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

Lianna Hrycyk
LumenVox GmbH
Lianna.Hrycyk@LumenVox.com

Alessandra Zarcone
HumAIn, Fraunhofer IIS
zce@iis.fraunhofer.de

Luzian Hahn
Semantic Audio Processing, Fraunhofer IIS
hal@iis.fraunhofer.de

Abstract

Incremental intent classification requires the assignment of intent labels to partial utterances. However, partial utterances do not necessarily contain enough information to be mapped to the intent class of their complete utterance (correctly and with a certain degree of confidence). Using the final interpretation as the ground truth to measure a classifier's accuracy during intent classification of partial utterances is thus problematic. We release inCLINC, a dataset of partial and full utterances with human annotations of plausible intent labels for different portions of each utterance, as an upper (human) baseline for incremental intent classification. We analyse the incremental annotations and propose entropy reduction as a measure of human annotators' convergence on an interpretation (i.e. intent label). We argue that, when the annotators do not converge to one or a few possible interpretations and yet the classifier already identifies the final intent class early on, it is a sign of overfitting that can be ascribed to artefacts in the dataset.

1 Introduction

In non-incremental spoken dialogue systems (SDS), modules process complete utterances sequentially: the Automatic Speech Recognition (ASR) module must detect an end of turn before the transcribed speech can be processed by the Natural Language Understanding (NLU) module, in which utterances are often assigned an intent label. The sequential execution of such systems not only increases response latency, but also affects the perceived naturalness of the interaction. Natural conversations typically proceed incrementally: people rely on multiple cues to build partial interpretations of incomplete input, check if the communication is successful, and adapt their production accordingly (Clark, 1996), sometimes completing the interlocutor's turn, barging in, or responding before the turn is over (Jaffe and Feldstein, 1970; Brady, 1968).

Incremental dialogue systems, on the other hand, make use of incremental processors to enable more efficient, flexible, and effective interactions with users (Schlangen and Skantze, 2011). For example, the ASR module of an incremental, task-oriented SDS continuously processes incoming speech signals and posts hypothesized transcriptions. Then, the downstream NLU module analyses the increasingly longer portions of the final utterance, before the user finishes their turn. These early hypotheses produced by an incremental NLU (iNLU) module can in turn be consumed by downstream modules as soon as they are posted. Thanks to incremental processing, an SDS can approximate various characteristics of human dialogue, such as timing backchannel responses or optimizing turn-taking between speakers (Baumann and Schlangen, 2011; Skantze and Hjalmarsson, 2013; Lala et al., 2017; Khouzaimi et al., 2018, *inter alia*). Incremental processing can also improve the computational efficiency of an SDS, for example by accessing time-consuming external services (e.g. database queries) before the end of a user's turn.

What's more, incremental processing enables new types of interactions between a dialogue system and its users. For example, Kennington and Schlangen (2016) developed a personal assistant that could communicate its incremental understanding of the user's ongoing speech and its prediction states by graphically displaying a branching tree. At the start of the interaction, a tree with one branch per supported intent is displayed. Once the user's intent is recognized by the system, the tree is adjusted and its associated slots are displayed as subbranches to its node. This visual feedback increases the transparency of both the system's capabilities and current understanding; it also guides the user to provide values for all required slot for the task.

iNLU is not just relevant to spoken dialogue systems: recent work has enabled incremental processing in the NLU pipeline of RASA, a widely-

Proceedings of the Third Workshop on Natural Language Processing for Conversational AI, pages 52–67

used, open-source framework for building chat and voice-based virtual assistants (Rafla and Kennington, 2019; Bocklisch et al., 2017). When equipping a text-based dialogue system with incremental capabilities, iNLU processing is not limited by the rate at which the ASR posts intermediate results: user text can be processed as it is typed.

Intent classification, a common task assigned to the NLU module of a task-oriented SDS, poses the problem of identifying the point of the utterance where the intent has been classified with a given degree of certainty. An early identification of the correct intent, however, is not necessarily a sign of an effective classifier: when an iNLU module identifies the correct intent label before a human can (for example, after processing a not yet informative partial utterance such as "I'd like to"), then its "success" may likely be caused by the presence of artefacts in the training set.

In order to provide an upper baseline for incremental intent classification, we present inCLINC[1], a dataset of crowd-sourced incremental intent annotations, where utterances are broken into increasingly longer partial utterances by identifying peaks and troughs in surprisal (Shannon, 1948; Hale, 2001) as boundaries. We then compare the performance of human annotators to that of a Transformer-based classifier (Vaswani et al., 2017). We propose entropy reduction (between the different intent interpretation hypotheses) as a measure of uncertainty reduction during human incremental intent identification. We show that, for a substantial amount of the partial utterances, the final intent label is not yet identifiable to humans, and that a reduction in uncertainty (as annotators converge on an interpretation) is typically associated with an increase in accuracy for the annotators. We argue that, when the human annotators do not converge to one or a few possible interpretations and yet the classifier already identifies the final intent class early on, it is a sign of overfitting that can be ascribed to artefacts in the dataset.

1.1 Previous Work

The NLU module of a task-oriented SDS is commonly tasked with intent classification, where an utterance's most likely intent label is predicted. Intent can be operationalised as the main goal that one wants to achieve with a particular speech act (Cohen, 2019; Allen and Perrault, 1980). Such information is found on a pragmatic level: it reflects the overall meaning communicated by a person in a particular context. Accordingly, incremental intent classification is typically framed as a *predictive* task: based on the information present in a partial utterance, a classifier must predict what information will be present in the complete utterance.

	Word	Predicted Intent	Correct?
w_1	I	SearchCWork	No
w_2	want	SearchCWork	No
w_3	to	SearchCWork	No
w_4	hear	PlayMusic	Yes
w_5	any	PlayMusic	Yes
w_6	tune	PlayMusic	Yes
w_7	from	PlayMusic	Yes
w_8	the	PlayMusic	Yes
w_9	Twenties	PlayMusic	Yes

Table 1: Intent of SNIPS utterance incrementally predicted by a DistilBERT classifier. *SearchCreativeWork* is abbreviated as *SearchCWork*. "Correct?" indicates whether the predicted label would be considered as accurate in a typical incremental intent classification evaluation framework, where the complete utterance's label is assigned to each of its partial utterances.

