Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

# A post-editing approach to machine translation evaluation at the document-level

David Dunne
*Supervised by Prof. Yvette Graham*

April 19, 2022

A dissertation submitted in partial fulfilment
of the requirements for the degree of
MAI (Computer Engineering)

# Declaration

I hereby declare that this dissertation is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at `http://www.tcd.ie/calendar`.

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at `http://tcd-ie.libguides.com/plagiarism/ready-steady-write`.

Signed: _____ Date: _____

# Abstract

Machine translation has become ever more prominent in society with Google translate acknowledging over 500 million users in 2016. With this reliance on machine translation, it is important to determine quality and whether human parity has been reached. Automatic metrics like BLEU scores can be used to estimate the quality of machine translations but these methods are often considered unreliable. Therefore, manual evaluations campaigns are carried out including at the annual workshop on statistical machine translation (WMT). There has been an appeal in the machine translation community to start evaluating at the document-level to give annotators more context. As such, the direct assessment method used at WMT was adapted to suit a document-level evaluation instead of isolated sentences and segments. However, the method has fallen short at providing full document context while also failing to give annotators a chance to edit previous segments.

The research undertaken aimed at addressing this problem of providing full document-level context while also ensuring high-levels of quality control for the crowd-sourced assessments. This project is the first to explore the use of post-editing at the document-level for machine translation with crowd-sourced annotations. The benefits of this approach over existing evaluation methodologies are that it gives full document context, removes subjective numerical judgements and it can specify where MT systems go wrong. Evaluations were carried out by researchers as well as crowd-sourced workers. In addition, inter-annotator agreements levels were calculated against these post-edits using Cohen's kappa.

The results demonstrated some correlation in terms of machine translation system rankings with official results that were obtained at WMT 20. In the crowd-sourced evaluation campaign, 2 cluster groups of rankings emerged compared to 4 cluster groups of rankings from official WMT results. In terms of inter-annotator agreement for post-editing, it was found that annotators showed moderate agreement when choosing to edit or not edit a segment ($\kappa = 0.4821$) but this agreement was not as strong when crowd-sourcing ($\kappa = 0.2826$). Furthermore, the QC mechanisms proved effective and allowed for a manual inspection of submitted data, something that current machine translation evaluations can not feasibly do. Overall, the work carried out demonstrated that crowd-sourcing document-level post-edits is a feasible approach to machine translation evaluation.

# Acknowledgements

I would like to thank my supervisor, Prof. Yvette Graham, who helped me throughout the course of the year. I would also like thank members of Prof. Graham's natural language processing group who carried out evaluations, in particular Chenyang Lyu. The work undertaken would not have been possible without necessary funding and so I would like to acknowledge WMT.

I would like to thank my brother for his support and for inspiring me to pursue Computer Engineering.

Finally, I would like thank my family and friends, especially my parents who have supported me throughout my academic journey.

# List of Figures

# List of Tables

# Contents

# Nomenclature

| | |
|---|---|
| AMT | Amazon Mechanical Turk |
| DA | Direct Assessment |
| DPEA | Document Post-editing Application |
| HIT | Human Intelligence Task |
| MT | Machine Translation |
| NMT | Neural Machine Translation |
| QC | Quality Control |
| RR | Relative Ranking |
| SR+DC | Segment-level rating with document context |
| SR-DC | Segment-level rating without document context |
| UI | User Interface |
| WMT | Workshop on Statistical Machine Translation |

# 1 Introduction

Translation is a process that enables those that do not have a common language to communicate. Traditionally, translation has been performed by human interpreters or translators but this came at the cost of time, money and availability of translators. In 1949, Warren Weaver presented a set of proposals for machine based translations [5], marking what one might consider the birth of machine translation (MT). MT did not have a strong start, a report in 1966 by the Automatic Language Processing Advisory Committee concluded that "there is no immediate or predictable prospect of useful machine translation" and that MT was slower, less accurate and more expensive than humans [6]. However, popularity for MT grew, notably when Google Translate was launched in 2006, recording over 500 million users in 2016 [7].

As more MT systems are developed, it proves a challenge in determining the quality of each MT system and whether a MT system has reached human parity. Furthermore, the evaluation of MT can be considered a subjective process which creates complexities in determining quality.

## 1.1 Background & Motivation

WMT is an annual workshop on statistical machine translation where research teams submit MT systems for evaluation. The news translation task requires the translating of news articles from a source to a target language. At WMT 20, the into English language pairs were evaluated using direct assessment (DA), an MT evaluation method inspired by Graham et al. [1].

For the method of DA, annotators evaluate translations using a 0-100 point Likert scale, segment by segment (similar to sentence) in a random order through Amazon Mechanical Turk (AMT). However, since WMT 19 there has been an appeal to move towards a document-level approach instead of this existing segment-level approach. This would result in showing the entire document to the annotator as opposed to segments in isolation.

The reason for this appeal for document-level evaluation was due to the findings of Läubli

et al. [8] and Toral et al. [9]. MT evaluation experiments were conducted [8], concluding that human raters strongly preferred human translations over MT. However, this level of preference was not as obvious when presenting the raters with isolated sentences. This raised the question whether MT evaluations should include the context of the entire document in order to provide a more reliable evaluation. It also puts a question mark on conclusions drawn from any prior work that was conducted at the segment-level or sentence-level.

It was therefore considered of "high practical relevance" to test whether MT evaluation could be carried out effectively at the document-level. A number of methods for MT evaluation at the document-level have been trialed, most notably methods from Castilho [10], Barrault et al. [2] and Bojar et al. [11]. In spite of that, these methods have not gained much traction as WMT have still been appealing for improved MT evaluation methods at the document-level. In addition, Castilho [10] highlights that human-evaluation of MT at the document-level is in its infancy and it is essential to test more methodologies, while also pointing out that there is a "lack of [a] proper tool able to handle different MT evaluation methodologies."

## 1.2 Project Objective

This project has aimed to develop a tool and methodology for MT evaluation at the document-level that is more reliable than the existing solutions. With this in mind, the feasibility of this approach can be assessed by answering the following questions:

(i) Can machine translations be reliably evaluated at the document-level with crowd-sourced post-edits?

(ii) How effective are the quality control techniques which have been developed for this task?

(iii) How strongly do the results of this newly proposed method of evaluation correlate with official results of past WMT evaluation campaigns?

If these criterion proved successful, then this approach could be considered as a method for official evaluation at WMT. It would be beneficial if this method could provide information as to where MT systems go wrong. In addition, high inter-annotator and intra-annotator agreement levels would be desirable.

## 1.3 Challenges

MT evaluations can be carried out using automatic metrics like BLEU [12] scores which are cheap and fast but are often considered unreliable methods [13, 14]. As such, manual

human evaluation is often carried out but this can lead to other difficulties and challenges. Crowd-sourcing monolingual evaluations is cost effective but requires quality control checks [1] to ensure the reliability of the results. Furthermore, it was discussed that monolingual evaluation can introduce a reference bias [15], but this particular claim has been dis-proven [3]. Similarly, bilingual evaluations can introduce a bias as the annotator generally has different language skills for different languages [3]. Getting annotators to agree with one another as well as themselves has also proven troublesome [1]. These are just some of the challenges which make MT difficult to evaluate.

## 1.4 Approach

The approach chosen to address document-level evaluation was monolingual post-editing with human-targeted metrics. A similar setup was carried out at the segment-level yielding positive results in [16] but came under question in [3]. As such, it would be of interest in performing similar experiments but at the document-level. The project also seeks to address the problem of developing a tool that can be easily used to perform document-level evaluations. The post-edits produced from this MT method could train NMT systems further, ultimately leading to improvements in MT. In addition, this method would give research teams who participated in the translation tasks clarity as to where their MT systems went wrong, something that current methodologies do not offer.

## 1.5 Roadmap

The remainder of the paper is broken down as follows.

- **Chapter 2 - Background:** outlines MT evaluation methods and challenges associated with MT evaluation campaigns.
- **Chapter 3 - Design & Methodology:** illustrates the configuration and implementation of the approach.
- **Chapter 4 - Results & Evaluation:** analyses the QC mechanisms, IAA experiments and MT system rankings obtained from the proposed MT evaluation method.
- **Chapter 5 - Conclusions & Future Work:** discusses the feasibility to the approach with suggestions for future experiments.

# 2 Background

Shared translation tasks are conducted for the annual Workshop on Machine Translation (WMT), including the news translation task. For this task, research teams are provided with news text that they must translate from a source language to a target language, e.g. German-English. The research teams tend to use NMT models to translate this text and afterwards, they submit these translations to WMT for evaluation. As such, the quality of MT must be assessed reliably and fairly. This section discusses some of these evaluation methods.

## 2.1 Automatic Metrics

MT can be evaluated using automatic metrics like BLEU [12] scores which are quick to calculate, inexpensive and language-independent. Although there is some evidence for correlation with BLEU scores and human judgements, this correlation and method has come under question [13, 14]. Callison-Burch et al. [13] demonstrate using two counter examples that "an improved BLEU score is neither necessary nor sufficient for achieving an actual improvement in translation quality". It has also been discussed that BLEU scores are based on lexical similarity and not on meaning [1]. Moreover, there are strong advantages in using automatic metrics like BLEU but ultimately, they are often considered unreliable methods of evaluation [13, 14]. As such, manual human evaluation is carried out, including for the annual Workshop on Machine Translation (WMT).

## 2.2 Direct Assessment (DA)

Graham et al. [1] pointed out an alternative MT evaluation method compared to the relative preference method that was being used by WMT to evaluate at the time. This proposed method involved collecting crowd-sourced assessments of MT quality through Amazon Mechanical Turk (AMT). Human Intelligence Tasks (HITs) were posted on AMT whereby workers could complete these HITs in exchange for a fee. These HITs could evaluate both adequacy and fluency, require only monolingual annotators and were cheap to carry out.

Annotators used a 0-100 point Likert scale to capture the quality of the MT, see figure 2.1. In this case, translations were considered in isolation to minimize the kind of bias created

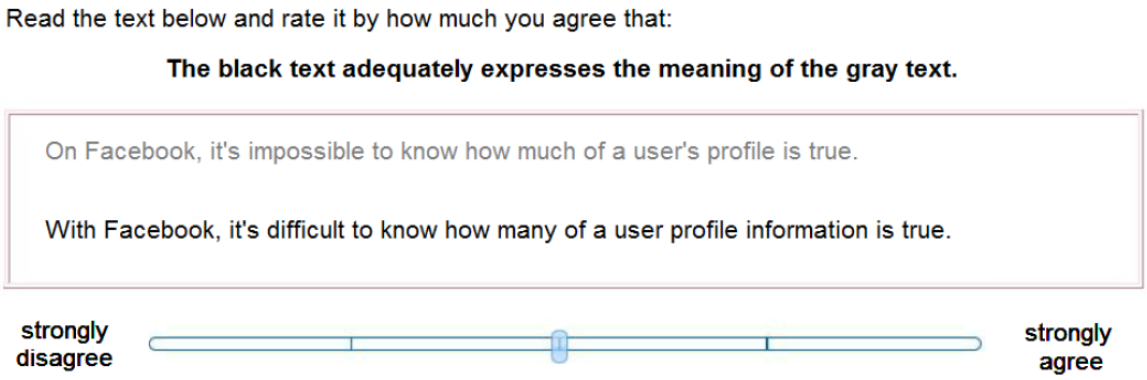


Figure 2.1: Screenshot of the adequacy assessment interface. Screenshot taken from [1]

when performing the relative preference method [17], where multiple translations are shown.

The use of monolingual annotators meant that there was a larger pool of potential assessors. Bilingual evaluations (source-based) are often bottle-necked by the availability and cost of expert translators. Thus, more assessments can be collected when carrying out monolingual evaluation (reference-based) and at a cheap cost. However, the use of crowd-sourcing for monolingual evaluation runs the risk of annotators trying to "game" the system, which would ultimately contaminate the results. Therefore, quality control checks were put in place, taking up approximately 30% of the data that would be collected per HIT.

The results from this 0-100 point scale were standardized to remove any individual biases. The paper concludes that this method is scalable, highly efficient and cost effective. Moreover, it proved to be more reliable than the method of relative preference. This method proposed by Graham et al. [1] replaced the relative preference method, becoming the standard for MT evaluations at WMT since 2017, now known as direct assessment (DA).

From a more critical standpoint, the paper notes that some workers might lack the necessary literary skills to complete evaluations effectively. In addition, it notes that this method was feasible for into English translations but not other language pairs such as into Czech translations. A further drawback with DA is that the results don't specify where the candidate translations go wrong, they only present the best to worst order of the machine translation systems.

The setup used in this paper carried out evaluations at the sentence-level and not the document-level. Sentence-level evaluation meant that 100 sentences could be placed inside

one HIT which was practical in terms of the time to complete a HIT, the cost of a HIT and the ability to incorporate effective quality control checks. To make this method document-level compatible, one could try insert 100 documents into a single HIT but then this would increase the time and cost to evaluate each system. If one was to try a setup with less documents, for example 10 documents in a single HIT, then standardising scores and quality control could become problematic. This demonstrates the difficulties and challenges that can arise when performing MT evaluation, particularly as the community moves towards this document-level evaluation approach. Therefore it is important to find a balance and trade-off in MT evaluation methods.

Due to this demand for document-level evaluation, DA has been trialled with different formulations since WMT 19 in order to come up with a more reliable MT evaluation. Document-level rating with document context (DR+DC) was performed at WMT 19 where a single score was given for an entire document. It was concluded in [18] and [19] that the use of DR+DC was problematic in terms of statistical power due to the reduced sample size of documents, while also producing inconclusive ties. Therefore, this particular method was not pursued the following year at WMT 2020 [2]. Segment-level rating with document context (SR+DC) proved more promising as it allowed for "a sufficient sample size of translations but importantly keeping the context available to human assessors" [11]. For this SR+DC setup, segments (similar to sentences) were presented in sequential order, screen by screen, but it appears that there was no ability for an annotator to go back and edit a previous score or go back/forward to see a previous segment to get context. Evidently, this setup is somewhat questionable as it doesn't give the annotator the full document context in one screen while also giving no opportunity to change prior segments of a document. This drawback was dealt with for the out of English evaluations at WMT 2020 [2], enabling annotators to see the entire document on a screen broken down by individual segments one-by-one. This gave annotators the ability to go back and change their scores for prior segments, see figure 2.2. In spite of this, the into English configuration for SR+DC did not apply this configuration at WMT 20. The DA method used for into English evaluation has yielded reliable results in recent years at the segment-level but this method may no longer be as feasible at the document-level for crowd-scoured evaluation campaigns.

## 2.3 Post-editing (PE)

Both monolingual and bilingual annotators can carry out post-editing in which they perform edits on a candidate translation. Post-editing time, post-editing distance and post-editing effort scores can be used to determine the quality of MT [20]. To measure post-editing time, a clock is normally run throughout the evaluation of a translation [10]. It can not be guaranteed that an annotator's focus is entirely invested throughout the evaluation and so

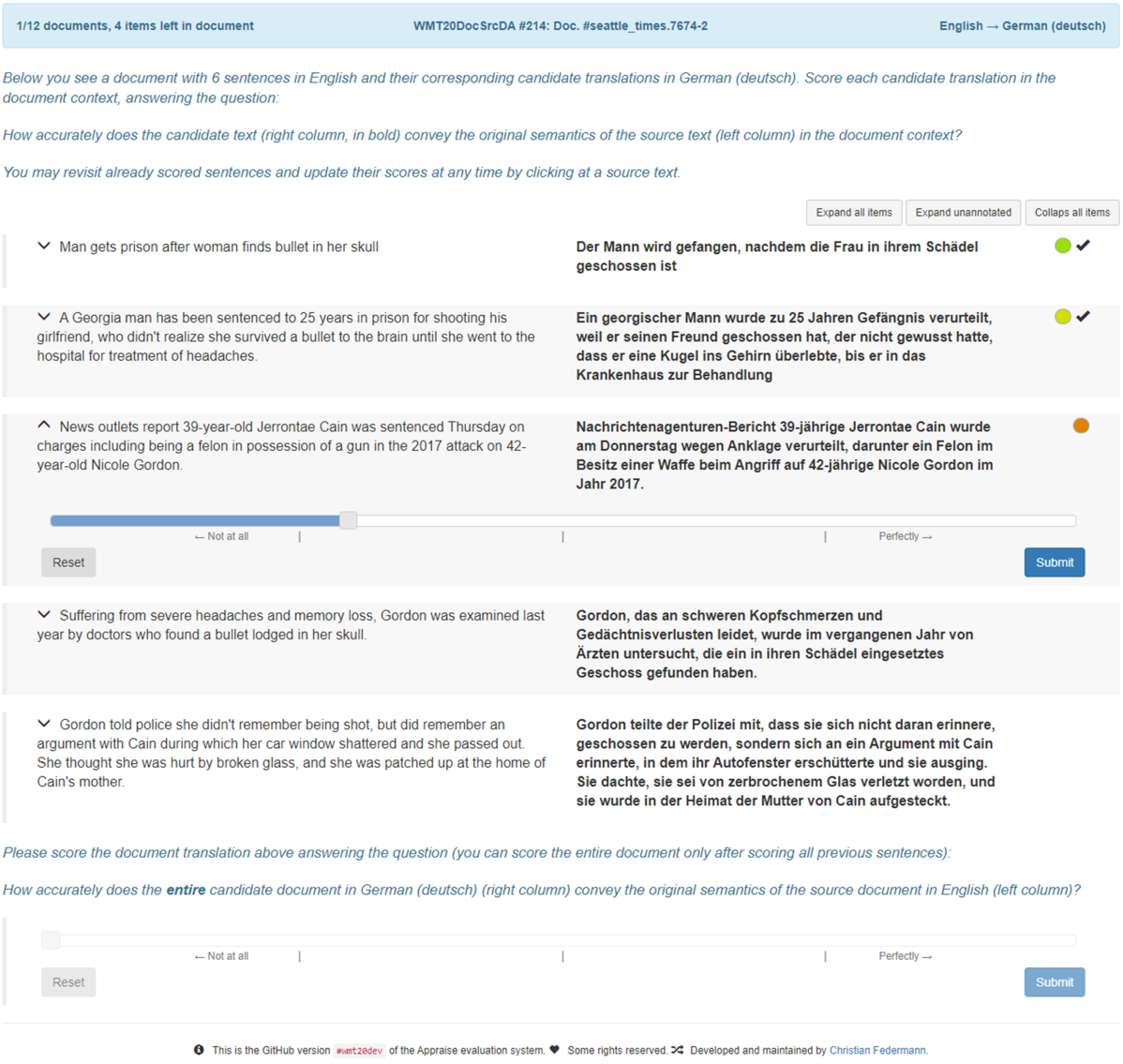


Figure 2.2: Screen shot of the document-level out-of-English DA configuration in the Appraise interface, taken from WMT 20 [2]

this raise question about the reliability of this measurement. Similarly, post-editing effort scores may not be an ideal form of assessment as annotators literary skills can vary in different areas. For example, an annotator finds spelling more difficult than grammar and so their literary ability would influence their post-editing effort scores. In conclusion, post-editing distance through human targeted metrics tend be more widely used [3, 16, 21].

