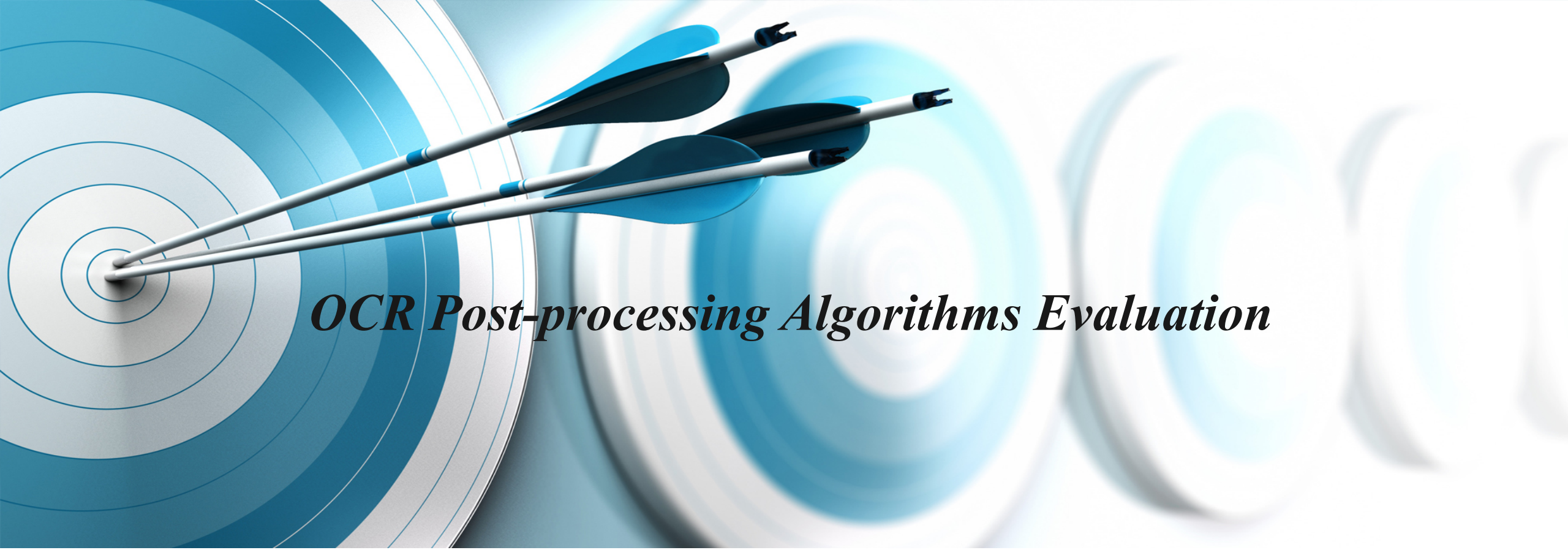


OCR Post-processing Algorithms Evaluation

Group 7

Shen, Yu He, Yuting Ma, Qiaozhen Lin, Nelson Zeng, Yiyang

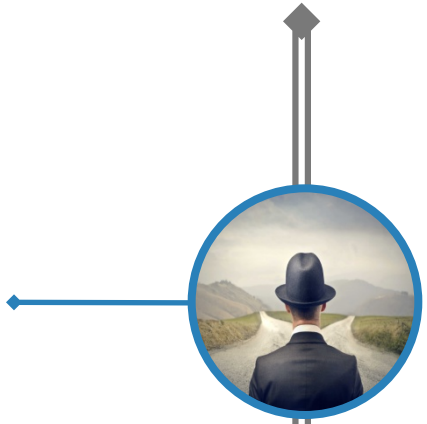
Step1: Pre-processing

Step2: Word recognition

Step3: Post-processing

Post-processing

Error Detection



Error Correction

Error Detection

C1:Rule based

#If the number of punctuation characters in a string is greater than the number of alphanumeric characters, it is garbage.
Example: ?3//la‘

#Ignoring the first and last characters in a string, if there are two or more different punctuation characters in the string, it is garbage.
Example: b?bl@bjk.1e.322

#A string composed of more than 20 characters is garbage. *Example: iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii...*

#If there are three or more identical characters in a row in a string, it is garbage. *Example: aaaaaBlE*

#If the number of uppercase characters in a string is greater than the number of lowercase characters, and if the number of uppercase characters is less than the total number of characters in the string, it is garbage. *Example: BBEYaYYq*

#If all the characters in a string are alphabetic, and if the number of consonants in the string is greater than 8 times the number of vowels in the string, or vice-versa, it is garbage. *Example: jabwqbpP*

