MULTI-LANGUAGE INTENT PREDICTION

A PREPRINT

Yuehao Wu
yuehaowu@stanford.edu

Junhua Zhao
junhua@stanford.edu

Rongbin Li
rongbli@stanford.edu

December 9, 2019

ABSTRACT

Nowadays Machine Learning has been widely used in customer service area. It significantly reduced total cost of ownership via reducing human interactions. In this scenario the first and the most useful step is to predict the intent of a customer contact. The problem we are trying to solve in this project is to detect intent from an utterance in different language, including English, French, German, Spain, and Chinese. The input of our model in training phase is in one language - English, and that in inferencing phase is sentences from other 4 languages. The output is a ranking of a list of predefined intents, such as greeting, negative feedback, taking to agent or asking for refund status, etc.

***K*eywords** NLP · Transfer Learning · BERT · Multi-Language

1 Introduction

Nowadays AI powered chatbot is widely used in the area of customer service. It has significantly improved customer service agent's efficiency and reduced human contact as well as total cost. Usually the first and the most useful step is to predict the intent of customer contact. The accuracy of this prediction has direct impact to customer satisfaction as the time spent on resolving the case. Now that we have developed a couple of models that can handle single language such as English, French, and Chinese pretty well, however, as a global company it can have business in 200+ countries and regions, and the languages spoken in chat are very diverse. One solution is to have different models for different language, however it depends on the extra step of language detection, and it does not scale from model development perspective because it requires training data labeling and preparation in all languages.

Inspired by BERT multilingual model, we think having one model that understands multiple languages could be a good direction to explore for industry needs. This project aims to develop a scalable solution for multiple language intent prediction model training and inference that can be potentially applied in real customer service department in a global company.

This requires reasonable model training cost, model analytics performance, system performance and scalability in inferencing.

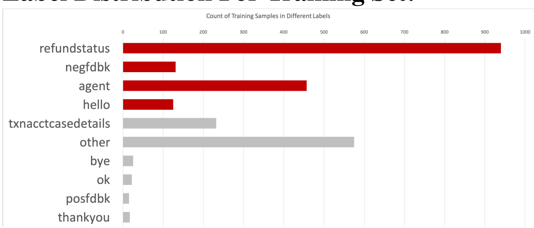
2 Datasets

2.1 Training data

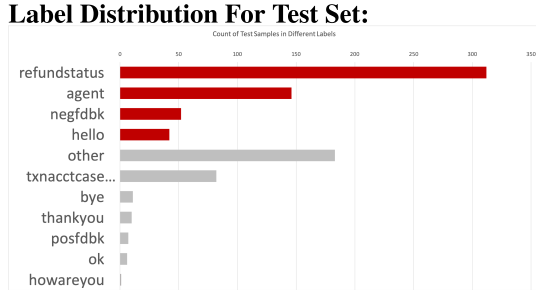
We are using customer support data collected from a real customer service channel with sensitive information removed. This data set has about 3.5k sentences and the corresponding intent labeled by human. The data is split into 2.7k training set and 800 test set. There are 11 different intents in this dataset.

2.2 Data Sample and Label Distribution

Label Distribution For Training Set:



Label Distribution For Test Set:



Training Data Sample:

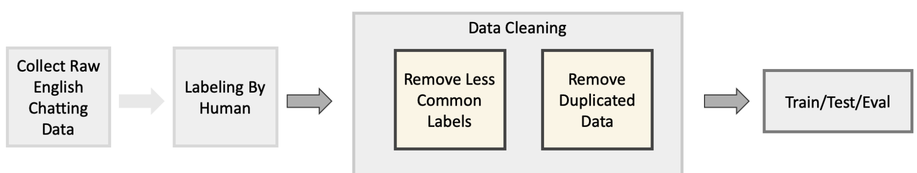
Sentence	Label
so do i my account statement said that i got the refund in my balance said the amount is visible	refund
you make it easy to contact	negfdbk
i have a current case open that i wanted to add comments to	other
Hello how are you	greeting

Testing Data Sample:

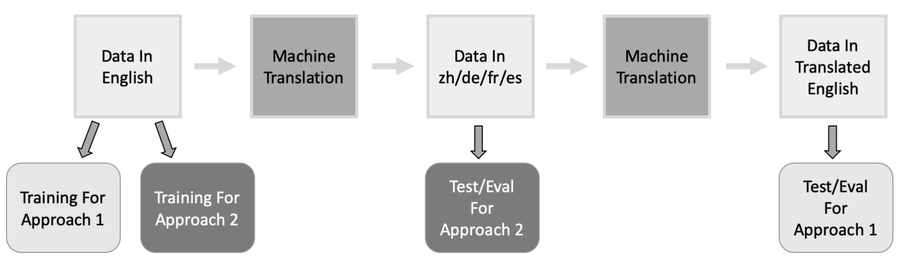
Language	Sentence	Label
english	i received an email from merchant about a refund but not showing in my accout	refund
franch	vous facilitez le contact	negfdbk
spanish	facilitas el contacto	negfdbk
germany	Ich habe eine E-Mail von über eine Rückerstattung erhalten, die aber nicht in meinem Konto angezeigt wird	refund

3 Data Processing

We did data pre-processing in following steps to make sure the training data is in a good shape



We leveraged google translation service to generate testset in other language because we only have human labeled data in English



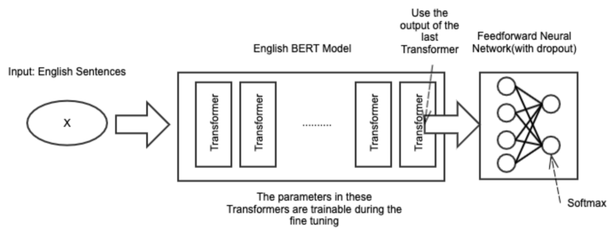
4 Methodology

We've tried two approaches on this multi-class intent prediction problem:

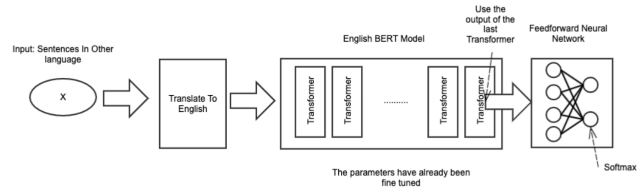
4.1 Approach 1: English Bert + Translation

We use BERT Base Model as embedding layer and added FC layers to do a fine tune. The model we created can only understand English input data. Later on in prediction phase, we use Google machine translation to translate the sentences of different languages into English and then use our model to predict the intent.

Training:



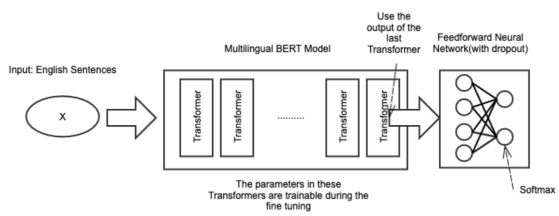
Inference:



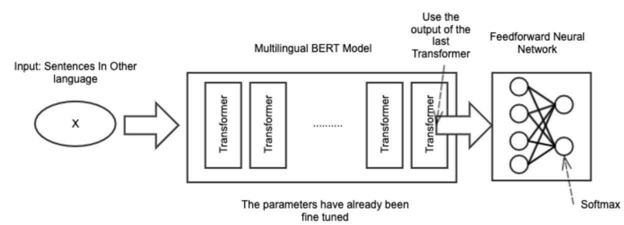
4.2 Approach 2: Multi-lang Bert + Fine Tuning

We use BERT Multilingual model which was trained on multi-language corpus. In training phase we apply English training data to create a model that can understand our problem in multiple language. Finally we pass in test data in other language and let the model predict the intent directly. This is a zero shot learning.

