Global Under-Resourced MEedia Translation (GoURMET)

H2020 Research and Innovation Action
Number: 825299
D5.1 – Evaluation Plan

Nature	Report	**Work Package**	WP5
Due Date	31/06/2019	**Submission Date**	28/06/2019
Main authors	Andrew Secker (BBC), Alexandra Birch (UEDIN), Peggy van der Kreeft (DW), Felipe Sánchez-Martínez (ALAC)		
Co-authors			
Reviewers	Mikel L. Forcada (ALAC)		
Keywords	Evaluation		
Version Control			
v0.1	**Status**	Draft	17/06/2019
v1.0	**Status**	Final	28/06/2019

Contents

List of Figures

1 Introduction

This document describes the methods and metrics that will be used to evaluate the machine translation technologies developed within GoURMET.

The evaluation methodology for GoURMET is two-tiered and as such this document is divided into two primary sections.

First, the data-driven evaluation techniques that will be undertaken are described, and detail is given of the test sets that will be created and used for this. The academic partners lead that evaluation, including automated metrics and shared tasks.

The second part on this document describes two further levels of use-case-driven testing, focusing on the primary use cases of the two media partners, i.e. media monitoring and content creation. This second tier includes firstly deliberately controlled metrics such as gap filling or post-editing speed. The second stage of this tier is subjective evaluation, e.g. opinion rating and user interviews. Insights here will be used to validate and augment the results from the data-driven evaluation. It will support the detailed use cases identified in deliverable D5.2, show the usefulness of the translation models to be developed, and reinforce the introduction of MT into the media workflow and extend CAT to low-resourced languages.

Thus, the different types of methodology used cover a continuum of evaluation, from the rigorous and strictly reproducible data driven evaluation, to the more nuanced opinion-driven subjective evaluation.

Ultimately the evaluation methods described herein will allow all project partners to understand the quality and efficacy of the developed machine translation technologies in their particular field of interest.

2 Scope

The scope of the evaluation extends only to the testing of the (objective) quality and (subjective) usefulness of the translations produced by the research technologies themselves. The evaluation of any prototype tools (as described in deliverable D5.2) and their user interfaces, which are used to expose and evaluate the MT technologies under the user-centred evaluation methods (Section 5), are out of scope.

3 Overall Evaluation Methodology

Evaluation begins as soon as the first language pairs are available (i.e directly after Milestone MS1). Evaluation will take place on the release of new languages and upon a significant update of an existing language. There are four delivery and evaluation cycles over the period of the project.

Evaluation will be undertaken for each language pair developed in the project and will take two forms – data-driven evaluation and user-centred (or "human") evaluation.

Data-driven metrics are constantly used to drive the development of the translation models. There will be an official evaluation of the models delivered to the user partners using all automatic metrics for each iteration of each language pair (Section 4.2.1).

Complementing the data-driven evaluation, user-centred evaluation will take place for all language pairs. The amount and level of evaluation may differ per language pair depending on the data available, the quality of the translation output and the availability of editorial testing staff. Some languages are only covered by one of the two user partners and may be targeted specifically towards one use case. In the extreme case where no user partner is able to support user-centred evaluation, outsourcing or crowd-sourcing may be used. Like the data driven evaluation, user-centred evaluation methods and metrics formats will be consistent, allowing for comparison of output results. The methodology and detail of the proposed user-centred evaluation is described in Section 5.

3.1 Importance of the Selected Evaluation Methods

While data-driven metrics are used to understand the raw accuracy of the translations, for the project's media partners it is essential to understand how the values of the metrics revealed by these apply in the real world. It is hypothesised that data-driven metrics are proxies for real-world usefulness, but the project seeks to understand this relationship.

Given that this project is motivated by real-world use cases of content creation and media monitoring, GoURMETmust seek to:

1. understand how the subjective measures of MT quality correlates with scores generated by the data-driven methods (Section 4) and

2. understand how the nuances of errors may affect the usefulness of the translated output on a language by language basis.

Ultimately, user-centred evaluation allows us to answer the question "*To what extent is this translation technology good enough to allow me to do what I want to do*?". For example, in some (gisting) cases, the inaccurate rendering of a plural as a singular or vice versa will not affect the perceived usefulness of the system. In this case it might be more important that the meaning of the translation is true to the original. For instance, the omission (or spurious addition) of 'not' would be a more serious problem and affect fidelity. However, in the content-creation use case, frequent morphological errors might cause the journalist to abandon post-editing and start translating from scratch.

The combination of evaluation methodologies described in the remainder of this document will allow the project to draw informed conclusions in this regard.

