RoBERTa vs BERT for intent classification

Gabriel GUAQUIERE
ENSAE Paris
gabriel.guaquiere@ensae.fr

Anna NGUYEN THANH SON
ENSAE Paris
anna.nguyen-thanh-son@ensae.fr

Abstract

Intent classification is an essential task in natural language processing, which aims to identify the intention or purpose behind a user's utterance. This task has become increasingly important in the development of conversational agents and chatbots, as they need to understand user requests to provide relevant and accurate responses. In this paper[1], we will look at which algorithm, between BERT and RoBERTA, seems more adapted to the prediction of dialogue act (DA) or sentiment and emotion (S/E). We will use the SILICONE dataset which seems to fit the task. It contains corpora including an utterance and the associated DA or S/E. We observe that RoBERTa outperforms BERT in prediction, especially for DA. For the mrda dataset, it even manages to reach an accuracy of 89%. For the prediction of S/E, its performance is better than BERT, nevertheless its predictivity rate is low.

1 Problem Framing

The identification of both Dialog Acts (DA) and Emotion/Sentiment (E/S) in spoken language plays a crucial role in enhancing the quality of automated dialog systems [1, 2]. By identifying DAs, such as questions or statements, and E/S, such as happiness or sadness, these systems can generate more appropriate and context-specific responses.

To identify DAs and E/S in spoken language [3], sequence labeling systems trained in a supervised manner are used [4]. These systems are trained on large annotated datasets of spoken language to learn how to associate different linguistic features, such as sentence structure and intonation, with specific DAs and E/S.

Overall, identifying DAs and E/S [5, 6] in spoken language is an essential step toward developing more effective and context-aware automated dialog systems. By improving the ability of these systems to understand and respond to user needs and emotions, we can enhance the overall quality of human-machine interactions.

BERT and ROBERTA are two such models, which have set the standard for SOA intent classification due to their performance and advanced transformer architectures. These models have achieved good results in various domains and are widely used for web search, chatbots, or voice assistants [7, 8].

The purpose of this project is to compare the performance of BERT and ROBERTA on similar datasets with the goal of identifying which model is better suited for DA and E/S identification.

2 Experiments Protocol

We will use two pre-trained models: BERT and RoBERTa and apply them on SILICON datasets to predict DA or E/S. Afterwards, we will compare the results of the two models.

2.1 Presentation of SILICONE data

The collection of sequence labelling tasks known as SILICONE comprises both DA and E/S annotated datasets, which are derived from existing datasets that are highly regarded for their complexity and interest to the research community.

We choose to work on this data after having been inspired by the SOA paper about Sequence Labelling [9] in which the authors obtained better accuracy on this dataset than BERT or ROBERTA.

2.1.1 DA Datasets

ICSI MRDA Corpus (MRDA) [10] contains transcripts of multi-party meetings handannotated with DA (110k utterances).

[1] https://github.com/Zingwompe/NLP-project-

DailyDialog Act Corpus (DyDAa) [11] contains multiturn dialogues, supposed to reflect daily communication by covering topics about daily life (102k utterances).

HCRC MapTask Corpus (MT) [12] To build this corpus, participants were asked to collaborate verbally by describing a route from a first participant's map by using the map of another participant (27k utterances).

Bt Oasis Corpus (Oasis) [13] contains the transcripts of live calls made to the BT and operator services (15k utterances).

2.1.2 S/E Datasets

DailyDialog Emotion Corpus (DyDAe) has been previously introduced and contains eleven emotional labels.

Multimodal EmotionLines Dataset (MELD) [14] To build this corpus multiple speakers participated in the dialogues. There are two types of annotations MELDs and MELDe: three sentiments (positive, negative and neutral) and seven emotions (anger, disgust, fear, joy,neutral,sadness and surprise).

IEMOCAP database (IEMO) [15] is a multimodal database of ten speakers. It consists of dyadic sessions where actors perform improvisations or scripted scenarios.

SEMAINE database (SEM) [16] comes from the Sustained Emotionally coloured Machine human Interaction using Nonverbal Expression project. This dataset has been annotated on three sentiments labels: positive, negative and neutral.

