WATSON ASSISTANT

- Watson Assistant is part of IBM Watson’s suite of enterprise-ready AI business solutions.
- Allows business users to build their own domain-specific chatbot—a customizable “Siri”

PROBLEM STATEMENT

- Machine learning models form relationships between features and outputs in a purely pragmatic manner, based on what produces the best overall performance.
- The most powerful models currently available—including those used by Watson Assistant—do not offer interpretability, meaning the validity of the relationships a model forms cannot be confirmed.

In order to increase end-user confidence and long-term performance of the end-user’s Watson Assistant, we aimed to derive explainability from non-interpretable models to allow domain experts on the client side to assess and calibrate the validity of classifier relationships.

OBJECTIVE

- Users request information from Watson Assistant in what is known as the user’s utterance, and Watson Assistant classifies the objective of the user’s request, which is known as the user’s intent.

Utterance Example: ‘When do I need to submit my performance assessment?’

Corresponding Intent: ‘AssessmentDueDate’

DATASETS: GROUND TRUTH & LOG DATA

- User ‘Intent’ classifications
- Labeled ‘ground truth’ utterances, used for training multiclass model
- Unlabeled log data utterances consisting of real-life user inputs, used for bias-removal and testing performance

DATA OVERVIEW

- 32
- 2,900
- 14,300

BIAS REMOVAL

- Focus was on improving the precision of a specific intent using log data
- Objective: build a cascade with superior performance to baseline multiclass model in order to reduce bias

BIAS REMOVAL IMPLEMENTATION

- The bias removal models implemented involve down-weighting or filtering the labeled log training data based on identified biased features.
- The various bias removal models’ precision across the labeled intents in log data test set can be seen below:

<table>
<thead>
<tr>
<th>Model Type</th>
<th>GET-manager</th>
<th>AssessmentDueDate</th>
<th>Information Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Multiclass</td>
<td>0.68</td>
<td>0.45</td>
<td>64.60%</td>
</tr>
<tr>
<td>Single Feature Cascade</td>
<td>0.69</td>
<td>0.41</td>
<td>67.80%</td>
</tr>
<tr>
<td>Basic Filter Cascade</td>
<td>0.68</td>
<td>0.42</td>
<td>61.20%</td>
</tr>
<tr>
<td>Basic Filter Cascade w. Score</td>
<td>0.69</td>
<td>0.41</td>
<td>60.12%</td>
</tr>
<tr>
<td>Full Downweight</td>
<td>0.80</td>
<td>0.51</td>
<td>80.00%</td>
</tr>
<tr>
<td>Partial Downweight</td>
<td>0.72</td>
<td>0.50</td>
<td>70.12%</td>
</tr>
<tr>
<td>Baseline Unigram Cascade</td>
<td>0.72</td>
<td>0.52</td>
<td>78.00%</td>
</tr>
<tr>
<td>Baseline Bigram Cascade</td>
<td>0.72</td>
<td>0.52</td>
<td>78.00%</td>
</tr>
<tr>
<td>Best Epoch (%)</td>
<td>0.76</td>
<td>0.49</td>
<td>80.60%</td>
</tr>
<tr>
<td>Best Final Prediction</td>
<td>0.80</td>
<td>0.51</td>
<td>80.12%</td>
</tr>
</tbody>
</table>

- The influence of cascade training data size on final precision can be seen below on intent ‘GET-manager’

EXPLAINABILITY

- Able to identify the most significant features (words) in each classification as well as their influence on the final confidence using SHAP values
- Quickly noticed a trend of intent-related keywords comprising the majority of the confidence for a particular classification
- Produced aggregate plots of feature influence across subsets of data (true and false positives) to identify most influential features, and if the classifier relationship appeared to be valid
- Presence of keywords propels the confidence of a classification quite high, while the absence of the same term does not reduce confidence by nearly as much

APPROACHING EXPLAINABILITY

- Red indicates presence of word in utterance; blue indicates absence
- Positive values add confidence; negative reduces

FEATURE INFLUENCE FORCEPLOT

- CHI-SQUARE ANALYSIS

<table>
<thead>
<tr>
<th>Term</th>
<th>Chi-Sq Value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>date</td>
<td>27.81</td>
<td>1.0e-02</td>
</tr>
<tr>
<td>deadline</td>
<td>22.34</td>
<td>8.9e-02</td>
</tr>
<tr>
<td>AssessmentDueDate</td>
<td>27.21</td>
<td>1.0e-02</td>
</tr>
<tr>
<td>time</td>
<td>26.86</td>
<td>1.0e-02</td>
</tr>
<tr>
<td>deadline</td>
<td>25.83</td>
<td>1.0e-02</td>
</tr>
</tbody>
</table>

- Chi-Square feature selection produced similar results to SHAP forceplots
- Confirms statistical significance of highly correlated terms
- Also used to surface terms more prevalent in misclassifications

FEATURING ENGINEERING

- Unigram and bigram embeddings to preserve interpretability
- Limited preprocessing to preserve outliers

CONCLUSIONS

- All cascading classifiers produce a statistically significant performance lift over the multiclass model across intents
- Explainability is actionable by a domain expert to improve performance and reduce bias
- Improved performance occurs regardless of the size of the training set; cascade training data as low as 140 utterances can still offer significant lift
- Domain experts not required to invest significant time in labeling log data to see an improvement in precision, meaning the cascade solution is likely to be adopted by end-users
- Explainability manages to play an important role in certain cases where a problematic keyword needs to be down-weighted, eliminated, or used as a filter in the cascade model’s training
- Different cascade variants may be more effective given characteristics of the intent, such as the percentage of occurrences of problematic terms

MULTICLASS MODEL

- Selected as surrogate classifier due to speed and performance
- To attain optimal hyperparameters

10,662