Abstract Meaning Representation for Human-Robot Dialogue

Abstract Meaning Representation for Human-Robot Dialogue

Claire Bonial[1], Lucia Donatelli[2], Jessica Ervin[3], and Clare R. Voss[1]

[1]U.S Army Research Laboratory, Adelphi, MD 20783
[2]Georgetown University, Washington D.C. 20057
[3]University of Rochester, Rochester, NY 14627
claire.n.bonial.civ@mail.mil

Abstract

In this research, we begin to tackle the challenge of natural language understanding (NLU) in the context of the development of a robot dialogue system. We explore the adequacy of Abstract Meaning Representation (AMR) as a conduit for NLU. First, we consider the feasibility of using existing AMR parsers for automatically creating meaning representations for robot-directed transcribed speech data. We evaluate the quality of output of two parsers on this data against a manually annotated gold-standard data set. Second, we evaluate the semantic coverage and distinctions made in AMR overall: how well does it capture the meaning and distinctions needed in our collaborative human-robot dialogue domain? We find that AMR has gaps that align with linguistic information critical for effective human-robot collaboration in search and navigation tasks, and we present task-specific modifications to AMR to address the deficiencies.

1 Introduction

A central challenge in human-agent collaboration is that robots (or their virtual counterparts) do not have sufficient linguistic or world knowledge to communicate in a timely and effective manner with their human collaborators (Chai et al., 2017; She and Chai, 2017). We address this challenge in ongoing research directed at analyzing robot-directed communication in collaborative human-agent exploration tasks, with the ultimate goal of enabling robots to adapt to domain-specific language.

In this paper, we choose to adopt an intermediate semantic representation and select Abstract Meaning Representation (AMR) (Banarescu et al., 2013) in particular for three reasons: (i) the semantic representation framework abstracts away from surface variation, therefore the robot will only be trained to process and execute the actions corresponding to semantic elements of the representation (ii) there are a variety of fairly robust AMR parsers we can employ for this work, enabling us to forego manual annotation of substantial portions of our data and facilitating efficient automatic parsing in a future end-to-end system; and (iii) the structured representation facilitates the interpretation of novel instructions and grounding instructions with respect to the robot's current physical surroundings and set of executable actions. The latter motivation is especially important given that our human-robot dialogue is physically situated. This stands in contrast to many other dialogue systems, such as task-oriented chat bots, which do not require establishing and acting upon a shared understanding of the physical environment and often do not require any intermediate semantic representation (see §6 for further comparison to related work).

Our paper is structured as follows: First, we present background both on the corpus of human-robot dialogue we are leveraging (§2), and on AMR (§3). §4 discusses the implementation and results of two AMR parsers on the human-robot dialogue data. §5 assesses the semantic coverage of AMR for the human-robot dialogue data in particular. We then discuss related work that informs the current research in §6. Finally, §7 concludes and presents ideas for future work.

2 Background: Human-Robot Dialogue Corpus

We aim to support NLU within the broader context of ongoing research to develop a human-robot dialogue system (Marge et al., 2016a) to be used onboard a remotely located agent collaborating with humans in search and navigation tasks (e.g., disaster relief). In developing this dialogue system, we

are making use of portions of the corpus of human-robot dialogue data collected under this effort (Bonial et al., 2018; Traum et al., 2018).[1] This corpus was collected via a phased 'Wizard-of-Oz' (WoZ) methodology, in which human experimenters perform the dialogue and navigation capabilities of the robot during experimental trials, unbeknownst to participants interacting with the 'robot'(Marge et al., 2016b).

Specifically, a naïve participant (unaware of the wizards) is tasked with instructing a robot to navigate through a remote, unfamiliar house-like environment, and asked to find and count objects such as shoes and shovels. In reality, the participant is not speaking directly to a robot, but to an unseen Dialogue Manager (DM) Wizard who listens to the participant's spoken instructions and responds with text messages in a chat window or passes a simplified text version of the instructions to a Robot Navigator (RN) Wizard, who joysticks the robot to complete the instructions. Given that the DM acts as an intermediary passing communications between the participant and the RN, the dialogue takes place across multiple conversational floors. The flow of dialogue from participant to DM, DM to RN and subsequent feedback to the participant can be seen in table 1.

The corpus comprises 20 participants and about 20 hours of audio, with 3,573 participant utterances (continuous speech) totaling 18,336 words, as well as 13,550 words from DM-Wizard text messages. The corpus includes speech transcriptions from participants as well as the speech of the RN-Wizard. These transcriptions are time-aligned with the DM-Wizard text messages passed either to the participant or to the RN-Wizard.

