Natural Language Generation

Evaluation

Claire Gardent

Advanced Language Processing Winter School, 17-22 January 2021

Part 1: Basics of Neural NLG

The Encoder-Decoder Framework

- The Recurrent Encoder-Decoder

Better Decoding

- Attention, Copy and Coverage

Encoders

- Improved Recurrent Neural Network (RNN): LSTM, biLSTM, GRU
- Convolutional Neural Network (CNN)
- Transformer

Evaluation

- BLEU, ROUGE, PARENT, BLEURT, BERTScore

Evaluating Generated text

- The same content can be formulated in different ways
- Human evaluation is costly and difficult to manage
- Different evaluation focus for different generation tasks
  - Task based dialog: dialog efficiency, slot filling ratio ...
  - Summarisation: key information
  - Simplification: lexical, syntactic, discourse level simplificity level
  - ...
- All generated text must be fluent and faithful to the input

BLEU, ROUGE, BERTScore, BLEURT, PARENT

BLEU

Bilingual Evaluation Understudy

- Precision metric
- Compares an automatically generated sentence with one or more reference(s) (the expected output)
- Used to evaluate Machine Translation and NLG models
- Key idea: Compare the n-grams of the generated string to those of the reference

Papineni et al. ACL 2002

Precision

Nb of correct n-grams in generated text / Nb of n-grams in generated text

	Text	1-gram Precision
Reference	the cat is on the mat	
Output 1	the the the the the the the	7/7
Output 2	the cat is mat the in	5/6

- Output 1 and 2 have high 1-gram precision but are not good output
- BLEU in fact uses clipped precision, combined n-gram sizes and brevity penalty

Clipped Precision

Only accept as many correct n-grams as actually appear in the reference

	Text	1-gram Precision
Reference	the cat is on the mat	
Output 1	the the the the the the the	2/7

Modified N-Gram Precision

- BLEU computes individual clipped precision scores for 1 to 4 grams (not just 1 gram) and takes their geometric mean
- The modified n-gram precision decays roughly exponentially with n (the size of the n-gram)
- To take this exponential decay into account, BLEU uses a weighted average of the logarithm of clipped precisions

$$exp(\sum_{n=1}^{N} w_n \ log \ p_n)$$

Brevity Penalty

- Penalises sentences that are shorter than the reference (as these are likely to score high on precision)
- Computed on the whole corpus
  - r sum of the best match lengths for each candidate sentence in the corpus
  - c total length of the candidate sentences

$$BP = 1 \text{ if c} > r \text{ else } = e^{1-r/c}$$

ROUGE

Recall Oriented Understudy for Gisting Evaluation

Nb of correct n-grams in generated text / Nb of n-grams in reference text

- Recall metric
- Used to evaluate Summarisation models
- 1-ROUGE, 2-ROUGE ...
- L-ROUGE: checks for Longest Common Subsequence instead of n-gram overlap
- Key idea: Compare the n-grams of the generated string to those of the reference

Paraphrasing

GEN
This point will certainly be the subject of subsequent further debates in the council

REF
This is a point that will undoubtedly be discussed later in the council.

BERTScore

Semantic rather than string similarity

- Computes the similarity of two sentences as a sum of cosine similarities between their(BERT) embeddings.
- Evaluated on the outputs of 363 machine translation and image captioning systems.
- Correlates better with human judgments and provides stronger model selection performance than existing metrics (BLEU, MLETEOR, ROUGE etc.)

Zhang et al. ICLR 2020

PARENT

considers both reference text and data input

		BLEU	ROUGE	PARENT
Reference:	Michael Dahlquist (December 22 , 1965 -- July 14 , 2005) was a drummer in the Seattle band Silkworm .			
Candidate 1:	Michael Dahlquist (December 22 , 1965 -- July 14 , 2005) was a drummer in the California band Grateful Dead .	**0.79**	0.77	0.76
Candidate 2:	Michael Dahlquist (December 22 , 1965 -- July 14 , 2005) was a drummer .	0.71	**0.79**	0.82
Candidate 3:	Michael Dahlquist (December 22 , 1965 -- July 14 , 2005) was a drummer from Seattle, Washington .	0.73	0.70	**0.84**

Michael Dahlquist

Birth name	Michael Dahlquist
Born	December 22, 1965 Seattle, Washington
Died	July 14, 2005 (aged 39) Skokie, Illinois
Genres	Male
Occupation(s)	Drummer
Instruments	Drums