### 2.3.1 Human-targeted Metrics

A human-targeted metric known as HTER (derived from Translation Edit Rate (TER)) has been used to measure post-editing distance and has demonstrated high correlations with human judgements [16]. TER is defined as "the minimum number of edits needed to change a hypothesis so that it exactly matches one of the references", given by the following formula:

$$TER = \frac{\#\ of\ edits}{average\ \#\ of\ reference\ words}$$

Snover et al. [16] argue that one of the difficulties in MT evaluation is that humans have to assign "a subjective numerical judgement to a translation". A further challenge is achieving high levels of inter-annotator agreement [1]. This problem arises from the fact that annotators struggle on what factors of a translation they consider "good" or "bad" [21]. In the HTER setup, humans do not apply a numerical judgement to translations, instead this numerical value is determined by a human-targeted metric. Therefore one could argue that this setup is less subjective. Snover et al. [16] also outline that HTER correlated better with average human judgements than human judgements did with each other.

As previously discussed, the use of automatic metrics is often considered unreliable. Thus, the use of HTER is somewhat non-ideal as it relies on the TER metric. However, one could argue that the reference bias is eliminated in the HTER setup as the candidate translation is calculated against the post-edited translation. A further concern with human-targeted metrics is that different metrics have shown to correlate more strongly with human judgments for certain tasks and language pairs [3, 21]. For example, HBLEU has shown higher correlations with human judgements than HTER for German to English but not for English to Spanish [3]. Therefore, this raise the question on which human-targeted metric to use when conducting post-editing evaluations.

### 2.3.2 Comparing Direct Assessment & Human-targeted Metrics

As previously stated, DA is the official evaluation method that is used for the WMT news translation task. This method of DA was compared against the use of human-targeted metrics such as HTER and HBLEU [3]. Annotators were asked to apply the minimum number of edits to a candidate translation to make it have the same meaning as the

reference segment and be grammatically correct. These post-edits were then processed with various human targeted metrics and subsequently compared to human judgements calculated with DA. The results highlighted that there was a relatively strong degree of divergence between system rankings when applying HTER or DA. For example, the system that ranked 7th with DA, came out with a rank of 2nd place when evaluated with HTER. Although HTER achieved high levels of correlation with DA, this level of disparity between system rankings is concerning. As such, DA human assessment was recommended for future evaluation of quality estimation. It is worth noting that these experiments were carried out at the segment-level and not the document-level which raises the question as to what would happen in a document-level setup.

The paper includes a data visualisation of human scores vs HTER scores, see figure 2.3. This visualisation shows that some segments scored a "perfect" HTER score of 0 but spanned a wide range of different human adequacy levels. A "perfect" score for HTER is derived when no edits are made to a candidate translation. Therefore, it is likely that provided more text, e.g. a full document, that at least one or more edits would be made, resulting in less "perfect" scores. Ultimately, it would be of interest running a similar series of experiments but at the document-level.

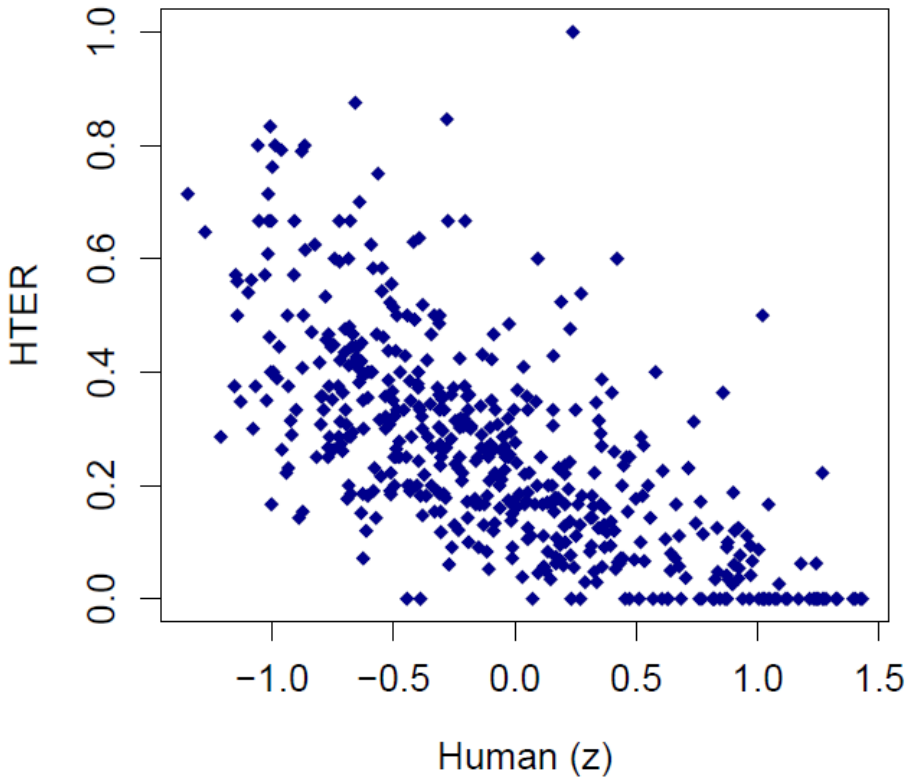


Figure 2.3: Correlation of HTER with DA. Figure taken from [3]

Scarton et al. [22] researched the difference in post-editing at the sentence-level (denoted

PE1) versus the document-level (denoted PE2). The correlation between PE1 and PE2 in terms of HTER varied from -0.14 and 0.39. The negative correlation was indicative of disagreements between PE1 and PE2, i.e. annotators disagreed with each other when post-editing at the sentence-level versus post-editing at the document-level. Therefore, this lack of agreement demonstrates that the work of Graham et al. [3] which was carried out at the segment-level, could yield different result if done at the document-level. In addition to the recent work identifying a need for document-level evaluation [8, 9], it was concluded in this paper that the "results showed that several issues could only be solved with paragraph[document]-wide context" [22].

## 2.4 Relative Ranking (RR)

A relative ranking (RR) method had been popular for evaluating the news translation task up to WMT 16 [23]. A RR HIT consisted of a source sentence and a human reference translation alongside 5 anonymised candidate translations. The goal of the annotator was to rank the candidate translations from 1 to 5, see figure 2.4. From these rankings, pairwise

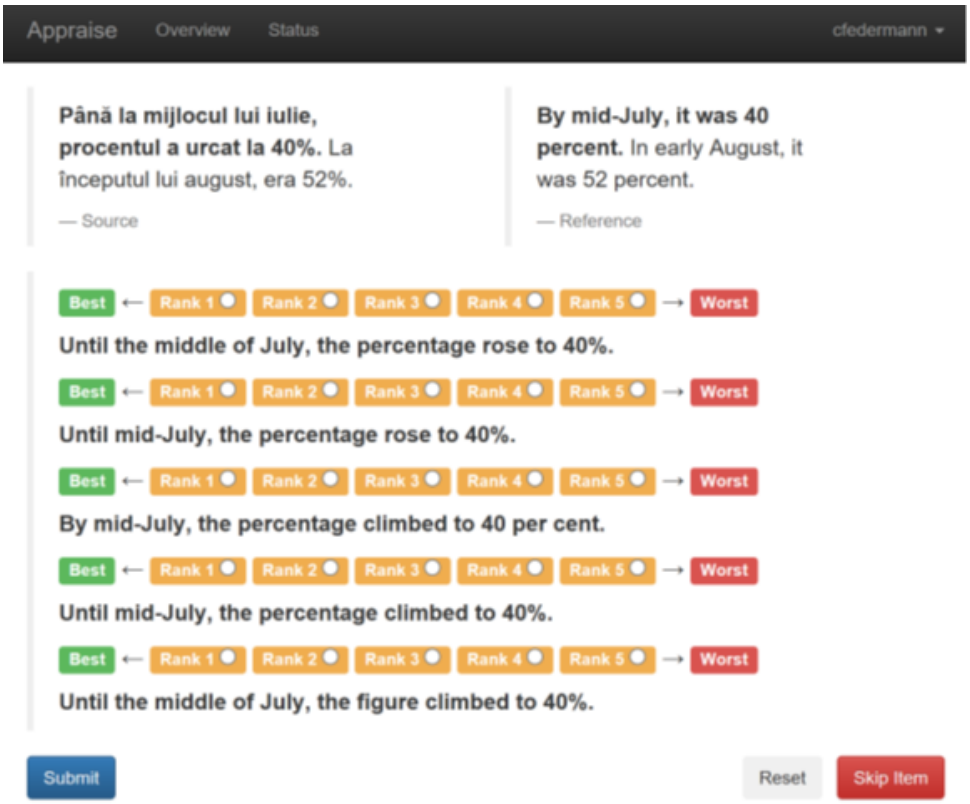


Figure 2.4: Relative Ranking - Figure taken from WMT 16 [4]

translation comparisons were produced, and the TrueSkill algorithm applied to produce system rankings [23]. It was discussed at WMT 17 [23] that "RR does not provide any information about the absolute quality of system translations". In contrast, DA can provide

this level of absolute translation quality through its use of the 0-100 point Likert scale. Therefore, these benefits of DA led to DA replacing RR as the official evaluation method for the news translation task since WMT 17 [23].

The WMT 20 Biomedical Translation Shared Task [24] used a form of RR to carry out its evaluation campaign. Evaluation was rendered at the sentence-level and abstract-level (similar to document-level). The bilingual annotators were provided with a source input alongside two candidate translations. They could then choose which was 'better' or if they were of similar quality. After the ratings were collected, a points-based system was applied, allocating 3 points for a superior translation, 1 point for a draw and 0 points for an inferior translation. Furthermore, these point-based rules were determined by the organisers and therefore one might argue that a bias exists here. In conclusion, performing RR evaluation at the document-level would likely prove problematic due to the amount of text an annotator would have to read and assess.

## 2.5 Other Document-level Methods

Due to the appeal for document-level evaluation, different methodologies have since been employed, including the work of Castilho [10]. In this research, 5 professional English to Brazilian Portuguese translators evaluated translations in terms of fluency, adequacy and error mark-up using the PET tool [25]. They also carried out further evaluations using pairwise ranking on a Google spreadsheet. For the error mark-up evaluation, translators could choose between four error categories: Mistranslation, Untranslated, Word Form, and Word order. The PET tool did not allow word-level tagging and so each error category could only be assigned once.

The experiments undertaken were only done for one language pair with just 5 translators, a relatively small sample size. These bilingual annotators were required to perform error annotation, assessing coherence types of errors. Therefore, the availability of potential annotators is limited, thus questioning the feasibility of this evaluation method. Moreover, a number of other evaluation methodologies were trialled as part this research. These experiments gave insights into how future document-level evaluations should be run. It was discovered that annotators found assessing an entire document at once tiring and they would rather evaluate single sentences. Furthermore, breaking a document into sentences/segments while also providing the full document context, could be a viable approach to assessing MT, similar to the out of English setup used at WMT 20 [2]. The paper concluded that human-evaluation of MT in document-level setups is in its infancy and it is essential to test more methodologies, while also pointing out that there is a "lack of [a] proper tool able to handle different MT evaluation methodologies."

A document-level evaluation was explored by Rysová et al. [26] where bilingual annotators were provided with the source input on the left hand side and the candidate translation on the right. The evaluations were carried out for the English into Czech language pair by trained linguists. The annotators were asked whether the "given expression/phrase in the source fulfills the function of a connective". If they answered yes, then they were posed a further questionnaire asking if the translation was:

- adequate and correctly placed
- adequate but incorrectly placed
- omitted and it does not harm the output
- omitted and it harms the output
- not adequate

This further questionnaire could only be filled out by trained linguists familiar with this level of linguistic detail. Thus, questioning the feasibility of the MT evaluation method due to the pool of available annotators. Overall, these document-level approaches [10, 26] have since not been used as official evaluation methods at WMT and there still exists an appeal for an improved MT evaluation method.

## 2.6 Summary

MT evaluation is a subjective process that has evolved over the years. The shift of the MT community towards document-level evaluation has raised question about conclusions drawn from previous sentence/segment-level evaluations and experiments. It has been noted that document-level evaluation is still in its "infancy, and therefore, it is essential to test which methodologies will be best suited for different tasks and domains" [10]. Assessing documents as a whole and providing one score across the whole document has proven problematic due to smaller sample sizes [11] and annotators finding it more tiring than assessing single sentences [10]. Therefore a setup giving the entire document broken down into segments or sentences has been popular, most notably SR+DC [2, 11].

The into English setup for SR+DC however has limited functionality, with no ability for annotators to go back and edit scores. This setup could provide the entire document with a segment-level breakdown per screen but then this would create problems with the QC checks like repeat segments. A further issue with SR+DC is that it does not specify exactly where MT systems go wrong, it merely states the rank of the MT system for that language pair.

In terms of monolingual evaluation (reference-based) and bilingual evaluation (source-based),

monolingual evaluation has a larger pool of potential annotators over bilingual evaluation. This makes monolingual evaluation cheap, efficient and quick to perform on crowd-sourcing platforms [1]. However, the literary skills of monolingual annotators may be weaker than that of bilingual annotators. The use of crowd-sourcing also introduces the problem of unreliable and careless evaluations getting submitted from annotators who seek to maximise profit, known as "gaming". To combat this, strict quality control checks have to be put in place which effectively makes some of the data collected redundant.

Post-editing MT evaluation has demonstrated high correlation with human judgement [16] but the divergence in system rankings demonstrated by Graham et al. [3] is concerning. As previously stated, both of these conclusions were based off sentence/segment-level experiments and so it would be of interest to rerun similar experiments at the document-level.

These are just some of the considerations one has to make when coming up with an MT evaluation method. These considerations demonstrate the importance of finding a balance with necessary trade-offs in terms of cost, time and reliability. For example, whether to use monolingual or bilingual annotators. It is these considerations that make MT evaluation a challenging and complex process.

# 3 Design & Methodology

This section discusses the MT evaluation approach chosen, its implementation and the experiments that were intended to be run.

## 3.1 Project Methodology

After considering the requirements for an improved manual MT evaluation system, it was evident that the proposed method would need to be at the document-level and reliable. In light of these requirements, it was decided to explore the use of post-editing with crowd-sourced monolingual annotators at the document-level. The setup is to display an entire document broken down into segments for evaluation. Similar setups have been used before and they have proven successful [2]. In terms of post-editing, annotators will be provided with the reference translation, candidate translation and a box to edit the candidate translation. Human-targeted metrics can then be applied to these resulting post-edits.

This method gives the full document context while also providing an ability to edit previous segments, something which the current into-English WMT evaluation setup doesn't offer [2]. A further benefit to this approach is that these post-edits can specify exactly where MT systems go wrong in the eyes of human annotators. This level of detail is not currently available with the method of DA.

The method of DA has demonstrated the feasibility of running evaluation campaigns on crowd-sourcing platforms [1]. They are reliable, fast and one can obtain a larger sample of results. These evaluations were therefore collected from Amazon Mechanical Turk (AMT), a crowd-sourcing platform. However, appropriate quality control mechanisms had to be implemented to avoid people 'gaming' the system, something which has proven challenging in previous evaluation campaigns [1].

The fact that the work undertaken involved paid human participants and the storing of data meant that ethics approval was needed. The project was covered under the broader ethics application submitted by Prof. Graham entitled "Evaluating Natural Language Processing Systems".

## 3.2 Web App

### 3.2.1 User Interface (UI)

The user interface (UI) for this setup would need to display the entire document, broken down as segments, showing the reference translations and candidate translations. Furthermore, there needed to be a text box to perform edits on the candidate translation while also keeping the original candidate translation available. For example, if an annotator deleted all the text in the candidate translation, they would still have the original available.

Adobe XD was used to design the user interface for this evaluation setup. Any user instruction on the UI had to be short and clear so that the average AMT worker would be able to read and understand it. Colours were used to highlight boxes and text to aid the user experience. After consultation with Prof. Graham, a final UI design which achieved these specifications was chosen, see figure 3.1.

