Comparison Between ATA Grading Framework Scores and Auto Scores

Evelyn Yang Garland
Acta Chinese Language Services LLC, egarland@actalanguage.com
Carola F Berger
CFB Scientific Translations LLC, info@cfbtranslations.com
Jon Ritzdorf
Procore, jon@ritzdorfacademy.com

Question

- How much **agreement** is there between **human evaluation scores** and **auto evaluation scores** when they are used to evaluate human translations and MTs?

Methodology

- Exploratory study
  - Data from a previous study
  - Not specifically designed to test the hypothesis question
- 2 source passages, English-into-Chinese, general

- Passage A: 263 words
  - 8 human translations (HTAs)
    - No use of MT
    - 6 professional translators and 2 students
  - 2 reference translations (for auto scoring)
    - RefA1: plain, error-free reference
    - RefA2: fancy, few errors

- Passage B: 264 words
  - 8 human translations (HTBs)
    - One MT provided for reference
    - Free to use other MTs
    - Same 6 professional translators and 2 students
  - 3 MTs (including the reference MT)
  - 2 reference translations (for auto scoring)
    - RefA1: plain, error-free reference
    - RefA2: fancy, few errors

Methodology (cont'd)

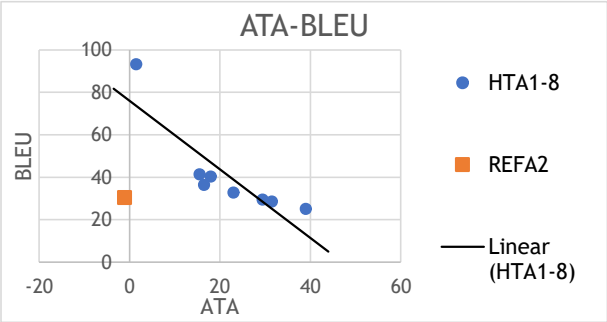
- Human evaluation
  - ATA Grading Framework
  - Graded by 2 ATA certified translators
  - Average of the 2 graders' scores
- Auto evaluation
  - BLEU
  - TER
  - COMET (wmt20-da)
  - COMET no reference (wmt20-da, wmt20-da v2, wmt21-mqm)

Result 1

- Auto scores that rely on reference translations depend heavily on which reference is used

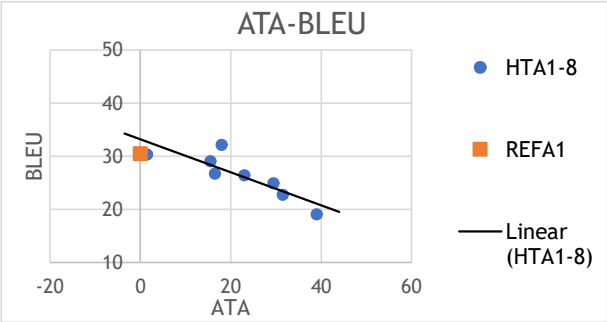
Result 1 – Passage A, BLEU, TER

RefA1: “plain”

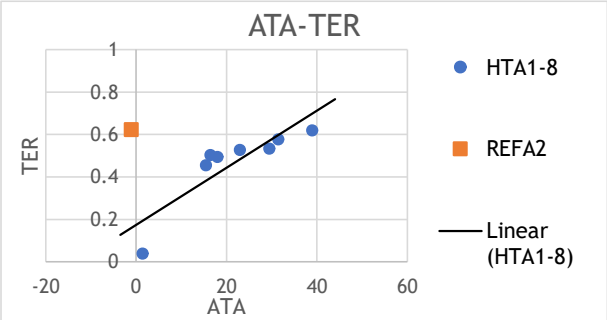


Pearson: -0.857, p: 0.006

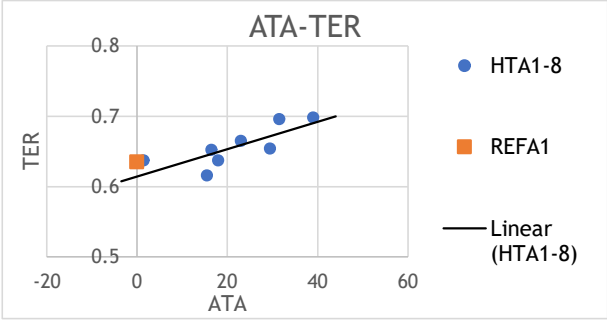
RefA2: “fancy”



Pearson: -0.853, p: 0.007



Pearson: 0.866, p: 0.005

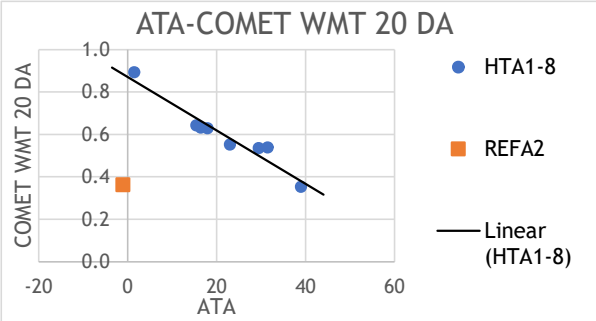


Pearson: 0.786, p: 0.021

- ATA, TER: higher quality = lower score; BLEU: higher quality = higher score
- Trendline based on HTA1-8