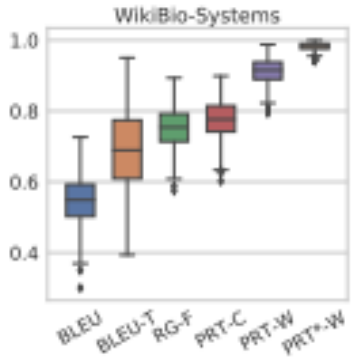
Figure 1: A table from the WikiBio dataset (right), its reference description and three hypothetical generated texts with scores assigned to them by automatic evaluation metrics. Text which cannot be inferred from the table is in **red**, and text which can be inferred but isn't present in the reference is in **green**. PARENT is our proposed metric.

- Candidate 1 is semantically incorrect (it contains information that is not in the input)
- Candidate 2 is semantically correct but does not verbalise all input
- Candidate 3 is semantically correct and mentions more information than Candidate 2

Dhingra et al. ACL 2019

PARENT

Correlates better with human scores than BLEU and ROUGE



- Data-to-Text
  No strict overlap between input data and output text
  E.g., Data: "from Seattle", Text: "in the Seattle band"
- Use entailment rather than token overlap to decide if a text n-gram is entailed by the table
- Precision + Recall
- Precision on union of input data and output text
  Rewards correct information missing from the reference
- Recall on intersection of input data and output reference text
  Penalise incorrect information present in the reference

BLEURT

Trained to predict a Human Score

Transformer model with regression objective

- BLEURT, a **learned evaluation metric** based on BERT that can **model human judgments** with a few thousand training examples
- Learn a function $f : (x, \tilde{x}) \rightarrow y$
  x: reference sentence, $\tilde{x}$: generated sentence with human rating y
- Uses a **novel pre-training scheme** that learns from millions of **synthetic examples** to help the model generalize
- Provides state-of-the-art results on the last three years of the **WMT Metrics shared task** and the **WebNLG** Competition dataset.

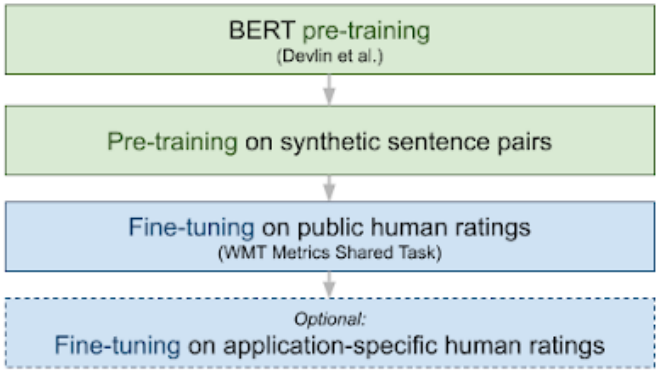
Sellam et al. ACL 2020.

BLEURT Training

Trained in 4 steps

Pre-training

- BERT pre-trained model (LM objective)
- On large quantities of synthetic sentences (NLG objective)



Fine tuning

- On WMT metrics rating (human ratings for translations)
- On Task-specific ratings (e.g., WebNLG human ratings)

BLEURT

- Pre-training
  - of the sentence representations using BERT
  - using synthetic data
- Fine tuning
  The linear layer and the BERT parameters are fine-tuned on the supervised data (reference, prediction, human rating)
- Predict the human rating
  A linear layer on top of the Transformer model is used to predict the rating

Pre-Training on Synthetic Data

- (sentence1, sentence 2, BLEU/ROUGE/...)
- Sentence pairs are created automatically using backtranslation, random deletion and substitution (1.8M segments from Wikipedia)
- Mimick errors of NLG systems proedduce (e.g., omissions, repetitions, non-sensical substitutions)
- Automatic scores (rather than human ratings): 15 objective functions e.g., BLEU, ROUGE, BERTscore, Backtranslation Likelihood, Textual Entailment etc.
- Multi-task loss

BERT Pre-training + Fine tuning on WMT data

- Test Data: Several thousand pairs of sentences with human ratings from the news domain.
- Pearson's correlation and Kendall's tau to measure agreement between the automatic metrics and the human ratings
- Compare BLEURT to participant results from the shared task and automatic metrics
- BLEURT yields state-of-the art performance for all years of the WMT Metrics Shared task

End of Part 1