.g., Source PLS: “the Canadian Institutes of Health Research”, MT: “la Canadian Institutes of Health Research” ‘the Canadian Institutes of Health Research’, Oracle: “la Canadian Institutes de recherche en santé de recherche” ‘The Canadian Institutes of research in health of research’, PE: “les instituts de recherche en santé du Canada” ‘the institutes of research in health of Canada’).

#### 4. Which kinds of residual errors could be potentially

#### resolved by the high-performance system given its training data?

We also performed a manual analysis of the oracle improvements to the “worst” translated unit occurrences within the target groups ( $\Delta M$  of about 25%, see Table 8). They mostly concern:

1. grammatical errors (change of article or preposition for the N and NP groups, e.g., Source: “with taxanes”, MT: “avec taxane” ‘with taxane’, PE: “avec les taxanes” ‘with the taxanes’, oracle output corresponds to PE; tense changes for the V group, e.g., Source: “were excluded”, MT: “ont été exclues” ‘have been excluded’, PE: “étaient exclues” ‘were excluded’, oracle output corresponds to PE);
2. certain reformulations (e.g., Source: “the trial ... showed a clear benefit”, MT: “l’essai ... a montré un bénéfice clair” ‘the trial ... has shown a clear evidence’, PE: “l’essai ... a mis en évidence un bénéfice clair” ‘the trial ... has highlighted a clear evidence’, oracle output corresponds to PE);
3. some terminological precision errors, including terminological construction translated by composition (e.g., Source: “alternative treatments”, MT: “d’autres traitements” ‘other treatments’, PE: “des traitements alternatifs” ‘alternative treatments’, oracle output corresponds to PE; Source: “wound management properties”, MT: “la prise en charge de la plaie propriétés” ‘the management of the wound any properties’, PE: “les propriétés” ‘the properties’, oracle output corresponds to PE);

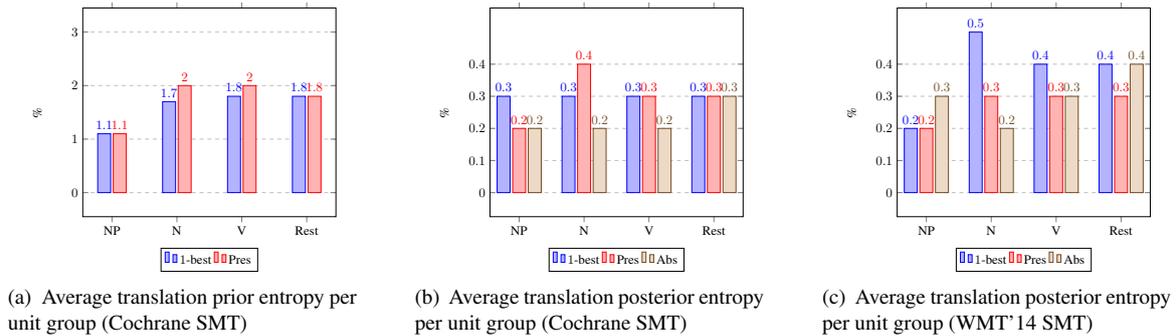


Figure 4: Entropy Estimations

Source	However, the evidence for survival improvement is still lacking.
MT	Cependant, les preuves d'amélioration de la survie est encore manquantes. 'However, the proofs of the improvement of survival is still missing.'
Oracle	Cependant, les preuves d'amélioration de la survie, il manque toujours de la. 'However, the proofs of the improvement of survival, it misses still the.'
PE	Cependant, il manque toujours de données probantes sur l'amélioration de la survie. 'However, it still misses the proving data on the improvement of survival.'

Table 9: Sentence restructuring example

- minor (rarely major) reformulations and restructurings (e.g., Source: "a one-day training course on **how to resuscitate newborn babies**", MT: "un schéma d'évolution de formation **sur la façon de réanimer des nouveau-nés**" 'a scheme of development of training on the way to resuscitate newborns', Oracle: "un schéma d'évolution de formation **sur la réanimation des nouveau-nés**" 'a scheme of development of training on the resuscitation of newborns', PE: "une formation d'un jour **sur la réanimation des nouveau-nés**" 'a training of one day on the resuscitation of newborns').

As a summary, we can enumerate the following main translation difficulties faced by our Cochrane MT system:

- term and professional jargon translation precision;
- translation of complex terminological constructions;
- translation of source-specific syntactic/stylistic constructions requiring target-side reformulation;
- translation of verbs (grammatical/stylistic variant).

We tend to relate those difficulties to the nature of the medical translation task, since they are not specific to the high-performance system. They are caused by the original corpus limitations (absence of the "correct" translation in the training data), as well as to the limitations of SMT in general. Those limitations include the inability to resolve structural differences between languages or to take the more distant context into account.

The indicated issues can be partially solved by *ad hoc* solutions (fine-tuning of the system parameters to improve scoring, model separation to resolve stylistic differences, rewriting of source sentences, etc.), though their final resolution requires professional human knowledge.

## 6. Conclusion

In this article, we have introduced a fine-grained analysis methodology for high-quality narrow-domain SMT, which are typical situations where automatic error metrics prove not informative enough to guide the improvement of systems. Such levels of high performance, however, require adapted solutions.

The novelty of the proposed approach consists in diagnosing high-performance MT by finding an interconnection between residual errors, source phenomena and system parameters, such as original corpus quality and system scoring procedure, and using post-editing traces and Quality Estimation techniques. Thus, this approach provides some necessary hints to better detect translation difficulties and identify their reasons.

It can be used as an effective means to explore a system's potential with the perspective of improving it further.

We have demonstrated the usefulness of such an analysis on the example of the high-quality medical Cochrane SMT system. We found that its residual errors most significantly concern terminology and professional jargon, which are caused by the original corpus limitations, as shown by oracle estimations. The other main difficulty is the syntactic and stylistic peculiarities of the source language, often requiring reformulations on the target side. Those difficulties are related to the nature of the medical translation task and are not specific to the high-performance MT, as confirmed by our comparative study.

The described analysis procedure can be further extended by introducing an algorithm that will make a decision on the translation difficulty of a text given a system. This final decision can be provided as a difficulty score.

We also presented a high-quality English-French parallel corpus of Cochrane systematic review abstracts, which can be used for a variety of NLP tasks. We provided a description of the corpus (human translated and PE parts), as well as the translation challenges related to the genre of medical reviews with its internal stylistic variety (popular scientific vs. scientific style).

## Acknowledgments

The work of the first author is supported by a CIFRE grant from the French ANRT.

## 7. Bibliographical References

- Bannard, C. and Callison-Burch, C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, Ann Arbor, Michigan, June. Association for Computational Linguistics.
- Berka, J., Bojar, O., Fishel, M., Popović, M., and Zeman, D. (2012). Automatic MT error analysis: Hjerson helping Addicter. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2158–2163, Istanbul, Turkey, May.
- Boguraev, B., Manandise, E., and Segal, B. (2015). The bare necessities: Increasing lexical coverage for multiword domain terms with less lexical data. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 60–64, Denver, Colorado, June.
- Bojar, O. (2011). Analyzing error types in English-Czech machine translation. In *The Prague Bulletin of Mathematical Linguistics*, April.
- Costa-jussà, M. R., Farrús, M., and Pons, J. S. (2012). Machine translation in medicine. a quality analysis of statistical machine translation in the medical domain. In *Proceedings of the 1st Virtual International Conference on Advanced Research in Scientific Areas*, pages 1995–1998, December.
- de Gispert, A., Blackwood, G. W., Iglesias, G., and Byrne, W. (2013). N-gram posterior probability confidence measures for statistical machine translation: an empirical study. *Machine Translation*, 27(2):85–114.
- Denkowski, M. and Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.
- Irvine, A., Morgan, J., Carpuat, M., III, H. D., and Munteanu, D. (2013). Measuring machine translation errors in new domains.
- Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June.
- Max, A., Crego, J. M., and Yvon, F. (2010). Contrastive lexical evaluation of machine translation. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. pages 311–318.
- Popovic, M. and Ney, H. (2011). Towards automatic error analysis of machine translation output. *Computational Linguistics*, 37:657–688.
- Porter, M. F. (2001). Snowball: A language for stemming algorithms.
- Schmid, H. (1995). Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop*, pages 47–50.
- Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.
- Snover, M., Madnani, N., Dorr, B. J., and Schwartz, R. (2009). Fluency, adequacy, or HTER?: Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, pages 259–268, Stroudsburg, PA, USA.
- Sokolov, A., Wisniewski, G., and Yvon, F. (2012). Computing lattice BLEU oracle scores for machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 120–129, Avignon, France, April. Association for Computational Linguistics.
- Specia, L. and Giménez, J. (2010). Combining confidence estimation and reference-based metrics for segment-level MT evaluation. In *Ninth Conference of the Association for Machine Translation in the Americas*, AMTA, Denver, Colorado.
- Specia, L., Raj, D., and Turchi, M. (2010). Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50.
- Specia, L., Paetzold, G., and Scarton, C. (2015). Multi-level translation quality prediction with QuEst++. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 115–120, Beijing, China, July.
- Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.
- UMLS. (2009). UMLS reference manual. *Multidisciplinary Information Retrieval*.
- Wang, L., Lu, Y., Wong, D. F., Chao, L. S., Wang, Y., and Oliveira, F. (2014). Combining domain adaptation approaches for medical text translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 254–259, Baltimore, Maryland, USA, June.
- Wisniewski, G., Pécheux, N., Allauzen, A., and Yvon, F. (2014). LIMS submission for WMT'14 QE task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 348–354, Baltimore, Maryland, USA, June.

## Blues for BLEU: Reconsidering the Validity of Reference-Based MT Evaluation

Arle Lommel

E-mail: arle.lommel@gmail.com

### Abstract

This article describes a set of experiments designed to test whether reference-based machine translation evaluation methods (represented by BLEU) (a) measure translation “quality” and (b) whether the scores they generate are reliable as a measure for systems (rather than for particular texts). It considers these questions via three methods. First, it examines the impact of changing reference translations and using them in combination on BLEU scores. Second, it examines the internal consistency of BLEU scores, the extent to which reference-based scores for a part of a text represent the score of the whole. Third, it applies BLEU to human translation to determine whether BLEU can reliably distinguish human translation from MT output. The results of these experiments, conducted on a Chinese>English news corpus with eleven human reference translations, bring the validity of BLEU as a measure of translation quality into question and suggest that the score differences cited in a considerable body of MT literature are likely to be unreliable indicators of system performance due to an *inherent imprecision* in reference-based methods. Although previous research has found that human quality judgments largely correlate with BLEU, this study suggests that the correlation is an artefact of experimental design rather than an indicator of validity.

**Keywords:** machine translation, BLEU, translation evaluation and assessment

### 1. Introduction

Determining the quality of translation is a fraught and complex task, in part due to the lack of any single, widely accepted definition of what “translation quality” is. In the human translation world, quality is generally assessed by bilingual subject-matter experts who review translations to identify errors. This defect-driven approach (exemplified by systems such as the LISA QA Model and SAE J2450) is quite common in translation-production environments, but it is relatively expensive and time-consuming. For MT developers who may need to rapidly test multiple system configurations, the time factor is a significant barrier.

Prior to the early 2000s, evaluation of machine translation (MT) in particular had been largely ad hoc and driven by the needs of particular evaluation tasks. Because human evaluation is expensive and time consuming there was a push to develop automatic methods of evaluation that could automatically provide results. The first prominent automatic method of this type was BLEU (Papineni, 2002). BLEU provided an automatic measure of similarity between a *translation hypothesis* and one or more *reference translations*. The assumption is that the more a translation is like a human reference the more likely to be of a higher quality.

BLEU has since become the most widely used reference-based method for evaluating MT quality. Other reference-based methods, such as METEOR, Word Error Rate (WER), and NIST, have appeared since this time, but BLEU has maintained a prominent role, with many MT-related papers using BLEU-score improvements to evaluate systems and changes.

### 2. Validity: Do Reference-Based Methods Evaluate “Quality”?

Reference-based methods for evaluation provide a mechanically determined score of string similarity between the translation hypothesis (the output from the system to be evaluated) and the reference translation. If the hypothesis contains the same tokens in the same order as the source, it will receive a high score. If the hypothesis contains other tokens or if they appear in a different order it will receive a lower score.

The implicit assumption is that quality can be measured based on similarity to human translation and that a mechanical measure of similarity is adequate to evaluate quality. Thus for reference-based methods, similarity to human output is assumed to be a reliable *proxy* for quality: Researchers recognize that it is not a direct measure.

One consequence of this mechanical model for evaluation is that reference-based models are sensitive to the particular references chosen. For example, consider the following source, reference, and hypothesis:

- Source (Hungarian): Ő a ferfi amit láttam
- Reference (English): That’s the man I saw.
- Hypothesis (English): The chap I saw is him

Three of the hypothesis tokens (*the*, *I*, and *saw*) appear in the reference and the word order is quite different. Consequently this hypothesis would receive a low reference-based score (around 50%), even though it is a good (albeit colloquial) translation of the Hungarian source.

Papineni et al. (2002) recognized the variability of human translation, and so BLEU from the beginning has allowed for the use of multiple reference translations. The expectation was that with multiple references, BLEU would be able to account for the variation of possible human translation and thereby not penalize a hypothesis just because it does not happen to look like a particular reference.

