



Global Under-Resourced MEedia Translation (GoURMET)

H2020 Research and Innovation Action

Number: 825299

D5.4 – Initial Progress Report on Evaluation

Nature	Report	Work Package	WP5
Due Date	30/06/2019	Submission Date	30/06/2019
Main Authors	Andrew Secker (BBC), Julie Wall (BBC), Mikel L. Forcada (UA), Barry Haddow (UEDIN), Antonio Miceli Barone (UEDIN), Susie Coleman (BBC), Anna Blaziak (BBC), Peggy van der Kreeft (DW)		
Co-authors			
Reviewers	Mikel L. Forcada (UA)		
Keywords	evaluation, gap filling, direct assessment		
Version Control			
v0.1	Status	Draft	20/06/2020
v1.0	Status	Final	30/06/2020



Contents

1	Introduction	5
1.1	Workpackage 5 Context	5
2	Evaluation Methodologies	7
2.1	Automatic Evaluation	7
2.1.1	Evaluation architecture	8
2.1.2	Comparison with Google Translate	8
2.2	Human Evaluation	8
2.2.1	Evaluation levels	9
2.2.2	Gold Standard Evaluation	9
2.2.3	Gap Filling and Direct Assessment Evaluations	10
2.2.4	Evaluators	10
2.2.5	Data Preparation	11
2.2.6	Translators	11
2.2.7	Direct Assessment: Details	12
2.2.8	Gap Filling: Details	16
3	Interfaces for Human Evaluation	20
3.1	Direct Assessment Evaluation Tool	21
3.2	Gap-Filling Evaluation Tool	23
3.3	Open-source Releases	24
4	Results of Data Driven Evaluation	26
4.1	Test Sets	26
4.2	Summary of the results	28
4.3	Language details	30
4.3.1	Bulgarian	30
4.3.2	Turkish	30
4.3.3	Swahili	30
4.3.4	Gujarati	30
4.3.5	Tamil	30
4.3.6	Serbian	31
4.3.7	Amharic	31
4.3.8	Kyrgyz	31

5	Results of Human Evaluation	32
5.1	Gap Filling	32
5.1.1	Investigation into the Postprocessing of Evaluation Scores	33
5.1.2	Summary of Results	34
5.1.3	Bulgarian	35
5.1.4	Turkish	35
5.1.5	Swahili	36
5.1.6	Gujarati	36
5.1.7	Serbian	38
5.1.8	Kyrgyz	39
5.2	Direct Assessment	41
5.2.1	Bulgarian	41
5.2.2	Turkish	44
5.2.3	Swahili	44
5.2.4	Gujarati	48
5.2.5	Serbian	50
6	Research Outputs	52
6.1	Publications	52
6.2	Software	52
7	Conclusion	53

List of Figures

1	Custom Direct Assessment interface.	13
2	DA evaluator feedback screen	16
3	Custom Gap Filling interface.	17
4	Architecture for DA evaluation tool	20
5	Direct-assessment tool — submitting test dataset	22
6	Direct-assessment tool — exporting results	22
7	Gap filling tool - upload test data screen	24
8	Gap Filling tool - export results screen	24
9	Accuracy results for GF evaluation for bg → en	36
10	Results of GF evaluation for tr → en	37
11	Results of GF evaluation for sw → en	38
12	Results of GF evaluation for gu → en	38
13	Results of GF evaluation for sr → en	39
14	Results of GF evaluation for ky → en	40
15	Results of DA evaluation for en → bg , mean scores for question Q1 – x-axis lists the users per organisation and the y-axis is the scoring by users	42
16	Results of DA evaluation for en → tr , mean scores for question Q1 – The x-axis lists the user per organisation, the y-axis gives user scoring of MT output (blue bar) and user scoring of the ground truth sentence (orange bar)	45
17	Results of DA evaluation for en → tr , mean scores for question Q2	46
18	Results of DA evaluation for en → sw , mean scores for question Q1	47
19	Results of DA evaluation for en → sw , mean scores for question Q2	48
20	Results of DA evaluation for en → gu	49
21	Results of DA evaluation for en → sr	51

1 Introduction

This document describes the strategy, the plan and the methodology for evaluation, and interim results for both automatic and human evaluations are presented. This document follows on from the previously published Evaluation Plan (GoURMET Deliverable D5.1, Secker et al. (2019a)).

Evaluation of the research technologies is a key aspect of the GoURMETproject. The project has two complementary aims in this regard

- Understand the quality of the translation in a data driven manner such that the translation models developed can be directly compared with other translation systems using shared test sets.
- Understand the quality of the translations from the user’s perspective. using strategically designed tests to determine the quality of translations using the pinions of human experts.

The remainder of this document is structured as follows:

The overarching evaluation strategy is initially described in Section 2, which is subdivided into the methodology for automatic evaluation and human evaluation. These are described in detail in Sections 4 and 5 respectively.

The interfaces which have been created by the project to support the human evaluation are described, along with a short technical overview, in Section 3.

The results of the evaluations are then presented in Sections 4 and 5, broken down by language pair to give a complete record of the initial results.

1.1 Workpackage 5 Context

The aim of GoURMET WP5 is to take the output of the research partners in the form of translation models, make them available to a media partners’ production environment for the production of prototype tools and experiences, and evaluate the quality of the translation models. Human evaluation should take place using the tools produced under WP5 or model the two use cases of global content creation and media monitoring in a more controlled environment but still in domain.

This deliverable is an interim report on Tasks T5.4 and T5.5 from the GoURMET description of work. Not all parts of the tasks are covered in this interim deliverable. These will be in the updated deliverable, Deliverable D5.6, due M36.

Task T5.4: Media monitoring: user evaluation

This task is led by DW —but covered by both user partners— and is targeted towards evaluating machine translation for monitoring and discovering news in lower-resourced languages. It will encompass both specific assessment of the different components and technologies, from a user point of view, in close coordination with the technical partners, as well as evaluation of the overall, integrated platform. It will involve real users covering a variety of functions dealing with or benefiting from monitoring news, in the language departments of the GoURMET language package, as well as other consuming language departments. Usefulness, usability and user-friendliness

as well as specific language quality control will be assessed. The methods applied will be coordinated by both users partners, under the leadership of BBC, as described in Task T5.5

Task T5.5: Global content creation: user evaluation

This task is led by the BBC, as they focus primarily on the global content creation use case. However, both the BBC and Deutsche Welle will participate in the evaluation tasks, provide feedback and contribute to the evaluation deliverable. Feedback will be provided on all testing. Specific user validation workshops will be organised for final prototype validation. There will be close collaboration among user partners as to user evaluation in order to provide consistent user feedback to the research partners.

The BBC shall lead the creation of a detailed evaluation plan at the early stages of the project, defining the testing and validation methodology of the platform based on the technical and functional requirements defined in [Deliverable D5.2 Secker et al. (2019b)]. Different target groups are addressed: multilingual journalists, monitors, analysts, editors, and other media professionals involved in the content creation or media monitoring process will be attracted as target test user groups for evaluating the research.

Evaluation shall be split between MT language-based trials and field trials. MT language-based trials represent an important means of accurately calibrating the performance of specific technologies. Wherever possible, we shall use standard data sets and evaluation protocols, participating in appropriate international technology evaluation whenever possible such as the annual shared task in the Workshop on Statistical Machine Translation (<http://www.statmt.org/wmt18/>) also organised by UEDIN. Field trials are a step beyond the main prototyping cycle of development and testing. With the field trials integrated and tested user interfaces will be taken and trialled with real users in the actual working environment for limited periods of time. Field trials may be limited in functionality, languages, research subjects/areas or content types. Field trials will be an opportunity to assess the impact of the new tools on the larger workflow and operation. From this we will gain insights into how translation technologies such as these are used in a real-world environment. For example, we will attempt to ascertain how accurate MT must be for it to be useful across our varied use cases. It is expected that this would vary between both uses cases and languages.

2 Evaluation Methodologies

This section describes the overarching evaluation strategy. Details of the individual methodologies for the data driven and human evaluations are the described in detail in sections 4 and 5 respectively.

2.1 Automatic Evaluation

Automatic evaluation assesses the quality of a machine translation system by automatically comparing its output translations to reference translations. This enables a quick, cost-effective and reproducible evaluation of a system since, unlike human evaluation, it does not require annotators to directly assess the outputs of the system. However, the ultimate goal of a translation is to fluently and accurately convey the meaning of the source text to users in a language they understand, which is most accurately assessed by human evaluation protocols rather than automated tests. Therefore, automatic evaluation does not replace, rather it complements human evaluation.

Automatic evaluation requires a choice of a test set and an evaluation metric. The test set is a set source sentences with one or more reference translations. Using multiple references can in principle improve the correlation between the evaluation and the value to the user, but in practice obtaining multiple references is expensive and therefore it is not often done in machine translation research. In the GoURMET project in particular we have access to limited amounts of data, therefore we use single reference translations.

The evaluation metric is a function that computes a text similarity score between the generated and reference translations. In this project we use two standard metrics: BLEU and chrF.

- BLEU (Papineni et al., 2002) is the most common automatic evaluation metric for machine translation reported in the scientific literature. Despite its age and simplicity, BLEU still correlates fairly well with human quality judgements, therefore it is still widely used as the primary, and often unique, evaluation metric in most research papers. It is based on a modified precision computed on word n -grams and corrected by a brevity penalty.
- chrF (Popović, 2015) is a metric based on weighted F-scores computed on character n -grams. It has been found to strongly correlate with human quality judgements consistently over different languages and test sets (Ma et al., 2019). Because it is based on characters, chrF is able to give partial scores to word forms which are not in the same morphological form as the reference translation. This is beneficial in the case of morphologically-rich languages (such as Turkish and Kyrgyz) of interest to the project.

Different implementations of these metrics exist that can produce slightly different results depending on sentence segmentation, word tokenization and other details. In order to maximise reproducibility and consistency with the scientific literature, we use the implementation of BLEU and chrF provided by the SacreBLEU tool (Post, 2018) which has been designed specifically for reproducibility and is widely used.

Other evaluation metrics have been proposed in the machine translation literature and they are being evaluated each year in the WMT Automatic Metric shared task (Ma et al., 2019), in some cases obtaining higher correlation with human judgements than BLEU or chrF, but they have drawbacks such as high computational cost, lack of publicly available implementations, limited supported

languages, use of machine learning to train the metric (which calls into question their ability to generalise out of their training distributions), and so on. BLEU and chrF are simple, robust, fast, applicable to any language, have standard highly reproducible implementations in SacreBLEU and are widely used, therefore we chose them for the GoURMET project.

2.1.1 Evaluation architecture

In order to perform automatic evaluation, we collected a repository of test sets for the GoURMET project language pairs. The tests sets consist of internal parallel sets of sentences with their translations extracted from the BBC and Deutsche Welle websites and validated by human annotators, as well as public test sets where available. The details of the tests sets are described in deliverable D5.3 (Secker et al. (2020)), section 2.

We collected our test sets in the SFTP data repository hosted on the "Valhalla" cluster of the University of Edinburgh. We translated each test set using the Translation Service System Architecture described in deliverable D5.3 section 5 (Secker et al., 2020).

We queried the system using the same API designed for production in order to make sure that our evaluation results are as consistent as possible with the actual use case.¹ Specifically, we sent untokenized source text and received untokenized translations, letting the Translation Service handle tokenization and detokenization internally. Finally, we computed BLEU and chrF scores on the automated translations using the SacreBLEU tool.

2.1.2 Comparison with Google Translate

We compare our system with the commercial machine translation system provided by Google, at a cost of approximately \$20 per million characters. We submit our test sets to the Google Translate service using their API and we compute BLEU and chrF scores using the SacreBLEU tool.

When comparing the Google scores with our system, it must be noted that we cannot exclude that our test sets were contained in the training sets used by Google, since they were extracted from data publicly available on the web. This could lead to artificially inflated scores for the Google system.

2.2 Human Evaluation

Human evaluation indicators involve the participation of humans and either collect subjective feedback on the quality of translation or measure human performance in tasks mediated by machine translation.

Wherever possible, manual evaluation undertaken within the GoURMET project uses in-domain data, i.e. test data derived from news sources.

¹ except for the English–Tamil system which is still in development and has not been integrated in the Translation Service at the time of this writing.

2.2.1 Evaluation levels

Within the project it has been agreed to divide the human evaluation of the MT models into 3 ascending levels — bronze, silver and gold. The levels relate to increasing amounts of relevance to the media partners’ work, at increased effort placed on evaluating at that level.

Bronze standard evaluation requires gap-filling and direct-assessment exercises (see section 2.2.3 below) to be undertaken with the minimum number of evaluators required to produce a meaningful result.

Silver standard evaluation requires gap-filling and direct-assessment exercises to be undertaken with more evaluators, increasing the confidence of the result.

Gold standard evaluation is the evaluation of post-editing effort, with indicators such as post-editing time, number of edits, or a related metric to be defined.

All language pairs are subjected to at least a bronze standard evaluation. Bronze or silver standards differ only in the number of evaluators available to evaluate the translation models. Small language services with few staff may only be able to support bronze evaluation. Note there is no difference in the actual methodology between bronze and silver evaluations. Only the number of times a specific sentence is evaluated.

Additionally:

- Where possible (i.e., that language is used by both media partners), both media partners will contribute to the evaluation
- Best efforts will be made to reach silver evaluation by both media partners.²

Details of the parameters for bronze and silver evaluation follow in Sections 2.2.7 or 2.2.8 for the selected gap-filling and direct-assessment human evaluations.

2.2.2 Gold Standard Evaluation

This section describes the plan and motivation for gold standard evaluation. Gold standard evaluation will be evaluation within a realistic news production environment. One of the stated aims of the project is to continue to investigate how the metrics found in the literature, both for data-driven (automatic) and human evaluation, translate into the real world. A stated aim in this regard from Deliverable D5.1 Secker et al. (2019a) is to establish how less expensive³ evaluation methods may be used as proxies for real-world usefulness of MT models. The gold-standard evaluation will estimate real-world usefulness by using post-edit metrics as an indicator. This process may build upon the general findings in Scarton et al. (2019).

If the bronze and or silver evaluations have revealed the language pair to be of sufficient quality and thus of realistic benefit in a production workflow, a gold standard evaluation, which involves looking at post-editing metrics, may follow.

For gold standard evaluation we will create a realistic workflow scenario and compare manual translation from scratch with machine translation post-editing. Metrics for this evaluation are not

² If having met the bronze evaluation level it becomes clear that translations are not sufficiently meaningful, no further human evaluation of that MT model is done until the model is updated.

³ expensive in the sense of time/effort

yet defined, but likely to include measurements such as time taken, number of key strokes as well as subjective reviewing of quality of writing and storytelling in the output. Prototype tools and experiences will be designed to capture these metrics.

The production of prototypes, deployed by the media partners and capable of gathering post edit metrics requires both MT models with sufficient quality to be useful and the foundations of technical work in the form of a translation service to be in place. The translation service is described in Deliverable D5.3 Secker et al. (2020). Prototypes capable of supporting the capturing of post-edit metrics (gold evaluation) will be developed using the GoURMET translation service over the remainder of the project, and gold standard results will be presented in the subsequent evaluation deliverable (D5.6, due M36).

2.2.3 Gap Filling and Direct Assessment Evaluations

Two methods of evaluation will be applied to all MT models in order to generate the most relevant insight as they a) closely match the two project use cases of global content creation and media monitoring and b) are established evaluation methods from the literature, ensuring a certain amount of comparison is possible with other MT systems:

Direct assessment (DA) (Graham et al., 2016a,b, 2013) is used to test translation from English into the non-English language. This corresponds to the content creation use case which will use translation predominantly in this direction, and where the correctness of the translation is key.

Gap filling (GF) (Forcada et al., 2018b) is used to test translation from the non-English language into English. This corresponds to the media monitoring use case which will use translation almost exclusively in this direction and where getting the gist of the meaning of a sentence is enough to fulfil the use-case, perfect translation of sentence structure is less important.

Thus, each language pair will be subject to both DA and GF assessments. However, GF will only apply from non-English into English and DA will only apply from English into non-English.

2.2.4 Evaluators

This subsection defines the agreed guidelines for engaging evaluators across all evaluation types.

Evaluators are recruited from within the media partner organisations to complete the evaluation tasks. There may be rare occasions when the media partners are unable to supply the number of staff required to complete the evaluations, especially in the cases where the language services are small. In these instances evaluators will be sourced from outside the organisation. Care will be taken to ensure external evaluators meet the criteria applied to internal evaluators.