#If there are four or more consecutive vowels in the string or five or more consecutive consonants in the string, it is garbage.
Example: buauub

#If the first and last characters in a string are both lowercase and any other character is uppercase, it is garbage.
Example: awwgrapHic

Error Correction

D4: Probability scoring with contextual constraints

Step1 ： Correcting misspelled words

Step2: Calculating scores

Correcting misspelled words

- 4 ways a word is misspelled:
  - **Deletion**: A character is deleted and we can add a character to create a correctly spelled word
  - **Insertion**: A character is inserted and we can delete a character to create a correctly spelled word
  - **Substitution**: A character is substituted and we can substitut a character to create a correctly spelled word
  - **Reversal**: A character is reversed and we can reverse a character to create a correctly spelled word

- **A big limitation of these techniques is that they can only work for words off by one character**

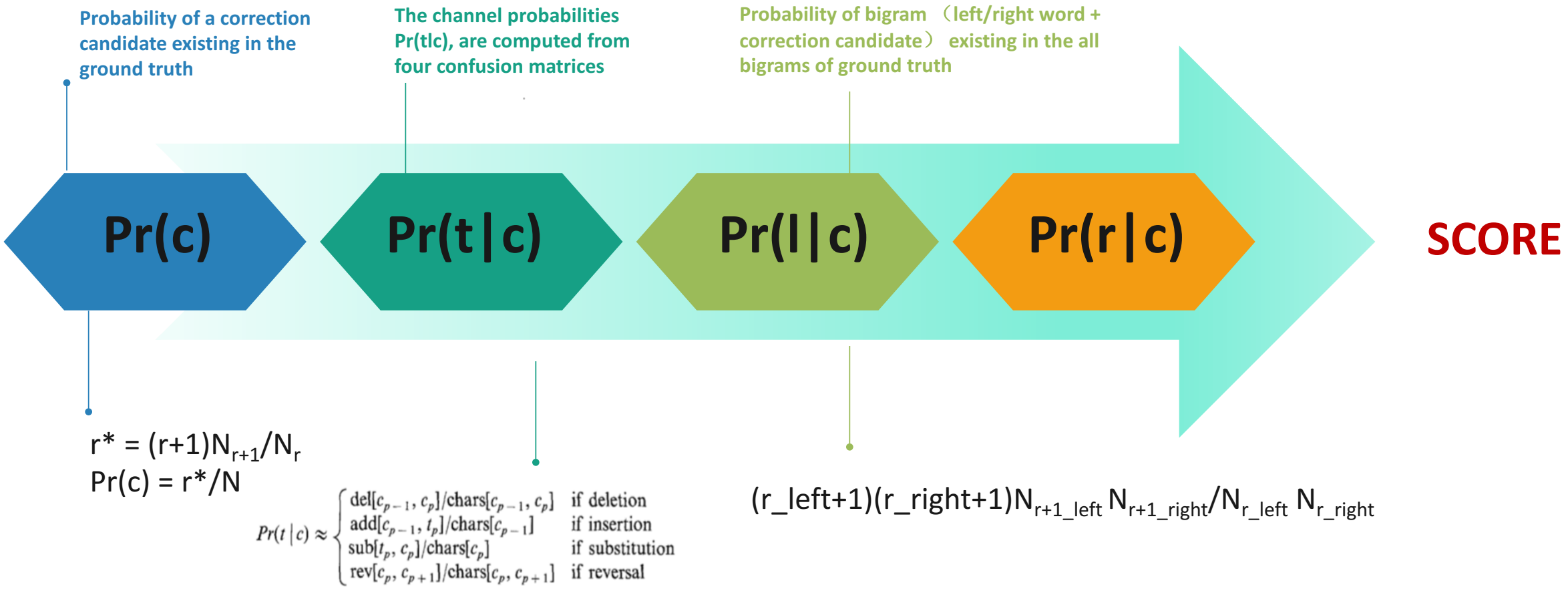
Calculating Scores

MLE The MLE method is uninformative in nearly half of the cases because either Pr(l|c) or Pr(r|c) is zero.

