

Intermediate Progress Report: Natural Language Interaction Project

Cristian Bassotto - u210426

Nikolay Kormushev - u234126

Adrian Gheorghiu - u234621

Giuseppe Cianci - u234127

Task 1: Domain Identification/Dialog Act Prediction

Approaches Tried

- **XGBRFClassifier:** Extreme Gradient Boosted Random Forest Classifier
 - Word unigram frequencies + POS unigram frequencies features: Accuracy: 63.3%, Precision: 73.2%, Recall: 52.8%, F1: 59.4
 - TF-IDF (term frequency-inverse document frequency) features: Accuracy: 65.7%, Precision: 72.1%, Recall: 54.5%, F1: 60.7%
 - TF-IDF + POS unigram frequencies: Accuracy: 65.9%, Precision: 73.9%, Recall: 55.1%, F1: 61.6%
 - TF-IDF + Spacy embeddings: Accuracy: 64.4%, Precision: 69.5%, Recall: 53.4%, F1: 59.2%
- **SVM:** Support Vector Machine with rbf kernel using fasttext embeddings and lemmatization
 - TF-IDF + fastText embeddings + top-k feature selection based on ANOVA score: Accuracy: 76.9%, Precision: 69.4%, Recall: 64.1%, F1: 66.5%
- **MLP:** Multi Layer Perceptron with BatchNorm and Dropout (and unifying non-hotel or restaurant intents into 'other' class):
 - TF-IDF + fastText embeddings + top-k feature selection based on ANOVA score: Accuracy: 79.1%, Precision: 91.2%, Recall: 72.4%, F1: 79.5%
- **LSTM** Stacked bidirectional Long Short-Term Memory architecture with Effective Number of Samples class loss weighting to combat class imbalance (and unifying non-hotel or restaurant intents into 'other' class):
 - fastText embeddings: Accuracy: 77.7%, Precision: 91.9%, Recall: 70.4%, F1: 77.7%
 - BERT embeddings: Accuracy: 77.7%, Precision: 90.2%, Recall: 69.2%, F1: 75.1%
- **Finetuning Pre-trained Transformer Models:** Utilized BERT for fine-tuning on our specific dataset (and unifying non-hotel or restaurant intents into 'other' class):
Accuracy: 86.6 %, Precision: 89.8 %, Recall: 89.2 %, F1: 89.5 %

Possible Improvements

- Make use of utterances and intent histories
- Experiment with LSTMs
- Experiment with other variants of BERT (RoBERTa, DistilBERT) for better contextual understanding.
- Integrate linguistically motivated rule-based post-processing to refine predictions.

Task 2: Content Extraction from User Utterances (Semantic Frame Slot Filling)

Approaches Tried

- **Finetuning Pre-trained Transformer Models:**

- BERT for token classification fine tuned for sequence to sequence BIO tagging, with punctuation removal preprocessing, splitting on spaces then running tokenizer on every split word: Accuracy: 98 %, Precision: 91 %, Recall: 95 %, F1-Score: 93 %
- RoBERTa for token classification fine tuned for sequence to sequence BIO tagging, with Effective Number of Samples class loss weighting to combat class imbalance and running tokenizer on whole sentence: Accuracy: 97%, Precision: 76.6%, Recall: 96.8%, F1: 84.2%

Possible Improvements

- Make use of utterances and intent histories
- Do mapping of tagged spans to actual slot values (e.g. 'two' becomes '2')
- Otherwise, the performances are already very good, the focus should be on the other tasks first.

Task 3: Agent Move Prediction

Not started yet. First idea to try is a rule-based approach of predicting the frame by matching the filled slots to the frames existent in the train dataset. After that, generalization can be tested by trying doing multi-label classification on the slots.