Evaluators are required to have an excellent level of comprehension in the language(s) relevant to the evaluation they are undertaking. In the case of gold standard, the evaluator must also be proficient at translation, that is, able to assess how good a translation of the source the target is. Wherever possible, evaluators will be employed or contracted in a journalistic role, with experience of writing and/or subediting⁴ content that is published or broadcast in the relevant language.

⁴ Subediting is the processes of checking written content before publication, typically for typographical errors or mistakes in spelling and grammar.

2.2.5 Data Preparation

This subsection defines the agreed guidelines for preparation of evaluation data across all relevant evaluation types.

Wherever possible, evaluation data sets are prepared using articles previously published by BBC or DW. In limited cases it may be necessary for a research partner to source evaluation data. This should also be drawn from the news domain, ensuring consistency across testing.

For the creation of evaluation data sets, two techniques have been used thus far. When curating the evaluation data, each sentence is either a) translated into English by a multilingual journalist employed by BBC or DW with excellent comprehension of the source and target languages or b) automatically identified in a parallel text and validated by a language expert to ensure the quality of sentences is equivalent to that of a human translator. In rare cases where this might not be possible, a professional translator will be employed for the task. Above all, whichever method is used, this step requires the highest translation quality possible.

Evaluation data are used across both GF and DA evaluations. Some sentences may appear in both evaluations, however no evaluators completed both the GF and DA for the same language pair so any one sentence was seen at most once by a single evaluator.

- Text will be sourced from five or more different stories.
- Source material will come either from BBC, DW, or an equivalent News output in the unlikely event this is required.
- BBC and DW will share the evaluation data and translation effort to generate them (as journalist availability for each language allows).
- Care will be taken to source evaluation data such that it can be guaranteed that no evaluation data are included in the training data.

2.2.6 Translators

Individuals who are tasked with translation either for the generation of evaluation data or for future gold standard evaluation are:

- Fluent in the non-English language of the pair (or one of the two languages where English is not part of the pair).
- Fluent in, and able to write to media partner output standards in, English (or second language of the pair where there is no English)
- Employed or contracted by a media partner or alternative with equivalent standard of output, in a journalistic role with experience of writing and subediting. news based content that is published or broadcast on a media partner's platform.

2.2.7 Direct Assessment: Details

Direct assessment entails asking a human evaluator, fluent in the target language, to rate the, readability and understandability and correctness of a machine translated sentence compared to a reference translation.

Evaluation material is prepared as described previously, to create a pool of high-quality parallel sentence data consisting of a minimum of 205 parallel sentences, for most languages derived from DW or BBC content. This produces two sets of 100 sentences each, plus 5 control sentences. 205 sentences are required as this is the number used by the DA tests. The GF use a subset of these, i.e. the evaluation data are shared between GF and DA.

All evaluators are asked to rate the quality of the machine translated sentence compared with a comparison on a sliding scale from 0% to 100% for two criteria according to the statement “*For the pair of sentences below read the text and state how much you agree that:*”

With the criteria being:

Q1 *The black text adequately expresses the meaning of the grey text*

Q2 *The black text is a well written phrase or sentence that is grammatically and idiomatically correct.*

DA, as run by Graham et al. (2016b), etc., only asks a single question (Q1), and, therefore, it is acknowledged that the evaluation methodology here departs slightly from common usage. However, the wording of these questions is chosen in order to elicit opinion on two related but distinct attributes of the translated sentence which correlate with the project’s two primary application use cases. It is also specifically directed towards editorial users.

The opinion expressed against the *Q1* criterion can be correlated with the expectations of a user in the *media monitoring* use case. That is, for a journalist working in media monitoring, the ability to gain an understanding of the meaning of a sentence is often sufficient. In reality, the exact wording of that sentence does not have to be particularly good for a journalist to be able to understand whether something requiring further research is necessary, and take the appropriate action.

The opinion expressed against the *Q2* criteria can be combined with *Q2* and correlated with the expectations of a user in the Global Content Creation use case. In this use case, it is expected the end user will be using the translation for direct use in news articles to be published. These articles need to be grammatically and idiomatically perfect upon publication, and, therefore, the closer the translations can get to this standard, the better, as this will reduce the effort required by the journalist to clean the translated text to the required standard.

Figure 1 shows the interface by which the opinions for Q1 and Q2 are given. Sentence pairs are evaluated on a sliding scale with 0% and 100% marked at the extremes. No other score indication or guidance is shown, other than quarter point marks of the slider scale for the evaluator’s reference.

Evaluators are expected to complete a language set (containing both series for that language) in a single sitting. The evaluations will not be timed and there will be no time limit. Evaluators are not be asked to evaluate a test set containing a sentence pair where they produced the manual translation.

At the end of the test there will be an opportunity for evaluators to provide free comments on the machine translation set.

For the pair of sentences below: Read the text and state how much to agree that:

Wakati mji hii anapotimua mbio kukwepa wanaomuwinda hukusanya nguvu na mwendokasi wa kumwezesha kukimbia akitumia miguu yake ya nyuma.

Wakati ilizard hii ni waliokimbia kutoka predators ni inakusanya kasi na kubadilika na kukimbia juu ya miguu yake miwili ya nyuma.

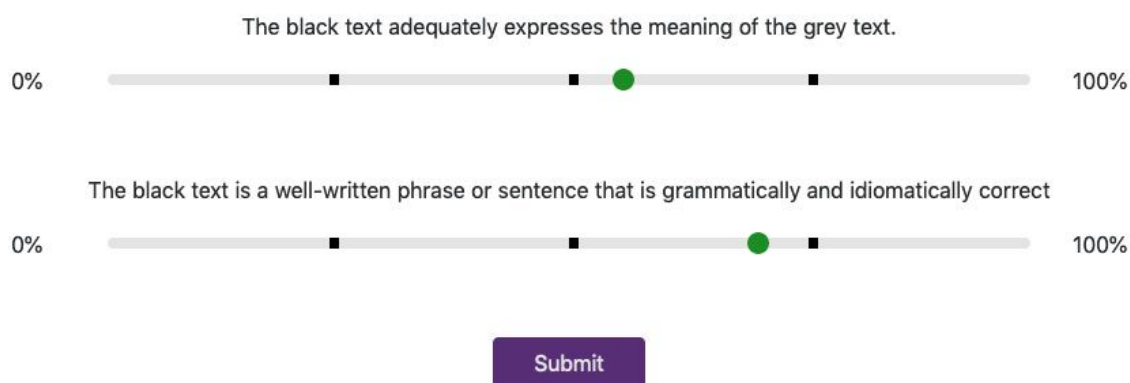


Figure 1: Custom Direct Assessment interface.

Evaluation levels for Direct Assessment

The differences between levels are as follows:

Bronze standard: Each pair will be rated exactly twice (400 responses)

Silver standard: Each pair will be rated three or more times (600 responses or more)

Data for Direct Assessment

For all DA evaluations, the data set shown to each evaluators consist of:

- practice examples
- true test example
- calibration examples.

To create the evaluation data set, 205 sentences are drawn at random from 5 ore more different news articles. From this set, 5 of these are chosen at random and set aside as practice example, leaving 200 true test sentences.

A further 5 sentences are written in the target language by a human and used as *calibration* examples (see below). Thus, each test set for direct assessment contains a total of 210 sentences.

The sentences were shown to the evaluators as follows:

Sentences 1–5: Practice sentences. The opinion expressed by the evaluator for these is discarded.

Sentences 6–110: Evaluation sentences 1-100 presented in a random order, plus 5 calibration sentences presented at random points throughout.

Break

Sentences 111–115: Practice sentences (repeated). The opinion expressed by the evaluator for these is discarded.

Sentence 116–220: Evaluation sentences 101-200. presented in a random order, plus 5 calibration sentences (repeated) presented at random points throughout.

As such,

- The 5 practice examples will be seen twice
- The 5 calibration examples will be seen twice
- The 200 true evaluation sentences will be seen only once each

Calibration data

Calibration data are used to ascertain a dynamic range for an individual evaluator’s stated opinions. That is, it allows to control between individual evaluators where individuals may rate particularly high or low, based on their prior expectations, and thus give a more consistent method of comparing translation quality *between* languages. The calibration data allows indication of how individual evaluators are scoring when there is an excellent linguist able to manually create pairs to provoke an expected result.

Note that Graham et al. (2016b) used z -standardisation to normalize scores between evaluators. This is useful when working with a large number of diverse, crowd-sourced annotators, comparing several systems, with an unbalanced assignment of annotators to systems. z -standardisation ensures that (e.g.) if a system’s output is disproportionately evaluated by a harsh evaluator, it is not penalised. However, here we are interested in collecting opinions on a single system from a small number of reliable annotators, so we report absolute scores.

It can be seen how evaluators score in relation to the expected result – high, low, seemingly random.

If evaluators seem not to correspond with expected results for the calibration pairs, when others do, it might suggest that the evaluator has not understood the test. We always assume the evaluator is doing his or her best.

During evaluation, the calibration sentences are shown to the evaluator twice in order to give some indication of how consistently the evaluator is rating.

The calibration sentences are sourced as follows.

The calibration data are created in non-English only, as DA only took place with the target as non-English. Experts in the target language (i.e., an experienced news journalist in a non-English language service) are asked to select an item from the target language output, and select five different sentences.

For three sentences chosen at random, the language expert is requested to write an alternative sentence (a paraphrase) that conveys the same information and detail in perfect grammar. A sentence which meets the evaluation criteria (Q1 and Q2 above) would be expected to score close to 100% for both criteria.

An example in English for sentence pairs 1–3. The source sentence:

Provisional figures released by the World Meteorological Organization (WMO) suggest this year is on course to be the second or third warmest year ever.

Is re-written as:

This year is on course to be the second or third warmest year ever, according to provisional figures released by the World Meteorological Organization (WMO).

Once written, the language expert is then requested to provide the expected score on the two criteria, as it may be impossible to re-write the sentence and still score 100% on both.

The above examples are given in English. However, as stated previously, the DA task replicates translation from English into a non-English language and as such, all calibration sentences will be generated in non-English.

Evaluators

Evaluators are recruited from within the media partners wherever possible according to the guidelines described previously in the parent section.

For DA, evaluators must be fluent in the non-English target language. This does impose some restrictions on the pool of evaluators available from the media partners, especially for smaller languages. As the number of evaluators available for DA is limited, the number of sentences each is required to evaluate is large in order to compensate and ensure the results are meaningful.

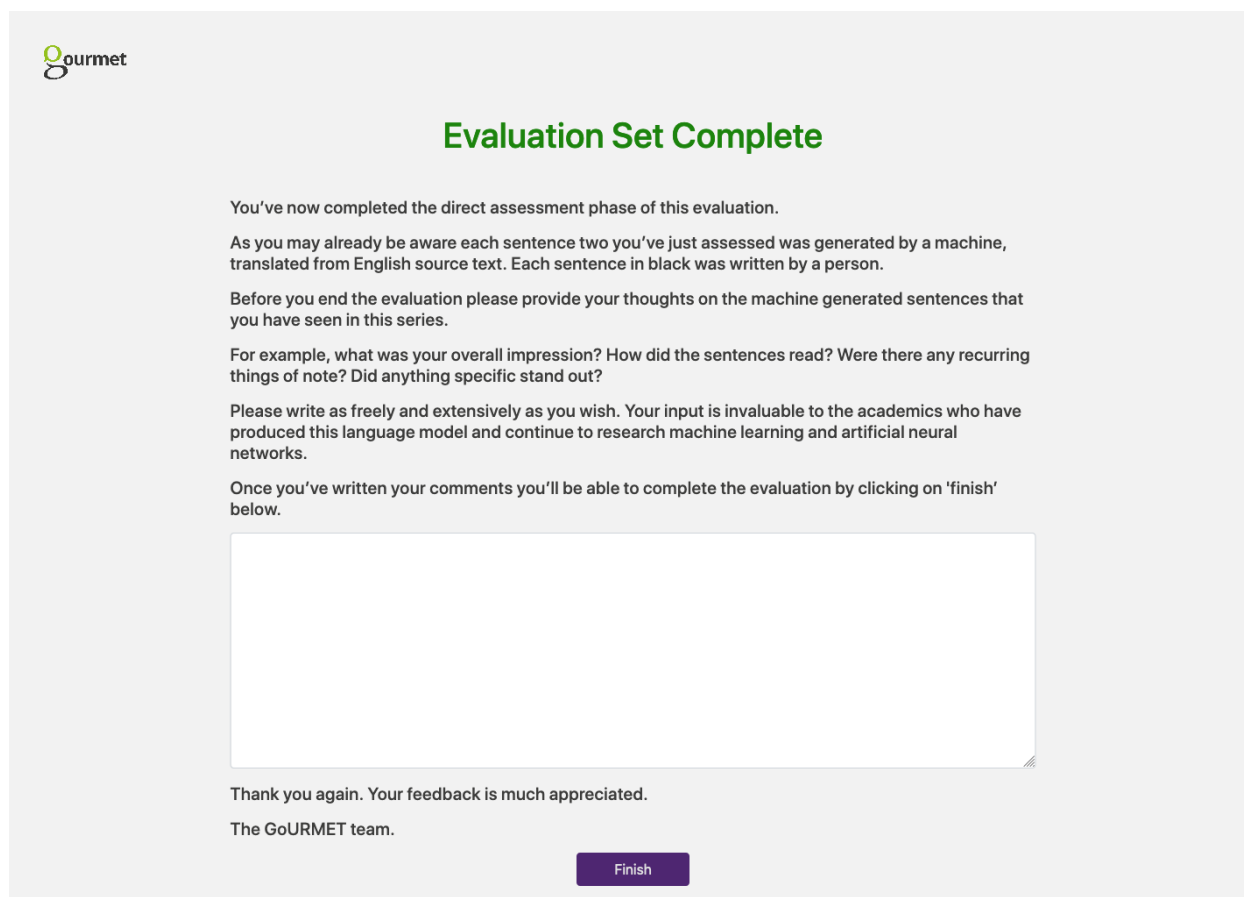
Evaluators are instructed to complete the task according to the following guidelines:

- For each pair you will be asked to use a slider to rate how strongly you agree or disagree that the alternative sentence accurately expresses the meaning of the model sentence and is well written.
- The final part of the evaluation is an open feedback section where you can share your thoughts about the quality and style of the sentences you have rated.
- The evaluation process is not timed and there is no time limit. However each series should be completed in an unbroken single sitting.
- Please start the series when you will be able to give it your full attention to completion.
- If you have been asked to complete multiple series you don't need to do them in one continuous sitting.

Gathering of Subjective Opinions At the end of each direct-assessment task, evaluators are asked to give their opinion on the translation they have seen (Figure 2). This is obtained via a free-text field presented by the evaluation tool.

The length of the feedback is not constrained and the evaluator is not compelled to write any feedback. It is acknowledged therefore that the sample is self-selecting.

Summaries of these opinions are presented in the relevant results subsections.



The screenshot shows a web interface for the GoURMET project. At the top left is the GoURMET logo. The main heading is "Evaluation Set Complete" in green. Below this, the text informs the user that they have completed the direct assessment phase. It explains that sentences were generated by a machine from English source text, with some in black being human-written. The user is asked to provide feedback on the machine-generated sentences. Examples of feedback questions are provided, such as "For example, what was your overall impression? How did the sentences read? Were there any recurring things of note? Did anything specific stand out?". The user is encouraged to write freely and extensively, noting that their input is valuable to the project. A large text area is provided for comments. Below the text area, a "Finish" button is visible. The screen concludes with a thank you message and the GoURMET team's name.

gourmet

Evaluation Set Complete

You've now completed the direct assessment phase of this evaluation.

As you may already be aware each sentence two you've just assessed was generated by a machine, translated from English source text. Each sentence in black was written by a person.

Before you end the evaluation please provide your thoughts on the machine generated sentences that you have seen in this series.

For example, what was your overall impression? How did the sentences read? Were there any recurring things of note? Did anything specific stand out?

Please write as freely and extensively as you wish. Your input is invaluable to the academics who have produced this language model and continue to research machine learning and artificial neural networks.

Once you've written your comments you'll be able to complete the evaluation by clicking on 'finish' below.

Thank you again. Your feedback is much appreciated.

The GoURMET team.

