# SubER: A Metric for Automatic Evaluation of Subtitle Quality

**Patrick Wilken**
AppTek
Aachen, Germany
pwilken@apptek.com

**Panayota Georgakopoulou**
Athena Consultancy
Athens, Greece
yota@athenaconsultancy.eu

**Evgeny Matusov**
AppTek
Aachen, Germany
ematusov@apptek.com

## Abstract

This paper addresses the problem of evaluating the quality of automatically generated subtitles, which includes not only the quality of the machine-transcribed or translated speech, but also the quality of line segmentation and subtitle timing. We propose SubER - a single novel metric based on edit distance with shifts that takes all of these subtitle properties into account. We compare it to existing metrics for evaluating transcription, translation, and subtitle quality. A careful human evaluation in a post-editing scenario shows that the new metric has a high correlation with the post-editing effort and direct human assessment scores, outperforming baseline metrics considering only the subtitle text, such as WER and BLEU, and existing methods to integrate segmentation and timing features.

## 1 Introduction

The use of automatically created subtitles has become popular due to improved speech recognition (ASR) and machine translation (MT) quality in recent years. Most notably, they are used on the web to make content available to a broad audience in a cost-efficient and scalable way. They also gain attraction in the media industry, where they can be an aid to professional subtitlers and lead to increased productivity.

In this work, we address the problem of measuring the quality of such automatic subtitling systems. We argue that existing metrics which compare the plain text output of an ASR or MT system to a reference text are not sufficient to reflect the particularities of the subtitling task. We consider two use cases: 1) running speech recognition on the audio track of a video to create subtitles in the original language; 2) translating existing subtitle files with an MT system. For the first case, the word error rate (WER) of the ASR system is a natural choice for quality control. For MT there exist a wider range of automatic metrics such as BLEU (Papineni et al., 2002), TER (Snover et al., 2006), chrF (Popović, 2015) and, more recently, learned metrics like BertScore (Zhang et al., 2019) and COMET (Rei et al., 2020).

These existing metrics are suited to measure the quality of ASR and MT in terms of recognized or translated content only. However, subtitles are defined by more than just their textual content: they include timing information, as well as formatting with possible line breaks within a sentence in syntactically and semantically proper positions. Figure 1 shows examples of subtitle files in the common SubRip text (SRT) format. Evidently, it differs from plain text, in particular:

- The text is segmented into blocks. These blocks are distinct from sentences. A sentence can span several blocks, a block can contain multiple sentences.
- A block may be further split into lines.
- Start and end times define when text is displayed.

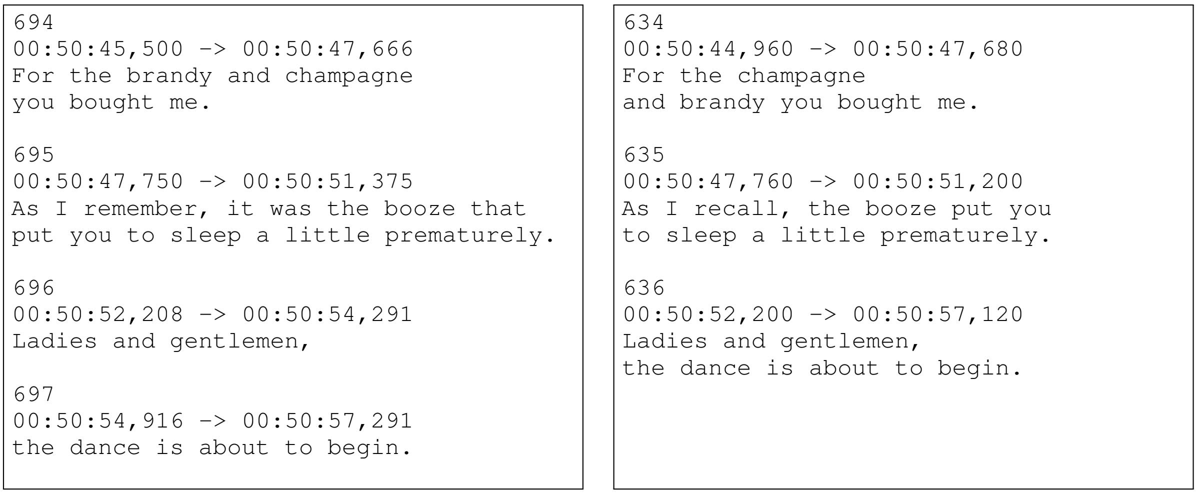
All of these additional characteristics are crucial for the viewers' comprehension of the content. Professional subtitlers check and possibly improve them as part of the machine-assisted process of subtitle creation.