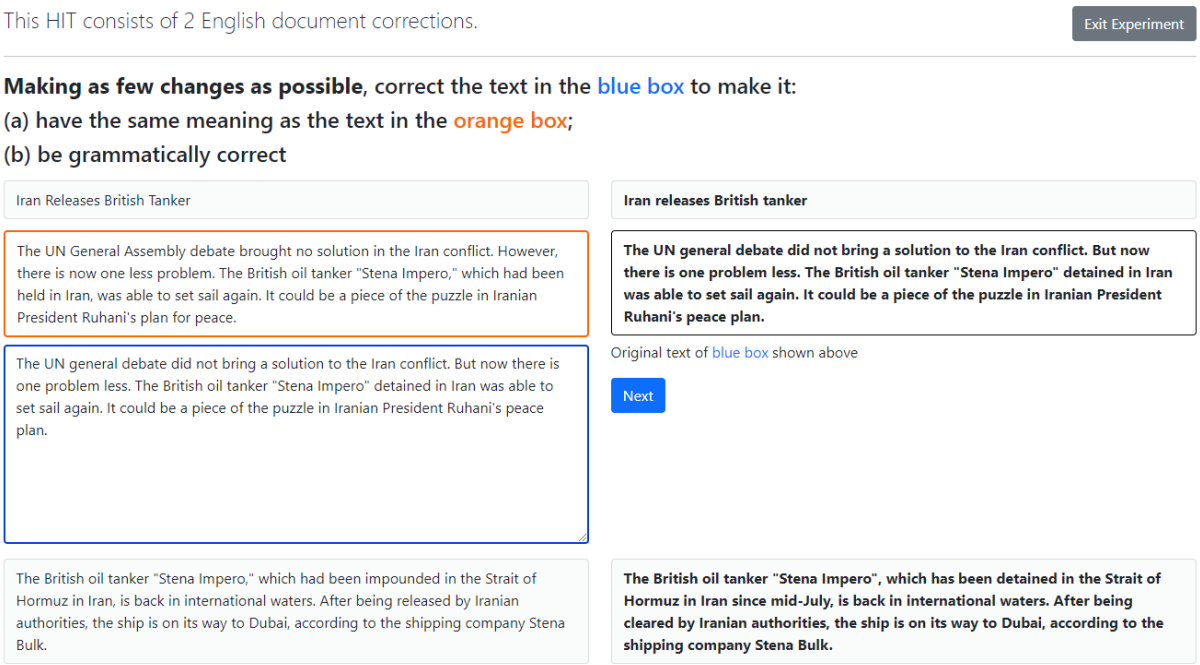


Figure 3.1: Screen shot of the User Interface (UI) for a Human Intelligence Task (HIT)

### 3.2.2 A Document Post-editing Application (DPEA)

Since the appeal for a document-level evaluation of MT, different attempts have been made to address this problem. It had been noted that as the area is relatively new and that there are not many tools readily available to perform document-level evaluation [10]. The Appraise Evaluation Framework [27] supports document-level DA but does not support

document-level post editing. The 'PET: Post-Editing Tool' evidently supports post-editing but the tool has not been maintained with no changes since 2015. It has been documented that although the tool is helpful, it lacks in functionality [10]. The MT community is thus lacking in modern document-level post-editing software. Therefore, as part of this research, a MT evaluation web application has been developed and will subsequently be made open-source. The requirements of such a web application included:

- Document-level post-editing interface.
- User authentication system.
- Ability to store results.
- Easy to deploy the application.
- Upload datasets for evaluating.

This web application will hereby be referred to as the 'document post-editing application' (DPEA).

A number of web frameworks were considered to build the DPEA but it was the Django Web Framework [28] that stuck out due to its security benefits, rapid-development and its ability to sync with SQL databases. A simplified diagram of the relational database is available in figure 3.2. The 'Dataset' model contains a file with all the HITs inside of it. This file can then be parsed by the 'Human Intelligent Task' model which is passed to the frontend for rendering to the end user. The post-edited documents store the content of the post-edit in the 'Post-edited Segment' model.

The DPEA connected to a local SQLite database when running in local development mode. For the production environment, the DPEA connected to an Amazon RDS MySQL database. This isolation of databases per environment meant that work could be carried out in development mode while not affecting the production database. To minimise the cost of the application, the Amazon Web Services free-tier allowed for this MySQL database to be used free of charge.

When the Django application was ready to be deployed, the cloud architecture had to be designed. A more traditional approach would involve the use of a Linux virtual machine, something like an EC2 instance. This approach would require an Apache Web Server be setup with an SSL certificate, a firewall be configured and a bunch of dependencies be installed. Carrying out an upgrade or update of the application could also prove disruptive with this approach. Therefore, after researching other methods of deployments, it was decided to make use of Microsoft Azure's App Services.

App Services takes care of firewalls and SSL certificates alongside preparing dependencies in the form of a docker container. One of the key advantages to App Services is that it supports

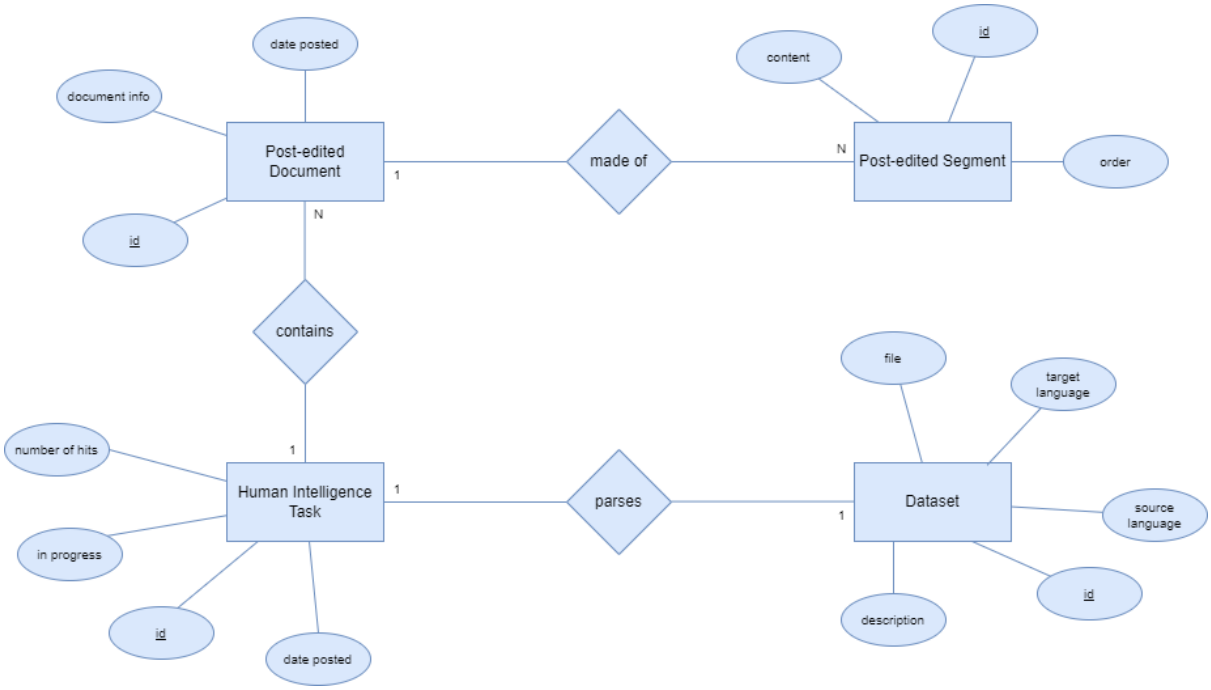


Figure 3.2: Relational database diagram for DPEA.

GitHub Actions (CI/CD) out of the box. For example, when a commit is pushed to the 'main' branch of the DPEA on GitHub, the application is redeployed to App Services.

Constants and secrets were heavily used during deployments which makes this application relatively simple for someone else to deploy. The WMT dataset files were stored on Azure Blob storage which is similar to Amazon S3. A diagram of the cloud architecture is available in figure 3.3. The DPEA also supported user authentication, contained an admin panel and could return results in a single csv using a temporary file.

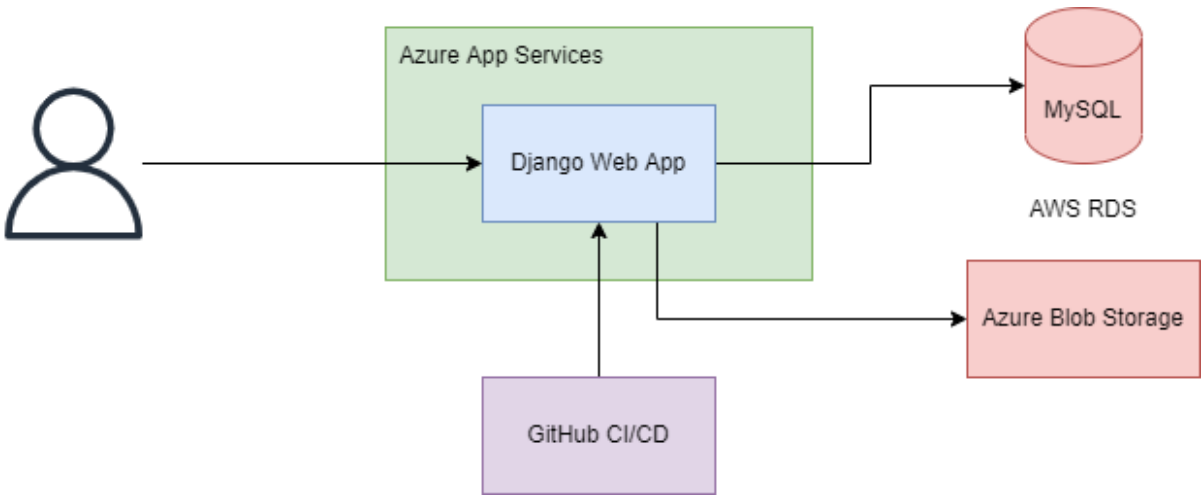


Figure 3.3: Cloud architecture of DPEA.

### 3.2.3 Security Considerations

The Django Web Framework that was used to develop this web application offered "security protection out of the box" [29] including:

- Cross site scripting protection
- Cross site request forgery protection
- SQL injection protection
- Clickjacking protection
- SSL/HTTPS
- Host header validation

This level of security ultimately protects the website from attackers, mitigating the risk of data breaches. Conversely, the documentation notes that the developer should have an awareness about security and properly deploying to a web server. Accordingly, the website runs only with encrypted traffic on port 443 (TLS) and users who try to access port 80 are redirected to 443. In addition, the MySQL database needed to be protected and so the credentials of the database were accessed through secret keys and not hard-coded. As previously stated, this use of secrets and constants also enables another developer to a deploy this DPEA out of the box given they have keys for their own MySQL database, Azure Blob Storage and Azure App Services.

Two user roles were setup for accessing the website, a super user role with unlimited access and a role with post-editing evaluation functionality only. This prohibited the standard users from accessing material like the administration page. The only globally accessible web-page that did not require authentication was the 'login page'. In fact, there was no registration page and therefore one could be very selective about what accounts were created on the website.

The passwords for the accounts were hashed into the database using PBKDF2 [30] (and a SHA256 hash) which is essentially a key derivation function used to reduce brute force due to the computational costs. To ensure passwords were somewhat complex, password constraints were put in place including: passwords must be at least 7 characters, passwords must be entirely numeric and passwords cannot be commonly used. Password hashing and password complexity constraints thus reduce the risk of accounts being compromised. Overall, this implementation of the DPEA lowered the risk of an attack.

### 3.2.4 Amazon Mechanical Turk (AMT) Frontend

The UI provided on AMT was identical to that of the one provided on the DPEA, see figure 3.1. The UI was developed in html and the HITs were rendered with jQuery [31]. The CSS styling for the UI was done through Bootstrap v5.0 [32]. Further design requirements included an information sheet, informed consent form and participant debriefing sheet. These sheets were displayed as Bootstrap modals. It was considered whether to use AMT to host the single html web page or to host something like a Node server that would interact with the AMT external API. A single html file has been popular in the past for AMT MT Evaluation campaigns [1, 2] as it was easy to configure. As such, this was perceived as the most appropriate choice.

## 3.3 Preparing Human Intelligence Tasks (HITs)

HITs were prepared in the form of a CSV file that was compatible with AMT. HIT preparation code was provided by Prof. Graham in Python 2. However, the code had to refactored for the use case in question where full documents were made available, broken down into segments. It was also necessary to upgrade this code to Python 3, to create quality control scripts and to make use of configuration files. To safely make the code compatible with Python 3, the python 2to3 [33] program was used. As well as refactoring existing scripts, other modules were created such as for creating the quality control documents and compiling the HITs.

As previously stated, it was important to employ configuration files for constructing these HITs. This made the process of performing an MT evaluation campaign highly customizable. For example, in this configuration file, one could set the dataset, source language, target language and the number of words for a HIT, see listing 3.1. The quality control dictionary of the configuration file enables the ‘common_words_file‘ and ‘alphabet_file‘ to change based on the language. This enables the preparation scripts to support other languages as previously they had only been used for into English evaluation campaigns. Other features exist in the configuration files but have been removed from listing 3.1 for clarity. It is also possible to make a bilingual (source-based) evaluation campaign by setting the key ‘bilingual‘ to a value of ‘true‘.

Listing 3.1: JSON configuration file example

```
{
    'data': 'newstest2020',
    'src': 'de',
    'trg': 'en',
    'max_words_per_hit': 1000,
```

```
    'qc': {
        'common_words_file': 'common-words-en.json',
        'alphabet_file': 'alphabet-en.json',
        'num_words_per_qc_item': 20,
        'min_word_count': 400,
        'max_word_count': 1000
    }
}
```

When compiling the HITs for AMT, a CSV file was used where each line represented a separate HIT. As each HIT had differing numbers of documents and segments, it was not ideal to use conventional CSV columns as each HIT would have a different number of columns. Therefore the contents of a HIT were embedded into a single CSV cell, with segments marked for separation by '||' and documents separated by '__'. It would have been more suitable if AMT supported HITs formatted in JSON but this approach used was sufficient.

## 3.4 Processing Human Intelligence Tasks (HITs)

The responses of the HITs needed to be processed to extract the necessary data, apply quality control checks and generate results. Python scripts were written that similarly supported the use of configuration files. Features to the configuration file of HIT processing included:

- Minimum precision to pass quality control.
- Minimum recall to pass quality control.
- Minimum time spent to pass quality control.

### 3.4.1 Extracting the Edits

The post edits were extracted from the CSV, converted into a JSON segment-level breakdown and combined with the unedited original system output. The next stage to this processing required the exact edits to be extracted from the text. For example, the words inserted or the words deleted from the candidate translation. Initially, the edits made to the document were extracted using the in-built python module difflib [34]. This module has a sequence matcher that was derived from the 'Gestalt Pattern Matching' [35] algorithm. The algorithm takes 2 strings as inputs ($S_1, S_2$) and finds their common substrings. $K_m$ is the number of matching characters from these substrings. As such, a similarity ratio can be

computed and is denoted in equation 1, where $0 <= D_{ro} <= 1$.

$$D_{ro} = \frac{2K_m}{|S_1| + |S_2|} \tag{1}$$

This algorithm provides the basis for the difflib module which can also determine what substrings have been inserted, removed or replaced from $S_1$. A table could thus be constructed of the number and type of edits applied to a document.

When initially evaluating the quality control results, it was observed that there was a higher than expected number of false positives. Accordingly, the sequence matching algorithm was looked at closely and what one might call 'undesirable features' were discovered. The underlying reason for this derived from the fact that the module used 'replacements'. These 'replacements' were then interpreted as $S_1$ deletions and $S_2$ insertions in the HIT processing code. For example, if 'red' was replaced with 'black', then that would produce 3 character deletions and 5 character insertions. To demonstrate the undesirable behaviour of the difflib module, an example can be seen below where 'replacements' are highlighted in yellow.

(Original) The US government wishes appointment to concentrate on two...

(Post-edit) The US government wishes to concentrate on two...

Evidently, the fact that 'to concentrate' was included as part of the replacement was unnecessary. The research undertaken was focused on the minimum number of edits between 2 segments and the behaviour of the difflib module thus came under question. To uncover why the algorithm behaved this way would require a deep dive into the difflib module, something which was beyond the scope of this project but nonetheless, would be worth exploring in future work. To overcome this challenge, an alternate sequence matcher was used known as diff-match-patch [36], developed by Google. This library implements the Myer's diff algorithm [37] which proved more promising than difflib. The same extract of text showed one word deletion when using this sequence matcher. The deletion is highlighted in yellow in the extract below.

(Original) The US government wishes appointment to concentrate on two...

(Post-edit) The US government wishes to concentrate on two...

As it was of prime importance that this new sequence matcher behaved as expected, unit testing was implemented. The in-built unittest [38] python framework was used for a variety of test cases on how the system should measure true positives and false positives for a document.

### 3.4.2 Applying Metrics

Human-targeted metrics were calculated using sacreBLEU [39] which supported the calculation of BLEU and TER scores. If one was to assume that the post-edit of a document was a 'perfect' translation of the source input, then the word-level matches between the system output and the post-edit can be denoted as the true positives, i.e. the post-edit and machine output word-level matches. Moreover, these true positives provided insight into classical recall and precision utilising the formulas in (2). As previously stated, unit-tests were put in place to ensure the reliability of the computation of true positives.

$$Recall = \frac{TP}{\#\ words\ in\ system\ output}, \ Precision = \frac{TP}{\#\ words\ in\ post\text{-}edit} \tag{2}$$

Additional experiments look at the most commonly inserted/removed words from a MT system, the position of the edits in a segment and the types of QC error corrections.

## 3.5 Configuring an Evaluation Campaign

The WMT 2020 news task dataset was chosen to carry out evaluations for English to German. A summary of the evaluation dataset is available in table 3.1.

| MT Evaluation Dataset | | | |
|---|---|---|---|
| **Language Pair** | **Reference Words** | **Reference Documents** | **Systems** |
| German to English | 33781 | 118 | 13 |

Table 3.1: Language pair datasets taken from WMT 20 [2]

To effectively plan the evaluation campaign, it was necessary to evaluate an equivalent number of assessments to that of WMT 2020. For the WMT 2020 evaluation campaign, it worked out there were roughly 10,000 assessments collected per language pair, where a segment rating is counted as one assessment. In this case, every word in the document can be regarded as an assessment as post-editing gives back more data points than a single score. Furthermore, it works out that 35 documents would need to be evaluated per system.

The next challenge was to determine the number of documents to put inside a single HIT. For the MT evaluation method DA, typically 30% of the data in a HIT is for quality control purposes. This method uses 100 segments in a single HIT, where the quality control checks can be randomly placed throughout. In this project, it was not possible to randomly place quality control segments because full documents were evaluated and it was important to maintain the entire document context. As AMT tends to suffer from people 'gaming' the system, it was decided to keep the duration of a HIT short, consisting of 1 quality control document and 1 system output document (denoted as REAL). This would result in roughly

50% of the data collected being QC documents. However, this ultimately gave more confidence in the reliability of the REAL results collected. A summary of the target data collection is available in table 3.2. Furthermore, AMT workers were paid minimum wage.

| Language Pair | WMT Assessments | Target Assessments | Target Documents | HITs |
|---|---|---|---|---|
| German to English | 14,303 | 10,000 | 455 | 455 |

Table 3.2: Target data collection

### 3.5.1 Inter-annotator Agreement Experiments

It is well documented that achieving high levels of inter-annotator agreement has been challenging at WMT [1]. The DA method relies on humans assigning "a subjective numerical judgement to a translation" [16] where annotators can struggle on what factors of a translation are considered "good" or "bad" [21]. Thus, a possible reason for lower levels of inter-annotator agreement. Furthermore, human annotators do not apply a numerical judgement in the post-editing setup proposed. As such, it would be of interest in performing an inter-annotator agreement experiment for this post-editing setup.

