A Multi-level Framework for Understanding Spoken Dialogue Using Topic Detection

Andrew P. Valenti, Ravenna Thielstrom, Felix Gervits, Michael Gold, Derek Egolf, Matthias Scheutz
Human-Robot Interaction Laboratory
Tufts University, Medford, MA 02155
{andrew.valenti, ravenna.thielstrom, felix.gervits,
michael.gold, derek.egolf, matthias.scheutz}@tufts.edu

Abstract

Spoken Dialogue Systems (SDS) are used to interact with intelligent agents through natural language. Speech processing errors may cause the system to fail to generate an appropriate response. In this paper, we present a novel framework for understanding spoken dialogue in which utterance analysis is escalated through a multi-level system *according to the feedback retrieved at the syntactic, semantic, and contextual/topic level*. Analysis is applied incrementally at each level as the system attempts to resolve the uncertainty surrounding utterance interpretation. We demonstrate how our multi-level approach can be integrated with other SDS components to improve its ability to recognize spoken task commands. We evaluate this by comparing the interpretation accuracy of utterances from two task domains given as input to an SDS, under two experimental conditions: one with the multi-level framework and one without.

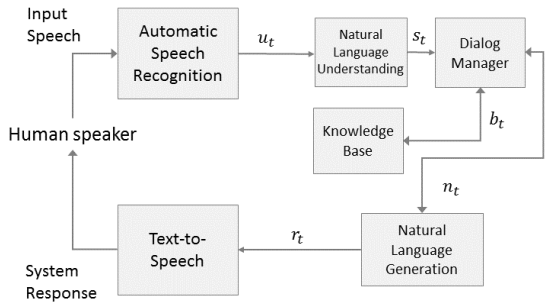


Figure 1: Typical spoken dialogue system. At each turn t, input speech is converted to an utterance, u_t, which the Natural Language Understanding component maps to an internal representation, s_t of the human's intent. The Dialogue Manager uses this to update the agent's belief state in the Knowledge Base, b_t and then infers a natural language form, n_t from Natural Language Generation which initiates a response, r_t to the Text-to-Speech component.

1 Introduction

Humans use Spoken Dialogue Systems (SDSs) to interact with intelligent agents using speech-based natural language (De Mori, 1997; Zue and Glass, 2000; Jokinen and McTear, 2009). Figure 1 shows components typically found in such systems (Scheutz et al., 2019; Young et al., 2013). Here, the Automatic Speech Recognizer (ASR) recognizes the human's utterance, u_t, and sends it to the Natural Language Understanding (NLU) component. However, problems may arise in the ASR component which can propagate through the system and cause it to fail to generate an appropriate response. For example, the ASR may recognize the word "Iraq" instead of "a rock" (Sarma and Palmer, 2004), or it may hear a novel word it has not yet learned (Scheutz et al., 2017). Alternatively, the user may believe the system to be capable of retrieving the weather report when its domain is retrieving movie listings; in such a case, the system will need to respond to the user's out-of-domain (OOD) request (Tur et al., 2014). Finally, in a multilingual environment the SDS may switch between different languages, e.g., a robot that a human can query in English or Japanese to initiate a Wikipedia search (Wilcock and Jokinen, 2015).

The ASR cannot recognize what it does not know about, and in the cited examples the researchers solved this problem by extending the ASR vocabulary or by adjusting the prior probability of the hypothesized word sequences. However, detecting and interpreting the user's true intention, and selecting an appropriate response given noisy human speech and ASR transcription errors, requires a method for communicating between SDS components. For example, the NLU, Knowledge Base, and Dialogue Manager components can request the recognition subsystem to reinterpret the utterance in the event of say, a processing failure.

In this paper, we show a novel framework for understanding spoken dialogue in which utterance analysis is escalated through a multi-level system *involving interpretation on syntactic, semantic, and contextual/topic levels* (see Figure 2). Analysis is applied incrementally at each level as the system attempts to resolve the uncertainty surrounding utterance interpretation. Links to other SDS components from each of the levels can affect the agent's beliefs and, conversely, other components can signal the framework to reinterpret the utterance. This may occur in the context of, for example, a new topic. To our knowledge, no other approach has demonstrated the use of such a multi-tiered system for improving accuracy in the SDS' ability to recognize spoken task commands. This paper proceeds as follows. In Section 2, we discuss prior approaches to resolving out-of-domain requests, using context to improve ASR and parser performance, and learning novel words. Section 3, situates these approaches in our framework. We also discuss how *topic detection* is used to determine context. Section 4, discusses the framework demonstration and evaluation, using utterances drawn from two task domains in two conditions: one with the framework and one without. Finally, Section 5, examines the advantages, disadvantages, and limitations of this approach and framework improvements.