In practice, however, most BLEU scores are calculated against a single reference translation. Researchers justify this practice because Coughlin (2003) and subsequent research found that human adequacy and fluency judgments correlate quite well with the output of BLEU (and other reference-based methods), even when a single reference is used. If BLEU can reliably predict human judgment from a single reference, researchers do not need to solicit multiple human translations.

Stating that BLEU corresponds to human judgment, however, runs into a fundamental issue: judgment about what? BLEU is correlated *not* against the evaluations of professional translators (or even bilinguals) who understand the languages and subject matters under consideration. Coughlin’s experiment (and most subsequent research) relied on the judgment of *monolingual* evaluators, individuals who could evaluate *adequacy* only with respect to a reference translation:

We suggest that when [monolingual] human evaluators are forced to make decisions without sufficient context or domain expertise, they fall back on strategies that are *not unlike determining n-gram precision*. (2003:23, emphasis added)

This finding is not surprising. The human evaluators could base their decisions only on how similar the hypotheses were to the single reference translation: They did not have the linguistic or domain skill to evaluate the hypotheses as translations independent of the particular references they had. Accordingly those translations that were most similar to the reference would be evaluated as the most adequate (i.e., they convey the same information as the reference).

Although Coughlin is very clear that it appears that humans utilized BLEU-like strategies, the claim frequently heard is that BLEU corresponds to human judgment, not that human judgment can correspond to BLEU in a certain (rather artificial) experimental setting.

Understanding the limitations of claims based on monolingual adequacy is important. BLEU and similar methods *are* predictive of human quality assessments, if human quality assessments are determined in a fashion that encourages BLEU-like assessment. The claim then is circular unless it can be confirmed that both *also* correspond to other methods of human evaluation.

Although concerns about BLEU and similar metrics have long been voiced (e.g., Callison-Burch et al., 2006), issues of practicality and professional convention have kept them central to the field and they remain in widespread use. They remain in use because there is no alternative.

Nevertheless, the validity of reference-based methods has yet to be demonstrated. Reference-based methods assume that similarity to a given reference is a valid measure of quality and the tests designed to demonstrate that validity bias the results because they use a similar method with human evaluators who cannot independently evaluate the translations without the references that are under consideration.

If “quality” can be measured as similarity to a reference then reference-based methods evaluate quality. How well those judgments correlate to what bilinguals or translators would understand as “quality” is another issue that has not been fully explored. As will be seen below, however, there are reasons to believe that reference-based methods do not measure what would generally be understood as translation quality.

### 3. Reliability: Are Reference-Based Methods a Reliable Measure of Quality?

Setting aside the issues raised in the last section, let us assume that what BLEU measures is actually “quality.” The next question is whether BLEU is *reliable* in measuring it.

First off, BLEU is not an *absolute* measure of quality. Researchers are well aware that a score of 43.1 for one system, for instance, does not mean that the system performs better than one that obtains a score of 42.0. Instead they use them as a *relative* measure of change with respect to a given set of references. I do not question this usage at one level, but as I will demonstrate, the devil is in the details. Researchers assume that relatively small changes (as little as a quarter point on a 100-point scale) are meaningful. However, this paper demonstrates that such changes are *not* reliable as an indicator of improvement. Larger changes *may* be, but small changes with respect to a particular reference cannot be so easily interpreted.

At one level, BLEU is obviously reliable. Because it is mechanical, for a given set of references and a hypothesis BLEU will always generate the exact same score. When the hypothesis changes the score will perfectly reflect the differences. BLEU does not depend on the judgment of an annotator. This reliability is very attractive to researchers who want a way to measure change. So it is clearly reliable in measuring *something* that has some relationship to translation quality, however complex that relationship may be.

However, at another level, this reliability is highly contingent. If the references change the score will change. If a different portion of an engine’s output is evaluated, even if taken from the *same* source text, the score will change. Reference-based methods evaluate a *particular* text, not an engine. To some extent this problem is unavoidable because it reflects the inconsistency of MT engines: An engine may translate one piece of text very well but perform poorly for another, and a measure of quality *should* distinguish between the two.

But if an engine is working on an internally coherent body of text, the scores for one part should be relatively similar to the scores for another part. If they are not then the scoring method is unreliable as a way to evaluate the engine. This issue is crucial because if researchers find changes smaller than the inherently expected variability of scores those changes are unlikely to be significant and they will be unreliable as a measure of changes in quality. As will be seen, there is also good reason to conclude that BLEU is not a reliable measure, at least at the normal thresholds for significance in MT research literature.

#### 4. Experimental Setup

In the experiment described in this paper, I used the standard multi-bleu.perl script<sup>1</sup> to examine whether BLEU is valid and reliable as a measure of translation quality. To do so I performed three experiments, two of which focused on reliability and one of which focused on validity. The experiments used a Chinese>English corpus containing eleven reference translations, each with 993 segments of news data. It was derived from the *Multi-Translation Chinese Corpus* (Huang et al. 2002). To prepare it I tokenized the text, changed the encoding to UTF-8, and checked alignment.

(In addition, I performed smaller-scale experiments with English>German translations taken from the QTLep project and text&form, a Berlin-based language service provider. The results from these experiments are not included here but correlated well with the larger Chinese>English corpus study.)

The three experiments examined the following:

- (1) the impact of using multiple combinations of references on BLEU score
- (2) the internal consistency of BLEU scores for translations
- (3) how well BLEU serves to evaluate human translations.

<sup>1</sup> <https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl>

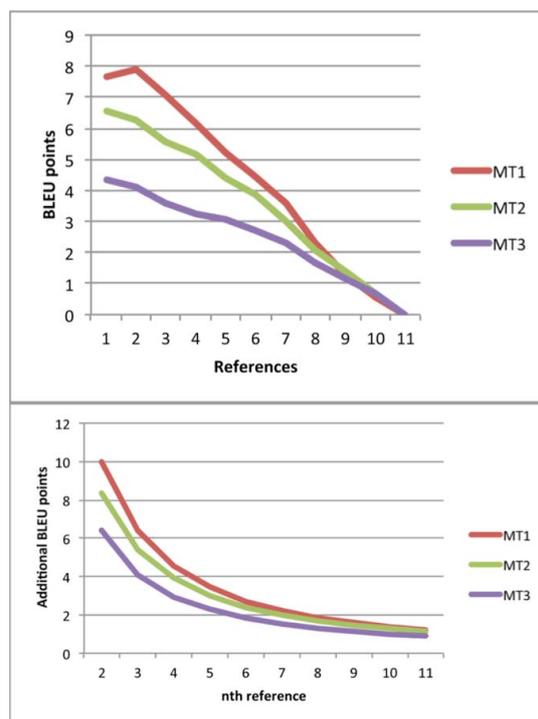


Figure 1. Range of BLEU (max – min) scores for  $n$  references (top) and increases in average BLEU score from adding the  $n$ th reference (bottom)

The first two experiments address reliability. The third address validity.

The hypotheses used in these experiments were generated using two online SMT systems and one online RbMT system. (The smaller-scale English>German test had used five systems—two SMT, two RbMT, and one hybrid, but not all of these systems covered Chinese>English.)

The particular setup and results for each experiment are described below.

#### 5. Impact of Multiple References

In this experiment, I compared the hypotheses against every possible combination of the available references, from a single reference up to all possible references. To make this comparison I created a shell script that ran through all the combinations and fed them into multi-bleu.perl and recorded the resulting BLEU scores. Based on the results I then calculated the range of scores for a given number of references and the average impact of adding the  $n$ th reference.

As can be seen, adding additional references results in a substantial BLEU score increase. The average increase for the  $n$ th reference and the span in scores for  $n$  references are shown in **Figure 1**. The actual scores

No. Refs	MT1 (online SMT)				MT2 (online SMT)				MT3 (online RbMT)			
	Avg.	Range	Sdev.	Diff	Avg.	Range	Sdev.	Diff	Avg.	Range	Sdev.	Diff.
1	18.11	7.64	2.39		14.89	6.55	1.92		12.38	4.34	1.45	
2	28.13	7.88	1.88	10.01	23.27	6.27	1.50	8.38	18.81	4.10	1.04	6.44
3	34.55	7.06	1.50	6.42	28.69	5.56	1.20	5.43	22.91	3.60	0.81	4.10
4	39.11	6.17	1.20	4.56	32.61	5.17	0.97	3.91	25.86	3.23	0.67	2.94
5	42.55	5.20	0.96	3.44	35.62	4.41	0.80	3.01	28.12	3.04	0.57	2.27
6	45.27	4.45	0.76	2.72	38.03	3.90	0.66	2.41	29.95	2.73	0.49	1.83
7	47.47	3.57	0.60	2.21	40.03	3.01	0.55	2.00	31.47	2.30	0.42	1.52
8	49.32	2.30	0.47	1.84	41.72	2.08	0.44	1.69	32.76	1.65	0.35	1.29
9	50.88	1.27	0.35	1.57	43.18	1.35	0.34	1.46	33.87	1.16	0.28	1.12
10	52.24	0.58	0.23	1.35	44.46	0.67	0.24	1.28	34.85	0.66	0.21	0.98
11	53.42			1.18	45.60			1.14	35.73			0.88

Table 1. Average BLEU scores by number of references used, with range of scores (highest – lowest scores), standard deviation of scores, and difference from adding the *n*th reference.

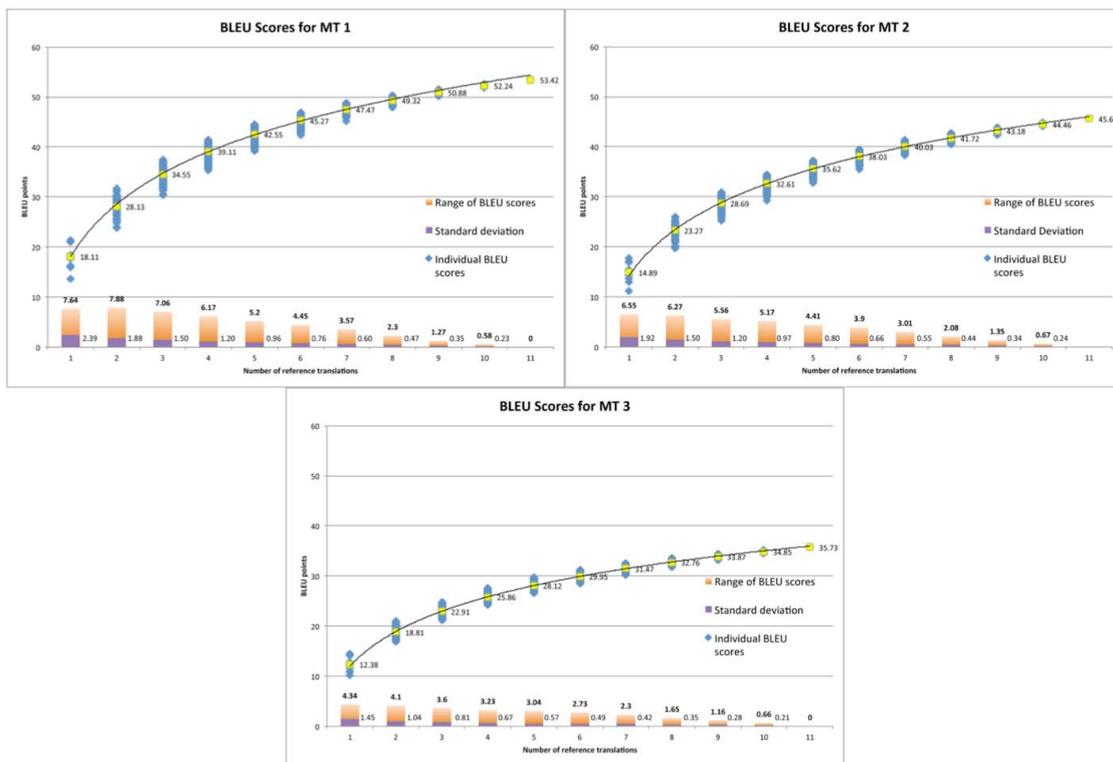


Figure 2. Impact of adding multiple references on various online MT systems (MT1 and MT2: SMT; MT3: RbMT)

appear in **Figure 2** and a summary of the data behind these graphs is in **Table 1**.

The magnitude of the score change that comes from adding each additional reference is striking. Adding a second reference increase average BLEU score across all systems by 8.28 points. Adding each additional reference provides a declining increase, but adding the 11th reference still provides an average BLEU score increase across all systems of 1.07 points.

On the one hand, the substantial increases here are particularly troubling for BLEU. System changes that show an increase of .25 points are often considered worthy of publication. Adding references translations, even when there are already more references than any normal evaluation would use, creates a larger score increase than most system changes. It is thus not an exaggeration to say that the fastest route to BLEU score improvements would be to simply use more references.

This finding means that specifying a “BLEU score” without also specifying the number of references results in a meaningless figure. BLEU can only be interpreted with respect to the number of references used.

This dependence of the score on the number of references should not be surprising since multiple references provide a larger number of potential targets for that the hypothesis can match against. Although the magnitude of the increase is striking, one response would be that as long as the same number of references is used across calculation, BLEU scores should still correlate with human evaluation.

