Software architectures for incremental understanding of human speech

Gregory Aist[*,1,3], *James Allen*[1,3], *Ellen Campana*[1,2], *Lucian Galescu*[3],
Carlos A. Gómez Gallo[1], *Scott C. Stoness*[1], *Mary Swift*[1], *Michael Tanenhaus*[2]

[1]Computer Science Department, University of Rochester, USA
[2]Brain and Cognitive Sciences Department, University of Rochester, USA
[3]Institute for Human and Machine Cognition, Pensacola, Florida, USA

*Contact address: gsa@gregoryaist.com

Abstract

The prevalent state of the art in spoken language understanding by spoken dialog systems is both modular and whole-utterance. It is modular in that incoming utterances are processed by independent components that handle different aspects, such as acoustics, syntax, semantics, and intention / goal recognition. It is whole-utterance in that each component completes its work for an entire utterance prior to handing off the utterance to the next component. However, a growing body of evidence suggests that humans do not process language that way. Rather, people process speech by rapidly integrating constraints from multiple sources of knowledge and multiple linguistic levels incrementally, as the utterance unfolds. In this paper we describe ongoing work aimed at developing an architecture that will allow machines to understand spoken language in a similar way. This revolutionary approach is promising for two reasons: 1) it more accurately reflects contemporary models of human language understanding, and 2) it results in empirical improvements including increased parsing performance.
Index Terms: dialogue systems, speech understanding, psycholinguistics, parsing, incremental understanding.

1. Introduction

Computational Natural Language Understanding (NLU), after decades of research, remains one of the many areas of Artificial Intelligence that is easy for people yet profoundly difficult for computers. The major reason that language understanding is so difficult for computers to understand is that ambiguity is rampant; each input is locally consistent with multiple interpretations, and each of those interpretations, in turn, is locally consistent with a number of potential inputs. This ambiguity occurs simultaneously at all levels of processing. For instance, the speech signal is locally consistent with multiple word sequences and each such word sequence is locally consistent with multiple speech inputs. Likewise, each word sequence is locally consistent with multiple syntactic structures, each of which is locally consistent with other possible word sequences. In order to cope with this tremendous complexity in real-time, some simplifying assumptions are needed to subdivide NLU into smaller, more tractable, sub-problems, and to constrain how each of those sub-problems might be solved. Spoken dialogue system researchers tend to assume:

Standard Simplifying Assumption 1: Speech and linguistic information can be treated as independent of other inputs and knowledge sources.

Standard Simplifying Assumption 2: Speech and language processing can be divided into a small number of levels. Each level depends only on the final utterance-level output of the previous level.

These simplifying assumptions were once consistent with linguistic and psycholinguistic models of how humans understand language. However, a growing body of evidence suggests that humans process spoken language incrementally. New models have been developed to account for this data, and the common simplifying assumptions outlined above are not consistent with these models. In the next section we briefly overview the data and models of human language processing.

1.1. Incremental Human Language Understanding

In recent years, psycholinguists have begun to use more fine-grained tools and metrics to investigate language. This change has made it possible for researchers to investigate spoken language in more or less natural contexts. This body of research has demonstrated that as an utterance unfolds, listeners take advantage of both linguistic and extra-linguistic information to arrive at interpretations more quickly than they could with language alone. For instance, listeners have been shown to use visual information about the scene (Tanenhaus et al., 1995 & 2000), the goals and perspectives of their partners (Hanna & Tanenhaus, 2003), and spatial / embodied constraints about how objects in the world can be manipulated (Chambers et al., 2004.) during language understanding to restrict the set of potential interpretations that are explored. Similarly, information from different levels of processing such as phonology, syntax, semantics and discourse / reference can be combined by listeners to constrain the set of potential interpretations that are explored (Altmann & Kamide, 1999).

1.2. Incremental Computer Language Understanding

The previous section described current models of how humans process spoken language – incrementally, rapidly integrating information from multiple sources and levels to arrive at partial / local interpretations. We aim to develop an architecture that will allow machines to process spoken language in a similar way. Where possible, we will leverage existing technologies and components. Thus, we propose replacing the standard simplifying assumptions with these, which are more consistent with incremental models of human language understanding:

Proposed Simplifying Assumption 1: Speech and linguistic information can be treated as independent of other inputs and knowledge sources – except that non-linguistic knowledge and information can be used online as it becomes available to improve search during speech / linguistic processing.

