COSC478

Special Topic: Generative Artificial Intelligence Using Large Language Models

Lecture 5 – Metrics for Evaluating Large Language Models

Tuesday 30 July 2024

Semester 2, 2024

Trevor Nesbit

Today

- Completing the example using BLEU and recapping what BLEU is all about
- Exploring ROGUE and going through a worked example
- If time permits, and introduction to what the “Human Evaluation” metric is all about

Recap on BLEU

- BLEU is used for evaluating the quality of text which has been machine-translated from one natural language to another

- We used the following sentence in English: “I went to University to drink coffee”

- We consulted with a fluent speaker of Te Reo Māori who translated it as: “Haere ahau ki te whare wānanga ki te inu kawhe”
- We then used a large language model to do the translation and it came up with: “I whakawhiti ahau ki te whare wānanga ki te inu kawhe”

Recap on BLEU

- We determined:
  - the number of unigrams in the text generated by the model (commonly called the model text)
  - the number of unigrams in the text generated by the fluent speaker (commonly called the reference text)
  - the number of unigrams that are in common across the model text and the reference text

- We calculated the precision for unigrams:

= number of unigrams in common/unigrams in text generated by model

Recap on BLEU

- We repeated the same for bigrams – what were they again?

- The precision for unigrams was 9/11 = 0.818
- The precision for bigrams was 8/10 = 0.800

- The total BLEU score is the geometric mean of the n-gram precisions

= sqrt (0.818 * 0.800) = 0.809

- The final note was we can do this for 3-grams and 4-grams too…

BLEU 3 grams

Reference text (fluent transalator version)

Haere ahau ki te whare wānanga ki te inu kawhe

3 grams (8 in total)

- Haere ahau ki
- ahau ki te
- ki te whare
- te whare wānanga
- whare wānanga ki
- wānanga ki te
- ki te inu
- te inu kawh

BLEU 3 grams

Model text (fluent transalator version)

I whakawhiti ahau ki te whare wānanga ki te inu kawhe

3 grams (9 in total)

- I whakawhiti ahau
- whakawhiti ahau ki
- ahau ki te
- ki te whare
- te whare wānanga
- whare wānanga ki
- wānanga ki te
- ki te inu
- te inu kawhe

BLEU 3 grams

3 grams in common (7 in total)

- ahau ki te
- ki te whare
- te whare wānanga
- whare wānanga ki
- wānanga ki te
- ki te inu
- te inu kawhe

BLEU 3 grams

Precision for 3 grams =

= number of 3grams in common/3grams in text generated by model

= 7/9 = 0.7780,

- The total BLEU score is the geometric mean of the n-gram precisions

= cubed root (0.818 * 0.800 * 0.778) = 0.798

BLEU 4 grams

Reference text (fluent transalator version)

Haere ahau ki te whare wānanga ki te inu kawhe

4 grams (7 in total)

- Haere ahau ki te
- ahau ki te whare
- ki te whare wānanga
- te whare wānanga ki
- whare wānanga ki te
- wānanga ki te inu
- ki te inu kawhe

BLEU 4 grams

Model text (fluent transalator version)

I whakawhiti ahau ki te whare wānanga ki te inu kawhe

4 grams (8 in total)

- I whakawhiti ahau ki
- whakawhiti ahau ki te
- ahau ki te whare
- ki te whare wānanga
- te whare wānanga ki
- whare wānanga ki te
- wānanga ki te inu
- ki te inu kawhe

BLEU 4 grams

4 grams in common (6 in total)

- ahau ki te whare
- ki te whare wānanga
- te whare wānanga ki
- whare wānanga ki te
- wānanga ki te inu
- ki te inu kawhe

BLEU 4 grams

Precision for 3 grams =

= number of 3grams in common/3grams in text generated by model

= 6/8 = 0.750,

- The total BLEU score is the geometric mean of the n-gram precisions

= fourth root (0.818 * 0.800 * 0.778 * 0.750) = 0.786

A question about BLEU

- Suppose you did two sets of BLEU calculations
- What factors could result in you getting two quite different sets of results?

