Empathetic BERT2BERT Conversational Model Project Proposal

Supervised By: Prof. Arya.

TA. Pouya Khodaee

Group17. Team Members

Last Name, First Name
Abdelazim, Sara
Fahem, Noha
Abdullah, Alaa
Mousa, Naser

Problem formulation:

Our goal for this project is to develop an empathetic Arabic conversational chatbot. As far as we know there are challenges still in Natural Language Understanding regarding Arabic language and this is due to the lack of appropriate datasets available for training. So, in this project, we will try to use a transformer-based encoder-decoder model with AraBERT in an attempt to improve the performance of generating Arabic text responses for the user.

Methodology:

We will try to follow the same approaches they used in the paper for better results with the datasets. These approaches are AraBERT, labeling emotions by using Parrott's classification, seq2seq, and Bi-LSTM.

Data Description and Data Sources:

We will use the ArabicEmpatheticDialogues dataset, which was translated from an English version by Rashkin in 2019. Our dataset contains 3 columns with 36628 row. Our columns are the emotion, context which is supposed to be the one user sends, and the response that is generated by the chatbot.

Evaluation Methods:

Numerical Evaluation: The models were evaluated using automated metrics such as Perplexity (PPL) and Bilingual Evaluation Understudy (BLEU) scores. These metrics provide quantitative measures of model performance in terms of language fluency and similarity to reference responses.

Human Evaluation: To assess the models' ability to exhibit empathetic behavior, human evaluation was conducted. Native Arabic speakers were asked to rate the generated responses on three aspects: Empathy, Relevance, and Fluency. Ratings were collected on a scale of 0 to 5, with 0 indicating poor performance and 5 indicating excellent performance. The average ratings were calculated for each model based on the responses received.

These evaluation methods provide both quantitative and qualitative assessments of the models' performance, taking into account language quality, relevance to the input utterance, and the ability to convey empathy in the generated responses

Expected Results:

1) **Proposed BERT2BERT Model:**
    - The BERT2BERT model is expected to leverage knowledge transfer and show enhanced performance in empathetic response generation.
    - By initializing the model's encoder and decoder with AraBERT pre-trained weights, it is expected to achieve improved performance compared to the baseline model.

- The proposed model is expected to exhibit empathy while generating relevant and fluent responses in open-domain settings.

2) **Evaluation Metrics:**
    - Numerical evaluation using perplexity (PPL) and Bilingual Evaluation Understudy (BLEU) scores is conducted to compare the proposed BERT2BERT model with benchmark models.
    - The lower the PPL score and the higher the BLEU score, the better the model's performance is considered.
3) **Benchmark Models:**
    - Baseline Model: A Seq2Seq Bi-LSTM model with attention, following the previous state-of-the-art model.
    - EmoPrepend Model: The emotion labels are prepended to the input utterances before feeding them into the baseline model.
    - BERT2BERT-UN Model: A regular transformer-based encoder-decoder model without initialization with pre-trained weights.
4) **Experimental Setup:**
    - The proposed BERT2BERT model is trained for 5 epochs with a batch size of 32.
    - The training and evaluation are conducted on common data splits of the ArabicEmpatheticDialogues dataset.
5) **Numerical Evaluation Results:**
    - The BERT2BERT model consistently outperforms the benchmark models in terms of lower perplexity (PPL) scores and higher BLEU scores.
    - The EmoPrepend model shows improvements compared to the baseline model but still has a relatively high PPL score.
    - The BERT2BERT-UN model performs poorly, indicating the importance of pre-training with AraBERT weights.
6) **Human Evaluation Results:**
    - Human evaluation is conducted to assess the models' empathy, relevance, and fluency in the generated responses.
    - The BERT2BERT model receives higher average ratings for empathy, relevance, and fluency compared to the baseline and EmoPrepend models.
7) **Performance on Inputs with Neutral Emotional States:**
    - The BERT2BERT model struggles to handle regular chit-chat utterances with neutral emotional states, generating empathetic responses instead of regular responses.
    - This limitation is attributed to the training data, which primarily focuses on emotional contexts, and the AraBERT pre-training, which lacks chit-chat samples.