4 Data-Driven Evaluation

4.1 Overview

In order to develop research into low-resource machine translation, the project will target reproducible and low-cost evaluation, and for that automatic evaluation is largely exploited. Automatic evaluation relies upon test sets which are ideally in-domain human translations. Automatic evaluation is an integral part of research in machine translation. It is used to both evaluate research methodology, to guide the training of the models and preventing models from overfitting. Two of the most prominent MT metrics will be used, one which acts at a word level, and one at a character level. In order to establish the quality of our research internationally, the project will compete in open shared-task competitions.

4.2 Evaluation Methods

4.2.1 Automatic Metrics

Automatic evaluation is essential to machine translation. It is low-cost, and can quickly evaluate large test sets in a reproducible fashion. Even though good metrics have been shown to correlate as much as possible with human judgements, they have their limitations. Firstly they require human translated test sets, which are expensive to create. Secondly, and more importantly, they have to balance being interpretable vs. having obvious limitations.

During our research, MT performance will be evaluated continuously in terms of commonly used automatic metrics, and in particular the BLEU score (Papineni et al., 2002) and CHRF (Popović, 2015). Using these two metrics, results will be well understood by the machine translation community while additionally different aspects of translation performance can be captured.

The BLEU score is a precision-based metric which rewards overlapping sequences of 1 to 4 words, between the machine translation output and the human translation. It captures both semantic and syntactic qualities, but it is obviously limited by not being able to reward synonyms or correct alternative word orderings. The CHRF score is a variant of the BLEU score which acts at the character level. This means that it can reward getting parts of words correct and this makes it more meaningful in evaluating morphologically rich target languages.

4.2.2 Shared Tasks

In order to definitively establish the quality of our systems, the project will participate in open shared task competitions. Specifically, the Shared Task on News Translation at the annual Conference on Machine Translation (WMT). Although automatic scores for individual systems are calculated, the final rankings are based on human evaluation. The main method for evaluating the ranking is Direct Assessment (DA) (Graham et al., 2017). Human evaluation is expensive and noisy, and DA scores have shown to be highly repeatable in self-replication experiments.

This metric focuses on fidelity (correctness, reliability of the information) of a translated text. The advantages of the metric are that it is simple and fast. Annotators are asked to rate the accuracy of the translation on a sliding scale of 0%–100%. They need no training and can perform the task quickly. It has been shown to be stable given sufficient number of judgements. It has been

designed for crowd-sourcing and when unreliable judgements are filtered, the results are reliable. The number of judgements needed scales linearly with the number of translation systems which are being compared, which is not the case with ranking-based evaluation systems. Finally, the scores are absolute which makes them interpretable.

Overall, human evaluation is expensive and noisy. DA provides some solutions to these two issues, as it has shown to be highly repeatable in self-replication experiments, leading to more efficient use of funding and consistency.

An example interface for gathering DA assessments from human evaluation measure this metric is shown in Figure 1.

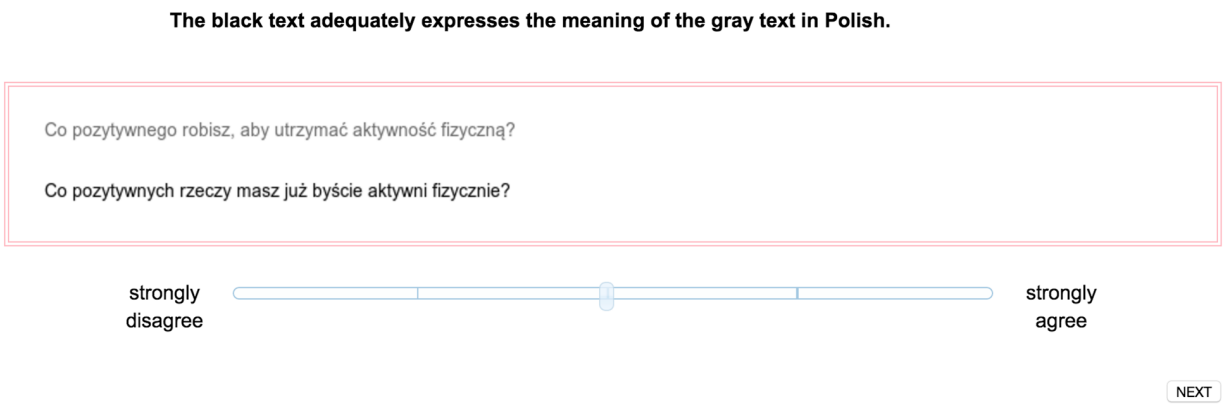

Figure 1: Example interface to allow feedback for direct assessment. The slider at the bottom returns scores ranging between 0 and 100.