Finish

Figure 2: DA evaluator feedback screen

2.2.8 Gap Filling: Details

Gap filling measures the level of comprehension of the translated document (Forcada et al., 2018a). The methodology using Gap filling for the GoURMET project proceeds as follows. Each evaluator is presented with a series of sentences drawn from news articles which have been translated by a human from the source language into English. A given percentage of content words (nouns, adjectives, verbs) have been automatically removed from this sentence. This evaluator is asked to fill in the gaps. A hint may be provided to aid the evaluator in doing this, where the hint is the same sentence but automatically translated. The quality of the machine translation is measured by the evaluator's success in filling in the gaps correctly using the hint. The more correct information retained by the automatic translation, the more gaps the evaluator will be able to fill out which match the human translation. A baseline is established by not providing any hint at all, and in this situation the evaluator must guess the correct word to go in the gap with no context. Previous studies Forcada et al. (2018a); O'Regan and Forcada (2013) have shown that around 25% correctness is to be expected. The correctness of the gaps filled using the hints is expected to be 25%-100%, dependent on the quality of the MT system under evaluation.

Gap filling is used as a proxy for reading comprehension questionnaires (Forcada et al., 2018a; Scarton and Specia, 2016) which have been shown to be time-consuming to undertake (questions have to be prepared from the source text and translated to the target language). The results of a gap-filling exercise indicate how much of the key information in a text is translated in a way that

humans can understand. These exercises require evaluators who are fluent in the target language (e.g. native speakers), no knowledge of the source language is needed.

It has previously been studied if, when processing the results of a gap filling exercise, synonyms of the correct word should be counted as a match when filled into a gap by the evaluator. (O'Regan and Forcada, 2013). Evidence suggests that this will increase the overall accuracy reported, but that the accuracies of each MT system under test will be increased by the same amount relative to their base accuracy. As such, when comparing across systems, this step is unnecessary. Evidence from GoURMET evaluation presented in Section 5.1.1 supports this simplification.

Gap filling is carried out when English (or if necessary, German) is the target language.

Gap filling took place using the custom user interface pictured in Figure 3, analogous to the one used by Forcada et al. (2018a).

Please fill in the gaps in the sentence below with a single word.

If a hint is provided please use this sentence to inform your decision on the most appropriate word.

Hint: It also promises strong economic cooperation, which includes coordinated environmental policies and climate change.

It also pledges stronger economic integration, which includes coordinated

and change

Submit

Figure 3: Custom Gap Filling interface.

Data

GF evaluation data for each language pair consists of 30 sentences, selected across at least 5 different articles originally published in the non-English language by one or both of the media partners. Preparation of the evaluation data proceeds as follows.

Once translated, each candidate sentence must be 15 words or more in length. Sentences which are too short are replaced by longer ones.

Each sentence from the non-English language is translated into English by the GoURMET MT system and, if a silver evaluation is being undertaken, by Google Translate (see below). The GoURMET API (Secker et al. (2020)) was used to generate the translations using the GoURMET system. Sentences were manually submitted to Google Translate to generate the translated sentences using that system. This was managed manually as each set of evaluation data for the GF exercise consists of only 30 sentences.

For each sentence translated into English, a certain percentage of the content words are removed at random making sure there are no two consecutive gaps; in our experiments, this meant typically leaving between 1 and 8 missing words in each sentence. The Python `stop_words` library was used. Gaps are not allowed at stop-words or punctuation, and two gaps are never consecutive or separated only by stop-words or punctuation.

Due to the fact that sentences have different lengths and each language has a different number of stopwords, the actual percentage of gaps in the test set for each language happened to vary around 10%, see Table 1.

Language	Gap percentage	Average sentence length
Bulgarian	9.3%	23.4
Gujarati	13.2%	31.5
Kyrgyz	9.7%	22.6
Serbian	9.0%	19.4
Swahili	9.1%	21.4
Turkish	10.8%	21.7

Table 1: Gap percentage and sentence length for each language evaluated via gap-filling.

Note that there have been previous studies into whether a metric such as entropy could be used to select the hardest-to-guess content words for removal and thus increase the efficacy of the test Forcada et al. (2018a), but this did not significantly change the result of the evaluation.

Evaluation levels for Gap Filling

Bronze standard: Each gap will be filled 3 times in 2 scenarios. The scenarios will be:

- With a translated hint sentence generated by the GoURMETmodel
- No hint

Silver standard: Each gap will be filled 3 or more times in 3 scenarios The scenarios will be:

- With a translated hint sentence generated by the Gourmet model
- With a translated hint sentence generated by Google translate
- No hint

We will go for the silver standard per default, i.e. each gap will be filled in 3 or more times in 3 scenarios, actually by both partners, resulting in 18 results. The bronze standard will only apply in case there is no Google Translate engine available for that language.

Evaluation procedure

For each of the 30 sentences, evaluation occurs two or three times for each scenario. Evaluators only fill gaps in one scenario per sentence. That is to say, evaluators only ever see a single sentence in one particular context. Doing otherwise would bias the results when the same sentence is seen subsequently to its initial presentation but in a different context.

The GF evaluation requires the evaluator to fill in the missing words using the hint, or no hint (the latter to establish a baseline). The hint is the translated output of the MT system under test for this gap fill instance.

Each of the 30 sentences is evaluated in three different ways: one evaluator sees the gapped sentence with no hint; one evaluator sees the gapped sentence with the GoURMET MT output as a hint; finally, for silver level evaluation only, one evaluator sees the gapped sentence with the Google Translate output as a hint.

The accuracy of the translation is a *success rate*: the fraction of gaps correctly filled.

Evaluators

Evaluators are recruited from within the media partners wherever possible according to the guidelines described previously in the parent section.

Given the requirements above, the minimum number of evaluators required to complete the gap filling task is 6 for the bronze standard and 9 for silver.

As the evaluation sentences will be different between languages, evaluators will be able take part in multiple gap filling evaluation rounds across the project (where each evaluation round concerns a different source language).

Evaluators are asked to complete the exercise according to the following guidelines

- You will be shown a series of X sentences. Each sentence will have one or more words missing. Missing words will be replaced by a gap. Your task is to try to recreate the original sentences by typing the word that you believe is missing into each gap.
- Some evaluation sentences contain hints to help you fill the gaps. If you are provided with these please refer to them to aid your decision.
- In others you'll be on your own with just the rest of the sentence as context to guide you. Make your best guess.
- You can of course use your knowledge to fill the gaps where no hint is provided. However, please do not conduct research or speak to others to help you fill gaps. We are not assessing you, we are comparing scenarios. Additional external input may invalidate results.
- Please fill all gaps.
- If you have no hint and no clear idea just give it your best shot. Please try to make a sentence you think may have been published and avoid using words you know or suspect are not correct just because it will give an amusing outcome.

3 Interfaces for Human Evaluation

Two custom user interfaces were created to support the human evaluations as described previously; see figures 1 and 3 for DA and GF, respectively.

This section describes the background and technical aspects of those interfaces. The way they are actually used in order to support the human evaluation has previously been covered in Sections 2.2.7 and 2.2.8.

Note that any deviations from the general methodology along with further specifics (such as number of evaluators for each specific test and so on) will be noted where relevant under Section 5.

Evaluation takes place using two cloud-based web interfaces, one for GF and one for DA. These are accessible to all parties taking part in the evaluation. This allows administration and collection of results to take place in a centralised location, and therefore means that there is no need for each institution taking part in the evaluation to deploy their own version of the interface. Deploying the interfaces as web applications provides the added benefit that the task can be completed and the data sets administrated through a web browser rather than the participant or administrator needing to install dedicated software.

Both applications run on the general architecture as shown in Figure 4

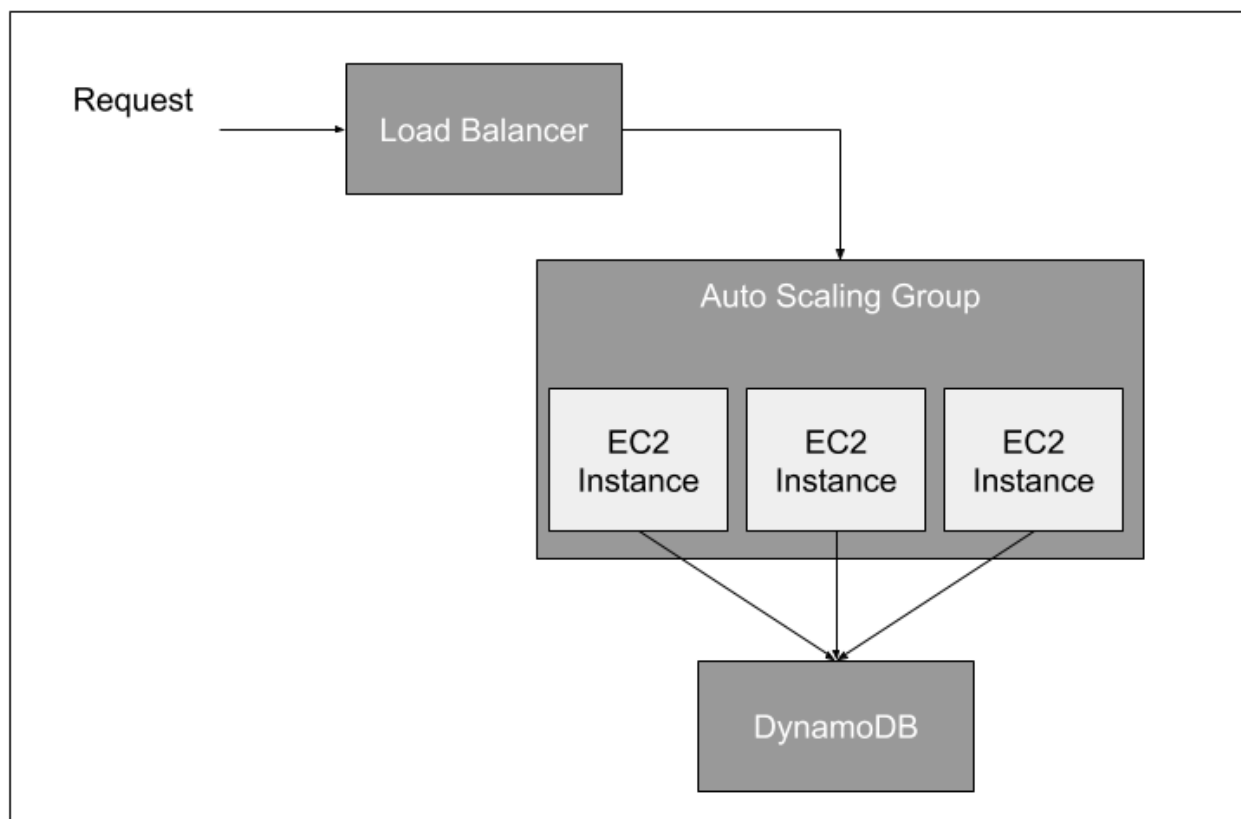


Figure 4: Architecture for DA evaluation tool

The application uses HTML, CSS, JavaScript and the Handlebars templating language to create the client facing interface. This is served from a back-end application running on an EC2 instance. This back-end application uses Node with the Express framework and is written in TypeScript.

DynamoDB is used for persistent data storage. The database is secured using AWS IAM Roles to restrict access so that records can only be accessed via the server side application.

Further, more detailed, technical information regarding the building and deployment of the software, the required database structure, and so on, can be found in the documentation contained in the open-source repository of the appropriate tool and will not be repeated here. The repository documentation also contains detailed instructions for administrators and users, see Section 3.3.

The descriptions in the following subsections are presented in order to give the reader an indication of how the tools are used in practice, mainly from an administrative/evaluation coordinator perspective, and should be read in conjunction with the description of how the evaluators should use the tools as described in Sections 2.2.7 and 2.2.8.

3.1 Direct Assessment Evaluation Tool

To administer the DA testing, test data is uploaded to the tool manually by the evaluation coordinator. The source of the test data are the manual translations as described in Section 2.2.7. The test data set must be converted to a simple JSON file according to the definition below before upload:

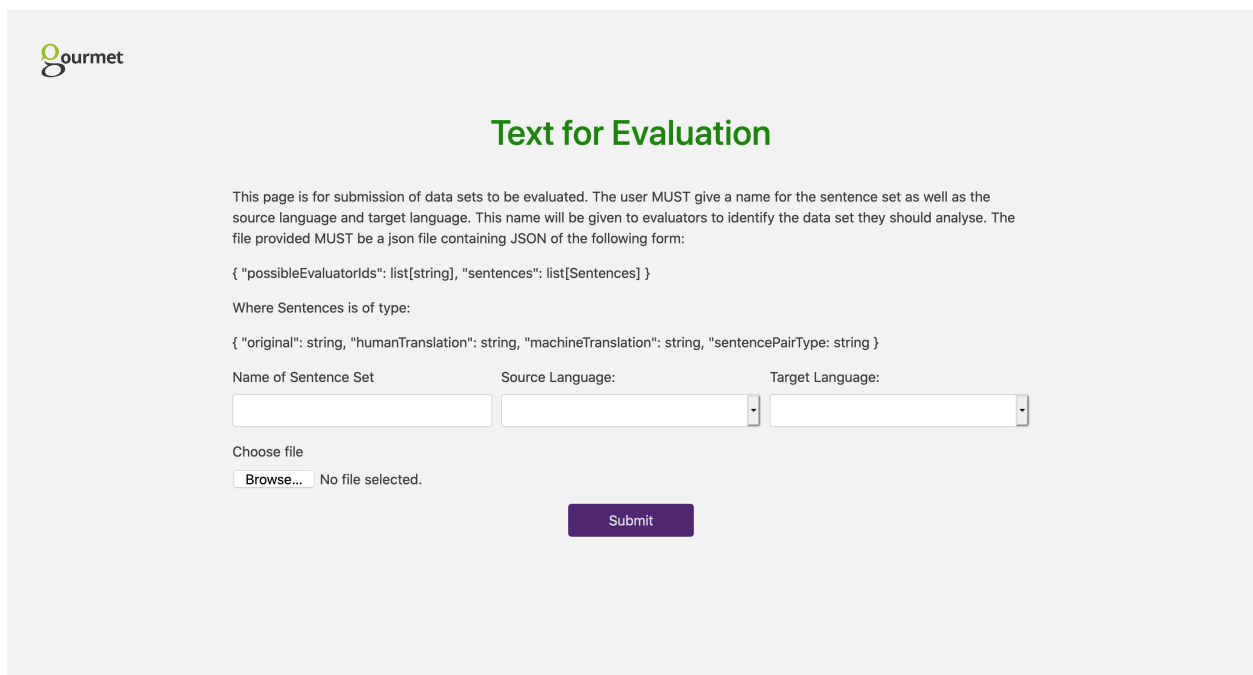
```
{
  "possibleEvaluatorIds": list[string],
  "sentences": list[Sentences]
}
```

Where Sentences have the following properties:

```
{
  "original": string,
  "humanTranslation": string,
  "machineTranslation": string,
  "sentencePairType": string
}
```

The upload of the test data can then take place via the main user interface 'submit data sets' page as shown in Figure 5;

The *View Results* page allows the evaluation coordinator to export the direct-assessment scores by language as a CSV file. It also shows which participants have *started* the direct-assessment Task. See Figure 6.