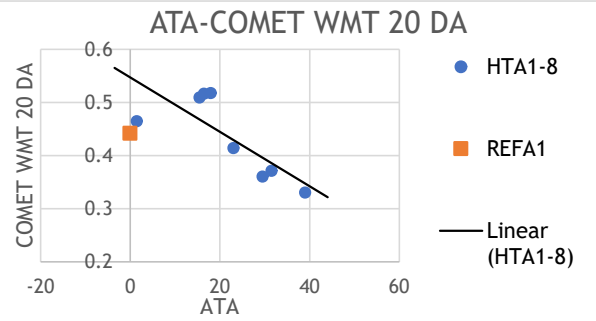
Result 1 – Passage A, COMET

RefA1: similar to HTA1-8

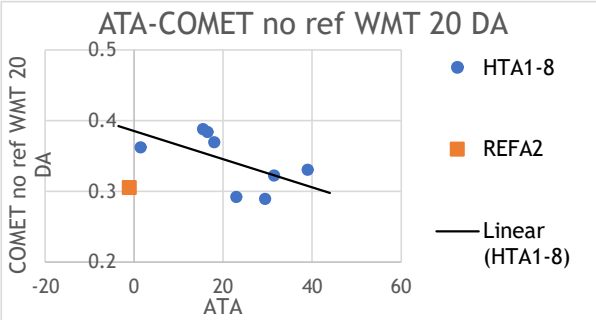


Pearson: -0.965, p: 0.0001

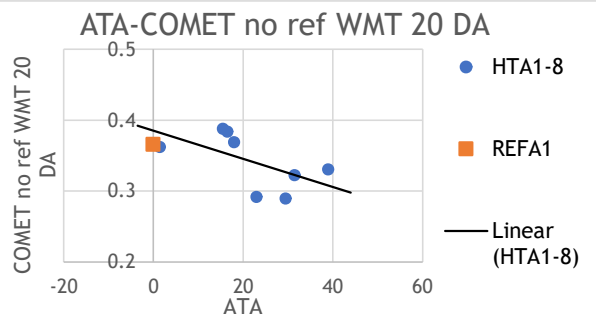
RefA2: independent



Pearson: -0.777, p: 0.0023



Pearson: -0.588, p: 0.126

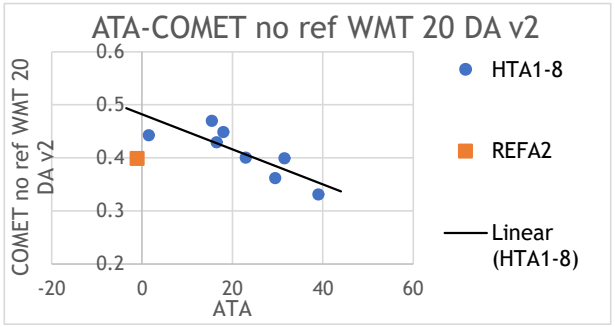


Pearson: -0.588, p: 0.126

- ATA: higher quality = lower score; COMET: higher quality = higher score
- Trendline based on HTA1-8

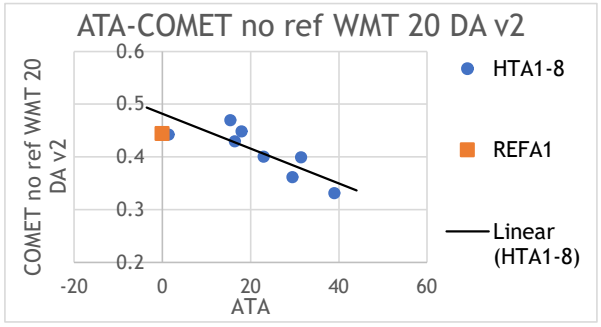
Result 1 – Passage A, COMET(cont'd)

RefA1: "plain"

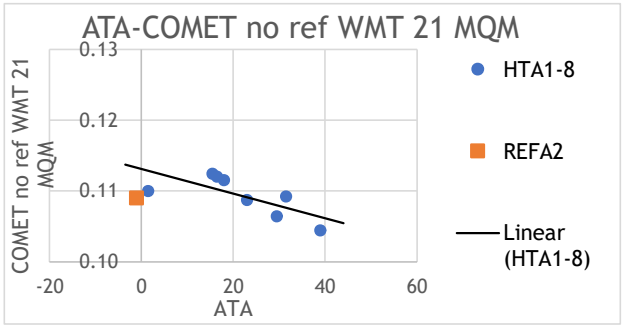


Pearson: -0.824, p: 0.012

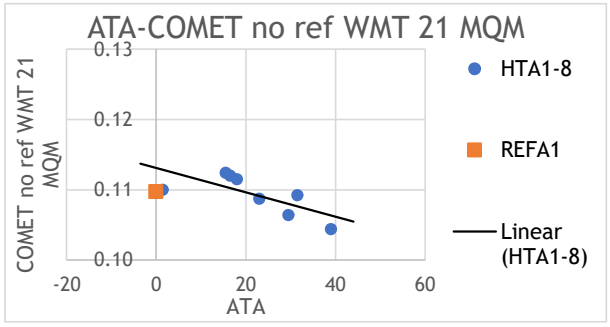
RefA2: "fancy"