Thus, our participation in these shared tasks will benchmark our systems using human evaluation. It will also allow us to determine how the MT systems produced by GoURMETcompare to the world's best research labs and to the most widely used commercial translation systems.

4.2.3 Robustness to Low-Resource Scenarios

The main focus of research in the project is on how robust our models are to low-resource scenarios. Because of this, our evaluations will focus on how well our models perform under low-resource conditions, but for some language pairs experiments will be scaled up and show how our methods perform as the training data becomes larger. Evaluation will also look at the ability to predict rare words, and unseen or rarely seen morphological variants of words in the source and target languages. This will be done by looking specifically at the translation accuracy of words in the test set that have occurred with low frequency in the training set. Looking at different types of low frequency words (morphological variants, named entities etc.) will allow us to determine what aspects of translation the MT technologies can successfully model, whilst which remain out of reach.

4.2.4 Translation Speed

Although speed is not foreseen to be a big issue in this project, there is a lower bound in performance which would make machine translation models unusable. See Deliverable D5.2 which

describes efficiency requirements. A maximum time of 500 ms per sentence of 80 words maximum is set as a minimum requirement. Our machine translation engines will be benchmarked with respect to translation speed. The evaluation will measure the number of words and sentences translated per second, and compare this to a number of other popular machine translation decoders. Trade-offs between model quality and speed will be measured to determine good settings for parameters that affect speed and translation quality.

4.3 Test Sets

Automatic metrics need human translated test sets to benchmark the system performance. The project will specifically create three novel, in-domain test sets. These are expensive to produce and so the maximum use will be made of them by deploying each as a low-resource track in the WMT annual shared-task competition. This spurs on the whole MT community to focus on problems this project is motivated by. This has already been done in 2019. A Gujarati test set was created by selecting news reports and translating them using a professional translation service. Participation in the Gujarati track[1] was good with 27 systems submitted in the Gujarati–English track[2] and 20 in the English–Gujarati track.[3]

The project will create some test sets specifically for some of the covered language pairs. These will be supplemented by other test sets. BBC and DW will provide some test sets from their repositories. The volume of such data sets will differ per language. Also externally available data sets will be explored. For instance, the project will therefore try to select existing test sets which have already had publications that use them and for which state-of-the-art performance can be determined. These will often come from competitions. However, there may either be no such test sets available, or they might not be in the correct domain (eg. biomedical test sets). In which case the most appropriate sentences can be selected from any existing parallel data, and create a GoURMET test set from these.

For those language pairs for which post-editing is conducted (see Section 5.2.1), post-editions will be used to expand the test sets to be used for the automatic evaluation.

[1] http://www.statmt.org/wmt19/translation-task.html
[2] http://matrix.statmt.org/matrix/systems_list/1904
[3] http://matrix.statmt.org/matrix/systems_list/1911

5 User-Centred Evaluation

5.1 Overview

User-centred evaluation emphasises the role of the user rather than the system and considers the needs of the end users. It focuses upon testing in a near-real-life scenario by giving test persons realistic tasks in a staged, yet realistic environment.

The responsibility for conducting user evaluation primarily falls to the media partners. The BBC and DW will ensure that, for the evaluation tasks described herein, a consistent evaluation process will be implemented, which is shared between both partners. i.e. user interfaces for performing the tests, online feedback forms, etc. are shared and identical between partners. The user evaluation process needs to be flexible enough to fit around live operations with as little interference as possible and as such a number of *options* are described here. It is not feasible for all tests to be undertaken for each language.

This section covers two broad types of user-centred evaluation which are informally referred to here as structured and unstructured evaluation methods. Structured evaluation methods use specially constructed user interfaces designed to allow the user to complete a single, contrived, task, examples are shown in Figures 1 and 2. This task is repeated numerous times to build confidence in the result. As a great deal of control over the evaluation can be exercised, results are directly comparable between media partners and across languages.

Unstructured evaluation will consist of a user (typically one or more journalists at one of the media partners) using a prototype tool to complete their everyday work, then metrics will be gathered regarding the efficacy of the MT. This may be done automatically (Section 5.2.3) or opinion may be purposefully solicited (Section 5.2.2). This latter evaluation type is closely aligned with Field Testing (Section 5.3) as both involve the use of prototypes.