Moving on to ROGUE

- Interesting acronym
- What does is stand for?
- Recall-Oriented-Understudy for Gisting Evaluation

- What is “gisting”?
  - In a language processing and machine translation context it refers to machine translation to foreign text to get an understanding of the meaning (sounds like BLEUR?)
  - It can also be taking as referring to the process of creating a summary – and this is what it means in this context

Moving on to ROGUE

- What is ROGUE?
- A group of metrics that can be used to evaluate summaries of text by comparing them to reference summaries
- So, we have an original piece of text (can be about anything)
- A reference summary of that text that has been produced by an human expert in the field
- A summary that has been automatically generated (by a LLM)

What does ROGUE do?

- Compares
- Quantifies
- Produces scores:
  - ROGUE-N overlap on n-grams
  - ROGUE-L longest common sub-sequence (LCS)
  - ROGUE-W weighted LCS-based stats
  - ROGUE-S Skip-bigram based
  - ROGUE-SU Skip-bigram and unigram based

- The higher the score the more similarities there are between the reference summary and the generated summary
- Can be used for comparing two generated summaries

An example of ROGUE-N

- Assume we had an original text that was all about things that are done by cats
- An expert in cats produced us a reference summary of “The cat sat on the mat”
- We used a LLM to generate another summary which was “The cat is sitting on the mat”
- From here on it is very much like BLEUR
- Reference unigrams – “The”, “cat”, “sat”, “on”, “the”, “mat” (6 in total)
- Generated unigrams - “The”, “cat”, “is”, “sitting”, “on”, “the”, “mat” (7 in total)
- Unigrams in common - “The”, “cat”, “on”, “the”, “mat” (5 in total)
- ROGUE-1 score = ***unigrams in common/reference unigrams = 5/6 = 0.833***

An example of ROGUE-N

- Reference bigrams – “The cat”, “cat sat”, “sat on”, “on the”, “the mat” (5)
- Generated bigrams - “The cat”, “cat is”, “is sitting”, “sitting on”, “on the”, “the mat” (6)
- Bigrams in common - “The cat”, “on the”, “the mat” (3 in total)
- ROGUE-2 score = ***unigrams in common/reference unigrams = 3/5 = 0.600***

- Reference 3grams – “The cat sat”, “cat sat on”, “sat on the”, “on the mat” (4)
- Generated 3grams - “The cat is”, “cat is sitting”, “is sitting on”, “sitting on the”, “on the mat” (5)
- 3grams in common - “on the mat” (1 in total)
- ROGUE-3 score = ***unigrams in common/reference unigrams = 1/4 = 0.250***

What does ROGUE-L look like?

- ROGUE-L looks at the longest common sub-sequence (LCS) between the reference summary and the generated summary
  - The longest common sub-sequence is the longest sequence of words that are the same between the two summaries and can be in any order
- Let's use the same example
  - Reference summary: "The cat sat on the mat"
  - Generated summary: "The cat is sitting on the mat"
- LCS = "The cat on the mat"
- Length of the LCS = 5
- Precision (P) = Length of LCS/Length of Generated Summary = 5/7 = 0.71
- Recall (R) = Length of LCS/Length of Reference Summary = 5/6 = 0.83
- F-measure = 2PR/(P+R) = 2 * .71 * .83 / (.71+.83) = 0.765

Comparing ROGUE-N and ROGUE-L

- What does it mean if they are both high (>0.7)
- What does it mean if they are both low (<0.5)

- What does it mean if ROGUE-N is low, and ROGUE-L is high?
  - could mean that the system summary has many of the same words as the reference summary, but they are arranged differently or used in different contexts. Might suggest that the system is capturing the main ideas of the reference summary but expressing them in a different way.

- What does it mean if ROGUE-N is high, and ROGUE-L is low?
  - Could mean that the system summary has many of the same phrases as the reference summary, but overall uses different words. Might suggest that the system is capturing the specific wording or phrasing of the reference summary but is missing some of the main ideas or content.

What is the Human Evaluation metric all about?

- Involves having human evaluators assess the quality of generated text:
    1. Scoring adequacy and fluency
    2. Ranking of outputs
    3. Post editing measures
    4. Subject evaluation judgement
    5. Correlation with human scores

- Does it really fit with our understanding of what a metric is?

- Why is it commonly seen as one of the main metrics to use for evaluating LLMs?

Why is the Human Evaluation metric used?

- Better at capturing context
- Humans can assess quality better (at the moment?)
- Evaluating creativity and novelty
- Identifying errors
- Benchmarking

What are the issues with the Human Evaluation metric?