Not So Fast, Classifier: Accuracy and Entropy Reduction in Incremental Intent Classification

Lianna Hrycyk, Alessandra Zarcone & Luzian Hahn

inCLINC available at: http://dx.doi.org/10.24406/fordatis/140

Incremental Intent Classification (IC)

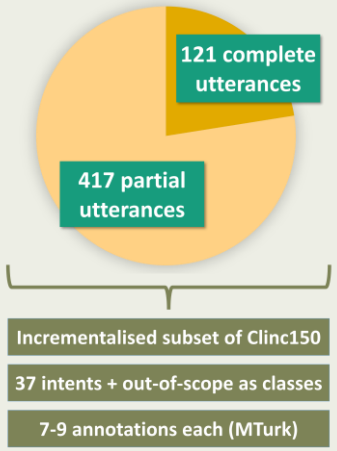
- **What's the advantage for a spoken dialogue system?** Reduce response latency, time backchanneling, optimize turn-taking...
- **What's the challenge?** Partial utterances do not necessarily contain enough information to be mapped to the intent class of their complete utterance

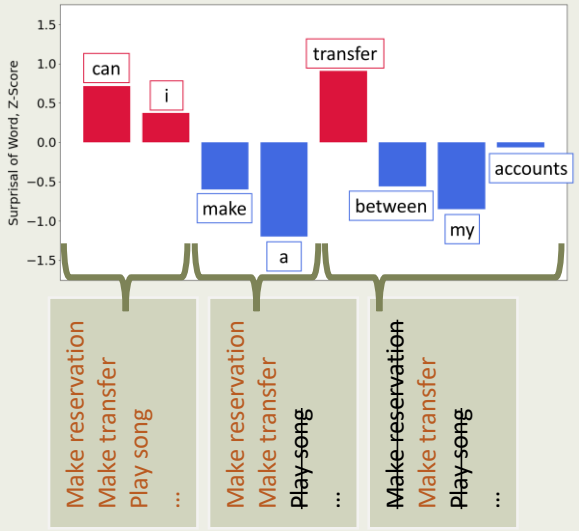
Q1
How to adapt intent labels to incremental IC?

Q2
What is the human upper bound for incremental IC? How does a SOTA classifier compare?

Q3
What is the relationship between ER and the point where the final intent can be identified?

inCLINC Dataset





Surprisal: conditional probability of a linguistic unit (word) given the context (inversely proportional to a word's information content)

Results

Q1

	Accuracy, IC complete utterances	Accuracy, incremental IC	Edit Overhead	Word Chunk Savings
Annotators	N/A	66.43%	0.39	2.43
DistilBERT*	94.56%**	56.35%	0.45	1.94

* Pretrained DistilBERT + classification head (1 FC layer, [CLS] as input), fine-tuned 2 epochs on Clinc150 complete utterances
** On held-out complete utterances from Clinc150 dataset

Q2

Classifier Overfitting

"i have to..."
"on the..."
"tell my..."

Classifier Under-Fitting

"i need milk..."
(update shopping list vs. ~~place an order~~ vs. ...)
"get reservations..."
(make reservation vs. ~~accept a reservation~~ vs. ...)

Q3

Entropy Reduction (ER): how much processing a given word decreases the amount of uncertainty about (the overall intent of) an ongoing utterance

	↑ Accuracy	↓ Accuracy
ER < 0	85	143
ER ≥ 0	10	179

Utterances following a **reduction in entropy (ER <0)** are **more frequently* associated with ↑ accuracy** (of annotators) compared to those with ER ≥ 0
*p <0.001, one-sided

Conclusions & Future Work

- Assigning ground-truth labels to incomplete utterances for incremental IC over-simplifies the task
- Correct early predictions for only DistilBERT signal overfitting (due to artefacts in the data)
- Correct early predictions for only humans signal areas of improvement for the classifier
- **Open challenge: predicting a set of plausible labels and identifying the point where interpretations converge (ER shows potential as a useful metric for this task)**