CSCI 6364 – Machine Learning

Spring 2022

Final Project

Jesse Robles-Alcaraz

<https://github.com/jesserobles/query-intent-classifier>

# Comparison of BERT and DIET on Intent Classification and Named Entity Recognition

## Abstract

User expectations regarding their interactions with search systems and dialog systems (e.g., chatbots, Siri, Cortana) have significantly increased in the last decades beyond simple keyword matching and rule-based responses. A main driver for this has been large corporations such as Google, Amazon, and Microsoft’s significant investment in research and development of machine learning algorithms and hardware. This has led to the development of embedding techniques and language models, followed by the development of the Transformer architecture which further pushed the state of the art in natural language understanding. Notable among these is the Bidirectional Encoder Representations from Transformers (BERT) model. These models are trained on general corpora of text data (e.g., new articles, Wikipedia) and can be fine-tuned to specific use cases to achieve better performance. This architecture has also been adapted to more domain specific applications such as search and conversational AI, such as the Rasa Dual Intent and Entity Transformer (DIET). This project compares the out-of-the-box performance of BERT, DIET, and DIET with a language model for both query intent classification and Named Entity Recognition (NER) tasks on six datasets. The results on these datasets show that model performance is very close for all three, although BERT slightly outperformed both DIET models in NER. For query intent classification, DIET with a language model and BERT each outperformed the other on 3 of the datasets, but BERT had higher overall averages in the evaluation metrics. The DIET model without language model also outperformed BERT on one of the datasets. Additionally, the DIET architecture provides the added benefit of a single model for both tasks and a significantly smaller resulting model size, making it a compelling choice overall.

## Background

Some recent research indicates that for specific domains where the language style differs from the data used to train BERT, training a transformer architecture from scratch can approach the performance of these larger architectures, some instances outperforming them. This project is based on one of the papers (Bunk, et al. 2020) reviewed, which showed that training a custom model, named the Dual Intent and Entity Transformer (DIET) from scratch outperforms BERT with fine tuning on intent classification and entity recognition. However, this was after what appears to be some fine tuning on the DIET model. One key feature of the DIET architecture is that it allows combining dense embedding representations with sparse features and character level n-grams easily. Below is a graphic representing the DIET architecture:

A screenshot of a computer

Description automatically generated with medium confidence

The goal of this project is to compare the out-of-the box performance of each of these models from a practitioner’s perspective. Thus, part of this project was to identify some of the datasets that were used in that paper (SNIPS, ATIS) and other datasets to train and compare these two models. The BERT model used was trained using the Huggingface library, and the DIET model was trained using the Rasa python conversational AI library. The datasets are described in the section and table below.

## Data Sets and Data Preprocessing

The datasets were sourced from the GitHub repository at <https://github.com/jianguoz/Few-Shot-Intent-Detection> and from Kaggle at <https://www.kaggle.com/datasets/joydeb28/nlp-benchmarking-data-for-intent-and-entity>. These datasets are somewhat similar in that they are intended to be used in a conversational assistant or search setting. The ATIS and SNIPS datasets were used in (Bunk, et al. 2020), and although the NLU-Benchmarking dataset from that paper was not used in this project, the other datasets (CLINC150, HWU64) contain examples from that dataset. Table 1 shows a brief description of each dataset.

Table 1: Descriptions of Datasets

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset Name** | **Has Entities** | **Original Format** | **Description** |
| ATIS | Y | IOB | Airline travel information systems. Contains intents and entities. Examples: "Find cheapest flights from BWI to MCO" |
| BANKING77 | N | Sample/Labels | Intents in the banking domain. No entities. Examples: "Still waiting on my new card" |
| BENCHMARKING\_DATA | Y | JSON | Kaggle dataset for benchmarking intents and entities. Examples: "Book restaurant", "Add song to playlist" |
| CLINC150 | N | Sample/Labels | A dataset with 10 domains and 150 intents. No entities. Style is similar to personal assistant. Examples "how do you say dog in spanish" |
| HWU64 | N | Sample/Labels | Personal assistant data with 64 intents and several domains. No entities. Examples: "remind me about the meeting" |
| SNIPS | Y | IOB | Personal assistant data with 7 intents. Includes entities. Examples: rate this current novel 1 stars" |

Most of the NER datasets were in [IOB format](https://en.wikipedia.org/wiki/Inside%E2%80%93outside%E2%80%93beginning_(tagging)), and one (BENCHMARKING\_DATA) was in JSON format. The first challenge was to preprocess the datasets so that each model type can accept the data. It made more sense to handle a single format, so the JSON parser processes the JSON files and outputs in the same IOB format as the other datasets. Additionally, all but the BENCHMARKING\_DATA dataset were split into train, test, validate datasets. That dataset was split into the same format as it was processed into IOB format.