2 Related Work

Research in improving NLU in task-oriented dialogue systems and intelligent agents can be motivated as follows. One way to ensure reliable performance of speech recognition for SDSs is to make a *closed-world* design assumption, and limit their operation to well-defined domains (for examples, see Lane et al. (2005)). This could be accomplished by representing the dialogue model as a finite state system using a pre-defined state transition network, which assumes that the dialogue is known in advance (McTear, 1998). This approach is not resilient to input outside the agent's domain, and so frame-based dialogue systems have been proposed. In these, the model attempts to fit the dialogue into frame slots (i.e., a "form") corresponding to an action or utterance (Xu and Rudnicky, 2000). However, these systems struggle when utterances fail to fit into a frame. Finally, agent-based systems have been developed for more advanced dialogue management with interactive agents (Allen and Perrault, 1980; Cohen and Perrault, 1979). These approaches typically categorize utterances into *speech acts* (Searle and Searle, 1969) and perform inference or planning based on a mental model of the interlocutor. Such systems make use of robust symbolic representations of an agent's beliefs, desires, intentions, and other properties, allowing for mixed-initiative interaction (Rao and Georgeff, 1991). This is the dialogue management approach used in the present system.

It is desirable that the human be able to communicate in a natural and flexible manner with the agent. To enhance usability, NLU systems are built on *open-world* assumptions. In these systems, the user may provide both in-domain and OOD inputs, the latter of which may be unsupported by the system. Accepting OOD inputs could lead to errors propagating through the system, which may lead to undesirable responses unless it can reliably distinguish between the two and process them accordingly. Context detection is one approach researchers have used for OOD. Veale et al. (2013) discuss a method for applying top-down contextual bias based on the expected dialogue turn to a neural speech recognition system to improve its performance. Sarma and Palmer (2004) compute the likely contexts of all words in an ASR system vocabulary by performing a lexical co-occurrence analysis using a large corpus of output from the speech system. This is used to find the likely context for query words, and the system uses this to identify similarly-sounding, but erroneous query words.

Topic detection may also be used to infer context. Lane et al. (2006) proposed a detection framework which makes use of the classification confidence scores of multiple topics and applies a linear discriminant model to perform in-domain verification. Lane et al. (2005) describe an architecture which combines topic detection with topic-dependent language models for use in a multi-domain SDS. According to the researchers, their approach allows the user to freely switch among domains while maintaining a high-level of accuracy.

However, topic approaches use a *bag-of-words* which, along with those that are feature-based (Tur et al., 2014), have difficulty dealing with unknown words, e.g., rarely used expressions and neologisms. To overcome this problem, Oh et al. (2018) describe a method in which OOD sentences occurring in a dialogue are detected based on sentence