Text for Evaluation

This page is for submission of data sets to be evaluated. The user MUST give a name for the sentence set as well as the source language and target language. This name will be given to evaluators to identify the data set they should analyse. The file provided MUST be a json file containing JSON of the following form:

```
{ "possibleEvaluatorIds": list[string], "sentences": list[Sentences] }
```

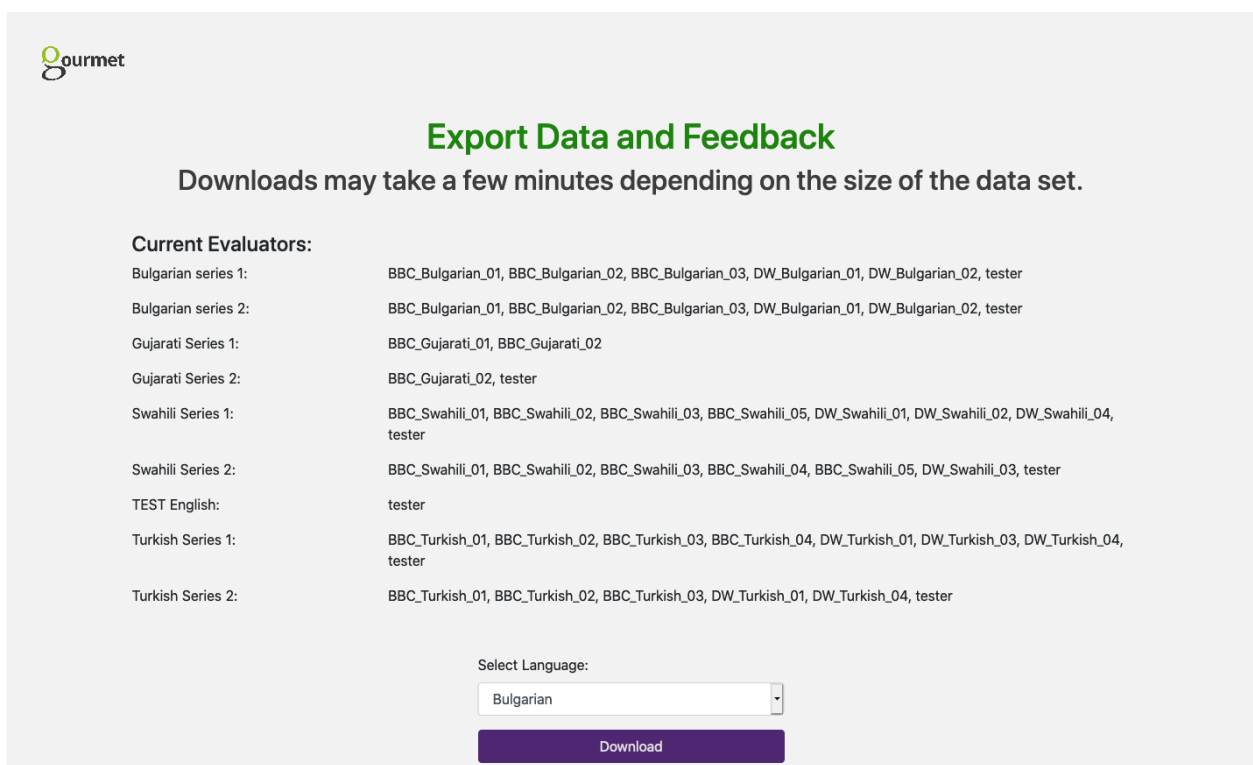
Where Sentences is of type:

```
{ "original": string, "humanTranslation": string, "machineTranslation": string, "sentencePairType": string }
```

Name of Sentence Set: Source Language: Target Language:

Choose file
 No file selected.

Figure 5: Direct-assessment tool — submitting test dataset



Export Data and Feedback

Downloads may take a few minutes depending on the size of the data set.

Current Evaluators:

Bulgarian series 1:	BBC_Bulgarian_01, BBC_Bulgarian_02, BBC_Bulgarian_03, DW_Bulgarian_01, DW_Bulgarian_02, tester
Bulgarian series 2:	BBC_Bulgarian_01, BBC_Bulgarian_02, BBC_Bulgarian_03, DW_Bulgarian_01, DW_Bulgarian_02, tester
Gujarati Series 1:	BBC_Gujarati_01, BBC_Gujarati_02
Gujarati Series 2:	BBC_Gujarati_02, tester
Swahili Series 1:	BBC_Swahili_01, BBC_Swahili_02, BBC_Swahili_03, BBC_Swahili_05, DW_Swahili_01, DW_Swahili_02, DW_Swahili_04, tester
Swahili Series 2:	BBC_Swahili_01, BBC_Swahili_02, BBC_Swahili_03, BBC_Swahili_04, BBC_Swahili_05, DW_Swahili_03, tester
TEST English:	tester
Turkish Series 1:	BBC_Turkish_01, BBC_Turkish_02, BBC_Turkish_03, BBC_Turkish_04, DW_Turkish_01, DW_Turkish_03, DW_Turkish_04, tester
Turkish Series 2:	BBC_Turkish_01, BBC_Turkish_02, BBC_Turkish_03, DW_Turkish_01, DW_Turkish_04, tester

Select Language:

Figure 6: Direct-assessment tool — exporting results

3.2 Gap-Filling Evaluation Tool

The input data for the gap-filling tool is generated from the raw evaluation data (curated as described in Section 2.2.8) using a script created by project partner ALAC and available as open source. See Section 3.3. The script processes the raw evaluation data into gap-filling evaluation data by performing the following steps:

- Single word gaps to be created in each human translated sentence so that the sentence has a gap density of 20%; gaps are only created where content words are, and there must be at least one content word between gaps.
- Create nine JSON files. These will contain gap fill problems with one the three variants of hint type and result in each variant being evaluated a minimum of three times for each sentence. Each file will contain a mixture of the three hint types where the three different hint types are:
 - no hint
 - a hint generated by GoURMET
 - a hint generated by Google Translate

It should be noted that the translations required as hints in the above are computed offline and used as one of the source inputs to the script. The script itself has no means at this time to directly query either the GoURMET API or the Google Translate API for translations.

The output of the pre-processing script is nine JSON files in the JSON format below, and as such can be directly uploaded to the tool via the UI for evaluation.

The GF tool accepts evaluation data according to the following JSON structure:

```
{
  "id": "string",
  "sourceLanguage": "string",
  "targetLanguage": "string",
  "evaluatorIds": ["string"]
}
```

Where a SegmentObject has the following form:

```
{
  "id": "string",
  "translationSystem": "string",
  "source": "string",
  "translation": "string",
  "hint": "string",
  "problem": "string",
  "gapDensity": "string",
  "context": "string",
  "entropyMode": "string",
  "correctAnswers": ["string"]
}
```

The JSON file can then be uploaded by the evaluation coordinator via the interface pictured in Figure 7.

The screenshot shows the 'Text for Evaluation' interface. At the top left is the GoURMET logo. The title 'Text for Evaluation' is centered in green. Below the title, there is a form with two main sections. The left section has a label 'Name of Segment Set' above a text input field, and a purple 'Submit' button below it. The right section has a label 'Choose file' above a 'Choose File' button, which is currently disabled and shows 'No file chosen'.

Figure 7: Gap filling tool - upload test data screen

The *View Results* page, pictured in Figure 8, allows the evaluation coordinator to export the results by language as a CSV file. It also shows which participants have *started* the gap-filling task.

The screenshot shows the 'Export Data' interface. At the top left is the GoURMET logo. The title 'Export Data' is centered in green. Below the title, a message states: 'Downloads may take a few minutes depending on the size of the data set.' Underneath, there is a section titled 'Current Evaluators:' followed by a list of evaluators grouped by language: BG actual: BBC_3_GU, BBC_4_GU, BBC_9_GU, DW_7_GU, tester; BG series 1: BBC_1_BG, BBC_1_GU, BBC_2_GU, BBC_3_BG, BBC_3_GU, DW_6_GU, DW_7_GU, tester; test: ab, new, tester; tr: DW_7_BG, tester. Below this list is a 'Select Language:' dropdown menu currently set to 'Bulgarian'. At the bottom is a purple 'Download' button.

Figure 8: Gap Filling tool - export results screen

3.3 Open-source Releases

The DA and GF tools described herein are open-sourced under the GNU General Public License (GPL), version 3 and made freely available.

The direct-assessment evaluation tool is available from <https://github.com/bbc/gourmet-sentence-pairs-evaluation> and the gap-filling evaluation tool is available from <https://github.com/bbc/gourmet-gap-fill-evaluation>.⁵

⁵ The scripts used by partner ALAC to generate the gap-filling data in the format accepted by the latter and some post-processing scripts are kept at <https://gitlab.com/mlforcada/bbc-dw-gf>.

The client applications are both available as a Docker images from the above repositories in order to ease deployment.

The GA pre-processing script is also open sourced under GPLv3 and is available from <https://gitlab.com/mlforcada/bbc-dw-gf>.

The repository locations cited above all contain documentation and notes on the process for accepting updates and maintenance requests.

4 Results of Data Driven Evaluation

In this section we evaluate the translation models which have been deployed via the platform API. More details about the models evaluated can be found in D5.3 Initial Integration Report Appendix A Translation Model Details.

4.1 Test Sets

In this section we describe the test sets used for the automatic evaluation in the next section.

Swahili The development and test sets were obtained from the GlobalVoices parallel corpus. 4 000 parallel sentences were selected from the concatenation of GlobalVoices-v2015 and GlobalVoices-v2017q3, and randomly split into two halves (with 2 000 sentences each), which were used respectively as development and test corpora. The half reserved to be used as test corpus was further filtered to remove the sentences that could be found in any of the monolingual corpora.

Gujarati The development and test sets are the official sets provided by the WMT19 news shared task (Barrault et al., 2019). The development set contains 1988 sentences. There is a separate test set for each language direction (en–gu and gu–en), so that the source side of each test set is the original text and the target side sentence are the human translations. The en–gu test set contains 998 sentences and the gu–en test set contains 1016 sentences.

Turkish The development and test sets were obtained from the WMT18 news shared task (Bojar et al., 2018). We combined `newtest2016` and `newtest2017` for development, a total of 7,008 sentence pairs, and reserved `newtest2018` for test, a total of 3,000 sentence pairs. DW editors also carefully curated a dataset of 210 sentences for the user evaluation.

Bulgarian For the test and dev set, we took 4000 sentences from the end of the SETIMES2 corpus (from OPUS). The first 2000 were the test set, and the second 2000 were the test set. The preprocessing was Moses normalisation, tokenisation, truecasing, then BPE with 50k merges learnt separately on each side of the training set.

Tamil The models are currently evaluated using the official WMT20 development and test sets. The development set consists of 1989 sentences. The test sets (separate for each language direction) consist of 1000 sentences for en–gu and 997 sentences for gu–en.

We also created GoURMET development and test set by aligning data from BBC dumps using a modified version of Bitextor (Esplà-Gomis and Forcada, 2010). The models will be tested on these sets at a later date. For document alignment we used an existing MT system to translate all Tamil articles into English and align them based on a TF/IDF score. The alignment was restricted so that only documents originally published within a 30-day time frame of each other are aligned. Segment alignment was done using Bleualign⁶, producing a score for each segment pair. For the dev and test set, we took the pairs with a Bleualign score over 0.24 where both source and target

⁶ <https://github.com/rsennrich/Bleualign>

sentence contain more than 5 tokens. The sentence pairs were shuffled and split into a dev set of 1916 sentence pairs and a test set of 1917 sentence pairs. Data size is described in Table 2.

Corpus	Sents	en tokens	sr/ta tokens
En-Sr dev	2100	53112	49877
En-Sr test	2100	51933	48762
En-Ta dev	1916	36993	30573
En-Ta test	1917	36940	31180

Table 2: Size of the dev and test sets used for the development and evaluation of the English-Serbian and English-Tamil models. Token counts reported were calculated on raw text, non-tokenised and before BPE segmentation.

Serbian The development and test set for English-Serbian were obtained from the crawled DW corpus. The crawling procedure is described in Deliverable D1.2. From the full En-Sr corpus, we extracted 4200 sentence pairs with a Bicleaner score over 0.8, where both sentences contained more than 10 tokens and the source-to-target length ratio was between 0.8 and 1.1. The Bicleaner model used was an English-Croatian model released with Bicleaner⁷. Half of these sentence pairs formed the dev set, while the other half formed the test set. Data size is described in Table 2.

Amharic The development and test set for English-Amharic were obtained from the GoURMET English-Amharic crawled parallel corpus. There was no provided split for the evaluation sets and therefore we randomly sample sentences. We sample randomly 3,000 unique sentences for each evaluation set.

Kyrgyz A development set and part of the test set were obtained from the GoURMET English–Kyrgyz crawled parallel corpus as follows. First, the crawled corpus was ranked with Bicleaner (?), whose model was trained on all the publicly parallel corpora for this language pair. Sentence pairs with a score lower than 0.5 were discarded. Then, all the sentences extracted from the news website <https://24.kg/> were reserved for the test set. From the remaining sentences, those with a score higher than 0.7, which are very likely to be parallel, were selected to build the development set and the rest was used as training data. The test set was further enlarged with parallel sentences extracted from documents provided by project partner BBC. The number of sentences and words of the test set obtained from each source are depicted in Table 3.

Corpus	Sents	en tokens	ky tokens
GoURMET crawled	144	2 499	1 830
BBC	1 117	19 811	15 749
total	1 261	22 310	17 579

Table 3: Distribution of data among the sources used to build the English–Kyrgyz test set

⁷ <https://github.com/bitextor/bicleaner-data/releases>

4.2 Summary of the results

We report the BLEU and chrF translations scores for our systems and Google translate. The comparison with Google translate is essential for our user partners, to help them to calibrate the research models. It is not scientifically valid to compare our models with Google. The Google models are not documented and therefore they are not reproducible. We do not know the details of the architecture that they use, or the data that they use for training. We do not know when they upgrade or change their translation models from one version to another. One of the most concerning differences with our systems, is that in order for the experiments to be valid, test data must not be included in training data or the model’s performance is inflated. All our systems hold the test set apart from training, but they may very well be part of Google’s training resources, and this would give the Google systems a large advantage. We therefore provide these evaluation results with caution. The more scientifically rigorous comparisons will always be the results from the annual WMT competition, as WMT use novel test sets produced each year. We have produced Tamil and Gujarati test sets for WMT for use in their evaluation campaign and for convincing evidence of our success for the GoURMET project.

		BLEU				chrF			
Language pair	Test set	GoURMET		Google		GoURMET		Google	
		avg		avg		avg		avg	
Bg→En	SETIMES2	50.40		48.97		0.72		0.71	
Tr→En	WMT18	17.26		29.35		0.45		0.58	
Sw→En	GlobalVoices	29.70		30.28		0.55		0.55	
Gu→En	WMT19	22.17		30.58		0.50		0.58	
Ta→En	BBC	28.60	21.91	33.91	28.59	0.57	0.51	0.61	0.57
	WMT20 (dev)	22.21		23.27		0.51		0.52	
Sr→En	BBC	67.03	47.79	71.98	50.03	0.82	0.69	0.85	0.71
	Deutsche Welle	28.55		28.08		0.56		0.57	
Am→En	GoURMET public	11.79		22.76		0.31		0.42	
Ky→En	BBC	17.50	17.83	22.76	22.21	0.46	0.46	0.50	0.49
	GoURMET public	18.16		21.65		0.45		0.48	

Table 4: Scores for translation into English

In Table 4 we report results for translation models which translate into English, and in Table 5 we report out of English results. We observe that BLEU and chrF scores result in consistent rankings of systems for most language pairs and translation directions, hence the two metrics validate each other. This provides evidence that the evaluation methodology is sound and the rankings reflect a reasonable measure of quality rather than depending on the quirks of a specific metric.

Systems that translate into English mostly obtain higher scores than systems that translate from English. This is expected because there is much more in-domain monolingual data for English than any of the low-resource languages we consider, and monolingual data is most effectively used to improve target-language fluency by means of backtranslation. Furthermore, the difference is more pronounced for the BLEU scores than the chrF score, which is also expected because English is morphologically simpler than most of the considered languages, which facilitates exact

		BLEU				chrF			
Language pair	Test set	GoURMET		Google		GoURMET		Google	
		avg		avg		avg		avg	
En→Bg	SETIMES2	50.08		44.17		0.71		0.69	
En→Tr	WMT18	11.17		21.05		0.46		0.56	
En→Sw	GlobalVoices	27.53		23.27		0.55		0.52	
En→Gu	WMT19	16.31		24.27		0.48		0.57	
En→Ta	BBC	11.63	10.84	12.80	12.94	0.54	0.53	0.56	0.56
	WMT20 (dev)	12.59		13.08		0.53		0.55	
En→Sr	BBC	51.30	36.54	47.94	33.03	0.73	0.63	0.72	0.61
	Deutsche Welle	21.78		18.11		0.52		0.50	
En→Am	GoURMET public	6.66		9.06		0.18		0.26	
En→Ky	BBC	9.61	10.55	9.67	10.81	0.43	0.43	0.43	0.43
	GoURMET public	11.49		11.94		0.43		0.43	

Table 5: Scores for translation from English

word matching to reference translations, and as discussed in section 2.1, chrF is more robust on morphologically rich languages since it allows for partial word matches.

