Decalog 2007: Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue, pages 149–154.
Trento, Italy, 30 May – 1 June 2007. Edited by Ron Artstein and Laure Vieu.

Incremental understanding in human-computer dialogue and experimental evidence for advantages over nonincremental methods

Gregory Aist[1] (gregory.aist@asu.edu), James Allen[2] (james@cs.rochester.edu), Ellen Campana[2,3,4,5] (ecampana@bcs.rochester.edu), Carlos Gomez Gallo[2] (cgomez@cs.rochester.edu), Scott Stoness[2] (stoness@cs.rochester.edu), Mary Swift[2] (swift@cs.rochester.edu), and Michael K. Tanenhaus[3] (mtan@bcs.rochester.edu)

[1]Computer Science and Engineering Arizona State Tempe AZ USA

[2]Computer Science University of Rochester Rochester NY USA

[3]Brain and Cognitive Sciences University of Rochester Rochester NY USA

[4]Arts, Media, and Engineering Arizona State Tempe AZ USA

[5]Department of Psychology Arizona State Tempe AZ USA

Abstract

Current dialogue systems generally operate in a pipelined, modular fashion on one complete utterance at a time. Converging evidence shows that human understanding operates incrementally and makes use of multiple sources of information during the parsing process, including traditionally "later" aspects such as pragmatics. We describe a spoken dialogue system that understands language incrementally, gives visual feedback on possible referents during the course of the user's utterance, and allows for overlapping speech and actions. We present findings from an empirical study showing that the resulting dialogue system is faster overall than its nonincremental counterpart. Furthermore, the incremental system is preferred to its counterpart – beyond what is accounted for by factors such as speed and accuracy. These results are the first to indicate, from a controlled user study, that successful incremental understanding systems will improve both performance and usability.

1 Introduction

The standard model of natural language understanding for dialogue systems is pipelined, modular, and operates on complete utterances. By pipelined we mean that only one level of processing operates at a time, in a sequential manner. By modular, we mean that each level of processing depends only on the previous level. By complete utterances we mean that the system operates on one sentence at a time.

There is, however, converging evidence that human language processing is neither pipelined nor modular nor whole-utterance. Evidence is converging from a variety of sources, including particularly actions taken while speech arrives. For example, natural turn-taking behavior such as backchanneling (uh-huh) and interruption occur while the speaker is still speaking. Evidence from psycholinguistics also shows incremental language understanding in humans (Tanenhaus et al. 1995, Traxler et al. 1997, Altmann and Kamide 1999) as evidenced by eye movements during language comprehension.

Many different sources of knowledge are available for use in understanding. On the speech recognition side, commonly used sources of information include acoustics, phonetics and phonemics, lexical probability, and word order. In dialogue systems, additional sources of information often include syntax and semantics (both general and domain-specific.) There are also however some sources of information that are less frequently programmed. These include such linguistics as morphology and prosody. Knowledge-based features are also available, such as world knowledge (triangles have three sides), domain knowledge (here there are two sizes of triangles), and task knowledge (the next step is to click on a small triangle. And, there is also pragmatic information available from the visual context (there is a small triangle near the flag.)

Here we discuss some of the progress we have made towards building methods for incremental understanding of spoken language by machines, which incorporates pragmatic information at the early stages of the understanding process. We also present a controlled experimental evaluation of our incremental system vs. its nonincremental counterpart.