Pearson: -0.824, p: 0.012



Pearson: -0.722, p: 0.043

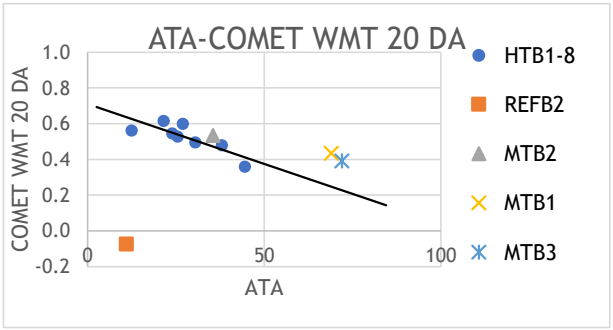


Pearson: -0.722, p: 0.043

- ATA: higher quality = lower score; COMET: higher quality = higher score
- Trendline based on HTA1-8

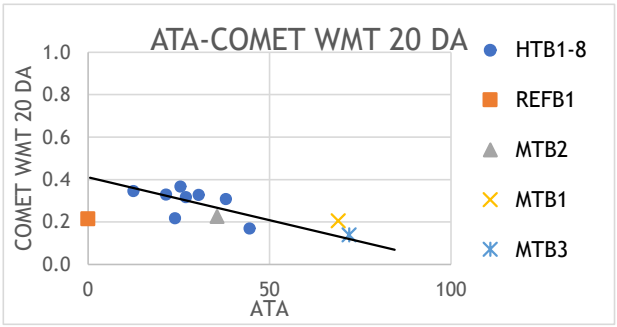
Result 1 – Passage B, COMET

RefB1: “plain”

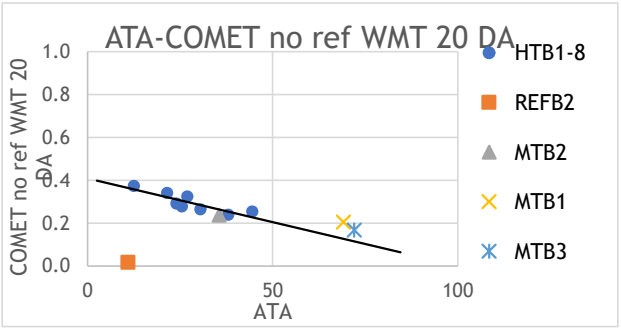


Pearson: -0.819, p: 0.013

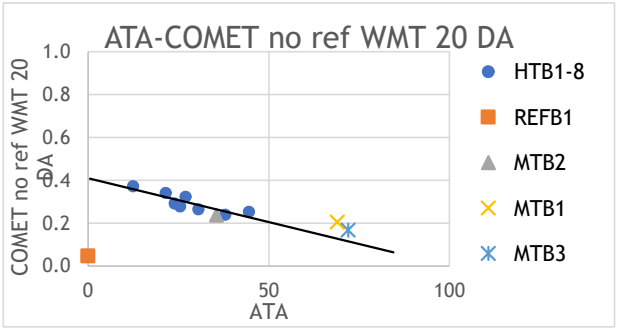
RefB2: “fancy”



Pearson: -0.588, p: 0.125



Pearson: -0.873, p: 0.005



Pearson: -0.873, p: 0.005

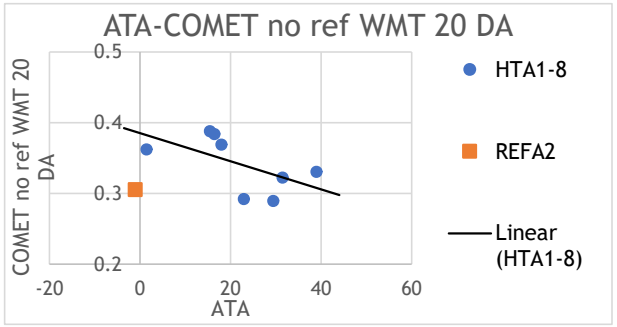
- ATA: higher quality = lower score; COMET: higher quality = higher score
- Trendline based on HTB1-8
- BLEU, TER, and COMET no ref WMT 21 MQM: no significant agreement;
- COMET no ref WMT 20 DA v2: did not obtain

Result 2

- Referenceless COMET seems promising when it is used to evaluate translations of short passages (~250 English words)