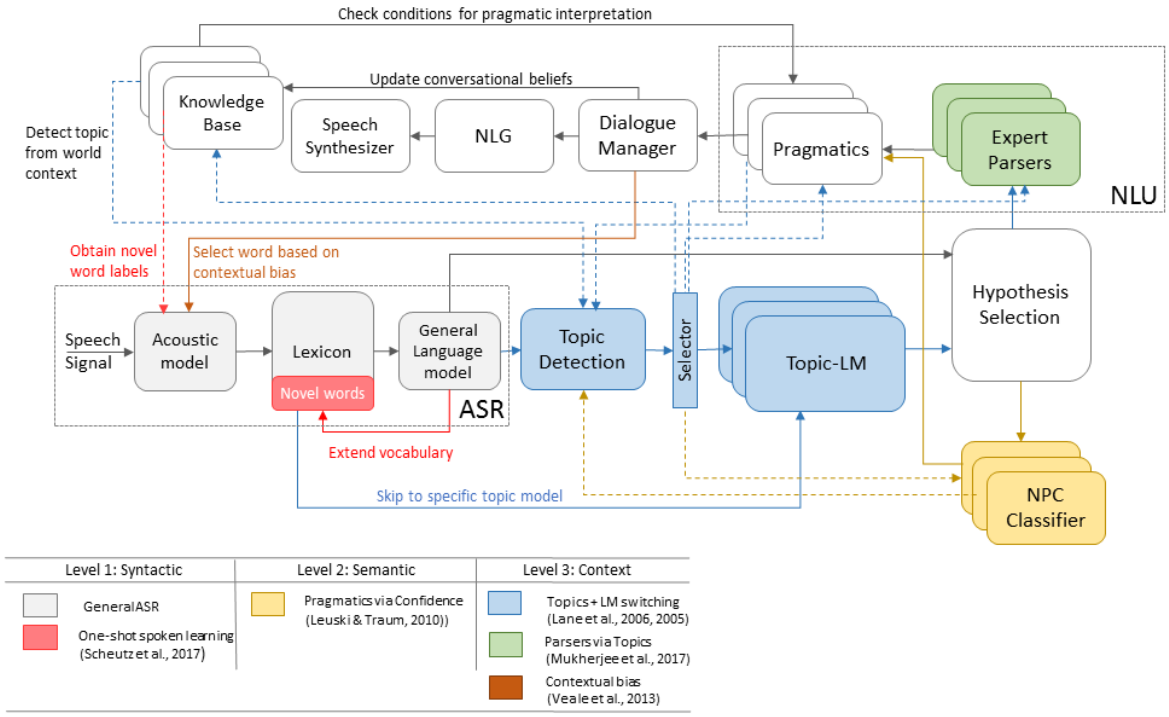


Figure 2: Multi-level framework for understanding spoken dialogue encompassing prior approaches.

distances. The distances are measured by sentence embedding vectors using RNN (Recurrent Neural Network) encoders and incorporate an attention mechanism.

Alternatively, Scheutz et al. (2017) describe a mechanism for detecting the intentional use of novel words in a *one-shot learning* system. Here, the ASR is modified such that when an unknown token is generated by the acoustic model, its corresponding word-level unit is discovered from the acoustic features. A nearest-neighbor classifier is used to determine whether the discovered unit represents the first member of a new word-class of the vocabulary and, if so, the class and example are added; otherwise it is added to an existing class.

In addition to using context to switch among language models, topic modeling can be applied to syntactic SDS components. Mukherjee et al. (2017) use Latent Dirichlet Allocation (LDA) to improve parser performance across multiple domains. LDA is used to find the topic structure in a document, which is a single sentence here. The sentence is assigned to the most likely topic and an "expert" parser for the topic is trained for syntactic analysis.

For situations where the domain is constrained, yet the user will be using natural language with its attendant disfluencies and irregularities, the ASR is likely to not recognize domain-specific commands. For this type of system, Leuski and Traum (2010) describe a statistical classification component which, in order to automate natural and flexible human-agent dialogue, identifies the best response to user input by estimating the closest possible match within its training set if a precise match cannot be found.

Finally, Chen et al. (2013) describe how information from multiple non-ASR components in their conversational spoken language translation system can be combined with strong baseline ASR error detector features and used to improve overall ASR error rate. The system contains built-in error detection modules that pinpoint regions in the input where the ASR is likely to fail, including a confidence estimator of the language translation (i.e., English-Iraq). In addition, the posterior word probabilities from a named entity detector is used to improve out-of-vocabulary word recognition.

In this section, we reviewed literature representing the main approaches to resolving OOD inputs and improving ASR performance. The contextual approaches (e.g., topic modeling, word co-occurrence, statistical classification) have the effect of changing the prior probabilities of the trained ASR by making a selection from multiple language models (Mukherjee et al., 2017; Lane et al., 2006, 2005; Sarma and Palmer, 2004), biasing the ASR

word hypothesis (Veale et al., 2013), or discriminating among similar interpretations (Leuski and Traum, 2010). An alternative approach is to extend the ASR vocabulary when a novel instance of a word class is detected as in Scheutz et al. (2017). Chen et al. (2013) used a combination of the two approaches. In the following section, we will discuss how these approaches have been integrated with some, but not all, of the components of the SDS framework.

3 Multi-level Framework

Our framework consists of three possible levels of utterance interpretation: syntactic, semantic, and context. The purpose of the first level is to generate an interpretation of the user's meaning using a *syntactic* analysis of the utterance. For example, an utterance such as "pick up the red box" can be easily parsed along strict syntactic rules, generating a semantic form for the utterance by means of these syntactic rules. However, a more complicated, non-grammatical utterance such as "the red box, pick that up maybe" may violate expected syntactic rule structure. Thus, we attempt to handle unexpected utterances with a second level of interpretation. This level bypasses any syntactic assessment entirely; it attempts to guess directly at the *semantics* of the user's intent by finding which group of semantically-linked utterances is most similar in word content to the target utterance. Here, a classifier is used to select the most likely interpretation based on the utterance's similarity to previous utterances which the classifier has been trained on. For example, "the red box, pick that up maybe" is assessed as most similar to "pick up the red box", and is assumed to have the same meaning (and therefore the same semantic form). Both of these levels return a confidence score to allow the Hypothesis component to select between the semantic form produced by the classifier and the semantic form produced by syntactic analysis (in this case, the semantic form may be in predicate logic, i.e., *pickUp(self,box(red))*). At the third level, the *context* of the utterance is analyzed and used to restart and inform connected components to reinterpret the utterance using, for example, a new language model, classifier, or parser.