For translation into English, Google Translate obtains higher or equal scores to our systems for most source languages, but for translation from English our systems are much more competitive with Google Translate, surpassing it for several target language. These differences might be again attributed to the large amount of English monolingual text that was presumably used by Google to train their system. We shall remark that we can’t exclude training-test set contamination for Google Translate, especially for test set that we scraped from the web and are not part of standard training-test splits (BBC, Deutsche Welle and GoURMET public), hence the scores Google Translate might overestimate its quality.

4.3 Language details

4.3.1 Bulgarian

We evaluate on the publicly available SETIMES2 test set. Our systems obtain high scores for both translation directions, surpassing Google Translate.

The user evaluation dataset consists of DW translated content specifically curated for this purpose, resulting in a set of 210 fully parallel sentences.

4.3.2 Turkish

We evaluate on the WMT18 news translation task test set. Our results are underwhelming, especially for the En→Tr direction, but it should be noted that these refer to the baseline (parallel data-only) systems. The research systems described in deliverable D5.3 but not yet integrated in the GoURMET API are more promising. For English-Turkish, in particular, we are aware of additional training resources that we did not exploit (e.g., over 40 million sentence pairs in the subtitling domain) and which Google Translate might use for training. We plan to integrate a stronger model in July 2020.

4.3.3 Swahili

We evaluate on the GlobalVoices test set. Our systems are quite strong, performing similarly to Google Translate for the Sw→En direction and better for the En→Sw direction.

The user evaluation dataset consists of DW translated content specifically curated for this purpose, resulting in a set of 210 fully parallel sentences.

4.3.4 Gujarati

We evaluate on the WMT19 news translation task test set. Our systems obtain scores lower than Google Translate, but they are still of reasonable quality especially for the Gu→En direction.

4.3.5 Tamil

We evaluate on a test set extracted from the BBC website and validated by human annotators, as well as the development set of the WMT2020 news translation shared task.

At the moment of this writing, the WMT2020 test set has not been released, therefore it should be noted that since we use the development set for hyperparameter tuning and early stopping, the scores might be higher compared to a truly independent test set.

For both our systems and Google Translate, the En→Ta direction obtains worse BLEU scores than the Ta→En direction, however the chrF scores are similar for both directions and in line with those of other language pairs. We observe that our systems generally achieve lower scores than Google Translate, although the differences are small in terms of chrF. The low BLEU scores might be caused by tokenization issues with the SacreBLEU tool applied to the Tamil script. Being character-based, chrF is presumably more robust to such issues.

4.3.6 Serbian

We evaluate on test sets extracted from the BBC and Deutsche Welle websites and validated by DW human annotators for technical as well as human evaluation.

Serbian can be written in both Latin and Cyrillic scripts, in the GoURMET project we made the decision to use the Latin script since both the BBC and Deutsche Welle websites use it. Google Translate, however, returns Serbian written in the Cyrillic scripts, therefore in order to evaluate it we transliterate it into Latin using the Python *transliterate* library ⁸.

In the technical evaluation, we observe much higher scores for the BBC test set compared to the Deutsche Welle test set, which might indicate some degree of training/test overlap for both our systems and Google Translate. Our systems are quite strong, comparable to Google Translate for the Sr→En direction and better for the En→Sr direction.

4.3.7 Amharic

We evaluate on the test portion of the GoURMET public corpus. Our systems and Google Translate obtain poor results, especially for the En→Am direction. According to these results, Amharic is a very hard language to translate from and into, presumably due to the very low amounts of available training data and its linguistic distance from most languages commonly addressed in machine translation. Note that the Am-En system integrated into the API employs a different recaser than described in deliverable D5.3 due to difficulties in integrating the original recaser into a Docker.⁹ The En-Am system described in deliverable D5.3 obtains more promising results but is not yet integrated into the GoURMET API.

4.3.8 Kyrgyz

We evaluate on human-validated test sets extracted from the BBC website and the GoURMET public corpus. The results for our systems are not particularly strong, but very similar to Google Translate, in particular for the chrF scores which are very appropriate when evaluating translations into a language as morphologically rich as Kyrgyz. It should be noted that English–Kyrgyz Google Translate MT systems have been updated recently. As a consequence, the Google Translate scores reported in deliverable D5.3, which were computed during the development of the systems, do not match those reported in this document.

⁸ <https://pypi.org/project/transliterate/>

⁹ We used a custom recaser where we map tokens to their most frequent form as found in the training set.

5 Results of Human Evaluation

The results of human evaluation (Gap Filling and Direct Assessment) are contained in this section. As this is an interim report, not all languages pairs have completed evaluation to a level at which results can meaningfully be reported at the time of writing. Results are reported in the following sections where it is possible to do so. Language pairs for which a translation model is available, but no results have been shown will be reported in the subsequent evaluation deliverable.

5.1 Gap Filling

This section contains results of the gap-filling (GF) exercise. All GF evaluations take place with the non-English language as the source and English as the target. GF can very well replicate the media monitoring scenario and as such, translation in this direction is relevant.

For each language-pair evaluation, the results below summarise the rate at which the word used to fill the gap by the evaluator matched the word which was removed from the original test, using different systems to provide the hint. This is also compared to a baseline where no hint was given (i.e., the evaluator must guess the correct word).

Each MT system built in GoURMET is compared to Google Translate¹⁰ for language pairs supported by that service at the time of the evaluation. This includes all language pairs in this deliverable, but in the subsequent M36 evaluation deliverable, it is likely that GoURMET will support language pairs not covered by Google Translate.

This comparison is made for the benefit and interest of the media partners. It cannot, however, be reported as a scientifically rigorous comparison, as the evaluation method cannot control for Google Translate. There is no way to know what parallel texts Google translate accessed to train, and therefore, it is possible that training and testing data may overlap. The testing of the GoURMET language models is controlled such that no training data will be present in any test set. Nevertheless, it seems reasonable to compare the results obtained by GoURMET from existing public sources and via our own crawling to a popular general-purpose machine translation system.

Given the importance of the adoption of GoURMET translation technologies in the media partners' The comparison with Google Translate in this context is important. Media partners will be able to use this information to make real-world informed choices regarding the technologies to use for gisting for media monitoring purposes.

The sets of evaluation data were created as described in Section 2.2.8. 30 sentences of 15 or more words each are used as the basis of the evaluation set.

From the 30 evaluation sentences, 9 problem sets are derived. In each problem set there will be a combination 10 GoURMET translation hints, 10 Google translation hints and 10 with no hint, such that each sentence will be evaluated 3 times in each of the 3 'states' and if each evaluator only sees one problem set, they shall only see each sentence once.

In all completed evaluations below, all 9 problem sets have been evaluated (at least) twice. That is, each each sentence has been evaluated in each state, 6 times.

Incomplete evaluations are reported where a reasonable number of results have been gathered. These are clearly indicated and the aim will be to evaluate to the same level as those marked

¹⁰<http://translate.google.com>

”complete”.

The following subsection contains a brief description of our strategy for post-processing the raw GF results and a justification for making this decision. Subsequent subsections contain firstly a summary of results across languages (Section 5.1.2), followed by result details for each language pair.

5.1.1 Investigation into the Postprocessing of Evaluation Scores

The raw results used to create the summaries in this section were not post-processed before the summary was created.

The summaries of results below have been created by counting a correct gap fill if the word removed from each sentence in the gap-filling scenario has been filled with an exact (case-insensitive) match by the human evaluator. Data cleaning could be applied to these matches. Data cleaning in this context can be the correction of spelling mistakes such that an accidental mismatch is corrected, and additionally the inclusion of synonyms as a positive match.

However, previous tests of cleaning in the literature Forcada et al. (2018a) indicate the use of raw results in this manner is sufficient when comparing across languages (as is the case in this section) and no data cleaning is required.

The literature indicates that while corrections for spelling, synonyms, etc. may slightly raise the absolute scores achieved by the models under test, (a) this increase is unlikely to be significant and (b) all models are raised by roughly the same amount and thus relative comparisons of models within a single language and across multiple languages remain unchanged.

An informal validation of the above assertion was undertaken with the results for **bg**→**en**. Table 6 enumerates all instances where a spelling correction or synonym match could apply.

Of 937 raw results, only the 37 above were candidates for manual correction.

Of these, only three were true spelling mistakes (*consequences/consuquences*, *homosexuality/homsexuality*, *Instagram/Instangram*), there were four further substitutions between the literal number 2 and the word *two*. If corrected, a total of seven spelling adjustments would not make a significant difference to the final results.

The remaining corrections were judged to be synonyms by a human expert. It is noted that there was some debate regarding what constitute a synonym for a particular word and which alternate was too “distant” from the original. A case in point would be the inclusion of *gig/concert* as a legitimate substitution but the judgement that *gig/performance* and *gig/event* were less specific and therefore not allowable. While this judgement can be debated, it can therefore be noted that the corrections of synonyms may introduce artificial noise into the final results.

It is further noted that candidate corrections are spread fairly evenly across all three systems under test. No model is likely to be disproportionately affected compared with others.

The suggestions from the literature regarding correcting raw results are therefore seen to be reasonable. No manual post processing corrections are applied to any of the results described in this section.

Model under test	Missing word	Answer given	Correction type
google	2	two	Spelling
gourmet	meteorologists	metereologists	Spelling
gourmet	world	globe	Synonym
NONE	globe	world	Synonym
NONE	globe	world	Synonym
NONE	world	planet	Synonym
gourmet	globe	world	Synonym
gourmet	globe	world	Synonym
NONE	gig	concert	Synonym
NONE	globe	world	Synonym
gourmet	jail	gaol	Synonym
gourmet	jail	prison	Synonym
gourmet	2	two	Spelling
NONE	globe	world	Synonym
gourmet	globe	world	Synonym
gourmet	globe	world	Synonym
NONE	jail	prison	Synonym
NONE	jail	prison	Synonym
gourmet	consequences	consuquences	Spelling
gourmet	globe	world	Synonym
gourmet	globe	world	Synonym
google	globe	world	Synonym
google	deity	god	Synonym
google	gig	concert	Synonym
gourmet	Instagram	Instangram	Spelling
google	jail	prison	Synonym
gourmet	jail	prison	Synonym
gourmet	jail	prison	Synonym
NONE	homosexuality	homsexuality	Spelling
gourmet	profit	profits	Synonym
google	globe	world	Synonym
gourmet	deity	god	Synonym
NONE	gig	concert	Synonym
gourmet	gig	concert	Synonym
NONE	2	two	Spelling
NONE	2	two	Spelling
gourmet	gig	concert	Synonym

Table 6: Example spelling mistakes and synonym corrections in raw evaluation output for Bulgarian.

5.1.2 Summary of Results

Table 7 shows a summary of the gap fill success rates per language primarily allowing a comparison across languages. All evaluations are from the non-English language to English. GF success rate

is the fraction of gaps which are filled with the correct word given different hint types.

Target Language	Hint Type		
	None	GoURMET	Google
Bulgarian	39.9%	65.3%	68.1%
Turkish	29.7%	60.4%	69.9%
Swahili	27.6%	55.9%	61.9%
Gujarati	19.0%	44.9%	51.8%
Serbian	32.6%	86.2%	88.6%
Kyrgyz	34.0%	60.9%	68.9%

Table 7: GF success rates (%)

5.1.3 Bulgarian

Status: Silver standard - Complete

The statistics for the GF evaluation for $\text{bg} \rightarrow \text{en}$ are shown in Table 8. The detailed box plot of results is shown in Figure 9.

Unique Evaluators	18
Number of unique gaps	72
Average number of gaps per sentence	2.4
Evaluations per gap-configuration	6.0
Number of evaluations by hint type: NONE	432
Number of evaluations by hint type: Gourmet	432
Number of evaluations by hint type: Google	432

Table 8: Summary of GF evaluation for $\text{bg} \rightarrow \text{en}$

The box plot for this language pair indicates broad similarity, with the upper limits closely aligned. The relative positions of the lower quartiles indicate that on average the translation from Google will be slightly better than the GoURMET MT solution. Both GoURMET and Google are clearly far superior to the no-hint baseline.

5.1.4 Turkish

Status: Silver standard - Complete

The statistics for the GF evaluation for $\text{tr} \rightarrow \text{en}$ are shown in Table 9. The detailed box plot of results is shown in Figure 10.

Google appears to fair better than the GoURMET MT solution for this language pair, however the median for Google does coincide with the upper quartile for GoURMET.

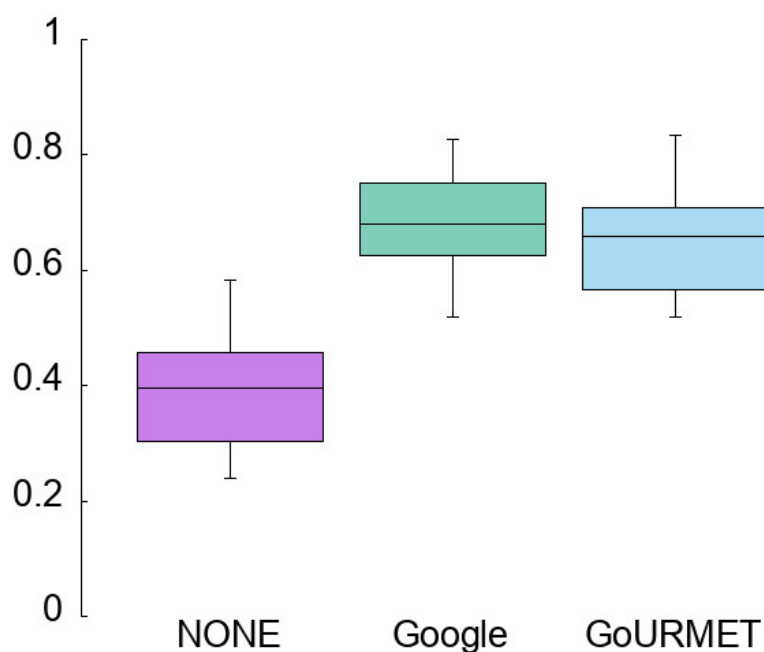


Figure 9: Accuracy results for GF evaluation for bg→en

Unique Evaluators	18
Number of unique gaps	83
Average number of gaps per sentence	2.8
Evaluations per gap-configuration	6.0
Number of evaluations by hint type: NONE	498
Number of evaluations by hint type: Gourmet	498
Number of evaluations by hint type: Google	498

Table 9: Summary of GF evaluation for tr→en

5.1.5 Swahili

Status: Silver standard - Complete

The statistics for the GF evaluation for sw→en are shown in Table 10. The detailed box plot of results is shown in Figure 11.

As may be seen, the boxes for Google and GoURMET clearly overlap, meaning that the difference in usefulness is not significant. However we also notice a slight overlap between the GoURMET box and the maximum success-rate of the baseline (NONE); this overlap does not occur with Google Translate.

5.1.6 Gujarati

Status: Silver standard - Ongoing

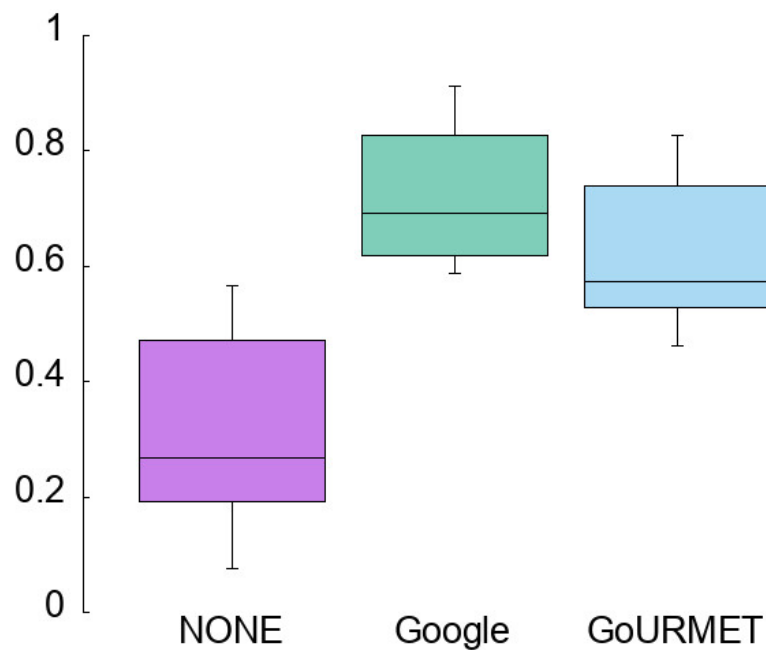


Figure 10: Results of GF evaluation for tr→en

Unique Evaluators	17
Number of unique gaps	70
Average number of gaps per sentence	2.3
Evaluations per gap-configuration	5.7
Number of evaluations by hint type: NONE	406
Number of evaluations by hint type: Gourmet	388
Number of evaluations by hint type: Google	396

Table 10: Summary of GF evaluation for sw→en

The statistics for the GF evaluation for gu→en are shown in Table 11. The detailed box plot of results is shown in Figure 12.