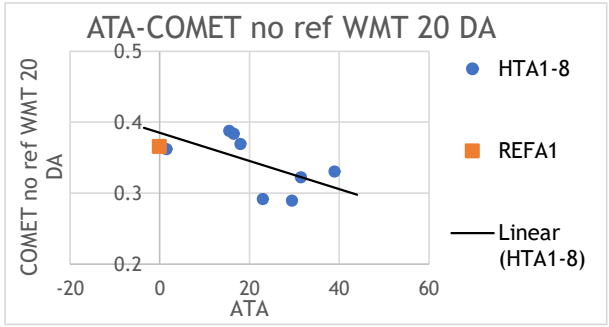
Result 2 – Passage A, COMET no ref

RefA1: “plain”

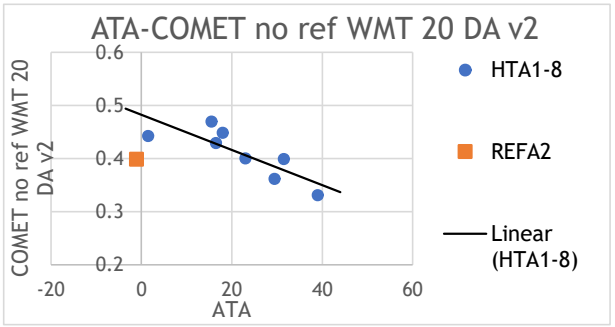


Pearson: -0.588, p: 0.126

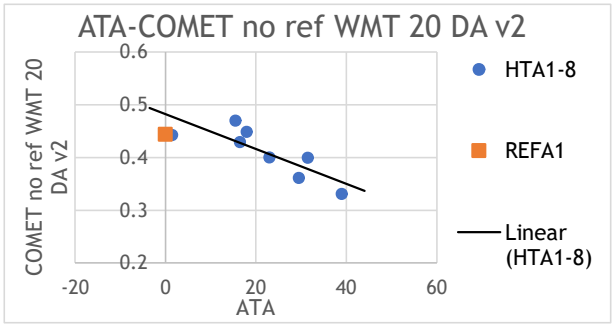
RefA2: “fancy”



Pearson: -0.588, p: 0.126



Pearson: -0.824, p: 0.012

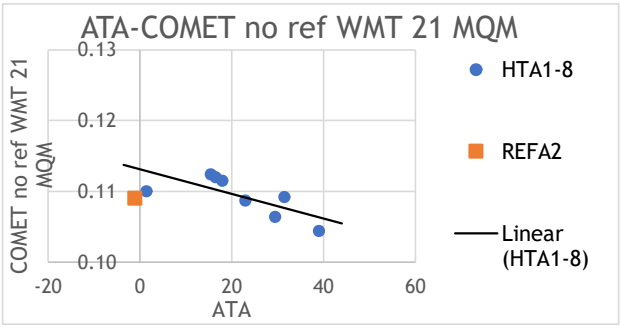


Pearson: -0.824, p: 0.012

- ATA: higher quality = lower score; COMET: higher quality = higher score
- Trendline based on HTA1-8

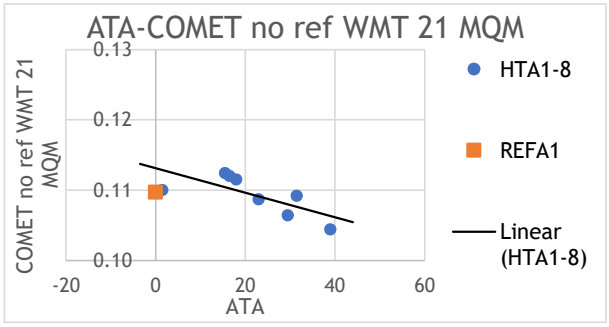
Result 2 – Passage A, COMET no ref (cont'd)

RefA1: "plain"



Pearson: -0.722, p: 0.043

RefA2: "fancy"

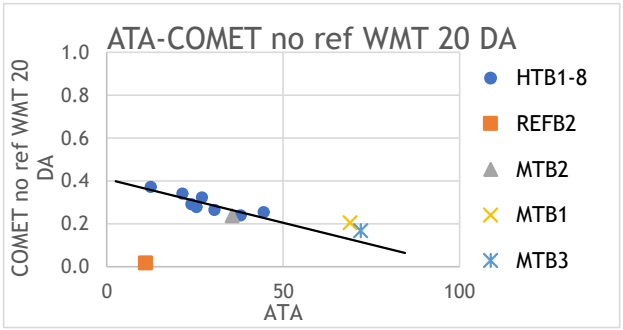


Pearson: -0.722, p: 0.043

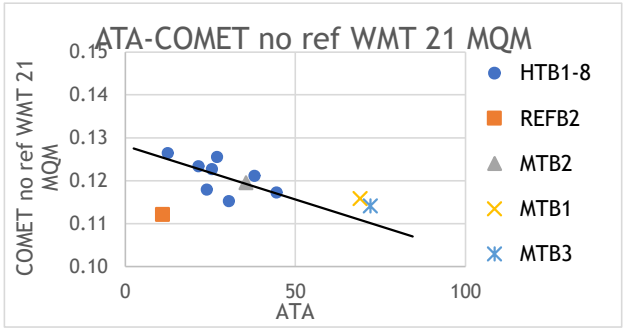
- ATA: higher quality = lower score; COMET: higher quality = higher score
- Trendline based on HTA1-8