In the legend of Figure 2, we situate selected prior work in the framework, assigning them to the syntactic, semantic, or context levels in accordance with their approach to improving SDS performance. We place Scheutz et al. (2017) "One-shot Spoken Learning" in the syntactic level as they assume an unrecognized speech token may be a novel word. After a pattern analysis of the acoustic features, their system attempts to place the new token in the vocabulary and interprets its grammar function within the syntactic context of the rest of the sentence. This flow is shown by the solid red connections in the figure. However, to recover the label of the word so that the agent's text-to-speech system can say it back, the phonemic sub-units within the word feature must be recovered and mapped to the pronunciation dictionary. The dotted red line indicates the required connection for this capability.

We situate the NLU system described by Leuski and Traum (2010) in the second, semantic level as it makes no assumption that the syntactic form is correct. Level 2 uses a statistical classifier to generate multiple similar interpretations of the utterance, selecting the one with the highest confidence score and sending it to the Pragmatics component for intention analysis; these connections are shown by the solid gold lines in the figure. The classifier component could ask for a back-off and reinterpretation of the utterance if the highest confidence score falls below a specific threshold; the connections for this additional capability are indicated by the dashed gold lines.

One particular benefit of this multi-level framework is that it allows for the combination of a rule-based approach to NLU (the parser) with a machine-learning approach to NLU (the classifier). Alone, a rule-based approach hardly ever reaches the broad accuracy of state-of-the-art systems that use machine learning. However, with lack of explainability being a key weakness of machine learning, the machine-learning approach by itself rarely offers any clear introspective explanation of how or why it arrived at its results, unlike a rule-based approach which can provide the rules and syntactic definitions behind the reasoning of any particular parse. With a trained classifier able to outperform the strict limitations of the rule-based parser by finding the closest possible expected sentence to the user input, and a parser able to provide syntactic explanation for any expected sentence, this multi-level architecture can surpass the weaknesses of both approaches.

The basis for the context level 3 is its ability to use topic detection to infer utterance context and thus we situate the hierarchical topic classification

of Lane et al. (2005) in that level. In their implementation, the researchers describe a system which can detect in- and out-of-domain utterances, and freely switch among several topic-dependent language models. In the figure, the connections and components for this system are shown in blue; however, we indicate by the blue dashed line that there could be additional connections that could further improve the interpretation. New connections from the Knowledge Base, Pragmatics, and Classifier components allow them to request a back-off and reinterpretation of the utterance by inferring its context through topic detection. Connections from the selector back to those components can signal that an alternative, topic-dependent classifier, pragmatics, KB model should be used.

We also situate in level 3, the system described in Mukherjee et al. (2017) which creates topic-specific datasets that are then used to train expert parsers. This system is shown in the green box in the figure without a solid line connection to the Selector because the researchers have evaluated the expert parsers individually and do not specify a method for freely selecting from among language models. The green dashed lines show the connections to Topic Detection and from the Selector to indicate this added capability.

Finally, we situate the dialogue contextual bias signal system described in Veale et al. (2013) in level 3. Rather than using a topic model to infer bias, the authors use the knowledge of common dialogue exchange patterns contained in the dialogue Manager to develop a bias signal for the ASR component (shown as a solid brown line in the figure). This is used to change the words' prior probabilities in the ASR, influencing word selection according to dialogue context. The authors describe this system as a biologically plausible cognitive model based on human perceptual decision making. As such, it provides an interesting avenue for further research into human-like ways to improve speech recognition.

3.1 Implementation

We implemented the multi-level framework shown in Figure 2 in the DIARC cognitive robotic architecture (Scheutz et al., 2019). The implementation consists of the ASR, Topic Detection, Selector, Topic-Dependent Language Models, Parsers, Classifier components, and Pragmatics. The ASR is based on the chain model developed using Kaldi for the ASpIRE Challenge and trained on Fisher English (Harper, 2015; Povey et al., 2011). For parsing, we used a symbolic, rule-based parser, and for the classification component, an implementation which is part of the NPCEditor platform (Leuski and Traum, 2010).

Figure 3 shows that the processing flow begins with a spoken utterance that is transcribed imperfectly by the ASR component. In a single-level system, the utterance follows the blue line to the syntactic parser, which attempts to parse the meaning into a semantic form and send it along the blue line to the pragmatics component for extraction and further processing. In the multi-level system, the utterance is sent both to the parser and classifier (brown lines), which then each send the pragmatics component their own interpretation of the utterance's meaning along with a confidence score. For the classifier, this confidence score rates the utterance's similarity to the closest utterance in the classifier's training set (though if this score is below a predefined confidence threshold (0.6) it rounds the score to 0). For the parser, the confidence score is either 1.0 to signify that it was able to find a valid parse, or 0, in cases where no parse was able to be found.