In the case of DW, integration with news.bridge and SUMMA platforms is envisaged (for further details, see deliverable D5.2). SUMMA [4] is a cross-lingual media monitoring platform installed at DW, under further development and in beta testing at DW. It provides a cross-lingual overview of (text and video) content and provides ingestion, transcription, translation, summarisation and other data analysis in a fully automated manner. It currently works for 9 languages, with English as single target language. WD is aiming at expanding it to other languages, including GoURMET engines via API.

news.bridge [5] is a platform co-developed by Deutsche Welle and provides automated transcription, translation, subtitling and voice-over for selected video items in a large number of languages. DW is also aiming at enriching this platform with customised engines, such as those developed in GoURMET. Access of the platform to the GoURMET engines will be via API.

The BBC will build prototypes in order to address a subset of the use cases described in deliverable D5.2, which can also be found in this document, Section 5.3. Assessment will take place through these tools using techniques detailed in this section.

Thus, users will evaluate the MT quality and output through various channels/platforms/prototypes and the opinions of the users involved will be gathered in a range of ways as described in the following sections. Evaluation in this manner is an important step for the media partners as it

[4] http://summa-project.eu/
[5] http://newsbridge.eu/

indicates how *useful* the translation technology is in the real world.

5.2 User Assessment

These evaluations cover the process of asking users how they feel about the quality or other aspects of the translation.

It is important to make a distinction between different uses of the translated text and evaluate accordingly. For the GoURMET content creation use case, a major task will be post-editing machine translation of content - or assessing it for reuse purposes - so that it is an adequate translation of the source. For the GoURMET monitoring use case, the task is whether users can extract the gist of the meaning, or grasp the overall message, of the translated article. The evaluation will ask them to rate the reliability of the information: is it only good enough to provide the gist, or can all/most of the details in the text be relied upon, independent of grammatical mistakes and fluency? The use of machine translation for gisting purposes is very different when compared with (re)publication purposes requiring post-editing efforts.

The evaluation methods presented in this section are gathered under the headings of the two primary use cases according to how applicable they are in these two scenarios.

The third use case, International Business News Analysis, will be developed later on in the project and is envisaged to use similar methods as the one set for media monitoring.

Machine Translation Quality for Media Monitoring: Gisting and Understanding

Gisting is the use of a machine translated text to gain a general idea or understanding of the topics and focus of an original text. Flawless grammar in the machine translation results will not be the focus for this particular metric and full fluency and spotless grammar are not required. Evaluation of the quality of a translated text for gisting will be undertaken in two ways: opinion rating by the user and the (more objective) gap-filling methodology.

Ranking - DA using a Likert scale for fluency and accuracy

This entails opinion rating of readability and understandability and will confront monitors, journalists, and other media professionals with a translation of the item without looking at the source text and judge the translation to be readable and understandable and get a gist of the message. For this, testers do not necessarily need to master the source language, just the target language. For details on which specific aspects are envisaged to be assessed, go to Section 5.3.1)

Metric: Likert scale ranking readability and understandability.

Ranking - DA using a Likert scale for fidelity and informativeness

This will be expanded by an opinion rating of accuracy by comparing source and target text. Media professionals mastering both the source and target language (native in one of the two) will assess the accuracy in conveying the message accuracy from a monitoring/editorial point of view. This covers fidelity and informativeness. For details on which specific aspects are envisaged to be assessed, go to Section 5.3.1)

Metric: Likert scale ranking accuracy in conveying the message.

Gap Filling

Gap filling measures the level of comprehension of the translated document (Forcada et al., 2018). Document-level comprehension tasks are usually measured by reading comprehension questionnaires (RCQ). RCQ are very time consuming to create, and result in a low coverage of the documents content. It may also be quite laborious to evaluate the results of comprehension questions. Gap filling is seen as a more efficient and reliable way of extracting document-level comprehension scores. The gap-filling task starts with an evaluator reading a machine-translated article. Then they are presented with one or more human translated sentences from the same article where important content words are removed. They are asked to fill in the gaps. The quality of the machine translation is measured by the evaluator's success in filling in the gaps correctly, and how long it takes them. This will allow us to provide a quantitative measure for how much of the key information in an article is translated in a way that humans can understand. These exercises will require test users who are fluent in the target language (e.g. native speakers), no knowledge of the source language is needed. Specific languages will be tested in this way when the relevant language pairs have been released.

Figure 2 shows an example interface which can be used to test understanding of a text via gisting. The resulting metric is the accuracy of the user in guessing the gap correctly. Synonyms can optionally be allowed.