Unfortunately, however, when the range of scores in **Figure 1** is considered, it is also apparent that BLEU scores are highly dependent on *which* references are used: A translation that would score very highly against one set of references might score poorly against another set. The problem is particularly acute when the number of references is small because it is impossible to know where in the spread of potential scores a particular score stands. Consider that for MT1, there is a range of 7.88 BLEU points between the highest- and lowest-scoring combinations of two references. If one system’s output happens to achieve a low score against a particular reference (or even two or three) while another system’s output happens to score well, a researcher might conclude that the first system outperforms the second system, even though a different set of references could produce a different result.

The average standard deviation in BLEU scores for all three systems tested in this experiment when using a single reference is 1.92. This result indicates that BLEU increases of less than this number cannot be considered significant increases for determining how systems perform in general. (They can be considered significant with respect to a particular set of references, but as this experiment shows, significance against particular references does not demonstrate real-world performance increase.)

The results of this study call the reliability of reference-based methods into question, at least when a small number of references are available. Reliability increases with the number of references, but even with 11 references the inherent imprecision in BLEU is larger than the effects observed in many MT experiments.

One possible response to this criticism is that BLEU can be used for ranking systems with respect to one another. Certainly, the averages shown in **Figures 1** and **2** would rank the systems in a certain order in every case (whether that order reflects their quality is, as noted above, very much in question). However, note that the spread of scores for MT 1 and MT 2 overlap for up to three references: for a single reference in 8 of the 11 cases MT 2 scores higher than the minimum for MT 1; for two

references MT 2 outscores the minimum for MT 1 in 23 of 54 cases; and for three references in 14 of 165 cases. Because it is impossible to know where particular scores fall in the possible range, close rankings (where they differ by 1 or 2 points) may reflect chance rather than actual performance differences.

## 6. Internal Consistency

In the second experiment, I wrote a script that took a random slice of one half of the segments of each hypothesis and calculated the BLEU score for that half versus the remaining half. I then found the difference between the scores to see how different the two halves were. Because the segments in each half were selected at random they should not be biased by the particular news sources used in particular portions of the corpus.

I repeated this procedure 10,000 times for two online SMT engines. This test allowed me to see how consistent BLEU was in ranking a particular engine’s output (versus in ranking the complete translation). If BLEU is really providing a quality evaluation of the engine rather than the particular translation, the difference in scores between the halves should be quite low; by contrast, if the difference is high, it indicates either that (a) BLEU is limited for evaluating engines rather than particular outputs, or (b) the engines are very inconsistent in their output.

**Figure 3** (overleaf) shows the results categorized into bands of 0.1 BLEU point difference. The red column marks the average and the orange the difference within the standard deviation.

For MT1 the average difference between the halves was 1.92 BLEU points (standard deviation = .50) with a minimum difference of 0.24 and a maximum of 3.95.

For MT2 the curve is rather different. It is much more likely to show absolute differences closer to 0 than was MT1. At the same time the range of difference was considerably greater. The average difference was 0.97 (with a standard deviation of = .78). The minimum of 0.00 and a maximum of 5.00.

These results suggest that reference-based scores are actually *not* terribly consistent at evaluating system performance. Instead they evaluate a particular set of strings consistently, but selecting a different set of strings for evaluation, *even from the same corpus*, can result in substantial changes in BLEU score. This finding is important when evaluating changes in scores that result from system tweaks: A tweak that results in a relatively large positive change (e.g., a full BLEU point) for one text might result in a negative change for another text, even if taken from the same corpus with a reference from the same translator. Larger corpora should reduce the variability, but would not eliminate it.

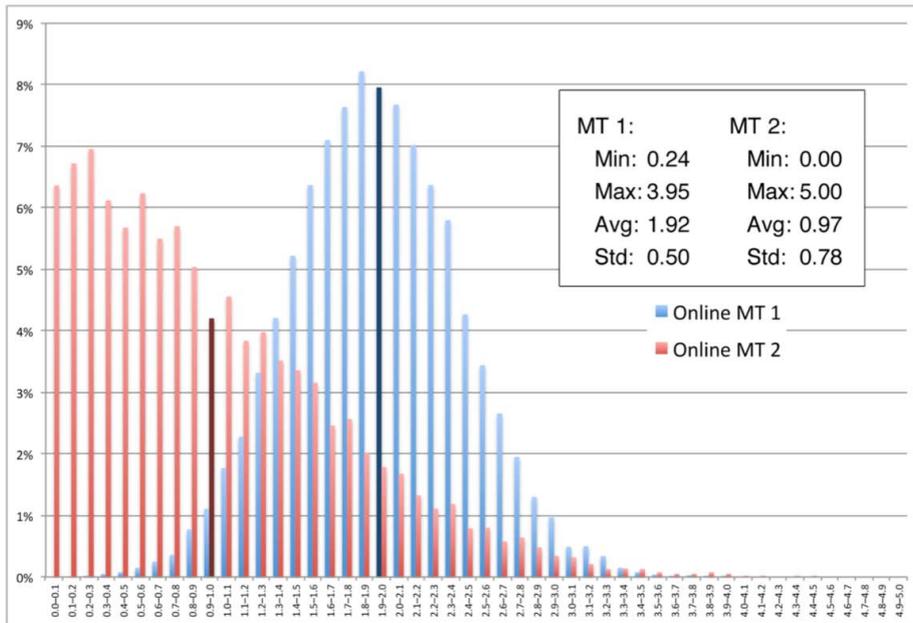


Figure 3. Comparison of difference in BLEU scores between random halves of corpus using 11 references and 10,000 repetitions. Y axis values are the percentage of results in each band in the x axis. Darker columns mark the average for each system.

This result also shows that some engines are more consistent than others (at least in terms of BLEU). MT2 was more consistent (by almost a full BLEU points) on average than MT1. We see that an absolute change in BLEU score of less than roughly 2.4 (the average + the standard deviation) for MT1 and 1.8 (for MT2) cannot reliably reflect system change because it is within the inherent “noise” of the system with respect to BLEU. A smaller change in BLEU score indicates that the translation *changed*, but cannot indicate with any certainty whether the change is an actual systematic change in system performance absent additional evidence.

As with the previous experiment, this result suggests that reliability, particularly for changes of BLEU score of less than 2 points, is a major concern and that changes evaluate particular texts rather than particular engines. (However, larger differences for the same text between engines are likely to be significant, so these results do not suggest that BLEU is inherently useless for comparing engines, but rather that its precision and reliability for low values is limited.)

### 7. Evaluating Human Translation

The final experiment was designed to test validity. It replicated the first experiment but used each of the human reference translations as a hypothesis, treating them in the same way MT would normally be evaluated. Accordingly each professional human translation was compared against possible all combinations of *n* references (*n*=1 to 10, with a maximum of 10 because one reference was always set aside for testing).

Figure 4 (overleaf) shows the results for one of the reference sets (due to space constraints, the results for only one translation can be shown).

Not surprisingly, the curves shown look very similar to those from the first experiment. What is surprising, however, is that the BLEU scores are so low: the maximum BLEU score for any of the translations against a single reference was 25.23, a BLEU score that would normally indicate relatively poor performance. In other words BLEU scores seem to indicate that human translations are worse than many MT systems’ output.

And in fact, if we compare the results of the first experiment with this experiment, we find that, in fact, the BLEU scores do seem to indicate that MT may be better than human translation. As shown in Figure 5 (overleaf), one of the SMT systems (MT1) out-performed eight of the eleven reference translations in terms of BLEU and the other (MT2) outperformed one of them. Only the RbMT system (MT3) fell below the human references in each case.

If BLEU determines translation quality, the developer of MT1 could say that it has created a system that outperforms human translators 73% of the time and we would have to conclude that one of the professional human translators just barely managed to exceed the quality of the lowest-performing MT system.

This conclusion is clearly nonsense (as even a casual perusal of the MT hypotheses demonstrates). Rather, it demonstrates that BLEU scores, no matter how well they

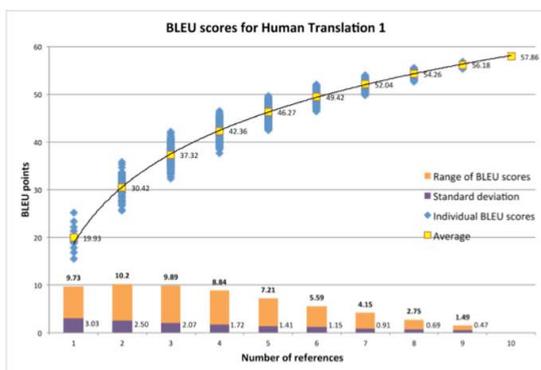


Figure 4. BLEU scores for one human translation compared against the others as references.

correlate to the judgment of monolingual evaluators comparing MT output to a reference translation, are not determining quality in a sense that is meaningful for comparison with human translation. If a human reference translation, which is considered the gold standard in MT evaluation, can score as low as 11.25 (the lowest score for a human translation against one other reference), then whatever BLEU may be evaluating, it is not useful for determining how MT will perform in any circumstances outside of experimental conditions.

## 8. Does Human Evaluation Perform Any Better?

If BLEU cannot reliably measure something that can be reasonably understood to be “quality,” what is the alternative? Human evaluators have their own major problems with reliability. In the QTLaunchPad (<http://qt21.eu/launchpad>) and QT21 (<http://qt21.eu>) projects, investigators found major disagreement in the number and type of errors. Human reviewers frequently disagree with each other about how good particular translations are. They are inconsistent with themselves from one day to the next. So reliability problems are hardly unique to reference-based methods.

This paper is not meant to suggest that human evaluation can replace BLEU. It is expensive and inconsistent. Although projects like QT21 are trying to learn from human annotation, relatively little work has been completed in this area and we do not yet know how well MT can learn from human annotation and scoring. So I certainly do not mean to indicate that all is gloom and doom for reference-based methods and sunshine and butterflies for human evaluation.

Part of the problem is that reference-based methods provide consistent and seemingly precise scores. While MT researchers are aware of the limits of reference-based assessment, they do not always convey this awareness. When they present a 0.5 BLEU-point increase as significant, others may interpret the research as being more precise and reliable than it is. If

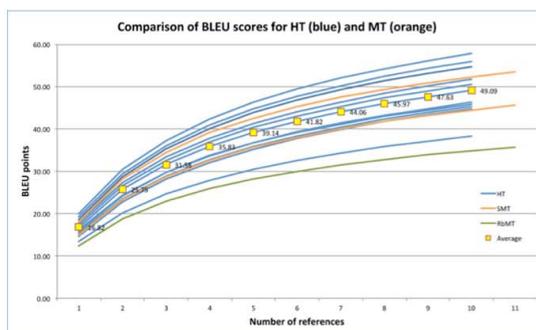


Figure 5. Comparison of BLEU scores for HT versus MT.

researchers claimed that a 0.5 point difference (on a 100 point scale) as determined by human translators were significant, the problem would be clear: Humans are simply not that precise. But, as shown here, reference-based methods are not that precise either, even if they appear to be because that can repeatedly generate the same result.

## 9. What Are the Alternatives?

If reference-based methods are problematic, what is the alternative? Unfortunately there are no good alternatives. This paper points out severe limitations in how MT researchers use and understand reference-based scores, but it cannot suggest a replacement.

However, it does suggest some ways to improve how researchers use BLEU:

1. **Use multiple references.** Single references are too variable and misleading. While using 11 references for most research is impractical, using the average of three or four would improve reliability and would prevent situations in which one system happens to perform well against a particular reference and another does not from skewing results.
2. **Do not over-interpret small differences.** MT researchers should take care not to convey the idea that BLEU values have a precision of less than a few points. Touting 0.5 or even 1.0 point score increases as significant simply overstates what BLEU can actually tell us.
3. **Use multiple texts and generate independent scores for them.** If score differences are consistent across multiple texts (and not just as an aggregate) that will indicate consistent performance and help increase confidence that score differences matter.

Ultimately, however, what we need is a better understanding of what translation quality is and how to understand and measure it. Until we have a sound

theoretical understanding of what we are trying to measure, it is likely that any alternative measures will fall short as well.

## 10. Conclusion

The findings detailed in this paper should be troubling. They call into question the significance of a considerable body of MT research that relies on the use of small differences in BLEU (or other reference-based) scores to demonstrate system improvement or to compare systems. Because BLEU is simply a measure of string similarity to a particular reference, it is not evaluating “quality” in any sense that really corresponds to human understanding (even if we see some correlation in judgments). If changing the reference or adding references can change scores so dramatically, then the scores are too sensitive to input. If an increase of two or even three BLEU points falls within the inherent noise of BLEU, then BLEU is unreliable as a measure of quality as it is used today.

These results also suggest that the oft-cited correlation between human judgment and scores from reference-based systems is epiphenomenal to the experimental setup used to measure such correlation. As Coughlin herself suggested in her seminal paper, the correlation in judgment is likely due to the monolingual reviewers considering whether the same words seems to be found in the translation hypothesis and the reference. Because they cannot evaluate the hypothesis on its own terms, they are effectively recreating the judgments of the reference-based approach and their results correlate to BLEU rather than BLEU correlating to any real understanding of translation quality.

These results call into question fundamental approaches in MT development and should be replicated rather than relied upon as is. Unfortunately, there are few corpora with sufficient numbers of references to be used in such studies. Generating reference translations is expensive. Nevertheless, if MT research is to claim that its methods for evaluating quality are valid and reliable, they must be rigorously tested and underlying assumptions must be questioned. The results of these experiments suggest strongly that this bar has not yet been met.