2.2 BERT

BERT [17] is a deep neural network that uses a transformer-based architecture, which is a type of recurrent neural network (RNN) that is designed to handle long-range dependencies in sequential data, such as natural language text. BERT is pre-trained on a large corpus of text data (Wikipedia) using an unsupervised learning approach.

During pre-training, BERT is trained to predict masked words in a given sentence (MLM) or to predict the next sentence (NSP) in a pair of sentences. This pre-training process allows the model to learn rich representations of the meaning and relationships between words in a sentence, which can then be fine-tuned on specific NLP tasks such as intent classification.

To fine-tune BERT for intent classification, a dataset is required that includes labeled examples of user queries and their corresponding intents. The model is then trained on this dataset using supervised learning, where the weights and biases of the model are adjusted to minimize the difference between the predicted intent labels and the true labels in the dataset.

During training, the input text is first tokenized into a sequence of subword tokens using the WordPiece tokenization algorithm. The tokens are then passed through several layers of transformer blocks, each of which applies multi-head attention and feedforward transformations to the input tokens. The output of the last transformer block is a sequence of encoded vectors, one for each token in the input text.

The encoded vectors for the special [CLS] token, which is added to the beginning of the input text, are then used as input to a final classification layer, which produces a probability distribution over the possible intent labels. The model is trained to minimize the cross-entropy loss between the predicted probability distribution and the true labels in the training data.

During inference, the model takes a new user query as input, tokenizes it using the same WordPiece algorithm, and passes the resulting sequence of subword tokens through the same transformer blocks as during training. The encoded vector for the [CLS] token is then fed into the final classification layer, which produces a probability distribution over the possible intent labels for the query.

One of the key advantages of BERT for intent classification is its ability to capture the context and relationships between words in a sentence, which allows it to handle complex queries and variations in user language. Additionally, because BERT is pre-trained on large amounts of text data, it can be fine-tuned on small amounts of labeled data for specific tasks, making it a highly flexible and effective model for intent classification.

2.2.1 BERT uncased L12 H768 A12

This project utilizes a specific version of the BERT model architecture, namely BERT uncased L12 H768 A12.

The name "uncased" indicates that the model was trained on text that has been converted to lowercase, resulting in the treatment of uppercase and

lowercase letters as equivalent.

The variant's name also contains several other important details about its architecture.

"L12" denotes the number of layers in the model, which is 12 in this case. These layers consist of transformer blocks that help the model understand the relationships between different words in a given piece of text.

"H768" indicates the size of the hidden layer in the model, with this variant featuring 768 hidden units in each layer.

"A12" refers to the number of attention heads in the model, with this variant having 12 attention heads. Attention is a mechanism that allows the model to focus on specific parts of the input sequence that are most relevant for a given task.

2.3 RoBERTa

RoBERTa [18]: Robustly Optimized BERT-Pretraining Approach proposed in Liu et. al. is an extension to the original BERT model. Like BERT, RoBERTa is a natural language processing model based on Transformer neural networks. Recently work suggest that BERT is under-trained and Liu et. al. will therefore fine-tune it in four aspects :
BERT's performance increases greatly on larger datasets. So they will train it on a larger dataset (160 GB of uncompressed text).

Previous work has shown that training with very large mini-batches can both improve optimization speed and performance [19]. Recent work has shown that this also applies for BERT [20]. RoBERTa will therefore be trained on a higher batch size.

RoBERTa uses a different token masking strategy than BERT. Instead of statically masking certain tokens, RoBERTa randomly masks some tokens at each pass, forcing the model to learn to use all available contextual information to predict the masked tokens

In the paper by, the authors hypothesized that NSP loss is an important factor in BERT training. Nevertheless, recent studies have questioned the necessity of NSP loss. RoBERTa therefore excludes NSP.