Proposed Simplifying Assumption 2: Speech / Language processing can be divided into a small number of levels that operate on partial information in parallel. Each level can be treated as independent – except that dynamically updated outputs of other levels can be used to improve search.

We have implemented a system based on the TRIPS architecture (Allen et al. 2001), as modified to make use of the proposed simplifying assumptions. In the remainder of this paper we describe the architecture in detail, and demonstrate that the new incremental architecture provides not only theoretical advantages, but empirical advantages too.

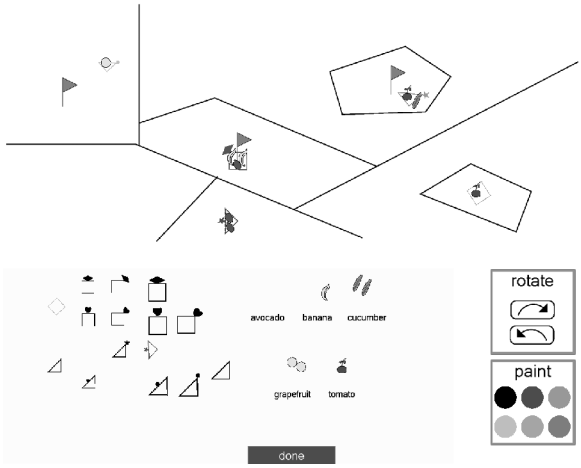

Figure 1 Fruit carts domain – example screen.

2. Human-human conversation (in testbed)

For this paper, we used the Fruit Carts domain, a testbed we developed to explore issues of incremental understanding. Subjects are given a map showing a number of shapes placed on the map, with varying colors, locations, angles, and contents. Their task is to describe how to replicate this map, giving instructions either to another person (for human-human dialog) or to a computer (for human-computer dialog). The main screen is shown in Figure 1. (The subjects have access to a "key" which has names for the regions.) Possible actions include selecting a shape, moving it to a region, painting it, turning it, and filling it with other objects (the fruit).

We used human-human conversations collected in this domain (Aist et al 2005) to form the basis for formalizing various aspects of incremental understanding, and for gauging the behavior of the spoken dialog system that we built to operate in this domain. To evaluate parser performance in incremental understanding mode compared to standard utterance by utterance interpretation we developed a gold standard corpus of parsed output for a sample dialogue. For each utterance the gold standard includes complete and correct syntactic analysis, word sense disambiguation, semantic role assignment and surface speech act analysis, as well as timing results and number of constituents produced during the parse. The high level of ambiguity in this domain often presents the parser with multiple possible interpretations, and the correct one is not always the first choice of the parser in standard mode. We have also developed a parsed corpus based on transcripts from experimental sessions to use as training data for new system components such as the VP advisor.

3. Incremental dialog system

We now describe TRIPS - a state-of-the-art platform that has been used to study a variety of domains such as emergency management, equipment purchasing, and learning-from-instruction - as redesigned for incremental understanding.

We have in this system components that are required for a dialogue system: speech input, parsing, and so on. The system presented here differes from conventional architectures such as Galaxy in two ways. First, as a TRIPS-based system, it uses general-purpose components with domain-specific models. For example, it uses a general reasoner (the Behavioral Agent) rather than a domain-specific dialogue manager. Likewise, a general process invokes domain specific models to construct interpretations from the parser output (Dzikovska, Allen & Swift, 2003). Second, as an incremental system, components process language as it arrives and send messages to other components when new information is calculated. For example, the Segmenter passes along advice about pragmatically relevant fragments as they are detected in the incoming speech.

