Dense retrieval system for general court laws

Adriana NICOARA, Dalila LADLI, Silviya SILWAL

Institut des Sciences du Digital
Pôle Herbert Simon
13 rue Michel Ney
54000 Nancy, France

February 1st, 2023

Overview

- human evaluation - not very efficient
- domain of the dataset - too specific
- after analyzing more question - contexts pairs we concluded that the result are not satisfactory
- eliminate the human evaluation

Automatic Evaluation

- Using embedding based scoring system, similar to BERTScore[2], DepthScore[1]
- Compute similarity of query and contexts as a average of cosine similarity between their embeddings
- Use different model sentence transformer model[1] unseen by our system to get embeddings
- Compute average score from a list of queries to get overall score
- Compare results among the three models
- Preliminary result based on 14 questions

Models	Mean Similarity Score	Std.
Civile-law-IR	0.88	0.018
STSB	0.83	0.022
DR-Baseline	0.86	0.019

Table: Overall Mean similarity score and Standard deviation for 3 dense retrieval models

[1] https://huggingface.co/nreimers/MiniLM-L6-H384-uncased

Automatic Evaluation

- selected questions not exposed to the model during training
- because we are averaging all of the semantic similarity scores using the same model, the ranking of the contexts is not taken into account
- To address this issue, we only consider the top 15 contexts from a total of 100 that are retrieved and re-ranked by each model
- semantic similarity reflects good performance, but doubts based on human eval

Models	Mean Similarity Score	Std.
Civile-Law-IR	0.91	0.03
STSB	0.87	0.05
DR-Baseline	0.90	0.04

Table: Overall Mean similarity score and Standard deviation on synthetic-nli dataset

Limitations

- Database
- No gold data
- Lack of legal knowledge

Future work

- Include the rest of CASS dataset
- Obtain a database that covers new fields
- Improve the Q/A relevance

References

G. Staerman, P. Mozharovskyi, P. Colombo, S. Clémençon, and F. d'Alché Buc.
A pseudo-metric between probability distributions based on depth-trimmed regions, 2021.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi.
Bertscore: Evaluating text generation with BERT.
CoRR, abs/1904.09675, 2019.

Thank you for your attention !