The corpus also includes a dialogue annotation scheme specific to multi-floor dialogue that identifies initiator intent and signals relations between individual utterances pertaining to that intent (Traum et al., 2018).The design of the existing annotation scheme allows for the characterization of distinct information states by way of sets of participants, participant roles, turn-taking and floor-holding, and other factors (Traum and Larsson, 2003). *Transaction units (TU)* identify utterances from multiple participants and floors into units according to the realization of an initiator's intent, such that all utterances involved in an exchange surrounding the successful execution of a command are grouped and annotated for the relations they hold to one another. *Relation types (Rel)* signal how utterances within the same TU relate to one another in the context of the ultimate goal of the TU (e.g. "ack-done" in table 1, shortened from "acknowledge-done," signals that an utterance acknowledges completion of a previous utterance; for full details on Rel types, see Traum et al. (2018)). *Antecedents (Ant)* specify which utterance is related to which. An example of a TU may be seen in table 1. It is notable that the existing annotation scheme highlights *dialogue structure* and does not provide a markup of the semantic content of participant instructions.

[1]This corpus is still being collected and a public release is in preparation.

3 Background: Abstract Meaning Representation

The Abstract Meaning Representation (AMR) project (Banarescu et al., 2013) has created a manually annotated "semantics bank" of text drawn from a variety of genres. The AMR project annotations are completed on a sentence-by-sentence basis, where each sentence is represented by a rooted directed acyclic graph (DAG). For ease of creation and manipulation, annotators work with the PENMAN representation of the same information (Penman Natural Language Group, 1989). For example:

```
(w / want-01
   :ARG0 (d / dog)
   :ARG1 (p / pet-01
             :ARG0 (g / girl)
             :ARG1 d))
```

Figure 1: PENMAN notation of *The dog wants the girl to pet him.*

In neo-Davidsonian fashion (Davidson, 1969; Parsons, 1990), AMR introduces variables (or graph nodes) for entities, events, properties, and states. Leaves are labeled with concepts, so that `(d / dog)` refers to an instance (`d`) of the concept *dog*. Relations link entities, so that `(w / walk-01 :location (p/ park))` means the walking event (`w`) has the relation `location` to the entity, park (`p`). When an entity plays multiple roles in a sentence (e.g., `(d / dog)` above), AMR employs re-entrancy in graph notation (nodes with multiple parents) or variable re-use in PENMAN notation.

AMR concepts are either English words (`boy`), PropBank (Palmer et al., 2005) role-

	Left floor		Right Floor		Annotations		
#	**Participant**	**DM → Participant**	**DM → RN**	**RN**	**TU**	**Ant**	**Rel**
1	move forward 3 feet				1		
2		ok			1	1	ack-wilco
3			move forward 3 feet		1	1	trans-r
4				done	1	3	ack-done
5		I moved forward 3 feet			1	4	trans-l

Table 1: Example of a Transaction Unit (TU) in the existing corpus dialogue annotation, which contains an instruction initiated by the participant, its translation to a simplified form (DM to RN), and the execution of the instruction and acknowledgement of such by the RN. TU, Ant(ecedent), and Rel(ation type) are indicated in the right columns. (Traum et al., 2018)

sets (`want-01`), or special keywords indicating generic entity types: `date-entity`, `world-region`, `distance-quantity`, etc. In addition to the PropBank lexicon of role-sets, which associate argument numbers (ARG 0–6) with predicate-specific semantic roles (e.g., ARG0=*wanter* in ex. 1), AMR uses approximately 100 relations of its own (e.g., `:time`, `:age`, `:quantity`, `:destination`, etc.).

The representation captures *who is doing what to whom* like other semantic role labeling (SRL) schemes (e.g., PropBank (Palmer et al., 2005), FrameNet (Baker et al., 1998; Fillmore et al., 2003), VerbNet (Kipper et al., 2008)), but also represents other aspects of meaning outside of semantic role information, such as fine-grained quantity and unit information and parthood relations. Also distinguishing it from other SRL schemes, a goal of AMR is to capture core facets of meaning while abstracting away from idiosyncratic syntactic structures; thus, for example, *She adjusted the machine* and *She made an adjustment to the machine* share the same AMR. AMR has been widely used to support NLU, generation, and summarization (Liu et al., 2015; Pourdamghani et al., 2016), machine translation, question answering (Mitra and Baral, 2016), information extraction (Pan et al., 2015), and biomedical text mining (Garg et al., 2016; Rao et al., 2017; Wang et al., 2017)

4 Evaluating AMR parsers on Human-Robot Dialogue Data

To serve as a conduit for NLU in a dialogue system, the ideal semantic representation would have robust parsers, allowing the representation to be implemented efficiently on a large scale. There have been a variety of parsers developed for AMR; two parsers using very different approaches are explored in the sections to follow.

4.1 Parsers

To automatically create AMRs for the human-robot diaogue data, we used two off-the-shelf AMR parsers, JAMR[2] (Flanigan et al., 2014) and CAMR[3] (Wang et al., 2015). JAMR was one of the first AMR parsers and uses a two-part algorithm to first identify concepts and then to build the maximum spanning connected subgraph of those concepts, adding in the relations. CAMR, in contrast, starts by obtaining the dependency tree (in this case, using the Charniak parser[4] and Stanford CoreNLP toolkit (Manning et al., 2014)) and then uses their algorithm to apply a series of transformations to the dependency tree, ultimately transforming it into an AMR graph. One strength of CAMR is that the dependency parser is independent of the AMR creation, so a dependency parser that is trained on a larger data set, and therefore more accurate, can be used. Both JAMR and CAMR have algorithms that have learned probabilities from training data in order to execute their algorithms on novel sentences.