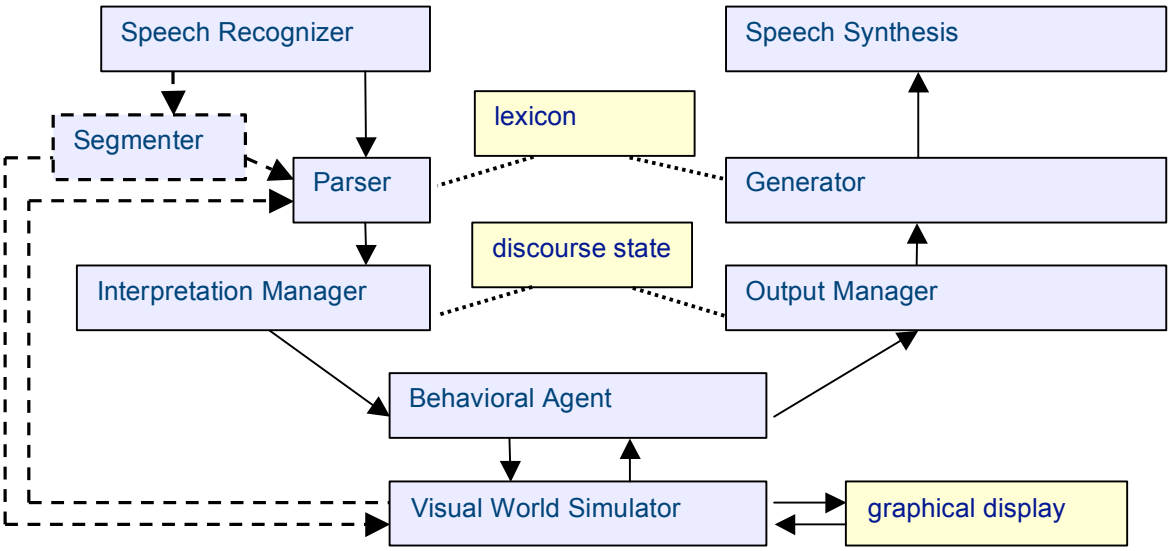

Figure 1. Changes to spoken dialogue system architecture to allow incremental understanding. Boxes show components; lines show message flow. In both types of systems, the lexicon and the discourse state are resources shared by input and output. Components and connections new to the incremental system are shown in **dashed lines**. Incremental understanding also places requirements on the speech recognizer (production of partial hypotheses), the parser (incremental construction of charts), the interpretation manager and behavioral agent (handling partial interpretations and actions), and the visual world simulator (incorporation of semantic models of partial actions) which are also important to the overall functioning of the system. This paper focuses on incremental *understanding* and thus the changes are to the *understanding* aspects of the dialogue system, including action-taking as representing direct evidence of understanding.

2 Traditional vs. Incremental Systems

Figure 1 shows a diagram of a traditional dialogue system architecture, with additional components and connections to added to support incremental understanding. Incremental language processing as we conceive it involves a number of fundamental and inter-related changes to the way in which processing occurs:

(a) input sentences are processed before the user turn ends, as opposed to processing only when turn is finished;

(b) components of the architecture operate asynchronously with several operating simultaneously, in contrast to a serial one where only one module at a time can be active;

(c) knowledge resources are available to several components at the same time, in contrast to a "pipeline" architecture where knowledge is sent from module to module;

(d) there is overlapping speech and graphical output ("action"), in contrast to presenting speech and other output sequentially;

(e) system and user turns and actions can overlap as appropriate for the dialogue.

We discuss some of these distinctions in more detail.

In a traditional dialogue system architecture, each component processes input from other components one utterance at a time. In our incremental architecture, each component receives input from other components as available, on a word-by-word basis.

In a traditional system, each component feeds forward into other components. In our incremental architecture, each component advises other components as needed – and advice can flow both "forward" in the traditional directions and "backward" from traditionally later stages of processing (such as pragmatics) to traditionally earlier stages of processing (such as parsing.) In a traditional system, the internal representations assume a strict division of time according to what's happening – the system is speaking, or the user is speaking, or the system is acting, and so forth. In our incremental architecture, representations can accommodate multiple events happening at once – such as the system acting while the user is still speaking.

In addition to these overall changes, our system incorporates a number of specific changes.

1. A Segmenter operates on incoming words, identifies pragmatically relevant fragments, and announces them to other system components such as the parser and the visual world simulator.

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 2. Example human-human dialogue in the fruit carts domain.

2. Pragmatic information is provided to the parser in order to assist with ongoing parses.
3. Modeling of actions and events is done by means of incremental semantics, in order to properly represent partial actions and allow for overlapping actions and speech.
4. Visual feedback is provided to the user about possible referents while the user is speaking.

