MT Summer Term 2021 Ex6: MT Evaluation (Basics)

1. In your own words please describe the differences between:

- Human – automatic evaluation
- Scoring – ranking evaluation
- Intrinsic – extrinsic evaluation
- Quality – diagnostic evaluation

2. Given the following

		Predictions	
		true	false
Ground Truth	true	tp	fn
	false	fp	tn

$$P = \frac{tp}{tp + fp}$$

$$R = \frac{tp}{tp + fn}$$

$$F = \frac{2 \times P \times R}{P + R}$$

please state what each of tp, fp, and fn are in the following MT evaluation example:

Reference	Israeli officials are responsible for airport security
System A	Israeli officials responsible of airport security

3. In your own words, please define precision, recall and f-score in automatic machine translation evaluation.

4. Why is precision on its own not a good measure of MT output quality?

5. Why is recall on its own not a good measure of MT output quality?

6. Why is f-score a "conservative" measure?

7. Can f-score (as defined above) ever be lower than the lowest of its component measures P or R?

8. Please use precision, recall and f-score to evaluate

Reference	Israeli officials are responsible for airport security
System A	Israeli officials responsible of airport security
System B	Israeli officials are in charge of airport security
System C	security airport are officials for responsible Israeli

9. In your own words, please describe BLEU. Compare with f-score, what is the motivation for BLEU, which part is precision focused, which part approximates recall?

$$BLEU = \min\left(1, \exp\left(1 - \frac{|reference|}{|output|}\right)\right)\left(\prod_{n=1}^{4} n - gram\ precision\right)^{\frac{1}{4}}$$

10. Please use BLEU

$$BLEU = \min\left(1, \exp\left(1 - \frac{|reference|}{|output|}\right)\right)\left(\prod_{n=1}^{4} n - gram\ precision\right)^{\frac{1}{4}}$$

to compute evaluations for

Reference	Israeli officials are responsible for airport security
System A	Israeli officials responsible for airport security
System B	Israeli officials are in charge of airport security
System C	for airport security Israeli officials are responsible

11. Can you use BLEU to evaluate translations of single sentences? Does BLEU correlate well with human quality assessments? Can BLEU be used without question to compare e.g. RBMT with PBSMT systems?

12. Please compute the BLEU score between

Reference:	Yesterday John resigned from his job
System A:	John quit his job yesterday

What does this say about BLEU? Can you think about ways of improving BLEU to attempt to capture some of this?

13. For what kinds of languages could a character- rather than a word token-based automatic evaluation be a good idea?

14. Please explain why BLEU is not a great sentence level evaluation metric (in the sense that you should not be using it to rate an individual sentence but rather 100s or better 1000s of them)?

15. In your own words, what are the advantages and disadvantages of human evaluation?

16. In your own words, what are the advantages and disadvantages of an automatic evaluation such as BLEU?

17. What is the big difference between automatic MT evaluation and automatic MT quality estimation?