Measuring Incremental Performance Accuracy and word savings are the most commonly reported metrics in studies on incremental intent classification in the literature. An incremental prediction is typically evaluated as *accurate* when the predicted label for a partial utterance matches the ground truth label of its complete utterance. *Word savings* are then used to show the point in an ongoing utterance at which a classifier first makes an accurate prediction. If a complete utterance has 12 words and a classifier successfully predicts its label after 10 words (w_{10}), then two words are saved. Additionally, Schlangen and Skantze (2011) and Baumann et al. (2011) present metrics specifically for incremental processors. As incremental intent classification involves the prediction of a single label (i.e. one "information unit"), *edit overhead* (EO $\in [0, 1]$) is most relevant[2]. It describes the ratio of unnecessary changes in label predictions to the total number of changes.

In the literature, different types of models have been applied to the task of intent classifi-

[1] https://fordatis.fraunhofer.de/handle/fordatis/213

[2] Other metrics are less relevant when only one IU is predicted. For example, correctness is equivalent to accuracy on the set of complete and partial utterances in this case.

cation for incremental NLU (e.g. DeVault et al., 2009; Manuvinakurike et al., 2018; Constantin et al., 2019; Coman et al., 2019; Madureira and Schlangen, 2020, *inter alia*). A typical approach is to segment complete utterances into increasingly longer partial utterances. Each partial utterance is then assigned the ground truth label of the complete utterance. This method, however, is overly simplistic, because a partial utterance does not necessarily contain enough information to be mapped to the given intent class of its complete utterance. Table 1 presents an utterance from the popular SNIPS dataset (Coucke et al., 2018). The classifier predicts *PlayMusic* for partial utterance w_4 = "I want to hear". However, this partial context is arguably not restrictive enough to be predictive of the class *PlayMusic*, especially for a SDS which may play back different kinds of information beyond music. It is not until the mention of "tune" in w_6 that the utterance arguably has enough semantic information to reasonably belong to the class *PlayMusic*. Until then, utterances beginning with "I want to hear" could easily be assigned to other intent classes. However, the phrase "I want to hear" is exclusively found in utterances belonging to the *PlayMusic* class in SNIPS. Rather than an important semantic distinction between intents, this characteristic reflects an artefact of the dataset, arising from an arbitrary choice of intent labels.

Existing approaches in the literature would view w_4 as correctly classified, as it matches the complete utterance's ground truth label. This approach inflates the assessment of classifier's accuracy and other performance metrics; the external validity of its performance on utterances outside of the dataset should be questioned. The example in Table 1 emphasises how *a*) adopting the complete utterance's label is inappropriate for many partial utterances and *b*) the performance of models in studies that do so risk being inflated by over-fitting.

1.2 Incremental Processing in Humans

Communication between human speakers and listeners is incremental on many levels. A speaker delivers information incrementally by speaking words one after the other. Exchanges of information between speakers and listeners unroll rapidly: human listeners interpret these incoming linguistic signals incrementally (Tanenhaus et al., 1995), rapidly forming both partial hypotheses based on what they hear, as they hear it (Marslen-Wilson, 1973), as well as expectations about the next signal.

The view of human language processing as being expectation-based has gained considerable support over the last 30 years in psycho-linguistic research. Research has shown that comprehenders form expectations at different levels of granularity (Zarcone et al., 2016): about the next upcoming word (Ehrlich and Rayner, 1981; McDonald and Shillcock, 2003), about its semantic category (Federmeier and Kutas, 1999), about the next event to follow in an ongoing sequence (Chwilla and Kolk, 2005), about verb selectional restrictions (Altmann and Kamide, 1999), and more. Crucially, hypotheses about the correct syntactic parse of the ongoing sentence are not only revised each time a new word is encountered; new predictions are also proactively made about the upcoming syntactic structure once this new information is integrated (Hale, 2001; Levy and Jaeger, 2007). Highly-predictable input is easier to process, as it matches the comprehender's expectations, but is also less informative. Conversely, input conveying more information given the context requires higher cognitive effort to process (Hale, 2001; Jaeger and Tily, 2011). More effort is required to process input when it results in a revision of these hypotheses than that which conforms with prior expectations (e.g. Sturt et al., 1999). The relationship between information content and processing difficulty has been described on a computational level (Marr, 1982) using Surprisal and Entropy Reduction.

Surprisal Surprisal defines the predictability of a linguistic unit (e.g. a word) in terms of its conditional probability given the context in which it appears (Shannon, 1948; Hale, 2001). Specifically, the Surprisal S of word w_t in a sentence given preceding words $(w_1, ..., w_{t-1})$ is equal to:

$$S(w_t) = -\log P(w_t|(w_1, ..., w_{t-1}, Ctxt)) \quad (1)$$

where $Ctxt$ represents extra-sentential information (e.g. visual cues, Knoeferle et al., 2008). Surprisal has been shown to be an effective complexity metric for the prediction of human sentence comprehension difficulty (Boston et al., 2008; Demberg and Keller, 2008; Frank, 2013; Levy and Jaeger, 2007; Levy, 2008). As a word's predictability is inversely proportional to its information content, we adopt Surprisal as a measure of information content at the word level - information which can contribute to the intent interpretation of an utterance.

Entropy Reduction Entropy is defined as the average amount of uncertainty at a given state associated with a random variable's possible outcomes (Shannon, 1948). If I is the set of all possible interpretations of a sentence, then the entropy of all possible interpretations can be expressed as:

$$H(I) = -\sum_{i \in I} P(I) \log_2 P(I) \quad (2)$$

Entropy is high when an ongoing sentence has many probable interpretations and is maximal when all possible interpretations have the same probability (the sentence is ambiguous). As words come in, they are "either helpful or unhelpful in narrowing down the interpretation" (Yun et al., 2015). Each new word w_t carries a certain amount of information, which can be used to *a*) revise existing hypotheses about the correct interpretation of an ongoing sentence and *b*) predict the interpretation of the remainder of the sentence (Hale, 2001; Levy and Jaeger, 2007).

Entropy reduction is then a measure of how much a given word w_t decreases the amount of uncertainty about the ongoing sentence being processed (Hale, 2003, 2006). It can be expressed as the difference in the entropy at state $t-1$ and the entropy at state t, i.e. before and after w_t:

$$\Delta H(w_t) = H(t-1) - H(t) \quad (3)$$

A greater reduction in uncertainty at a given step results in higher processing difficulty (Hale, 2006; Yun et al., 2015). Entropy reduction been shown to predict processing difficulty independently from Surprisal (Frank, 2013; Linzen and Jaeger, 2016). More specifically, two types of uncertainty can be identified, namely *a*) uncertainty about the next prediction step and *b*) uncertainty about the full sentence. Linzen and Jaeger (2016) investigated how both types impacted readers' parsing performance and showed that increased reading times were correlated with the *reduction* of uncertainty about the overall structure of an ongoing sentence but not with an *increase* in uncertainty about the next prediction step.

