uOttawa

CSI 5180: AI Virtual Assistants

Project Definition

Group 6

Daniel Lobo (300319498)

Romario Vaz (300308477)

12 March 2023

"No Alexa, I don't want to book a flight; I want the book 'The Flight': Improving Intent Detection for Virtual Assistants"

Short description of goals and limitations

The project will tackle the problem of intent detection. The dataset we are using is the CLINC150 dataset [1] which consists of 23,700 queries, of which 22,500 queries are associated with one of 150 intents and the rest are out-of-scope intents. The main limitation of the approach used by the authors is that it does not accurately identify out-of-scope queries, and their best approaches only achieve a recall score of about 0.5 on out-of-scope examples. We aim to improve upon the overall accuracy metrics through a combination of data augmentation, clustering, and additional models.

At the end of the project, our goal is to have an intent detection system that obtains a higher overall accuracy and a higher recall on out-of-scope detection, compared to the approaches presented in the paper. Our prior knowledge of intent detection is limited to the material covered in class, but the team members have previous knowledge in using deep learning models like BERT and other NLP and machine learning techniques. We plan to learn how to use our prior experience to tackle a task that is highly relevant in the context of virtual assistants. We will be performing a brief literature review before embarking on the task in order to ensure that we are familiar with other techniques used in this domain.

For data augmentation, we plan on using ChaGPT to synthesize relevant data to augment the current dataset, which only has about 150 examples per intent. Due to the automated nature of this approach, we will need to manually validate some examples to ensure that the outputted data is relevant to the task and of the correct format. The test sets will not be modified, to ensure that we can properly evaluate the effect of the data augmentation.

We plan on using clustering algorithms like K-means/hierarchical clustering to tackle the out-of-scope problem. The intuition behind this is that the out-of-scope examples should form a large cluster while examples belonging to other intents will form smaller clusters. Due to the unsupervised nature of the algorithms, we will need to investigate the clusters to analyze how well the clustering has partitioned the data.

For the additional models, we plan to use a more advanced model than the ones presented in the paper. Tentatively, this would be RoBERTa or GPT-2, or some other large transformer-based model. A brief error analysis of the models will be done after their implementation.

The final deliverable will be a comparative study of the different approaches, to see which approach worked well on which tasks. Using this information, if we have improved upon the baseline work, we will create a simple end-to-end pipeline that can be used for intent detection. This can be directly used as an intermediate building block by people trying to build a virtual assistant system.

In order to ensure that the project can be completed in 25*2=50 hours, we are using automation (ChatGPT) to generate data for the data augmentation step, because manual generation of data would be too time-consuming. Further, rather than developing our own metrics, we are using the same metrics that were proposed in the dataset's paper to ensure that we have comparable results and to save time. There are probably more nuanced metrics that weigh different intents differently since not all of them are equally important, but given the time constraints, we have decided to stick with what was proposed in the paper.

Description and justification of the software platform/programming involved/dataset involved

Dataset

The CLINC150 dataset is used to assess the performance of intent classification systems under challenging conditions. In particular, it is designed to evaluate the ability of these systems to handle out-of-scope (OOS) queries, which are those that do not fit into the predefined intent classes. The dataset consists of 150 intent classes, and it is in English. The dataset was created using crowdsourcing, with crowd workers asked to paraphrase seed phrases or respond to different scenarios to generate the data (eg. “What would you ask to improve your credit score?”). Additionally, the dataset only contains single-intent samples.

The dataset is available in JSON format with six partitions, including validation, training, and testing data, as well as separate partitions for OOS validation, OOS training, and OOS testing. In total, there are 23,700 queries, with 22,500 of these being in-scope queries spread over the 150 intent classes. The remaining 1,200 queries are out-of-scope queries, which are essential for evaluating the performance of systems when handling OOS conditions.

Programming Involved

For this project, we will be using Python as the programming language to implement the intent classification system. We plan to use the scikit-learn library to implement the clustering approach and the HuggingFace Transformers library to implement state-of-the-art transformer-based models such as BERT, RoBERTa, and others.

Planned Activity Table

Activity	**Why?**	**Time Planned**	**Deliverable**
Read Paper for the Dataset	Gather knowledge about the dataset	1h	Summary of the article with important ideas highlighted
Literature Review	Gather information about existing work	5h	Review of literature on the dataset and related approaches
Explore Dataset	Understand the	2h	Documentation of dataset

	structure of the data		characteristics and properties
Preprocessing data	Prepare data for training and evaluation	3h	Cleaned and transformed dataset ready for training
Develop Baseline from paper	Establish a benchmark for comparison	4h	Implementation of paper's approach on the dataset
Generate Data for Augmentation	Increase dataset size	3h	Augmented dataset for better performance
Validate data and check for errors	Ensure data quality	4h	Cleaned and validated dataset free of errors
Build updated dataset and data splits	Prepare data for training and evaluation	2h	Updated dataset ready for model training and evaluation
Re-implement baseline on new data	Evaluate baseline on new dataset	1h	Performance of baseline approach on updated dataset
Develop a new approach - A (maybe clustering?) + Analysis	Propose and evaluate a new approach	6h	Implementation of clustering approach and analysis
Develop a new approach - B (maybe RoBERTa?) + Analysis	Propose and evaluate a new approach	3h	Implementation of RoBERTa-based approach and analysis
Comparative analysis	Compare performance of approaches	3h	Comparison of performance of different approaches
Prepare Demo	Showcase the project	3h	Demo of the implemented approach
Prepare Report	Document the project	10h	Comprehensive report detailing the project and its results
		Total = 50h	

References

[1] S. Larson *et al.*, ‘An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction’, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.