3.1. Speech recognition and segmentation

We used the human-human conversations described above to construct a specialized statistical language model for this domain using techniques as in Galescu, Ringger, and Allen (1998). This language model was used as one of the inputs to the speech recognizer - Sphinx 2, 3, or 4 depending on the system configuration (Lamere et al. 2003). The results from the speech recognizer are fed to the parser and to a separate segmentation module (the Segmenter) which uses a small top-down fragment grammar to incrementally make predictions about the presence of interaction-relevant fragments such as verb phrase prefixes ("we need to move") and referring expressions ("a large triangle"). The Segmenter passes its advice on to the Parser and also (for referring expressions) to the GUI, so as to allow the highlighting of possible referents.

3.2. Parser and Interpretation Manager

From speech recognizer output, the parser produces detailed semantic representations for analysis by the system reasoners. The incremental parser is different from its nonincremental counterpart in that not only does it builds arcs as words arrive, it also takes advice from other components such as the VP-advisor and the real-world KB as described below. Noun phrases are judged with respect to contextual reference resolution, and verb phrases by the likelihood of the argument structure co-occuring with the head verb in the domain; the incremental feedback from such advisors is used in the parser's search, biasing it towards globally more likely hypotheses. The parser also passes constituents on to other components for real-time feedback such as reference resolution (Stoness et al. 2005). Constituents are passed to the intention manager as they are constructed, which in turn consults other components for feedback. When probalistic judgments are received by the parser, constituent scores on the chart are modified to reflect these judgements.

The Interpretation Manager (IM) takes the analysis from the parser and refines the semantic interpretations. The IM also mediates between the Parser and advice agents such as the VP Advisor and the Simulator/KB, as described below.

3.3. Behavioral Agent; Output Components

The Behavioral Agent produces decisions about what to do, based on input from the Interpretation Manager. Each decision is passed onwards (to components such as the Simulator, Sequencer, and GUI) and results in actions such as highlighting an object or moving it to a new location.

4. Results: VP Advisor

Even though the Fruit Carts domain allows users to use free style language, the set of actions that can be performed on objects provide us with well defined constructions we can exploit. Table 1 summarizes the actions with all their possible thematic roles expressed in the data at one time or another.

Table 1. *Actions and their prototypical arguments*.

Action	arguments
Select	Verb-object
Rotate	Verb-object-angle-heading
Move	Verb-object-distance-heading-location
Paint	Verb-object-color

Due to common elliptical constructions in speech dialogue (Fernandez, Ginzburg, and Lappin 2004), examples of all cases where there was a missing verb or any verb argument were seen. Nevertheless, certain constructions were more likely than others, knowledge which might help the parser arrive at a more accurate analysis with less effort.

To this end an initial set of six dialogues were manually annotated with verb and verb argument type labels. Then statistics that measured how often a verb argument appears given the verb were collected. Table 2 is an example for the statistics found for the action MOVE. For example the most likely MOVE action is performed by giving the verb, object and location which is intuitively correct. However this only occurs 66% of the time; MOVE actions are also done by stating a location only. The object is presumably already in the context by a previous SELECT action. This is the case of object elision.

Table 2. *Statistics for MOVE action*.

args	Probs
-ver-obj-loc	0.658
-ver-obj-hea	0.109
-ver-obj	0.061
-ver-obj-dis	0.049
-ver-loc	0.037
-ver	0.037
-ver-obj-dis-hea	0.036

The mechanism works as follows: when the parser is constructing a VP, it asks the VP advisor how likely the construction under consideration is in this domain. This advice is taking place after the logical form of the utterance has been translated into our domain specific semantics. Therefore we can think of the advice as a way to encode semantic restrictions for each verb. The parser then modifies the probability of the constituent in the chart and puts it back into the agenda.

Experimental results show us that on average the number of constituents built by the parser decreases with the VP advice. The best result can be seen on sentences as complicated as the following: “take the box in morningside and put it into pine tree mountain on the bottom of the flag”; here, the number of constituents was decreased by as much as 19%. On less complex sentences such as "and then change it to brown" there is no difference in number of constituents since the standard parser already finds a spanning parse efficiently.

5. Results: Simulator / Real-World KB

Knowledge about the current state of the world is key for undertanding. Consider "put the square near the flag in the park". It is inherently ambiguous between an interpretation where there is a flag in a park, and one in which there is a square in the proximity of a flag. Any parser without access to real world knowledge must allow both, or choose at random.