We adopt entropy reduction as a measure of how much a given word decreases the amount of uncertainty about the possible intent interpretation of the complete utterance. While Surprisal defines a word's information content from its conditional probability (estimated from co-occurrences), Entropy Reduction measures the changes in the human hypotheses at the (higher) intent level, at different stages in the utterance, and is estimated from the (partial) intent interpretations.

1.3 Uncertainty in Neural Networks

Uncertainty is not only relevant to human language processing; the estimation of uncertainty in deep neural networks is also an important field of research. For completeness, we will briefly introduce uncertainty and its estimation in this context. We refer the interested reader to a recent survey by Gawlikowski et al. (2021) for more information.

At its core, machine learning is interested in using data to extract models to then make predictions about unseen data (Hüllermeier and Waegeman, 2019). Such predictions are accompanied by predictive uncertainty, which can be distinguished as *aleatoric* and *epistemic* uncertainty (Kiureghian and Ditlevsen, 2009; Hüllermeier and Waegeman, 2019). For example, consider a neural network intent classifier that approximates an intent class $c \in C$ as a region in its sentence embedding space. Aleatoric uncertainty arises when such regions belonging to different intents in C overlap, whereas epistemic uncertainty is high for utterances that occur in regions in the input space sparsely populated by training instances. Points with high epistemic uncertainty could constitute an outlier or an out-of-scope utterance.

Perhaps the simplest way to estimate the predictive uncertainty of a deep neural network is to interpret its softmax output as a probability distribution. However, the softmax output distributions of deep neural networks are often poorly calibrated (Guo et al., 2017). Monte Carlo Dropout (MCD) is an alternative method for use in networks trained with dropout (Gal and Ghahramani, 2016). It interprets dropout as a Bayesian optimization approach that samples from the approximate posterior distribution of the model's parameters given the training data. Essentially, applying different dropout masks to drop different neurons from a single network can be viewed as creating an ensemble of different networks, which are treated as Monte Carlo samples from the space of all possible models for the task. By enabling dropout during inference, a prediction is generated by each network in this ensemble. The distribution of these predictions can be analysed to determine the predictive mean and the associated predictive uncertainty (e.g. the variance of this distribution) of an unseen sample at test time. Finally,

deep ensembles constitute another sampling-based approach, where an ensemble of neural networks with the same architecture are trained after being initialised with different values (Lakshminarayanan et al., 2016). As with MCD, the softmax outputs across all models is aggregated to quantify the uncertainty of the prediction.

2 Proposed Approach

We propose the theory of Entropy Reduction as a lens through which to view the problem of incremental intent classification. The theory suggests that more cognitive effort is needed to process an encountered word that greatly reduces the probable interpretations of the complete sentence, as compared to a word which does not. Like human listeners, iNLU modules form (multiple) early predictions (in this case, intent predictions) for partial utterances during incremental processing, which can be revised or revoked as more words arrive.

Of course, the computational cost of a typical iNLU module does not vary as a function of an input's Surprisal or the Entropy Reduction that it triggers. However, identifying which parts of an utterance trigger a considerable reduction in the set of plausible intent interpretations may be helpful when evaluating the performance of such a module. When the set of interpretations is too open (i.e. a partial utterance could conceivably belong to almost any intent), identifying one correct answer (and potentially acting upon it) does not make sense. As this set narrows, comparing the human hypotheses with the classifier's hypotheses may help understand the classifier's decisions.

Intent labels for complete utterances do not show how intent interpretations by humans change as an utterance progresses. To the best of our knowledge, however, there is no (publicly-available) dataset with incremental annotations of utterances with intent labels. Our first contribution is a dataset of partial and full utterances with human annotations of plausible intent labels at different portions of each utterance, which can provide an upper baseline for incremental intent classification. As a second contribution, we use the collected annotations for an analysis of the performance of a Transformer-based classifier on an equivalent task of incremental intent classification (Vaswani et al., 2017). We hypothesise that humans will outperform the classifier, achieving a higher overall accuracy and higher word savings due to correct early predictions. Responses from the humans are then used to examine over-fitting versus under-performance of the classifier.

Lastly, to showcase the application of our dataset, we present an analysis of the relationship between entropy reduction and increases in accuracy during incremental intent classification. Based on previous research, such as that by Linzen and Jaeger (2016), we hypothesise that a reduction in entropy between two subsequent partial utterances are more frequently associated with increases in accuracy than between partial utterances with no changes/increases in entropy.

3 Methods

inCLINC Dataset Clinc150 is a challenging intent classification dataset with utterances from 150 classes spanning across 10 different domains, plus out-of-scope (OOS) queries (Larson et al., 2020). Adapting it to an incremental setting, we created *inCLINC*, with utterances spanning a smaller number of intents (37 intents plus OOS) so that human annotators could become familiar with the task more quickly. OOS utterances were also included proportionally. Some intent names were modified to increase transparency, while avoiding the addition of key words that appear in the utterances themselves and could bias participants (see Appendix A). inCLINC was created from Clinc150's published test set[3]. Complete utterances were first split based on the presence of white space characters. An incrementalised set of partial utterances was then created for each full utterance w_n of length *n*: w_n was segmented into a set of *n* partial utterances, such that a partial utterance w_t contained *t* words. The complete utterance w_n was also included as a control. As this method produced a large set of stimuli (> 9000), we sought to identify a smaller set of partial utterances that were likely to introduce relevant (based on Surprisal) and thus potentially intent-relevant information, so that they could be shown to participants as stimuli.

3.1 Annotation Study

Stimulus selection We first computed the Surprisal of each token w_t in inCLINC by estimating their conditional probability using a pre-trained DistilGPT-2 model. We restricted the set of complete utterances by removing outliers (extremely

[3]The complete utterances from the original training and validation subsets (100 and 20 per intent, respectively) were reserved for training our intent classifier.