To assess the quality of automatically created subtitle files, it is beneficial to have a *single* metric that evaluates the ASR/MT quality and the quality of the characteristics listed above.

The main contributions of this work are:

1. A novel segmentation- and timing-aware quality metric designed for the task of automatic subtitling.
2. A human evaluation that analyzes how well the proposed metric correlates with human judgements of subtitle quality, measured in

*Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 1 - 10
May 26-27, 2022 ©2022 Association for Computational Linguistics



```
694
00:50:45,500 -> 00:50:47,666
For the brandy and champagne
you bought me.

695
00:50:47,750 -> 00:50:51,375
As I remember, it was the booze that
put you to sleep a little prematurely.

696
00:50:52,208 -> 00:50:54,291
Ladies and gentlemen,

697
00:50:54,916 -> 00:50:57,291
the dance is about to begin.
```

```
634
00:50:44,960 -> 00:50:47,680
For the champagne
and brandy you bought me.

635
00:50:47,760 -> 00:50:51,200
As I recall, the booze put you
to sleep a little prematurely.

636
00:50:52,200 -> 00:50:57,120
Ladies and gentlemen,
the dance is about to begin.
```

Figure 1: Two examples of subtitles in SRT format for the same video excerpt. Note the different line and block segmentation. Also note that subtitles on the right have been condensed for improved readability.

post-editing effort as well as direct assessment scores.

3. The publication of a scoring tool to calculate the proposed metric as well as many baseline metrics, directly operating on subtitle files: `https://github.com/apptek/SubER`

## 2 Subtitle Quality Assessment in the Media Industry

Related to this work are subtitling quality metrics used in the media industry. The most widely used ones to date are NER (Romero-Fresco and Pérez, 2015) and NTR (Romero-Fresco and Pöchhacker, 2017) for live subtitle quality, the former addressing intralingual subtitles or captions and the latter interlingual ones.

Offline interlingual subtitles have traditionally been assessed on the basis of internal quality guidelines and error typologies produced by media localization companies. To address this gap, the FAR model (Pedersen, 2017) was developed and there have also been attempts to implement a version of MQM[1].

None of the above metrics, however, are automatic ones. They require manual evaluation by an expert to categorize errors and assign appropriate penalties depending on their severity. This makes their use costly and time-consuming. In this work we therefore address automatic quality assessment of subtitle files by comparing them to a professionally created reference.

## 3 Automatic Metrics for Subtitling

### 3.1 Baseline Approaches

When subtitling in the original language of a video, the baseline quality measurement is to calculate word error rate (WER) against a reference transcription. Traditionally, WER is computed on lower-cased words and without punctuation. We show results for a cased and punctuated variant as well, as those are important aspects of subtitle quality. Because of the efficiency of the Levenshtein algorithm, WER calculation can be done on the whole file without splitting it into segments.

For translation, automatic metrics are usually computed on sentence level. Karakanta et al. (2020a) and other related work assumes hypothesis-reference sentence pairs to be given for subtitle scoring. However, in the most general case we only have access to the reference subtitle file and the hypothesis subtitle file to be scored. They do not contain any explicit sentence boundary information. To calculate traditional MT metrics (BLEU, TER and chrF), we first define reference segments and then align the hypothesis subtitle text to these reference segments by minimizing the edit distance ("Levenshtein alignment") (Matusov et al., 2005). Two choices of reference segments are reasonable: 1) subtitle blocks; 2) sentences, split according to simple rules based on sentence-final punctuation, possibly spanning across subtitle blocks. Only for the case of translation from a subtitle template, which preserves subtitle timings, there is a third option, namely to directly use the parallel subtitle blocks as units without any alignment step. This makes the metric sensitive to how translated

[1] Multidimensional Quality Metrics (MQM) Definition http://www.qt21.eu/mqm-definition/definition-2015-12-30.html

sentences are distributed among several subtitles, which is a problem a subtitle translation system has to solve.

To evaluate subtitle segmentation quality in isolation, Alvarez et al. (2017); Karakanta et al. (2020b,c) calculate precision and recall of predicted breaks. Such an analysis is only possible when the subtitle text to be segmented is fixed and the only degree of freedom is the position of breaks. We however consider the general case, where subtitles that differ in text, segmentation and timing are compared and evaluated.

### 3.2 Line Break Tokens

A simple method to extend the baseline metrics to take line and subtitle breaks into account is to insert special tokens at the corresponding positions into the subtitle text (Karakanta et al., 2020a; Matusov et al., 2019). Figure 2 shows an example. The automatic metrics treat these tokens as any other word, e.g. BLEU includes them in n-grams, WER and TER count edit operations for them. Therefore, subtitles with a segmentation not matching the reference will get lower scores.