4.2 Gold Standard Data Set

In order to evaluate parser performance on our data set, we hand-annotated a subset of the participant's speech in the human-robot dialogue corpus to create a gold standard data set. We focus on only participant language because it is the natural language that the robot will ultimately need to process and act on. This selected subset comprises 10% of participant utterances from one phase of the corpus including 10 subjects. The resulting sample is 137 sentences, equally distributed across the 10 participants, who tend to have unique speech patterns. Three expert annotators familiar with both the human-robot dialogue data and AMR independently annotated this sample, ob-

[2]https://github.com/jflanigan/jamr
[3]https://github.com/c-amr/camr
[4]https://github.com/BLLIP/bllip-parser

taining inter-annotator agreement (IAA) scores of .82, .82, and .91 using the Smatch metric.[5]

After independent annotations, we collaboratively created the gold standard set. Notable choices made during this process include the treatment of "can you" utterances, re-entry of the subject in commands using motion verbs with independent arguments for the mover and thing-moved, and handling of disfluencies; each is described here.

In "can you" utterances, there is an ambiguity as to whether it is a genuine question of ability, or a polite request. This difference determines whether the sentence gets annotated with `possible-01` (used in AMR to convey both possibility and ability), or just as a command (figure 2). It also determines whether the robot should respond with a statement of ability, or perform the action requested. To resolve this ambiguity, we referred back to the full transcripts of the data, and inferred based on context. In our sample, only one of these utterances ("can you go that way") was deemed to be a genuine questions of ability, while the remaining 13 (e.g. "can you take a picture," "can you turn to your right") were treated as commands. Those that were commands were annotated with `:polite +`, in order to preserve what we believe to be the speaker's intention in using the modal "can."

```
(p / possible-01
    :ARG1 (p2 / picture-01
        :ARG0 (y / you))
    :polarity (a / amr-unknown))

(p / picture-01 :mode imperative
    :polite +
    :ARG0 (y / you))
```

Figure 2: Two different AMR parses for the utterance "can you take a picture," convey two distinct interpretations of the utterance: the top can be read as "Is it possible for you to take a picture?," the bottom as a command for a picturing event.

With commands like "move" or "turn," it is implied that the robot is the agent impelling motion and the thing being moved. Therefore, we used re-entry in those AMRs to infer the implied "you" as both the `:ARG0`, mover, and the `:ARG1`, thing-moved (figure 3). This is consistent with AMR's goal of capturing the meaning of an utterance, independent of syntax—all arguments that can be confidently inferred should be included in the AMR, even if implicit.[6]

```
(m / move-01 :mode imperative
    :ARG0 (y / you)
    :ARG1 y
    :direction (f / forward)
    :extent (d / distance-quantity
        :quant 3
        :unit (f2 / foot)))
```

Figure 3: Gold standard AMR for "move forward 3 feet," showing the inferred argument "you" as `:ARG0` (mover) and `:ARG1` (thing-moved).

Although the LDC AMR corpus[7] does not include speech data, AMR does offer guidance for disfluencies—dropping the disfluent portion of an utterance in favor of representing only the speaker's repair utterance.[8] We followed this general AMR practice and dropped disfluent speech for the surprisingly infrequent cases of disfluency in our gold standard sample.

4.3 Results & Error Analysis

Having created a gold standard sample of our data, we ran both JAMR and CAMR on the same sample and obtained the Smatch scores when compared to the gold standard. As seen in Table 2, CAMR performs better on both precision and recall, thus obtaining the higher F-score. However, compared to their self-reported F-scores (0.58 for JAMR and 0.63 for CAMR) on other corpora, both under-perform on the human-robot dialogue data.

	Precision	**Recall**	**F-score**
JAMR	0.27	0.44	0.33
CAMR	**0.33**	**0.51**	**0.40**

Table 2: Parser performance on human-robot dialogue data.

Of the errors present in the parser output, many come from improper handling of light verb construction, imperatives, inferred arguments, and requests phrased as "can you" questions. "Take a picture" is an example of a frequent light verb construction, in which the verb ("take") is not semantically the main predicating element of the sentence. The correct parse is shown in Figure 4,

[5] IAA Smatch scores on AMRs are generally between .7 and .8, depending on the complexity of the data (AMR development group communication, 2014).

[6] See "implicit roles" in AMR guidelines: https://github.com/amrisi/amr-guidelines/blob/master/amr.md

[7] https://catalog.ldc.upenn.edu/LDC2017T10

[8] https://www.isi.edu/~ulf/amr/lib/amr-dict.html#disfluency

followed by the AMR that both parsers consistently create. They both incorrectly place "take" as the top node (using the grasping, caused motion sense of the verb), when in reality it should be dropped from the AMR completely according to AMR practice for light verbs. Light verb constructions occur in 60 utterances in our sample, with 59 of those being some variation on "take a picture" or "take a photo."

```
(p / picture-01 :mode imperative
    :ARG0 (y / you))

(x1 / take-01
    :ARG1 (x3 / picture))
```

Figure 4: Gold standard AMR for "take a picture" (top), followed by parser output.