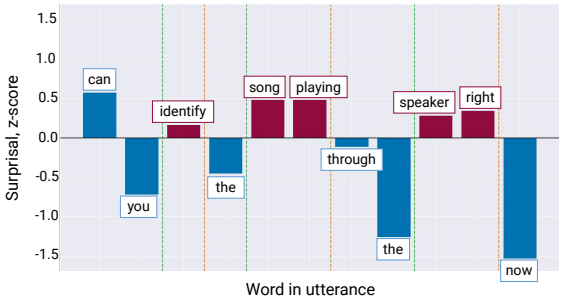

Figure 1: Segmenting an utterance into stimuli (six partial utterances plus one complete utterance) based on peaks in Surprisal.

high-surprisal words $>$ 2.5 SD from the standardized mean), utterances with less than 4 words, utterances starting with a keyword (belonging to a list of lemmatized intents names) and utterances that did not contain at least one peak in Surprisal (see below). We randomly selected four utterances per intent from this set.

We identified peaks in Surprisal to help select a subset of partial utterances (see example in Figure 1). Peaks were detected as (consecutive) words with a positive Surprisal z-score. The onset of a peak was marked as the transition point between a negative to positive z-score, whereas its offset was marked during the switch from positive to negative. The first word in an utterance always had high Surprisal: as such, it was only included in a peak if the second word also has a positive z-score. We selected partial utterance stimuli by breaking the utterances before and after each peak's onset.

In the end, a total of 630 partial and complete utterances were presented to participants, representing 152 distinct complete utterances across 37 intent classes, plus a OOS class.

Data Collection Stimuli were randomly divided across 17 batches of 37-38 stimuli each. Stimuli from the incrementalised set of the same complete utterance were put in separate batches. Each batch contained eight to nine complete utterances as controls. For each batch, a questionnaire was created using SoSciSurvey (Leiner, 2014), presenting one stimulus per page, along with a table listing the possible intents (see screenshot in Appendix B, Figure 3). The order of presentation of the stimuli and of the intent categories was randomized across participants and between stimuli.

Participants were instructed to predict the most likely intent of the complete query, based on the incomplete text provided. An attention check was included to confirm participants read the instructions. To familiarize them with the task, two example items were also presented. The intent of the first example was clear and was used as a control question for data cleaning. The second ("are you able to") was used to illustrate how multiple possible intents could match its complete query.

Participants Participants were recruited through Amazon Mechanical Turk[4] and redirected to the annotation questionnaires. They were paid for their participation and were informed that they were free to stop the task and delete their data at any point. They were asked to confirm their English fluency: those who did not were excluded from participation. Participants could complete multiple batches. We collected nine annotations per batch.

Data Cleaning We removed stimuli belonging to ambiguous complete utterances. A complete utterance was deemed ambiguous if the classifier's prediction did not match the ground truth label and/or $< 50\%$ of participants selected the same intent for a given complete utterance. We also excluded all answers from participants who either failed the attention checks and/or selected an incorrect intent for ≥ 3 unambiguous control stimuli.

3.2 Intent Classifier

A DistilBERT[5] Transformer model with a linear layer classification head was fine-tuned to classify the intent class of the pooled output (CLS token) (Sanh et al., 2019). DistilBERT was selected to mimic a feasible set-up for an online, incremental processing setting. All weights were trained for 2 epochs (for more details, see Appendix C).

3.3 Annotated Dataset

After data cleaning, the inCLINC dataset included 121 distinct utterances in their complete form (121) and in partial form (417), for a total of 538 annotated utterances. After data cleaning, each utterance (partial or complete) had six to nine annotations. Each utterance was then assigned a predicted human label as the response with the highest number of votes. In case of a tie, we assigned the label of the complete query or, if this was not among the list of intent labels with the most votes, then the predicted intent label was randomly selected from

[4] https://www.mturk.com/

[5] Checkpoint from *Hugging Face* library https://huggingface.co/transformers/

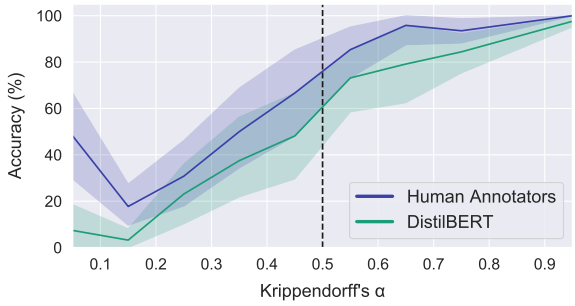

Figure 2: Mean accuracy at different levels of agreement. Stimuli were divided into 10 equal-sized bins based on $\alpha \in [0, 1]$ (shadow shows standard deviation).

	Accuracy	EO	WC Savings
Annotators	66.43%	0.39	2.43
DistilBERT	56.35%	0.45	1.94

Table 2: Performance on partial utterance stimuli.

the top-voted list. inCLINC with its annotations (majority-vote labels, as well as single labels) has been made publicly available[6]. 365 unique words appear in the inCLINC dataset. For descriptive statistics of the stimuli, see Table 5 in Appendix A.

4 Results

Annotation reliability To verify response reliability, we measured participant agreement on the set of complete utterances using Krippendorff's α (Krippendorff, 2011). On these control stimuli, participant responses reached $\alpha = 0.80$, reflecting substantial to perfect agreement (Artstein and Poesio, 2008). Figure 2 shows a positive trend between α and accuracy for both participants' and DistilBERT's predictions, supporting task validity: the more participants agree on the label for a partial utterance, the more often they identify the complete utterance's intent label.

Accuracy For the complete utterances in Clinc150's test set[7], DistilBERT achieved an accuracy of 94.56%. However, for inCLINC, Table 2 shows that the annotators outperformed DistilBERT by over 10% for partial utterances. What's more, the annotators "only" reached an accuracy of 66%: for many partial utterances, the complete utterance's intent is not discernible.

[6] https://fordatis.fraunhofer.de/handle/fordatis/213

[7] Utterances from the original test subset that belong to a class included in inCLINC.

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

Table 3: Frequency table showing entropy reduction (ER) of stimuli and associated increases/decreases in accuracy, as compared to the previous partial utterance.

Word Chunk (WC) Savings Complete utterances were chunked based on Surprisal peaks and troughs, and thus sequential partial utterances differed by a variable number of words. As such, rather than the absolute number of words saved, we examine *word chunk savings* (WC savings), i.e. how many stimuli earlier than the complete utterance was the final intent first predicted. As shown in Table 2, human annotators achieved a higher mean WC savings than DistilBERT. Despite their superior performance, human annotators met an upper bound: they had zero WC savings for about 6% of utterances (13% for DistilBERT), whereas only 1 WC was saved for about 28% of utterances (28% for DistilBERT).

