CHRF & CHRF++

MT Evaluation Measurement

- Quality (Chatzikoumi 2020)
  - Fluency: grammaticality and naturalness
  - Adequacy: semantic equivalence between source and target
  - Compliance: audiences' need being met

CHRF Introduction

- A metric for machine translation evaluation

- Character-level n-gram F score (usually n = 4)

- $\mathbf{CHRF}\beta = (1+\beta^2)\frac{\mathbf{CHRP}\cdot\mathbf{CHRR}}{\beta^2\cdot\mathbf{CHRP}+\mathbf{CHRR}}$
  - CHRP = precision (P); CHRR = recall (R)
  - CHRF3 is recommended (R is 3x more important than P)
  - β = parameter representing β times more importance assigned to recall than to precision. If $\beta == 1$, then same importance.

Beta Parameter as Weight for Recall

- $\text{CHRF}\beta = (1 + \beta^2) \frac{\text{CHRP} \cdot \text{CHRR}}{\beta^2 \cdot \text{CHRP} + \text{CHRR}}$
- EX: CHRP = 0.8, CHRR = 0.1
  - beta = 1 => CHRF = 0.17
  - beta = 3 => CHRF = 0.11
  - beta = 5 => CHRF = 0.1
- EX: CHRP = 0.1, CHRR = 0.8
  - beta = 1 => CHRF = 0.17
  - beta = 3 => CHRF = 0.47
  - beta = 5 => CHRF = 0.63
- The higher the beta, the lower the importance of CHRP

Why Recall May Be More Important

- beta == 2 or beta ==3 are recommended
- Why recall may be more important
  - Precision does not reflect adequacy
    - y = the cat is on the mat
    - y_hat = the the the the the
    - Precision = 5/5 = 100%; Recall = 2/6 = 33.3%
- Recall does reflect adequacy
  - We want reference to be translated as completely as possible
  - Higher recall means more parts of reference are in the hypothesis

Spearman's Rank Correlation Coefficient

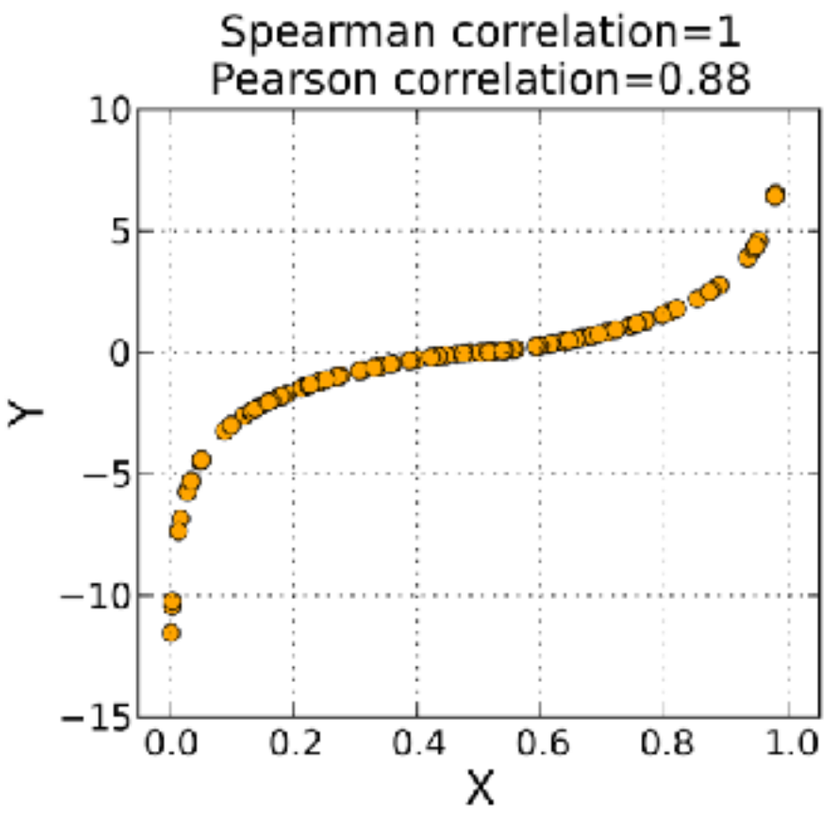
$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)},$$