Result 2 – Passage B, COMET no ref

RefB1: “plain”

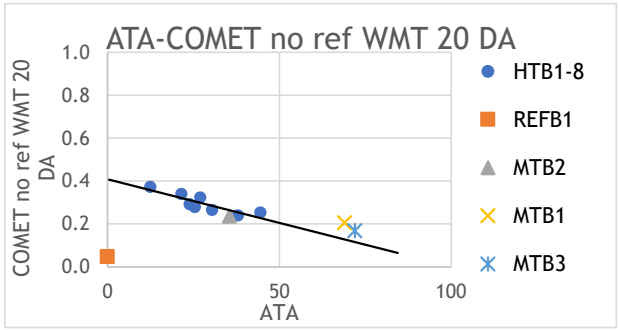


Pearson: -0.873, p: 0.005

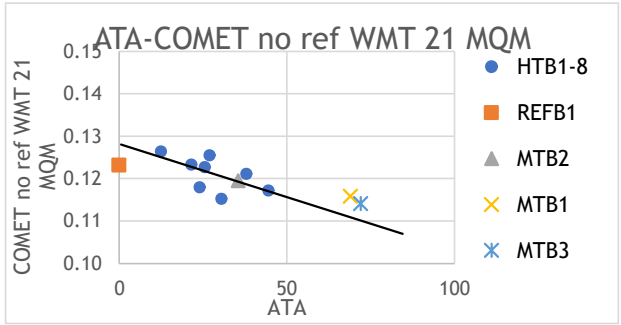


Pearson: -0.610, p: 0.108

RefB2: “fancy”



Pearson: -0.873, p: 0.005



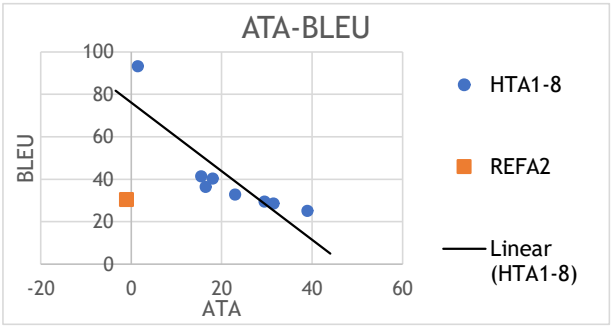
Pearson: -0.610, p: 0.108

- ATA: higher quality = lower score; COMET: higher quality = higher score
- Trendline based on HTB1-8
- COMET no ref WMT 20 DA v2: did not obtain

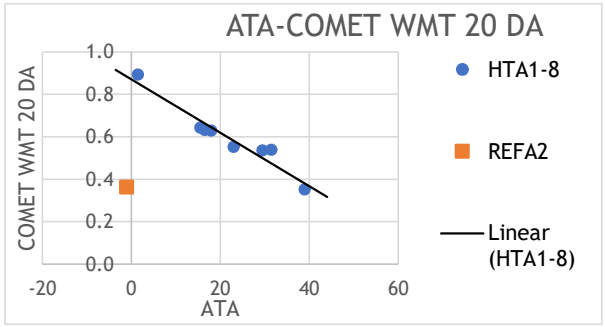
Result 3

- Good agreement between the ATA-Framework score and auto scores within a middle range, but the relationship becomes non-monotonic beyond the middle range

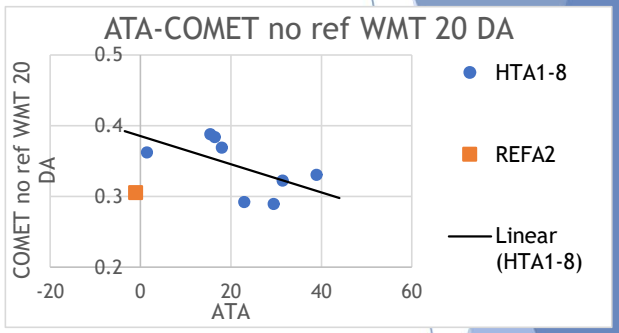
Result 3 – Passage A, ref = RefA1 (“plain”)