Unique Evaluators	14
Number of unique gaps	143
Average number of gaps per sentence	4.8
Evaluations per gap-configuration	4.7
Number of evaluations by hint type: NONE	677
Number of evaluations by hint type: Gourmet	666
Number of evaluations by hint type: Google	687

Table 11: of GF evaluation for gu→en

The number of gaps per sentence in the evaluation data appears to be significantly higher than the other languages evaluated at this time. It is likely this is down to the average sentence length being

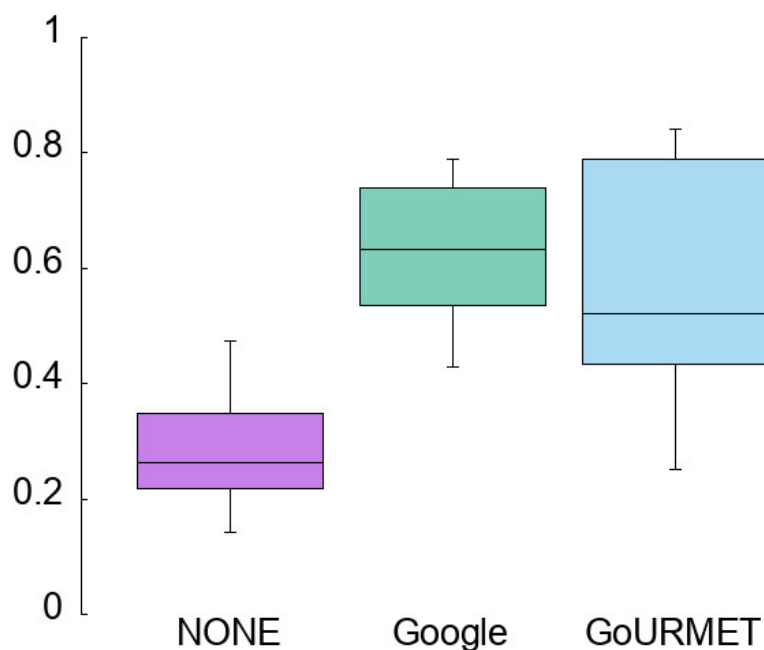


Figure 11: Results of GF evaluation for sw→en

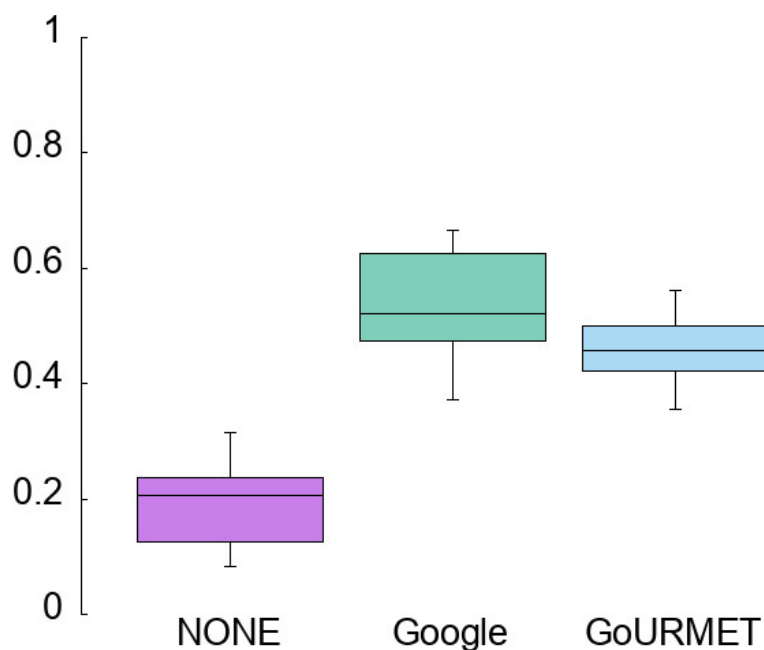


Figure 12: Results of GF evaluation for gu→en

longer. For this language pair the variation around the median for the GoURMET MT system is surprisingly small and warrants further investigation. The boxes for GoURMET and Google are coincident, but the minor extent of this suggests Google is likely to perform slightly better overall.

5.1.7 Serbian

Status: Silver standard - Ongoing

The statistics for the GF evaluation for $sr \rightarrow en$ are shown in Table 12. The detailed box plot of results is shown in Figure 13.

Unique Evaluators	10
Number of unique gaps	69
Average number of gaps per sentence	2.3
Evaluations per gap-configuration	3.3
Number of evaluations by hint type: NONE	230
Number of evaluations by hint type: Gourmet	232
Number of evaluations by hint type: Google	228

Table 12: Summary of GF evaluation for $sr \rightarrow en$

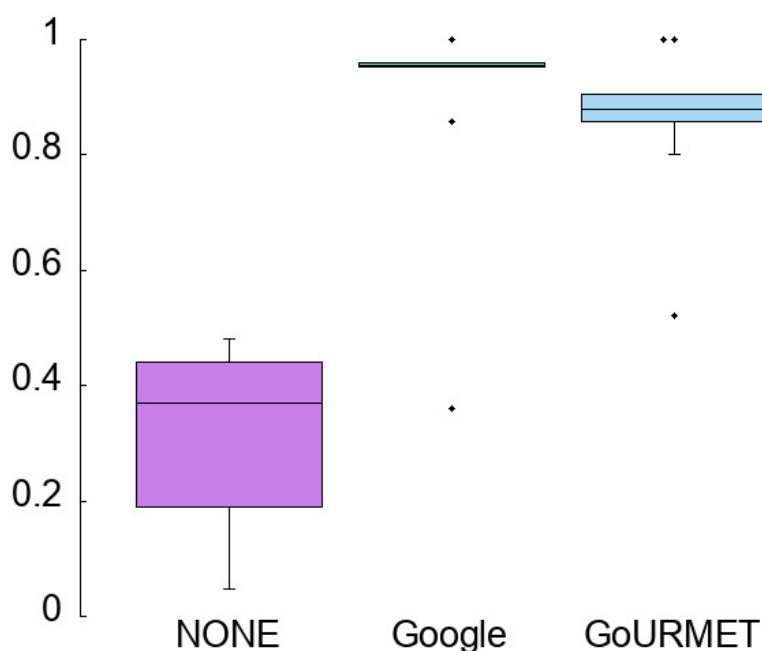


Figure 13: Results of GF evaluation for $sr \rightarrow en$

With a smaller number of evaluators than other languages, it is not possible to draw any solid conclusions at this time. However, the absolute scores for both the GoURMET and Google Translate systems for language pair $sr \rightarrow en$ appear very high indeed compared with other language pairs considered in this section.

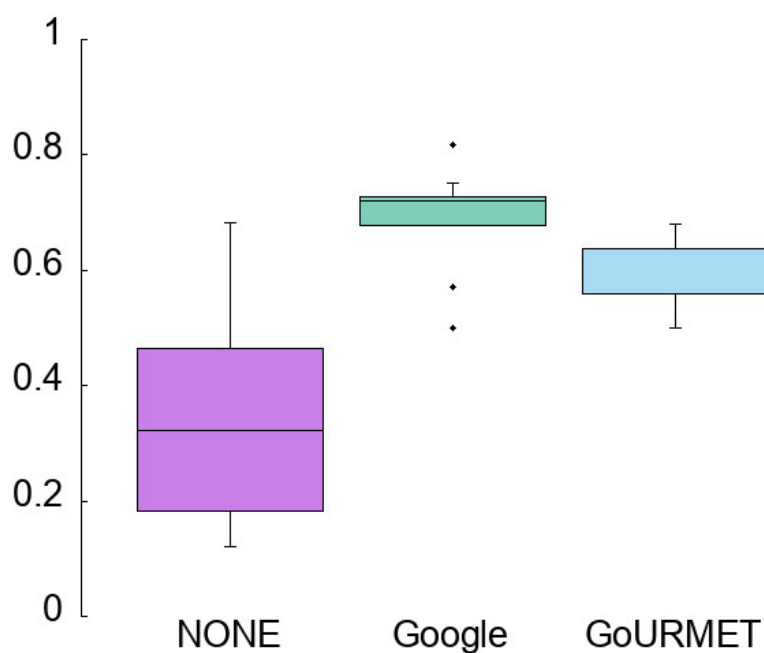
5.1.8 Kyrgyz

Status: Silver standard - Ongoing

The statistics for the GF evaluation for $ky \rightarrow en$ are shown in Table 13. The detailed box plot of results is shown in Figure 14.

The evaluation of this language pair is ongoing. Preliminary results suggest a fair overlap between the result for Google Translate and GoURMET, although at this stage, Google Translate is showing

Unique Evaluators	9
Number of unique gaps	74
Average number of gaps per sentence	2.5
Evaluations per gap-configuration	3.0
Number of evaluations by hint type: NONE	225
Number of evaluations by hint type: Gourmet	225
Number of evaluations by hint type: Google	225

Table 13: Summary of GF evaluation for ky→en**Figure 14:** Results of GF evaluation for ky→en

a significantly higher median at this stage. Further evaluation results are needed to have confidence in this result.

5.2 Direct Assessment

Direct assessment is carried out when English is the source language, machine-translated into a non-English target.

For each language pair, the results of the direct-assessment evaluations are summarised below. Mean scores and confidence intervals are shown for each annotator–language–question combination. Scores are shown for the MT system output (blue bars) and for human-created paraphrases (used for calibration, see page 14) of the references (orange bars). Since the paraphrases are considered to be perfect translations, we expect that they should achieve a maximum score. In one case the evaluation did not include paraphrases, so the orange bars are missing, see note below.

The 95% confidence intervals are estimated using bootstrap resampling, taking 10000 samples (with replacement) for each annotator–language–question combination and using the *Bias-corrected accelerated (BCA)* method to calculate the intervals, as implemented by `scikits.bootstrap`¹¹. We do not show confidence intervals for the paraphrase scores, as there are too few measurements.

Questions Q1 and Q2 are as stated previously in Section 2.2.7.

For each language pair, a summary of the free text feedback gathered from the screen shown in Figure 2 is given.

5.2.1 Bulgarian

Status: Silver standard - Complete

The results of the DA evaluation for language pair en→bg are shown in Figure 15 for evaluation question Q1.

Note that the methodology for en→bg differed from that adopted for the evaluation of the remainder of the DA evaluations. en→bg was the first language pair to be evaluated using DA under GoURMET WP5. After analysing the results and subsequently evaluating the methodology, the project partners decided to evolve the basic process used herein and a) introduce a second question (Q2) and b) introduce calibration sentences. Thus, evaluation of language pair en→bg required only the answering of Q1 and did not require evaluators to evaluate calibration sentences. The results below are therefore presented slightly differently from the other presented in this section.

User comments

Representative user comments for en→bg are as follows.

A number of evaluators commented on the difficulty of retaining details from the source to the target language:

- The weakness is often in the details - sometimes odd words or odd combination of words or words at odd places in the sentence. Sometimes the sentence would be almost correct if it wasn't for the lack of a crucial word, such as where the action is happening.
- Most of them would give you a fairly good idea of what the text is all about.

¹¹<https://github.com/cgevens/scikits-bootstrap>

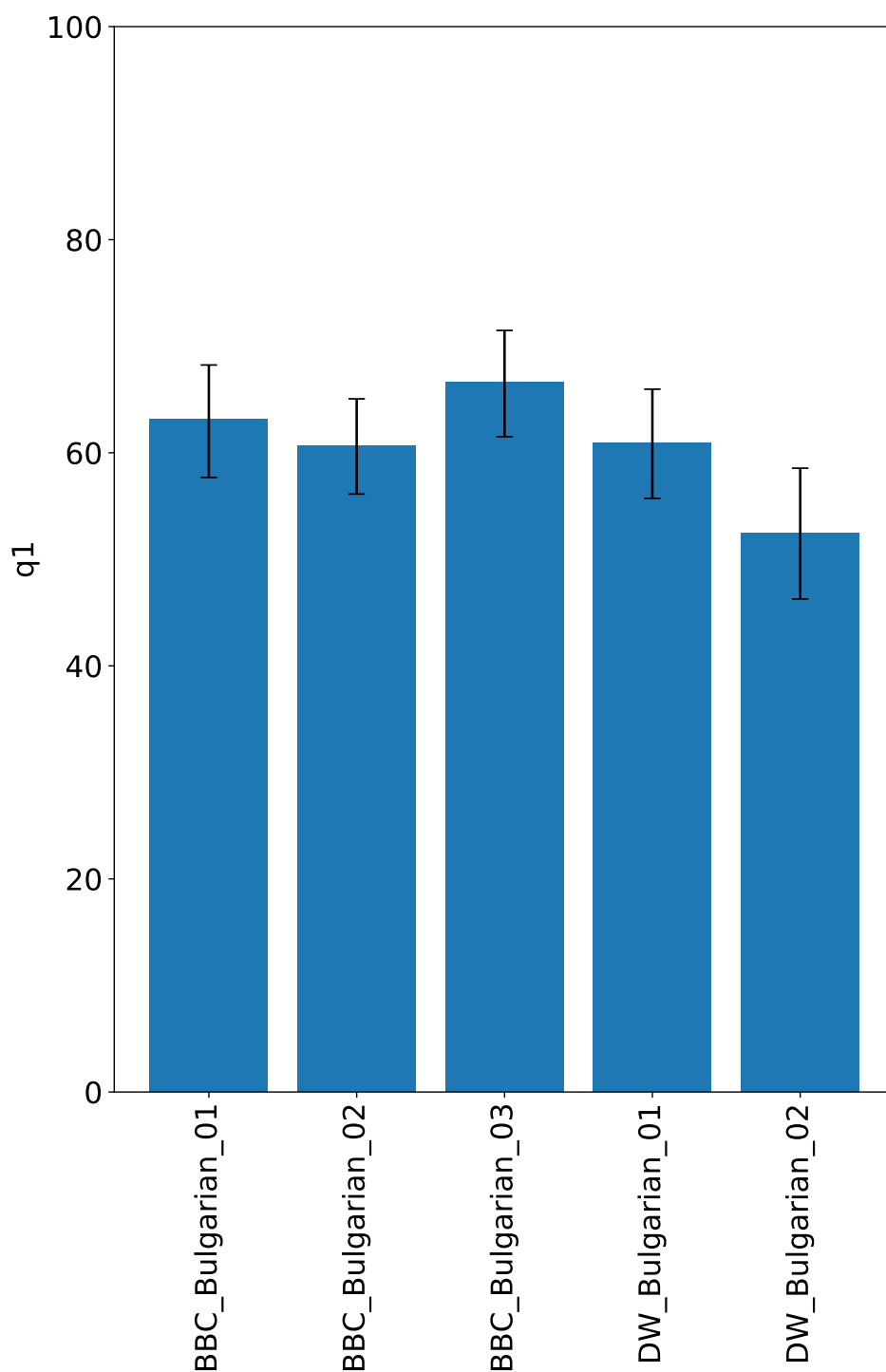


Figure 15: Results of DA evaluation for en→bg, mean scores for question Q1 – x-axis lists the users per organisation and the y-axis is the scoring by users

- Some of the translations are perfect. A lot of them however, while conveying the main message, contain odd expressions or grammar and often nuances are lost [...] while the translation may be perfect, the object of the main idea is not included.

A number of comments suggested some specific nuances of Bulgarian grammar which deviate from that of the source English were not adequately captured by the MT system:

- A few sentences completely misinterpreted the original sentence and gave a totally opposite meaning. (An example is someone saying that the Tirana authorities ”*have asked me for a bribe*” . Sentence 2 said: ”*they have given me a bribe*”.
- In one example, the first sentence quotes Andrei who says he was accused of wanting to kill his children “Андрей: Твърдят, че съм искал да убия децата си”, but the second, generated sentence suggests that the accusation is that he has already killed his children “Андрей: Обвинен съм, че съм убил децата ми”
- Sentence 1 said that government officials had *lined their pockets* with money, while, according to Sentence 2, they had *collected the money* — as if on behalf of the government
- The system repeatedly makes mistakes regarding the gender of nouns in Bulgarian. When one first sentence indicated it was about a female minister министърката, the second sentence used the masculine Министърът
- The system sometimes adheres to English-language grammatical rules about double negation. However, this is not necessarily appropriate in Bulgarian, where double negation is used to make claims, that, in English, would only require a single negative.