### 3.3 Timing-Based Segment Alignment

The time alignment method proposed in Cherry et al. (2021) to calculate t-BLEU is an alternative to Levenshtein hypothesis-to-reference alignment that offers the potential advantage of punishing mistimed words. It uses interpolation of the hypothesis subtitle timings to word-level. Mistimed words may get assigned to a segment without a corresponding reference word, or will even be dropped from the hypothesis if they do not fall into any reference segment.

In this work we consider translation from a template file, thus time alignment is equivalent to using subtitle blocks as unit. However, for the transcription task, where subtitle timings of hypothesis and reference are different, we analyze a variant of WER that operates on "t-BLEU segments", i.e. allows for word matches only if hypothesis and reference word are aligned in time (according to interpolated hypothesis word timings). We refer to this variant as t-WER.

### 3.4 New Metric: Subtitle Edit Rate (SubER)

None of the above-mentioned metrics considers *all* of the relevant information present in a subtitle file, namely subtitle text, line segmentation and timing. We therefore propose a new metric called subtitle edit rate (SubER) that attempts to cover all these aspects, and on top avoids segmentation of the subtitle files into aligned hypothesis-reference pairs as a pre-processing step.

We choose TER (Snover et al., 2006) as the basis of SubER because of its interpretability, especially in the case of post-editing. It corresponds to the number of edit operations, namely substitutions, deletions, insertions and shifts of words that are required to turn the hypothesis text into the reference. Also, it allows for easy integration of segmentation and timing information by extending it with break edit operations and time-alignment constraints.

We define the SubER score to be the minimal possible value of (read "#" as "number of"):

$$\mathrm{SubER} = \frac{\text{\# word edits} + \text{\# break edits} + \text{\# shifts}}{\text{\# reference words} + \text{\# reference breaks}}$$

where

- a hypothesis word is only regarded as correct (**no edit**) if it is part of a subtitle that overlaps in time with the subtitle containing the matching reference word (otherwise edits are required, e.g. deletion + insertion).
- **word edits** are insertions, deletions and substitutions of words, substitutions being only allowed if the hypothesis and reference word are from subtitles that overlap in time.
- **break edits** are insertions, deletions and substitutions of breaks, treated as additional tokens (`<eol>` and `<eob>`) inserted at the positions of the breaks. Substitutions are only allowed between end-of-line and end-of-block, not between a word and a break, and the same time-overlap condition as for word substitution applies.
- **shifts** are movements of one or more adjacent hypothesis tokens to a position of a matching phrase in the reference. Only allowed if all the shifted words come from a hypothesis subtitle that overlaps in time with the subtitle of the matching reference word. The shifted phrase may consist of any combination of words and break tokens.

We only consider subtitle timings present in the subtitle files, as opposed to interpolating timings of words as done by Cherry et al. (2021). This avoids hypothesis words "falling off the edges" of reference subtitles, e.g. in case the hypothesis subtitle

```
For the champagne <eol> and brandy you bought me. <eob>
As I recall, the booze put you <eol> to sleep a little prematurely. <eob>
Ladies and gentlemen, <eol> the dance is about to begin. <eob>
```

Figure 2: Example for usage of end-of-line (<eol>) and end-of-block tokens (<eob>) to represent subtitle formatting. Corresponds to right subtitle from Figure 1. Symbols are adopted from Karakanta et al. (2020b).

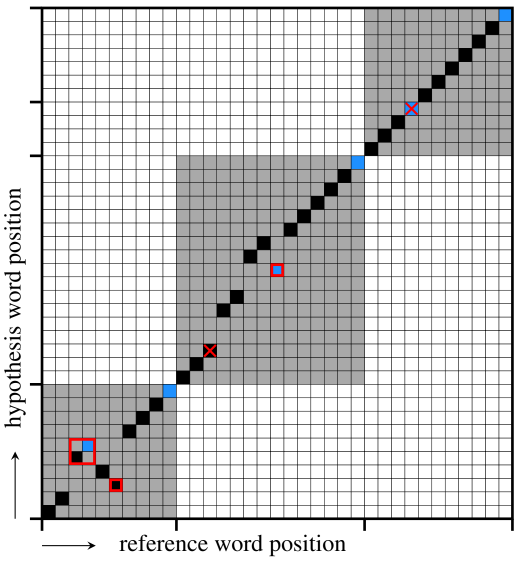


Figure 3: Visualization of SubER applied to the subtitles from Figure 1 (hypothesis left, reference right). Ticks on the axes indicate subtitle block boundaries. Grey areas show regions of time-overlapping reference and hypothesis subtitles. Word matches, substitutions and shifts are allowed only within those areas. Black squares represent word alignments, blue squares represent break token alignments. Red borders mark shifted phrases, red crosses indicate substitutions. 35 reference words (including breaks), 3 insertions, 2 substitutions, 3 shifts lead to a SubER score of $(3 + 2 + 3)/35 = 22.86\%$.

starts a fraction of a second early. It also prevents alignment errors originating from the assumption that all words have the same duration.