Edit Overhead (EO) Table 2 reports a lower Edit Overhead (EO) for annotators than DistilBERT. Participants not only predicted the final label earlier, but were also more consistent with their predictions throughout an utterance.

Entropy Reduction (ER) Entropy reduction (ER) was considered as an independent variable with categories ER < 0 and ER ≥ 0. We tested whether a difference in outcomes across these ER categories exists, where the possible outcomes were *a*) an increase in accuracy between processing the sequential partial utterances (i.e. predicted label was incorrect for the previous partial utterance before but was correct for the following partial utterance) *b*) a decrease/no change in accuracy. McNemar's test was performed using the binomial probability distribution (McNemar, 1947). The frequencies in Table 3 differed significantly across ER categories ($p < 0.001$, one-sided). More specifically, partial utterances characterised by Entropy Reduction (ER < 0) were more frequently accompanied by an increase in participant accuracy than by an increase in entropy/no change (ER ≥ 0).

5 Discussion

Humans as an Upper Baseline Bender and Koller (2020) debate how much *meaning* a trained

neural language model *understands* and argue that, when a system outperforms inter-annotator agreement, the task likely contains artefacts that do not represent meaning. As such, over-fitting in the context of incremental intent prediction can be assessed by examining the cases where DistilBERT predicts the complete label for an earlier partial utterance than the annotators. Conversely, areas where a model could be improved can be studied by looking at utterances where the annotators predicted the complete utterance's intent earlier. We looked at partial utterances whose complete label was predicted by either the annotators or the classifier, but not both. These disagreements, found in 16% of partial utterances, were used to examine over-fitting and under-performance.

Evidence for Overfitting For eight partial utterances, DistilBERT predicted the complete utterance's label, but the majority of participants did not. These partial utterances included: "tell my", "i have to", and "on the" (complete list in Table 7, Appendix D). None of these utterances can be clearly assigned to a specific intent: this prediction is a "lucky guess" based on artefacts distinguishing classes. Popular NLU intent benchmarks with notably fewer classes, such as ATIS and SNIPS (Hemphill et al., 1990; Coucke et al., 2018), may contain more such artefacts, which would speak against the generalisability of the results obtained on them. Furthermore, existing studies that label partial utterances with the complete utterance's label do not distinguish between such "lucky guesses" and points where the intent is identifiable: reported performance could be inflated by overfitting.

Evidence for Under-Performance 61 utterances were correctly predicted by the participants but not by DistilBERT. Of these, 24 had at least moderate agreement ($\alpha \geq 0.50$) between participants (values reported in Appendix D). These utterances do not represent the point where intent is identifiable by a vast majority of participants. Rather, they represent subtle differences in formulation of utterances that humans, but not the classifier, might associate with a certain intent class, which results in a considerable amount of (but not all) participants predicting the final intent.

A few interesting observations can be made. First, "i need milk" was correctly assigned to *Update/add to shopping list* by participants, while DistilBERT predicted *Place an order*. People are probably more likely to buy perishable items such as milk at a store in-person: a failure to "understand" this real-world knowledge might have caused the classifier to miss this prediction. Next, the familiarity of participants with the well-known expression "how's the weather" is visible in the agreement of $\alpha = 1.0$ for the partial utterance "how's the" and class *Get Weather*. The completion "how's the" to "how's the weather" suggests that "how's the" is a high-cloze phrase (Taylor, 1953) and "weather" is a highly-predictable continuation. Note that the missing word "weather" is also found in the label for its intent, and that "how's the weather" matches an utterance one would expect to hear in the context of a task-oriented SDS with inCLINC's supported intents. This example can be considered the human-equivalent of the artefact in Table 1. However, humans make such a prediction by generalising over their own statistical experience accumulated over a lifetime of language exposure, while a model only has access to the patterns it has learned to represent based on those found in the (domain-specific, limited) data on which it was trained.

Entropy Reduction During incremental intent classification, not all steps contribute equally to the final interpretation. We identified Surprisal peaks and troughs as relevant points to break the utterances into informative incremental chunks, as informed by the token's conditional probability. We then proposed ER (computed from the human interpretations) as a metric to demonstrate the incremental narrowing of the set of plausible final intents at these breaking points.Utterances accompanied by a reduction in entropy are more frequently associated with an increase in accuracy compared to those with no changes/increases in entropy. ER serves as a tangible metric of how much certain (chunks of) words, driven by information content, restrict the set of plausible final intents and is thus accompanied by increased prediction accuracy.

Limitations Changes in ER across partial utterances could be affected by *a*) different representations of the intent classes across participants and/or *b*) their differing language abilities. This limitations arises as our (relatively small number of) participants were recruited online and their English fluency was self-reported[8]. With six to nine annotations per item, each response has significantly influence the set of annotations for a given item. Collecting more annotations would be beneficial to

[8]Fluency was indirectly verified by excluding participants who gave incorrect responses for multiple control utterances.

reduce fluctuations in the distribution of responses triggered by individual participants and to allow for a more in-depth study of the role of Entropy Reduction in incremental intent classification (controlling for e.g. the Surprisal and the predictability of the next word, the position in which reductions of entropy are expected, etc.). Additionally, when the complete utterance's label was among those with the most votes, then this label was selected as the prediction: this is a biased selection of a favourable answer. 3.5% of stimuli were resolved using this heuristic and annotators' accuracy in Table 2 is arguably inflated by this amount. Finally, only one classifier was used in the presented work. Comparing different models' performance to our human upper baseline would be interesting.

6 Conclusion

Assigning an utterance's ground-truth label to its partial utterances is a common oversimplification of incremental intent classification. Existing studies report global performance metrics across partial utterances and fail to distinguish between those not containing enough information to be representative of the final intent and those who do. It is then unclear to what degree a classifier's performance is attributable to overfitting to the evaluated dataset versus to a generalisable representation of intent classes. We proposed an alternate lens to view this task: evaluating a model of incremental intent classification should not just be a matter of *getting the intent right*. Rather, it should be about predicting a set of plausible labels for an ongoing utterance and communicating with what certainty the final intent has been classified at this point.