Another common error shown in figure 4 is the notation `:mode imperative`, used to indicate commands, which is present in 127 sentences within our sample. Despite the prevalence of this feature in our gold standard, it is never present in parser output.

Another problematic omission on the parsers' part is a lack of inferred arguments. As discussed earlier, commands like "move forward 3 feet" have an implied "you" in both the `:ARG0` and `:ARG1` positions. However, the parsers don't include this variable, instead only including concepts that are explicitly mentioned in the sentence (figure 5).

```
(x1 / move-01
    :direction (x2 / forward)
    :ARG1 (x4 / distance-quantity
        :unit (f / foot)
        :quant 3))
```

Figure 5: CAMR output for "move forward 3 feet." The distance is mistaken for the `:ARG1` thing-moved and the implied "you"/robot is omitted. Compare with figure 3.

A fourth error made by the parsers was on "can you" requests. These were consistently handled by both parsers as questions of ability, annotated using `possible-01`, even when (as was usually the case) the utterances were intended to be polite commands (parser output is shown in the top parse of figure 2).

4.4 Discussion

The poor performance of both parsers on the human-robot dialogue data is unsurprising given the significant differences between it and the data the parsers were trained on. For both parsers, training data came from the LDC AMR corpus, made up entirely of written text, mostly newswire.[9] In contrast, the human-robot dialogue data is transcribed from spoken utterances, taken from dialogue that is instructional and goal-oriented. Thus, running the parsers on this data set shows that the differences in domain have significant effects on the parsers, and give rise to the systematic errors described above.

Improvements to the parser output could be obtained even by just adding a few simple heuristics, due to the formulaic nature of our data. Of 137 sample sentences, 25 were "take a picture," so introducing a heuristic specific to that sentence would be a simple way to make several corrections. To obtain broader improvements, however, it's clear that it will be necessary to retrain the parsers on in-domain data. Given that retraining requires a corpus of hand-annotated data, this gives us an opportunity to examine the current features of AMR in relation to our collaborative human-robot dialogue domain, and to explore possible additions to the annotation scheme to ensure that all elements of meaning essential to our domain have coverage. The findings of this analysis are described in the sections to follow.

5 Evaluating Semantic Coverage & Distinctions of AMR

We assess the adequacy of AMR for its use in an NLU component of a human-robot dialogue system both on theoretical grounds and in light of the results and error analysis presented in §4. To our knowledge, our research is the first to employ AMR to capture the semantics of spoken language; existing corpora are otherwise text-based.[10] Here, we discuss the characteristics of our data relevant to semantic representation and highlight specific challenges and areas of interest that we hope to address with AMR (§5.1); explore how to leverage AMR for these purposes by identifying (§5.2) and remedying (§5.3) gaps in existing AMR; and conclude with discussion (§5.4).

5.1 Challenges of the Data

Our goal in introducing AMR is to bridge the gap between what is annotated currently as *dialogue*

[9] https://catalog.ldc.upenn.edu/LDC2017T10

[10] Data released at https://amr.isi.edu/download/amr-bank-struct-v1.6.txt (*Little Prince*) and LDC corpus (footnote 5).

structure (Traum et al., 2018), and the semantic content of utterances that comprise such dialogue (not included in the current scheme). This goal follows from the understanding that *dialogue acts* are composed of two primary components: (i) *semantic content*, identifying the entities, events, and propositions relevant to the dialogue; and (ii) *communicative function*, identifying the ways an addressee may use semantic content to update the information state (Bunt et al., 2012). The existing dialogue structure annotation scheme of Traum et al. (2018) distinguishes two primary levels of pragmatic meaning important to dialogue that our research aims to maintain. The first, *intentional structure* (Grosz and Sidner, 1986), is equivalent to a TU[11]: all utterances that explicate and address an initiator's intent. The second, *interactional structure*, captures how the information state of participants in the dialogue is updated as the TU is constructed (Traum et al., 2018). These two levels of meaning stand apart from the basic compositional meaning of their associated utterances.We seek to represent these pragmatic levels of meaning and link them to their respective semantic forms.

We also seek to represent the temporal, aspectual, and veridical/modal nature of the robot's actions. Human instructions must be interpreted for actions: the robot may respond to such instruction by asserting whether or not such an instruction is possible given the robot's capabilities and the surrounding physical environment, and the robot may also communicate whether it is in the process of completing or has completed the desired instruction. Such information is implicitly linked to the intentional and interactional structure of the dialogue. For example, the act of giving a command implies that an event (if it occurs) will happen in the future; the act of asserting that an event has occurred signals that the event is past.