Can BLEU and other similar methods be used to produce valid results that withstand scrutiny? My results suggest that they can be used if the magnitude of change exceeds that of their inherent. An increase of 10 BLEU points, for example, would almost certainly indicate a real quality improvement in MT output. However, given that a change of 0.5 is within the standard deviation for seven reference translations and a change of 2.0 is within the standard deviation for a single reference, it is clear that changes smaller than a few points, no matter what  $p$  value is obtained, are not likely to represent real changes in how humans will perceive quality.

Adding additional references helps as well by reducing the likelihood that a change is relative only to a single reference translation. If three or four references are used and a system shows an improvement against each of them individually and in combination, it is likely to represent a real change. But in most cases a system change that shows a score increase with respect to certain references would show a decrease with respect to others. We do not yet have the tools to interpret such results in order to tell if apparent changes are meaningful.

I would also suggest that approaches that combine the judgment of professional human translators with machine evaluation are the only way to be certain about the meaning of changes. Such approaches are being pioneered in the QT21 project, but there is considerable scepticism about their value and utility among researchers. Because human evaluation is time-consuming (and noisy in its own right) researchers have sought more consistent and practical methods. But consistency and practicality are not enough if validity and reliability cannot be demonstrated. We do not know what shape evaluation will take in the future, but it is clear that reference-based methods on their own provide us imprecise and at-times misleading guidance.

## 11. Acknowledgements

Early stages of this research were funded by the QT21 (<http://qt21.eu>) project, which has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 645452. The conclusions of this research, however, are the author’s own and do not represent the opinions or positions of the QT21 consortium.

Special thanks for Maja Popović (Humboldt Universität zur Berlin), who ran early versions of the multi-reference experiments, and to Aljoscha Burchardt, Kim Harris, and Hans Uszkoreit (DFKI Berlin) for their valuable feedback. Any errors are, of course, my own.

## 12. Bibliographical References

- Callison-Burch, C., Osborne, M., Koehn, P. (2006). Re-evaluating the Role of BLEU in Machine Translation Research. *EACL 2006*, pp. 249–256
- Coughlin, D. (2003). Correlating Automated and Human Assessments of Machine Translation Quality. *MT Summit IX*, pp. 23–27.
- Huang, S., Graff, D., Doddington, G. (2002). *Multiple-translation Chinese corpus*. Linguistic Data Consortium.
- Papineni, K., Roukos, S., Ward, T., Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. *ACL-2002*, pp. 311–318.

# Can Quality Metrics Become the Drivers of Machine Translation Uptake? An Industry Perspective.

**Katrin Marheinecke**

text&form

Berlin

E-mail: [katrin\\_marheinecke@textform.com](mailto:katrin_marheinecke@textform.com)

## Abstract

Language service providers (LSPs) who want to make use of Machine Translation (MT) have to fight on several fronts. The skepticism within the language industry is still very high: End-customers worry about paying too much for translations that no human has interfered with. Translators refuse to get involved in post-editing activities because they fear that MT will take away their actual work, rendering themselves useless on the long run. And competitors try to outperform each other by either spoiling the market with low-quality MT offered on dumping rates or declining MT altogether for being inappropriate for commercial usage. This paper seeks to show that well-defined quality metrics can help all stakeholders of the translation market to specify adequate benchmarks for the desired translation quality, to use an agreed-upon consistent mark-up and to evaluate translation quality – MT and human translation output alike – accordingly. As a by-product, this professionally error-annotated MT output will help researchers to further improve MT quality, which in turn will help to make this technology more popular in the industry.

**Keywords:** Machine Translation, Translation Quality, Quality Evaluation, Benchmarking, Tool Integration

## 1. Introduction

The discussion on the usage of MT in daily life-scenarios of the translation business is still highly controversial. Whereas some sectors (e.g. the software sector) apply MT already for quite a while in their standard localization workflows, others tend to think that MT is not fit to end-clients' requirements.

Among the group of the second important stakeholder in the translation market – the human translators – the fear of abolishing their own jobs by optimizing MT output through offering post-editing services is prevalent. The pool of adequately trained post-editors is very limited and the actual qualification profile for post-editors remains blurred.

Language Service Providers (LSPs) themselves, have to cope with tough and narrow margins. Some in the business help themselves by offering low-quality MT at rock-bottom prices without bothering about systematic post-editing. This again lowers the esteem of MT in general in the sector as a whole as well as among end-customers.

But what do we mean when talking about quality at all? How can quality be measured? And how can the parties involved contribute to reach the goal of higher MT quality? This paper seeks to provide answers to these questions.

## 2. Obstacles LSPs are Facing when Using MT

Although significant progress in field of MT within the language service industry is to be observed, many LSPs still resist from the broad usage of MT in their daily routines. This is for various reasons.

### 2.1 Lack of Understanding

Since only larger LSPs can afford to run professional MT divisions in their companies, most LSPs have to rely on ready-made MT solutions that are on the market. Often

enough this means using untrained translation services that have no relevance to the domain and the end-users' area of application. To many translation providers, these solutions equal black-boxes that deliver output they can neither affect nor properly evaluate. Not having universal evaluation criteria in place, LSPs are often forced to have a human editor proof-read the complete material without really knowing what kind of issues to focus on. This makes an objective quality assessment difficult, if not impossible. What is worse, the steps are performed in different tools, breaking the commonly applied translation workflows and causing additional manual pre- and post-production steps.

Taking the above said into account, for many smaller LSPs the use of MT seems inefficient and costly, instead of saving them time and money.

### 2.2 Common criteria

What would help to raise acceptance for MT in the translation business, therefore, is a better understanding of the processes and of the anticipated output. If the providers and the requesters of MT were on common grounds concerning evaluation criteria both parties would benefit – the requesters would know what to expect and the providers where there is room for optimization in their MT results. The goal should be that both sides worked together more seamlessly, using the same vocabulary for common quality issue types in order to optimize MT engines accordingly, and thus resulting into improvements from translation to translation.

## 3. Quality Metrics: Why and How

The problem of how to evaluate the quality of translations is not new to the language service industry. Also for human translation, the question as to whether a translation is good or bad and by what this finding can be measured has been a matter of dispute as long as the professional translation sector exists. Despite many and improved ways of computer-aided checking methods the so-called "Four-eyes

principle” is still the method of choice. Even the latest version of the industry standard – the ISO 17100<sup>1</sup> – does not accept any other means of quality revision.

### 3.1 Learning from evaluation of HT

Evaluation performed by humans always harbors the risks of subjectivity and inconsistency. We cannot abandon these risks completely. By defining clear principles for error classes and by categorizing errors accordingly, these risks can be minimized significantly, though. The same goes for the evaluation of MT output. Therefore, a dedicated metrics system is key to a controlled quality assessment for both MT and HT.

### 3.2 Relevant metrics

Using some kind of metrics for the quality estimation is not new to the industry, either. There have been several approaches to the compilation of error scorecards to support objective human quality evaluation models (LISA QA<sup>2</sup>, SA J2450<sup>3</sup>). The problem with these approaches was, though, that they were either restricted to one domain (SAE J2450), or that they followed a “one-size-fits-all” approach (LISA QA and its predecessors).

What was missing for a very long time was an approach that allowed us to compile domain- or even end-customer specific error profiles and – based on those profiles – standards that have to be reached in order to rate a given translation as acceptable.

What acceptable quality means for different environments must be defined by the industries and businesses themselves. That means that the industries or even companies must specify for their textual domains which error classes and categories are relevant in their respective use cases. Only upon these specifications error categorization and annotation can be performed. This is a distinction that automatic evaluation scores obviously cannot deliver.

### 3.3 Industry’s requirements towards MT

For human translation, a translation job that is rated as unacceptable will be returned to its producer in order to have it fixed. Post-editing will (under usual circumstances) not be performed on translated material that is rendered deficient in many ways. The same principle goes for MT output: If the MT output is too far away from what is needed for a given translation scenario a human translation from scratch will be performed faster than the post-editing of a machine-translated text. In other words: In such a scenario, the usage of MT for a translation company is economically nonsense. An LSP will not incorporate such a workflow on the long run. The judgement as to whether a given translation serves its actual market purposes cannot be performed by the means of automatic assessment scores but only by human specialists who have linguistic and domain-specific knowledge.

Another argument that is applied in human translation revision scenarios must be taken into consideration in the

<sup>1</sup> See [http://www.iso.org/iso/catalogue\\_detail.htm?csnumber=59149](http://www.iso.org/iso/catalogue_detail.htm?csnumber=59149)

<sup>2</sup> See <https://www.w3.org/International/O-LISA.html>

<sup>3</sup> See <http://www.sae.org/standardsdev/j2450p1.htm>

MT context: After reviewing and – if necessary – reworking a translator’s work it is common practice to provide them with feedback on what they delivered. For the next assignment, the LSP will expect not to find the same kinds of errors again in the translator’s work. If they do – maybe repeatedly – it is very likely that the LSP will terminate the cooperation with this translator in the near future for obvious reasons: The translator seems unwilling or incapable to *learn*.

The same requirement is valid for MT output. If an LSP has to correct the same error types again and again in every MT workflow it is probably not worthwhile using it. Just as with the human translator, the LSP would expect the machine to learn from its previous mistakes i.e. have the MT engineers fixed what went wrong during the last translation circle.

## 4. MQM: A Recap

Before specifying categories for a quality metrics system we must define what we mean by “quality”.

### 4.1 The Idea of Quality

The underlying quality definition stated by the originators of Multidimensional Quality Metrics (MQM)<sup>4</sup> assumes that a quality translation “demonstrates required accuracy and fluency for the audience and purpose and complies with all other negotiated specifications, taking into account end-user needs” (Koby and Melby 2013). What is important about this definition (and what sets it apart from other translation quality assessment theories) is that the end-users, applying the metrics determine the relevance of a given category, rather than the metrics itself.

### 4.2 The MQM hierarchy

The complete MQM master lists all issue types that different existing metrics models contain and results in a comprehensive but rather confusing hierarchy:

<sup>4</sup> See <http://qt21.eu/mqm-definition>

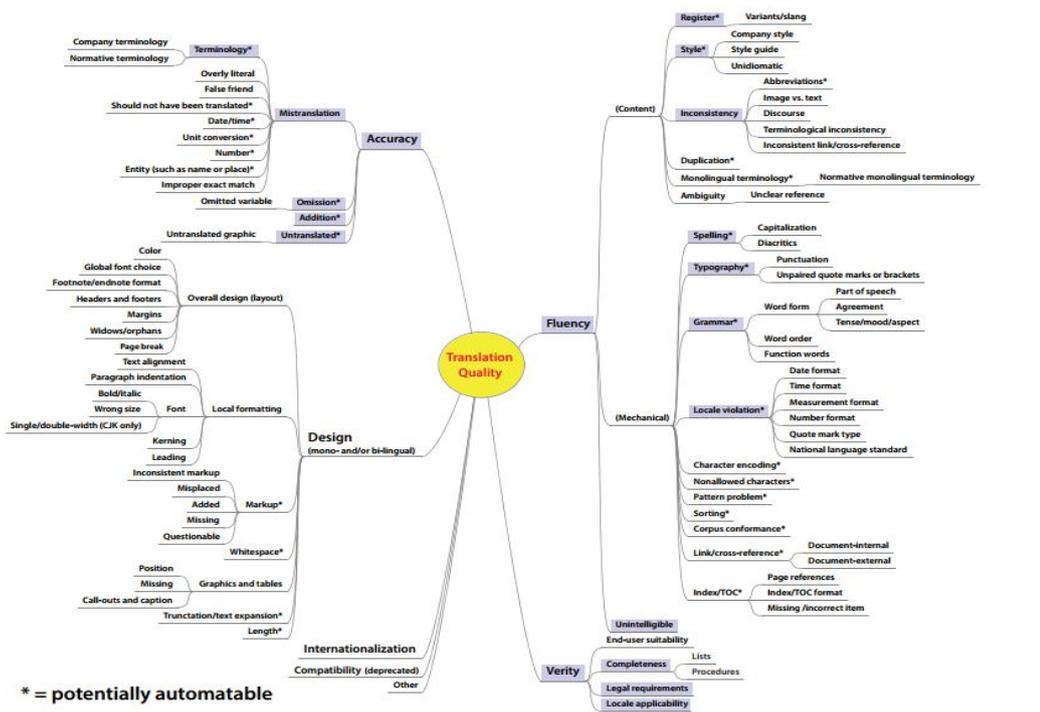


Figure 1 MQM Categories (full)

For actual post-editing purposes this hierarchy would be impossible to handle. What differentiates the MQM model from other approaches, though, is the fact that only those categories will be considered that are needed in the translation scenario in place. Applying the concept of the end users' requirements, requesters and providers can agree upon their relevant set of categories beforehand, making sure that only issue types are taken into account that matter to a given translation context. The MQM core (see figure 2) consists of 21 more commonly addressed issue types:

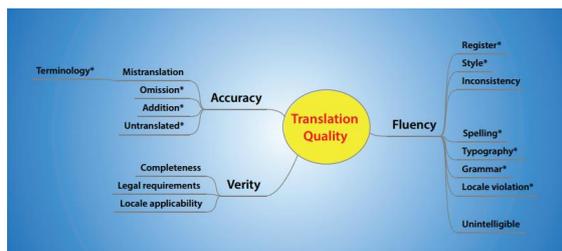


Figure 2: MQM Categories (core)

### 4.3 Learnings from MQM annotation

The MQM system applied to error annotation of MT output serves the evaluation on several levels. For LSPs it makes error types countable and allows to classify them. It not merely counts errors, though, but it enables industries to choose only those error categories that are relevant to them. These can obviously vary from domain and from

text type to text type. By that, certain error profiles for various use cases and sectors can be compiled. Moreover, errors can be weighted according to their severity in a given context. An error that does not affect the overall quality of a text in one domain can be a show-stopper in the other. For example, the usage of the curly quotation marks (“ ”) instead of straight ones (" ") do not affect the text quality of a technical documentation whereas in a software localization project where straight quotation marks function as a marker for UI options curly quotation marks can break the software strings and ruin the whole translation.