The pragmatics component determines the highest score between the parser and classifier and picks the respective interpretation for further processing by the rest of the dialogue system. If the parser fails and the classifier cannot find a semantic interpretation with $score > 0.6$, then the brown line on the "No" branch is followed to the Topic Detector. If the topic has changed, the Selector will switch to a generic language model that is a mixture of topic unigrams, and the utterance will be reinterpreted. If there was no topic change, the framework assumes there cannot be a valid interpretation, and will generate an appropriate response to the user. The generic LM is used so that the Topic Detector has a basic utterance to which it infers a topic distribution. The distribution is used to select a domain-specific LM, and the utterance is reinterpreted using the new LM.

3.2 Topic Detection and Model Training

Prior to evaluating the multi-level framework, we trained the Topic Detector and tested its ability to distinguish utterances drawn from one domain or the other as follows (see Section 4 for a discussion on the domains used). We used Latent Dirichlet

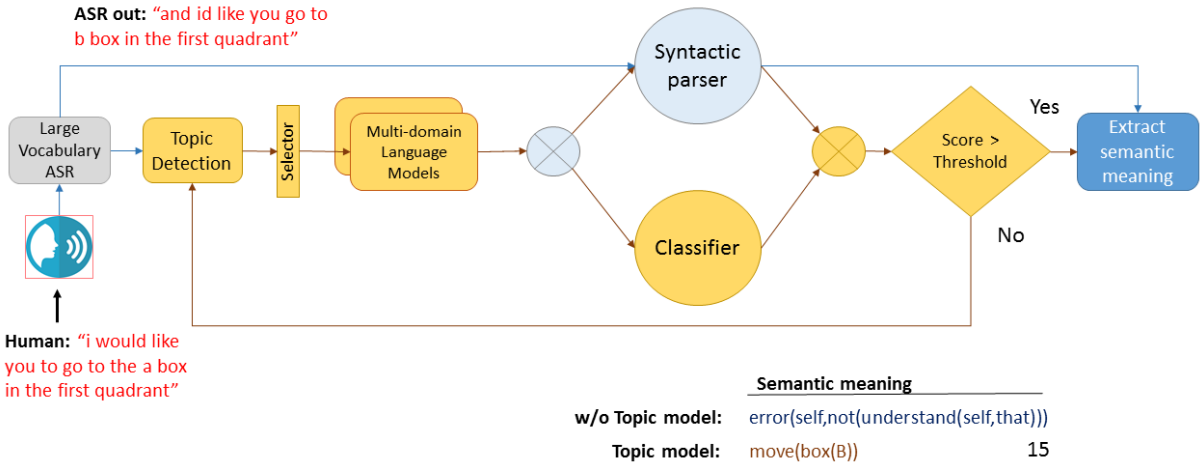


Figure 3: Multi-level processing flow. A single-level system assumes the utterance transcription received from the ASR is entirely accurate or has a single interpretation (blue lines). The multi-level system acknowledges ASR output may not be in exact expected form, and uses a classifier to find a close match. If the classifier fails, the topic is inferred and the language model and other components are switched to a new context (brown lines).

Allocation (Blei et al., 2003) to infer topics in a heterogeneous collection of textual data. We used the Gensim (Řehůřek and Sojka, 2010) implementation of LDA to train the Topic Detector and extract the topic distribution from the utterances. We used the default hyper-parameters and set the number of topics equal to the number of domains used in the training and test datasets, i.e., two.

We used the trained LDA model to create the Topic Detector. It detects a topic shift from one utterance to another by comparing the Kullback–Leibler (KL) divergence between the two topic distributions (Kullback and Leibler, 1951). If the difference is above a pre-determined threshold, a shift is indicated. This is used by the topic detection component to signal the Selector to switch to a new language model. As stated previously, this mechanism can be extended to select among alternative parsers, classifiers, knowledge base components, etc.

We used k-fold validation with $k = 5$ to train and test the Topic Detector's ability to distinguish utterances drawn from one domain or the other. Each fold contains utterances from both domains. To prepare the folds, we selected utterances from both domains, and placed them in 5 equally sized sets. We then trained and evaluated the topic detection model 5 times, selecting a different fold for evaluation every time and training on the other 4 folds. The training subsets consist of utterances from both domains.

The testing subset consists of at most 100 randomly paired utterances from the domains. Since the number of pairs increases exponentially with the size of the original list, this limit is imposed on the maximum number of sentence pairs to select. If a pair of sentences come from different files, then this is considered ground truth of a topic change, and vice versa. Sentence pairs are fed to the topic model, one after the other, and the predicted topic change is compared against the ground truth topic change. If the KL divergence of the two sentences is equal to or exceeds the input KL threshold (0.5), then this is considered a topic change, and otherwise not.