As an alternative to simply adopting the complete utterance's label, we presented a new dataset with annotated incrementalised utterances. We proposed a novel method for determining plausible intent annotations at relevant points – from an information-theoretic point of view – in an ongoing utterance. With humans as an upper baseline, the performance of incremental intent classifiers may be evaluated against our labeled annotations. The remainder of our work then focused on the evaluation of a Transformer-based intent classifier in an incremental setting. To analyse our dataset, we proposed ER as a metric to detect the incremental narrowing of intent interpretations by human annotators. Our analysis showed that ER more frequently accompanied increased prediction accuracy for the annotators compared to decreases/no changes in accuracy. ER has the potential to identify the earliest point of understanding, before which accuracy and word savings are not informative measures to evaluate the performance of an incremental classifier, as before this point the upper performance bound is unknown.

While our work did not focus on the training of incremental intent classifiers, inCLINC opens up several new possibilities. For example, its annotations could be used to assess and improve the confidence calibration (in the strong sense, see Vaicenavicius et al., 2019) of multi-class classifiers to ensure that their softmax output more reliably estimates the uncertainty about the (final) intent at a given point in an ongoing utterance. Providing strongly-calibrated confidence scores alongside predictions would help increase model interpretability, facilitate its integration into other probabilistic models (Guo et al., 2017), and help coordinated processing across modules in an incremental SDS (Schlangen and Skantze, 2011). Alternatively, the annotations could be used as labels in a multi-intent classification setting for partial utterances[9]. This approach would circumvent the problem of assigning a complete utterance's label to a partial utterance with not yet enough semantic information relevant to the target class: all plausible labels could be trained for partial utterances. Finally, ER in a SDS could be monitored across ongoing utterances to learn points at which the entropy of possible interpretations is low enough for the given task. A dataset with annotations such as inCLINC could provide useful labels for this paradigm. All in all, models which are able to overcome the challenges of incremental intent classification could have the potential to learn more robust and generalisable class representations, which will could conceivably improve their performance in a real-life applications.

Acknowledgements

This research was carried out while the first author was affiliated with the Fraunhofer IIS, as a member of the Semantic Audio Processing department. It was funded by the German Federal Ministry for Economic Affairs and Energy (BMWi) through the SPEAKER project (FKZ 01MK19011).

[9] E.g. "I need to cancel my" could represent many inCLINC intents, such as *Cancel a reservation*, *Update/add to reminders*, and *Update/add to to-list*.

References

James F. Allen and C. Raymond Perrault. 1980. Analyzing intention in utterances. *Artificial Intelligence*, 15(3):143–178.

Gerry T.M. Altmann and Yuki Kamide. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3):247–264.

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Timo Baumann, Okko Buß, and David Schlangen. 2011. Evaluation and optimisation of incremental processors. *Dialogue and Discourse*, 2(1):113–141.

Timo Baumann and David Schlangen. 2011. Predicting the micro-timing of user input for an incremental spoken dialogue system that completes a user's ongoing turn. In *Proceedings of the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 120–129.

Emily M Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2, pages 5185–5198.

Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open Source Language Understanding and Dialogue Management. *arXiv preprint arXiv:1712.05181*.

Marisa Ferrara Boston, John Hale, and Reinhold Kliegl. 2008. Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, 2(1):1–12.

Paul T Brady. 1968. A statistical analysis of on-off patterns in 16 conversations. *Bell System Technical Journal*, 47(1):73–91.

Dorothee J. Chwilla and Herman H.J. Kolk. 2005. Accessing world knowledge: Evidence from N400 and reaction time priming. *Cognitive Brain Research*, 25(3):589–606.

Herbert H Clark. 1996. *Using Language*. Cambridge University Press.

Philip R. Cohen. 2019. Foundations of collaborative task-oriented dialogue: What's in a slot? In *Proceedings of the - 20th Annual Meeting of the Special Interest Group Discourse Dialogue - Proceedings of the Conference (SIGDIAL)*, September, pages 198–209.

Andrei C. Coman, Koichiro Yoshino, Yukitoshi Murase, Satoshi Nakamura, and Giuseppe Riccardi. 2019. An incremental turn-taking model for task-oriented dialog systems. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 4155–4159.

Stefan Constantin, Jan Niehues, and Alex Waibel. 2019. Incremental processing of noisy user utterances in the spoken language understanding task. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 265–274.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips Voice Platform: an embedded Spoken Language Understanding system for private-by-design voice interfaces.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

David DeVault, Kenji Sagae, and David Traum. 2009. Can I Finish? Learning when to respond to incremental interpretation results in interactive dialogue. In *Proceedings of the 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 11–20.

Susan Ehrlich and Keith Rayner. 1981. Contextual Effects on Word Perception and Eye Movements during Reading. *Journal of Verbal Learning and Verbal Behavior*, 20(6):641–655.

Kara D. Federmeier and Marta Kutas. 1999. A Rose by Any Other Name: Long-Term Memory Structure and Sentence Processing. *Journal of Memory and Language*, 41(4):469–495.

Stefan L. Frank. 2013. Uncertainty Reduction as a Measure of Cognitive Load in Sentence Comprehension. *Topics in Cognitive Science*, 5(3):475–494.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *33rd International Conference on Machine Learning, ICML 2016*, 3:1651–1660.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. 2021. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 2130–2143.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 1–8.

John Hale. 2003. The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, 32(2):101–123.

John Hale. 2006. Uncertainty about the rest of the sentence. *Cognitive Science*, 30(4):643–672.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, pages 96–101.

Eyke Hüllermeier and Willem Waegeman. 2019. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *arXiv preprint arXiv:1910.09457*.

T Florian Jaeger and Harry Tily. 2011. On language 'utility': Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(3):323–335.

J. Jaffe and S. Feldstein. 1970. *Rhythms of dialogue*. Academic Press, New York, NY.

Casey Kennington and David Schlangen. 2016. Supporting spoken assistant systems with a graphical user interface that signals incremental understanding and prediction state. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 242–251.

Hatim Khouzaimi, Romain Laroche, and Fabrice Lefèvre. 2018. A methodology for turn-taking capabilities enhancement in Spoken Dialogue Systems using Reinforcement Learning. *Computer Speech and Language*, 47:93–111.

Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? Does it matter? *Structural Safety*, 31(2):105–112.

Pia Knoeferle, Boukje Habets, Matthew W. Crocker, and Thomas F. Münte. 2008. Visual scenes trigger immediate syntactic reanalysis: Evidence from ERPs during situated spoken comprehension. *Cerebral Cortex*, 18(4):789–795.

Klaus Krippendorff. 2011. Computing Krippendorff's Alpha-Reliability. Departmental Papers (ASC).

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2016. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*.