The Huggingface library data input format for Named Entity Recognition is like the IOB format in that it takes in tokenized text and the NER label for each token. The main preprocessing step is to encode the IOB format into numeric values. The input format for the Rasa DIET model is a YAML file that contains examples grouped by intent and with entities annotated in the text example. The CoNLL/IOB parser in the code base handles preprocessing the input files and outputting the proper format for each model type. For DIET, new YAML files are created.

## Model Training and Model Comparison

Since the Rasa DIET model performs both query intent classification and named entity recognition, two BERT models were trained on each of the datasets that contained both intents and entities. Although there are ways (such as the one described [here](https://towardsdatascience.com/how-to-create-and-train-a-multi-task-transformer-model-18c54a146240)) to use the Huggingface library to build a multi-task model, it is not supported out of the box and building that architecture is outside the scope of this project.

### BERT

In order to reduce model size and training time, the distilbert-base-uncased model was used as the base model for fine tuning, which is a smaller and faster version of BERT. The BERT models were trained for 5 epochs with a learning rate of 2e-5 and weight decay of 0.01, which is a form of regularization. These parameters were selected based on the Huggingface documentation. Training time on the BERT model ranged from 1-5 minutes on a 1080TI GPU.

### DIET

For the DIET models, each model was trained on 100 epochs with the other parameters set with default values. For each dataset, a default DIET model was trained, as well as a model that included a ConverTFeaturizer, which uses a pre-trained language model to convert the input text into embeddings before inputting them into the model. This was shown in the paper to improve performance. In terms of training time, the DIET models take longer because of the larger number of epochs. These epochs are necessary because the models are trained from scratch. To verify that 100 epochs was enough, the HWU64 dataset was also used to train the DIET model for 200 epochs to see if it improved model performance, but no notable improvement was observed. The training times for diet ranged from 2 to 15 minutes when trained on a 1080TI GPU. The ATIS dataset had to be trained on CPU for the DIET model because it caused the GPU to run out of memory. This is likely the result of the number of entity types and number of entities present in that dataset.

### Model Size

The difference in model sizes is also notable. The average size of the DIET models is approximately 86MB, while the BERT models are all approximately 235MB each. This might be an important consideration for limited resource environments such as mobile or edge devices.

## Comparison of Results

Below are two charts showing the precision, recall, F1 score and accuracy of each model in intent classification and named entity recognition, where applicable. The seqeval python library was used on the NER evaluation and the sklearn classification report was used on the query intent evaluation. On these datasets, the models were generally comparable to each other. For query intent classification, the BERT model slightly outperformed the default DIET model on 5 of the 6 datasets on all metrics but F1 score. The DIET with language model outperformed BERT on BANKING77, CLINC150, and HWU64. Further fine-tuning of the rasa model might close the gap between the models, but it is notable that the model performs almost as well as BERT with a significantly smaller dataset than even what distilbert is trained on. Additional data regarding the performance on each dataset is available in the Appendix, and on performance for specific intents and entity types is available in the [results](https://github.com/jesserobles/query-intent-classifier/tree/main/results) folder of the GitHub repository.

## Conclusions and Future Work

This project demonstrated that the BERT and DIET models provide comparable results out-of-the-box on the six datasets reviewed. BERT tends to outperform DIET in NER, which might be the result of more semantic information being available within the model itself that can help with understanding unseen examples. The DIET models showed promising results, particularly when adding a language model to it. Doing so managed to either approach or surpass the performance of BERT on query intent classification. However, it should be noted that these models were trained with mostly default hyperparameters and minimal tuning. The DIET performance is quite admirable given the fact that it is trained on a relatively small dataset compared to BERT. Additionally, it provides a very compelling performance to model size ratio. It is also important to note that the authors of (Bunk, et al. 2020) are Rasa employees, and the results of that work might be biased towards that company’s product.