2.3.1 Model

We worked with *Tensorflow* library on Python and used RoBERTa base model. This model has 12 layers of transformers, each with 768 hidden units and 12 attention heads. Each layer of the transformer consists of two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. The self-attention mechanism allows the model to attend to different parts of the input sequence when generating the output representation. The multi-head attention mechanism in each transformer layer allows the model to attend to different positions in the input sequence simultaneously, enabling more effective representation learning.

For the encoding of our statements, we used the Roberta tokenizer. It uses Byte Pair Encoding (BPE) to generate the vocabulary of subword units used for tokenization. BPE is a method for generating a vocabulary of subword units by iteratively merging the most frequently occurring pairs of characters in a corpus.

Roberta, being a transformer-based language model, uses the GELU (Gaussian Error Linear Unit) activation function in its feed-forward network layers. In addition to the GELU activation function, Roberta also uses the softmax activation function in its final layer for tasks such as text classification and sequence labeling, where the output is a probability distribution over a set of classes.

2.4 Loss

For the evaluation of our models, we used the cross-entropy (or log-loss) function. It is used when the variable to be predicted is categorical. It measures the performance of a classification model whose output is a probability value between 0 and 1. Cross-entropy loss increases as the predicted probability diverges from the actual label. It calculates a separate loss for each class label per observation and sum the result and is defined as follows:

$$L = -\sum_{c=1}^{M} y_{o,c} log(p_{o,c}) \tag{1}$$

where M is the number of label, $y_{o,c}$ is binary indicator if class label c is the correct classification for observation o and $p_{o,c}$ is the predicted observation o of class c.

2.5 Accuracy

In addition to the loss function, we used the accuracy on test set to evaluate our models and is defined as follows:

$$Accuracy = \frac{1}{|Testset|} \sum_{X_i in Testset} 1_{\hat{Y}_i = Y_i} \quad (2)$$

where X_i can either represent utterance, Y_i a list of DA or E/S and $\hat{Y}_i$ is the label predicted by our model,

3 Results

In this section we will compare the performance of BERT and ROBERTA in DA/ES identification tasks on similar datasets

3.1 DA classification

3.1.1 DA classification on all DA categories

The following are the outcomes we achieved on the previously mentioned datasets. Initially, we didn't perform any data cleaning; instead, we picked the entire dataset and divided it into test, validation, and train subsets. Afterward, we selected only the "utterances" and "dialog act" columns and applied the BERT and ROBERTA models. We selected the models that have low loss validation while having high accuracy.

Dataset	BERT val accuracy	RoBERTa val accuracy
dyda da	75%	87%
maptask	45%	66%
mrda	59%	89%
oasis	29%	72%
Average	52%	79%

Dataset	BERT val loss	RoBERTa val loss
dyda da	0.99	0.35
maptask	2.19	1.00
mrda	1.32	0.29
oasis	3.49	1.05
Average	2.00	0.68

We noticed that the ROBERTA model consistently performs better than BERT on raw SILICONE datasets, achieving higher accuracy and lower loss values. Moreover we observe that with RoBERTa, we do not suffer from the overlearning problem, the validation loss decreases with the number of epochs.

3.1.2 DA classification on most relevant DA categories

The following are the outcomes we achieved on the **Maptask** dataset but choosing only the most represented DA categories.
We choose the following DA : acknowledge, instruct, reply y, explain, check, ready, align, query y, those categories represents 80% of the whole dataset.

Dataset	BERT val accuracy	ROBERTA val accuracy
maptask	67%	71%

Dataset	BERT val loss	RoBERTa val loss
maptask	1.5	2.6

ROBERTA outperforms BERT in accuracy even when only 80% of the data is chosen based on the most common dialog acts. However, this time, ROBERTA's loss is more significant than that of BERT.

3.2 E/S classification

Dataset	BERT val accuracy	RoBERTa val accuracy
dyda_de	81%	81%
meld_e	46%	64%
meld_s	67%	70%
iemocap	24%	40%
Average	55%	64%

Dataset	BERT val loss	RoBERTa val loss
dyda_de	1.35	0.51
meld_e	1.7	1.00
meld_s	0.89	1.08
iemocap	1.91	1.60
Average	1.46	1.05

Our observation is that the RoBERTa model consistently outperforms BERT on SILICONE datasets when doing Emotion Sentiment classification, achieving higher accuracy and lower loss values. We have also noticed that RoBERTa does not suffer from overfitting, as the validation loss continues to decrease with each epoch.