Divesh Lala, Pierrick Milhorat, Koji Inoue, Masanari Ishida, Katsuya Takanashi, and Tatsuya Kawahara. 2017. Attentive listening system with backchanneling, response generation and flexible turn-taking. In *Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 127–136.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2020. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*, pages 1311–1316.

Dominik J Leiner. 2014. SoSci survey (version 2.4. 00-i).

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Roger Levy and T Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems (NIPS)*, volume 19, page 849–856. MIT Press.

Tal Linzen and T. Florian Jaeger. 2016. Uncertainty and Expectation in Sentence Processing: Evidence From Subcategorization Distributions. *Cognitive Science*, 40(6):1382–1411.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.

Brielen Madureira and David Schlangen. 2020. Incremental processing in the age of non-incremental encoders: An empirical assessment of bidirectional models for incremental NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 357–374, Online. Association for Computational Linguistics.

Ramesh Manuvinakurike, Trung Bui, Walter Chang, and Kallirroi Georgila. 2018. Conversational image editing: Incremental intent identification in a new dialogue task. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 284–295, Melbourne, Australia. Association for Computational Linguistics.

David Marr. 1982. *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. W. H. Freeman, San Francisco, CA, USA.

William Marslen-Wilson. 1973. Linguistic structure and speech shadowing at very short latencies. *Nature*, 244(5417):522–523.

Scott A. McDonald and Richard C. Shillcock. 2003. Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, 14(6):648–652.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Andrew Rafla and Casey Kennington. 2019. Incrementalizing RASA's Open-Source Natural Language Understanding Pipeline. *arXiv preprint arXiv:1907.05403*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *Proceedings of EMC2: 5th Edition Co-located with NeurIPS 2019*.

David Schlangen and Gabriel Skantze. 2011. A General, Abstract Model of Incremental Dialogue Processing. *Dialogue & Discourse*, 2(1):83–111.

C. E. Shannon. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423.

Gabriel Skantze and Anna Hjalmarsson. 2013. Towards incremental speech generation in conversational systems. *Computer Speech and Language*, 27(1):243–262.

Patrick Sturt, Martin J. Pickering, and Matthew W. Crocker. 1999. Structural Change and Reanalysis Difficulty in Language Comprehension. *Journal of Memory and Language*, 40(1):136–150.

Michael K. Tanenhaus, Michael J. Spivey-Knowlton, Kathleen M. Eberhard, and Julie C. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

Wilson L. Taylor. 1953. "Cloze Procedure": A New Tool for Measuring Readability. *Journalism Quarterly*, 30(4):415–433.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas B. Schön. 2019. Evaluating model calibration in classification. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3459—-3467.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5999–6009.

Jiwon Yun, Zhong Chen, Tim Hunter, John Whitman, and John Hale. 2015. Uncertainty in processing relative clauses across East Asian languages. *Journal of East Asian Linguistics*, 24(2):113–148.

Alessandra Zarcone, Marten Van Schijndel, Jorrig Vogels, and Vera Demberg. 2016. Salience and attention in surprisal-based accounts of language processing. *Frontiers in Psychology*, 7:844.

A Description of inCLINC

The original training/validation/test subsets published in Larson et al. (2020) were used to create inCLINC. Expanded intent names were created for some of the labels in inCLINC to clarify their meaning. The addition of key words that appear in the utterances themselves was avoided. As well, intent labels were grouped into six categories (plus OOS) for presentation to participants. Table 4 shows the mapping of the original labels to their expanded labels and their assigned category. Descriptive statistics for stimuli in inCLINC are presented in Table 5.

B Presentation of Task

Figure 3 shows how a partial utterance was displayed to participants during Experiment 2. The intent categories were randomly shuffled between the presentation of each stimulus.[10] The order in which intents within a given category were displayed was not shuffled. This design choice was made to reduce the cognitive load of participants: as participants were not familiar with the dataset, shuffling the order of items within the category could be frustrating and discourage participants from making thoughtful predictions about the intent for short utterances, especially when intent is not clear (e.g. for the utterance "I want").[11] This choice has the added advantage that similar intents are always grouped, encouraging participants to make fine-grained decisions between similar intents.[12]

C Training DistilBERT

The intent classifier used HuggingFace's implemented *DistilBertForSequenceClassification* architecture. The final linear layer of the model consisted of 38 output neurons: one for each 37 in-scope classes, plus an additional neuron for the out-of-scope class. The weights in the classifier's task-specific classification head were randomly initialized; the weights in the encoder layers were loaded from the pre-trained model. The weights in the dropout layer had a probability $p = 0.2$ of

[10] The intent categories are the headers of the blue boxes: *Shopping, Events & Tasks, Music, Out-of-Scope, Restaurant, Cooking, Tools & Utilities*.

[11] Participants were made familiar with the displayed intents in three practice rounds before the presentation of experimental stimuli.

[12] For example, *Update/add to calendar* is always presented under *Ask about calendar*.

inCLINC Category	Clinc150 Domain	inCLINC Label	Clinc150 Label
Cooking	Kitchen and Dining	Ask about calories	calories
Cooking	Kitchen and Dining	Ask about cook time	cook time
Cooking	Kitchen and Dining	How long food lasts	food last
Cooking	Kitchen and Dining	Ingredients for recipes	ingredient list
Cooking	Kitchen and Dining	Ingredient substitution	ingredient substitution
Cooking	Kitchen and Dining	Ask for meal suggestion	meal suggestion
Cooking	Kitchen and Dining	Nutrition information	nutrition info
Cooking	Kitchen and Dining	Get recipe	recipe
Events & Tasks	Home	Ask about calendar	calendar
Events & Tasks	Home	Update/add to calendar	calendar update
Events & Tasks	Home	Ask about reminders	reminder
Events & Tasks	Home	Update/add to reminders	reminder update
Events & Tasks	Home	Ask about to-do list	todo list
Events & Tasks	Home	Update/add to to-do list	todo list update
Events & Tasks	Home	Update/add to playlist	update playlist
Music	Tools and Utilities	Calculator	calculator
Music	Home	Next song	next song
Music	Home	Play music	play music
Music	Home	Identify song	what song
Restaurant	Kitchen and Dining	Accept a reservation	accept reservations
Restaurant	Kitchen and Dining	Cancel a reservation	cancel reservation
Restaurant	Kitchen and Dining	Confirm a reservation	confirm reservation
Restaurant	Kitchen and Dining	How busy is restaurant	how busy
Restaurant	Kitchen and Dining	Make a reservation	restaurant reservation
Restaurant	Kitchen and Dining	Ask for restaurant review	restaurant reviews
Restaurant	Kitchen and Dining	Ask for restaurant suggestion	restaurant suggestion
Shopping	Home	Place an order	order
Shopping	Home	Ask about order status	order status
Shopping	Home	Ask about shopping list	shopping list
Shopping	Home	Update/add to shopping list	shopping list update
Utilities	Tools and Utilities	Ask about date	date
Utilities	Tools and Utilities	Find phone	find phone
Utilities	Tools and Utilities	Make call	make call
Utilities	Tools and Utilities	Share location	share location
Utilities	Home	Smart home function	smart home
Utilities	Tools and Utilities	Text	text
Utilities	Tools and Utilities	Weather	weather
Out-Of-Scope	Out-Of-Scope	Out-Of-Scope	OOS

Table 4: Mapping of original labels from Clinc150 to categories and labels in inCLINC.