In a standard dialogue system pragmatics does not come into play until the parse has been selected. In the incremental system, however, pragmatic feedback is available during parsing (Figure 2; the bold arrow shows the feedback element.) The KB Advisor receives all referring noun phrases from the parser via the intention manager, and reports a judgment on whether or not they refer to something in the current state of the world. As an example, for the sentence above the standard parser chose "the square" as the direct object of the verb, building 197 constituents during the parse. The incremental parser, using a knowledge base with one selected flag but no squares near flags, arrived at the same interpretation after only 121 constituents, an efficiency improvement of almost 40%.

Operating in incremental mode doesn’t just improve the efficiency of the parser, but its accuracy as well. If, for example, the world KB features a square near a flag, but no flag in a park, and has no square selected, the favored interpretation would be the one in which “the square near the flag” is the direct object. The non-incremental parser cannot make this distinction, even in principle, and so to capture the multiple possible interpretations, each preferable in a different context, it is necessary for the parser to feed forward a number of complete parses at the completion of its processing. An incremental understanding parser, however, has at its disposal, incrementally and immediately, the same knowledge that would be used to disambiguate the complete parses in a non-incremental system. By using the real-world knowledge base and allowing the reference feedback to be incorporated into the parse, the incremental system finds the correct parse as its most likely candidate, while building only 131 constituents.

We have also run the system on the transcript of a complete dialogue from the corpus that we collected for this domain. Candidate NPs are sent forward through the Interpretation Manager to the Knowledge Base, which provided feedback on whether the NP was a reasonable candidate, taking into account both domain-specific knowledge and the current state of the world. Because the user’s utterances had to be interpreted relative to the state of the world that the user had been aware of during dialogue collection, a series of knowledge base updates were performed between sentences to ensure that the KB was an accurate reflection of what the user had seen. Overall, the incremental understanding parser only had to build 75% as many constituents as the standard parser in order to find its first complete parse of each utterance from the dialogue in Figure 3, using transcripts as input.

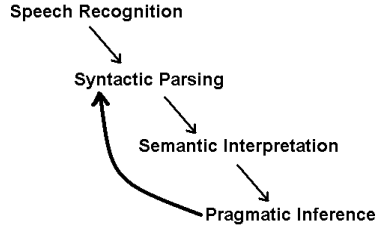

Figure 2 Pragmatic feedback affects incremental parsing.

1 okay so
2 we're going to put a large triangle with nothing into morningside
3 we're going to make it blue
4 and rotate it to the left forty five degrees
5 take one tomato and put it in the center of that triangle
6 take two avocados and put it in the bottom of that triangle
7 and move that entire set a little bit to the left and down
8 mmkay
9 now take a small square with a heart on the corner
10 put it onto the flag area in central park
11 rotate it a little more than forty five degrees to the left
12 now make it brown
13 and put a tomato in the center of it
14 yeah that's good
15 and we'll take a square with a diamond on the corner
16 small
17 put it in oceanview terrace
18 rotate it to the right forty five degrees
19 make it orange
20 take two grapefruit and put them inside that square
21 now take a triangle with the star in the center
22 small
23 put it in oceanview just to the left of oceanview terrace
24 and rotate it left ninety degrees
25 okay
26 and put two cucumbers in that triangle
27 and make the color of the triangle purple

Figure 3 Fruit carts domain – example dialogue.

6. Related and Future Work; Conclusion

Higashinaka et al. (2002) describe work on a process they term Incremental Sentence Sequence Search (ISSS), where both sentences and sentence fragments are used to update the dialog state. ISSS constructs multiple dialog states which can be decided upon as needed after any desired interval of speech. In a sense this can be viewed as a late binding process, whereas our work generally takes an earlier binding approach where information is brought to bear on the search as soon as possible. (In principle either system could no doubt be configured to perform late binding or early binding as desired.)

Rosé et al. (2002) describe a reworking of a chart parser so that "as the text is progressively revised, only minimal changes are made to the chart". They found that incrementally parsing incoming text allows for the parsing time to be folded into the time it takes to type, which can be substantial especially for longer user responses. Our current work operates on spoken input as well as typed input and makes extensive use of the visual context and of pragmatic constraints during parsing.