The representation of space and specific parameters that contribute to the robot's understanding of how human language maps on to its physical environment is also of interest to our work here.[12] As its capabilities are presented to the participant, the 'robot' in this research is capable of performing low-level actions with specific endpoints or goals: for example, *"move five feet forward," "face the doorway on your left,"* and *"get closer to that orange object."* This robot cannot successfully perform instructions that have no clear endpoint:*"move forward"* and *"turn"* will trigger clarification requests for further specification. In our planned system, however, we would ultimately like to give the robot an instruction such as *"Robot, explore this space and tell me if anyone has been here recently."* If the robot can learn to decompose an instruction such as *explore* into smaller actions, as well as how to identify signs of previous inhabitants, such instructions may become feasible.

Finally, much of the semantic content of the data in our experiment must be situated within the dialogue context to be properly interpreted. A command of *"Do that again"* is ambiguous in terms of which action it refers back to. Similarly, a negative command such as *"No, go to the doorway on the left"* negates specific information contained in a previous command. Implicit in natural language input, as well, are subtle event-event temporal relations, such as *"Move forward and take a picture"* (sequential actions) and *"Turn around and take a picture every 45 degrees"* (simultaneous actions).

5.2 Gaps in AMR

We focus on the three elements of meaning crucial to human-robot dialogue and currently lacking in AMR: (i) speaker intended meaning (as differing from the compositional meaning of the speaker's utterances);[13] (ii) tense and aspect (and vericity, by association); and (iii) spatial parameters necessary for the robot to successfully execute instructions within its physical environment.

The goal of the AMR project is to represent *meaning*, but whether such meaning is purely semantic or also captures a speaker's intended meaning is not specified. Here, we attempt to strike a balance between capturing speaker intended meaning and overspecifying utterances. We do this to stay faithful to existing experimental dialogue annotation practices, and to enable the robot

[11] Transaction Unit is described in §2.

[12] Though this mapping is outside the scope of our work, we see AMR as contributing substantially richer semantic forms to the NL planning research in robotics of others, such as (Howard et al., 2014) that entails such *grounding*, the process of assigning physical meanings to NL expressions.

[13] Speaker vs. compositional meaning is arguably annotated inconsistently in the current AMR corpus; see https://www.isi.edu/~ulf/amr/lib/amr-dict.html#pragmatics for some specific guidelines, but note that the released annotations seem to differ from this guidance in places.

to generalize the connections between such intention and underlying semantic content.

To illustrate the need for adding tense and aspect information to existing AMRs, compare the following utterances: (i) *"move forward five feet"* (uttered by the human participant); (ii) *"moving forward five feet..."* (sent via text by the robot to signal initiation of instruction); and (iii) *"I moved forward five feet"* (sent by the robot upon completion of the action). Although distinctions between these three utterances are critical for our domain, AMR represents these three utterances the same way:[14]

```
(m / move-01
    :ARG1 (r / robot)
    :extent (d / distance-quantity
        :quant 5
        :unit (f / foot))
    :direction (f / forward))
```

Figure 6: Without tense and aspect representation, current AMR conflates commands to move forward and assertions of ongoing and/or completed motion.

Spatial parameters of actions are represented in the current AMR through the use of predicate-specific ARG numbers, as outlined in the PropBank rolesets, or with the use of AMR relations, such as `:path` and `:destination`. Whether or not a relation or argument number is used, and which argument number, is specific to a predicate and therefore inconsistent across motion relations. We aim to make the representation of these parameters more consistent and enrich them with information about which are required, which are optional, and which might have assumed, default interpretations.

5.3 Proposed Refinements

We leverage the existing dialogue annotations to extract intentional and interactional meaning and, where relevant, map these annotations to proposed refinements described below.

Speech Acts. We add a higher level of pragmatic meaning to the propositional content represented in the AMR through frames that correspond to speech acts (Austin, 1975; Searle, 1969). We use the model of vocatives in existing AMR as our guide.[15] We also make use of the existing dialogue structure annotation scheme for our corpus that identifies dialogue types similar to speech acts. Using this scheme as a starting point, we create 36 unique speech acts and corresponding AMR frames: this consists of six general dialogue types (*command, assert, request, question, evaluate, express*) with 5 to 12 subtypes each (e.g., move, send-image). An example of a speech act template for *command:move* may be seen in figure 7.[16] Notably, by adding a layer of speaker-intended meaning to the content of the proposition itself, we are able to capture participant roles within the speech act (`:ARG0` and `:ARG1` of `command-02`). Future work will be able to reference these roles and model how participant relations evolve over the course of the discourse (Allwood et al., 2000).

```
(c / command-02
    :ARG0-commander
    :ARG1-impelled agent
    :ARG2 (g / go-02 :completable +
        :ARG0-goer
        :ARG1-extent
        :ARG3-start point
        :ARG4-end point
        :path
        :direction
        :time (a / after
            :op1 (n / now))))
```

Figure 7: Speech act template for *command:move*. Arguments and relations in italics are filled in from context and the utterance.