Repeat HITs can be posted on AMT and the edits made to the REAL documents can be analysed. Measuring the inter-annotator agreement can be calculated through Cohen's kappa [40], see equation 3. However, quantifying an agreement ($P(A)$) from the post-edits appears to be a mostly unexplored area.

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{3}$$

### 3.5.2 Summary

The MT evaluation setup proposed comprises of monolingual evaluations posted on AMT where assessment is carried out through post-editing. Most notably, this setup is done at the document-level which addresses the shortcomings in current MT evaluation techniques like SR+DC for into-English at WMT. Moreover, document-level evaluation is still in its "infancy" [10], and so currently available frameworks and tools like Appraise [27] and PET [25] do not support the evaluation method proposed. For this reason, an open-source tool (DPEA) has been developed to address this setup and so valuations will also be sourced from researchers on the DPEA.

The project additionally aims to assess inter-annotator agreement at the document-level for post-editing. Human-targeted metrics and parameters like recall can be calculated from this

post-editing setup. It will be closely assessed how these values correlate with the official results from WMT. Overall, the advantages to this setup are listed below.

- Document-level evaluation
- Specifies exactly where MT systems go wrong
- Cheap and fast due to the use of AMT, a crowd-sourcing platform
- Language-independent
- Backwards-compatible - for document-level datasets
- Removes subjective numerical judgements provided by annotators

# 4 Results & Evaluation

Annotations were sourced from researchers who used the DPEA website and by workers on AMT. German-English reference-based (monolingual) evaluation campaigns were performed on both the DPEA and on AMT. A source-based (bilingual) evaluation was also carried out for English-Chinese on the DPEA In addition, inter-annotator agreement (IAA) experiments were performed on both AMT and the DPEA.

## 4.1 Quality Controlling the Crowd

### 4.1.1 Why is Quality Control Needed?

Collecting data from crowd-sourcing platforms comes with the risk of annotators submitting low quality data. To this end, the MT community has often employed QC mechanisms to mitigate this concern [2, 11]. An underlying problem to the submission of low quality is from people 'gaming' the system. As previously outlined, this is where AMT workers try to complete as many HITs as possible in order to maximise profit. In light of this issue, QC mechanisms were configured for this method of MT evaluation to ensure the reliability of results.

### 4.1.2 Quality Control Set-up

To maintain full document-level context, it was decided to dedicate full documents to QC and not individual segments. As such, QC documents were generated from the reference documents. This generation of QC documents involved the insertion of words and letters into the reference documents. To make the setup language independent, words were inserted from a specified 'common_words' file and letters were added from a given 'alphabet' file. In addition, random words and random letters (from the specified 'alphabet' file) were deleted from the documents. These modifications can be summarised as follows.

- Word insertions
- Letter insertions

- Word deletions
- Letter deletions

Moreover, these modifications were recorded in a log file and compared against the submitted post-edited documents from AMT. It was important to have a variety of different quality control checks to make it difficult for someone to 'game' the system.

As these quality control documents were generated from the reference documents, an annotator could pass all the quality control checks if they were to just copy the reference document. Therefore, the ability to copy text from the reference translation was disabled in the UI. However, it was possible for annotators to still copy this text by viewing the raw HTML of the web page. Previous MT evaluation campaigns used images of translations to mitigate such a concern [11] but in this case, where the annotator would have access to a text box of the candidate translation, it was desirable to keep everything the same size and font. If images were displayed for the candidate and reference translations, then the scale of the text in the text box and the images would differ. This would prove problematic as it would make the evaluation more difficult to perform. In addition, spell checking was disabled from the HTML text boxes as it generally made QC edits easier to spot.

### 4.1.3 Processing Quality Control Data

As previously stated, there were 4 different types of QC items that were assessed including word insertion, letter insertion, word deletion and letter deletion. The QC documents contained these items and it was up to the annotator to successfully spot the errors and correct them. A true positive was recorded if any one of these errors was rectified. Otherwise, a false positive was recorded for every word or isolated substring that was edited in the document. To clearly demonstrate the behaviour of the quality control checking code, a sentence-level example is available below.

(Reference) A black cat sat on the wall in London.

(QC generated) A bzlack cat sat on airplane the wall in London.

(Post-edit) A black cat sat on the big wall in London.

The QC generated sentence contains a letter insertion of 'z' in the work 'black' and it contains a word insertion of 'airplane'. The post-edit successfully spotted and corrected the mistakes which would be the equivalent of 2 true positives. However, the post-edit also adds the word 'big' before 'wall' and so this would result in a false positive. These parameters facilitated the calculation of specific recall and precision scores, denoted QC recall and QC precision respectively. Note that QC recall and QC precision are different to classical recall and precision. For the case of quality control, these were given by the following

Table 4.1: QC document results from DPEA. 'Avg. Real Similarity' represents the average similarity of the real documents to the reference document, where 1 is indicative of an identical match to the reference document.

| Hit | True Positives | False Positives | QC Recall | QC Precision | F-score | Avg. Real Similarity |
|---|---|---|---|---|---|---|
| 0 | 4 | 0 | 80.00 | 100.00 | 88.89 | 0.64 |
| 1 | 15 | 0 | 88.24 | 100.00 | 93.62 | 0.55 |
| 2 | 8 | 1 | 80.00 | 88.89 | 84.26 | 0.71 |
| 3 | 15 | 0 | 100.00 | 100.00 | 100.00 | 0.64 |
| 4 | 8 | 1 | 72.73 | 88.89 | 80.21 | 0.45 |
| 5 | 14 | 0 | 87.50 | 100.00 | 93.62 | 0.61 |
| 6 | 20 | 0 | 95.24 | 100.00 | 97.44 | 0.59 |
| 7 | 5 | 0 | 83.33 | 100.00 | 90.71 | 0.45 |
| 8 | 21 | 1 | 87.50 | 95.45 | 91.37 | 0.44 |
| 9 | 6 | 0 | 85.71 | 100.00 | 92.47 | 0.50 |
| 10 | 5 | 0 | 100.00 | 100.00 | 100.00 | 0.57 |
| 11 | 14 | 0 | 93.33 | 100.00 | 96.37 | 0.69 |
| 12 | 11 | 1 | 100.00 | 91.67 | 95.83 | 0.73 |
| 13 | 9 | 0 | 90.00 | 100.00 | 94.74 | 0.62 |
| 14 | 17 | 0 | 100.00 | 100.00 | 100.00 | 0.52 |
| 15 | 8 | 0 | 88.89 | 100.00 | 94.18 | 0.58 |
| 16 | 11 | 0 | 91.67 | 100.00 | 95.83 | 0.52 |
| 17 | 2 | 0 | 100.00 | 100.00 | 100.00 | 1.00 |
| 18 | 11 | 7 | 84.62 | 61.11 | 71.03 | 0.62 |

formulas:

$$QC\ Recall = \frac{TP}{\#\ of\ QC\ items},\ QC\ Precision = \frac{TP}{TP + FP} \tag{1}$$

QC recall can be summarised as the number of corrected mistakes (TP) over the total number of mistakes inserted to the document. On the other hand, QC precision was the number of corrected mistakes (TP) over the total number of edits made by the worker.

### 4.1.4 Experiments

#### Evaluation by Researchers

The deployed DPEA allowed for an initial series of HITs to be posted in order to test the setup. Each HIT posted to the DPEA included real (machine-output) documents along with 1 QC document. A key factor to remember is that evaluations on the DPEA were annotated by researchers. It is also worth noting that at that stage of development, the QC document only included word insertions and letter insertions. The German-English WMT 20 news dataset was used to perform the monolingual (reference-based) evaluation by researchers.

Table 4.1 shows the results of QC from the DPEA where true positives are indicative of an expected correction like removing an incorrectly inserted word. False positives in this case

represent words or isolated substrings which were "incorrectly" deleted or inserted. False positives were not necessarily "incorrect" but they were not expected as these edits did not exist in the reference document. In light of table 4.1, it was observed that QC recall tended to be around 80% or higher. However, QC precision was 100% for 14 of the 19 QC documents. Despite these observations, the sample size of 19 documents was relatively small and so drawing conclusions at this stage would not be ideal. In one case, there were only 2 QC items present (Hit 17) in a document which could be difficult for an annotator to spot or over estimate the annotator's abilities if they spotted both items. Therefore, the QC for AMT was designed so that there would be more QC items present and the QC documents were sufficiently long. The configuration file was set so that there was a minimum of 350 words in the QC document and maximum of 800 words, producing between 20-30 QC items.

### Evaluation by AMT Workers

The first batch of HITs posted on AMT consisted of 1 QC document and 1 real document. This batch was posted to get a sense of workers' abilities and to establish parameters for approving or rejecting a HIT. A sample of the results of these HITs is available in table 4.2. It can be observed that there were a number of cases where annotators made no edits to the QC documents, producing 0 true positives (TP) and 0 false positives (FP). Furthermore, one might not expect these low quality submissions by looking at the time that had been spent on the task, in one case 54.45 minutes. In conclusion, this demonstrated that the QC mechanisms, such as QC recall, were useful at detecting low quality submissions of data.

As the post-edits could be compared against the QC text and reference text, a manual analysis was carried out on this first batch of HITs. During this analysis, the worker would be marked as diligent (D), not diligent (ND) or unsure (U). In addition, the analysis shed light on some issues to look out for when performing QC checks.

In the AMT interface, the annotator was provided with the reference translation and candidate translation. Moreover, it was possible that the annotator would solely use the reference translation when post-editing, i.e. they would copy the reference translation for the post-edit and disregard the candidate translation. This was evidently problematic as the annotators were asked to make the "minimum number of edits". Although the selecting and copying of the reference translation text was disabled on the UI, it was still possible to manually type out the reference translation. This brought about a concern that the annotator could manually copy the reference translation for both the QC and real documents and they would pass QC. Therefore, it was important to make sure this concern of 'reference copying' was alleviated.

Table 4.2: Sample of QC document results from first batch of HITs on AMT. 'TP' = True positives, 'FP' = False positives, 'RE Similarity' = Reference document similarity to post-edit, 'RS Similarity' = Reference document similarity to system output, Status: 'ND' = Not diligent, 'D' = Diligent, 'U' = Unsure.

| Minutes | TP | FP | QC Recall | QC Precision | F-score | RE Similarity | RS Similarity | Status |
|---|---|---|---|---|---|---|---|---|
| 11.62 | 29 | 0 | 100.00 | 100.00 | 100.00 | 0.52 | 0.60 | D |
| 16.60 | 26 | 1 | 96.00 | 96.00 | 96.00 | 0.45 | 0.45 | D |
| 28.87 | 21 | 1 | 88.00 | 95.00 | 91.00 | 0.52 | 0.56 | D |
| 20.52 | 30 | 3 | 91.00 | 91.00 | 91.00 | 0.40 | 0.39 | D |
| 20.70 | 20 | 2 | 91.00 | 91.00 | 91.00 | 0.53 | 1.00 | ND |
| 18.35 | 31 | 1 | 82.00 | 97.00 | 89.00 | 0.42 | 0.44 | D |
| 29.52 | 18 | 0 | 75.00 | 100.00 | 86.00 | 0.50 | 0.96 | D |
| 41.87 | 18 | 2 | 82.00 | 90.00 | 86.00 | 0.43 | 0.62 | D |
| 14.40 | 31 | 12 | 97.00 | 72.00 | 83.00 | 0.50 | 0.62 | D |
| 24.90 | 33 | 4 | 75.00 | 89.00 | 81.00 | 0.34 | 0.36 | D |
| 12.88 | 17 | 2 | 74.00 | 89.00 | 81.00 | 0.60 | 0.61 | D |
| 33.18 | 16 | 0 | 59.00 | 100.00 | 74.00 | 0.38 | 0.38 | D |
| 5.95 | 23 | 3 | 61.00 | 88.00 | 72.00 | 0.70 | 0.70 | D |
| 28.25 | 21 | 7 | 64.00 | 75.00 | 69.00 | 0.76 | 0.78 | D |
| 18.47 | 12 | 0 | 52.00 | 100.00 | 68.00 | 0.52 | 0.57 | U |
| 9.45 | 16 | 8 | 67.00 | 67.00 | 67.00 | 0.59 | 0.57 | D |
| 15.65 | 20 | 17 | 83.00 | 54.00 | 65.00 | 0.57 | 0.65 | D |
| 24.73 | 9 | 0 | 43.00 | 100.00 | 60.00 | 0.52 | 0.52 | U |
| 50.75 | 29 | 38 | 97.00 | 43.00 | 60.00 | 0.45 | 0.93 | U |
| 39.02 | 13 | 3 | 48.00 | 81.00 | 60.00 | 0.63 | 0.66 | U |
| 10.12 | 16 | 6 | 46.00 | 73.00 | 56.00 | 0.66 | 0.69 | U |
| 17.78 | 21 | 45 | 72.00 | 32.00 | 44.00 | 0.50 | 0.58 | U |
| 19.60 | 7 | 11 | 27.00 | 39.00 | 32.00 | 0.41 | 0.41 | ND |
| 12.48 | 4 | 3 | 16.00 | 57.00 | 25.00 | 0.66 | 0.66 | ND |
| 14.65 | 3 | 0 | 13.00 | 100.00 | 23.00 | 0.29 | 0.29 | ND |
| 14.72 | 5 | 20 | 17.00 | 20.00 | 18.00 | 0.47 | 0.45 | ND |
| 7.35 | 1 | 2 | 4.00 | 33.00 | 7.00 | 0.58 | 0.58 | ND |
| 28.50 | 1 | 2 | 4.00 | 33.00 | 7.00 | 0.45 | 0.45 | ND |
| 5.55 | 1 | 2 | 4.00 | 33.00 | 7.00 | 0.08 | 0.12 | ND |
| 54.55 | 1 | 0 | 3.00 | 100.00 | 6.00 | 0.13 | 0.22 | ND |
| 1.92 | 1 | 1 | 3.00 | 50.00 | 6.00 | 0.45 | 0.45 | ND |
| 4.62 | 1 | 3 | 3.00 | 25.00 | 5.00 | 0.53 | 0.53 | ND |
| 21.33 | 1 | 3 | 3.00 | 25.00 | 5.00 | 0.83 | 0.99 | ND |
| 4.78 | 0 | 0 | 0.00 | - | - | 0.39 | 0.39 | ND |
| 54.45 | 0 | 0 | 0.00 | - | - | 0.42 | 0.42 | ND |
| 14.13 | 0 | 0 | 0.00 | - | - | 0.42 | 0.42 | ND |
| 57.15 | 0 | 1 | 0.00 | 0.00 | - | 0.45 | 0.45 | ND |
| 21.92 | 0 | 1 | 0.00 | 0.00 | - | 0.38 | 0.38 | ND |
| 49.17 | 0 | 0 | 0.00 | - | - | 0.33 | 0.33 | ND |
| 32.10 | 0 | 0 | 0.00 | - | - | 0.57 | 0.57 | ND |
| 7.07 | 0 | 0 | 0.00 | - | - | 0.55 | 0.55 | ND |
| 3.95 | 0 | 0 | 0.00 | - | - | 0.20 | 0.20 | ND |
| 1.50 | 0 | 1 | 0.00 | 0.00 | - | 0.31 | 0.30 | ND |
| 4.95 | 0 | 0 | 0.00 | - | - | 0.19 | 0.19 | ND |
| 3.67 | 0 | 0 | 0.00 | - | - | 0.09 | 0.09 | ND |

It was observed during the manual analysis of documents, in certain cases, that the annotator was in fact copying the reference translation for their post-edit. Similarity scores were thus established using difflib [34], where a similarity score of 1 denotes a perfect match. This score was computed for the reference translation similarity to the post-edit (RE Similarity) and the reference translation similarity to the system output (RS Similarity). It was found from inspection that when the RE Similarity was 0.9 or higher, that the annotator was mostly copying the reference and disregarding the system output. However, it was possible that the system output already had a high similarity score to the reference translation. Therefore, it was decided to set thresholds whereby, if the RE similarity was above 0.9 and the RS similarity was below 0.6, then the annotator was guilty of 'reference copying' and so the HIT would be rejected. Conversely, there was a particular outlier case where the system output was in fact so weak that the most viable option was to just copy the reference. As stated, this applied to one system which was referred to as 'yolo' at WMT 20 and in fact finished last place in the DA system rankings. In summary, the similarity checks were ignored for this particular MT system.

### Low QC Precision

A density plot which visualises QC recall and QC precision for the AMT HITs is available in figure 4.1. HITs which achieved 0% QC recall or QC precision were excluded from this plot as these were cases where annotator made 0 edits. The graph shows that QC precision was generally lower than QC recall.