Figure 3: Stimulus Presentation in Experiment 2. The complete utterance is “i need buy a birthday gift for sue taken off my calendar”, belonging to the intent class *Update/add to calendar*.

	Stimuli	Mean	SD	Min	Max
Length (# words)	Full Utterances	8.83	3.02	5.00	20.00
	Partial Utterances	5.35	3.11	1.00	19.00
# of Partials Utterances	Full Utterances	3.45	1.74	1.00	12.00
	Partial Utterances	–	–	–	–
# of Annotations	Full Utterances	7.47	0.82	6.00	9.00
	Partial Utterances	7.41	0.83	6.00	9.00
Δ Entropy	Full Utterances	-0.23	0.46	-1.49	0.45
	Partial Utterances	-0.31	0.59	-1.95	1.15
Krippendorff's α	Full Utterances	0.89	0.16	0.46	1.00
	Partial Utterances	0.57	0.35	0.00	1.00
z-Surprisal	Full Utterances	0.12	1.02	-1.66	2.49
	Partial Utterances	0.06	1.01	-1.65	2.49

Table 5: Descriptive statistics for inClinc stimuli. Length refers to the number of whitespace-separated words in a stimulus. z-Surprisal refers to the Surprisal value for the last word in the stimulus, standardized based on all Surprisal values computed from utterances in the original Clinc150 dataset.

being dropped. All weights were fine-tuned using the complete utterances in the training subset.

Model fit during training was evaluated with respect to the NLL loss for the validation subset. The classifier was trained for 5 epochs, using an initial learning rate of $2e{-}5$ that was linearly decreased by a factor of 0.2. An AdamW optimizer was used, which performs gradient bias correction and weight decay to improve regularization (Loshchilov and Hutter, 2019). Optimization was performed with respect to the NLL on the validation subset.

Validation NLL was the lowest after epoch 2; this is the model that was retained for testing. After training, the classifier had a 94.56% accuracy on the complete utterances on the test set.[13]

D Disagreement between Participants and DistilBERT

The partial utterances whose complete label was either successfully predicted by either the coders or the classifier, but not both, were inspected. These disagreements were found in 69 presented stimuli out of 538 (12.82%). For 61 of these stimuli, the coders successfully predicted the correct complete label while the classifier did not; the agreement of 24 of these utterances surpassed the threshold of moderate agreement $\alpha = 0.50$. These 24 utterances are presented in Table 6. For the remaining eight of 69 disagreements, the classifier was the one with the successful prediction. These stimuli are presented in Table 7. Note that none of the stimuli reported in Table 6 involved ties.

[13]Here, the test set refers to all complete utterances from the original Clinc150 test split, before sampling the select stimuli to be presented to participants.

#	Partial Utterance	Predicted, DistilBERT	Label, Complete Utterance	α
1	i need milk	Place an order	Update/add to shopping list	0.51
2	put the dishes on my	Smart home function	Update/add to to-do list	0.51
3	please give me the name of a few good options	Ask for meal suggestion	Ask for restaurant suggestion	0.51
4	i'll be confirming	Share location	Confirm a reservation	0.52
5	please put chips	Ingredient Substitution	Update/add to shopping list	0.52
6	what is the solution	OOS	Calculator	0.58
7	what events do i have going on on	Ask about to-do list	Ask about calendar	0.65
8	can you help me hunt for my missing	Ask about reminders	Find phone	0.65
9	please give me the name of a few good options	Identify song	Ask for restaurant suggestion	0.65
10	do they serve good tacos	Ask for meal suggestion	Ask for restaurant review	0.65
11	what events	OOS	Ask about calendar	0.70
12	after how much time is it still safe	Ask about date	How long food lasts	0.70
13	get reservations	Accept a reservation	Make a reservation	0.70
14	i need a table	Cancel a reservation	Make a reservation	0.74
15	please identify	Share location	Identify song	0.74
16	will this recipe still be good if i use	Get recipe	Ingredient Substitution	0.74
17	please identify the	Share location	Identify song	0.74
18	can you identify	Share location	Identify song	0.74
19	i don't need mowing the lawn on my	Cancel a reservation	Update/add to to-do list	0.74
20	get reservations at	Accept a reservation	Make a reservation	1.00
21	get reservations at olive garden	Accept a reservation	Make a reservation	1.00
22	can i store bread	Ingredient Substitution	How long food lasts	1.00
23	how's the	OOS	Weather	1.00
24	can you identify this	Share location	Identify song	1.00

Table 6: Stimuli where only the majority of human annotators successfully predicted the final intent, but DistilBERT made an incorrect prediction. α refers to Krippendorff's α for the responses of the annotators for the presented stimuli. Only stimuli with $\alpha \geq 0.50$ are presented. OOS represents the "out-of-scope" class label.

#	Partial Utterance	Predicted Label, Humans	Label, Complete Utterance	α
1	tell my	Text	Share location	0.29
2	i have to	OOS	Cancel a reservation	0.31
3	set the	OOS	Smart home function	0.38
4	what things are on my	OOS	Ask about to-do list	0.38
5	tell me what tomorrow's	Weather	Ask about date	0.46
6	on the	OOS	Smart home function	0.51
7	i have a reservation for strip house for jennifer that i'd	Cancel a reservation	Confirm a reservation	0.70
8	put the dishes on my list	Update/add to shopping list	Update/add to to-do list	0.70

Table 7: Partial utterances where DistilBERT successfully predicted the complete utterance's label (column 4), but not the majority of human annotators. α refers to the agreement of human annotators for each stimuli. OOS represents the "out-of-scope" class label.