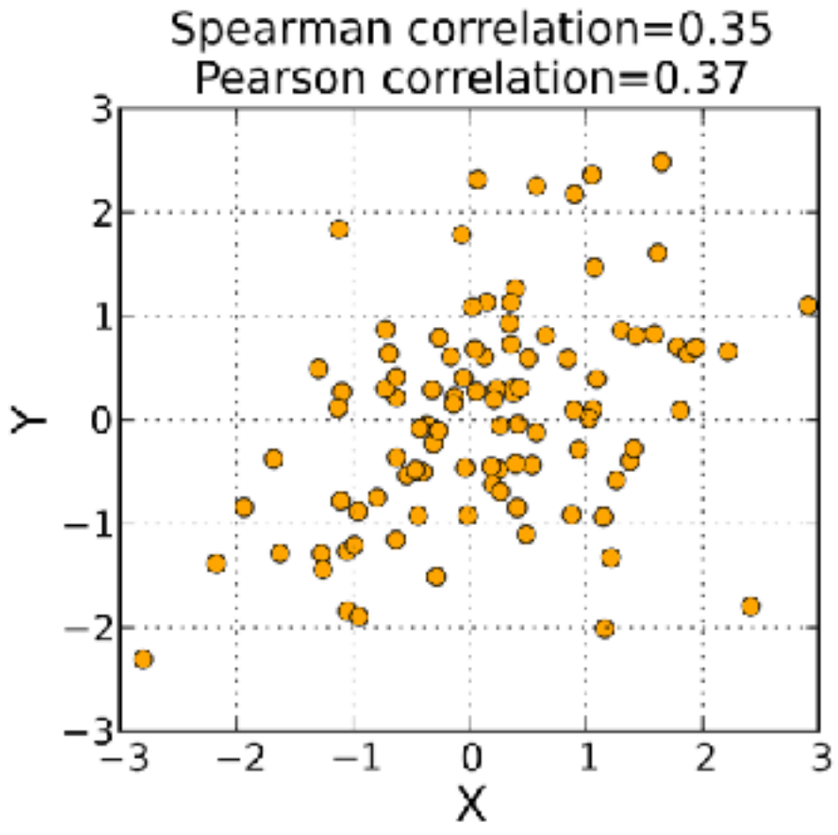
where

$$d_i = \mathrm{rg}(X_i) - \mathrm{rg}(Y_i)$$

year	WORDF	CHRF	CHRF3	BLEU	TER	METEOR
2014 (r)	0.810	0.805	0.857	0.845	0.814	0.822
2013 (ρ)	0.874	0.873	/	0.835	0.791	0.876
2012 (ρ)	0.659	0.696	/	0.671	0.682	0.690