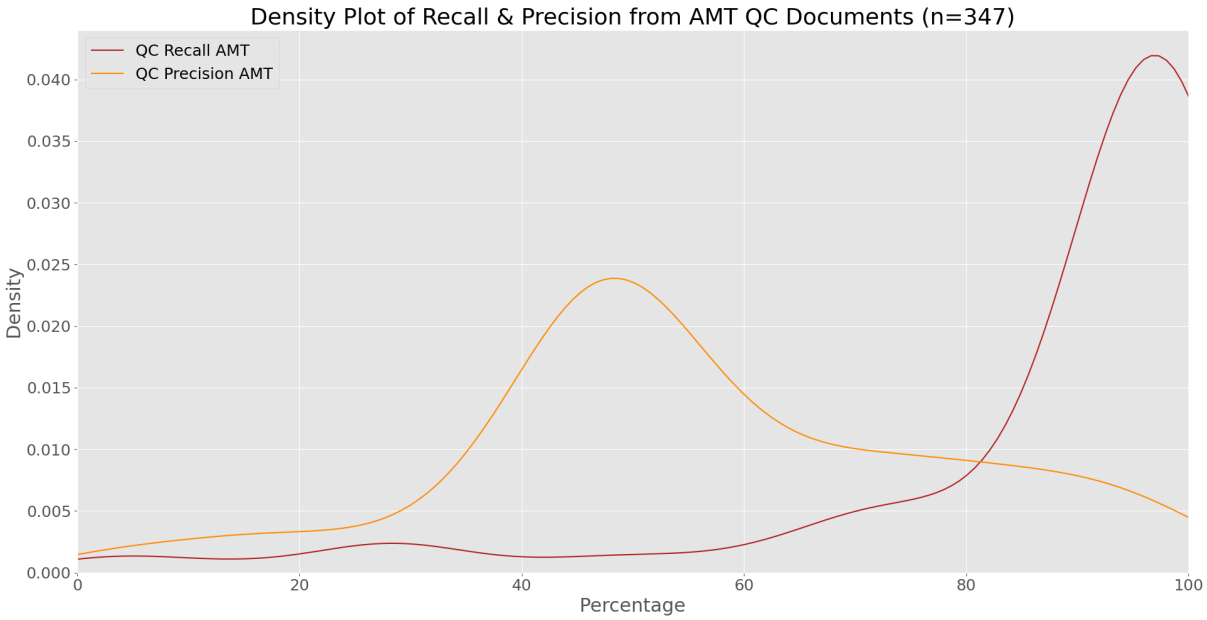


Figure 4.1: Density plot of the QC recall and QC precision from AMT. Values of 0 QC recall and QC precision omitted for clarity

Due to the distribution of QC precision, it was worth exploring classical precision on the QC

documents by comparing the post-edits to the reference translation, see figure 4.2. It bears repeating that the calculation of QC precision and precision were different. QC precision was based on the number of correctly spotted errors over the number of edits made whereas precision was calculated classically. The visualisation shows no anomalies and shows a level of correlation between precision and recall. This raised the question as to why QC precision peaked lower at 50% and had a broader distribution. It can be summarised that QC precision was based on edits made but precision was based on the entire document. Furthermore, QC precision had fewer data points and precision was initialised with pre-existing word-level matches. Therefore, these parameters do not share a much of a relationship.

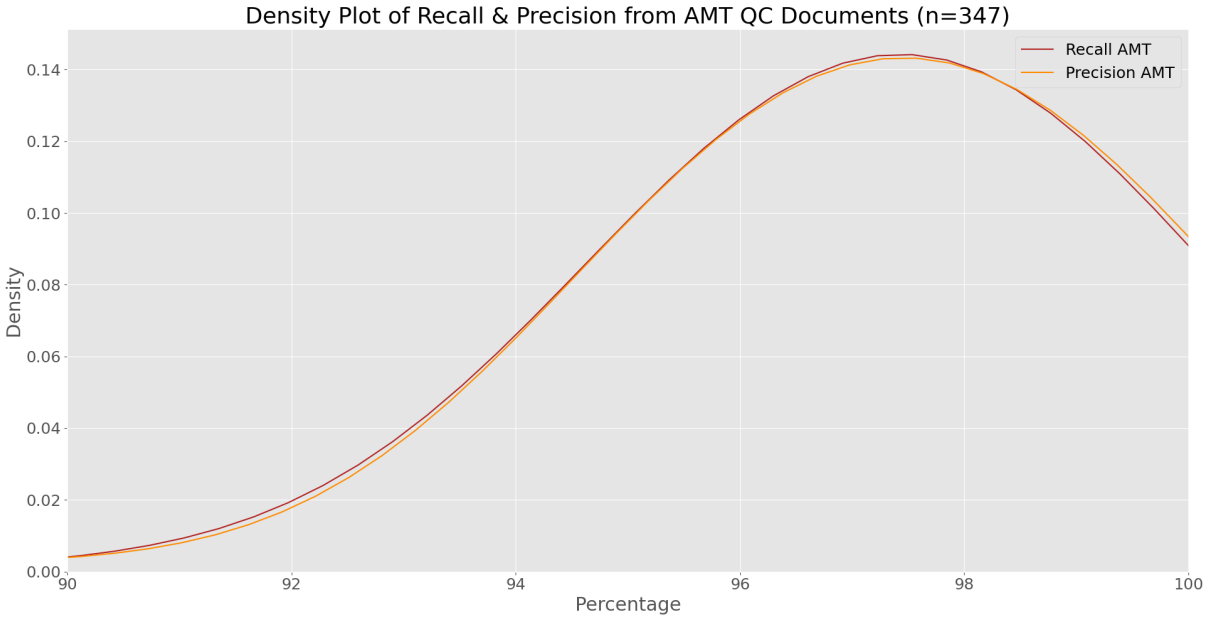


Figure 4.2: Density plot of recall and precision from AMT. Values of 0 recall and precision omitted for clarity

Nonetheless, it was still worth investigating why QC precision tended to be lower than expected and if the annotators were making unnecessary edits. Figure 4.3 shows a QC document which was edited by an annotator on AMT. On the right hand side, the reference document is available which represents an "expert translation". Throughout the post-edited document in figure 4.3, it can be seen that some excessive editing is made such as the moving of "in the future" and "primarily". The annotator also made the decision to change "United States" to "U.S.", something which again could be considered excessive. As previously stated, annotators were told to make "the minimum number of edits" and so ideally they shouldn't be making these edits. This particular example also shows some of the expected corrections that the annotator was supposed to spot like removing the letter 'f' from '2020' and removing the word 'estate' but failed to do so.

Upon studying individually submitted quality control documents, it was observed that some of the post-edits included the capitalisation of letters, the insertion of punctuation marks

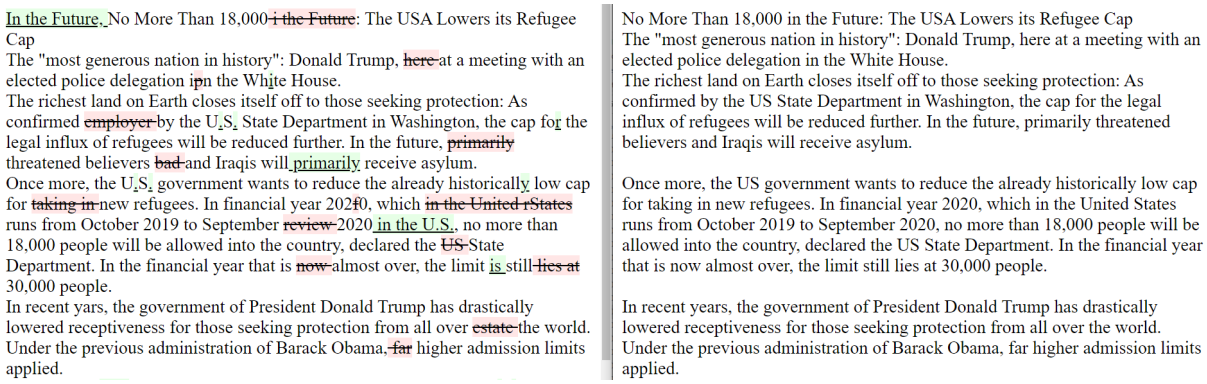


Figure 4.3: Example of QC edited document. Green highlighted text represents insertions, red highlighted text represents deletions. Left hand-side document is the post-edited document and the right hand-side is the reference document.

and leaving double spaces after removing words. These edits were not part of the quality control documents nor existed in the underlying reference document. Some of the edits were not necessarily right nor wrong but nonetheless, a form of tokenisation was implemented to avoid these anomalies. In essence, before submitted post-edits were processed, punctuation was removed, letters were set to lowercase and double spacing was replaced with single spacing.

Another example of what might consider excessive editing is available in figure 4.4. In this example, "the Greek coastguard" is changed to a "Greek Coast Guard official" and "to help overcome them" is changed to "to help in overcoming them". Conversely, the annotator inserts "German" before "Federal Minister" when referring to Horst Seehofer. In this case, it could be argued that this insertion was necessary as it provides more context. Overall, these examples shed light on why annotators can score highly on QC recall but less so on QC precision.

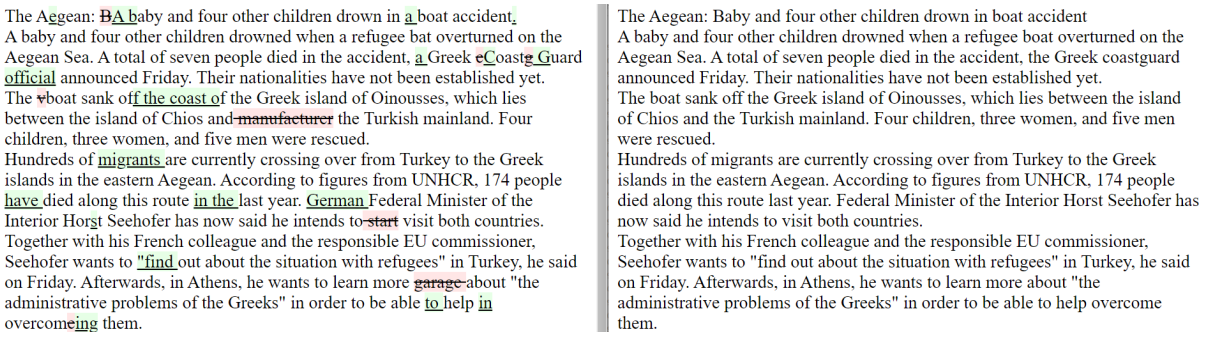


Figure 4.4: Example of QC edited document. Green highlighted text represents insertions, red highlighted text represents deletions. Left hand-side document is the post-edited document and the right hand-side is the reference document.

Figure 4.5 presents a case where an annotator added what appears to be arbitrary text to perhaps, make it seem like they were doing more work. In this example, the annotator thus

| | Word Insert | Word Delete | Char. Insert | Char. Delete |
|---|---|---|---|---|
| Passed QC | 97.68% | 92.85% | 97.47% | 94.00% |
| 0% <QC Recall <80% | 40.22% | 47.53% | 50.87% | 49.03% |

Table 4.3: Proportion of AMT annotator correcting QC item. E.g. Word Insert=2499/2558=97.68%

scored poorly on QC precision. Evidently, it was not desirable in including annotators who performed this type of editing but the QC checks were able to pick them out due to the lower QC precision.

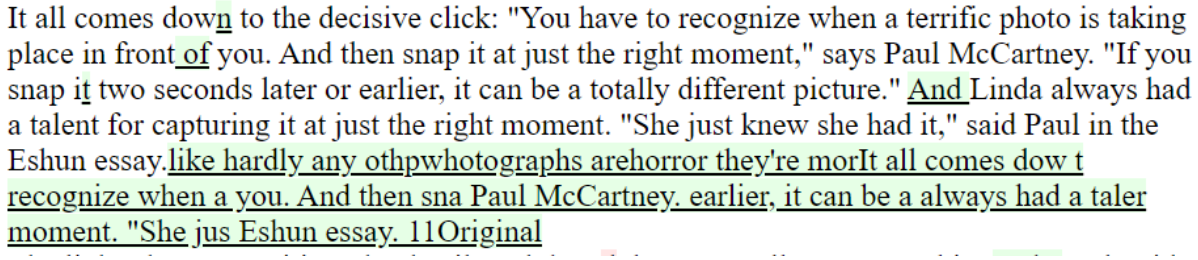
It all comes down to the decisive click: "You have to recognize when a terrific photo is taking place in front of you. And then snap it at just the right moment," says Paul McCartney. "If you snap it two seconds later or earlier, it can be a totally different picture." And Linda always had a talent for capturing it at just the right moment. "She just knew she had it," said Paul in the Eshun essay.like hardly any othpwhotographs arehorror they're morIt all comes dow t recognize when a you. And then sna Paul McCartney. earlier, it can be a always had a taler moment. "She jus Eshun essay. 11Original

Figure 4.5: Example of QC edited document. Green highlighted text represents insertions, red highlighted text represents deletions.

### Comparing QC Types

There were four different types of errors intentionally made for the QC document. Therefore, it was of interest in researching which types of QC corrections annotators were most likely to spot and edit. The total number of QC items present and the number of times each item was corrected was aggregated and is thus outlined in table 4.3. In cases where the annotator's passed QC, they were most likely to fix word insertions (97.68%) but least likely to spot word deletions (92.85%). In contrast, when examining annotators who scored 0% <QC Recall <80%, they were least likely to rectify a word insertion but most likely to correct a character insertion. A possible factor for this is that word insertions tended to be the most commonly used QC item.

### Setting the Standard

The manual analysis from table 4.2 gave an insight into what thresholds of QC recall and QC precision should be set to pass QC. The lowest recorded QC recall and QC precision that were regarded as diligent were 59.26% and 65.57% respectively. Workers were marked as 'unsure' (U) when QC recall and QC precision lay between 40% and 60%. Therefore, it seemed appropriate and fair to approve workers who achieved a minimum QC recall and QC precision of 40%. However, it was possible in final calculations to ignore those that scored lower in the 40% to 60% range if it was deemed necessary. A further condition was added to ensure workers did not complete 2 or more HITs at the same time.

The finalised parameters for approving a HIT were as follows:

- QC Recall $\geq$ 40%
- QC Precision $\geq$ 40%
- Time spent $\geq$ 4 minutes
- RE Similarity $\leq$ 0.9 if RS Similarity $\leq$ 0.6*
- HITs must not be completed simultaneously

*This did not apply for weaker MT system outputs.

It is worth noting that in the event where a worker was rejected and they had complained, their rejection was overturned and consequently told not to complete any more of these HITs. Overturned rejections were not included in the results.

### 4.1.5 Comparing Researchers and AMT Workers

A density plot of visualising QC recall and QC precision for the AMT HITs and the DPEA HITs is available in figure 4.6. This gives a comparison of performance between researchers and AMT workers. A vertical green line is plotted in addition to highlight the chosen pass rate, 40% or higher for QC recall and QC precision. It was ascertained that the QC precision for AMT had a lower peak and generally a broader distribution. The QC precision for AMT peaked at approximately 50% which was indicative of false positives, i.e. excess insertions or deletions on the QC document. Alternatively, the QC precision for the DPEA tended to be higher, near 100%. The annotators performed similarly in terms of QC recall and so it may be of interest increasing the QC recall threshold when finalising results.

## 4.2 Document-level MT Evaluation

Three evaluation campaigns were carried out, two of the evaluation campaigns were annotated by researchers using the DPEA website. The third evaluation campaign sourced more assessments as it was carried out on AMT. The sourced data is summarised in table 4.4. The target data collection for German-English was 10,000 word assessments in total but after assessing the cost and the sample size, this target was changed to 100 segment assessments per system. For the evaluations that were done on the DPEA, QC checks were not needed as the annotators were trusted researchers. However, in the case of AMT evaluations, table 4.4 includes data after QC checks had been applied. Most of the tables in this section include a Pearson correlation ($\rho$) which was derived using the system-level scores (avg. z) from WMT 20.

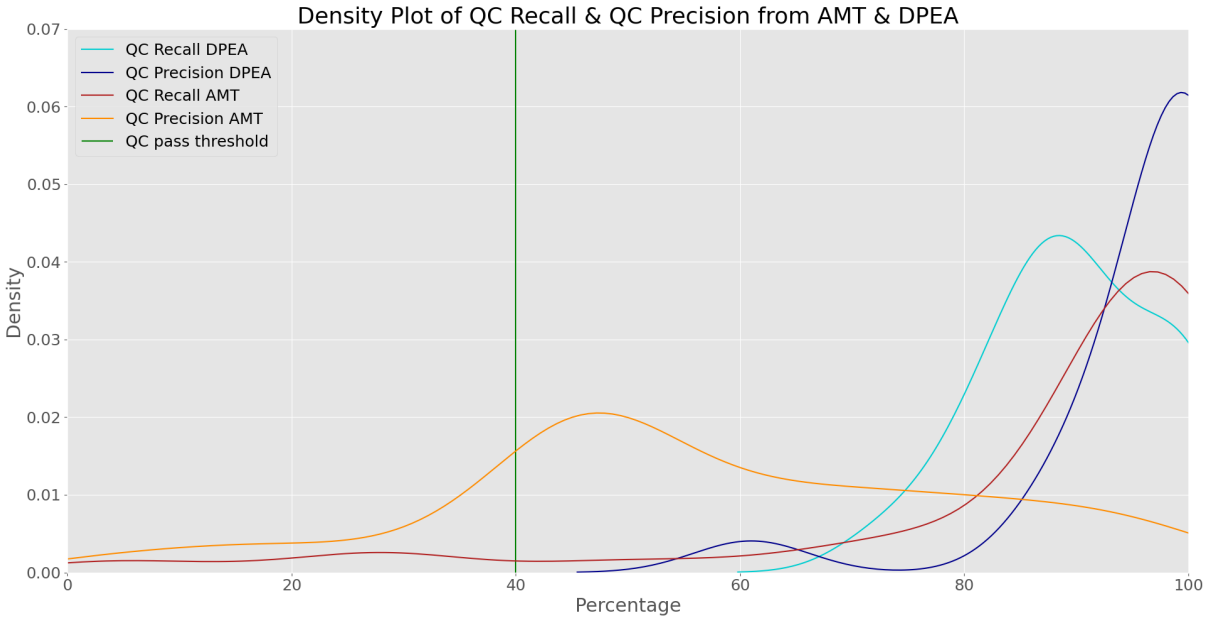


Figure 4.6: Density plot of the QC recall and precision from the DPEA (n=19) and AMT (n=347) experiments. Values of 0 QC recall and QC precision omitted for clarity

| Evaluation Type | Systems | Segments Assessed | Assessments/System |
|---|---|---|---|
| **German→English (DPEA)** | 13 | 253 | 19.5 |
| **German→English (AMT)** | 13 | 1953 | 150.2 |
| **English→Chinese (DPEA)*** | 14 | 503 | 35.9 |

Table 4.4: Summary of data collected. AMT evaluation summary includes data after QC had been applied.*Denotes a source-based (bilingual) evaluation.

| System | Words | TP | Recall | Precision | F-score |
|---|---|---|---|---|---|
| Tohoku-AIP-NTT | 431 | 427 | 99.072 | 99.072 | 99.072 |
| OPPO | 486 | 481 | 98.566 | 98.971 | 98.768 |
| HUMAN | 1176 | 1161 | 98.724 | 98.724 | 98.724 |
| VolcTrans | 697 | 686 | 98.422 | 98.422 | 98.422 |
| Online-G | 1326 | 1300 | 97.671 | 98.039 | 97.855 |
| UEDIN | 452 | 444 | 97.155 | 98.230 | 97.690 |
| Online-B | 983 | 955 | 97.648 | 97.152 | 97.399 |
| Promt-NMT | 470 | 457 | 97.441 | 97.234 | 97.338 |
| Online-A | 736 | 721 | 96.649 | 97.962 | 97.301 |
| Online-Z | 1590 | 1510 | 94.790 | 94.969 | 94.879 |
| zlabs-nlp | 536 | 498 | 93.084 | 92.910 | 92.997 |
| WMTBiomedBaseline | 944 | 878 | 92.616 | 93.008 | 92.812 |
| yolo | 375 | 50 | 12.594 | 13.333 | 12.953 |
| **Pearson Correlation ($\rho$)** | - | - | 0.991 | 0.991 | 0.991 |

Table 4.5: Unigram recall, precision and f-score taken from DPEA German-English.