Training:



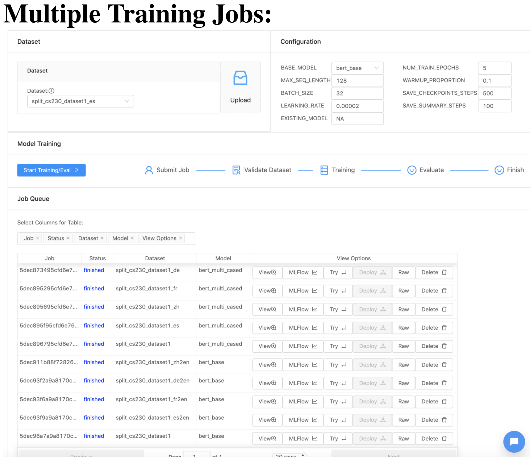
Inference:



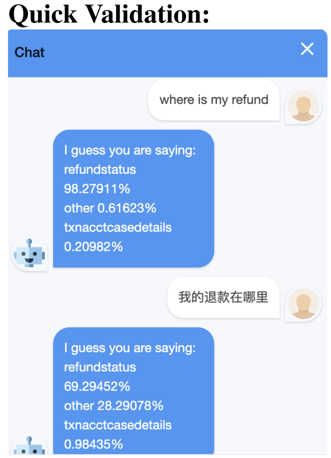
5 Rapid Model Creation

To speed up our experiment, we have built a UI based tool named NLP Model Factory which allows the users to choose base model (BERT in our example) for transfer learning / fine tune, specify hyper parameters, submit the job to GPU boxes, evaluate the result, generate reports and track jobs.

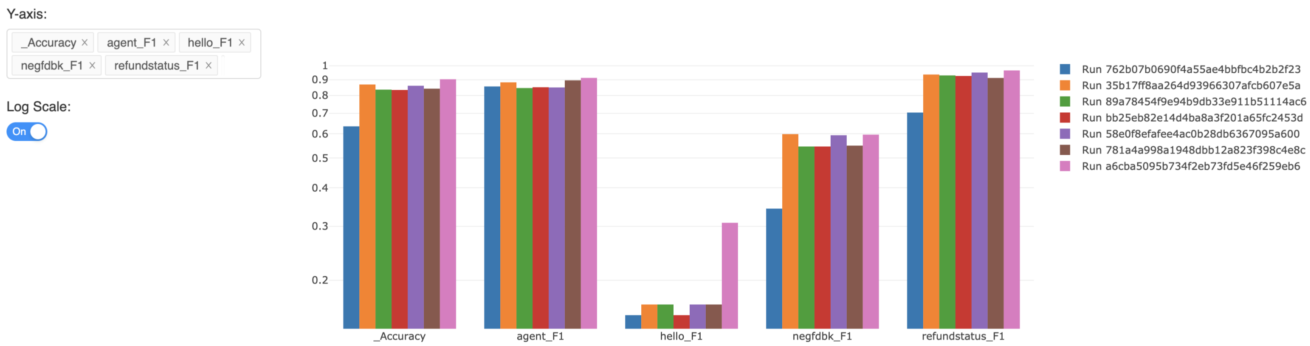
Multiple Training Jobs:



Quick Validation:

