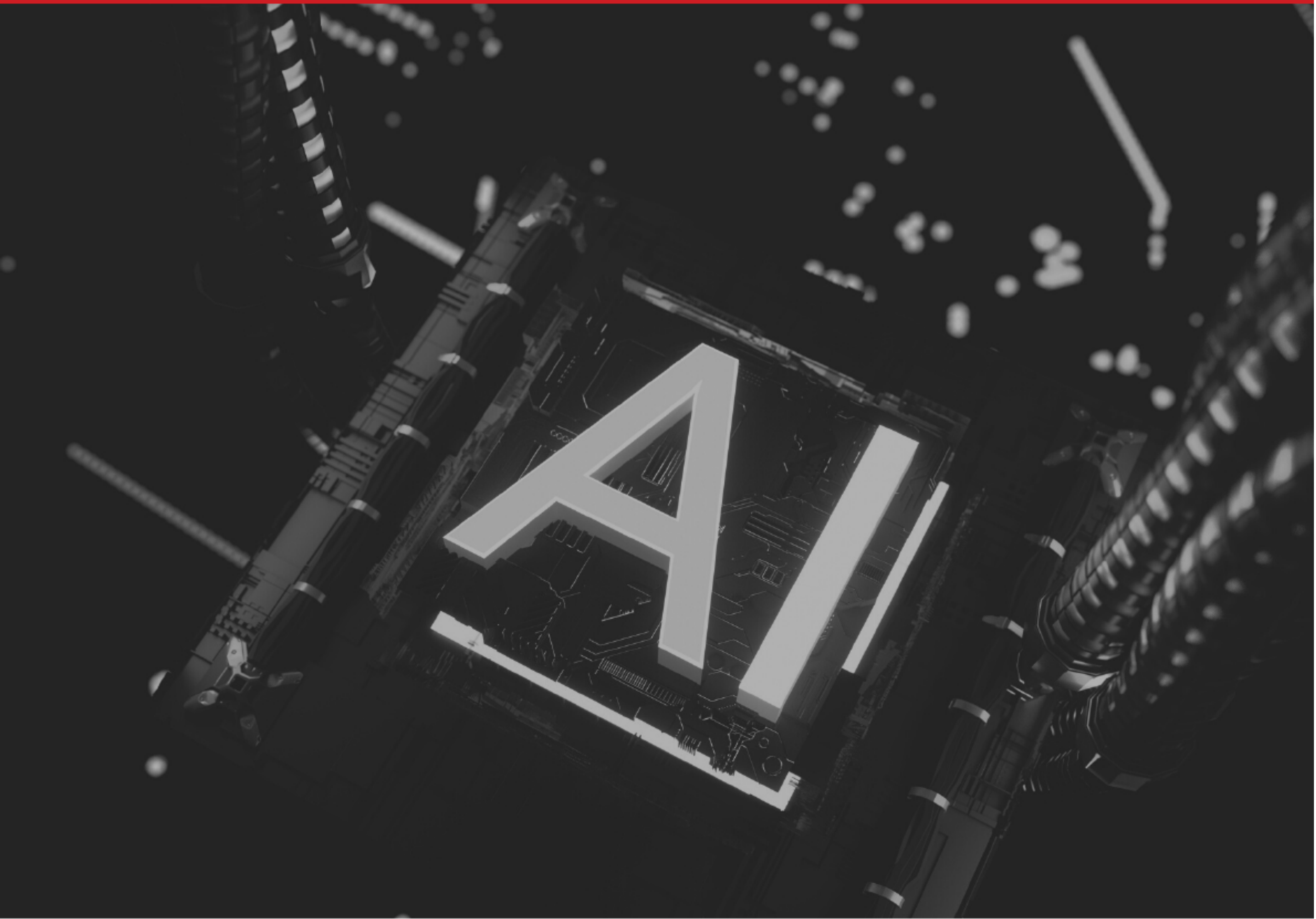
bitext we help AI understand humans

Benchmarking Rasa using LLMs



Introduction

Recently, large language models (LLMs) have shown state-of-the-art performance when used to solve different language tasks. The success of ChatGPT, which was initially based on the GPT 3.5 model, has drawn public attention to the language understanding and generation capabilities of LLMs. We at Bitext have worked on integrating our intent detection data for chatbots with different state-of-the-art LLMs.

To do this, we worked with Rasa's Dual Intent and Entity Transformer (DIET) architecture, which is a powerful and highly customizable Transformer-based architecture that accepts external pre-trained embeddings and allows additional fine-tuning on the training data using a masked language modeling objective. In this work, we trained different models with different settings and training data sizes.

To test our models, we collected an external customer support dataset of 715 examples. We found that training intent detection models with Bitext data leads to better performance. In the following sections, we give more details on the dataset used, the models, and our results.