Spearman's Rank Correlation Coefficient

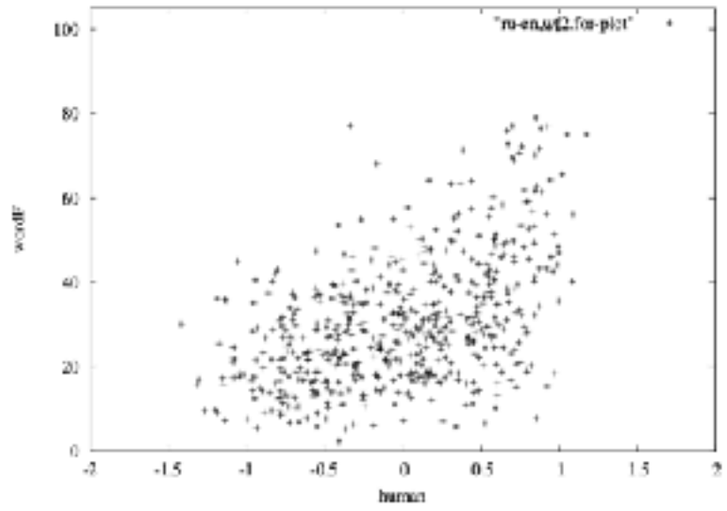




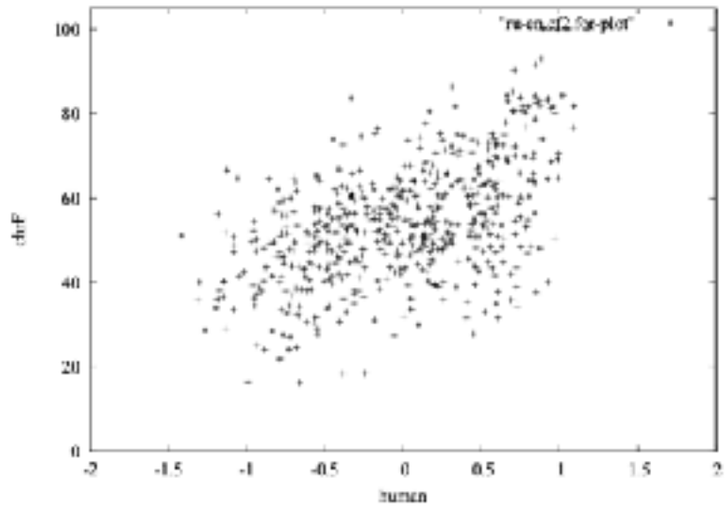
A Spearman correlation of 1 results when the two variables being compared are monotonically related, even if their relationship is not linear

CHRF++

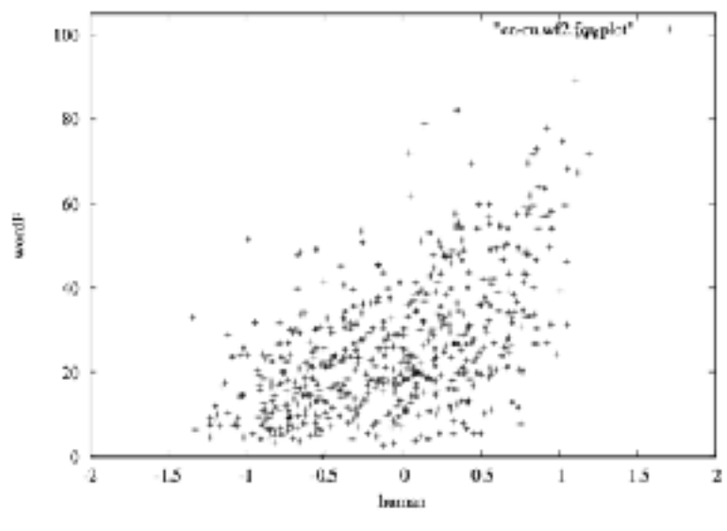
- Average of CHRF and WordF
  - Where WordF is word-level n-gram F score