Based on the relevant categories industries can define benchmarks that function as a delimiter for different quality levels. If an annotated test sample of a given output falls below the defined threshold then the translation is not suitable for an MT + post-editing workflow without making improvements on the engines that produced the output.

For MT engineers MQM helps to understand where their engine does well and where it fails. Whereas that may be true for automatic evaluation methods, this information alone leaves the improvement of engines to a large extent to the field of trial and error.

For an interpretation on why the engine fails in certain contexts and which patterns these errors show, a more detailed analysis will be necessary. This analysis (see figure 3) can be performed only by a trained linguist who has a deep understanding of both source and target language. The reliability and thoroughness of human annotation compensates for the higher effort compared to an automatic evaluation method.

Annotations			
bba/2012/12/01/303584-27_en_1			
Tucked away under motorway flyovers, it is in many ways a suburb of Mexico City itself.			
49 (unknown)	Versteckt unter der Autobahn flyovers, es ist in vielerlei Hinsicht ein Vorort von Mexiko-Stadt selbst.		3-4
AB	[Annotated] Versteckt unter der [[1] Autobahn flyovers][[2] ] [[3] es] ist in vielerlei Hinsicht ein Vorort von Mexiko-Stadt selbst. [Post-edit] Versteckt unter der Autobahnüberführung ist es in vielerlei Hinsicht ein Vorort von Mexiko-Stadt selbst.	3	1. <b>Mistranslation</b> [Autobahn flyovers] 2. <b>Typography</b> [ ] 3. <b>Word order</b> [es]
KH	[Annotated] Versteckt unter der Autobahn [[1] flyovers][[2] ] [[3] es] ist in vielerlei Hinsicht ein Vorort von Mexiko-Stadt selbst.	3	1. <b>Untranslated</b> [flyovers] 2. <b>Addition</b> [ ] 3. <b>Word order</b> [es]
KM	[Annotated] Versteckt unter der Autobahn [[1] flyovers][[2] ] [[3] es ist] in vielerlei Hinsicht ein Vorort von Mexiko-Stadt [[4] selbst]. [Post-edit] Versteckt unter den Autobahnüberführungen ist es in vielerlei Hinsicht ein Vorort von Mexiko-Stadt.	4	1. <b>Untranslated</b> [flyovers] 2. <b>Typography</b> [ ] 3. <b>Word order</b> [es ist] 4. <b>Extraneous function word</b> [selbst]

Figure 3: SMT translated segment with 3 annotation and post-editing varieties.

## 5. Easing the Collaboration

During the last years, not only the lack of common metrics and standards impeded a collaborative approach between language service business and language research. In order to make a cooperation between both parties work they also need work environments and tools that integrate well into the processes in place.

### 5.1 Tools and Applications

Since the times of LISA QA much has happened. Whereas in those days QA professionals had to fill in Excel spreadsheets with exact reproductions of found errors or to shoot screenshots to prove inconsistencies in translations, nowadays technological support by adequate applications is available. But, although there are many standalone tools in the market – open source and commercial – that offer useful functionalities that LSPs need for reasonable revision stages, one huge problem remains: Most of them break the industry’s common workflows for the handling of translation projects, HT and MT alike. That means existing translations have to be exported from the translation environment in use to be imported into the revision tool. After revision is done, the reworked material as well as the error descriptions have to be returned back from revision to translation tool. And metrics and evaluation results are most likely to be managed in yet another system like a translation management application or a translation resources database.

Every working step aside from the dedicated workflow path, though, costs the LSPs time and money and cause a vast management overhead. This renders the application of the given method user-unfriendly and uneconomic.

For MT researchers and engineers on the other hand it is important that the results from annotations and error mark-ups can be fed back easily into the MT engines in order to optimize output during the next optimization and translation round.

Future advances in the field, therefore, must not only focus on assessment methods but also on the development of suitable tools where methods interlock with translation and revision workflows. Only if functionality and accessibility

is well integrated into the translation and post-editing environment and if engines can “learn” easily from post-editors’ feedbacks added value for all parties will be generated.

### 5.2 Development and Progress

Fortunately, advances for a more feasible MT usage is underway, and huge progress has been made, recently. The TAUS Dynamic Quality Framework has developed a range of tools that support evaluation and benchmarking within the industry<sup>5</sup>. The underlying DQF Error Typology that has been harmonized with the MQM model in 2015 and that represents a subset of the MQM specifications provides a means to quantify translation errors. It can be integrated via customized plugins into many commercially available translation tools or will be in the very near future according to a TAUS press release as of March 2016.<sup>6</sup>

Old and new application vendors like *SDL*, *Memsource* or *MateCat* have brought up new functionalities and CAT tools that combine MT and translation memories for traditional computer-aided translation into well-integrated workflows. Many of them function in the cloud and offer real-time processing and interactive post-editing of suggested MT segments. The result is a “self-learning environment” that not only measures editing distance and errors but also incorporates required changes for future similar occurrences.

Although not all open questions are answered yet, these forward-looking developments in the field of language technology are encouraging and propose a real change to come in the translation market.

## 6. Conclusion

The usage of MT in many professional translation contexts bears many chances and future prospects for the translation industry. MT researchers, on the other hand, need large amounts of domain-specific data to train engines and qualified expert feedback that serves as a basis for further optimization.

If both parties bundle their knowledge and leverage it for the sake of high-quality MT not only the language service

<sup>5</sup> See <https://evaluate.taus.net/evaluate/dqf-tools>

<sup>6</sup> See: <https://www.taus.net/think-tank/news/press-release/dqf-tools-updated-with-dqf-mqm-error-types>

sector but also the field of MT research will have their merits. The analyses from real-life scenarios offer valuable insights into common error patterns and necessary approaches for the improvement of MT engines. By using common standards, consistent benchmarks and integrated tools all players will benefit from each other's work in order to reach better results.

## 7. Acknowledgements

This work has received support from the EC's Horizon 2020 research and innovation programme under grant agreement No. 645452 (QT21).

## 8. Bibliographical References

- Burchardt, A. and Lommel, A. (2014) *Practical Guidelines for the Use of MQM in Scientific Research on Translation Quality* (published at <http://www.qt21.eu/downloads/MQM-usage-guidelines.pdf>)
- House, J. (1997). *Translation Quality Assessment: A Model Revisited*. Tübingen: Gunter Narr Verlag
- Koby, G. S. and Melby, A. K. (2013). Certification and Job Task Analysis (JTA): Establishing Validity of Translator Certification Examinations. In *The International Journal of Translation and Interpreting Research* (5)1, pp. 174--210.
- Lommel, A.; Burchardt, A. and Uszkoreit, H. (2014) Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. In *Tradumàtica: tecnologies de la traducció*: 0 (12), pp. 455--463.
- Martínez Mateo, R. (2014). A Deeper Look into Metrics for Translation Quality Assessment (TQA): A Case Study. In *Miscelánea: A Journal of English and American Studies* 49, pp. 73—94
- O'Brien, S. (2012) Towards a Dynamic Quality Evaluation Model for Translation. In: *The Journal of Specialized Translation* (Issue 17), pp. 55--77

## 9. Websites

- ISO 17100:2015: Translation services – Requirements for translation services: [http://www.iso.org/iso/catalogue\\_detail.htm?csnumber=59149](http://www.iso.org/iso/catalogue_detail.htm?csnumber=59149) (consulted 31.03.2016)
- LISA QA Model 3.1: <https://www.w3.org/International/O-LISA.html> (consulted 31.03.2016)
- SAE J2450 Translation Quality Metric Task Force: <http://www.sae.org/standardsdev/j2450p1.htm> (consulted 31.03.2016)
- TAUS: <https://www.taus.net/> (consulted 31.03.2016)

## Using MT-ComparEval

Roman Sudarikov,<sup>α</sup> Martin Popel,<sup>α</sup> Ondřej Bojar,<sup>α</sup> Aljoscha Burchardt,<sup>β</sup> Ondřej Klejch<sup>α,γ</sup>

<sup>α</sup> Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

<sup>β</sup> German Research Center for Artificial Intelligence (DFKI), Language Technology Lab

<sup>γ</sup> Centre for Speech Technology Research, University of Edinburgh

{sudarikov, popel, bojar, klejch}@ufal.mff.cuni.cz, aljoscha.burchardt@dfki.de

### Abstract

The paper showcases the MT-ComparEval tool for qualitative evaluation of machine translation (MT). MT-ComparEval is an open-source tool that has been designed in order to help MT developers by providing a graphical user interface that allows the comparison and evaluation of different MT engines/experiments and settings. The tool implements several measures that represent the current best practice of automatic evaluation. It also provides guidance in the targeted inspection of examples that show a certain behavior in terms of n-gram similarity/dissimilarity with alternative translations or the reference translation. In this paper, we provide an applied, “hands-on” perspective on the actual usage of MT-ComparEval. In a case study, we use it to compare and analyze several systems submitted to the WMT 2015 shared task.

**Keywords:** Machine Translation Evaluation, Analysis of MT Output

### 1. Introduction

The MT *development* cycle is well supported by several sophisticated pipeline tools such as the Experiment Management System EMS (Koehn, 2010) that comes with the Moses toolkit. While these pipelines support intermediate evaluation steps, there has been a lack of versatile tool support for detailed, qualitative evaluation steps, in particular for:

- Systematically documenting *significant quantitative* changes in terms of various automatic measures over a potentially large number of different system types, combinations, and variants, and
- *Qualitatively* analyzing the effects of changes in the systems integrated with the above.

The evaluation interface of EMS shows some similarity with the tool to be described in this paper, but it is tightly connected with the training pipeline and optimized for the Moses statistical machine translation (SMT) scripts. MT-ComparEval in contrast provides more flexibility, since the evaluation interface can be run independently of the production of translation systems.<sup>1</sup>

The automatic quantitative evaluation of MT is supported by different metrics such as BLEU, Meteor, TER (Agarwal and Lavie, 2008). A separate issue with these metrics is that they are usually implemented in unrelated collections of scripts which, in addition to unnecessary burden of tool installation, easily leads to difficulties with replicability: we can get different outcomes for the same metric from different implementations. The bigger problem, however, is still the question of how to perform qualitative evaluation. MT-ComparEval addresses the problems described above. Klejch et al. (2015) have introduced the tool from a general perspective, focusing on how to deal with automatic measures. In this paper, we will present it from the “hands-on”

view of a researcher who compares several machine translation systems with the goal of getting deeper insights into these systems than “System A is better than system B by 1.5 BLEU score” and possibly also with the goal to improve some of the systems.

### 2. MT-ComparEval

MT-ComparEval, the open-source tool described in this article, has been designed in order to help MT developers by providing a graphical user interface that allows the comparison and evaluation of different MT engines/experiments and settings through the use of several measures that represent the current best practice. The user interface is web-based and backed by a server side of the tool.

This paper won’t dwell on the internal structure of the tool, but rather point out certain main features which will later be used in qualitative evaluation of machine translation systems. These features include:

- Integration of different evaluation metrics – by default MT-ComparEval configuration includes precision, recall and F-measure (all based on arithmetic average of 1-grams up to 4-grams), BLEU score and Brevity penalty (Papineni et al., 2002). It can produce also Hjerson (Popović, 2011) evaluation scores “out-of-the-box” when enabled in the configuration file.
- Focus on pairwise comparisons of MT systems – so strengths and weaknesses of one system are shown relative to another system.<sup>2</sup>

<sup>2</sup>We consider the pairwise comparisons a great advantage of MT-ComparEval (compared to other tools) because it focuses on the errors that are more likely to be repairable (because the second MT system was able to translate these correctly), instead of simply focusing on n-grams/sentences that are generally difficult to translate. If only one system is available, it is still possible to analyze it with MT-ComparEval by selecting the reference translation as the second “system”. We plan to promote this feature in the interface because it may be useful *per se* (even if more systems are available).

<sup>1</sup>See Klejch et al. (2015) for a more detailed discussion.



Figure 1: “Experiment” screen with overview of all tasks.

- Bootstrap resampling – MT-ComparEval automatically generates bootstrap samples and computes the p-value for systems comparison and confidence intervals for all the produced evaluation scores.
- Confirmed and unconfirmed n-grams – the tool presents top 10 n-grams (for n=1,2,3,4) where the two systems differ with respect to correctness of translation (as measured by the reference translation), that is n-grams that are responsible for the difference in BLEU scores. Full explanation is given in Section 3.4.
- Sentence comparison – MT-ComparEval provides a graphically rich interface for sentence by sentence comparison of systems’ outputs.
- Accessibility – this tool can be easily installed and run locally (see “Installation” section at <https://github.com/choko/MT-ComparEval>).

In further sections, we will use MT-ComparEval terms “Experiment” and “Task” to refer to whole comparison and each system’s output, respectively.