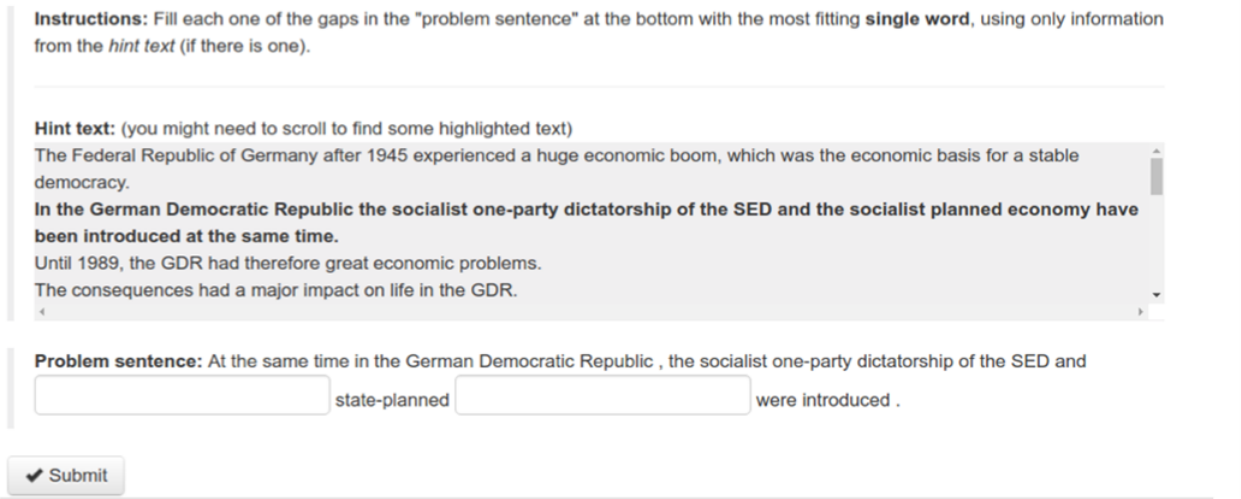

Figure 2: Example a gisting evaluation interface, with instructions, the translated text or hint text, and the problem sentence with the gaps that the annotator needs to fill

Metric: Accuracy in filling in the missing gaps. A factor that must be taken into account is the fact that some evaluators may be good at solving problems as a *puzzle* rather than as performing reading comprehension exercise, which may distort the actual measurement of usefulness.

5.2.1 Evaluation Methods for Content Creation

Machine Translation is used in the content creation workflow in a number of ways. The translated can be post-edited and then published as such, or parts of the content can be used in one or more new articles or serve as background material. Here fluency and fidelity are of extreme importance and thus the demands on the translation quality are higher. If it takes more time to post-edit it than to rewrite it or translate from scratch, MT has no use here.

The evaluation will measure MT quality for content creation in two ways: opinion rating by the user and post-editing usage metrics.

Ranking - DA using a Likert scale for fluency and accuracy

This entails an opinion rating of accuracy of translation for publication/reuse by comparing source and target text.

Thus, the first level of (subjective) assessment in this use case is to determine the accuracy of the translation by comparing source and target text and judge the suitability in terms of accuracy for publication in the target language or other reuse from an editorial point of view. This will be done by media professionals mastering both the source and target languages. For details on which specific aspects are envisaged to be assessed, go to Section 5.3.2)

Metric: Likert scale ranking accuracy and fluency of the translation for reuse.

Post-editing

The next level of assessment is task based using post-editing. Post-editing refers to the editing of mistranslations and grammatical errors in a machine-translated text to the point where it becomes a publishable translation of the source text. The idea is not to rewrite the text to one's own style, but correct to an acceptable, publishable quality. The evaluation will use automatically gathered metrics in terms of post-editing speed for instance, and a more subjective opinion rating on the usefulness and major obstacles. The evaluation will be performed by bilingual staff.

For the media partners, this assessment is crucial for understanding the usefulness of translated texts. It is hypothesised that the outcome of this assessment is likely to correlate with measures of translation quality reported from the tests described in Section 4.2.4. However, the strength of that correlation will be ascertained as part of this testing.

It is difficult to predict how this metric will behave. But the results are likely to vary from language to language as the machine translation algorithm makes different *types* of mistakes when translating different languages due to the differing grammar and/or structures of those languages. Different types of mistakes may be harder for a human editor to correct than others.

Two evaluation measurements will be applied in the post-editing process:

Automated usage analytics during post-editing will generate objective metrics. Post-editing speed can be measured automatically by the user interface used to perform the post-editing operation. Keystrokes or edit distance are alternative (objective) metrics here. The amount of text that has been changed will be measured.

