Machine Translation: Summer Term 2021

MT Evaluation (Basics)

Sangeet Sagar (7009050)
sasa00001@stud.uni-saarland.de

July 9, 2021

1. **In your own words please describe the differences between:**
   - **Human – automatic evaluation**
   - **Scoring – ranking evaluation**
   - **Intrinsic – extrinsic evaluation**
   - **Quality – diagnostic evaluation**

   1.

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

$$\text{Precision} : \frac{correct\ words\ in\ output}{total\ words\ in\ output} = \frac{3}{6} = 0.5$$

$$\text{Recall} : \frac{correct\ words\ in\ output}{total\ words\ in\ reference} = \frac{3}{7} = 0.43$$

$$\text{F-measure} : \frac{2 \times precision \times recall}{precision + recall} = 0.46$$

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

4. **Why is precision on its own not a good measure of MT output quality?**
   Consider the example:
   Reference Israeli officials are responsible for airport security
   System Israeli

   $$Precision : \frac{correct\ words\ in\ output}{total\ words\ in\ output} = \frac{1}{1} = 1$$

   Although the system produced just one correct word translation, the precision is 1. It is solely dependent on true positive and false positive values and the negatives are not taken into account. Therefore it's possible to be precise while being inaccurate at the same time.

5. **Why is a recall on its own not a good measure of MT output quality?**
   Consider the example:
   Reference Israeli officials are responsible for airport security
   System dump of all words in MT vocabulary

   $$Recall : \frac{correct\ words\ in\ output}{total\ words\ in\ reference} = \frac{1}{1} = 1$$

   It should be noted that recall deals with only true positives and false negatives. If in the system translation we just dumb the entire vocabulary generated from the training data, we get recall as 1 because the words in the reference sentence are a subset of the vocabulary. Hence, recall on its own is not a good evaluation metric in MT.

6. **Why is f-score a "conservative" measure?**
   F-score is conservative in the sense that its value is always closer to the lower of either precision or recall.

7. **Can f-score (as defined above) ever be lower than the lowest of its component measures P or R?**
   No. F-score is a geometric mean of precision and recall and hence always between precision and recall.

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

   $$\text{Precision} : \frac{5}{8} = 0.625$$
   $$\text{Recall} : \frac{5}{7} = 0.71$$
   $$\text{F-measure} : \frac{2 \times 0.625 \times 0.71}{0.625 + 0.71} = 0.66$$

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

$$\text{Precision}: \frac{5}{8} = 0.625$$
$$\text{Recall}: \frac{5}{7} = 0.71$$
$$\text{F-measure}: \frac{2 \times 0.625 \times 0.71}{0.625 + 0.71} = 0.66$$

Reference	Israeli officials are responsible for airport security
System C	security airport are officials for responsible Israeli

$$\text{Precision}: \frac{7}{7} = 1$$
$$\text{Recall}: \frac{7}{7} = 1$$
$$\text{F-measure}: 1$$

Here even with an F-score of 1, the translation is meaningless. Hence F-score fails to reflect word order.

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?**

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

BLEU (Bilingual Evaluation Understudy) algorithm compares consecutive phrases of the automatic translation with the consecutive phrases it finds in the reference translation, and counts the number of matches, in a weighted fashion. These matches are position-independent. A higher match degree indicates a higher degree of similarity with the reference translation, and a higher score. Intelligibility and grammatical correctness are not taken into account. [Source: Miscrosoft: What is a BLEU score?] Precision is captured by $\prod_{n=1}^{4} \textit{n-gram precision}^{\frac{1}{4}}$. Recall is captured by $min(1, exp(1 - \frac{|reference|}{|output|}))$. Here we are punishing the precision by a number smaller than 1 i.e. the recall component of the BLEU score.

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

System A:

$$\prod_{n=1}^{4} \textit{n-gram precision}^{\frac{1}{4}} = (\frac{5}{6} \times \frac{2}{5} \times \frac{0}{4} \times \frac{0}{3})^{\frac{1}{4}}$$
$$BLEU = 0$$

System B:

$$\prod_{n=1}^{4} n\text{-}gram\ precision^{\frac{1}{4}} = (\frac{5}{8} \times \frac{3}{7} \times \frac{1}{6} \times \frac{0}{5})^{\frac{1}{4}}$$
$$BLEU = 0$$

System C:

$$\prod_{n=1}^{4} n\text{-}gram\ precision^{\frac{7}{7}} = (\frac{7}{7} \times \frac{5}{6} \times \frac{3}{5} \times \frac{1}{4})^{\frac{1}{4}} = 0.5946$$
$$min(1, exp(1 - \frac{\mid reference \mid}{\mid output \mid})) = min(1, exp(1 - \frac{7}{7})) = 1$$
$$BLEU = 0.5946$$

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?**
BLEU is a terrible measure on sentence-level so we have to aggregate these over our test sentences, then compute the average. BLEU correlates with human quality assessments when computed on document level with a sentence in the range of 10^3 or 10^4.

    No, BLEU is not a standard metric to compare such systems as they have more manually framed rules to perform translations. Such systems rank high on a human score, but terrible on a BLEU score. It's better to use a metric to compare translations that are based on the same technology (e.g BLEU is ngram based evaluation metric and MT systems based on this technology are more suitable for comparison).

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?
The BLEU score is 0 since it has 0 tri-gram precision. The system output translated sentence is meaningfully aligned with the reference sentence but the fact is that BLEU does not consider the meaning. It has limited capability where it only compares n-grams with the reference sentence. This is the reason why human translators often score low in BLEU.

13. **For what kinds of languages could a character- rather than a word token-based automatic evaluation be a good idea?**
Character-based evaluation is a promising metric for comparing MT systems because it is language-independent, tokenization independent and it shows good correlations with human judgments both on the system- as well as on the segment-level [Popović, 2015]. In this paper, they used CHRF for French, German, Czech, Hindi, and Russian.

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)?**
BLEU is problematic since it can easily become zero at the sentence level. This is because of the product of n-gram precisions in the geometric mean of the precision component of BLEU: if some p_n is zero, the whole product will be zero. In particular, it is easy to see that BLEU will be zero for any hypothesis without 4-gram matches. This is undesirable and the score must be computed on a document level. [Nakov et al., 2012]

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

    - Indispensable because they present more accurate and trusted evaluation measurement,

    Disadvantages:

- Time consuming
- Expensive as it requires highly proficient language speakers
- Difficult to define and operationalise

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

    - Automatic evaluation algorithm. It's fast and easy to calculate,
    - Gives the same result when evaluated on same data, unlike human evaluation where we get new results every time.
    - Uses an average of n-gram precision

    Disadvantages:

    - A perfectly meaningful translation may result in 0 BLEU score if even a single n-gram is missing.
    - It doesn't handle morphologically rich languages well
    - It doesn't map well to human judgments

17. **What is the big difference between automatic MT evaluation and automatic MT quality estimation?**
    MT quality estimation is the estimation or prediction of translation quality without reference while in automatic MT evaluation, the reference translation is the base or the gold translation using which the system-generated translations are evaluated.

References

Preslav Nakov, Francisco Guzman, and Stephan Vogel. Optimizing for sentence-level BLEU+1 yields short translations. In *Proceedings of COLING 2012*, pages 1979–1994, Mumbai, India, December 2012. The COLING 2012 Organizing Committee. URL `https://aclanthology.org/C12-1121`.

Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL `https://aclanthology.org/W15-3049`.