Annotating Low-Confidence Questions Improves Classifier Performance

Stephanie Hernandez* and **Ron Artstein**
USC Institute of Creative Technologies
12015 Waterfront Drive, Playa Vista CA 90094-2536, USA
st3phy831@gmail.com artstein@ict.usc.edu

*Now at Hartnell College and CSU Monterey Bay

Abstract

This paper compares methods to select data for annotation in order to improve a classifier used in a question-answering dialogue system. With a classifier trained on 1,500 questions, adding 300 training questions on which the classifier is least confident results in consistently improved performance, whereas adding 300 arbitrarily selected training questions does not yield consistent improvement, and sometimes even degrades performance. The paper uses a new method for comparative evaluation of classifiers for dialogue, which scores each classifier based on the number of appropriate responses retrieved.

1 Introduction

Statistically trained dialogue systems can often be improved by adding annotated training data; when the system is deployed with real users, it often collects more interaction data than can be annotated, so prioritization of the data for annotation is required. This paper presents an experiment on selecting data for annotation using a dialogue system's internal confidence measure: it prioritizes annotation of those utterances for which the system is the least confident about how to react. This can be considered a form of active learning (Settles, 2010). Adding these utterances as training data (with appropriate annotations) improves the system's performance, whereas adding a comparable number of utterances that were arbitrarily selected does not improve performance to the same extent.

We use data from the Digital Survivor of Sexual Assault (Artstein et al., 2019), which is a system based on NPCEditor, a classifier trained on linked questions and answers (Leuski and Traum, 2011). For each new question, the classifier provides a score for every available answer, reflecting how well the classifier thinks it answers the question, and then returns the answers whose confidence exceeds a threshold (the list may be empty if all answers are below threshold). This experiment uses these scores to identify low-confidence questions to prioritize for annotation.

2 Method

2.1 Materials

For the baseline system we chose a very limited dataset, with 1,542 questions and only 1,517 links between questions and answers. Starting with an impoverished baseline allows room for measurable improvement with the addition of a small number of questions and links, whereas on a better trained baseline, the impact of additional training data is expected to be smaller. Also, the small baseline system left us with many questions that were already annotated and available for the experiment. The additional training data were taken from four datasets of questions annotated with links to appropriate answers (the four datasets were labeled "Alpha", "Beta", "Beta2" and "Windows", reflecting the development stage at which the questions were collected; see Artstein et al. 2019). All systems were tested using a fixed test set of 399 questions linked to appropriate answers.

2.2 Procedure

From each of the four annotated datasets, three sets of 300 questions (with corresponding links) were extracted: The "Duplicates" set simply selected 300 arbitrary questions, possibly including duplicate questions (that is, instances where the same exact question was asked by different users, though possibly annotated with different links). The "No-Duplicates" set also selected 300 arbitrary questions, ensuring that the 300 questions are all distinct from one another. Selection of questions for both the "Duplicates" and "No-Duplicates" sets was done through custom python scripts. A

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

third set of 300 questions was extracted by giving a full dataset of questions to the baseline classifier; the 300 questions for which the classifier returned the lowest confidence were chosen as the "Active Learning" dataset. Overall, we extracted 12 sets of 300 questions annotated with links (three from each of the four datasets).

Each of the 12 sets of annotated questions was added (separately) to the baseline classifier, and the resulting classifier was retrained; this resulted in a total of 13 classifiers (including the baseline). Each of the 13 classifiers was then run on the test set, returning an output of ranked responses for each question. A custom python script was then used to tabulate the outputs, pairing each question with the top-ranked three responses from each classifier (or fewer responses, if the classifier returned fewer than 3).

2.3 Evaluation

Traditional measures such as precision and recall are not well-suited for comparing the performance of ranked lists, because of the way responses are used in a dialogue system: the most common action of the system is to choose the top-ranked response, less commonly it chooses the second, then the third and so on. For this experiment, we chose to compare classifiers by the number of appropriate responses retrieved: for each of the test questions, a point was given to the classifier (or classifiers) with the highest number of appropriate responses within the top three. The total score of a classifier therefore reflects the number of questions for which it retrieved the highest number of appropriate responses, compared to the other classifiers.

3 Results

For 54 of the test questions, all classifiers gave the same output; these questions are excluded from further analysis. Figure 1 shows the classifier scores on the remaining 345 questions. These results show that the Active Learning classifiers score consistently above the baseline ($t(3)$=8.8, p=0.003); they also score signifantly higher than the Duplicates classifiers (Mann-Whitney U=16, p=0.03), and the difference from the No Duplicates classifiers approaches significance (U=15, p=0.06). No other differences were significant. Interestingly enough, many of the Duplicates and No Duplicates classifiers scored below the baseline, though these differences were not statistically significant.

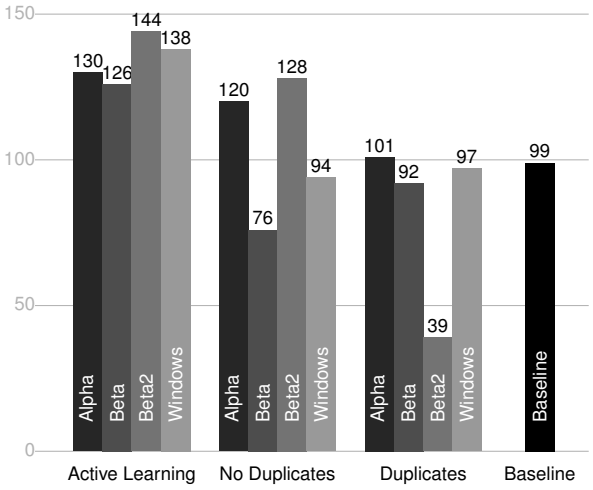

Figure 1: Classifier performance (points)

4 Discussion

The experiment shows that adding a small amount of low-confidence questions as training data can consistently improve the performance of a classifier, whereas adding the same amount of arbitrary questions does not lead to consistent improvement; this suggests that the classifier's confidence is a useful measure for prioritizing annotation. One limitation of this experiment is the impoverished baseline classifier, which reflects the very earliest stages of dialogue system development; at this stage, systems are usually not widely deployed and development budgets are still relatively large, so it is common to annotate all the available data anyway. It remains to be seen whether this method is useful at more mature stages of development, when the amount of available data exceeds the capacity for annotation. Another interesting observation is that some cases of added training data resulted in lower performance: this suggests that perhaps annotating all the available data is not the best approach, and that careful curation of data to be annotated needs to be explored.

Acknowledgments

The first author was supported by NSF award 1852583 "REU Site: Research in Interactive Virtual Experiences" (PI: Ron Artstein). The second author was sponsored by the U.S. Army Research Laboratory (ARL) under contract number W911NF-14-D-0005. Statements and opinions expressed and content included do not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

Ron Artstein, Carla Gordon, Usman Sohail, Chirag Merchant, Andrew Jones, Julia Campbell, Matthew Trimmer, Jeffrey Bevington, COL Christopher Engen, and David Traum. 2019. Digital survivor of sexual assault. In *IUI '19: Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 417–425, Marina del Rey, California. ACM.

Anton Leuski and David Traum. 2011. NPCEditor: Creating virtual human dialogue using information retrieval techniques. *AI Magazine*, 32(2):42–56.

Burr Settles. 2010. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.