In addition, opinion rating of the post-editing effort is an important measurement for the user partners, as it indicates whether the journalists/editors consider it a help or a nuisance. This will determine whether the tool will be applied in the workflow. For details on which specific aspects are envisaged to be assessed, go to Section 5.3.2)

It should also be noted that direct translation of news stories is not necessarily the default in the broadcasting sector. Deutsche Welle does translate some articles in full into other languages. Other articles are summarised or only partially translated or partly rewritten. Within the BBC, no news stories are ever directly translated from one language to another. As such, both media partners reversion content when translating from one language to another for publication. Reversioning is the addition or redaction of content, be that a large or small amount, after translating a text from one language to another for republication. Reversioning is common, and can be extremely important when translating news. News stories authored in one language are typically targeted at

one geographic region, culture, etc. As such there are many assumptions about the reader's prior knowledge, expectations, etc. inherent in the original text that may not hold in the translated text as this will be consumed by a reader who is unlikely to share the same above traits as someone reading the original text. As such, the journalist undertaking the reversioning operation not only translates the original text, but also seeks to add information or remove unnecessary detail to better serve the target audience. Reversioning in this manner is out of scope for the evaluation activites defined for the project because it is too difficult to control for. Instead, it is most reasonable for this evaluation process to consider only the translation step, within the reversioning process.

Metric: Average time taken to post-edit text, normalised by text length. Number of keystrokes. Amount of text changed. Likert scale ranking usefulness in terms of accuracy and fluency of the translation for reuse and the post-editing effort required.

5.2.2 Feedback Gathering

User observation

Observation sessions during scheduled sessions outside their operational hours for short periods of time. A member of the project team will observe the user and ask specific questions about the usability of the translations and the quality of the output (e.g. accuracy of proper nouns, omission of essential words, etc.). Test users will have experience in media monitoring or publication of news stories in the context of world news and will be fluent in the languages under consideration. The focus on particular languages depends on the order of language releases. Such observation can be combined with a brainstorming/thinking-aloud session, which could be recorded or at least feedback is noted.

Structured one-to-one interviews or surveys

Post-scenario evaluations include structured one-to-one interviews carried out to understand findings from the above observation sessions.

This are complemented by short surveys either via paper or online forms, requested after a user has performed a particular task. These can capture subjective opinions such as the usefulness of the translation and ease with which it was cleaned up (in the content creation case), but in a structured manner.

Feedback sheets will be identical across media partners and language pairs in order to promote consistency. Users will be asked a consistent set of questions and are required to answer from a consistent, fixed, 5-point Likert scale. A free text field may also be available.

Feedback forms allow the visualisation of results in the form shown in figure 3. This example is taken from a similar evaluation exercise which successfully evaluated the platform produced by the SUMMA[6] project.

Feedback can be gathered in a similar manner from within the user interfaces of the prototypes or tools developed. In this case a simple rating system will be used (e.g. rate the quality on a scale from 1 to 5) which can be integrated into the user interface for test purposes.

[6] http://summa-project.eu/

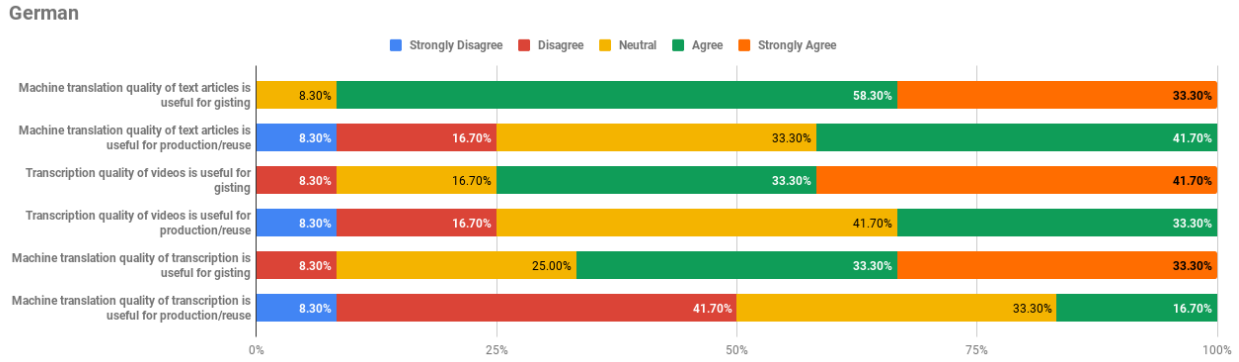

Figure 3: Example visualised result from feedback forms

5.2.3 Usage Analytics

Note that the above method of gathering feedback is distinct from *Usage analytics* as defined here. Opinion rating requires the user to consciously provide an opinion rating, whereas usage analytics gathers metrics automatically.