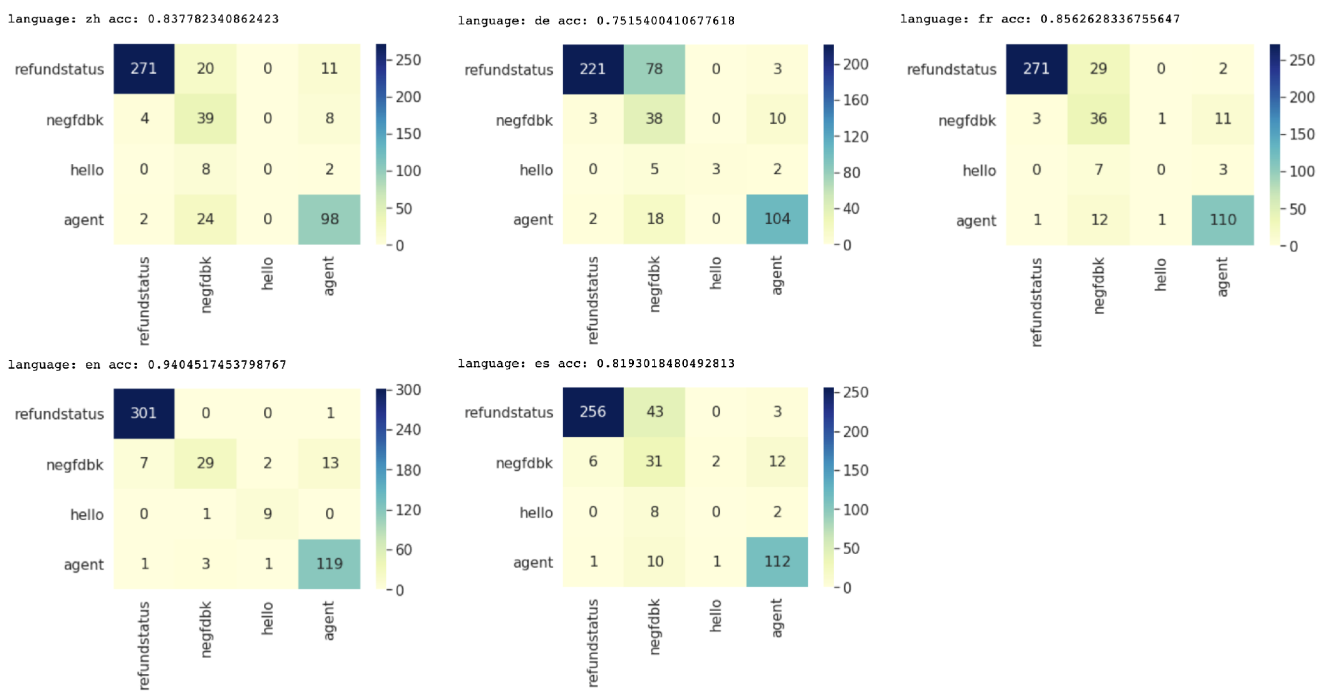


Tracking and Compare Using MLFLow:



6 Model Performance

We evaluate the model performance by primarily looking at overall accuracy, and the confusion matrix.



From the confusion matrix we generated a report for precision and recall of each intent.

Lang&Approach	Acc	Refund		Agent		Hello		Negfdbk	
		Pr	Re	Pr	Re	Pr	Re	Pr	Re
EN1	94.7	96.8	100.0	93.5	93.5	87.5	70.0	83.7	70.6
EN2	92.8	96.5	99.0	89.3	94.3	75.0	60.0	78.9	58.6
ZH1	94.0	97.7	98.7	92.2	95.2	85.7	60.0	76.7	70.6
ZH2	80.3	96.4	96.7	89.8	92.7	66.7	20.0	58.5	60.8
ES1	93.8	97.4	99.0	92.0	92.7	85.7	60.0	77.1	72.5
ES2	77.0	96.6	75.2	89.3	88.0	66.7	40.0	28.2	68.6
FR1	94.3	98.0	99.0	91.5	95.2	85.7	60.0	78.3	70.6
FR2	90.3	96.4	96.7	89.8	92.7	66.7	20.0	58.5	60.8
DE1	95.1	98.0	99.7	92.9	94.3	87.5	70.0	82.6	74.5
DE2	80.1	92.7	88.1	93.4	68.5	100.0	30.0	33.96	70.58

7 Conclusions

The result shows currently the Approach 1 has achieved better overall performance than Approach 2. We have done some analysis and here is our findings.

1. Single language model performs better in general than multi-language model for testset in a particular language that this model is training with.
2. We also noticed that the test data is from machine translation which is quite different from real human conversation in other language, however, it's friendly for machine translation engine to translate it back to English. This gives advantage in the evaluation for Approach 1 over Approach 2.
3. Multi-language model has shown solid result from zero-shot learning. This has significant advantage regarding to cost and time to market.

8 Future Works

As next steps, we have following proposals:

1. Validate the result with more data and more accurate data. one thing we can do is data augmentation, another is to collect real chat messages in other languages and use them as test data.
2. Architecture search. We can extract more features from BERT embedding layers, and try more architecture for FC part.
3. Pre-training on BERT model. there is some data that currently pre-trained model provided by google can not handle because they may not present in the corpus, such as some company or industry specific terminology and their translations.

8.1 Acknowledges

Special thanks to **Charles(Yancheng) Liu** who provided technical support of GPU based training environment.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[3] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert? *CoRR*, abs/1906.01502, 2019.