

Instantly share code, notes, and snippets.

shagunsodhani / [How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation.md](#)

Created 6 years ago


☆ Star

<> Code

🔗 Revisions 1

☆ Stars 1

Notes for paper titled "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation"

 [How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation.md](#)

How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation

Introduction

- The paper explores the strengths and weaknesses of different evaluation metrics for end-to-end dialogue systems(in unsupervised setting).
- [Link to the paper](#)

🔗 Evaluation Metrics Considered

Word Based Similarity Metric

BLEU

- Analyses the co-occurrences of n-grams in the ground truth and the proposed responses.

- BLEU-N: N-gram precision for the entire dataset.
- Brevity penalty added to avoid bias towards short sentences.

METEOR

- Create explicit alignment between candidate and target response (using Wordnet, stemmed token etc).
- Compute the harmonic mean of precision and recall between proposed and ground truth.

ROGUE

- F-measure based on Longest Common Subsequence (LCS) between candidate and target response.

Embedding Based Metric

Greedy Matching

- Each token in actual response is greedily matched with each token in predicted response based on cosine similarity of word embedding (and vice-versa).
- Total score is averaged over all words.

Embedding Average

- Calculate sentence level embedding by averaging word level embeddings
- Compare sentence level embeddings between candidate and target sentences.

Vector Extrema

- For each dimension in the word vector, take the most extreme value amongst all word vectors in the sentence, and use that value in the sentence-level embedding.
- Idea is that by taking the maxima along each dimension, we can ignore the common words (which will be pulled towards the origin in the vector space).

Dialogue Models Considered

Retrieval Models

TF-IDF

- Compute the TF-IDF vectors for each context and response in the corpus.
- C-TFIDF computes the cosine similarity between an input context and all other contexts in the corpus and returns the response with the highest score.

- R-TFIDF computes the cosine similarity between the input context and each response directly.

Dual Encoder

- Two RNNs which respectively compute the vector representation of the input context and response.
- Then calculate the probability that given response is the ground truth response given the context.

Generative Models

LSTM language model

- LSTM model trained to predict the next word in the (context, response) pair.
- Given a context, model encodes it with the LSTM and generates a response using a greedy beam search procedure.

Hierarchical Recurrent Encoder-Decoder (HRED)

- Uses a hierarchy of encoders.
- Each utterance in the context passes through an 'utterance-level' encoder and the output of these encoders is passed through another 'context-level' decoder.
- Better handling of long-term dependencies as compared to the conventional Encoder-Decoder.

Observations

- Human survey to determine the correlation between human judgement on the quality of responses, and the score assigned by each metric.
- Metrics (especially BLEU-4 and BLEU-3) correlate poorly with human evaluation.
- Best performing metric:
 - Using word-overlaps - BLEU-2 score
 - Using word embeddings - vector average
- Embedding-based metrics would benefit from a weighting of word saliency.
- BLEU could still be a good evaluation metric in constrained tasks like mapping dialogue acts to natural language sentences.