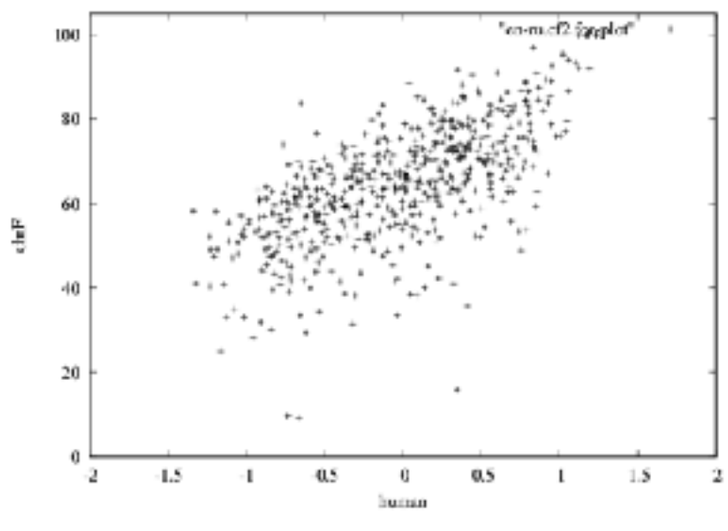
(a) Russian→English, WORDF



(c) Russian→English, CHRF

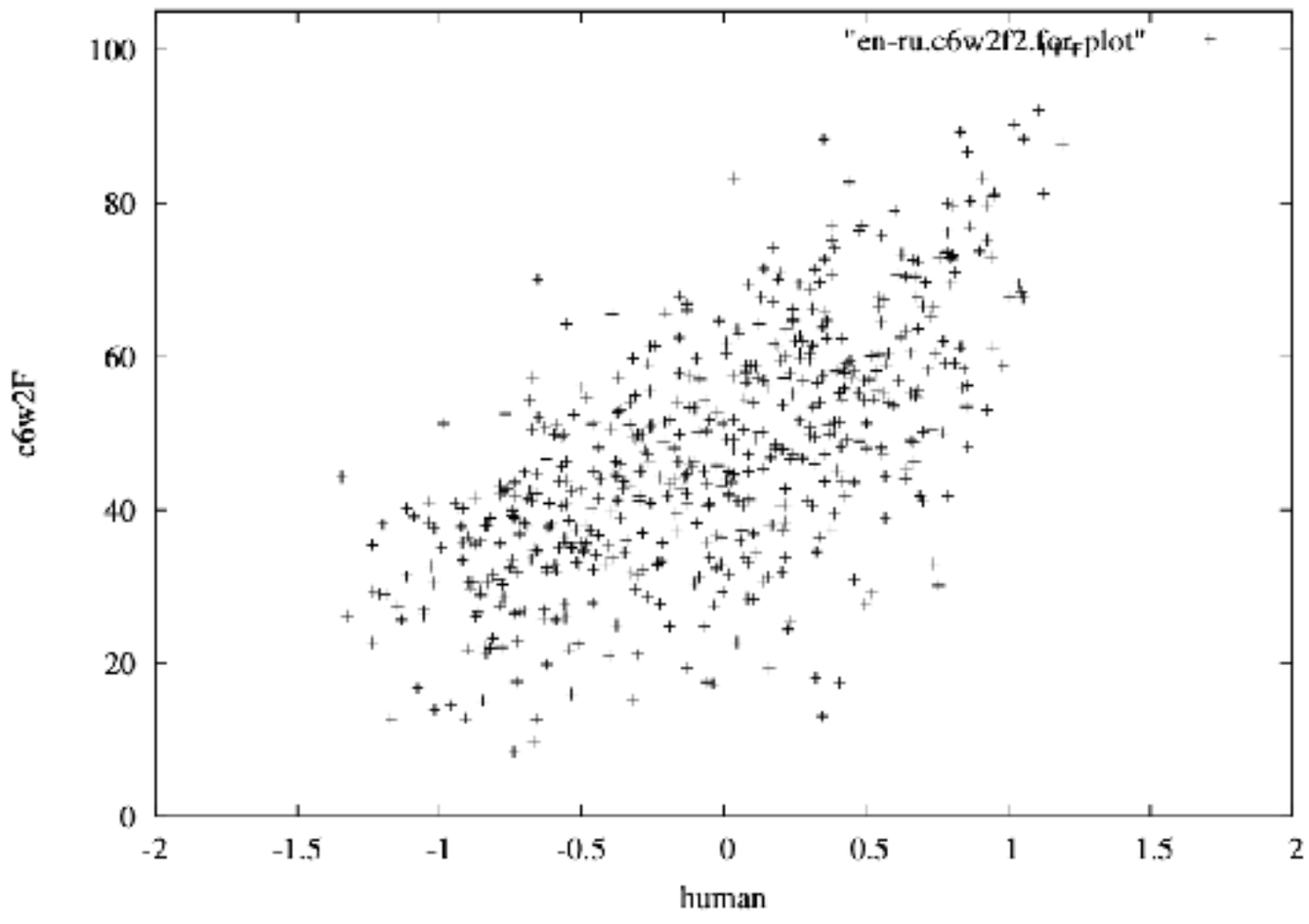


(b) English→Russian, WORDF



(d) English→Russian, CHRF

CHRF++



(d) CHRF++ (CHRF +word2F)

Thank You!