The time-overlap condition can be thought of as constraining the search space for Levenshtein-distance calculation. Figure 3 visualizes this for the subtitles from Figure 1. In the white areas no word matches are allowed, this can be exploited for an efficient implementation. The last two hypothesis subtitles overlap with the last reference subtitle and therefore form a single time-aligned region. The shifted 2-word phrase in the bottom left region is "champagne <eol>", showcasing that words and breaks can be shifted in a single operation. In the center region we see the substitution of "recall" with "remember", the inserted (i.e. unaligned) hypothesis words "it", "was" and "that", and a shift of the line break to a different position. The break substitution in the upper right region corresponds to the fact that the last block of the right subtitles in Figure 1 is split into two, i.e. end-of-line is replaced by end-of-block.

### 3.4.1 Implementation Details

We modify the TER implementation of SacreBLEU (Post, 2018) to implement SubER. We adopt the approximation of greedily searching for the best shift until no further reduction of the edit distance can be achieved (Snover et al., 2006). Break tokens (<eol> and <eob>) are inserted into the input text. String comparisons between hypothesis and reference words are replaced by a function additionally checking the time-overlap condition. To make SubER calculation feasible for large subtitle files we split hypothesis and reference into parts at time positions where both agree that no subtitle is displayed. The number of edit operations is then added up for all parts. By definition this does not affect the metric score, in contrast to e.g. segmenting into sentence vs. subtitle blocks when calculating BLEU (Section 3.1).

## 4 Human Evaluation

To analyze the expressiveness of SubER we conduct a human post-editing experiment on both subtitles automatically generated from audio, as well as automatic translations of subtitle text files. For each of the two post-editing tasks we employ three professional subtitlers with multiple years of experience in the subtitling industry. We evaluate how well automatic metric scores correlate with their post-editing effort and their MT quality judgements.

There exists previous work measuring the productivity gains from post-editing automatic subtitles under the aspect of MT quality (Etchegoyhen et al., 2014; Bywood et al., 2017; Koponen et al., 2020) and segmentation quality (Álvarez et al., 2016; Alvarez et al., 2017; Matusov et al., 2019), but to the best of our knowledge we conduct the first study with the goal of evaluating an automatic quality metric for subtitling.

### 4.1 Data

We perform our experiment using one episode from each of the following shows:

- *Master of None:* a comedy-drama series
- *Midnight Mass:* a supernatural horror series
- *Peaky Blinders:* an early 20th century British gangster drama

Each of the three videos has a duration of approximately 55 minutes. They are originally in English, for translation we choose Spanish as the target language. We use pre-existing English subtitles as template files for human translation, and also as the reference when scoring automatic transcriptions. Pre-existing Spanish subtitles, which follow the English template, are used as reference for MT output.

To gather data points for which we can compare post-editing effort with automatic scores, we manually split the videos into segments of roughly 1 minute, each containing 15 subtitle blocks and 103 words on average. We keep the first 15 minutes of each video as one large segment where we measure baseline speed of the subtitlers. Excluding these, we end up with 35, 38 and 37 segments for the videos, respectively, amounting to a total of 110 source-target reference subtitle pairs.

### 4.2 Automatic Subtitling Systems

For human post-editing, we create automatic English and Spanish subtitle files. We use several different subtitling systems to obtain evaluation data with a wider variety. The systems differ in ASR/MT, punctuation and segmentation quality.

We create a single automatic English and Spanish subtitle file for each video, each containing segments coming from different automatic subtitling systems. The subtitlers did not know about any of the details on how these files were created to avoid any bias.

#### 4.2.1 Transcription Systems

To create automatic English subtitles from the audio track of the video we use three different systems:

1. A hybrid ASR system, the output of which is punctuated and cased by a bi-directional LSTM model and then split into lines and subtitles using a beam search decoder that combines scores of a neural segmentation model and hard subtitling constraints, based on the algorithm proposed by Matusov et al. (2019);
2. same as 1., but without using a neural model for subtitle segmentation;
3. an online provider offering automatic transcription in SRT format.

We transcribe an equal number of video segments with each of the three systems and combine them into a single subtitle file which is delivered to the subtitlers for post-editing. The first segment of 15 minutes is not transcribed automatically. Instead, the subtitlers are asked to transcribe it from scratch to measure their baseline productivity.

