

**Industry Project Proposal**

| Name of the Organization | Cisco (CCAI) |
| --- | --- |
| Project location (city or timezone) | San Jose/ Milpitas, California |

**A. Project Description**

**Problem definition**

[50-100 word description of the problem which the candidates need to solve]

| The Cisco Agent Assist solution increases operational efficiency in a contact center by providing human agents the right information at the right time. During a call, the Agent Assist solution “listens” to the ongoing conversation and automatically surfaces resources from a company-specific knowledge base (KB) which are relevant to the customer’s query or issue. This feature’s entry point consists of a filter model that removes irrelevant parts of a conversation such as greetings or pleasantries and only sends queries closely related to the company domain for downstream processing.  The goal of this project is to build such a filter, a lightweight model capable of identifying which parts of a conversation are relevant and informative in the context of a contact center interaction. However, this project is not simply to develop a binary classifier. It will involve the acquisition and even generation of knowledgebases and datasets for training and evaluation.  This project can be divided into the following parts:  1. Data acquisition and generation  2. Model development  Data Generation and Knowledgebase Acquisition  A knowledge base is a data repository that persists organization-specific knowledge and documentation. Many companies have public customer service pages with articles, FAQs, and other documentation which can be scraped using a web crawler or off-the-shelf crawler tool like ParseHub to form example knowledgebases. The project team can select which domains and companies to target for populating these sample knowledgebases.  As a next step, it is required to generate synthetic examples of relevant queries for each knowledgebase. These queries can be generated using a natural language generation model. |
| --- |



| (e.g. T5 transformer-based question generation model). This data will be used for model training and evaluation.  The irrelevant queries, i.e. the queries not belonging to a knowledgebase can be sourced from here:  1. https://github.com/gunthercox/chatterbot  corpus/blob/master/chatterbot\_corpus/data/english/greetings.yml  2. https://github.com/zeloru/small-english-smalltalk-corpus  An example of irrelevant queries are greetings or pleasantries in a conversational setting.  Filter Model Development  For the second part of this project, the goal is to develop a model which is able to discriminate between relevant and irrelevant utterances in a conversation, given the domain context. The team can use either unsupervised or supervised algorithms to build the filter model. As this is a binary classification task, we are interested in the metrics of precision, recall, and F score. |
| --- |

**Key Research Questions/ Technological constraints that the Project will Answer**

| Tech Constraints:   * Target model latency <50ms   Key Challenges – Dealing With:   * Heterogeneous data sources * Noisy labels * Class imbalance (only 10% of conversations are relevant)   Key Questions   * Build a model that generalizes across industry domains or a set of models customized for specific domains? * How can we scale the process of synthetic data generation and validation of the generated data? * How can we understand and balance the bias-variance tradeoff for this model / use case? * How can we understand and balance the tradeoff between precision and recall for this model / use case and justify why one or the other should be emphasized for any model fine-tuning? |
| --- |



**Final deliverables at the end of the project**

[Please list the desired technical deliverables from the project team in as much detail as possible]

| - A python script for Data Generation  - A python script for any data labeling techniques used, if applicable. - A python script for model training and inference  - A python script for model evaluation  - Synthetic data generated in .csv or.json format |
| --- |

**Key activities/ technologies the project team may be expected to undertake/ work with**

[E.g. What kind of technology stack they will work with, the datasets they may need to work on, what kind of analysis they may be expected to undertake, etc.]

Technology stack:

* Python
* Pandas
* Keras or TensorFlow or PyTorch
* Scikit-learn
* Hugging Face Transformers

Web scraping and data generation:

* ParseHub
* BeautifulSoup

Machine Learning:

* Text classification
* Natural language generation

**Expected learning outcomes**

[What do you expect the candidates to learn from the project. Please mention the technical skills they will imbibe over the project.]

| 1. Synthetic Data Generation process  2. Model training and optimization  3. Natural language processing techniques |
| --- |

| Desired Team Size (if any): | 1- 3 |
| --- | --- |