DeVault and Stone (2003) describe techniques for incremental interpretation that involve annotating edges in a parser's chart with constraints of various types that must be met for to the edge to be valid. That has a clean and nice simplicity to it, but seems to impose uniformity on the sorts of information and reasoning that can be applied to the parsing process. In our approach, advice to the parser is represented as modifications to the chart, and can thus be in any framework best for the source.

In terms of future directions: Here we've evaluated the parser with number-of-constituents; we would like to look at elapsed-time as well. Here also, while the system as a whole runs on speech input, we've evaluated these components on transcripts. We are working on methods for robust interpretation, such as fragment recombination, and would like to include that in future evaluations. It is promising that in preliminary work with word-mesh input for the dialogue in Figure 3, the incremental parser built 53% as many constituents.

In conclusion, we have presented a system architecture for incremental understanding of human speech. In addition, we have demonstrated empirical improvements that arise from the incremental understanding process, due to improvements in the search process by early use of such knowledge as verb phrase likelihood (see section 4; compare proposed assumption 2) and the visual world (see section 5; cf. proposed assumption 1). Incremental understanding is proving to be an exciting and productive area for spoken language in humans and machines.

7. References

[1] Aist, G.S., Campana, E., Allen, J., Rotondo, M., Swift, M., and Tanenhaus, M. Variations along the contextual continuum in task-oriented speech. CogSci 2005.

[2] Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L. and Stent, A. 2001. Towards conversational human-computer interaction. *AI Magazine* **22**(4): 27-38.

[3] Altmann, G.T.M., and Kamide, Y. 1999. Incremental interpretation at verbs: resticting the domain of of subsequent reference. *Cognition* **73**: 247-264.

[4] Chambers, C.G., Tanenhaus, M.K., & Magnuson, J.S., 2004. Actions and affordances in syntactic ambiguity resolution. *Jnl. of Experimental Psychology: Learning, Memory & Cognition* **30**: 687-696.

[5] Clark, H., and Wilkes-Gibbs, D. 1990. Referring as a collaborative process. In Cohen, P., Morgan, J. and Pollack, M. E., eds. *Intentions in Communication*, MIT Press. 463-493.

[6] DeVault, D., and Stone, M. Domain inference in incremental interpretation. ICOS 2003.

[7] Dzikovska, M.O., Allen, J.F., Swift, M.D. Integrating linguistic and domain knowledge for spoken dialogue systems in multiple domains. Workshop on Knowledge and Reasoning in Practical Dialogue Systems, IJCAI 2003.

[8] Fernandez, R., Ginzburg, J., and Lappin, S. Classifying ellipsis in dialogue: A machine learning approach. COLING 2004.

[9] Galescu, L., Ringger, E., and Allen, J. Rapid language model development for new task domains. LREC 1998.

[10] Hanna, J. E., and Tanenhaus, M.K. 2003 The effects of common ground and perspective on domains of referential interpretation. *Jnl. of Memory and Language* **49**: 43-61.

[11] Higashinaka, R., Miyazaki, N., Nakano, M., and Aikawa, K. A method for evaluating incremental utterance understanding in spoken dialogue systems. ICSLP 2002.

[12] Lamere, P., Kwok, P., Gouvêa, E., Bhiksha Raj, Singh, R., Walker, W., Warmuth, M., and Wolf, P. The CMU Sphinx-4 speech recognition system. ICASSP 2003.

[13] Rosé, C.P., Roque, A., Bhembe, D., and Van Lehn, K. An efficient incremental architecture for robust interpretation. HLT 2002.

[14] Stoness, S.C., Allen, J., Aist, G., and Swift, M. Using real-world reference to improve spoken language understanding. AAAI Workshop on Spoken Language Understanding 2005.

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

[16] Tanenhaus, M.K., Magnuson, J.S., Dahan, D., and Chambers, C. 2000. Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Jnl. of Psycholinguistic Research* **29**(6): 557-580.

This material is based upon work supported by the National Science Foundation under Grants No. 0328810 and BCS-0110676, and by the National Institutes of Health under Grants No. HD 27206 and NIDCD DC 005071. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the National Institutes of Health.