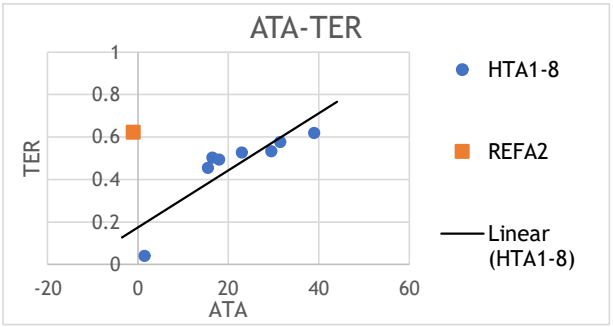
Pearson: -0.857, p: 0.006



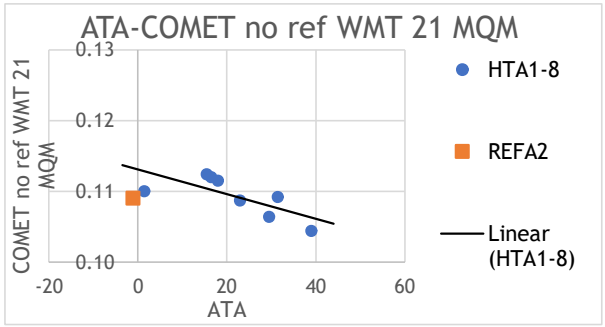
Pearson: -0.965, p: 0.0001



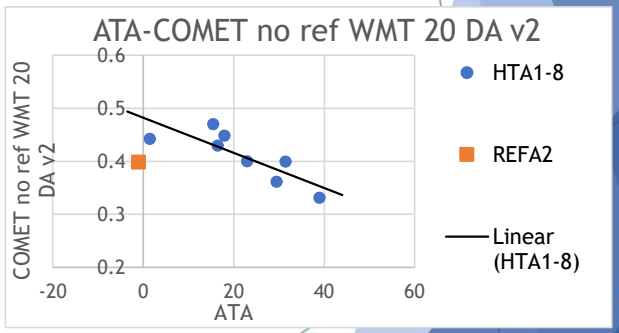
Pearson: -0.588, p: 0.126



Pearson: 0.866, p: 0.005



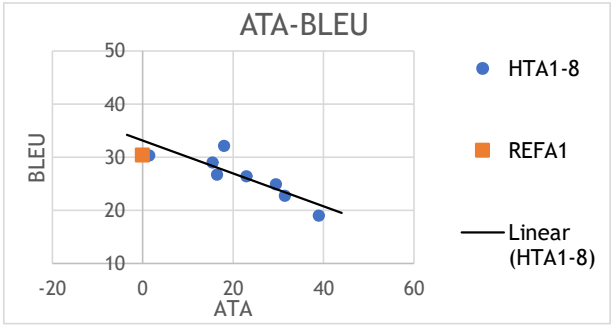
Pearson: -0.722, p: 0.043



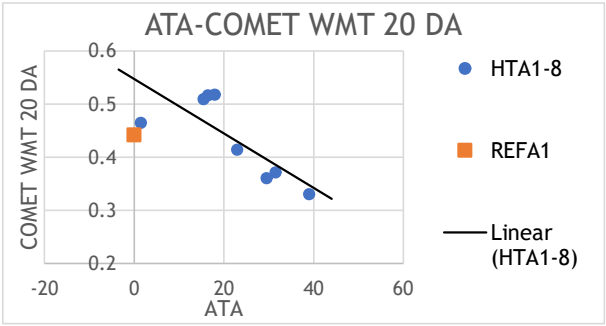
Pearson: -0.824, p: 0.012

- ATA, TER: higher quality = lower score; BLEU, COMET: higher quality = higher score
- Trendline based on HTA1-8

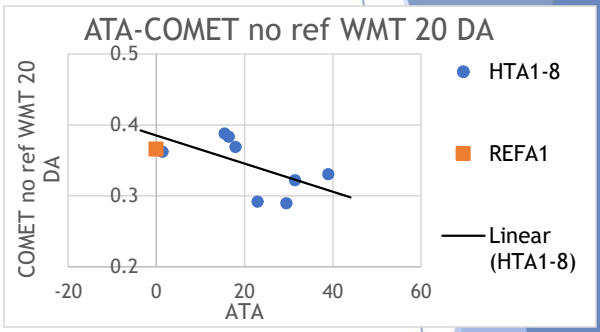
Result 3 – Passage A, ref = RefA2 (“fancy”)