4 Framework Demonstration

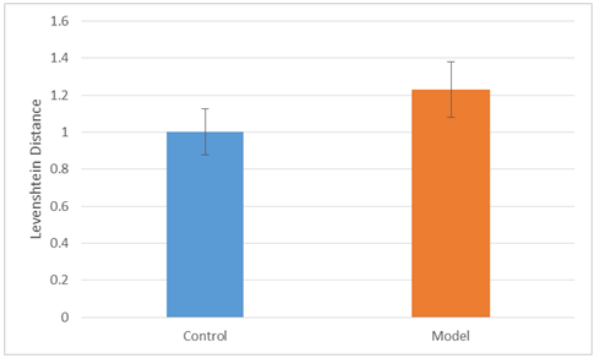
For model validation, we evaluated its accuracy in interpreting natural language utterances from different human-robot tasks. The goal was to compare our multi-level system to one which only used syntactic parsing. Using the data collected in other human-robot interaction experiments, we obtained two separate corpora of utterances used to instruct a robot in a specific task environment. For this implementation, we trained the topic detector on two topics. In the first domain, *SpaceStation*, participants ($n = 26$) gave commands in natural language to control several robots repairing components of a space station (Gervits et al., 2020). Out of 663 utterances, including duplicates, 363 unique utterances remained, out of which 50 were withheld for the test dataset and 313 were used to train the LDA model. In the second domain, *Diorama*, participants ($n = 33$) taught new skills to a robot learner

using natural language (Bennett et al., 2017). Out of 680 utterances, including duplicates, 525 unique utterances remained, out of which 50 were withheld for the test dataset and 475 were used to train the LDA model. The sentences from the SpaceStation domain comprised one document and those from the Diorama domain comprised another. The LDA model was trained on a union of the two document collections. We set up a pipeline for incoming utterances wherein each utterance would be processed in parallel by two different systems: our three-tier framework which included topic-switching and a bag-of-words classifier, and a control framework which possessed only the baseline tier of interpretation through syntactic parsing.

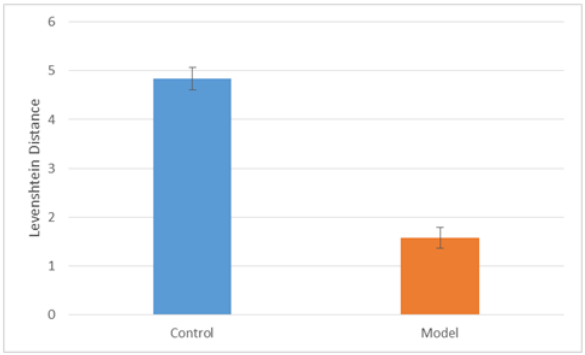
In both the control and multi-tier systems, parser rules were written by hand based on 100 utterances from each corpus, which were also used to train the ASR language models. For the control framework, these syntactic parsing rules were combined into one parser dictionary, whereas for the multi-tier system, the parser swaps between topic-specific dictionaries at the topic detector's signal. Similarly, the control system's language model was trained on the combined set of 200 utterances, while the multi-level system contained two separate models each trained on 100 utterances from the distinct tasks. For the multi-level system, two different classifiers were trained on the two different sets of utterance training data that had been hand-labeled with the correct semantic interpretation in predicate 1st-order logic form for each utterance. This pipeline was fed a test set of 50 utterances from each corpus (100 utterances total) that were withheld from the training data. Relevant output such as utterance transcription, semantic form, and in the multi-level framework's case, topic identification, were logged and manually annotated by the experimenter with the ground truth values of these variables: the correct transcription of each utterance, its intended interpretation in symbolic 1st-order logic form, and the task (topic) from which it originated.

4.1 Results

To investigate the differences in accuracy between the transcriptive and interpretive abilities of the multi-level and control frameworks, we compared the output of each framework per utterance to its respective ground truth value using a Levenshtein distance (LD) metric, which measures word dele-



(a)



(b)

Figure 4: (a) Mean token-based Levenshtein distance (LD) between each utterance and its ASR transcription. (b) LD between the ground truth meaning of an utterance in predicate form and the model interpretation.

tions, insertions, or substitutions. For transcription accuracy, we found the token-based LD between each utterance and the ASR transcription of that utterance[1]. In the multi-level framework, the $mean\ LD = 1.23$, and for the control, the $mean\ LD = 1.00$. We verified that the variances were homogeneous and then conducted a paired two-tailed t-test which showed no significant difference in transcription accuracy between frameworks ($t(99) = -1.707, p > .05$). Similarly, for interpretation accuracy, we found the LD between the ground truth meaning of an utterance in predicate form and the system's semantic interpretation of that utterance to measure our dependent variable: how close the predicate of the utterance interpretation was to the ground truth. For example, the ground truth utterance "pick up the red box" has a predicate form (semantic meaning) of *pickUp(self,box(red))*. In cases where the system was unable to come up with any interpretation, the LD defaulted to 6. In the multi-level framework, $mean\ LD = 1.57$ and for the control, $mean\ LD = 4.84$. We verified that the variances were homogeneous and then conducted a paired two-tailed t-test found a significant difference in these means ($t(99) = 12.425, p < .001$), demonstrating that the multi-level framework is significantly much more accurate than the control.