#### 4.2.2 Translation Systems

To create Spanish subtitles we translate the pre-existing English subtitles with 5 different systems:

1. A Transformer-based MT system, the output of which is split into lines and subtitles using a neural segmentation model and hard subtitling constraints;
2. same as 1., but without using a neural model for subtitle segmentation;
3. same as 1., but with additional inputs for length control and genre, similarly to the systems proposed in (Schioppa et al., 2021; Matusov et al., 2020);
4. an LSTM-based MT system with lower quality than 1., but also using the neural segmentation model;
5. an online provider offering subtitle translation in SRT format.

Also here, we distribute the video segments among the systems such that each system contributes a roughly equal portion of the assembled MT subtitle file delivered to the translators. We extract full sentences from the source subtitle file based on punctuation before translation. The first 15 minute segment of each video is translated directly from the source template without access to MT output to measure baseline productivity of the translators.

### 4.3 Methodology

#### 4.3.1 Productivity Gain Measurement

For both transcription and translation, we ask the subtitlers to measure the time $t_n$ (in minutes) spent to post-edit each of the 110 video segments. As a measure of post-editing productivity $P_n$ we compute the number of subtitles $S_n$ created per minute of work for the $n$-th segment:

$$P_n = \frac{S_n}{t_n} \tag{1}$$

To make these values comparable between subtitlers we normalize them using the subtitler's baseline speed $P_{\text{base}}$. It is computed by averaging the productivity in the first 15-minute segment $P_1$, where the subtitlers work from scratch, over all three videos. Finally, we average the normalized productivities across the three subtitlers $h = 1, 2, 3$ per task to get an average post-editing productivity gain for segment $n$:

$$\hat{P}_n = \frac{1}{3} \sum_{h=1}^{3} \frac{P_{n,h}}{P_{\text{base},h}} \tag{2}$$

To evaluate the expressiveness of a given metric we compute the Spearman's rank correlation coefficient $r_s$ between the per-segment metric scores and $\hat{P}_n$ for all segments of all three videos. We choose Spearman's correlation in favour of Pearson's correlation because subtitle quality varies a lot for different video segments and different systems, and we don't expect the metrics to behave linearly in this range.

### 4.3.2 Direct Assessment

For the translation task we additionally gather direct assessment scores for each segment. For this we ask the translators to give two scores (referred to as $U_n$ and $Q_n$, respectively) according to the following descriptions:

1. "Rate the overall **usefulness** of the automatically translated subtitles in this segment for post-editing purposes on a scale from 0 (completely useless) to 100 (perfect, not a single change needed)."
2. "Rate the overall **quality** of the automatically translated subtitles in this segment as perceived *by a viewer* on a scale from 0 (completely incomprehensible) to 100 (perfect, completely fluent and accurate). The score should reflect how well the automatic translation conveys the semantics of the original subtitles, and should also reflect how well the translated subtitles are formatted."

These scores are standardized into $z$-scores by subtracting the average and dividing by the standard deviation of scores per translator. Finally, we average the $z$-scores across the three translators to get expected usefulness and quality assessment scores for each segment, which we will refer to as $\hat{U}_n$ and $\hat{Q}_n$, respectively.

## 4.4 Results

### 4.4.1 Post-Editing of English Transcription

The baseline productivities $P_{\text{base}}$ of the three subtitlers A, B and C when transcribing the first 15 minutes of each video from scratch are 3.4, 2.8 and 2.7 subtitles per minute of work, respectively. Post-editing changes their productivities to 3.9, 2.6 and 3.1 subtitles per minute on average for the other segments, meaning subtitlers A and C work faster when post-editing automatic subtitles, while subtitler B does not benefit from them.

Table 1 shows the analysis of the correlation between automatic metric scores and productivity gains, calculated for each of the 110 one-minute video segments. Word error rate (WER) can predict the averaged productivity gain $\hat{P}_n$ with a Spearman's correlation of $-0.676$. This confirms the natural assumption that the more words the ASR system recognized correctly in a given segment, the less time is necessary for post-editing. Subtitler A's post-editing gains are more predictable than those of the other two subtitlers. This indicates that the subtitlers have different workflows and do not make use of the automatic subtitles with the same consistency.

Row 2 shows that making WER case-sensitive and keeping punctuation marks as part of the words does not improve correlation consistently. Although we believe that casing and punctuation errors harm subtitle quality, these errors might not have a significant impact on post-editing time because correcting them requires changing single characters only. Row 3 shows that extending the original WER definition by simply inserting end-of-line and end-of-block tokens into the text does not lead to improvements either. This can be explained by the fact that the original WER algorithm allows for substitution of break symbols with words. Such substitutions have no meaningful interpretation. Also, it does not support shifts of break symbols, which leads to breaks at wrong positions being punished more than completely missing ones.