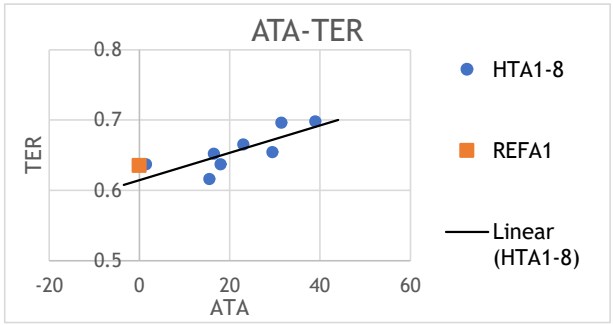
Pearson: -0.853, p: 0.007



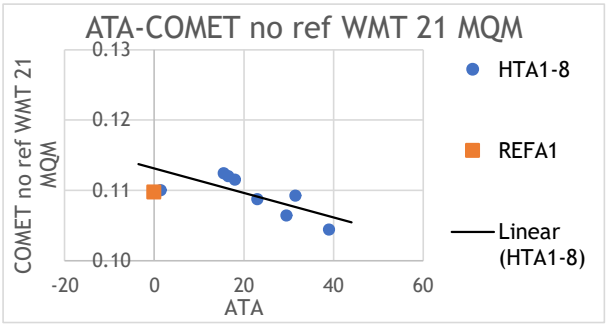
Pearson: -0.777, p: 0.0023



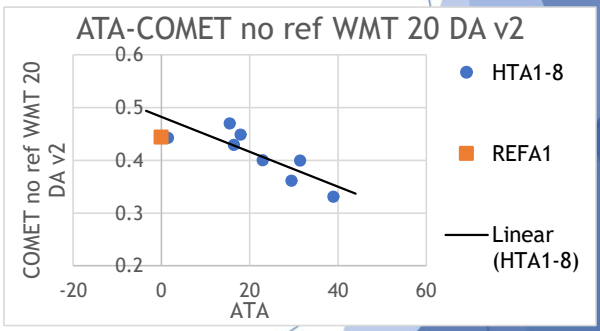
Pearson: -0.588, p: 0.126



Pearson: 0.786, p: 0.021



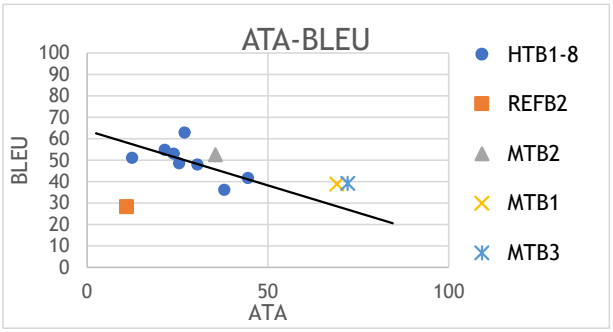
Pearson: -0.722, p: 0.043



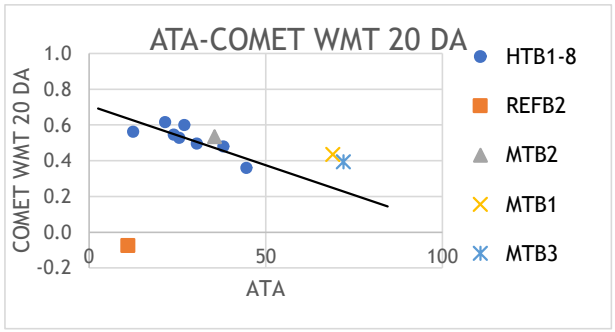
Pearson: -0.824, p: 0.012

- ATA, TER: higher quality = lower score; BLEU, COMET: higher quality = higher score
- Trendline based on HTA1-8

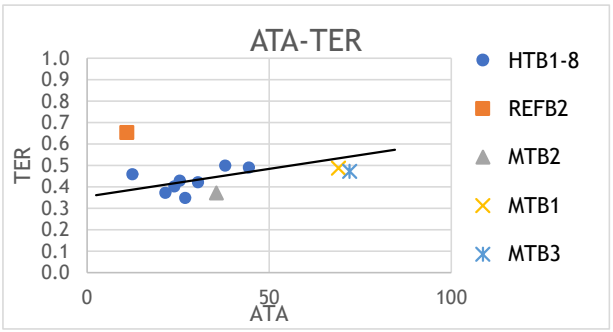
Result 3 – Passage B, ref = RefB1 (“plain”)