The control framework was generally only able to accurately identify the semantic form of the utterance in cases where the utterance matched exactly to the grammatical rules specified in the parser. In addition, it occasionally generated the correct semantic form in cases where it misrecognized an unexpected utterance as an expected utterance with the same meaning. For example, the utterance "hey robot one come fix this tube" was misrecognized as "fix the tube" by the control framework and thus correctly parsed with the meaning of *repairTube()*. In other cases of transcription inaccuracy, the control framework occasionally recognized portions of utterances from the wrong task corpus. For example, it recognized the utterance "robot one go to left four" as "robot one go left", thus interpreting the utterance as *move(left)*. While the diorama task has a need for this level of directability in the robot's movement, the space station does not, and the action *move(left)* is used exclusively in the diorama task. However, the control framework had no reason to suppose that this command was less likely to be uttered in this context, either at the speech recognition level or at the parsing level. In all other cases, the control framework was unable to parse the unexpected utterance, even if the ASR Component transcribed it completely accurately. In contrast, the multi-level framework, even if unable to get a wholly accurate transcription, was generally able to come up with a semantic form which, if not exact, was fairly close to the intended interpretation, on average only off by one or two arguments. The topic identifier in the multi-level framework also correctly identified the topic of the utterance 93% of the time.

[1]This was token-based instead of character-based because we did not wish to reward misrecognition of a shorter word over that of a longer word, i.e., "canned" and "can" vs "canned" and "tanned"

5 Discussion

We hypothesized that the multi-level framework would perform better in interpreting semantic meaning of utterances. The results show that there was a statistically significant difference between the semantic interpretation accuracy of the multi-level framework and the control framework, with semantic interpretation being more accurate in the multi-level framework (Figure 4b). In addition, there was no significant difference between the transcription accuracy of the multi-level framework and the accuracy of the control framework (Figure 4a). The control framework performed very well on expected utterances (speech that matched the syntactic structure of utterances from the training set), but very poorly on unexpected utterances. Its success at semantic interpretation was binary: either 100% or 0% certainty. In contrast, the multi-level framework was able to infer in uncertain situations due to the classifier, leading to greater overall interpretation success. Although executing a wrong command could be potentially worse than not understanding a command at all, having an uncertain estimate of what the user wants is better than no estimate. Rather than executing the command, further error recovery could begin based on information from other components. For example, if the agent's certainty regarding its interpretation is not above a certain threshold, its dialogue manager could initiate a confirmation or clarification request, or its knowledge database could be solicited for contextual information or dialogue history that might resolve uncertainty. This would be a direction to explore in future work.

We note that even in the control framework, the

LM was trained on a selective portion of data only containing the two topics. In contrast, the Aspire Chain Model default LM is used for general dialogue. Thus, there is not a substantial difference in the ASR word error recognition (WER) between the two. When run with the default, the control's WER is far worse. This will vary depending on how specific to the task the utterances are. For example, utterances like "go to left four" or "drive forward pushing box c" are transcribed as the irrelevant phrases "gonna last for" and "dry forward pushing box see", while "what are you doing right now" and "knock down the yellow tower" are recognized correctly. The success of the multi-level framework depends on whether or not the topic is identified correctly. If the topic is misidentified, all other output from the system will also be incorrect. Though a 93% success rate of topic identification appears good, the success rate might decrease as more topics are introduced. For this reason, focus should be placed on how other components of an autonomous dialogue system can be integrated into the process of topic identification, so that the burden is not solely placed on ASR. There are several examples of how this can be handled. For example, if the parser and classifier both fail to come up with an interpretation above some threshold of certainty, they could prompt the topic component to switch the topic to the second-place choice and attempt another pass at NLU, or additional information could be solicited from the system's knowledge base about the dialogue history, previous goals, goal status, or world state that may further assist with topic identification. The Dialogue Manager can also initiate clarification requests in case the above methods fail.

We used only classifier and parser confidence scores to evaluate interpretation, but other metrics might be considered. For example, the ASR's word-level transcription confidence could be used ahead of the interpretation components to signal an earlier switch to another language model. Chen et al. (2013) supplemented ASR confidence scores with additional metrics such as: LM perplexity, number of competing words, and acoustic model deviation from true scores. They also used acoustic-prosodic features for improved confidence in the ASR-hypothesized word boundary detection. This might have the greatest benefit for the framework when given OOD utterances. These utterances are often broken up into multiple in-domain words and thus, word insertions are frequent, making up about 40% of word errors.