Daset

The training dataset consists of Bitext data for intent detection, which has been proven to increase chatbots' language understanding capabilities. We used our customer support intent detection dataset that we make publicly available on our GitHub repository[1]. The dataset has 27 intents including: place order, change order, set up shipping address, contact human agent, and create account, among others.

Using the training data, we created 5 different sub-datasets where we sampled 5, 10, 50, 100, and 300 utterances per intent. The purpose of these sub-datasets is to test the effect of the training data size on models' performance.

To make our tests as realistic as possible and to avoid having biased results, we collected an external customer support dataset of 715 examples and evaluated our models on this dataset.

Models

To leverage different pre-trained LLMs for intent detection using Bitext data and to benchmark the resulting performance using different data sizes, we worked with Rasa's DIET architecture. DIET is a lightweight Transformer-based architecture that is highly configurable. It allows plugging in several pre-trained embeddings like BERT and fastText and is fast to train. It is also proven to achieve state-of-the-art results. We have developed the following models:

1. Basic-embed: This is the basic DIET model that doesn't use external embeddings.
2. Spacy-embed: We used Spacy's fastText-based English embeddings. We chose the en_core_web_md model which provides a good balance between performance and accuracy.

[1] https://github.com/bitext/customer-support-intent-detection-training-dataset-rasa

3. TOD-BERT-embed: Task-oriented dialogue BERT (TOD-BERT) is a BERT-based model that was pre-trained on dialogue data. This model was proven to achieve better performance on intent detection than the original BERT model.

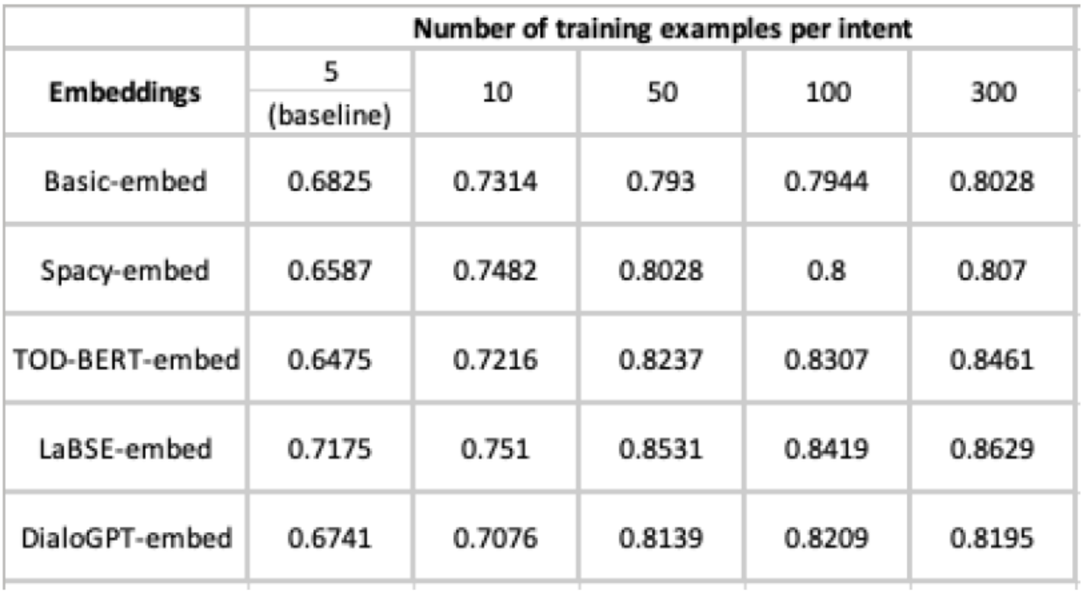
4. LaBSE-embed: LaBSE is a multilingual BERT-based model that was pre-trained on 109 languages. It achieves state-of-the-art results on intent detection.

5. DialoGPT-embed: DialoGPT is a GPT-based model that was developed by Microsoft. It was pre-trained on dialog data from Reddit.

For each model, we further fine-tune it on the training data using the masked language modeling objective. This can be easily implemented by configuring the DIET architecture.

Results and discussion

We report the accuracy of each model on the external test dataset. The accuracy is the percentage of examples that each model was able to successfully classify into their correct intents with high confidence. We list the results in the following table:



	Number of training examples per intent				
Embeddings	5 (baseline)	10	50	100	300
Basic-embed	0.6825	0.7314	0.793	0.7944	0.8028
Spacy-embed	0.6587	0.7482	0.8028	0.8	0.807
TOD-BERT-embed	0.6475	0.7216	0.8237	0.8307	0.8461
LaBSE-embed	0.7175	0.751	0.8531	0.8419	0.8629
DialoGPT-embed	0.6741	0.7076	0.8139	0.8209	0.8195

By looking at the results table, our first insight is that different embeddings can lead to different accuracy scores. This suggests that practitioners should take into consideration that evaluation results depend on each specific use case and that experimenting with different embeddings is an important first step before reaching an optimal choice.