Tense and Aspect. We also adopt the annotation scheme proposed by Donatelli et al. (2018) for augmenting AMR with tense and aspect. This scheme identifies the temporal nature of an event relative to the dialogue act in which it is expressed as past, present, or future. The aspectual nature of the event can be specified as atelic (`:ongoing -/+`), telic and hypothetical (`:ongoing -, :completable +`), telic and in progress (`:ongoing +, :complete -`), or telic and complete (`:ongoing -, :complete +`). A telic and hypothetical event representation can be seen in figure 7. This tense/aspect annotation scheme is specific to AMR and coarse-grained in nature, but through its use of existing AMR relations (`:time, before, after, now, :op`), it can be adapted to finer-grained temporal relations in future work.

Spatial Parameters. As seen in figure 7, all arguments of `go-02` as well as additional relations

[14] Utterance (i) would additionally annotate `(m / move)` with `:mode imperative` to signal a command. As noted in §4, parsers rarely capture this information.

[15] https://www.isi.edu/~ulf/amr/lib/amr-dict.html#vocative

[16] Annotation of tense and aspect, and the need for extra relations will be explained in continuation.

are made use of in the template. These positions correspond to parameters of the action, needed so the robot may carry out the action successfully. Having a template for each domain-relevant speech act will allow us to specify required and optional parameters for robot operationalization. In figure 7, this is either `:ARG1` or `:ARG4`, given that the robot currently needs an endpoint for each action; if both arguments are empty, this ought to trigger a request for further information by the robot. Note also that the same *command:move* AMR (including `go-02`) would be implemented for all realizations of commands for movement (e.g., *move, go, drive*), whereas these would receive distinct AMRs, headed by each individual predicate, under current practices.

5.4 Discussion

The refinements we present to the existing AMR aim to mimic a conservative learning process: we seek to provide just enough pragmatic meaning to assist robot understanding of natural language, but we do not provide specification to the point that there will be over-fitting in the form of one-to-one mappings of semantic content and pragmatic effect. As research continues and robot capabilities expand, we expect to augment the robot's linguistic knowledge based on general patterns in the current annotation scheme.

6 Related work

6.1 Semantic Representation

There is a long-standing tradition of research in semantic representation within NLP, AI, as well as theoretical linguistics and philosophy (see Schubert (2015) for an overview). In this body of research, there are a variety of options that could be used within dialogue systems for NLU. However, for many of these representations, there are no existing automatic parsers, limiting their feasibility for larger-scale implementation. A notable exception is combinatory categorical grammar (CCG) (Steedman and Baldridge, 2009); CCG parsers have already been incorporated in some current dialogue systems (Chai et al., 2014). Although promising, CCG parses closely mirror the input language, so systems making use of CCG parses still face the challenge of a great deal of linguistic variability that can be associated with a single intent. Again, in abstracting away from surface variation, AMR may offer more regular, consistent parses in comparison to CCG. Universal Conceptual Cognitive Annotation (UCCA) (Abend and Rappoport, 2013), which also abstracts away from syntactic idiosyncrasies, and its corresponding parser (Hershcovich et al., 2017) merits future investigation.

6.2 NLU in Dialogue Systems

Task-oriented spoken dialogue systems have been an active area of research since the early 1990s. Broadly, the architecture of such systems includes (i) automatic speech recognition (ASR) to recognize an utterance, (ii) an NLU component to identify the user's intent, and (iii) a dialogue manager to interact with the user and achieve the intended task (Bangalore et al., 2006). The meaning representation within such systems has, in the past, been predefined frames for particular subtasks (e.g., flight inquiry), with slots to be filled (e.g., destination city) (Issar and Ward, 1993). In such approaches, the meaning representation was crafted for a specific application, making generalizability to new domains difficult if not impossible. Current approaches still model NLU as a combination of intent and dialogue act classification and slot tagging, but many have begun to incorporate recurrent neural networks (RNNs) and some multi-task learning for both NLU and dialogue state tracking (Hakkani-Tür et al., 2016; Chen et al., 2016), the latter of which allows the system to take advantage of information from the discourse context to achieve improved NLU. Substantial challenges to these systems include working in domains with intents that have a large number of possible values for each slot and accommodation of out-of-vocabulary slot values (i.e. operating in a domain with a great deal of linguistic variability).

Thus, a primary challenge today and in the past is representing the meaning of an utterance in a form that can exploit the constraints of a particular domain but also remain portable across domains and robust despite linguistic variability. We see AMR as promising because the parsers are domain-independent (and can be retrained), and the representation itself is flexible enough for the addition of some domain-specific constraints. Furthermore, since AMR abstracts away from syntactic variability to represent only core elements of meaning, some of the variability in the input language can be "tamed," to give systems more

systematic input. With the proposed addition of speech acts to AMR described in §5.3, the augmented AMRs also facilitate dialogue state tracking.

Although human-robot dialogue systems often leverage a similar architecture to that of the spoken dialogue systems described above, human-robot dialogue introduces the challenge of physically situated dialogue and the necessity for symbol and action grounding, which generally incorporate computer vision. Few systems are tackling all of these challenges at this point (but see Chai et al. (2017)). A description of the preliminary human-robot dialogue system developed under the umbrella of this project, and where this research might fit into that system, is described in the next section.