6 Conclusion

Natural language interaction with SDSs can result in errors which propagate through the components, causing interpretation of utterance meaning to fail. We developed a multi-level framework in which utterance analysis is escalated according to feedback received at the syntactic, semantic, and topic level. We situated this framework in the context of prior research in improving speech recognition and natural language understanding and showed how they have been integrated with some, but not all, of the components of our framework. In a demonstration in which humans used natural language to initiate commands controlling robots in two separate domains, we showed how these approaches can be integrated with other SDS components. We found improved accuracy in the SDS' ability to interpret spoken task commands. By integrating multiple different potential routes for understanding into the dialogue system, we allow for better recovery across the system.

References

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

Maxwell Bennett, Tom Williams, Daria Thames, and Matthias Scheutz. 2017. Differences in interaction patterns and perception for teleoperated and autonomous humanoid robots. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6589–6594. IEEE.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Wei Chen, Sankaranarayanan Ananthakrishnan, Rohit Kumar, Rohit Prasad, and Prem Natarajan. 2013. ASR error detection in a conversational spoken language translation system. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7418–7422. IEEE.

Philip R Cohen and C Raymond Perrault. 1979. Elements of a plan-based theory of speech acts. *Cognitive science*, 3(3):177–212.

Renato De Mori. 1997. *Spoken dialogues with computers*. Academic Press, Inc.

Felix Gervits, Dean Thurston, Ravenna Thielstrom, Terry Fong, Quinn Pham, and Matthias Scheutz.

2020. Toward genuine robot teammates: Improving human-robot team performance using robot shared mental models. In *AAMAS '20: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems.

M. Harper. 2015. The automatic speech recogition in reverberant environments (ASpIRE) challenge. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 547–554.

Kristiina Jokinen and Michael McTear. 2009. Spoken dialogue systems. *Synthesis Lectures on Human Language Technologies*, 2(1):1–151.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Ian Lane, Tatsuya Kawahara, Tomoko Matsui, and Satoshi Nakamura. 2006. Out-of-domain utterance detection using classification confidences of multiple topics. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):150–161.

Ian R Lane, Tatsuya Kawahara, Tomoko Matsui, and Satoshi Nakamura. 2005. Dialogue speech recognition by combining hierarchical topic classification and language model switching. *IEICE transactions on information and systems*, 88(3):446–454.

Anton Leuski and David Traum. 2010. Practical language processing for virtual humans. In *Twenty-Second IAAI Conference*.

Michael F McTear. 1998. Modelling spoken dialogues with state transition diagrams: experiences with the CSLU toolkit. In *Fifth International Conference on Spoken Language Processing*.

Atreyee Mukherjee, Sandra Kübler, and Matthias Scheutz. 2017. Creating POS tagging and dependency parsing experts via topic modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 347–355.

Kyo-Joong Oh, DongKun Lee, Chanyong Park, Young-Seob Jeong, Sawook Hong, Sungtae Kwon, and Ho-Jin Choi. 2018. Out-of-domain detection method based on sentence distance for dialogue systems. In *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 673–676. IEEE.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

Anand S Rao and Michael P Georgeff. 1991. Modeling rational agents within a bdi-architecture. *KR*, 91:473–484.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Arup Sarma and David D Palmer. 2004. Context-based speech recognition error detection and correction. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 85–88. Association for Computational Linguistics.

Matthias Scheutz, Evan Krause, Brad Oosterveld, Tyler Frasca, and Robert Platt. 2017. Spoken instruction-based one-shot object and action learning in a cognitive robotic architecture. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*.

Matthias Scheutz, Thomas Williams, Evan Krause, Bradley Oosterveld, Vasanth Sarathy, and Tyler Frasca. 2019. An overview of the distributed integrated cognition affect and reflection DIARC architecture. In *Cognitive Architectures*, pages 165–193. Springer.

John R Searle and John Rogers Searle. 1969. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press.

Gokhan Tur, Anoop Deoras, and Dilek Hakkani-Tür. 2014. Detecting out-of-domain utterances addressed to a virtual personal assistant. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Richard Veale, Gordon Briggs, and Matthias Scheutz. 2013. Linking cognitive tokens to biological signals: Dialogue context improves neural speech recognizer performance. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35.

Graham Wilcock and Kristiina Jokinen. 2015. Multilingual wikitalk: Wikipedia-based talking robots that switch languages. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 162–164.

Wei Xu and Alexander Rudnicky. 2000. Task-based dialog management using an agenda. In *ANLP-NAACL 2000 Workshop: Conversational Systems*.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Victor W Zue and James R Glass. 2000. Conversational interfaces: Advances and challenges. *Proceedings of the IEEE*, 88(8):1166–1180.