In common with feedback for other language pairs, the inability to translate previously unseen words, often proper nouns, was jarring:

- Sometimes, English words with their original Latin spelling would appear in Sentence 2 (usually names of people or titles like *Rolling Stone* magazine, but also a word like *slimmer*.)
- I recognised that in Sentence 2 the names (especial personal names) were in Latin whereas in Sentence 1 they were always Cyrillic
- Generated sentences repeatedly failed in translating names, sometimes to confusing effect. In most cases, the name was simply not transliterated into Cyrillic and written out in Latin instead.
- The system even failed to provide Bulgarian equivalents of foreign names or words that have very similar and widely used equivalents in Bulgarian.

Finally and again in common with other language pairs, longer sentences were more prone to mistranslation:

- The machine is doing well only with the short sentences.
- Especially the short sentences were almost identical. [The human and machine translated sentences matched]

5.2.2 Turkish

Status: Silver standard - Complete

The results of the DA evaluation for language pair `en→tr` are shown in Figure 16 for evaluation question Q1 and Figure 17 for evaluation question Q2.

User comments

Representative user comments for `en→tr` are as follows.

There were some illuminating but consistent comments about common mistakes around transliteration (or lack thereof) of proper nouns and translation of numbers:

- Overall, the translations and sentence structures were much better than I assumed they would be. There are problems with special names like movie titles, which the system kept in English.
- The machine seems to have the most trouble with numbers. I don't think I've seen a single instance where they were right.
- I also noticed that the tool is very bad at recognising the numbers. It generally makes mistakes when it tries to use the numbers in a translated sentence.
- In some sentences, the private names (like street names) were misspelled. For instance, in the gray text, the name of Tiananmen Square was right, but in the UI text it was misspelled.

Feedback suggested the MT system was poor at translating long sentences:

- Shorter sentences had more accurate translations.
- Shorter sentences had better translations but more inaccurate grammar.
- There are very few times that long and complex sentences are translated accurately. Short sentences seem to work better.

5.2.3 Swahili

Status: Silver standard - Complete

The results of the DA evaluation for language pair `en→sw` are shown in Figure 18 for evaluation question Q1 and Figure 19 for evaluation question Q2.

User comments

Representative user comments for `en→sw` are as follows.

A number of comments suggested the MT system was creating overly long sentences with a degradation in quality:

- Some sentences are too long to sustain actual meaning when translated.
- Some sentences are too long so the translations lose meaning.

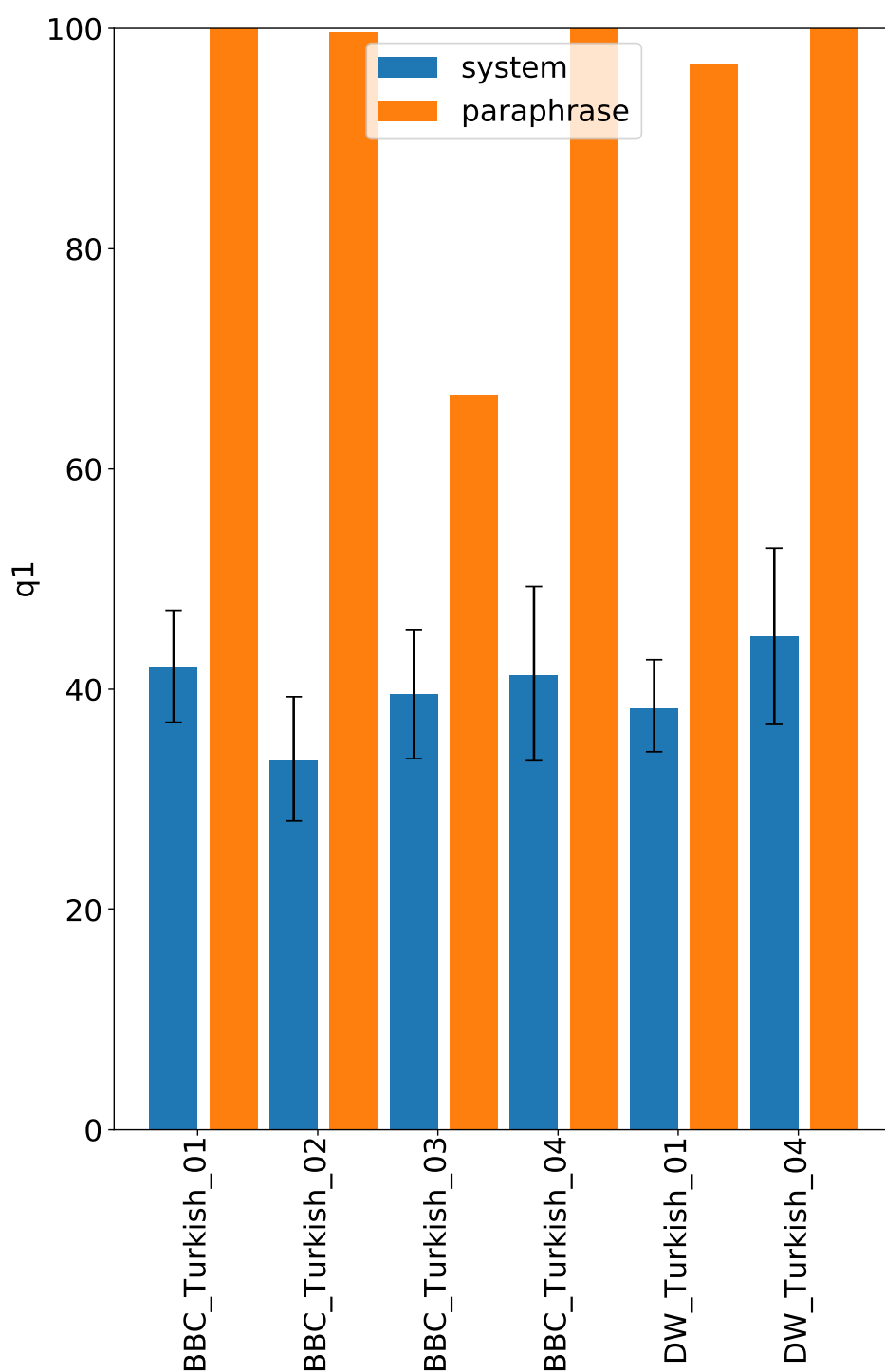


Figure 16: Results of DA evaluation for en→tr, mean scores for question Q1 – The x-axis lists the user per organisation, the y-axis gives user scoring of MT output (blue bar) and user scoring of the ground truth sentence (orange bar)

- The shorter the sentences were, the more correct they tended to be.

Some comments called into question the quality of the reference translations:

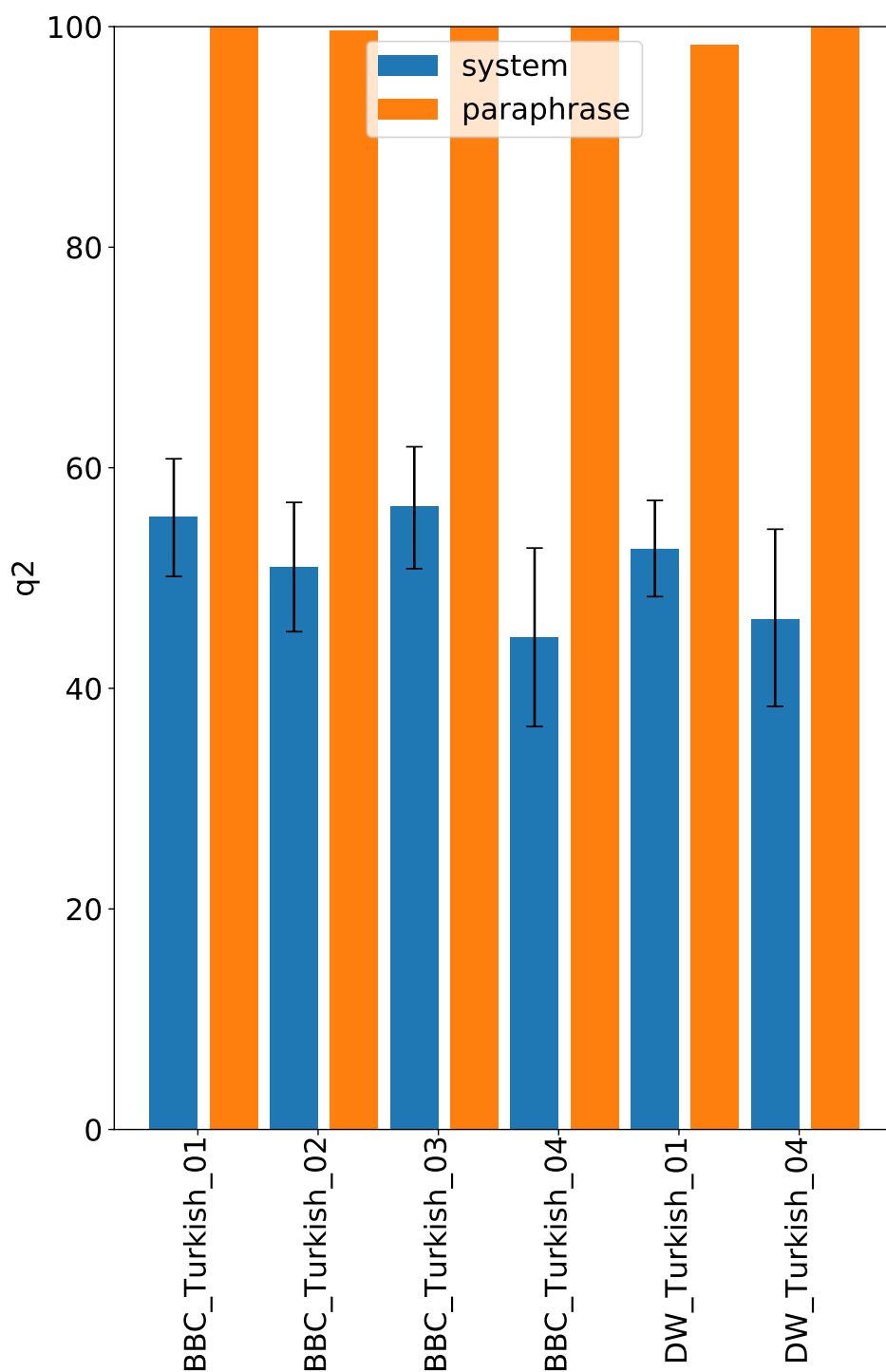


Figure 17: Results of DA evaluation for en→tr, mean scores for question Q2

- Although, according to the information above, sentences in black were written by persons, [...] in some cases they were either confusing, or made no sense at all.
- What I have observed is that it only a few cases where the two sentences carried the same meaning, at the same time both being grammatically correct. Although, according to the

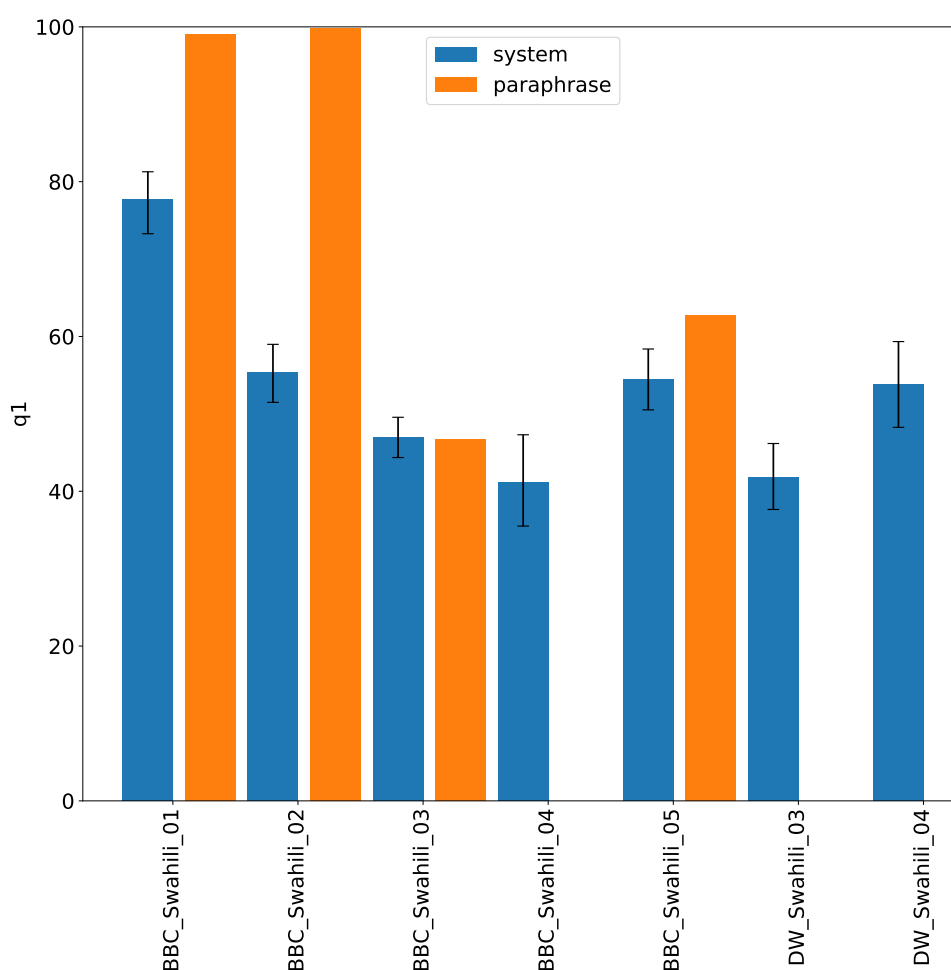


Figure 18: Results of DA evaluation for en→sw, mean scores for question Q1

information above, sentences in black were written by persons, they were the ones which mostly diverted from the original meaning, and in some cases they were either confusing, or made no sense at all.

General comments on the quality of the MT included:

- My general impression is that this machine is still very far from being accurate or at least coming close to correct and accurate translation. Most of the sentences were grammatically wrong and distorted the real meaning of what is really intended. There were too many distortions in the translations. I also noticed some sentences were mixing Kiswahili and English words; this does not make any sense at all in translation because the intended message is not delivered. I would not at the moment recommend the use of this translation engine but rather encourage further research with the aim of developing and improving it further.
- The general point of note for me is that some of the machine-generated sentences did not capture the meaning of the sentences that were written but people. But I generally found most of the machine-generated sentences to be grammatically and idiomatically correct.
- In many cases the machine tried to put everything in Kiswahili, where as the person in several

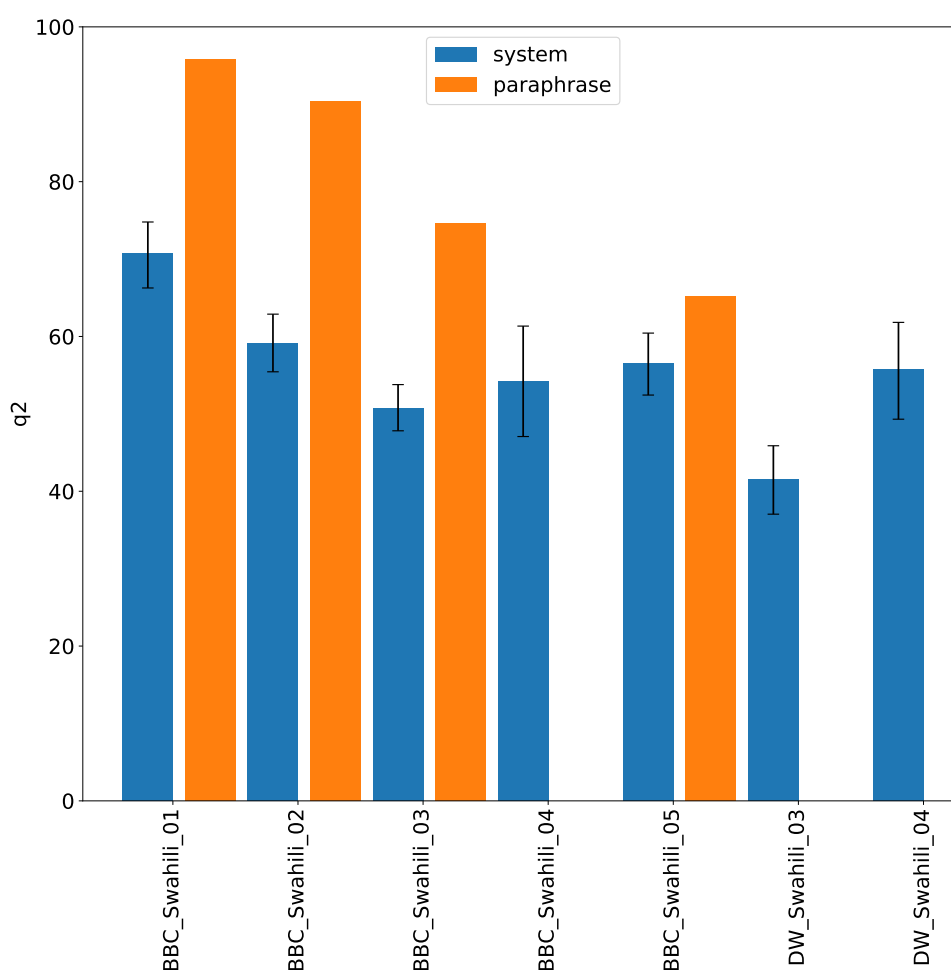


Figure 19: Results of DA evaluation for en→sw, mean scores for question Q2

cases used English words in sentences. At times the sentences [sic] completely differ but then the person writes a more meaningful sentence than a machine.

