Mandate-4 Contributions

Paraphrasing and Semantic matching

Implement an NLP based engine that can paraphrase an input document and says the same sentences in a different way. This engine should also be able to find semantic similarity between sentences of two paraphrased documents.

Datasets used

MSR corpus: https://www.microsoft.com/en-us/download/details.aspx?id=52398

Quora corpus: https://www.sbert.net/examples/training/quora_duplicate_questions/README.html

Google PAWS: https://github.com/google-research-datasets/paws

BART

- Bart uses a standard seq2seq/machine translation architecture with a bidirectional encoder (like BERT) and a left-to-right decoder (like GPT).
- The pretraining task involves randomly shuffling the order of the original sentences and a novel in-filling scheme, where spans of text are replaced with a single mask token.
- BART is particularly effective when fine-tuned for text generation but also works well for comprehension tasks.

Evaluation Metric: ROUGE

ROUGE is an evaluation metric used to assess the quality of NLP tasks such as text summarization and machine translation. Unlike BLEU, the ROUGE uses both recall and precision to compare model generated summaries known as candidates against a set of human generated summaries known as references. It measures how many of the n-grams in the references are in the predicted candidate.

References

1. https://huggingface.co/

2. https://huggingface.co/docs/transformers/training

3. https://pub.towardsai.net/fine-tune-bart-for-translation-on-wmt16-dataset-and-train-new-tokenizer-4d0fbdc4aa2e

4. https://simpletransformers.ai/

5. https://huggingface.co/docs/transformers/model_doc/bart

6. https://medium.com/@priyankads/rouge-your-nlp-results-b2feba61053a