CISC7021 Applied Natural Language Processing – Course Project

Comparison of Alternative Evaluation Approaches for Natural Language Applications

Group 37

M-C3-5171-2, CHAN CHI WENG

M-C3-5080-9, KE SIYUN

M-C3-5254-7, LEONG KA HOU

Objective

- This project aims to investigate and verify a few novel approaches for evaluating NLP applications (e.g. SMT/NMT models or similar text generation applications) and attempts to analyze the advantages or limitations of the novel methods.

Evaluation algorithms

- BLEU: It compares a candidate translation to one or more reference translations, measuring the precision of n-grams in the translated text.
- BARTScore: An evaluation model based on BART, it evaluates the quality of generated text by measuring how likely the model can reconstruct the generated text.

Evaluation algorithms

- BERTScore: a metric for evaluating text generation which leverages the pre-trained contextual embeddings from BERT (Bidirectional Encoder Representations from Transformers)

- COMET: (Crosslingual Optimized Metric for Evaluation of Translation) is a machine learning-based framework for the automatic evaluation of machine translation.

Dataset for experiments

- Chinese and English parallel sets from IWSLT 2014
- Process the Chinese text through third-party translation
  - Google Translation API
- Machine-generated English texts → candidate texts
- Original English texts → reference texts

Treatments to MT result

- **Original** : Establish a reference point or a baseline.
- **Word shuffled**: Allow to measure the extent of the scoring deviation introduced by the disruption of sentence structure.
- **4-gram reordered**: Make a significant change in the actual meaning and readability of these English sentences.
- **Synonym**: Aimed to maintain the semantic integrity of the sentences while altering their lexical composition.

Experiments

Experiments	Expectation
a) Original standard English texts vs. Original translated English texts	Expected to generate a relatively high score on this pair to acknowledge a good translation.
b) Original standard English texts vs. Shuffled translated English texts	Expected to return a low score on this pair to represent the loss of meaning in sentences.
c) Original standard English texts vs. 4- gram reordered translated English texts	The result shall be similar to result in b) but the score could be a little higher.
d) Original translated English texts vs. 4- gram reordered translated English texts	Expected to report a relatively low score on this pair to represent the loss of meaning in sentences
e) Original standard English texts vs. 4- gram reordered standard English texts	The result shall be similar to the result in experiments in d)
f) Original standard English texts vs. Standard English texts with synonyms replaced	Expected to provide a high rating similar to the original dataset.

Result - BLEU

#	Ref	Can	Score
a	*1*	*3*	*0.184*
b	*1*	*3.1*	*0.004*
c	*1*	*3.2*	*0.113*
d	*3*	*3.2*	*0.6598*
e	*1*	*1.2*	*0.6704*
f	*1*	*1.3*	*0.004*

Note- The list of dataset #s used in the experiments:
(1) Original standard English texts
(1.2) 4-grams reordered standard English texts
(1.3) Synonyms replaced standard English texts
(3) Original translation texts
(3.1) Shuffled translation texts
(3.2) 4-grams reordered translation texts

Result - BARTScore

#	Ref	Can	Score
a	*1*	*-3.174*	*-3.174*
b	*1*	*-7.389*	*-7.389*
c	*1*	*-5.298*	*-5.298*
d	*3*	*-3.915*	*-3.915*
e	*1*	*-3.817*	*-3.817*
f	*1*	*-7.016*	*-7.016*

Result - BERTScore

#	Ref	Can	Score (F1)
a	*1*	*3*	*0.932*
b	*1*	*3.1*	*0.855*
c	*1*	*3.2*	*0.882*
d	*3*	*3.2*	*0.913*
e	*1*	*1.2*	*0.911*
f	*1*	*1.3*	*0.853*

Result - COMET

#	Ref	Can	Score (F1)
a	*1*	*3*	*0.825*
b	*1*	*3.1*	*0.541*
c	*1*	*3.2*	*0.653*
d	*3*	*3.2*	*0.751*
e	*1*	*1.2*	*0.732*
f	*1*	*1.3*	*0.488*

Conclusion

- All these models did not show acceptable capacity for replaced synonyms.
- All three novel methods have better performance than BLEU in all capacities we tested.
- BLEU might be a practical alternative for teams without GPU access.

Thank you for listening