For future work, it would be interesting to build a multi-task Huggingface model to perform sequence and token classification (i.e., NER), train the model on this dataset and compare the performance of the models. Furthermore, additional fine tuning of the DIET model would be interesting to see whether it can close the gap with BERT on NER. Furthermore, training the models on additional datasets and comparing the results would also be enlightening.

# References

Bunk, Tanja, Daksh Varshneya, Vladimir Vlasov, and Alan Nichol. 2020. "DIET: Lightweight Language Understanding for Dialogue Systems."

Guo, Weiwei, Xiaowei Liu, Sida Wang, Huiji Gao, Ananth Sankar, Zimeng Yang, Qi Guo, et al. 2020. "DeText: A Deep Text Ranking Framework with BERT." *CoRR* abs/2008.02460. https://arxiv.org/abs/2008.02460.

Larson, Stefan, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, et al. 2019. "An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction."

Tejaswini Mallavarapu, Ying Xie, and Simon Hughes. 2022. "CatBERT: An Incrementally Trained Language Representation Model for E-Commerce Applications." *Proceedings of the International Workshop on Interactive and Scalable Information Retrieval methods for eCommerce (ISIR-eCom).*

# Appendix

Below are tables showing the different performance metrics on the validation/evaluation dataset for both BERT and DIET.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Intent Classification** | | | | |
| **Model** | **precision** | **recall** | **f1-score** | **accuracy** |
| BERT | 0.953135 | 0.952023 | 0.898755 | 0.952023 |
| DIET | 0.941358 | 0.937397 | 0.936648 | 0.937397 |
| DIET (w/ Language Model) | 0.947146 | 0.944301 | 0.943761 | 0.944301 |
|  |  |  |  |  |
| **NER** | | | | |
| **Model** | **precision** | **recall** | **f1-score** | **accuracy** |
| BERT | 0.947154 | 0.955963 | 0.951536 | 0.975778 |
| DIET | 0.741434 | 0.819272 | 0.777427 | 0.900818 |
| DIET (w/ Language Model) | 0.741452 | 0.823024 | 0.779096 | 0.902911 |

Table 2: Overall Results (Averages Across Datasets)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Intent Classification** | | | | | | |
| **Model** | **Type** | **precision** | **recall** | **f1-score** | **support** | **accuracy** |
| **BERT** | macro avg | 0.883003572 | 0.883660131 | 0.875604635 | 500 | 0.978 |
| weighted avg | 0.978767642 | 0.978 | 0.977314352 | 500 |
| **DIET** | macro avg | 0.724491942 | 0.734692067 | 0.726138321 | 500 | 0.962 |
| weighted avg | 0.96440009 | 0.962 | 0.961849211 | 500 |
| **DIET (w/ Language Model)** | macro avg | 0.843443848 | 0.800303455 | 0.817195899 | 500 | 0.962 |
| weighted avg | 0.966214332 | 0.962 | 0.963383521 | 500 |
|  |  |  |  |  |  |  |
|  | **NER** | | | | |  |
|  | **Model** | **precision** | **recall** | **f1-score** | **accuracy** |  |
|  | **BERT** | 0.959240648 | 0.964085297 | 0.961656871 | 0.9874 |  |
|  | **DIET** | 0.938399539 | 0.953774137 | 0.946024376 | 0.9809 |  |
|  | **DIET (w/ Language Model)** | 0.934595525 | 0.953188999 | 0.943800695 | 0.9828 |  |

Table 3: ATIS Dataset Results

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Intent Classification** | | | | | | |
| **Model** | **Type** | **precision** | **recall** | **f1-score** | **support** | **accuracy** |
| **BERT** | macro avg | 0.884329492 | 0.885714286 | 0.881055176 | 1540 | 0.885714286 |
| weighted avg | 0.884329492 | 0.885714286 | 0.881055176 | 1540 |
| **DIET** | macro avg | 0.911498555 | 0.906493506 | 0.906296759 | 1540 | 0.906493506 |
| weighted avg | 0.911498555 | 0.906493506 | 0.906296759 | 1540 |
| **DIET (w/ Language Model)** | macro avg | 0.931394505 | 0.925324675 | 0.925604791 | 1540 | 0.925324675 |
| weighted avg | 0.931394505 | 0.925324675 | 0.925604791 | 1540 |