Usage analytics will be incorporated to capture actual usage statistics from the prototypes. This will add automated quantitative feedback to the qualitative focus of these tests. This can be easily gathered by the online tool which is used to allow the user to edit and clean the translated text. An illustrative example may include the automated gathering of how long it took a user to clean a piece of automatically translated text to a state where it is suitable to be published, i.e. post editing speed. Another examples is the tool PET[7], which logs every single keystroke.

5.3 Field Testing

For the avoidance of doubt, the evaluations that come under this heading can and will be undertaken in two ways. Bespoke user interfaces can be created to support one specific evaluation task (i.e. as shown in 1) while journalistic tools with the appropriate feedback mechanisms (automatically gathered usage analytics, user observations, etc.) will also be used. While list below enumerates the previously identified options for prototyping of journalistic tools, only a subset of these will be created and used to gather user evaluations. There are too many listed for all to be created, and the choice of which ones to use will be made by each media partner in collaboration with the research partners to balance a number of criteria such as business need within the media partner, availability of specialist staff for evaluation tasks, consistency across language pairs, particular insights required by the research partners, and so on.

The language proficiency required of the test users depends on the type of evaluation and the purpose. This is summarised as follows:

- Ranking (DA - Likert Scale 1-5)
  - For comprehension in monitoring - target language knowledge sufficient
  - For fidelity in monitoring - knowledge of source and target languages needed

[7] https://github.com/ghpaetzold/PET

  - For content creation - knowledge of source and target languages needed

- Task-Based Assessment
  - Gap filling (reading comprehension) - target language knowledge sufficient
  - Post-editing (content creation) - bilingual testers needed

5.3.1 Media Monitoring Use Case

These use cases are evaluated for gisting purposes primarily.

USE CASE A: Improving internal visibility - BBC

- gap-filling
- opinion ranking - Likert scale
  - rating usefulness for gisting
  - rating fidelity
  - rating understandability
  - rating readability
  - rating informativeness

Example A1: Azeri service access to BBC Persian website
Example A2: Everything accessible to everyone in a single bureau
Example A3: Optimising visibility in internal content systems
Example A4: Allowing experienced journalists to feed back
Example A5: Facilitating greater ability to share
Example A6: A system of alerting journalists to stories doing well across languages

USE CASE C: Editorial oversight - BBC

- gap-filling
- opinion ranking - Likert scale
  - rating usefulness for gisting
  - rating fidelity
  - rating understandability
  - rating readability
  - rating informativeness

Example C1: Editorial oversight (online news)
Example C2: Editorial oversight of compliance (TV)
Example C3: Editorial oversight (Social media)

USE CASE D: Media insight - BBC

- gap-filling
- opinion ranking - Likert scale
  - rating usefulness for gisting
  - rating fidelity
  - rating understandability
  - rating readability
  - rating informativeness

Example D1: Kurdish
Example D2: North Korean

USE CASE G: Translation for Cross-Lingual Media Monitoring - DW

- opinion ranking - Likert scale
  - rating usefulness for gisting
  - rating fidelity
  - rating understandability
  - rating readability
  - rating informativeness
- gap-filling

Example G1: Internal monitoring
Example G2: Enabling monitoring of DW content externally

5.3.2 Content Creation Use Case

USE CASE B: Increased workflow efficiency for reversioning output - BBC

- opinion ranking - Likert scale
  - rating usefulness for reuse
  - rating fidelity
  - rating understandability

  - rating readability
  - rating informativeness
  - rating post-editing effort
  - rating grammatical correctness
  - rating idiomatic correctness
- post-editing usage analytics
  - post-editing speed
  - number of changes made in post-editing, e.g. keystrokes or edit distance

Example B1: Reducing time to Broadcast
Example B2: Efficient subtitling of video content
Example B3: Subtitle and dubbing script computer-assisted translation for TV

USE CASE E: Research and experimentation with automatically produced content - BBC

- opinion ranking - Likert scale
  - rating usefulness for reuse
  - rating fidelity
  - rating understandability
  - rating readability
  - rating informativeness
  - rating post-editing effort
  - rating grammatical correctness
  - rating idiomatic correctness
- post-editing usage analytics
  - post-editing speed
  - number of changes made in post-editing, e.g. keystrokes or edit distance

