

Evaluating NLP Text: A small Practical Guide to WER Metrics with Hugging Face

Metric: wer

Word error rate (WER) is a common metric of the performance of an automatic speech recognition system.

The general difficulty of measuring performance lies in the fact that the recognized word sequence can have a different length from the reference word sequence (supposedly the correct one). The WER is derived from the Levenshtein distance, working at the word level instead of the phoneme level. The WER is a valuable tool for comparing different systems as well as for evaluating improvements within one system. This kind of measurement, however, provides no details on the nature of translation errors and further work is therefore required to identify the main source(s) of error and to focus any research effort.

This problem is solved by first aligning the recognized word sequence with the reference (spoken) word sequence using dynamic string alignment. Examination of this issue is seen through a theory called the power law that states the correlation between perplexity and word error rate.

Word error rate can then be computed as:

$$\text{WER} = (S + D + I) / N = (S + D + I) / (S + D + C)$$

where

S is the number of substitutions, D is the number of deletions, I is the number of insertions, C is the number of correct words, N is the number of words in the reference ($N=S+D+C$).

This value indicates the average number of errors per reference word. The lower the value, the better the performance of the ASR system with a WER of 0 being a perfect score.

Word error rate can then be computed as:

$$\text{WER} = (S + D + I) / N = (S + D + I) / (S + D + C)$$

where

S is the number of substitutions,

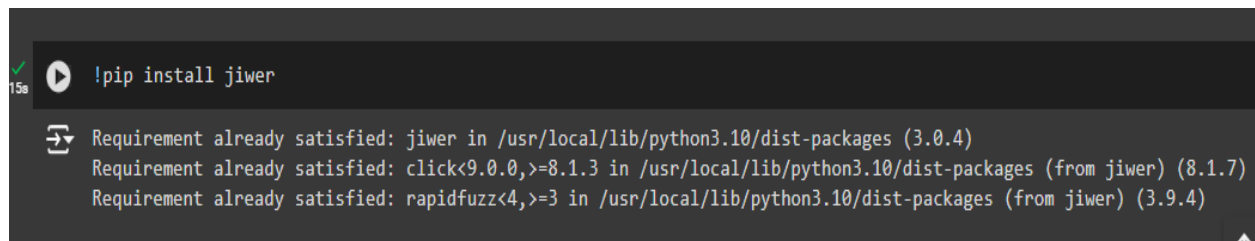
D is the number of deletions,

I is the number of insertions,

C is the number of correct words,

N is the number of words in the reference ($N=S+D+C$).

How to use:



```
15s !pip install jiwer
Requirement already satisfied: jiwer in /usr/local/lib/python3.10/dist-packages (3.0.4)
Requirement already satisfied: click<9.0.0,>=8.1.3 in /usr/local/lib/python3.10/dist-packages (from jiwer) (8.1.7)
Requirement already satisfied: rapidfuzz<4,>=3 in /usr/local/lib/python3.10/dist-packages (from jiwer) (3.9.4)
```

The metric takes two inputs: references (a list of references for each speech input) and predictions (a list of transcriptions to score).

```
from evaluate import load
wer = load("wer")
wer_score = wer.compute(predictions=predictions, references=references)
```

Output values

This metric outputs a float representing the word error rate.

```
print(wer_score)
0.5
```

This value indicates the average number of errors per reference word.

The **lower** the value, the **better** the performance of the ASR system, with a WER of 0 being a perfect score.

Examples

Perfect match between prediction and reference:

```
[2] from evaluate import load
    wer = load("wer")
    predictions = ["hello world", "good night moon"]
    references = ["hello world", "good night moon"]
    wer_score = wer.compute(predictions=predictions, references=references)
    print(wer_score)
```

0.0

Partial match between prediction and reference:

```
from evaluate import load
wer = load("wer")
predictions = ["this is the prediction", "there is an other sample"]
references = ["this is the reference", "there is another one"]
wer_score = wer.compute(predictions=predictions, references=references)
print(wer_score)
```

0.5

No match between prediction and reference:

```
[4] from evaluate import load
    wer = load("wer")
    predictions = ["hello world", "good night moon"]
    references = ["hi everyone", "have a great day"]
    wer_score = wer.compute(predictions=predictions, references=references)
    print(wer_score)
```

↔ 1.0

Limitations and bias

WER is a valuable tool for comparing different systems as well as for evaluating improvements within one system. This kind of measurement, however, provides no details on the nature of translation errors and further work is therefore required to identify the main source(s) of error and to focus any research effort.

References: <https://huggingface.co/spaces/evaluate-metric/wer>, <https://www.clari.com/blog/word-error-rate/#:~:text=Put%20simply%2C%20WER%20is%20the,are%20a%20little%20more%20nuanced>, <https://www.rev.com/blog/resources/what-is-wer-what-does-word-error-rate-mean>.