### 3. Using MT-ComparEval Step by Step

As a running example for the rest of the paper, we use a set of systems from the WMT2015 shared task (Bojar et al., 2015), Czech→English translation task.<sup>3</sup> All the described observations were done using the public MT-ComparEval server with WMT translations <http://wmt.ufal.cz>.<sup>4</sup> We encourage the readers to navigate to the “Newstest 2015 cs-en” experiment and try all the described steps.

#### 3.1. Experiment Screen

Figure 1 shows the main screen of an “experiment”, which lists the results of all “tasks” (MT systems’ outputs) in this experiment. Figure 2 shows the same screen, where we

<sup>3</sup><http://www.statmt.org/wmt15/translation-task.html>

<sup>4</sup>The buttons for uploading and deleting experiments and tasks are disabled at <http://wmt.ufal.cz>. Local installation of MT-ComparEval can be configured to show these buttons or to permanently monitor a data directory for new experiments and tasks, which is suitable for integrating MT-ComparEval into an MT development pipeline.

## Newstest 2015 cs-en

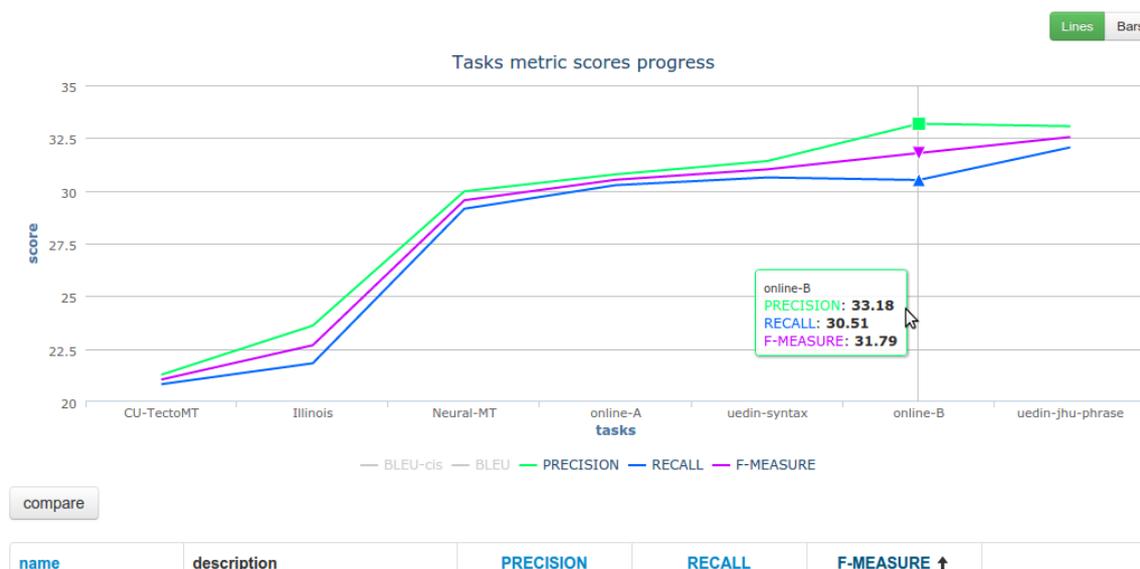


Figure 2: “Experiment” screen with Precision, Recall and F-measure (which was used for sorting the tasks).

have a) clicked on “F-MEASURE” in the table header, so the tasks are sorted according to this metric, b) clicked on the two variants of BLEU score under the graph to hide these metrics, so only Precision, Recall and F-measure are shown (and the graph y-axis is rescaled), and c) switched the graph type from bars to lines, so we can better see the differences between the metrics and check e.g. if some line segments are sloped down, which means disagreement with the F-measure used for sorting.

**Metrics disagreement** This actually happens for the top three systems, where Online-B is the best system according to Precision, second according to F-measure and third according to Recall.

**Precision and Recall mismatch** This is related to the fact that Online-B has Precision notably higher than Recall, while other systems have the difference much smaller. This may indicate that Online-B produces shorter translations and prefers to skip parts where the translation is not certain. This hypothesis can be checked by selecting Online-B and UEdin-jhu-phrase for pairwise comparison (see Section 3.2.) and looking at the sentences sorted according to RECALL or BREVITY-PENALTY (or even the default BLEU). See Figure 3 with two sentences where phrases *couldn't* and *high schools* were omitted in the Online-B translation.

**Casing problems** Figure 1 also shows that case-insensitive BLEU (BLEU-cis) is slightly lower than case-sensitive BLEU for all systems. The biggest difference is for Online-A (almost 1.5 BLEU points).<sup>5</sup> This indicates a

<sup>5</sup>This could be better seen when switching to the line graph and showing only BLEU and BLEU-cis. Online-A has one of the biggest differences in these two metrics also in other translation directions in WMT15: de-en, hi-en, fr-en and ru-en.

problem with upper-casing.

### 3.2. Sentences Pane

MT-ComparEval focuses on comparing two tasks (systemA and systemB). After marking the tasks’s checkboxes in the Experiment screen and clicking “Compare”, a screen with four panes is shown: Sentences, Statistics, Confirmed n-grams and Unconfirmed n-grams, which are described in the following subsections.

The Sentences pane (Figure 3) shows all sentences from the given testset sorted according to the differences in the chosen sentence-level metric scores. This means that the sentences shown at the top are those where systemB outperforms systemA the most.<sup>6</sup> Such a view is very useful when checking for regressions of new versions of an MT system against a baseline or a previous version of the same system, but it is useful also when comparing different systems.

**Color highlighting** A set of checkboxes allow to highlight differences between the two systems in several ways:

- **Confirmed n-grams** are n-grams occurring both in the system output and in the reference.<sup>7</sup> These are marked with light yellow (Online-B) and blue (UEdin-jhu-phrase) background. The confirmed n-grams are highlighted also in the reference, where light green color marks n-grams occurring in both system (e.g. “Why” in the first sentence in Figure 3).

<sup>6</sup>The metric used for sorting and the increasing/decreasing ordering can be changed in the upper right corner.

<sup>7</sup>If a given n-gram occurs e.g. three times in the system output and only twice in the reference, a heuristic algorithm (based on the longest common subsequence) is used to select two occurrences of the n-gram that will be marked as confirmed in the system output.

online-B uedin-jhu-phrase RECALL

Sentences Statistics Confirmed n-grams Unconfirmed n-grams

### Sentences

#### Options

N-grams highlighting options	Diff highlighting options	Sentences visibility options
<input checked="" type="checkbox"/> Highlight confirmed n-grams	<input type="checkbox"/> Show diff with reference	<input checked="" type="checkbox"/> Show source
<input checked="" type="checkbox"/> Highlight improving n-grams	<input checked="" type="radio"/> Show diff for online-B	<input checked="" type="checkbox"/> Show reference
<input checked="" type="checkbox"/> Highlight worsening n-grams	<input type="radio"/> Show diff for uedin-jhu-phrase	<input checked="" type="checkbox"/> Show online-B
	<input type="checkbox"/> Show diff with each other	<input checked="" type="checkbox"/> Show uedin-jhu-phrase
		<input type="checkbox"/> Show sentence level metrics

Source	Proč Strážci galaxie nedokázali zachránit tržby
Reference	Why the Guardians of the Galaxy couldn ' t save the box office
online-B	Why not rescue Rangers Galaxy sales
uedin-jhu-phrase	Why the Guardians of the Galaxy couldn ' t save sales

Source	Rušení gymnázií a středních škol nedává smysl .
Reference	Canceling high schools and secondary schools doesn ' t make sense .
online-B	Interference secondary schools and makes no sense .
uedin-jhu-phrase	Cancellation of grammar schools and secondary schools doesn ' t make sense .

Figure 3: Online-B shortens translations.

- **Improving n-grams** are confirmed n-grams occurring in only one of the systems. These are highlighted in the system outputs with darker yellow and blue (“the Guardians of the” and “couldn’t save” is present only in UEdin-jhu-phrase).
- **Worsening n-grams** are unconfirmed n-grams (i.e. probably wrong translations) occurring in only one of the systems. These are highlighted with red (e.g. “not rescue Rangers”).
- **Diff** of the reference and one of the systems: words in the longest common subsequence of the two sentences can be underlined in green, other words in red – this was switched off in Figure 3 to keep it uncluttered.

**Finding example sentences** MT researchers often need to find a nice example where their system outperforms another system due to a given linguistic phenomenon. They can hide everything except for the colored reference translation (so more sentences fit one screen) and quickly search for a long enough blue-highlighted phrase exhibiting the phenomenon.

### 3.3. Statistics Pane

This pane focuses on quantitative evaluation and shows all document-level metric scores for the two systems compared and four area charts. The bottom two charts show (*non-paired*) *bootstrap resampling* (Koehn, 2004) for the two systems, to assess BLEU (or other selected metric) confidence intervals for the individual systems. We will focus on the two upper charts, depicted in Figure 4, where we compare Neural-MT and Online-A.

The left chart shows sentence-level BLEU-cis difference (y-axis) for all the 2656 sentences in the testset (x-axis): about half of the sentences are translated better by Neural-MT (green region) and half by Online-A (red region). Even if the red and green regions have the same area (which seems to be the case here), it does not imply that the document-level BLEU-cis are the same: document-level BLEU-cis is influenced more by longer sentences, moreover, it is not decomposable to sentence-level scores due to brevity penalty etc. (Chiang et al., 2008). Nevertheless, it is interesting to see what portion of sentences is better in one system with a given sentence-level BLEU margin compared with the other system.

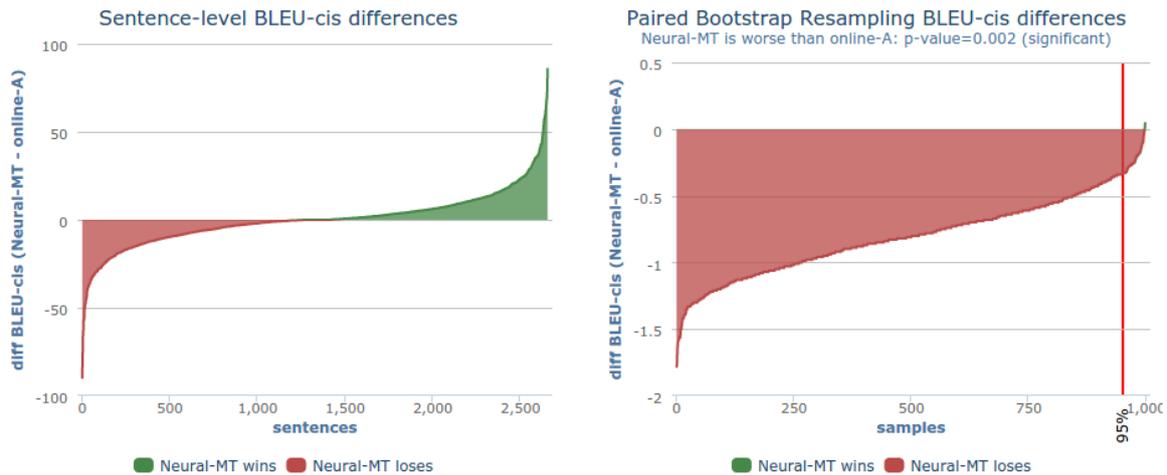


Figure 4: Statistics pane comparing Neural-MT and Online-A systems using Sentence-level BLEU differences graph (left) and Paired bootstrap resampling BLEU graph (right).

**Significance** MT researchers often need to know whether the difference between two systems in a given metric is significant or not. If the confidence intervals for the individual systems (in the bottom charts, not shown here) are not overlapping, it implies a significant difference, but the opposite implication does not hold. We need to use a paired test for checking the significance.

The right chart in Figure 4 shows *paired bootstrap resampling* (Koehn, 2004), where the x-axis lists 1,000 resamples of the testset and the y-axis is the difference in (document-level) BLEU-cis between the two systems for the given resample. One-tailed p-value is reported in the chart header:  $p = 0.002$ . This means that in 2 cases out of the 1000 resamples, Neural-MT had higher BLEU-cis than Online-A (this corresponds to the tiny green area on right, above the zero line). In the remaining 998 cases, Online-A had better scores, so we can conclude that Online-A is significantly better than Neural-MT in BLEU-cis (on the standard 95% confidence level).<sup>8</sup> When we change the metric to BLEU (case sensitive variant), we can see that Online-A is still better, but insignificantly ( $p = 0.415$ ).

### 3.4. Confirmed and Unconfirmed n-grams Panes

Figure 5 shows Confirmed and Unconfirmed n-grams panes, focusing on unigrams only and merging the two panes into one figure for space reasons. We selected TectoMT and UEdin-jhu-phrase (the worst and the best system according to BLEU) for comparison. After clicking on any n-gram, a Sentence pane is opened showing all sentences with this n-gram (which is highlighted).

<sup>8</sup>For better reliability, the number of resamples can be increased in the configuration file, in option `bootstrapSampler`. MT-ComparEval uses 1000 resamples by default in order to import quickly new tasks. It also uses a random seed, so replicating the experiment may lead to slightly different p-values, e.g. the current version at `wmt.ufal.cz` has  $p = 0.004$ .