7 Conclusions & Future Work

Overall, we find results to be mixed on the feasibility of AMR for NLU within a human-robot dialogue system. On one hand, AMR is attractive given that there are a variety of relatively robust parsers available for AMR, making implementation on a larger scale feasible. However, our evaluation of two parsers on the human-robot dialogue data demonstrates that retraining on domain-relevant data is necessary, and this will require a certain amount of manual annotation. Furthermore, our assessment of the distinctions made in AMR reveal gaps that must be addressed for effective use in collaborative human-robot search and navigation. Nonetheless, these AMR refinements are tractable and may also be valuable to the broader community.

Thus, we have several paths forward in ongoing and future work. First, we plan to use heuristics and manual corrections to CAMR parser output to create a larger in-domain training set following existing AMR guidelines. We plan to combine this training set with other AMRs from various human-agent dialogue data sets being annotated in parallel with this work. In addition to expanding the training set for dialogue, this will allow us to explore the extent to which our findings, with respect to AMR gaps, may also apply to other human-agent dialogue domains.

Second, we will consider how to implement AMR into the existing, preliminary dialogue system called 'Scout Bot,' which has been developed as part of our larger research project (Lukin et al., 2018). For NLU, Scout Bot makes use of the NPCEditor (Leuski and Traum, 2011), a statistical classifier that learns a mapping from inputs to outputs. Currently, the NPCEditor currently relies on string divergence measures to associate an instruction with either a text version to be sent forward to the RN-Wizard or a clarification question to be returned to the participant. However, some of the challenging cases we analyzed in §5.1 suggest that an intermediate semantic representation will be needed within the NLU phase. Specifically, because the instructions must be grounded within physical surroundings and with respect to an executable set of robot actions, a semantic representation provides the structure needed to interpret novel instructions as well as ground instructions in novel physical contexts. Error analysis has demonstrated that the current Scout Bot system, by simply learning an association between an input string and a particular set of executed actions, cannot generalize to unseen, novel input instructions (e.g, *"Turn left 100 degrees,"* as opposed to a more typical number of degrees like 90) and fails to interpret instructions with respect to the current physical surroundings (e.g., the destination of *"Move to the door on the left"* will be interpreted differently depending where the robot is facing). The structure of the semantic representation provided by AMR will allow the system to interpret 100 degrees as a novel extent of turning, and allow destination slots like *"door on the left"* to be grounded to a location in the current physical context with the help of the robot's sensors.

Thus, in future iterations of the dialogue system incorporating AMR, we will retrain or reformulate the NPCEditor to take the automatic AMR parses as input and output the in-domain AMR templates described in §5.3. The Dialogue Manager will act upon these templates with either a response/question to the participant or pass the domain-specific AMR along to be mapped to the behavior specification of the robot for execution. Specific steps on this research trajectory will include (i) development of graph to graph transformations to map parser output to the domain-refined AMRs and (ii) an assessment of how well the domain-refined AMRs map to a specific robot planning and behavior specification, which will facilitate determining what other refinements may be necessary to effectively bridge from natural language instructions to robot execution.

References

Omri Abend and Ari Rappoport. 2013. Universal Conceptual Cognitive Annotation (UCCA). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 228–238.

Jens Allwood, David Traum, and Kristiina Jokinen. 2000. Cooperation, dialogue and ethics. *International Journal of Human-Computer Studies*, 53(6):871–914.

John Langshaw Austin. 1975. *How to do things with words*, volume 88. Oxford University Press.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The Berkeley FrameNet project. In *Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

Srinivas Bangalore, Dilek Hakkani-Tür, and Gokhan Tur. 2006. Introduction to the special issue on spoken language understanding in conversational systems. *Speech Communication*, 3(48):233–238.

Claire Bonial, Stephanie Lukin, Ashley Foots, Cassidy Henry, Matthew Marge, Kimberly Pollard, Ron Artstein, David Traum, and Clare R. Voss. 2018. Human-robot dialogue and collaboration in search and navigation. In *Proceedings of the Annotation, Recognition and Evaluation of Actions (AREA) Workshop at LREC, 2018*.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David R Traum. 2012. Iso 24617-2: A semantically-based standard for dialogue annotation. In *LREC*, pages 430–437. Citeseer.

Joyce Y. Chai, Rui Fang, Changsong Liu, and Lanbo She. 2017. Collaborative Language Grounding Toward Situated Human-Robot Dialogue. *AI Magazine*, 37(4):32.

Joyce Y Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Littley, Changsong Liu, and Kenneth Hanson. 2014. Collaborative effort towards common ground in situated human-robot dialogue. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 33–40. ACM.

Yun-Nung Chen, Dilek Hakkani-Tür, Gökhan Tür, Jianfeng Gao, and Li Deng. 2016. End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding. In *INTERSPEECH*, pages 3245–3249.

Donald Davidson. 1969. The individuation of events. In *Essays in honor of Carl G. Hempel*, pages 216–234. Springer.