Our baseline models were trained using 5 utterances per intent. To benchmark models' performance depending on the size of training data, we trained each model with 10, 50, 100, and 300 utterances per intent and we noticed the following:

- Training with 10 utterance per intent increases the accuracy between ~3.5 to ~9% depending on the embedding choice.
- Training with 50 utterance per intent increases the accuracy between ~11 to ~17.5%.
- Training with 100 utterance per intent increases the accuracy between ~11.5 to ~18.5%.
- Training with 300 utterance per intent increases the accuracy between ~12 to ~20%.

We notice that both Basic-embed and Spacy-embed models achieve similar performance, where Spacy-embed is slightly better than Spacy-embed. Also, the performance of these two models are considerably lower than the other three LLMs-based models.

The performance of DialoGPT-embed is lower than both LaBSE-embed and TOD-BERT-embed.

This could be due to the fact that DialoGPT-embed is a GPT-based model that was pre-trained using the next token prediction objective, as opposed to the other two models which are BERT-based and were pre-trained using the masked language modeling objective which is more suitable for intent detection tasks.

The LaBSE-based model achieves the best performance reaching ~86% when training with 300 utterances per intent.

Conclusion

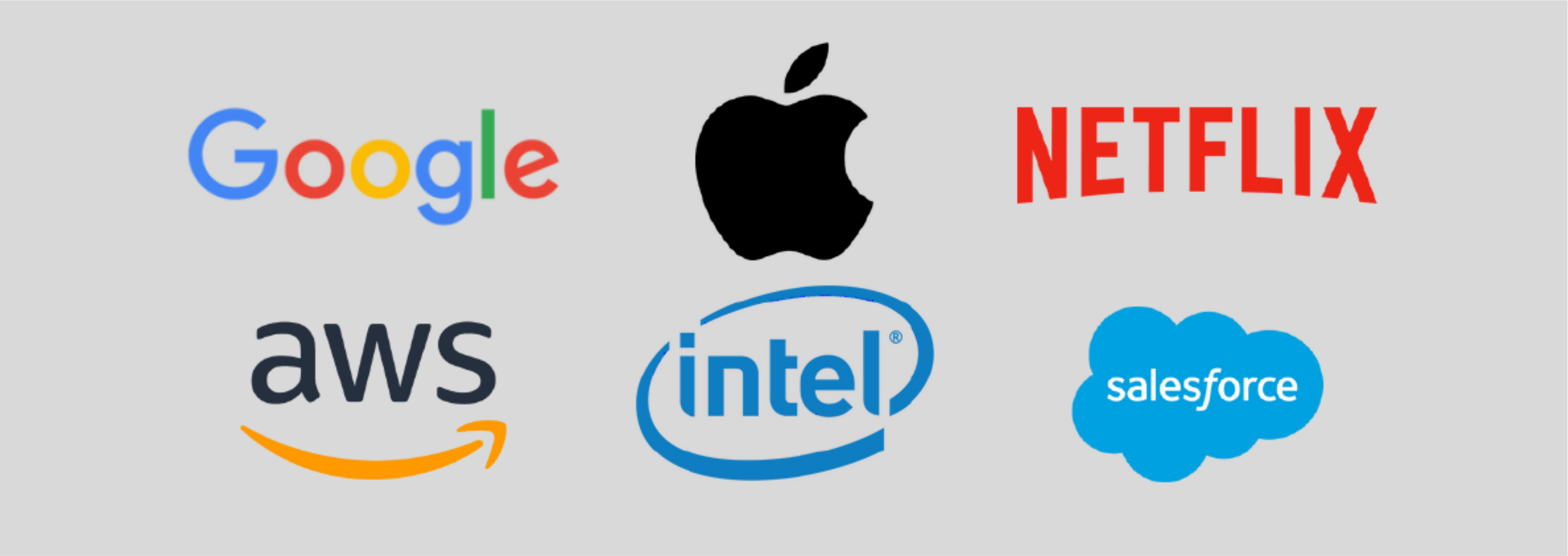
In this whitepaper, we benchmarked Rasa's performance for intent detection using different embeddings including LLMs-based ones.

To achieve this, we utilized the Rasa's DIET architecture. We trained different models using 5, 10, 50, 100, and 300 utterances per intent to test the effect of the size of training data on performance. When training, we further fine-tune the embeddings on the training data using the masked language modeling objective.

Our results suggest that leveraging LLMs-based embeddings and fine-tuned it on Bitext intent detection leads to considerable improvement in performance.

Finally, taking into consideration that LaBSE is a multilingual model, training it on datasets in languages other than English could be an interesting future direction.

Select Customers


Google
NETFLIX
aws
intel
salesforce