### 4.2.1 Monolingual German to English Evaluation (DPEA)

The reference-based German→English evaluation on the DPEA consisted of 5 researchers. True positives (TP) were computed using word-level (unigram) matches. The recall, precision and f-score could then be calculated across the entire system, see table 4.5. It was expected that the system 'yolo' would finish poorly because it was observed during the evaluation that its translations tended to bare no similarity to the reference translation. In fact, a researcher assumed the DPEA website was experiencing a 'bug' when evaluating this system. In essence, this system only achieved 50 word-level matches and scored significantly lower for recall, precision and f-score. All things considered, the sample size for the German→English DPEA evaluation was relatively small (19.5 assessments/system) and so it would be too early to draw any conclusions.

Upon analysing table 4.5, it raised the question whether it was effective to compute results across the entire system data. For example, recall was calculated using all of the true positives and all the system output words. As such, in the case where one annotator tended to edit considerably more than another annotator, this could skew the results and thus create a bias. Therefore, to remove this kind of bias, recall, precision and f-score were calculated as an average for each segment score, see table 4.6. It is worth noting that table 4.5 and 4.6 are sorted by f-score and average f-score respectively. The systems 'Online-Z', 'zlabs-nlp', 'WMTBiomedBaseline' and 'yolo' rank the same in each table. Conversely, other systems ranked differently in each table with 'Tohoku-AIP-NTT' finishing first in table 4.5 but 'HUMAN' finishing first in table 4.6. Similar scores for recall, precision and f-score were determined with 2-grams (2 word sequences) and are available in appendix tables A1.2 and A1.1.

| System | Avg. Recall | Avg. Precision | Avg. F-score |
|---|---|---|---|
| Tohoku-AIP-NTT | 99.333 | 99.333 | 99.333 |
| HUMAN | 99.037 | 99.148 | 99.086 |
| VolcTrans | 98.647 | 98.765 | 98.704 |
| UEDIN | 98.000 | 98.667 | 98.327 |
| OPPO | 98.143 | 98.429 | 98.278 |
| Online-B | 97.870 | 97.565 | 97.701 |
| Promt-NMT | 97.071 | 97.214 | 97.116 |
| Online-A | 96.944 | 97.333 | 97.101 |
| Online-G | 96.951 | 97.195 | 97.066 |
| Online-Z | 94.618 | 95.559 | 94.957 |
| zlabs-nlp | 90.417 | 90.917 | 90.629 |
| WMTBiomedBaseline | 89.375 | 90.792 | 89.874 |
| yolo | 9.636 | 10.091 | 13.391 |
| **Pearson Correlation ($\rho$)** | 0.995 | 0.994 | 0.995 |

Table 4.6: Average unigram recall, precision and f-score taken from DPEA German-English.

Although there are issues with the WMT 20 official document-level evaluation method as previously outlined, it was still of interest in comparing the system-level scores. And so, it can be observed that averaging recall, precision and f-score tended to result in a slight increase in correlation than compared to assessing these scores against the entire system as in table 4.5.

Human targeted-metrics were calculated alongside standard automatic metrics, e.g. BLEU. A summary of these findings is available in table 4.7, sorted by HTER. All of these metrics agree in ranking 'yolo' in last and ranking 'Online-Z', 'zlabs-nlp', 'WMTBiomedBaseline' in 10th to 12th places. Another key point is that all human-targeted metrics were in agreement that 'Tohoku-AIP-NTT' system has come in 1st place. However, there was still ambiguity and disagreements as to other system rankings due to the scores being very close.

The fact standard automatic metrics like TER, BLEU and chrF also possessed high correlations with the WMT 20 results raised questions. The underlying problem for these high correlation is due to the number of observations, in this case there are 13 systems and so 13 observations. However, it is recommended to have at 25 or more observations when calculating the Pearson correlation [41]. Therefore, one cannot really make conclusions from these correlations. It was considered retrieving the segment-level scores taken from WMT 20 instead of using system-level rankings but due to individual biases from segment-level scores and the fact the scores were not derived with full document-level context, this approach was also not optimal.

| System | HTER | TER | HBLEU | BLEU | HchrF | chrF |
|---|---|---|---|---|---|---|
| Tohoku-AIP-NTT | 0.926 | 33.408 | 97.885 | 48.667 | 98.636 | 71.466 |
| HUMAN | 1.699 | 43.894 | 97.209 | 42.499 | 98.334 | 67.264 |
| VolcTrans | 1.717 | 46.029 | 96.215 | 43.027 | 98.143 | 69.022 |
| OPPO | 2.045 | 42.331 | 97.350 | 46.540 | 98.629 | 69.457 |
| UEDIN | 3.282 | 44.828 | 94.752 | 40.329 | 97.606 | 66.542 |
| Promt-NMT | 3.625 | 48.225 | 94.435 | 38.545 | 96.898 | 63.538 |
| Online-A | 3.753 | 47.865 | 94.119 | 37.458 | 96.365 | 66.189 |
| Online-B | 3.760 | 42.455 | 94.378 | 41.732 | 96.638 | 66.636 |
| Online-G | 4.185 | 43.783 | 93.752 | 44.387 | 96.734 | 70.024 |
| Online-Z | 7.825 | 51.377 | 89.539 | 36.004 | 94.345 | 66.372 |
| zlabs-nlp | 9.515 | 49.033 | 88.221 | 33.403 | 92.010 | 60.211 |
| WMTBiomedBaseline | 10.609 | 52.560 | 86.373 | 32.321 | 92.157 | 61.496 |
| yolo | 98.492 | 98.737 | 0.575 | 0.579 | 19.435 | 19.418 |
| **Pearson Correlation ($\rho$)** | 0.993 | 0.966 | 0.994 | 0.963 | 0.993 | 0.991 |

Table 4.7: Human-targeted metrics for DPEA German-English. Sorted by HTER.

### 4.2.2 Monolingual German to English Evaluation (AMT)

An evaluation campaign was performed on Amazon Mechanical Turk (AMT), a crowd-sourcing platform. The same dataset was used to that of the DPEA German to English setup but this time more assessments could be sourced. In total, 710 HITs were posted on AMT as a monolingual (reference-based) evaluation for the German to English language pair. However, only 39.7% of HITs passed QC checks which compared to 25.9% passing QC for the into English WMT 20 campaign. The number of HITs that passed QC might be considered low and the main factor for this is that some annotators on AMT were "gaming" the system. These individuals would generally complete 20+ HITs and these HITs would all fail QC. Although these HITs sometimes achieved the minimum required QC recall and QC precision, they were often completed in a short period of time, less than the threshold of 4 minutes that was set. In addition, this type of annotator would usually complete HITs simultaneously, something that was clearly outlined as prohibited in the HIT title. In summary, the QC thresholds were mostly able to filter out people "gaming" the system.

Unigram (word-level) matches were determined as an average of each segment assessment, see table 4.8. In this setup, the system 'Online-A' ranked in 12th place whereas at WMT 20 it fell in the rank 1-9 cluster and so there is some evidence of disagreement. The average recall, precision and f-score were also measured at 4-grams, see table 4.9. When increasing the number of grams, it is worth mentioning that the average f-score increased for the 'yolo' system while its average recall and precision went down. This type of behaviour was thus investigated. The f-score cannot be computed in cases where recall and precision for a segment sum to 0 because one cannot divide by 0 and so these segments are excluded from

| System | Avg. Recall | Avg. Precision | Avg. F-score |
|---|---|---|---|
| Online-B | 96.806 | 96.900 | 96.818 |
| OPPO | 96.737 | 96.251 | 96.438 |
| Online-G | 96.194 | 96.204 | 96.163 |
| Tohoku-AIP-NTT | 96.098 | 96.268 | 96.105 |
| VolcTrans | 95.459 | 95.300 | 95.339 |
| UEDIN | 94.857 | 94.943 | 94.856 |
| HUMAN | 95.057 | 94.581 | 94.754 |
| Online-Z | 94.198 | 94.631 | 94.362 |
| Promt-NMT | 93.396 | 94.550 | 93.892 |
| zlabs-nlp | 93.303 | 92.440 | 92.750 |
| WMTBiomedBaseline | 92.827 | 92.418 | 92.520 |
| Online-A | 92.213 | 92.976 | 92.473 |
| yolo | 74.840 | 74.056 | 76.771 |
| **Pearson Correlation ($\rho$)** | 0.974 | 0.985 | 0.976 |

Table 4.8: Average unigram recall, precision and f-score taken from AMT German-English.

determining the average f-score. There were numerous 0 recall and 0 precision scores which are reflected in average recall and average precision but these scores are not represented in the average f-score. The scores that did qualify for measuring the average f-score included a number of 100% scores. These were segments where the AMT annotator made no edits. The reason they made no changes is because they were 'gaming' the system and managed to pass QC or that they were confused by the translations and so decided to make no edits. In light of this, both scenarios are problematic but on the other hand, this is somewhat an edge-case because of the low quality of this particular system as previously discussed. The scores were similarly computed for 2-grams and 3-grams and are available in the appendix, see A1.4, A1.5 respectively.

A reliable method for comparing the results of this evaluation campaign with that of WMT 20 was by performing Wilcoxon rank-sum tests for $p < 0.05$. Bootstrap resampling was thus carried out on the unigram f-score segment-level scores which gave the upper and lower bounds of a 95% confidence interval. These Wilcoxon rank-sum test results are available in table 4.10. A similar Wilcoxon rank-sum test was also carried out for the DPEA setup and is available in the appendix, see A1.3.

A total of 2 cluster groups emerged for ranks 1-12 and 13 in table 4.10. For the WMT 20 results, there were cluster groups of ranks for 1-9, 10, 11-12 and 13. In view of this, 1 of the cluster groups (13) was exactly the same for these experiments and that of WMT 20. This was the system 'yolo' which as previously discussed was a weak MT system. Unlike WMT 20, this setup did not categorise the system 'Online-Z' in its own cluster nor 'WMTBiomedBaseline' and 'zlabs-nlp', instead grouped them in the ranks of 1-12.

In addition to not finishing in the same cluster, one might expect that 'Online-Z',

| System | Avg. Recall | Avg. Precision | Avg. F-score |
|---|---|---|---|
| OPPO | 91.202 | 90.896 | 92.614 |
| Online-B | 90.062 | 90.130 | 90.937 |
| Online-G | 88.855 | 88.743 | 90.475 |
| yolo | 66.826 | 66.326 | 90.053 |
| Tohoku-AIP-NTT | 88.110 | 88.384 | 88.698 |
| zlabs-nlp | 84.780 | 84.284 | 88.491 |
| VolcTrans | 86.186 | 86.018 | 86.587 |
| UEDIN | 85.676 | 85.686 | 86.457 |
| HUMAN | 85.856 | 85.413 | 86.409 |
| Online-Z | 82.162 | 82.667 | 86.263 |
| Promt-NMT | 83.381 | 83.774 | 86.102 |
| Online-A | 80.528 | 81.006 | 84.867 |
| WMTBiomedBaseline | 79.027 | 78.282 | 79.944 |
| **Pearson Correlation ($\rho$)** | 0.974 | 0.985 | 0.976 |

Table 4.9: Average 4-grams recall, precision and f-score taken from AMT German-English.

| | Online-B | OPPO | Online-G | Tohoku-AIP-NTT | VolcTrans | UEDIN | HUMAN | Online-Z | Promt-NMT | zlabs-nlp | WMTBiomedBaseline | Online-A | yolo |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Online-B | - | 0.00 | 0.01 | 0.01 | 0.01* | 0.02* | 0.02* | 0.02* | 0.03* | 0.04* | 0.04* | 0.04* | 0.20* |
| OPPO | -0.00 | - | 0.00 | 0.00 | 0.01 | 0.02 | 0.02* | 0.02* | 0.03* | 0.04* | 0.04* | 0.04* | 0.20* |
| Online-G | -0.01 | -0.00 | - | 0.00 | 0.01 | 0.01* | 0.01* | 0.02* | 0.02* | 0.03* | 0.04* | 0.04* | 0.19* |
| Tohoku-AIP-NTT | -0.01 | -0.00 | -0.00 | - | 0.01 | 0.01 | 0.01* | 0.02* | 0.02* | 0.03* | 0.04* | 0.04* | 0.19* |
| VolcTrans | -0.01 | -0.01 | -0.01 | -0.01 | - | 0.00 | 0.01 | 0.01 | 0.01 | 0.03* | 0.03* | 0.03* | 0.19* |
| UEDIN | -0.02 | -0.02 | -0.01 | -0.01 | -0.00 | - | 0.00 | 0.00 | 0.01 | 0.02* | 0.02* | 0.02* | 0.18* |
| HUMAN | -0.02 | -0.02 | -0.01 | -0.01 | -0.01 | -0.00 | - | 0.00 | 0.01 | 0.02 | 0.02* | 0.02* | 0.18* |
| Online-Z | -0.02 | -0.02 | -0.02 | -0.02 | -0.01 | -0.00 | -0.00 | - | 0.00 | 0.02 | 0.02 | 0.02 | 0.18* |
| Promt-NMT | -0.03 | -0.03 | -0.02 | -0.02 | -0.01 | -0.01 | -0.01 | -0.00 | - | 0.01 | 0.01 | 0.01 | 0.17* |
| zlabs-nlp | -0.04 | -0.04 | -0.03 | -0.03 | -0.03 | -0.02 | -0.02 | -0.02 | -0.01 | - | 0.00 | 0.00 | 0.16* |
| WMTBiomedBaseline | -0.04 | -0.04 | -0.04 | -0.04 | -0.03 | -0.02 | -0.02 | -0.02 | -0.01 | -0.00 | - | 0.00 | 0.16* |
| Online-A | -0.04 | -0.04 | -0.04 | -0.04 | -0.03 | -0.02 | -0.02 | -0.02 | -0.01 | -0.00 | -0.00 | - | 0.16* |
| yolo | -0.20 | -0.20 | -0.19 | -0.19 | -0.19 | -0.18 | -0.18 | -0.18 | -0.17 | -0.16 | -0.16 | -0.16 | - |
| **Rank** | 1-12 | 1-12 | 1-12 | 1-12 | 1-12 | 1-12 | 1-12 | 1-12 | 1-12 | 1-12 | 1-12 | 1-12 | 13 |

Table 4.10: Head to head comparison for German→English systems from AMT. *Signifies the system lies outside the 95% confidence interval.

| System | Rank | | |
|---|---|---|---|
| | WMT20 DA | Avg. F-score (DPEA) | Avg. F-score (AMT) |
| **VolcTrans** | 1 (1-9) | 3 (1-10) | 5 (1-12) |
| **OPPO** | 2 (1-9) | 5 (1-10) | 2 (1-12) |
| **HUMAN** | 3 (1-9) | 2 (1-10) | 7 (1-12) |
| **Tohoku-AIP-NTT** | 4 (1-9) | 1 (1-10) | 4 (1-12) |
| **Online-A** | 5 (1-9) | 8 (1-10) | 12 (1-12) |
| **Online-G** | 6 (1-9) | 9 (1-10) | 3 (1-12) |
| **PROMT-NMT** | 7 (1-9) | 7 (1-10) | 9 (1-12) |
| **Online-B** | 8 (1-9) | 6 (1-10) | 1 (1-12) |
| **UEDIN** | 9 (1-9) | 4 (1-10) | 6 (1-12) |
| **Online-Z** | 10 (10) | 10 (1-10) | 8 (1-12) |
| **WMTBiomedBaseline** | 11 (11-12) | 12 (11-12) | 11 (1-12) |
| **zlabs-nlp** | 12 (11-12) | 11 (11-12) | 10 (1-12) |
| **yolo** | 13 (13) | 13 (13) | 13 (13) |

Table 4.11: Comparison of MT System Rankings for German→English. Brackets indicate the ranking cluster from Wilcoxon rank-sum test.

'WMTBiomedBaseline' and 'zlabs-nlp' may finish in the same rank but they did not. For example, 'Online-Z' finished in 8th position from this campaign but 10th at WMT 20 in the cluster group of rank 10. A summary of the AMT and DPEA MT system rankings compared to that of WMT 20 is available in table 4.11. In terms of comparing the DPEA results and the AMT results, they similarly agree on 1 cluster group which is that of last place. However, it is perceived that there is generally a lack of agreement for other system rankings from 1-12.

It was considered adjusting the parameters for what HITs to use in the final analysis. For example, to change the QC recall threshold from 40% to 80%. However, when adjusting these parameters, it was observed that the number of observations significantly decreased. Furthermore, there is a strong case that adjusting such parameters could be guilty of over-fitting and so this idea was not pursued. Overall, it should be considered that the AMT evaluation campaign which identified a cluster ranking group for 1-12, could in fact be a better estimate of quality as the evaluations were done entirely at the document-level.