**Quotes style** In the second row of the table of confirmed unigrams where UEdin-jhu-phrase “wins”, we can see (vertical) double quotes with numbers  $590 - 188 = 402$ . This means that this token was present 590 times in UEdin-jhu-phrase and confirmed by the reference, while TectoMT had only 188 confirmed occurrences of this token. The n-grams in the table are sorted according to the difference in the number of confirmed occurrences. In the 8th row of the table of unconfirmed unigrams where TectoMT loses, we can see (typographic) lower double quotes with numbers  $247 - 0 = 247$ . Lower quotes are used in Czech as opening quotes, but they should not be present in the English translations. The 0 means that UEdin-jhu-phrase did not produce any lower quotes (unconfirmed by the reference). TectoMT had 247 such occurrences and also 225 occurrences of unconfirmed (typographic) upper quotes. Thanks to MT-ComparEval, TectoMT developers were able to detect this error and fix it (simply by `s / [ , “ ” / g`). In a similar way, un/confirmed n-grams are useful for quick spotting of various encoding problems (which may be more important for the translation quality than quotes style).

**Definite and indefinite articles** Figure 5 also reveals a problem with articles in TectoMT output. Table 1 summarizes the relevant numbers and computes the total number of occurrences of “the” and “a” in the systems’ outputs.

We see that UEdin-jhu-phrase better uses “the” (confirmedDiff =  $2076 - 894 = 1182$ ), and it may seem that TectoMT better uses “a” (confirmedDiff =  $655 - 575 = 80$ ). It is important to always check also the Unconfirmed n-grams to prevent misleading conclusions. In Unconfirmed n-grams, we see that the seeming strengths of the systems are also their weaknesses: UEdin-jhu-phrase has 949 more unconfirmed “the”s than TectoMT, and TectoMT has 1065 more unconfirmed “a”s than UEdin-jhu-phrase.

We conclude that a) TectoMT produces in total fewer articles than UEdin-jhu-phrase. b) TectoMT prefers “a”, while UEdin-jhu-phrase prefers “the”. c) For “the”, TectoMT has higher precision than UEdin-jhu-phrase; for “a” vice versa

MT-ComparEval Newstest 2015 cs-en		MT-ComparEval Newstest 2015 cs-en	
<b>n-grams confirmed by the reference</b>		<b>n-grams unconfirmed by the reference</b>	
1-gram		1-gram	
CU-TectoMT wins		uedin-jhu-phrase wins	
not	245 - 160 = 85	the	2076 - 894 = 1182
a	655 - 575 = 80	"	590 - 188 = 402
he	163 - 110 = 53	,	1945 - 1617 = 328
will	149 - 97 = 52	'	230 - 76 = 154
.	2405 - 2363 = 42	The	201 - 60 = 141
they	93 - 59 = 34	that	341 - 204 = 137
an	74 - 52 = 22	his	97 - 22 = 75
we	151 - 132 = 19	at	103 - 28 = 75
did	33 - 14 = 19	t	74 - 0 = 74
it	222 - 205 = 17	in	701 - 635 = 66
CU-TectoMT loses		uedin-jhu-phrase loses	
a	1410 - 345 = 1065	the	1463 - 514 = 949
,	1135 - 776 = 359	The	247 - 90 = 157
it	528 - 169 = 359	'	165 - 29 = 136
he	416 - 103 = 313	that	309 - 201 = 108
an	336 - 43 = 293	"	82 - 9 = 73
not	400 - 126 = 274	at	90 - 23 = 67
of	1000 - 728 = 272	is	298 - 234 = 64
"	247 - 0 = 247	his	74 - 13 = 61
'	225 - 0 = 225	t	59 - 0 = 59
will	326 - 114 = 212	s	73 - 16 = 57

Figure 5: Confirmed and Unconfirmed n-grams panes (showing problems with quotes and articles in TectoMT).

	"the"		"a"	
	UEdin	TectoMT	UEdin	TectoMT
confirmed	2076	894	575	655
unconfirmed	1463	514	345	1410
total	3539	1408	920	2065
% confirmed	59%	63%	63%	32%

Table 1: Comparison of "the" and "a" usage in TectoMT and UEdin-jhu-phrase.

(see the last row in Table 1). d) Based on the number of confirmed n-grams, we see that for "the", UEdin-jhu-phrase has much higher recall than TectoMT; for "a", TectoMT has slightly higher recall than UEdin-jhu-phrase. e) With regards to the precision-recall balancing, TectoMT should produce more definite articles, but fewer indefinite ones.

**Untranslated chunks** Now, we will focus on Neural-MT and compare it with UEdin-jhu-phrase. Figure 6 shows unconfirmed unigrams and in the "Neural-MT loses" table, we can see "na" and "se", which are often erroneously produced by Neural-MT (58 and 57 times, respectively), but never by UEdin-jhu-phrase. These tokens are frequent Czech words (prepositions).<sup>9</sup> If we click on these tokens, we will see sentences where Neural-MT left untranslated these tokens, quite often within longer untranslated phrases. For example, in Figure 7 we see untranslated phrases "První jarní den" (first day of spring) and "na letišti na letišti" (on airport on airport). Also in many other sentences found with "se", "na" or "v", we can see untranslated phrases consisting of easy-to-translate common words. We hypothesize that this is a peculiarity related to recurrent-neural-network

<sup>9</sup>Also "s" and "v", which are listed in the table, are Czech prepositions, but these are sometimes erroneously produced also by UEdin-jhu-phrase. Due to the tokenization in MT-ComparEval (taken from BLEU), "he's" is tokenized as "he ' s" and thus token "s" may appear in English translations.

MT-ComparEval Newstest 2015 cs-en	
<b>n-grams unconfirmed by the reference</b>	
1-gram	
Neural-MT loses	
,	996 - 776 = 220
had	229 - 48 = 181
'	307 - 165 = 142
the	1571 - 1463 = 108
s	172 - 73 = 99
v	85 - 1 = 84
of	807 - 728 = 79
.	175 - 99 = 76
na	58 - 0 = 58
se	57 - 0 = 57
uedin-jhu-phrase loses	
is	298 - 179 = 119
The	247 - 205 = 42
which	129 - 89 = 40
in	398 - 365 = 33
his	74 - 43 = 31
and	165 - 136 = 29
It	80 - 53 = 27
will	114 - 90 = 24
but	46 - 22 = 24
just	36 - 15 = 21

Figure 6: Untranslated Czech prepositions in Neural-MT.

nature of Neural-MT (Jean et al., 2015), which could be easily fixed (at least with an automatic post-processing).

**Other Neural-MT peculiarities** We noticed that the English translations contain not only untranslated Czech phrases, but also Czech phrases which were not in the source sentence, e.g. "které byly" (which were) in Figure 7. We also noticed many mistakenly repeated words or phrases (both translated and untranslated), e.g. "na letišti" (on airport). MT-ComparEval does not have any specialized tool for finding such repeated phrases, but the red highlighting in Sentence pane helps to spot them. Also the top unconfirmed Neural-MT 4-gram is ". . . .", originating from a translation with a dot repeated 59 times.

Source	První jarní den poznamenán deštěm a bouřkami , které měly dopad na let na letišti v Adelaide
Reference	First day of spring marked with wet and blustery conditions impacting Adelaide Airport flights
Neural-MT	První jarní den by rain a storms , které byly impact <span style="border: 1px solid black; padding: 0 2px;">na</span> letišti <span style="border: 1px solid black; padding: 0 2px;">na</span> letišti v Adelaide

Figure 7: Example of Neural-MT output with untranslated phrases and unconfirmed unigram “na” highlighted.

#### 4. Conclusion

In this paper, we have presented MT-ComparEval, an open-source tool that provides a graphically rich environment to perform quantitative and qualitative evaluation and deep analysis of machine translation outputs. We have presented its usage in the comparison and improvement of several systems.

While the developers of the underlying MT systems may already be familiar with many of the issues in their systems’ output, MT-ComparEval helps to integrate quantitative analyses including significance tests with qualitative analysis that can help to avoid the most frequent systematic errors. This is especially relevant when working on “difficult” languages where fixing issues can be very costly, and thus has to be prioritized and systematic.

We are convinced that the usage of tools like MT-ComparEval in general will lead to a more analytic approach to MT development and evaluation, getting away from the very superficial level of “System A is better than system B by 1.5 BLEU score”. It will help researchers to generate informed hypotheses for improvements and to increase the informativeness of publications as the graphical interface makes it easy to search for nice illustrating examples that fix certain issues under consideration (or lead to new issues to be fixed).

#### 5. Acknowledgment

This research was supported by the grants FP7-ICT-2013-10-610516 (QTLep), GA15-10472S (Manyla), SVV 260 224, and using language resources distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (LM2015071). We thank the three anonymous reviewers for useful comments.

#### 6. Bibliographical References

Agarwal, A. and Lavie, A. (2008). Meteor, M-BLEU and M-TER: Evaluation metrics for high-correlation with human rankings of machine translation output. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 115–118. Association for Computational Linguistics.

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. (2015). Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisboa, Portugal, September. Association for Computational Linguistics.

Chiang, D., DeNeefe, S., Chan, Y. S., and Ng, H. T. (2008). Decomposability of translation metrics for improved evaluation and efficient algorithms. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 610–619, Honolulu, Hawaii, October. Association for Computational Linguistics.

Jean, S., Firat, O., Cho, K., Memisevic, R., and Bengio, Y. (2015). Montreal neural machine translation systems for WMT’15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 134–140, Lisbon, Portugal, September. Association for Computational Linguistics.

Klejch, O., Avramidis, E., Burchardt, A., and Popel, M. (2015). MT-ComparEval: graphical evaluation interface for machine translation development. *The Prague Bulletin of Mathematical Linguistics*, (104):63–74.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In Dekang Lin et al., editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. ACL.

Koehn, P. (2010). An Experimental Management System. *The Prague Bulletin of Mathematical Linguistics*, 94:87–96.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL*, pages 311–318, Stroudsburg, PA, USA. ACL.

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

# CMT: Predictive Machine Translation Quality Evaluation Metric

Michal Tyszkowski, Dorota Szaszko

Centrum Lokalizacji CM

Parkowa 19, 51-616, Wrocław, Poland

E-mail: michal.tyszkowski@cmlocalization.eu, dorota.szaszko@cmlocalization.eu

## Abstract

Machine Translation quality is evaluated using metrics that utilize human translations as reference materials. This means that the existing methods do not allow to predict the quality of Machine Translation if no human translations of the same material exist. To use Machine Translation in the translation industry, it is essential to have a metric that allows to evaluate the quality of a newly created Machine Translation in order to decide whether or not it can increase productivity. As a translation company that uses Statistical Machine Translation to generate translation memories for many projects, we decided to develop a metric that can predict its usability on a project basis. This metric, called CMT, is a combination of human assessment and statistics comparable to existing metrics. Our investigations showed that the mean difference between CMT and BLEU is 9.10, and between CMT and METEOR it is 9.69, so it correlates with the existing metrics in more than 90%. CMT is very easy to use and allows to evaluate each translation memory, regardless of its size, in 5 minutes without any reference material.

**Keywords:** Machine Translation, MT evaluation, CMT metric, predictive MT evaluation metric

## 1. Introduction

Machine Translation at its early stage of development was mainly used for investigation purposes. The aim of these investigations was to assess the similarity of translations produced by computers with human translations. To make the assessments comparable, a number of metrics were developed, amongst which BLEU (Papineni, Roukos, Ward, & Zhu, 2002), NIST (Doddington, 2002) and METEOR (Lavie, Sagae & Jayaraman 2004) are most commonly used. All these metrics show better or worse the similarity between machine and human translation, but they are retroactive. This means that they are calculated against reference translations after both machine and human translation are completed, which is correct for the purpose they were created.

Machine Translation, however, is not used only for investigation purposes any more. It is now used also as a tool for rough understanding of a text written in an unknown language, and also as a productivity tool for technical translations. Retroactive evaluation metrics is useless for the latter purpose, because the translation industry needs to know whether or not Machine Translation has an acceptable quality before it is used.

Translation buyers often request using Machine Translation for their projects to get lower prices. MT proposals are useful and really increase productivity only if their quality is high, meaning that their post-editing requires less time than translating from scratch. Our experience shows that this condition is fulfilled if the average BLEU score of MT-generated database is not less than 50. The existing evaluation metrics allow to calculate this score after the human translation or post-editing is completed, however, we need to estimate it in advance to be able to decide if a customer's database is really usable and speeds up the process or it is better not to use it and translate from scratch. This is the reason for developing our own tool that could estimate Machine Translation quality before it is used and before the reference human translations exist.

This paper describes briefly the main concept of the

predictive MT evaluation methodology and presents the results of a comparison to the metrics calculated after post-editing that shows the high efficiency of our tool.

## 2. Methods

The investigation required 3 components:

### 2.1 Machine Translation Engine

An engine to generate MT proposals for as many translation projects as possible and use it for research purposes.

NiuTrans (Xiao, Zhu, Zhang & Li, 2012) an Open Source solution, was used as a Statistical Machine Translation engine. Based on the translation projects completed by our company, 2 corpora were prepared – the first one for IT translations, containing about 30 million of words, and the second one for medical translations, with about 15 million words. For both corpora, a phrase-based translation model was trained. Selected translation projects in the field of IT and medicine from last year were subjected to an MT-evaluation procedure that consisted in producing MT proposals for all segments that did have no matches in our Translation Memories, evaluating these proposals with the use of our proprietary tool, and, depending on the score they achieved, incorporation of MT proposals during the translation phase.

