**SBERT Clarification ranking Evaluation**

**Overview**

To evaluate the performance of our fined-tuned **SBERT model** the ranking of clarification question. A custom dataset is created with GPT containing **75 ambiguous queries.** These are diverse still of queries to test the model a wide varieties of topics like technology, environment, health, education, finance psychology and more. The diverse dataset tests the semantic flexibility of the model simulating real world ambiguity and model’s generalizability beyond it’s familiar domain. Each query were accompanied with their four clarifying questions. Out of which one was the most relevant one. The model was tested to see if I it could come up with the most contextually appropriate follow up and that it understands the intent of the query or not.

The model used here is a fine-tuned version of the **all-MiniLM-L6-v2**, it is a compact SBERT model which was pretrained for ranking clarifying questions by understanding the intent. During fine-tuning the model was trained on 196 positive Query-clarification pair from the Qulac dataset using the **MultipleNegativesRankingLoss**. This setup enabled the model to rank multiple clarifications for a single query, which makes it ideal for retrieval and dialog system like RASA.

**Evaluation Results**

All the 75 ambiguous queries were passed through the trained SBERT model, and the cosine similarity were computed between the query and associated clarifying questions. And the resultant clarification with the highest score was selected as the top prediction made by the model. The prediction is only considered as “confident” if the top similarity score ≥ 0.40 threshold**.**

|  |  |
| --- | --- |
| **Metric** | **Results** |
| **Total Queries** | **75** |
| **Confident predictions (≥ 0.40)** | **63(84%)** |
| **Not confident** | **12(16%)** |
| **Average cosine similarity score(TOP)** | **0.563** |
| **Average Cosine(non-confident)** | **0.38** |

These are the metric used and the results. The standard metrics like Precision, re-call and f1 score are unreliable in this case. As this is a ranking task rather than a classification one.

**Discussion and conclusion**

By the evaluation we can say that the SBERT model showed a strong capability in ranking the contextually appropriate clarification where the ambiguous queries were diverse and different from the original Qulac data the model was trained on. The **84% confident score** and the **0.563 percent** of average similarity shows that the model capable and it is understanding the semantics of the Query-clarification of a new diverse dataset. Also mentioning that the queries in which the model was not confident the average the similarity is around decision threshold. **(0.38)** which indicate the model’s ability to express it uncertainty when there is less semantic alignment. This reduces the risk of returning irrelevant clarifications**.**

Although the model is good at expressing any uncertainties, but it sometimes refrains from answering even when a reasonably close clarification exists. This behaviour can limit the system’s responsiveness in the practical deployment.

The clarification is more subjective and contextual unlike the summary retrieval task. The subjectivity makes it harder to evaluate the model’s performance using the standard classification metrics. This lack of having ground truth can introduce some ambiguity in the interpretation of the results and tuning threshold. This can be improved by training the a model to a massive dataset and it can be finetuned for a particular clarification task as well can be trained on a large amount of diverse data to increase the accuracy of the results overall.

**Reference**

1. **Cosine Similarity Utility – SBERT:** [**https://www.sbert.net/docs/package\_reference/util.html**](https://www.sbert.net/docs/package_reference/util.html)
2. **MultipleNegativesRankingLoss – SBERT loss function documentation.**

[**https://www.sbert.net/docs/package\_reference/losses.html#multiplenegativesrankingloss**](https://www.sbert.net/docs/package_reference/losses.html#multiplenegativesrankingloss)

1. **all-MiniLM-L6-v2 pretrained model. Hugging Face Model Card.** [**https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2**](https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2)
2. **Hugging Face – SentenceTransformers Documentation.** [**https://www.sbert.net/**](https://www.sbert.net/)