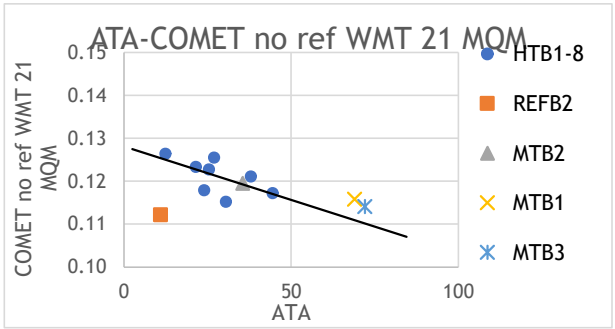
Pearson: -0.623, p: 0.099



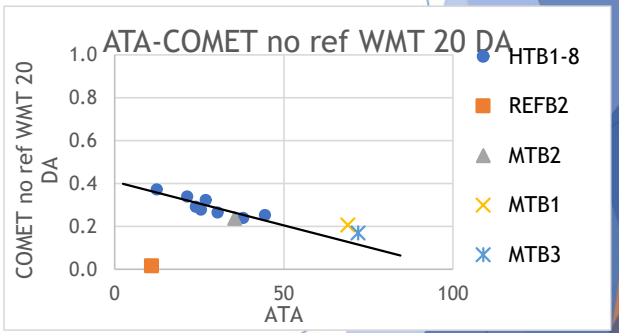
Pearson: -0.819, p: 0.013



Pearson: 0.479, p: 0.230



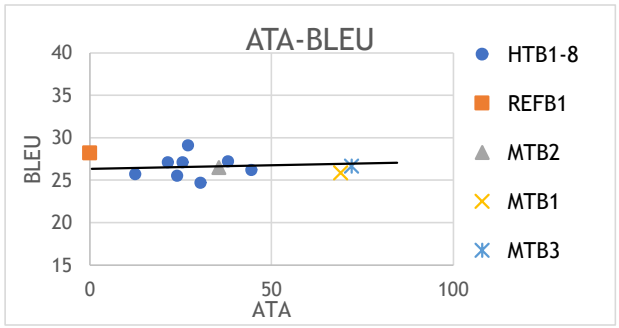
Pearson: -0.610, p: 0.108



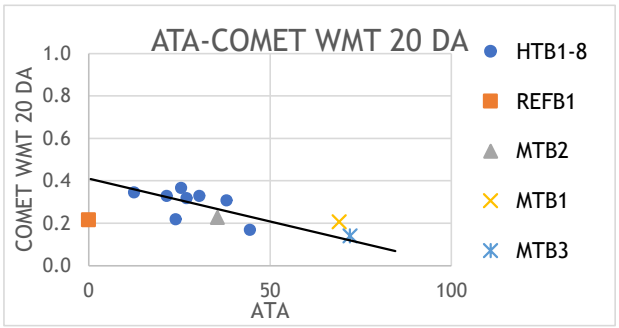
Pearson: -0.873, p: 0.005

- ATA, TER: higher quality = lower score; BLEU, COMET: higher quality = higher score
- Trendline based on HTB1-8

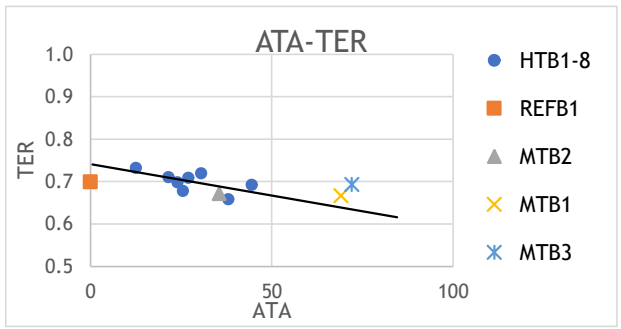
Result – Passage B, ref = RefB2 (“fancy”)



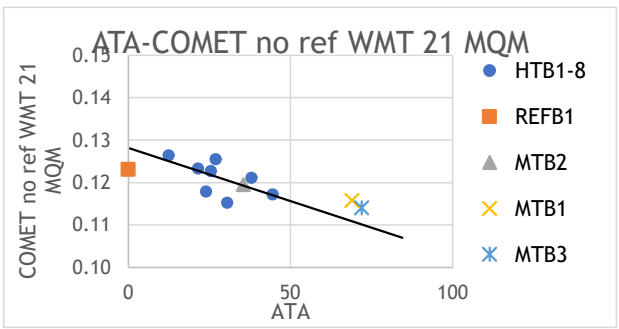
Pearson: 0.064, p: 0.881



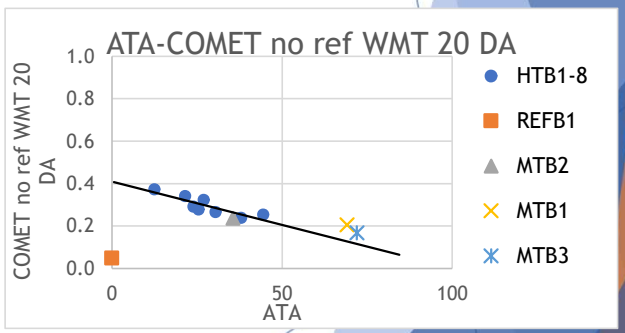
Pearson: -0.588, p: 0.125



Pearson: -0.621, p: 0.100



Pearson: -0.610, p: 0.108



Pearson: -0.873, p: 0.005

- ATA, TER: higher quality = lower score; BLEU, COMET: higher quality = higher score
- Trendline based on HTB1-8

Results & Conclusions

1. Auto scores that rely on reference translations depend heavily on which reference is used
   - Reference translation must be selected with care
2. Referenceless COMET seems promising when it is used to evaluate translations of short passages (~250 English words)
   - Potential of referenceless COMET as a QE tool (subject to limitation below)?
3. Good agreement between the ATA-Framework score and auto scores within a middle range, but the relationship becomes non-monotonic beyond the middle range
   - Auto scores do not work well beyond a middle range

Limitations

- Small sample size
- Exploratory study
- Only one evaluation criterion: quality score under the ATA grading framework
  - Time, productivity, or cost not measured

Acknowledgment

- Ji Chen, Achim Ruopp, Rony Gao, Jessie Lu & Tianlu Redmon
- Translators and graders who generously donated their time and expertise