- Some Sentence Twos [the MT sentence] had words missing.

5.2.4 Gujarati

Status: Bronze standard - Complete. Silver standard - Ongoing

Only early preliminary results are available for Gujarati at present, due to the small number of results received. The results here are illustrative only and should not be used for the purposes of comparison.

The results of the DA evaluation for language pair en→gu are shown in Figure 20a for evaluation question Q1 and Figure 20b for evaluation question Q2.

User comments With only a few evaluators at this stage, it is too early to draw themes from the free text comments, however, general comments echoed sentiments found previously with other language pairs:

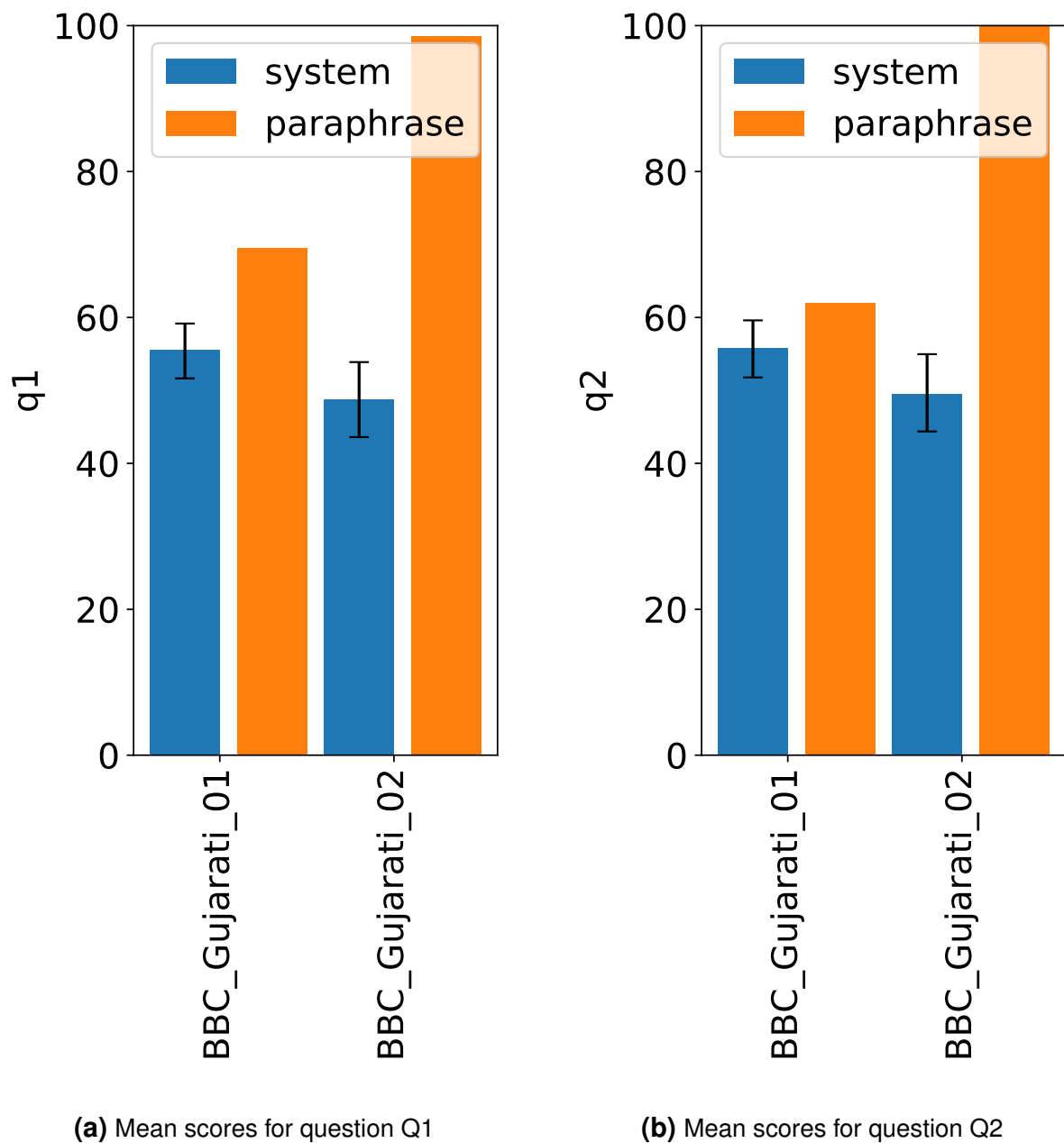


Figure 20: Results of DA evaluation for en→gu

- On the whole, the sentences were conveying what was written. Mainly instances of tense and the odd word being translated incorrectly. Often happening where there is a spelling of a name or building that doesn't exist as a word in Gujarati.
- Some of them were quite difficult to decipher, particularly the longer sentences.
- I think there were always discrepancies when it came to the translations of names and places. Also where English words were written out in Gujarati.

5.2.5 Serbian

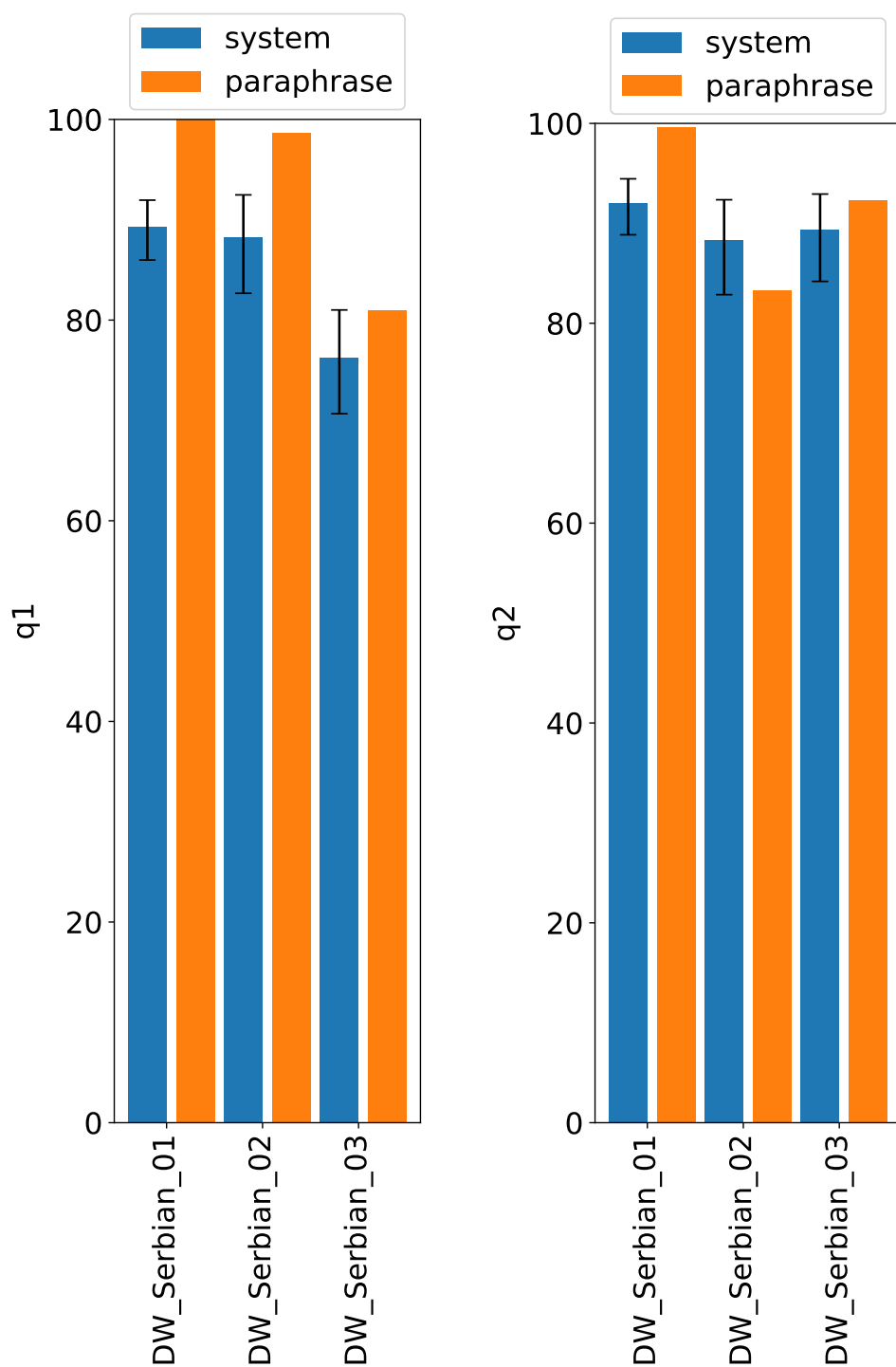
Status: Bronze standard - Complete. Silver standard - Ongoing

The results of the DA evaluation for language pair en→sr are shown in Figure 21a for evaluation question Q1 and Figure 21b for evaluation question Q2.

User comments

The comments revealed some issues where perhaps the differences in grammatical structure between the two languages were causing difficulties. It appeared that the gendered nature of the language appeared to cause a few issues as well as the way names are identified and translated. As is found to be common, idioms do not translate well.

- There are many sentences in which I couldn't tell which translation is better, without knowing the context. For instance, "I was" translates differently if the person speaking is male or female ("Bio sam", "bila sam").
- In a few cases the gender is translated in a wrong way
- The names of organisations should always be written after the type of organisation ("Univerzitet Oksford" rather than "Oksford univerzitet").
- The machine translation is often shorter and even better for a journalistic sentence. However, sometimes it leaves an important word out and some idioms are inaccurate.



(a) Mean scores for question Q1

(b) Mean scores for question Q2

Figure 21: Results of DA evaluation for en→sr

6 Research Outputs

6.1 Publications

This paper reports on evaluation research done under the GoURMET project.

- Felipe Sánchez-Martínez, Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, Mikel L. Forcada, Miquel Esplà-Gomis, Andrew Secker, Susie Coleman, and Julie Wall. An English-Swahili parallel corpus and its use for neural machine translation in the news domain. In *Proceedings of the 22th Annual Conference of the European Association for Machine Translation*, pages 299–308, Online Conference, November 2020

6.2 Software

The software tools used for running the gap-filling and direct-assessment evaluations have been released by the project as open source. See Section 3.3 for details.

- The direct-assessment evaluation tool <https://github.com/bbc/gourmet-sentence-pairs-evaluation>
- The gap-filling evaluation tool <https://github.com/bbc/gourmet-gap-fill-evaluation>
- The GA pre-processing script <https://gitlab.com/mlforcada/bbc-dw-gf>

7 Conclusion

This document has presented an interim view of the evaluation work undertaken by GoURMET. Methodologies for both automatic and human evaluation have been agreed by project partners. The methodologies have been chosen in order to best support the project use cases of media monitoring and global content creation primarily through the human evaluation, whilst allowing the quality of the translation models to be compared across GoURMET languages and with externally created systems via the automatic evaluation metrics.

The human evaluation makes use of the 3-tier system. Bronze and silver grade evaluations for direct assessment and gap filling have been presented here. These will continue to a complete set of results for languages presented in this document, but where interim results have been presented, and new languages delivered over the second half of the project.

During the second half of the project, the gold standard evaluation as described in GoURMET Deliverable D5.1 Secker et al. (2019a) will be tested on those languages which have performed particularly well at GF and/or DA. The translation service described in GoURMET deliverable D5.3 Secker et al. (2020) will allow evaluation based on post editing time, quality and so on, by allowing the creation of custom prototypes and tools which can be used to gather these metrics.

A complete set of results for both automatic and human evaluation will be presented at the end of the project in GoURMET deliverable D5.6.

References

- Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5301. URL <https://www.aclweb.org/anthology/W19-5301>.
- Ondrej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 272–307, Belgium, Brussels, October 2018. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/W18-6401>.
- M Esplà-Gomis and M.L. Forcada. Combining content-based and url-based heuristics to harvest aligned bitexts from multilingual sites with bitextor. *The Prague Bulletin of Mathematical Linguistics*, 93:77–86, 2010.
- Mikel L Forcada, Carolina Scarton, Lucia Specia, Barry Haddow, and Alexandra Birch. 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, 2018a.
- M.L. Forcada, C. Scarton, L. Specia, B. Haddow, and A. Birch. Exploring gap filling as a cheaper alternative to reading comprehension questionnaires when evaluating machine translation for gisting. *CoRR*, abs/1809.00315, 2018b. URL <http://arxiv.org/abs/1809.00315>.
- Y. Graham, T. Baldwin, M. Dowling, M. Eskevich, T. Lynn, and L. Tounsi. Is all that glitters in machine translation quality estimation really gold? In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3124–3134, Osaka, Japan, 2016a. URL <http://aclweb.org/anthology/C16-1294>.
- Y. Graham, T. Baldwin, A. Moffat, and J. Zobel. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30, 2016b.
- Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/W13-2305>.
- Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. Results of the wmt19 metrics shared task: Segment-level and strong mt systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy, August 2019. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/W19-5302>.

Jim O'Regan and Mikel L Forcada. Peeking through the language barrier: the development of a free/open-source gisting system for basque to english based on apertium. org. *Procesamiento del Lenguaje Natural*, (51):15–22, 2013.

Kishore Papineni, Salim Roukos, Todd Ward, John Henderson, and Florence Reeder. Corpus-based comprehensive and diagnostic mt evaluation: Initial arabic, chinese, french, and spanish results. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, page 132–137, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

Maja Popović. chrF: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, 2015.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/W18-6319>.

Carolina Scarton and Lucia Specia. A reading comprehension corpus for machine translation evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3652–3658, 2016.

Carolina Scarton, Mikel L. Forcada, Miquel Esplà-Gomis, and Lucia Specia. Estimating post-editing effort: a study on human judgements, task-based and reference-based metrics of mt quality, 2019.

Andrew Secker, Alexandra Birch, Peggy van der Kreeft, and Felipe Sánchez-Martínez. GoURMET: Deliverable 5.1 - evaluation plan, 2019a.

Andrew Secker, Julie Wall, Peggy van der Kreeft, and Susie Coleman. GoURMET: Deliverable 5.2 - use cases and requirements, 2019b.

Andrew Secker, Susie Coleman, Mikel L. Forcada, Anna Blaziak, Rachel Bawden, Radina Dobрева Felipe Sánchez-Martínez, and Víctor M. Sánchez-Cartagena. GoURMET: Deliverable 5.3 - initial integration report, 2020.

Felipe Sánchez-Martínez, Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, Mikel L. Forcada, Miquel Esplà-Gomis, Andrew Secker, Susie Coleman, and Julie Wall. An English-Swahili parallel corpus and its use for neural machine translation in the news domain. In *Proceedings of the 22th Annual Conference of the European Association for Machine Translation*, pages 299–308, Online Conference, November 2020.

ENDPAGE

GoURMET

H2020-ICT-2018-2 825299

D5.4 Initial Progress Report on Evaluation