USE CASE F: Translation and Adaptation for Content Creation - DW
These use cases are evaluated for the purpose of publication in the target language.
Assessment methods used:

- opinion ranking - Likert scale
  - rating usefulness for reuse
  - rating fidelity
  - rating understandability

  - rating readability
  - rating informativeness
  - rating post-editing effort
  - rating grammatical correctness
  - rating idiomatic correctness
- post-editing usage analytics
  - post-editing speed
  - number of changes made in post-editing, e.g. keystrokes or edit distance

Example F1: Translate English or German content for smaller language departments
Example F2: Translate content from low-resourced languages into English for reuse
Example F3: Support our new Turkish YouTube Channel
Example F4: DW distribution to partner broadcasters
Example F5: Distribution to other regions with translated subtitling

6 Time and language plan

This sections contains the timeplan of the evaluation and the list of languages which will be evaluated in the project.

Evaluation begins in Year 1, as soon as the first languages are available (Milestone MS1, around M7). Evaluation will take place regularly on the release of new languages and upon a significant update of an existing language. There are four delivery and evaluation cycles over the period of the project.

6.1 Time plan

The time plan for evaluation, shown in Table 1, is derived from the GoURMETproject plan, technical annex 3.

Months	**Related Milestone-/Deliverable**	**Activity**
M6	MS1	First release of translation models (Turkish, Swahili, Bulgarian and Gujarati)
M6–M9	MS3, D5.4, D5.6	Data-driven evaluation of Turkish, Swahili, Bulgarian and Gujarati translation models
M8–M18	D5.4, D5.6	User-centred evaluation of Turkish, Swahili, Bulgarian and Gujarati translation models
M15	MS15	Second release of translation models
M15–M18	MS8, D5.4, D5.6	Data-driven evaluation of second set of translation models
M16–M18	D5.4, D5.6	User-centred evaluation of second set of translation models and platform
M24	MS16	Third release of translation models including surprise language
M24–M27	MS10, D5.6	Data driven evaluation of surprise language
M26–M36	D5.6	User-centred evaluation of surprise language
M24–M27	MS18, D5.6	Data driven evaluation of third set of models
M26–M36	D5.6	User-centred evaluation of third set of models and platform
M33	MS21	Fourth release of translation models
M33–M36	MS22, D5.6	Data driven evaluation of fourth set of models
M35–M36	D5.6	User-centred evaluation of third fourth set of models

Table 1: Evaluation activity time plan, as derived from project plan

Note the user-centred evaluation of each set of models is delayed from the release of the models themselves, and therefore the start of the data driven evaluation. This is deliberate and is to allow time for integration of the models into the infrastructure that powers the shared interfaces, technical testing of the models and integration, and any updates to the user interfaces required to support the user-centred evaluation.

6.2 Languages

Four languages have been selected for the first development and evaluation phase: Turkish, Swahili, Bulgarian and Gujarati.

The second, third and fourth sets of language pairs will be decided prior to the data gathering activities to begin at months 10, 20 and 28 respectively.

Languages will be primarily selected from the set previously agreed and published in deliverable D1.1 Table 1. In exceptional cases, language pair not in this list may be selected if it is a) of particular strategic importance for one or both media partners, b) fulfils the project criteria of being low resource and c) is of particular interest and/or will drive particular high impact for one or more research partners. A ”surprise language”, not previously stated in deliverable D1.1 Table 1 and not previously agreed with research partners will be selected by media partners for the third cycle (i.e. just prior to M20).

7 Conclusion

This document contains the evaluation methodologies and plan for the GoURMET machine translation technologies. Evaluation will be an ongoing process throughout the lifetime of the project.

Different categories of evaluation were described, including data-driven evaluation using automated metrics, shared tasks, DA, and user evaluation and field testing.

The academically-focused evaluation will generate results for well-known metrics which will allow the comparison of translation quality and resulting usefulness between the GoURMET machine translation technologies and others in this field. Use-case-driven assessment will also take place which will allow the media partners to understand how well the translation technology is likely to work in the real world. After testing, it will be possible to determine how correlated the results of the objective tests of translation quality compare with the more subjective results of user tests.

References

Forcada, M. L., Scarton, C., Specia, L., Haddow, B., and Birch, A. (2018). Exploring gap filling as a cheaper alternative to reading comprehension questionnaires when evaluating machine translation for gisting. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 192–203.

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2017). Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

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

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

ENDPAGE

GoURMET

H2020-ICT-2018-2 825299

D5.1 Evaluation Plan