Evaluation protocol

Definition and examples

Herman Kamper

Passage and grid evaluations

```
can: ga u la e ka ha
ref: ga ga u e ka hi hi
hyp: ga u la i ka hi ho
```

- ASR evaluation: Word error rate (WER) (`wer_ref_hyp`)
- EGRA MAE: Mean absolute error (MAE) between `EGRA_ACC` and `ASR_EGRA_ACC` (call this `MAE_EGRA_ACC`)
  - Bias baseline
- Finer grained (explained below):
  - Mistake error rate (MER)
  - P, R, F1 for substitution, insertion, deletion and all mistakes

Complete example

```
can: ga u la e ka ha
ref: ga ga u e ka hi hi
hyp: ga u la i ka hi ho
```

Align reference to canonical (A):

```
can:     ga u  la e  ka    ha
     I   C  C  D  C  C  I   S
ref: ga ga u      e  ka hi hi
```

- `wer_can_ref` $= \frac{S+D+I}{N} = \frac{1+1+2}{6} = \frac{4}{6}$
- `acc_can_ref (EGRA_ACC)` $= \frac{C}{N} = \frac{4}{6}$

```
can: ga u la e ka ha
ref: ga ga u e ka hi hi
hyp: ga u la i ka hi ho
```

Align hypothesis to canonical (B):

```
can: ga u  la e  ka    ha
     C  C  C  S  C  I  S
hyp: ga u  la i  ka hi ho
```

- `acc_can_hyp (ASR_EGRA_ACC)` $= \frac{C}{N} = \frac{4}{6}$

```
can: ga u la e ka ha
ref: ga ga u e ka hi hi
hyp: ga u la i ka hi ho
```

Align hypothesis to reference:

```
ref: ga ga u     e  ka hi hi
     C  D  C  I  S  C  C  S
hyp: ga    u  la i  ka hi ho
```

- $\texttt{wer\_ref\_hyp} = \frac{S+D+I}{N} = \frac{2+1+1}{7} = \frac{4}{7} = 0.5714$

Compare the two predictions:

- $\texttt{MAE\_EGRA\_ACC} = |\ \texttt{acc\_can\_ref} - \texttt{acc\_can\_hyp}\ | = |\frac{4}{6} - \frac{4}{6}| = 0$

Align CSID sequences from A and B:

```
A: ref_to_can: I  C  C  D  C     C  I  S
               D  C  C  D  C  I  C  C  C
B: hyp_to_can:    C  C     C  S  C  I  S
```

- `mer` $= \frac{2+1}{8} = 0.375$
- Substitutions:
  - Precision $= \frac{1}{2}$
  - Recall $= \frac{1}{1}$
- Insertions:
  - Precision $= \frac{1}{1}$
  - Recall $= \frac{1}{2}$
- Deletions:
  - Precision $= \frac{0}{0}$
  - Recall $= \frac{0}{1}$
- All mistakes:
  - Precision $= \frac{2}{3}$
  - Recall $= \frac{2}{4}$

Isolated letters, syllables and non-words

- ASR evaluation: Word error rate (WER) (`wer_ref_hyp`)
- EGRA: Accuracy in correct / incorrect prediction
- Finer-grained:
  - P, R, F1 for correct mistake prediction
  - A mistake has a label 1
  - Scores for a majority baseline (predict everything as either correct or incorrect, whichever occurs most)

Example:

```
can: i
ref: i o
hyp: -
```

Summarising metrics in tables

- Passage and grid reading:
  - ASR WER: `wer_ref_hyp`
  - EGRA:
    - Correlation coefficient r (see next slide) between predicted (hyp–ref) and actual number of correct (can–ref) words per utterance (segment)
    - Scatter plot in code
    - MAE between correct counts
  - Finer grained:
    - Mistake error rate (MER)
    - Substitution P, R, F1
    - Insertion P, R, F1
    - Deletion P, R, F1
    - All mistakes P, R, F1
- Isolated letters, syllables and non-words:
  - ASR WER: `wer_ref_hyp`
  - EGRA accuracy $= \frac{\mathsf{TP}+\mathsf{TN}}{N}$
  - P, R, F1 for correct mistake prediction (label $1 =$ mistake)
  - P, R, F1 for majority baseline

Resources

- Link to notebook
- Correlation code:

```
import numpy as np

r = np.corrcoef(human_wcpm, automated_wcpm)[0, 1]
print("Correlation r:", r)

from sklearn.metrics import r2_score

R2 = r2_score(human_wcpm, automated_wcpm)
print("R2:", R2)

R2_from_r = r**2
print(R2_from_r)
```