### 2.2 Estimation Program

A computer program that estimates the quality of a newly generated MT database and calculates a score that can be then compared to a BLEU score or other metrics.

As already stressed in the introduction, no predictive MT evaluation metrics exists, so there is also no program that could calculate it. Our task was to create both: the metric and the program. We assumed that there is no possibility to create a fully automated metric that can evaluate translation quality with no reference. In fact, the existing metrics do not tell anything about translation quality, because it can only be assessed by a human. The metrics only measure the distance between a machine-translated and

human-translated sentence. If the reference translation is wrong, the BLEU value can be high, but the real quality is low. That is why we decided to base our tool on a human judgement, but built into software and rendered as a numeric value.

In our solution, a source text and machine-translated target are displayed in two columns. An evaluator looks at the source and target sentence and decides whether the target is:

- Correct (no errors) – the segment does not get any score,
- Almost correct (1–2 minor errors) – the segment scores 1 point,
- Acceptable (more errors, but still easily understandable) – the segment scores 2 points,
- Unacceptable – the segment scores 3 points.

The decision is taken by clicking one of 4 buttons. From our experience, an evaluator needs about 5 seconds per sentence to evaluate it. The sentences are chosen randomly, based on a selected number of words. For our investigation, 500 words per project were chosen.

The tool calculates the final score using the following formula:

$$100 - \frac{100 \sum s}{s_{max} n}$$

where:

**s** is a segment score with a value from 1 to 3

**s<sub>max</sub>** is a max segment score

**n** is a number of evaluated segments

### 2.3 Score Evaluation Program

A computer program that calculates MT evaluation metrics using our MT database and translated targets as a reference. To prove the usability of our score, it was compared with the results of MT evaluation performed using two metrics: BLEU and METEOR. For this investigation, we used machine-translated sentences matched with the same sentences translated by humans. 60 data sets saved as plain text files with sentences separated by line feed characters were processed by the iBLEU 2.6.2 (Madnani, 2011) program for BLEU metric calculation and METEOR 1.5 (Denkowski & Lavie, 2014) for METEOR calculation.

For each data set, a difference between our predictive metric and BLEU and METEOR metrics was calculated, as well as a mean value and standard deviation. The results are described below.

## 3. Results

The results of our research on comparison between CMT and known metrics BLEU and METEOR are shown in the Table 1:

Project	CMT	BLEU		METEOR		Usability diff.	Number of words
		iBLEU	Delta BLEU	METEOR	Delta METEOR		
IT01	32.70	17.90	14.80	24.96	7.74	No	1132
IT02	52.08	42.84	9.24	50.03	2.05	Yes	1671
IT03	56.25	35.65	20.60	39.92	16.33	Yes	9832
IT04	22.52	27.51	4.99	33.46	10.94	No	1491
IT05	18.33	23.60	5.27	29.37	11.04	No	912
IT06	33.33	22.20	11.13	30.52	2.81	No	939
IT07	19.61	17.12	2.49	25.75	6.14	No	228
IT08	20.16	27.07	6.91	32.72	12.56	No	1071
IT09	25.00	21.82	3.18	29.91	4.91	No	673
IT10	23.76	27.91	4.15	36.14	12.38	No	543
IT11	36.84	30.17	6.67	39.66	2.82	No	715
IT12	76.42	54.57	21.85	65.68	10.74	No	2343
IT13	62.37	41.62	20.75	45.78	16.59	Yes	8161
IT14	23.42	37.88	14.46	46.21	22.79	No	467
IT15	23.53	25.30	1.77	29.74	6.21	No	5407
IT16	52.52	29.41	23.11	39.91	12.61	Yes	567
IT17	73.33	68.38	4.95	71.78	1.55	No	7654
IT18	67.42	67.97	0.55	69.35	1.93	No	3423
IT19	71.27	60.61	10.66	64.40	6.87	No	6000
IT20	60.00	59.37	0.63	64.47	4.47	No	19,314
IT21	61.36	57.20	4.16	62.41	1.05	No	10,943
IT22	71.32	60.63	10.69	63.47	7.85	No	38,020
IT23	57.62	66.6	8.98	70.44	12.82	No	6099
IT24	47.62	43.2	4.42	49.14	1.52	No	1934
IT25	52.71	54.09	1.38	55.23	2.52	No	3051
IT26	61.81	59.05	2.76	62.55	0.74	No	446
IT27	48.41	47.03	1.38	53.88	5.47	No	5717
IT28	42.64	48.14	5.50	53.47	10.83	No	11,364
IT29	50.17	59.47	9.30	65.27	15.10	No	2861
IT30	50.00	58.56	8.56	60.12	10.12	No	3423
IT31	61.81	70.00	8.19	68.70	6.89	No	543
MED01	59.03	45.84	13.19	50.03	9.00	Yes	6708
MED02	41.18	38.62	2.56	45.35	4.17	No	3813

Project	CMT	BLEU		METEOR		Usability diff.	Number of words
		iBLEU	Delta BLEU	METEOR	Delta METEOR		
MED03	29.17	30.50	1.33	45.47	16.30	No	1217
MED04	17.65	26.84	9.19	32.65	15.00	No	9757
MED05	20.00	47.87	27.87	42.94	22.94	No	2076
MED06	35.24	31.37	3.87	36.96	1.72	No	1410
MED07	11.90	31.20	19.30	41.69	29.79	No	494
MED08	22.76	22.25	0.51	31.52	8.76	No	11,702
MED09	43.17	45.78	2.61	49.98	6.81	No	3119
MED10	39.61	43.88	4.27	48.20	8.59	No	3020
MED11	37.25	24.06	13.19	31.80	5.45	No	1054
MED12	40.14	36.52	3.62	42.94	2.80	No	1278
MED13	43.21	34.71	8.50	38.39	4.82	No	1238
MED14	39.13	46.51	7.38	50.24	11.11	No	1809
MED15	35.83	49.21	13.38	55.03	19.20	No	1107
MED16	33.33	45.23	11.90	52.07	18.74	No	2995
MED17	25.68	35.57	9.89	32.38	6.70	No	958
MED18	21.90	31.47	9.57	34.23	12.33	No	1958
MED19	54.50	39.04	15.46	42.10	12.40	Yes	3713
MED20	21.11	22.37	1.26	26.47	5.36	No	2152
MED21	25.49	29.71	4.22	31.48	5.99	No	373
MED22	6.06	24.74	18.68	32.74	26.68	No	2744
MED23	57.00	28.02	28.98	34.19	22.81	Yes	7111
MED24	24.76	25.96	1.20	28.00	3.24	No	836
MED25	29.00	21.55	7.45	26.30	2.70	No	6746
MED26	25.93	30.63	4.70	35.51	9.58	No	51,802
MED27	9.76	17.92	8.16	20.33	10.57	No	2375
MED28	19.19	32.08	12.89	25.81	6.62	No	12,375
MED29	57.75	30.28	27.47	34.95	22.80	Yes	12,875
Mean value			9.10		9.69	Total words:	315,759
Standard deviation			7.30		6.90		

Table 1. Comparison between CMT, BLEU, and METEOR metrics

31 IT and 29 medical translation projects were used for creating a machine-generated translation memory which was evaluated using the CMT metric. After human translation of these projects, the BLEU and METEOR metrics were calculated using human translation as a reference. Then, the results of the CMT, BLEU, and METEOR metrics were calculated by subtracting the values.

The highest difference between CMT and BLEU was 29.98, and the lowest difference was 0.51. The mean value of this difference was 9.10 and the standard deviation was 7.30.

The highest difference between CMT and METEOR was 29.79, and the lowest difference was 0.74. The mean value of this difference was 9.69 and the standard deviation was 6.90.

Apart from the difference between metrics, we also checked in how many cases the decision about the usability of Machine Translation for post-editing taken on the basis of the CMT metrics appeared to be wrong. As already mentioned, our practice shows that it is reasonable to use Machine Translation as a productivity tool only if the BLEU score is not less than 50. This threshold was obtained in an empiric way. Translators have always choice either post-edit an MT proposal or translate the sentence from scratch. While analyzing sentences that translators post-edited we noticed that their BLEU score was never below 50. This means that sentences with lower quality were not used for post-editing but translated from scratch. Because the CMT score corresponds to BLEU, the machine generated translations were used for post-editing only if the CMT score was 50 or more. After calculating the BLEU

score, it appeared that our decision was wrong only in 8 cases and it was right in 52 cases.

We also investigated whether or not the good results could be accidental. To verify this, the distribution of values was examined. The results are shown in Figure 1. The graph shows that the majority of values are in the ranges 0–2.8, 2.9–5.6, and 8.5–11.2, so the shape of this graph is far from the Gaussian curve.

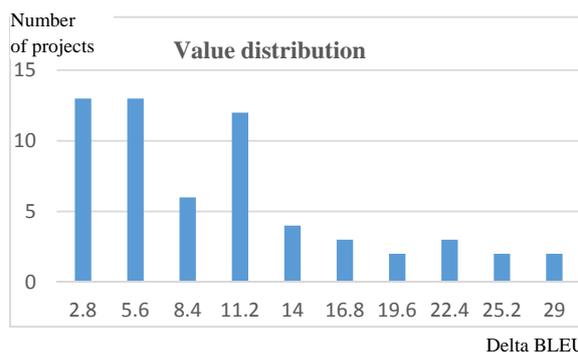


Figure 1. The distribution of the values of the difference between the CMT and BLEU scores

The last aspect that was checked was the dependency between the number of machine-translated words and the difference between the CMT and BLEU score. The results are illustrated in Figure 2. The graph shows that there is no significant dependency between these values, as the curves

have completely different shapes.

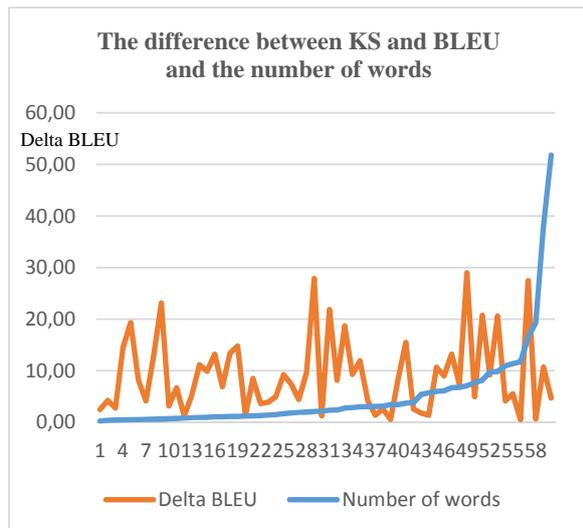


Figure 2. The dependency between number of words in MT and the difference between the CMT and BLEU scores

#### 4. Conclusions

CMT is the first predictive MT evaluation metrics which means that it is designed to evaluate the usability of Machine Translation without any reference material.

CMT is calculated with a dedicated software utilizing an algorithm that uses a human judgement and statistics. The value of CMT ranges from 0 to 100, which makes it comparable to known MT evaluation metrics such as BLEU and METEOR.

The software used for CMT calculation allows to evaluate any Machine Translation in about 5 minutes, regardless of its size.

The comparative research conducted using 60 translation projects with a total wordcount of 315,759 words showed that the mean difference between CMT compared to BLEU and METEOR was below 10 (9.10 for BLEU and 9.69 for METEOR), which means that the correlation between CMT and the metrics calculated using human translations as reference is above 90%.

The correlation between CMT, BLEU, and METEOR does not depend on the number of words evaluated and the most values placed in range 0–8.4, which means that the distribution is not normal, but shifted towards the smallest values.

Unlike BLEU, METEOR, and other retroactive metrics, CMT does not rely on the quality of reference materials, so it is much more comparable to the human judgement.

CMT is a score that can be used in the translation industry to support the decision whether or not it is reasonable to use Machine Translation as a productivity tool for a given translation project.

#### 5. References

Denkowski M. and Lavie A. (2014). Meteor Universal: Language Specific Translation Evaluation for Any Target

Language, *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*

Doddington, G. (2002) Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *Proceedings of the Human Language Technology Conference (HLT)*, San Diego, CA pp. 128–132

Lavie, A., Sagae, K. and Jayaraman, S. (2004). The Significance of Recall in Automatic Metrics for MT Evaluation in *Proceedings of AMTA 2004*, Washington DC. September 2004

Madnani, N. (2011). iBLEU: Interactively Debugging & Scoring Statistical Machine Translation Systems in *Proceedings of the fifth IEEE International Conference on Semantic Computing*, Sep 19-21, 2011

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

Tong Xiao, Jingbo Zhu, Hao Zhang and Qiang Li. (2012). NiuTrans: An Open Source Toolkit for Phrase-based and Syntax-based Machine Translation. In *Proc. of ACL, demonstration session*

Proceedings of the LREC 2016 Workshop  
“Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem”

24 May 2016 – Portorož, Slovenia

Edited by Georg Rehm, Aljoscha Burchardt, Ondřej Bojar, Christian Dugast, Marcello Federico, Josef van Genabith, Barry Haddow, Jan Hajič, Kim Harris, Philipp Koehn, Matteo Negri, Martin Popel, Lucia Specia, Marco Turchi, Hans Uszkoreit

<http://www.cracking-the-language-barrier.eu/mt-eval-workshop-2016/>

Acknowledgments: This work has received funding from the EU’s Horizon 2020 research and innovation programme through the contracts CRACKER (grant agreement no.: 645357) and QT21 (grant agreement no.: 645452).