### 4.2.3 Source-based English to Chinese Evaluation (DPEA)

A source-based (bilingual) evaluation was performed with one annotator for English to Chinese on the DPEA. The code that had been developed supported a latin-script alphabet and did not include Chinese tokenisation. However, automatic metrics like BLEU, TER and chrF support Chinese when configured with SacreBLEU. The results of the human-targeted metrics are available in table 4.12. Note that Human-A is also the reference translation thus producing perfect HTER, TER and chrF scores. In addition, the system rankings were

| System | HBLEU | BLEU | HTER | TER | HchrF | chrF |
|---|---|---|---|---|---|---|
| Human-A | 99.850 | 100.000 | 8.824 | 0.000 | 99.878 | 100.000 |
| Tencent-Translation | 99.484 | 48.839 | 4.000 | 100.000 | 99.422 | 43.451 |
| OPPO | 98.920 | 47.789 | 9.375 | 117.073 | 98.742 | 40.821 |
| NiuTrans | 98.773 | 50.074 | 0.640 | 2 272.000 | 98.427 | 44.069 |
| SJTU-NICT | 98.342 | 42.586 | 17.391 | 112.000 | 98.012 | 35.997 |
| VolcTrans | 97.929 | 46.663 | 8.387 | 171.163 | 97.691 | 40.208 |
| Huawei-TSC | 96.236 | 38.087 | 26.316 | 100.000 | 95.998 | 35.435 |
| dong-nmt | 94.392 | 29.223 | 10.837 | 438.202 | 92.604 | 26.680 |
| Human-B | 92.288 | 36.800 | 18.182 | 103.571 | 91.991 | 33.819 |
| Online-A | 91.218 | 49.874 | 52.632 | 133.846 | 89.320 | 43.446 |
| Online-B | 88.525 | 45.855 | 47.368 | 110.909 | 86.981 | 40.112 |
| Online-G | 88.347 | 36.036 | 53.488 | 107.692 | 86.977 | 32.755 |
| zlabs-nlp | 88.291 | 31.144 | 47.500 | 119.863 | 87.235 | 27.160 |
| Online-Z | 84.399 | 33.179 | 52.941 | 158.333 | 81.493 | 29.207 |

Table 4.12: English-Chinese DPEA results. Sorted by HBLEU.

compared for HTER and HBLEU in table 4.13. The brackets under "WMT 20 DA" denote the cluster group of rankings, to wit 8 cluster groups were established.

A key point is that this method was calculated with SR+DC at WMT 20. As it was an out of English language pair, entire documents were evaluated with a segment-level breakdown (SR+DC) where the annotators could go back and forth between segments, unlike the into English setup. The out of English setup did not suffer the same issues as the into English setup, problems like trying to display an entire document while also maintaining QC. The reason it has this flexibility to show entire documents is due to the annotations being made by trusted researchers and professional translators. In view of this, it can be considered that these WMT 20 results are more reliable. Table 4.13 generally demonstrates a divergence in system rankings, without an agreement on what system came first or last. For these reasons, the use of document-level post-editing with HTER or HBLEU may not be the way forward for out of English source-based evaluation.

## 4.3 Inter-annotator Agreement for Post-editing

Determining inter-annotator agreement (IAA) for post-editing machine translations has appeared to be a mostly unexplored area. The data collected is not necessarily categorical as it is post-edits, i.e. word insertions or word deletions. This created complexities with regards to using Cohen's kappa [40], problems like quantifying an agreement ($P(A)$) and quantifying the agreement by chance ($P(E)$). Therefore, four unconventional strategies were conducted to measure IAA:

| System | Rank | | |
|---|---|---|---|
| | WMT20 DA | HTER | HBLEU |
| **Human-B** | 1 (1) | 9 | 8 |
| **Human-A** | 2 (2) | 1 | 4 |
| **OPPO** | 3 (3) | 3 | 5 |
| **Tencent-Translation** | 4 (4-8) | 2 | 2 |
| **Huawei-TSC** | 5 (4-8) | 7 | 9 |
| **NiuTrans** | 6 (4-8) | 4 | 1 |
| **SJTU-NICT** | 7 (4-8) | 5 | 7 |
| **VolcTrans** | 8 (4-8) | 6 | 3 |
| **Online-B** | 9 (9) | 11 | 10 |
| **Online-A** | 10 (10) | 10 | 12 |
| **dong-nmt** | 11 (11-13) | 8 | 6 |
| **Online-Z** | 12 (11-13) | 14 | 13 |
| **Online-G** | 13 (11-13) | 12 | 14 |
| **zlabs-nlp** | 14 (14) | 13 | 11 |

Table 4.13: Comparison of MT System Rankings for source-based English to Chinese. Brackets indicate the ranking cluster from Wilcoxon rank-sum tests.

- Annotator's agree to make one or more edits to a segment or none at all ($\kappa_1$)
- Annotator's agree to delete the same word from a segment ($\kappa_2$)
- Annotator's agree to insert the same word to a segment ($\kappa_3$)
- Annotator's agree to edit (deletions and insertions) the same word for a segment ($\kappa_4$)

50 segments were evaluated by the same 2 annotators on the DPEA. To gauge a comparison of the IAA for researchers and crowd-sourced workers, 50 segments were also annotated twice on AMT.

### Agreeing to edit or not to edit ($\kappa_1$)

Some MT systems did not requiring any editing and so $\kappa_1$ deals with agreement on whether to make any edits to a segment or none at all. In summary, there were 4 total outcomes for $\kappa_1$ including $a_1$ (annotator 1) and $a_2$ agree to edit a segment, they agree not to edit a segment, $a_1$ edits a segment but not $a_2$ and $a_2$ edits a segment but not $a_1$. The agreement of annotators ($P(A)$) for $\kappa_1$ is available in equation 2. $P(E)$, agreement by chance, was computed using equation 3 where $N$ is the observations, i.e. the sum of each of the 4 outcomes. $n_{ki}$ is the number of times annotator $i$ chose category $k$, e.g. number of times an annotator did make any edits to a segment.

$$P(A) = \frac{\#\ Agreements}{\#\ Segments} \tag{2}$$

| $a_1 \setminus a_2$ | Remove "large" | Remove "here" | No edit |
|---|---|---|---|
| **Remove "large"** | 1 | 0 | 0 |
| **Remove "here"** | 0 | 0 | 0 |
| **No edit** | 0 | 1 | 0 |

Table 4.14: Inter-annotator agreement levels from German-English reference-based evaluations.

$$P(E) = \frac{1}{N^2} \sum_k n_{k1} n_{k2} \tag{3}$$

### Annotator's agree to delete the same words ($\kappa_2$)

$\kappa_2$ looks at all removed words for each segment and determines the agreement on which words to remove. For $\kappa_2$, an agreement was given by the number of times annotators deleted the same word over the number of words that were deleted by both annotators, see equation 4.

$$P(A) = \frac{\#\ Agreed\ Word\ Deletions}{\#\ Deleted\ Words} \tag{4}$$

An example is provided to below to demonstrate the behaviour of this method.

Candidate translation: The cat sat on the large mat here.

Post-edit of $a_1$: The cat sat on the mat here.

Post-edit of $a_2$: The cat sat on the mat.

Both annotator's agree to remove the word "large" but annotator 2 also removes the word "here". Each unique word represents a category and so a diagonal matrix can be created, see table 4.14. Agreement by chance is similarly computed by equation 3. In this case, $N$ represents the total number of words deleted which is 2 in the example. $n_{ki}$ is the number of times that word ($k$) was deleted by annotator $i$. So for each uniquely deleted word and also the third category which represents no edits being made, $n_{ki}$ for word "large" and $a_1$ would be equal to 1 as they deleted this word once.

### Annotator's agree to insert the same words ($\kappa_3$)

$\kappa_3$ is similar to $\kappa_2$ except it deals with inserted words by annotators instead. As such P(A) is given by equation 5. Agreement by chance is again computed by equation 3.

$$P(A) = \frac{\#\ Agreed\ Word\ Insertions}{\#\ Inserted\ Words} \tag{5}$$

| Evaluation Type | $\kappa_1$ | $\kappa_2$ | $\kappa_3$ | $\kappa_4$ |
|---|---|---|---|---|
| **DPEA** | 0.4821 | 0.1835 | 0.1673 | 0.1741 |
| **AMT** | 0.2826 | 0.1290 | 0.0635 | 0.0950 |

Table 4.15: Inter-annotator agreement levels from German-English reference-based evaluations.

#### Annotator's agree to insert and delete the same words ($\kappa_4$)

$\kappa_4$ is essentially an aggregation of $\kappa_2$ and $\kappa_3$, i.e. the agreement on what words to delete and insert. P(A) for $\kappa_4$ is given in equation 6. Evidently, agreement by chance is again computed by equation 3.

$$P(A) = \frac{\#\ \textit{Agreed Word Insertions \& Agreed Word Deletions}}{\#\ \textit{Inserted Words \&}\ \#\ \textit{Deleted Words}} \tag{6}$$

#### Results of IAA Experiments

Table 4.15 outlines the IAA for both DPEA and AMT across the 4 different methods discussed. As one would expect, the highest agreement is seen in $\kappa_1$ where annotators either edit or do not edit a segment. In comparison with relative ranking (RR) German-English, IAA has been 0.475 (WMT 16), 0.423 (WMT 15) and 0.368 (WMT 14). Therefore, it can be observed that $\kappa_1$ for the DPEA tends to perform around the same in terms of IAA to these WMT results. However, AMT tended to have lower IAA levels throughout compared to the DPEA. AMT annotators achieved $\kappa_3$=0.0635 which means that they were not able to find much agreement on which words to insert to a segment. Another key fact to remember is that an annotator can insert any word into a segment and theoretically speaking, they could insert any sequence of characters of any length. In this setup, annotators are not limited by discrete categories and thus although the IAA levels may seem lower, it is significant that they are still showing slight agreement.

## 4.4 Post-editing Analysis

After collecting data from 1,953 segments which had been annotated on AMT, it was of interest in visualising some of this data. In figure 4.7, an aggregation of all edits made across all systems can be seen where an edit is an insertion or deletion. This graph highlights that more edits were done towards the end of segments instead of at the start. It could be concluded that either the annotators preferred editing towards the end of a segment or all MT systems required more editing towards the end of the segment. A more detailed system-level breakdown for some systems is available in figure 4.8. In contrast, this graph does not show as much of an evident skew as the aggregated results.

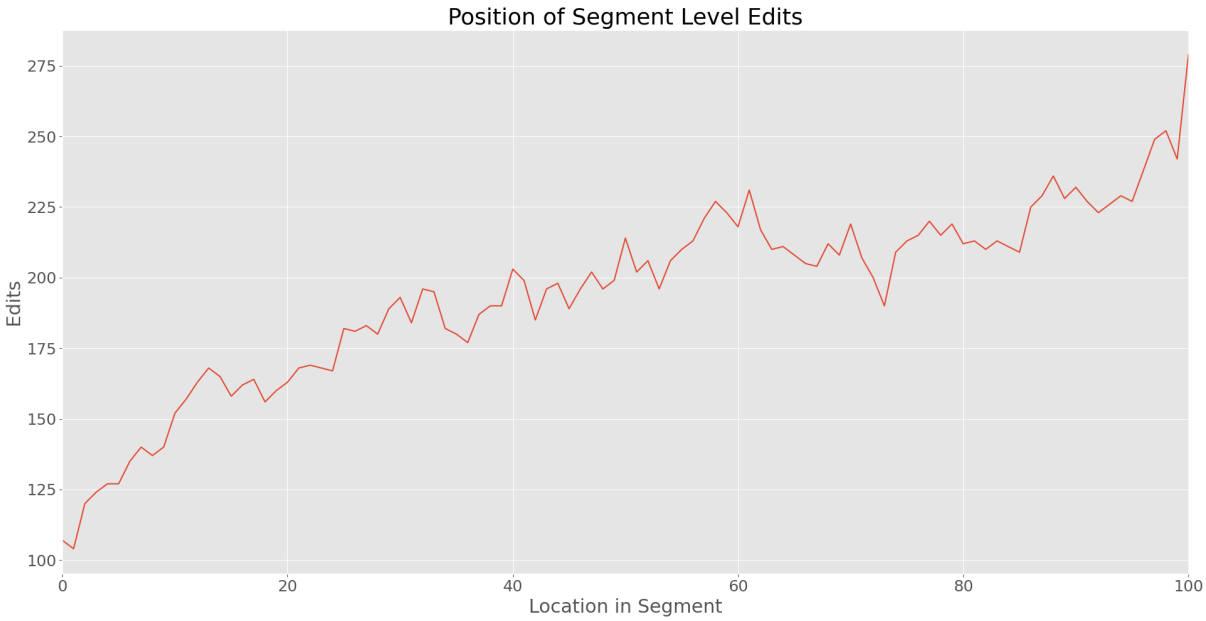


Figure 4.7: The position of where segments were edited by AMT annotators for German-English.

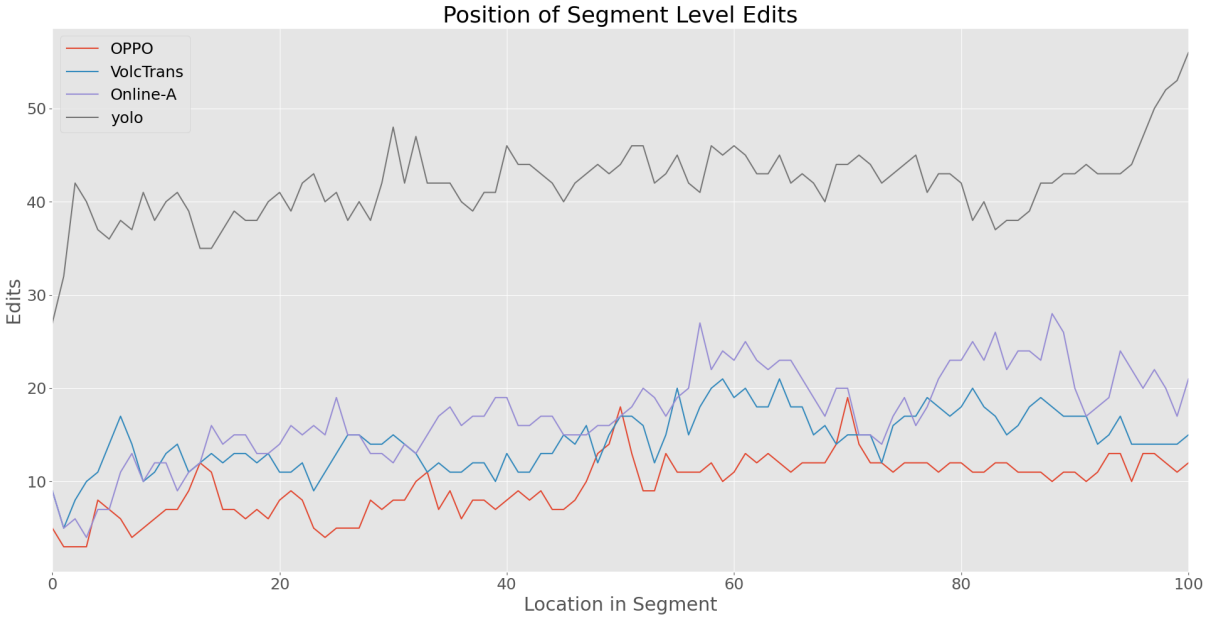


Figure 4.8: The position of where segments were edited by AMT annotators for German-English.

The most commonly edited (inserted/deleted) words is visualised in figure 4.9. This data is aggregated across all systems for the AMT German-English campaign. As expected, the most commonly edited words were also generally the most frequently written words in English, e.g. 'the', 'of' and 'to'. Overall, the word 'the' was the most edited word in the German-English AMT evaluation campaign.

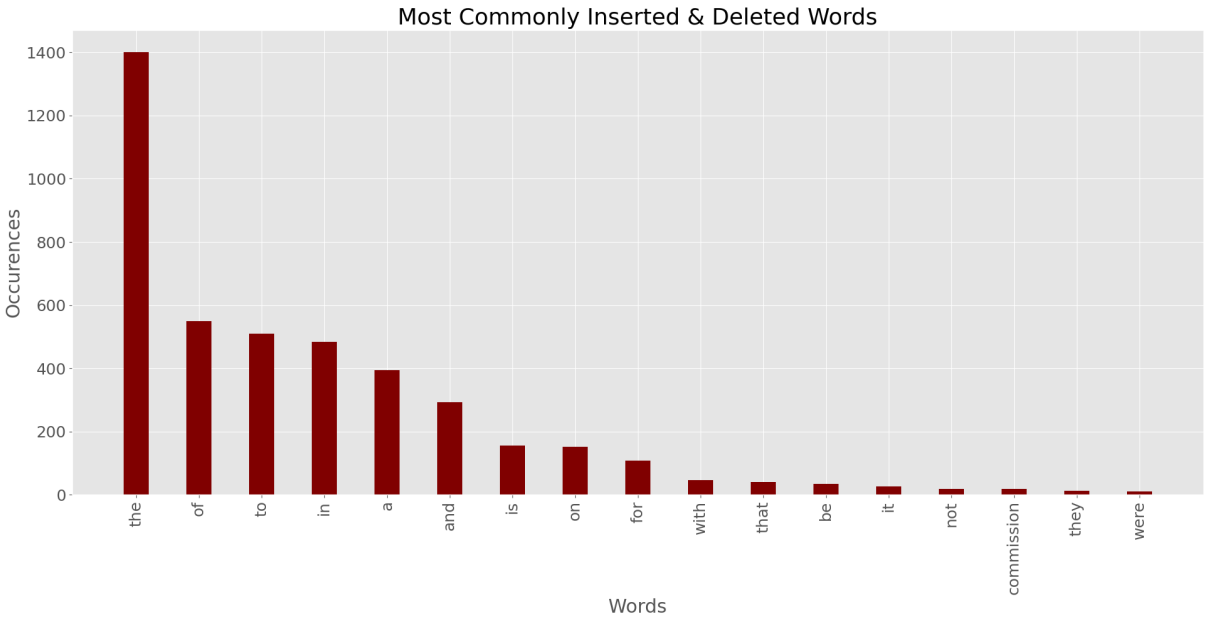


Figure 4.9: The most inserted & deleted words by AMT annotators for German-English

# 5 Conclusion

## 5.1 Summary of Work Completed

A new method of MT evaluation has been developed which is carried out at the document-level, annotated through post-editing. This method can be evaluated through the web application developed or through crowd-sourcing platforms like AMT. The method includes robust QC checking techniques in order to mitigate low-quality results. In total, 3 evaluation campaigns took place which included a source-based evaluation, reference-based evaluations and a crowd-sourced evaluation. Research was additionally undertaken to assess the inter-annotator agreement levels for post-editing at the document-level. Finally, a number of metrics, including newly developed scoring methods, have been proposed for estimating the quality of MT.