Lucia Donatelli, Michael Regan, William Croft, and Nathan Schneider. 2018. Annotation of tense and aspect semantics for sentential AMR. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions*, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Charles J Fillmore, Christopher R Johnson, and Miriam RL Petruck. 2003. Background to FrameNet. *International journal of lexicography*, 16(3):235–250.

Jeffrey Flanigan, Sam Thomson, Jaime Carbonell, Chris Dyer, and Noah A Smith. 2014. A discriminative graph-based parser for the Abstract Meaning Representation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1426–1436.

Sahil Garg, Aram Galstyan, Ulf Hermjakob, and Daniel Marcu. 2016. Extracting biomolecular interactions using semantic parsing of biomedical text. In *Proc. of AAAI*, Phoenix, Arizona, USA.

Barbara J Grosz and Candace L Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204.

Dilek Hakkani-Tür, Gökhan Tür, Asli Celikyilmaz, Yun-Nung Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang. 2016. Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. In *Interspeech*, pages 715–719.

Daniel Hershcovich, Omri Abend, and Ari Rappoport. 2017. A transition-based directed acyclic graph parser for UCCA. *arXiv preprint arXiv:1704.00552*.

Thomas Howard, Stephanie Tellex, and Nicholas Roy. 2014. A natural language planner interface for mobile manipulators. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2014)*.

Sunil Issar and Wayne Ward. 1993. Cmlps robust spoken language understanding system. In *Third European Conference on Speech Communication and Technology*.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of English verbs. *Language Resources and Evaluation*, 42(1):21–40.

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

Fei Liu, Jeffrey Flanigan, Sam Thomson, Norman Sadeh, and Noah A Smith. 2015. Toward abstractive summarization using semantic representations. In *Proc. of NAACL*.

Stephanie M Lukin, Felix Gervits, Cory J Hayes, Anton Leuski, Pooja Moolchandani, John G Rogers III, Carlos Sanchez Amaro, Matthew Marge, Clare R Voss, and David Traum. 2018. Scoutbot: A dialogue system for collaborative navigation. *arXiv preprint arXiv:1807.08074*.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Matthew Marge, Claire Bonial, Brendan Byrne, Taylor Cassidy, A. William Evans, Susan G. Hill, and Clare Voss. 2016a. Applying the Wizard-of-Oz Technique to Multimodal Human-Robot Dialogue. In *Proc. of RO-MAN*.

Matthew Marge, Claire Bonial, Kimberly A Pollard, Ron Artstein, Brendan Byrne, Susan G Hill, Clare Voss, and David Traum. 2016b. Assessing Agreement in Human-Robot Dialogue Strategies: A Tale of Two Wizards. In *International Conference on Intelligent Virtual Agents*, pages 484–488. Springer.

Arindam Mitra and Chitta Baral. 2016. Addressing a question answering challenge by combining statistical methods with inductive rule learning and reasoning. In *Proc. of AAAI*, pages 2779–2785.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Xiaoman Pan, Taylor Cassidy, Ulf Hermjakob, Heng Ji, and Kevin Knight. 2015. Unsupervised entity linking with Abstract Meaning Representation. In *Proc. of HLT-NAACL*, pages 1130–1139.

Terence Parsons. 1990. *Events in the Semantics of English*, volume 5. MIT Press, Cambridge, MA.

Penman Natural Language Group. 1989. The Penman user guide. *Technical report, Information Sciences Institute*.

Nima Pourdamghani, Kevin Knight, and Ulf Hermjakob. 2016. Generating English from Abstract Meaning Representations. In *Proc. of INLG*, pages 21–25.

Sudha Rao, Daniel Marcu, Kevin Knight, and Hal Daumé III. 2017. Biomedical event extraction using Abstract Meaning Representation. In *Proc. of BioNLP*, pages 126–135, Vancouver, Canada.

Lenhart K Schubert. 2015. Semantic representation. In *AAAI*, pages 4132–4139.

John Rogers Searle. 1969. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge University Press.

Lanbo She and Joyce Chai. 2017. Interactive Learning of Grounded Verb Semantics towards Human-Robot Communication. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1634–1644, Vancouver, Canada. Association for Computational Linguistics.

Mark Steedman and Jason Baldridge. 2009. Combinatory categorial grammar. nontransformational syntax: A guide to current models. *Blackwell, Oxford*, 9:13–67.

David Traum, Cassidy Henry, Stephanie Lukin, Ron Artstein, Felix Gervits, Kimberly Pollard, Claire Bonial, Su Lei, Clare Voss, Matthew Marge, Cory Hayes, and Susan Hill. 2018. Dialogue Structure Annotation for Multi-Floor Interaction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

David R Traum and Staffan Larsson. 2003. The information state approach to dialogue management. In *Current and new directions in discourse and dialogue*, pages 325–353. Springer.

Chuan Wang, Nianwen Xue, and Sameer Pradhan. 2015. A transition-based algorithm for AMR parsing. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 366–375.

Yanshan Wang, Sijia Liu, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Fei Liu, and Hongfang Liu. 2017. Dependency and AMR embeddings for drug-drug interaction extraction from biomedical literature. In *Proc. of ACM-BCB*, pages 36–43, New York, NY, USA.