Introducing a new 150-intent dataset for evaluating *out-of-scope* prediction performance of intent classification systems

An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction

Stefan Larson ◦ Anish Mahendran ◦ Joseph J. Peper ◦ Christopher Clarke
Andrew Lee ◦ Parker Hill ◦ Jonathan K. Kummerfeld ◦ Kevin Leach
Michael A. Laurenzano ◦ Lingjia Tang ◦ Jason Mars

stefan@clinc.com ◦ **www.clinc.com** ◦ **github.com/clinc/oos-eval**

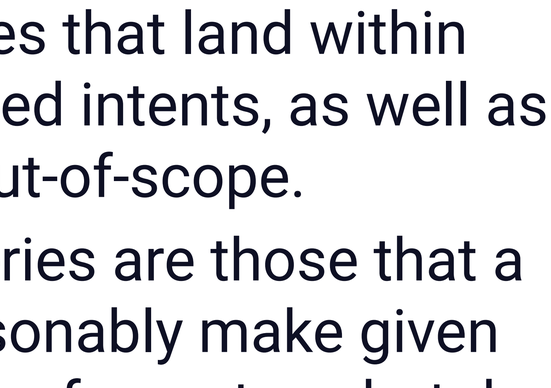
clinc

Task-driven dialog systems need to handle user queries that land within the set of supported intents, as well as queries that fall out-of-scope.

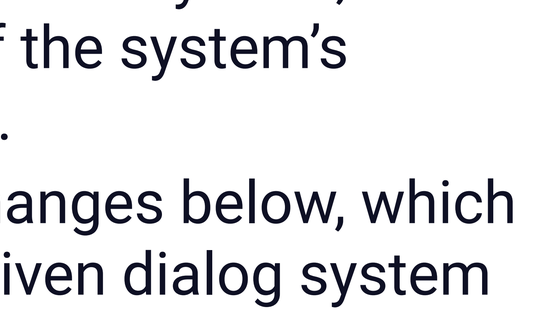
Out-of-scope queries are those that a user may (un)reasonably make given limited knowledge of a system, but do not fall into any of the system's supported intents.

Consider the exchanges below, which are from a task-driven dialog system for banking:

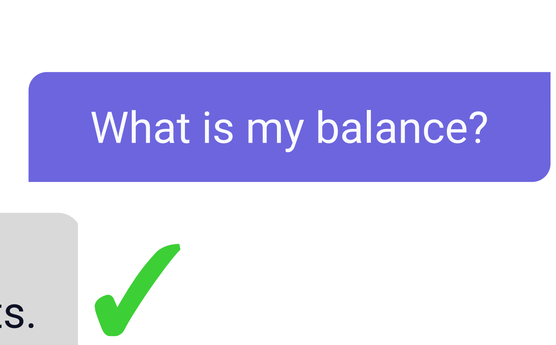
① correct in-scope classification



② out-of-scope mis-classified as in-scope

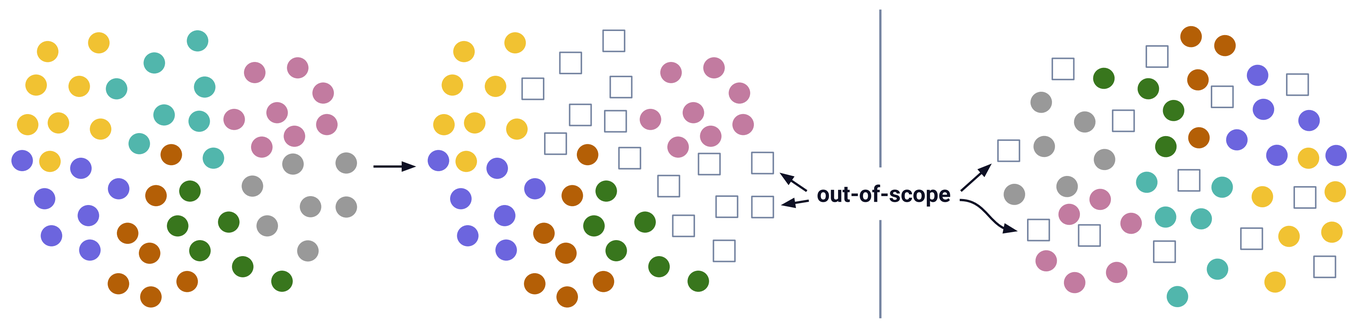


③ out-of-scope identified correctly



Our new dataset targets the task of intent classification in the presence of realistic, natural out-of-scope queries.

Other datasets and analyses either do not consider out-of-scope queries, or use held-out intents for out-of-scope.



Prior work creates out-of-scope data by holding out in-scope data, creating a single coherent out-of-scope intent.

Our work bootstraps a new dataset where the out-of-scope class is distributed, and not clearly defined.

Dataset Examples

Domain	Intent	Example
Banking	Balance	*how much do i have in my savings account*
Home	Todo List	*is vacuuming on my list of things to do*
Travel	Book Flight	*book a flight to los angeles from las vegas on american airlines*
Utility	Text	*i need leslie to be texted saying have a good day*
Work	PTO Used	*i want to know how many vacation days i have used*
Small Talk	Tell Joke	*can you tell me a joke about politicians*
Out-of-Scope	Out-of-Scope	*how many prime numbers are there between 0 and 100*
Out-of-Scope	Out-of-Scope	*how many students attend ucsb*
Out-of-Scope	Out-of-Scope	*change my seat to a window seat for my flight on monday*
Out-of-Scope	Out-of-Scope	*set a warning for when my bank account starts running low*

Benchmark

Classifier	In-Scope	OOS Recall
FastText	89.0	9.7
SVM	91.0	14.5
CNN	91.2	18.9
Rasa	91.5	45.3
DialogFlow	91.7	14.0
MLP + USE	93.5	**47.4**
BERT	**96.9**	40.3

github.com/clinc/oos-eval