4 Discussion/Conclusion

The objective of this project was to employ the two most prevalent NLP models for classification, and conduct intent classification and sentiment analysis using them. We compared the performance of these models. We conclude that RoBERTa outperforms BERT. Depending on the dataset, the accuracy can reach 89%. As a whole RoBERTa seems to have a good predictive power for the dialog act, its average accuracy is 79%. However, for the prediction of emotions or feelings, it seems to be less adapted, its average accuracy is only 64%. On the other hand, BERT seems to work better on Sentiment analysis, its average accuracy is 55% whereas it is 52% on Dialog act classification.

To enhance accuracy and reduce loss in our models, we could have improved our data preprocessing techniques or implement research papers that achieved state-of-the-art accuracy on the datasets we used. For instance, it is the case of **Guiding attention in sequence**

to-sequence models for dialogue act prediction from Colombo et. al. (2020) that have reached an accuracy of 85.5% on **SwDa** data, or the paper **A dual attention hierarchical recurrent neural network for dialogue act classification** from Li and al (2019) in which they have reached an accuracy of 92.2% on **mrda** data where we did at best 89%.

References

[1] Pierre Colombo*, Wojciech Witon*, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. Affect-driven dialog generation. *NAACL 2019*, 2019.

[2] Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloé Clavel. Code-switched inspired losses for spoken dialog representations. In *EMNLP 2021*, 2021.

[3] Tanvi Dinkar*, Pierre Colombo*, Matthieu Labeau, and Chloé Clavel. The importance of fillers for text representations of speech transcripts. *EMNLP 2020*, 2020.

[4] Emile Chapuis*, Pierre Colombo*, Matteo Manica, Matthieu Labeau, and Chloe Clavel. Hierarchical pre-training for sequence labelling in spoken dialog. *Finding of EMNLP 2020*, 2020.

[5] Alexandre Garcia*, Pierre Colombo*, Slim Essid, Florence d'Alché Buc, and Chloé Clavel. From the token to the review: A hierarchical multimodal approach to opinion mining. *EMNLP 2019*, 2019.

[6] Wojciech Witon*, Pierre Colombo*, Ashutosh Modi, and Mubbasir Kapadia. Disney at iest 2018: Predicting emotions using an ensemble. In *Wassa @EMNP2018*, 2018.

[7] Hamid Jalalzai*, Pierre Colombo*, Chloé Clavel, Éric Gaussier, Giovanna Varni, Emmanuel Vignon, and Anne Sabourin. Heavy-tailed representations, text polarity classification & data augmentation. *NeurIPS 2020*, 2020.

[8] Pierre Colombo. *Learning to represent and generate text using information measures*. PhD thesis, (PhD thesis) Institut polytechnique de Paris, 2021.

[9] Pierre Colombo*, Emile Chapuis*, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloe Clavel. Guiding attention in sequence-to-sequence models for dialogue act prediction. *AAAI 2020*, 2020.

[10] Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100, Cambridge, Massachusetts, USA, April 30 - May 1 2004. Association for Computational Linguistics.

[11] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset, 2017.

[12] Henry Thompson, Anne Anderson, Ellen Bard, Gwyneth Doherty-Sneddon, Alison Newlands, and Cathy Sotillo. The hcrc map task corpus: natural dialogue for speech recognition. 01 1993.

[13] Geoffrey Leech and Martin Weisser. Generic speech act annotation for task-oriented dialogues. 2003.

[14] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations, 2018.

[15] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359, 12 2008.

[16] Gary Mckeown, Michel Valstar, Roddy Cowie, Maja Pantic, and M. Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3:5–17, 08 2013.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[19] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*, 2019.

[20] Yang You, Jing Li, Jonathan Hseu, Xiaodan Song, James Demmel, and Cho-Jui Hsieh. Reducing bert pre-training time from 3 days to 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019.