ELE Badly overestimates the probability of a bigram that has not been seen in the training set, and consequently, it is wrong more often than it is right. $r^* = r+0.5$

GT $r^* = (r+1)N_{r+1}/N_r$

Calculating Scores

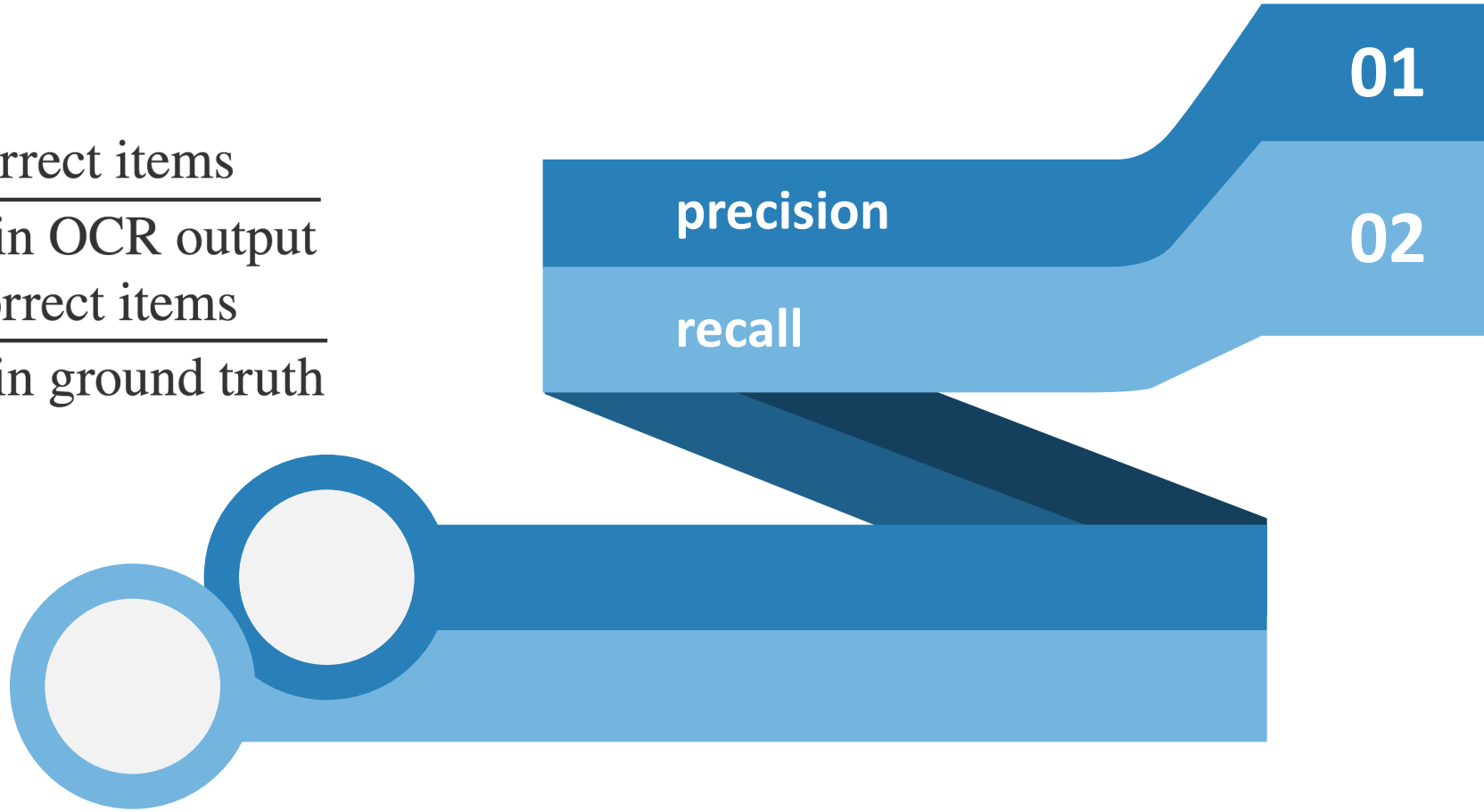


r = word frequencies/bigram frequencies
N = # of words in ground truth
N_r = # of words that has frequency r in ground truth.

Performance Measure

$$\text{precision} = \frac{\text{number of correct items}}{\text{number of items in OCR output}}$$

$$\text{recall} = \frac{\text{number of correct items}}{\text{number of items in ground truth}}$$



01 precision

02 recall

	Tesseract	Tesseract with Post-Processing (GT)
Word-wise recall	0.627530	0.632168
Word-wise precision	0.653725	0.658557
Character-wise recall	0.894004	0.894957
Character-wise precision	0.931114	0.932107

Reason for low recall & precision

01

The way we use to detect errors is not precise, most of the misspelled words we detect have more than one wrong characters..

02

The method we correct words is only useful for those misspelled words that have only one incorrect character

03

So, most of the errors we detect are not mendable using the correction method we are assigned.