Our proposed metric SubER achieves the overall best correlation of $-0.692$. We attribute this in part to a proper way of handling segmentation information: without it, as shown in the last row

| Metric | Subtitler A | Subtitler B | Subtitler C | Combined |
|---|---|---|---|---|
| WER | -0.731 | -0.494 | -0.499 | -0.676 |
| + case/punct | -0.671 | **-0.512** | -0.509 | -0.650 |
| + break tokens | -0.725 | -0.494 | -0.512 | -0.678 |
| t-WER | -0.661 | -0.440 | -0.476 | -0.625 |
| TER-br | -0.573 | -0.489 | -0.434 | -0.562 |
| SubER (ours) | **-0.746** | -0.506 | **-0.517** | **-0.692** |
| + case/punct | -0.670 | -0.507 | -0.500 | -0.645 |
| - break tokens | -0.741 | -0.495 | -0.502 | -0.682 |

Table 1: Spearman's correlation $r_s$ between automatic metric scores and post-editing productivity gains $P_n$ on all 110 video segments for the **English transcription task**. The last column shows correlation to the productivity gain averaged across subtitlers $\hat{P}_n$.

of Table 1, the correlation is lower. Unfortunately, for the same reasons as for the case of WER, we have to apply SubER to lower-cased text - as it is the default setting for the TER metric - to avoid a drop in correlation.

Correlations for t-WER (see Section 3.3) suggest that a word-level time-alignment using interpolation may result in misalignments which are punished too harsh in comparison to which mistimings are still tolerated by the post-editors. This supports our design choice of using subtitle-level timings for SubER.

Finally, we include TER-br from Karakanta et al. (2020a) in the results. It is a variant of TER + break tokens where each real word is replaced by a mask token. Given that the metric has no access to the actual words it achieves surprisingly high correlations. This shows that the subtitle formatting defined by the number of subtitle blocks, number of lines and number of words per line is in itself an important feature affecting the post-editing effort.

### 4.4.2 Post-Editing of Spanish Translation

Baseline productivities $P_{\text{base}}$ of the translators D, E and F are 1.9, 1.8 and 1.1 subtitles per minute, respectively. On average, their productivity changes to 1.6, 2.0 and 1.1 when post-editing, meaning only subtitler B gains consistently. Subtitler A is more productive on one of the videos, but slows down significantly for the other two.

Table 2 shows performances of the different MT metrics. In addition to post-edit effort, we show how well the metrics agree with human judgments of the usefulness and quality (see Section 4.3.2) for each of the 110 one-minute video segments.

Overall, the correlation of productivity gains is much lower than for the transcription task. This can be explained by the fact that a translator has more freedom than a transcriber. The translator's word choices are influenced by clues outside the scope of the translated text, like the style of language and references to other parts of the plot. Sometimes even research is required (e.g. bible verses for *Midnight Mass*). Despite this, the subjectively perceived usefulness $\hat{U}_n$ of the automatic subtitles for post-editing can be predicted from automatic scores with a Spearman's correlation of up to $-0.591$. The quality judgement $\hat{Q}_n$ shows even higher correlations of up to $0.659$.

We compare the baseline MT metrics BLEU and TER when applied to the subtitle block-level vs. the sentence-level. We note that BLEU on subtitle-level is identical to t-BLEU (Cherry et al., 2021) for the considered case of template translation, where timestamps in hypothesis and reference are identical. Overall, BLEU and TER perform similarly. For both, evaluation on subtitle-level outperforms evaluation on sentence-level. This is because the sentence-pairs extracted from the subtitle files preserve no formatting information, while using subtitle blocks as units is sensitive to how words of a sentence are distributed among subtitles after translation, especially in case of word re-ordering.

Extending BLEU and TER with break tokens to take subtitle segmentation into account shows only minor improvements for the subtitle-level, but significantly improves correlations for the sentence-level. This could be attributed to the extended context after end-of-block tokens that is not available for scoring on subtitle-level. Especially the way "BLEU + break tokens" punishes n-grams that are disrupted by an erroneous line break seems to lead to good results.

Our proposed metric SubER consistently outperforms all considered baseline metrics except for sentence-level BLEU with break tokens, which has a higher correlation for $\hat{Q}_n$ and for the scores given by subtitler F. For this subtitler we also observe