## 5.2 Key Findings

Performing post-edits on candidate translations has shown reasonable agreement amongst annotators. The IAA experiments have proven that annotators find moderate agreement on when to edit or not to edit a segment ($\kappa_1 = 0.4821$) although this agreement was not as high for crowd-sourced annotators ($\kappa_1 = 0.2826$). Furthermore, there is evidence of slight agreement for deleting ($\kappa_2 = 0.1835$) and inserting ($\kappa_3 = 0.1673$) words with the latter being significant due the fact that annotators were not limited by discrete categories but were still showing slight agreement.

In assessing objective (ii), the QC results were mostly positive. The QC documents were doable but it was still necessary for annotators to put effort into the HITs in order to pass QC. It was also feasible manually analysing the QC results for an AMT annotator and so after analysing ≈50 HITS, specific thresholds could be established to determine which HITs to pass or fail. Looking closely at QC items is not feasible for current methods like DA where there is somewhat of a black box in determining which HITs pass or fail QC. Nonetheless, this QC method proposed is not perfect as there were cases whereby annotators passed QC but then proceeded not to make edits to a known low quality MT

system output. All things considered, the QC method proposed provided more meaningful data-points for assessing quality than existing methods and to some degree successfully eliminated low quality submissions of data.

Objective (i) and (iii) dealt with the feasibility to the approach and correlation with previous WMT evaluation campaigns. The German-English monolingual evaluations that were sourced from researchers showed similar cluster rankings with that of the WMT evaluation campaign. However, for the crowd-sourced evaluations, only 2 cluster groups emerged compared to 4 cluster groups for the WMT evaluation campaign. It is possible that the change in cluster rankings was due to the fact that annotators were given the full document context. Nonetheless, the method proposed deals with the problem of providing document-level context on crowd-sourcing platforms and has still shown some correlation with official WMT results. In conclusion, it has been demonstrated that machine translation can be evaluated at the document-level through the use of post-editing on crowd-sourcing platforms.

## 5.3 Limitations

It was expected that the results from the 2 German-English evaluation campaigns would be somewhat correlated. Conversely, there seemed to be a sense of divergence in system rankings which could be down to a number of factors:

- The sample size for the DPEA was relatively small with 19.5 assessments/system compared to 150.2 assessments/system for AMT.
- The QC checks for AMT were unsuccessful in eliminating people 'gaming' the system, leading to low-quality results.
- The developed unigram f-score metric is not a reliable quality estimation.

Although QC mechanisms were put in place, there was still evidence of people not correctly completing the task. For example, one of the known weaker MT systems experienced issues where annotators made a total of 0 edits while still passing QC. It is not possible to place QC checks everywhere in a HIT but this finding may be indicative of a need for stricter QC mechanisms. Another key limitation to this approach is the use of human-targeted metrics. However, this was a necessary trade-off so that annotators were presented with entire documents.

As the SR+DC out of English setup tends to deal with providing full document-level context, this method could be assumed a 'gold standard'. The English-Chinese source-based evaluation campaign did not produce much agreement with categorising systems in similar ranks to that of the cluster ranks from WMT. To this end, the use of document-level

post-editing with HTER or HBLEU may not be the forward for out of English source-based evaluation. On the other hand, the annotations were only performed by one annotator and so the sample size was relatively small in terms of the number of documents evaluated.

When initially planning the project, the aim was to run 5 evaluation campaigns on AMT for 5 different languages. However, the work undertaken required ethics approval and so as previously stated, due to unforeseen circumstances, ethics approval was not granted until March 2022 and thus the scope of the project was cut.

## 5.4 Future Work

With the collection of over 2,000 segment post-edits, it would be possible to manually explore and investigate where individual MT systems went wrong. Furthermore, these post-edits could be provided to the research teams who submitted MT system to train NMT systems. As there is an automatic metric component to the work carried out, it may be of interest testing different types of metrics that were not used here. In terms of the sequence matching algorithms, the behaviour seen when extracting post-edits with difflib [34] was not desirable and so it would be worthwhile researching these anomalies.

A lot of work has been invested in the development of the frameworks necessary for carrying out post-editing at the document-level. In fact, the code developed for the research undertaken is already being used in another evaluation campaign which assesses video capturing. These tools could be used to run evaluation campaigns for other language pairs and datasets. Above all, it would be of interest to rerun the evaluation campaign for the German-English dataset and to see if these results are reproducible.

# Bibliography


[1] Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, FirstView:1–28, 1 2016. ISSN 1469-8110. doi: 10.1017/S1351324915000339. URL http://journals.cambridge.org/article_S1351324915000339.

[2] Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, et al. Findings of the 2020 conference on machine translation (wmt20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–54, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.wmt-1.1.

[3] Yvette Graham, Timothy Baldwin, Meghan Dowling, Maria Eskevich, Teresa Lynn, and Lamia Tounsi. Is all that glitters in machine translation quality estimation really gold standard? In *Proceedings of the 26th International Conference on Computational Linguistics*, Osaka, Japan, 2016.

[4] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany, August 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W16/W16-2301.

[5] Warren weaver memorandum, july 1949. *MT News International*, 1999.

[6] Automatic Language Processing Advisory Committee. Language and machines. *A report by the Automatic Language Processing Advisory Committee*, 1966.

[7] Barak Turovsky. Ten years of google translate. https:

//blog.google/products/translate/ten-years-of-google-translate/. Accessed: 2022-04-06.

[8] Samuel Läubli, Rico Sennrich, and Martin Volk. Has machine translation achieved human parity? a case for document-level evaluation. *arXiv preprint arXiv:1808.07048*, 2018.

[9] Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. Attaining the unattainable? reassessing claims of human parity in neural machine translation. *arXiv preprint arXiv:1808.10432*, 2018.

[10] Sheila Castilho. On the same page? comparing inter-annotator agreement in sentence and document level human machine translation evaluation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1150–1159, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.wmt-1.137.

[11] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André FT Martins, et al. 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. URL http://www.aclweb.org/anthology/W19-5301.

[12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[13] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

[14] Yvette Graham, Nitika Mathur, and Timothy Baldwin. Randomized significance tests in machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 266–274, 2014.

[15] Marina Fomicheva and Lucia Specia. Reference bias in monolingual machine translation evaluation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 77–82, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2013. URL https://aclanthology.org/P16-2013.

[16] Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA, August 8-12 2006. Association for Machine Translation in the Americas. URL https://aclanthology.org/2006.amta-papers.25.

[17] Ondřej Bojar, Miloš Ercegovčević, Martin Popel, and Omar Zaidan. A grain of salt for the WMT manual evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. URL https://aclanthology.org/W11-2101.

[18] Yvette Graham, Barry Haddow, and Philipp Koehn. Translationese in machine translation evaluation. *arXiv preprint arXiv:1906.09833*, 2019.

[19] Yvette Graham, Barry Haddow, and Philipp Koehn. Statistical power and translationese in machine translation evaluation. In *EMNLP (1)*, pages 72–81, 2020.

[20] Lucia Specia. Exploiting objective annotations for minimising translation post-editing effort. In *Proceedings of the 15th Annual conference of the European Association for Machine Translation*, 2011.

[21] Michael Denkowski and Alon Lavie. Choosing the right evaluation for machine translation: an examination of annotator and automatic metric performance on human judgment tasks. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*, Denver, Colorado, USA, October 31-November 4 2010. Association for Machine Translation in the Americas. URL https://aclanthology.org/2010.amta-papers.20.

[22] Carolina Scarton, Marcos Zampieri, Mihaela Vela, Josef van Genabith, and Lucia Specia. Searching for context: a study on document-level labels for translation quality estimation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 121–128, Antalya, Turkey, May 2015. URL https://aclanthology.org/W15-4916.

[23] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W17-4717.

[24] Rachel Bawden, Giorgio Di Nunzio, Cristian Grozea, Iñigo Unanue, Antonio Yepes, Nancy Mah, David Martinez, Aurélie Névéol, Mariana Neves, Maite Oronoz, et al. Findings of the wmt 2020 biomedical translation shared task: Basque, italian and russian as new additional languages. In *5th Conference on Machine Translation*, 2020.

[25] Wilker Aziz, Sheila Castilho, and Lucia Specia. Pet: a tool for post-editing and assessing machine translation. In *LREC*, pages 3982–3987. Citeseer, 2012.

[26] Kateřina Rysová, Magdaléna Rysová, Tomáš Musil, Lucie Poláková, and Ondřej Bojar. A test suite and manual evaluation of document-level nmt at wmt19. *arXiv preprint arXiv:1908.03043*, 2019.

[27] Christian Federmann. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico, August 2018. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/C18-2019`.

[28] Django Software Foundation. Django. URL `https://djangoproject.com`.

[29] Security in django. `https://docs.djangoproject.com/en/4.0/topics/security/`. Accessed: 2022-04-03.

[30] B. Kaliski. Pkcs #5: Password-based cryptography specification version 2.0. RFC 2898, RFC Editor, September 2000. URL `http://www.rfc-editor.org/rfc/rfc2898.txt`.

[31] jQuery. jquery : write less, do more, 2012. URL `http://www.jquery.com/`.

[32] Bootstrap v5.0. `https://getbootstrap.com/docs/5.0/getting-started/introduction/`. Accessed: 2022-04-03.

[33] 2to3 - automated python 2 to 3 code translation. `https://docs.python.org/3/library/2to3.html`, . Accessed: 2022-04-10.

[34] difflib. `https://docs.python.org/3/library/difflib.html`, . Accessed: 2022-04-04.

[35] David E. Metzener John W. Ratcliff. Pattern matching: The gestalt approach. *Dr. Dobb's Journal*, page 46, 1988. URL `https://www.drdobbs.com/database/pattern-matching-the-gestalt-approach/184407970?pgno=5`.

[36] diff-match-patch. `https://pypi.org/project/diff-match-patch/`, . Accessed: 2022-04-04.

[37] Eugene W. Myers. An o(nd) difference algorithm and its variations. *Algorithmica*, 1: 251–266, 1986.

[38] unittest — unit testing framework. `https://docs.python.org/3/library/unittest.html`, . Accessed: 2022-04-04.

[39] 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`.

[40] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37, 1960.

[41] Florence Nightingale David. *Tables of the ordinates and probability integral of the distribution of the correlation coefficient in small samples*. Cambridge University Press, 1938.

# A1 Appendix

## A1.1 DPEA German-English

## A1.2 AMT German-English

| System | 2-grams | TP | Recall | Precision | F-score |
|---|---|---|---|---|---|
| Tohoku-AIP-NTT | 422 | 414 | 98.104 | 98.104 | 98.104 |
| OPPO | 472 | 463 | 97.679 | 98.093 | 97.886 |
| HUMAN | 1149 | 1122 | 97.650 | 97.650 | 97.650 |
| VolcTrans | 680 | 657 | 96.618 | 96.618 | 96.618 |
| Online-G | 1285 | 1238 | 95.969 | 96.342 | 96.155 |
| Online-B | 960 | 914 | 95.707 | 95.208 | 95.457 |
| UEDIN | 443 | 425 | 94.866 | 95.937 | 95.398 |
| Promt-NMT | 456 | 433 | 95.165 | 94.956 | 95.060 |
| Online-A | 718 | 685 | 94.093 | 95.404 | 94.744 |
| Online-Z | 1556 | 1415 | 90.763 | 90.938 | 90.851 |
| zlabs-nlp | 524 | 464 | 88.719 | 88.550 | 88.634 |
| WMTBiomedBaseline | 920 | 811 | 87.771 | 88.152 | 87.961 |
| yolo | 364 | 3 | 0.777 | 0.824 | 0.800 |
| **Pearson Correlation ($\rho$)** | - | - | 0.991 | 0.991 | 0.991 |

Table A1.1: 2-grams recall, precision and f-score taken from DPEA German→English.

| System | Avg. Recall | Avg. Precision | Avg. F-score |
|---|---|---|---|
| Tohoku-AIP-NTT | 98.889 | 98.889 | 98.889 |
| HUMAN | 98.259 | 98.407 | 98.328 |
| VolcTrans | 97.353 | 97.353 | 97.352 |
| UEDIN | 96.222 | 97.111 | 96.660 |
| OPPO | 96.357 | 96.643 | 96.493 |
| Online-B | 96.130 | 95.826 | 95.961 |
| Promt-NMT | 94.786 | 94.857 | 94.795 |
| Online-A | 94.111 | 94.611 | 94.321 |
| Online-G | 94.073 | 94.341 | 94.199 |
| Online-Z | 91.118 | 92.000 | 91.426 |
| zlabs-nlp | 85.167 | 85.417 | 85.255 |
| WMTBiomedBaseline | 81.958 | 83.333 | 82.423 |
| yolo | 0.636 | 0.636 | 2.333 |
| **Pearson Correlation ($\rho$)** | 0.995 | 0.994 | 0.995 |

Table A1.2: Average 2-grams recall, precision and f-score taken from DPEA German→English.

| | Tohoku-AIP-NTT | HUMAN | VolcTrans | UEDIN | OPPO | Online-B | Promt-NMT | Online-A | Online-G | Online-Z | zlabs-nlp | WMTBiomedBaseline | yolo |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Tohoku-AIP-NTT | - | 0.00 | 0.01 | 0.01* | 0.01* | 0.02* | 0.02* | 0.02* | 0.02* | 0.04* | 0.09* | 0.09* | 0.86* |
| HUMAN | -0.00 | - | 0.00 | 0.01* | 0.01* | 0.01* | 0.02* | 0.02* | 0.02* | 0.04* | 0.08* | 0.09* | 0.86* |
| VolcTrans | -0.01 | -0.00 | - | 0.00 | 0.00 | 0.01* | 0.02* | 0.02* | 0.02* | 0.04* | 0.08* | 0.09* | 0.85* |
| UEDIN | -0.01 | -0.01 | -0.00 | - | 0.00 | 0.01 | 0.01* | 0.01* | 0.01* | 0.03* | 0.08* | 0.08* | 0.85* |
| OPPO | -0.01 | -0.01 | -0.00 | -0.00 | - | 0.01 | 0.01 | 0.01 | 0.01 | 0.03* | 0.08* | 0.08* | 0.85* |
| Online-B | -0.02 | -0.01 | -0.01 | -0.01 | -0.01 | - | 0.01 | 0.01 | 0.01 | 0.03* | 0.07* | 0.08* | 0.84* |
| Promt-NMT | -0.02 | -0.02 | -0.02 | -0.01 | -0.01 | -0.01 | - | 0.00 | 0.00 | 0.02 | 0.06* | 0.07* | 0.84* |
| Online-A | -0.02 | -0.02 | -0.02 | -0.01 | -0.01 | -0.01 | -0.00 | - | 0.00 | 0.02 | 0.06* | 0.07* | 0.84* |
| Online-G | -0.02 | -0.02 | -0.02 | -0.01 | -0.01 | -0.01 | -0.00 | -0.00 | - | 0.02* | 0.06* | 0.07* | 0.84* |
| Online-Z | -0.04 | -0.04 | -0.04 | -0.03 | -0.03 | -0.03 | -0.02 | -0.02 | -0.02 | - | 0.04* | 0.05* | 0.82* |
| zlabs-nlp | -0.09 | -0.08 | -0.08 | -0.08 | -0.08 | -0.07 | -0.06 | -0.06 | -0.06 | -0.04 | - | 0.01 | 0.77* |
| WMTBiomedBaseline | -0.09 | -0.09 | -0.09 | -0.08 | -0.08 | -0.08 | -0.07 | -0.07 | -0.07 | -0.05 | -0.01 | - | 0.76* |
| yolo | -0.86 | -0.86 | -0.85 | -0.85 | -0.85 | -0.84 | -0.84 | -0.84 | -0.84 | -0.82 | -0.77 | -0.76 | - |
| **Rank** | 1-10 | 1-10 | 1-10 | 1-10 | 1-10 | 1-10 | 1-10 | 1-10 | 1-10 | 1-10 | 11-12 | 11-12 | 13 |

Table A1.3: Head to head comparison for German-English systems. Data from DPEA whereby annotators were researchers. *Signifies the system lies outside the 95% confidence interval.

| System | Avg. Recall | Avg. Precision | Avg. F-score |
|---|---|---|---|
| OPPO | 94.206 | 93.794 | 94.492 |
| Online-B | 94.162 | 94.252 | 94.174 |
| Online-G | 93.347 | 93.292 | 93.290 |
| Tohoku-AIP-NTT | 93.152 | 93.293 | 93.145 |
| VolcTrans | 91.747 | 91.559 | 91.617 |
| UEDIN | 91.295 | 91.371 | 91.290 |
| HUMAN | 91.467 | 91.019 | 91.180 |
| zlabs-nlp | 89.661 | 88.963 | 90.873 |
| Promt-NMT | 89.485 | 90.396 | 89.867 |
| Online-Z | 89.514 | 89.946 | 89.681 |
| Online-A | 87.512 | 88.171 | 87.734 |
| WMTBiomedBaseline | 87.509 | 87.009 | 87.152 |
| yolo | 70.076 | 69.326 | 84.727 |
| **Pearson Correlation ($\rho$)** | 0.974 | 0.985 | 0.976 |

Table A1.4: Average 2-grams recall, precision and f-score taken from AMT German-English.

| System | Avg. Recall | Avg. Precision | Avg. F-score |
|---|---|---|---|
| OPPO | 92.132 | 91.770 | 93.523 |
| Online-B | 92.024 | 92.076 | 92.459 |
| Online-G | 90.958 | 90.833 | 92.167 |
| yolo | 68.319 | 67.681 | 91.136 |
| Tohoku-AIP-NTT | 90.457 | 90.634 | 91.020 |
| zlabs-nlp | 86.743 | 86.156 | 89.641 |
| VolcTrans | 88.385 | 88.160 | 89.295 |
| HUMAN | 88.365 | 87.942 | 88.949 |
| UEDIN | 88.476 | 88.543 | 88.463 |
| Promt-NMT | 85.970 | 86.589 | 87.263 |
| Online-Z | 85.658 | 86.135 | 85.850 |
| Online-A | 83.585 | 84.085 | 85.316 |
| WMTBiomedBaseline | 83.118 | 82.536 | 82.708 |
| **Pearson Correlation ($\rho$)** | 0.974 | 0.985 | 0.976 |

Table A1.5: Average 3-grams recall, precision and f-score taken from AMT German-English.