Table 4: BANKING77 Dataset Results

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Intent Classification** | | | | | | |
| **Model** | **Type** | **precision** | **recall** | **f1-score** | **support** | **accuracy** |
| **BERT** | macro avg | 0.998626374 | 0.998613037 | 0.998613005 | 1448 | 0.99862 |
| weighted avg | 0.998632065 | 0.998618785 | 0.998618752 | 1448 |
| **DIET** | macro avg | 0.986155184 | 0.98621604 | 0.986165909 | 1448 | 0.98619 |
| weighted avg | 0.986169902 | 0.986187845 | 0.98615917 | 1448 |
| **DIET (w/ Language Model)** | macro avg | 0.988465414 | 0.988276864 | 0.988310595 | 1448 | 0.98826 |
| weighted avg | 0.988457271 | 0.988259669 | 0.988298157 | 1448 |
|  |  |  |  |  |  |  |
|  | **NER** | | | | |  |
|  | **Model** | **precision** | **recall** | **f1-score** | **accuracy** |  |
|  | **BERT** | 0.93860405 | 0.949858295 | 0.944197638 | 0.9704 |  |
|  | **DIET** | 0.649005142 | 0.776618513 | 0.707100231 | 0.877 |  |
|  | **DIET (w/ Language Model)** | 0.647424512 | 0.780096308 | 0.707595244 | 0.876 |  |

Table 5: BENCHMARKING\_DATA Dataset Results

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Intent Classification** | | | | | | |
| **Model** | **Type** | **precision** | **recall** | **f1-score** | **support** | **accuracy** |
| **BERT** | macro avg | 0.958421166 | 0.954666667 | 0.954380975 | 3000 | 0.954666667 |
| weighted avg | 0.958421166 | 0.954666667 | 0.954380975 | 3000 |
| **DIET** | macro avg | 0.923669184 | 0.915666667 | 0.913385331 | 3000 | 0.915666667 |
| weighted avg | 0.923669184 | 0.915666667 | 0.913385331 | 3000 |
| **DIET (w/ Language Model)** | macro avg | 0.968550309 | 0.966333333 | 0.966232497 | 3000 | 0.966333333 |
| weighted avg | 0.968550309 | 0.966333333 | 0.966232497 | 3000 |

Table 6: CLINC150 Dataset Results

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Intent Classification** | | | | | | |
| **Model** | **Type** | **precision** | **recall** | **f1-score** | **support** | **accuracy** |
| **BERT** | macro avg | 0.894813464 | 0.894396473 | 0.891438594 | 1076 | 0.907992565 |
| weighted avg | 0.910906613 | 0.907992565 | 0.906564848 | 1076 |
| **DIET** | macro avg | 0.885742409 | 0.870893216 | 0.870584765 | 1076 | 0.875464684 |
| weighted avg | 0.883007143 | 0.875464684 | 0.87355475 | 1076 |
| **DIET (w/ Language Model)** | macro avg | 0.912716539 | 0.905107658 | 0.90548513 | 1076 | 0.92472119 |
| weighted avg | 0.925858206 | 0.92472119 | 0.922717492 | 1076 |

Table 7: HWU64 Dataset Results

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Intent Classification** | | | | | | |
| **Model** | **Type** | **precision** | **recall** | **f1-score** | **support** | **accuracy** |
| **BERT** | macro avg | 0.987752859 | 0.987142857 | 0.987149467 | 700 | 0.987142857 |
| weighted avg | 0.987752859 | 0.987142857 | 0.987149467 | 700 |
| **DIET** | macro avg | 0.979405547 | 0.978571429 | 0.978642794 | 700 | 0.978571429 |
| weighted avg | 0.979405547 | 0.978571429 | 0.978642794 | 700 |
| **DIET (w/ Language Model)** | macro avg | 0.984184133 | 0.982857143 | 0.982961141 | 700 | 0.982857143 |
| weighted avg | 0.984184133 | 0.982857143 | 0.982961141 | 700 |

Table 8: SNIPS Dataset Results