| **Metric** | **Subtitler D** | | | **Subtitler E** | | | **Subtitler F** | | | **Combined** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $P_n$ | $U_n$ | $Q_n$ | $P_n$ | $U_n$ | $Q_n$ | $P_n$ | $U_n$ | $Q_n$ | $\hat{P}_n$ | $\hat{U}_n$ | $\hat{Q}_n$ |
| **Subtitle-level** | | | | | | | | | | | | |
| BLEU | 0.03 | 0.34 | 0.52 | 0.22 | 0.21 | 0.39 | 0.07 | 0.58 | 0.49 | 0.172 | 0.541 | 0.595 |
| + break tokens | 0.04 | 0.35 | 0.53 | 0.22 | 0.24 | 0.43 | 0.12 | 0.58 | 0.46 | 0.210 | 0.554 | 0.595 |
| TER | 0.03 | -0.35 | -0.54 | -0.22 | -0.23 | -0.41 | -0.11 | -0.63 | -0.51 | -0.182 | -0.554 | -0.618 |
| + break tokens | 0.00 | -0.36 | -0.54 | -0.23 | -0.24 | -0.41 | -0.10 | -0.61 | -0.50 | -0.200 | -0.558 | -0.606 |
| **Sentence-level** | | | | | | | | | | | | |
| BLEU | -0.03 | 0.31 | 0.51 | 0.21 | 0.13 | 0.33 | 0.04 | 0.60 | 0.51 | 0.126 | 0.494 | 0.573 |
| + break tokens | 0.02 | 0.35 | 0.55 | 0.25 | 0.22 | 0.43 | 0.16 | 0.63 | **0.55** | 0.240 | 0.583 | **0.659** |
| TER | 0.07 | -0.32 | -0.52 | -0.22 | -0.14 | -0.34 | -0.07 | -0.59 | -0.48 | -0.133 | -0.484 | -0.559 |
| + break tokens | 0.00 | -0.36 | -0.55 | -0.25 | -0.19 | -0.38 | -0.13 | -0.58 | -0.45 | -0.218 | -0.515 | -0.574 |
| chrF | -0.09 | 0.26 | 0.52 | 0.21 | 0.10 | 0.28 | 0.04 | 0.64 | 0.51 | 0.104 | 0.483 | 0.556 |
| TER-br | 0.03 | -0.32 | -0.42 | -0.11 | -0.07 | -0.24 | -0.13 | -0.43 | -0.40 | -0.137 | -0.345 | -0.426 |
| SubER (ours) | -0.06 | **-0.38** | **-0.57** | **-0.27** | **-0.28** | **-0.47** | -0.16 | -0.61 | -0.52 | **-0.274** | **-0.591** | -0.651 |
| + case/punct | 0.00 | -0.36 | -0.56 | -0.25 | -0.23 | -0.42 | -0.15 | -0.61 | -0.49 | -0.237 | -0.554 | -0.612 |
| - break tokens | 0.02 | -0.34 | -0.54 | -0.24 | -0.25 | -0.44 | -0.11 | **-0.65** | **-0.55** | -0.197 | -0.572 | -0.645 |

Table 2: Spearman's correlation $r_s$ between automatic metric scores and $P_n$, $U_n$ and $Q_n$ on all 110 video segments for the **English→Spanish translation task**. $P_n$ are segment-wise productivity gains from post-editing measured in subtitles per minute of work. $U_n$ and $Q_n$ are segment-wise usefulness and quality scores, respectively, which the subtitlers assigned to the automatically generated subtitle segments.

that calculating SubER without break tokens improves results. In fact, subtitler F stated that moving around text is not a taxing procedure for him as he is very proficient with keyboard commands. For the other subtitlers, break tokens as part of the metric are shown to have a clear positive effect.

### 4.4.3 System-level Results

For both transcription and translation we have a pair of systems which differ only in subtitle segmentation (systems 1 and 2). We expect the system using a neural segmentation model to perform better overall. By definition, WER cannot distinguish between the transcription systems, scores for both are 40.6, 14.2 and 29.5 (%) for the three videos *Master of None*, *Midnight Mass* and *Peaky Blinders*, respectively. (High WER on *Master of None* is caused by colloquial and mumbling speech.) SubER scores for system 1 are 46.4, 20.3 and 33.1, for system 2 they are 47.3, 22.1 and 34.7. This means, for all videos SubER scores are able to reflect the better segmentation quality of system 1.

The same is true for translation: sentence-level BLEU scores are the same for systems 1 and 2, namely 18.9, 26.7 and 37.9 for the three videos. SubER scores for the system with neural segmentation are 65.1, 56.5 and 41.8, whereas the system without it gets worse scores of 67.4, 60.5 and 46.9.

## 5 Release of Code

We release the code to calculate the SubER metric as part of an open-source subtitle evaluation toolkit[2] to encourage its use in the research community as well as the media industry and to further promote research of automatic subtitling systems.

In addition to SubER, the toolkit implements all baseline metrics used in Table 1 and 2, as well as t-BLEU (Cherry et al., 2021). This includes implementations of hypothesis to reference alignment via the Levenshtein algorithm (Section 3.1) or via interpolated word timings (Section 3.3). We use the JiWER[3] Python package for word error rate calculations and SacreBLEU (Post, 2018) to compute BLEU, TER and chrF values.

All metrics can be calculated directly from SRT input files. Support for other subtitle file formats will be added on demand.

## 6 Conclusion

In this work, we proposed SubER – a novel metric for evaluating quality of automatically generated intralingual and interlingual subtitles. The metric is based on edit distance with shifts, but considers not only the automatically transcribed or translated text, but also subtitle timing and line segmentation information. It can be used to compare an automatically generated subtitle file to a human-generated one even if the two files contain a different number of subtitles with different timings.

A thorough evaluation by professional subtitlers confirmed that SubER correlates well with their transcription post-editing effort and direct assessment scores of translations. In most cases, SubER

[2] `https://github.com/apptek/SubER`

[3] `https://github.com/jitsi/jiwer`

shows highest correlation as compared to metrics that evaluate either the quality of the text alone, or use different approaches to integrate subtitle timing and segmentation information.

The source code for SubER will be publicly released for the benefit of speech recognition and speech translation research communities, as well as the media and entertainment industry.

## References

Aitor Álvarez, Marina Balenciaga, Arantza del Pozo, Haritz Arzelus, Anna Matamala, and Carlos-D. Martínez-Hinarejos. 2016. Impact of automatic segmentation on the quality, productivity and self-reported post-editing effort of intralingual subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3049–3053, Portorož, Slovenia. European Language Resources Association (ELRA).

Aitor Alvarez, Carlos-D Martínez-Hinarejos, Haritz Arzelus, Marina Balenciaga, and Arantza del Pozo. 2017. Improving the automatic segmentation of subtitles through conditional random field. *Speech Communication*, 88:83–95.

Lindsay Bywood, Panayota Georgakopoulou, and Thierry Etchegoyhen. 2017. Embracing the threat: machine translation as a solution for subtitling. *Perspectives*, 25(3):492–508.

Colin Cherry, Naveen Arivazhagan, Dirk Padfield, and Maxim Krikun. 2021. Subtitle translation as markup translation. *Proc. Interspeech 2021*, pages 2237–2241.

Thierry Etchegoyhen, Lindsay Bywood, Mark Fishel, Panayota Georgakopoulou, Jie Jiang, Gerard van Loenhout, Arantza del Pozo, Mirjam Sepesy Maučec, Anja Turner, and Martin Volk. 2014. Machine translation for subtitling: A large-scale evaluation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 46–53, Reykjavik, Iceland. European Language Resources Association (ELRA).

Alina Karakanta, Matteo Negri, and Marco Turchi. 2020a. Is 42 the answer to everything in subtitling-oriented speech translation? In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 209–219, Online. Association for Computational Linguistics.

Alina Karakanta, Matteo Negri, and Marco Turchi. 2020b. MuST-cinema: a speech-to-subtitles corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3727–3734, Marseille, France. European Language Resources Association.

Alina Karakanta, Matteo Negri, and Marco Turchi. 2020c. Point break: Surfing heterogeneous data for subtitle segmentation. In *CLiC-it*.

Maarit Koponen, Umut Sulubacak, Kaisa Vitikainen, and Jörg Tiedemann. 2020. MT for subtitling: User evaluation of post-editing productivity. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 115–124, Lisboa, Portugal. European Association for Machine Translation.

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005. Evaluating machine translation output with automatic sentence segmentation. In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

Evgeny Matusov, Patrick Wilken, and Yota Georgakopoulou. 2019. Customizing neural machine translation for subtitling. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 82–93, Florence, Italy. Association for Computational Linguistics.

Evgeny Matusov, Patrick Wilken, and Christian Herold. 2020. Flexible customization of a single neural machine translation system with multi-dimensional metadata inputs. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, pages 204–216, Virtual. Association for Machine Translation in the Americas.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

J. Pedersen. 2017. The FAR model: assessing quality in interlingual subtitling. In *Journal of Specialized Translation*, volume 18, pages 210–229.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

P. Romero-Fresco and F. Pöchhacker. 2017. Quality assessment in interlingual live subtitling: The NTR model. In *Linguistica Antverpiensia, New Series:*

*Themes in Translation Studies*, volume 16, pages 149–167.

P. Romero-Fresco and J.M. Pérez. 2015. Accuracy rate in live subtitling: The NER model. In *Audiovisual Translation in a Global Context. Palgrave Studies in Translating and Interpreting*. R.B., Cintas J.D. (eds), Palgrave Macmillan, London.

Andrea Schioppa, David Vilar, Artem Sokolov, and Katja Filippova. 2021. Controlling machine translation for multiple attributes with additive interventions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6676–6696, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. 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. Association for Machine Translation in the Americas.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.