3 Testbed Domain: Fruit Carts

To explore the effects of incremental understanding in human-computer dialogue, we devised a testbed domain (Figures 2, 3) where a person gives spoken instructions to a computer in order to reproduce a goal map. On the map, there are named regions, some of which contain flags as landmarks; the screen also has two kinds of objects: abstract shapes such as triangles and squares, and "fruit" of various kinds (avocados, bananas, cucumbers, grapefruits, and tomatoes.) In this domain, certain steps were taken in order to reduce complexity and increase the predictability of the spoken language. In particular, all objects and names of regions were chosen to be easy to name (or read) and easy for the speech recognizer to hear. In order to facilitate the study of incremental understanding of natural language

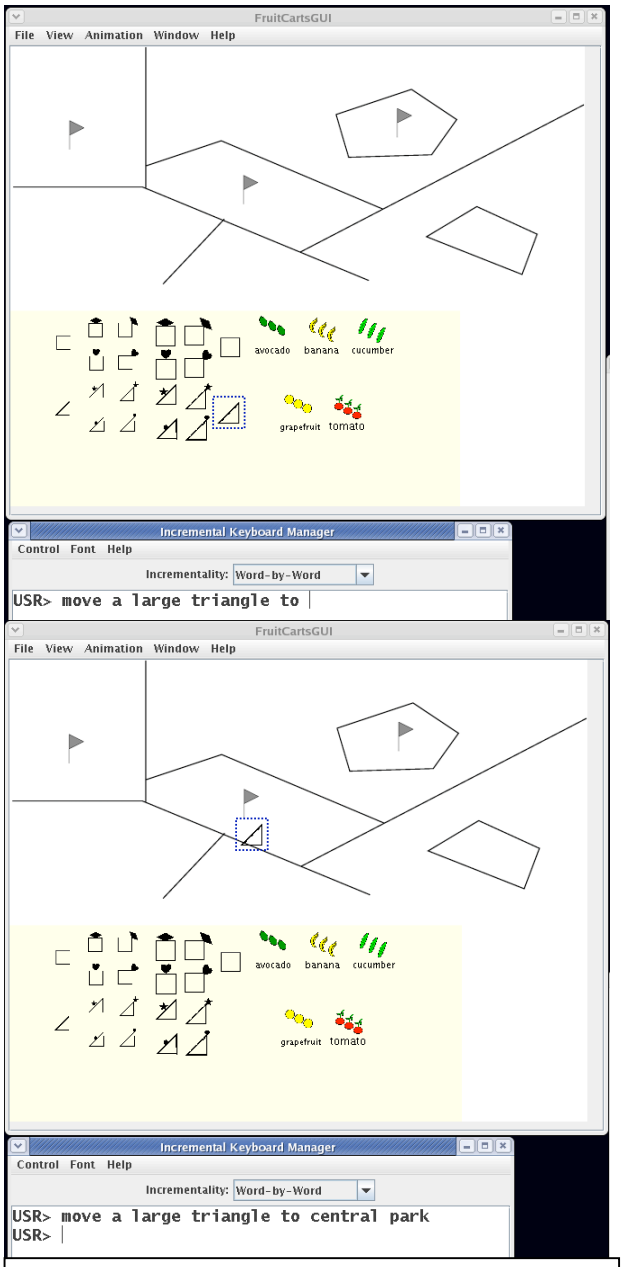

Figure 3. An example interaction with the incremental dialogue system. Note that in the top screenshot, halfway through the sentence, the large triangle is already highlighted. This figure shows typed input for clarity; the experiments used spoken input.

by machines, the Fruit Carts domain contains various points of disambiguation based on factors including object size, color, shape, and decoration; presence or absence of a landmark; and phonetic similarity of geographically close regions of the map (e.g. "Morningside" and "Morningside Heights" are close together.) For example, a square with stripes could also be referred to as "the stripey square", but a square with diamonds on the corner cannot be referred to as *"the corner-diamonded square". We thus chose a set of shapes such as "a small square with a diamond on the edge", "a large triangle with a star on the corner", "a small triangle with a circle on the edge", and so forth. Human-

human dialogue collected in this domain was used in the construction of the dialogue system.

We collected a set of dialogs from human-human conversation in this domain. Our observations included the following:

1. End-of-sentence boundaries tend to be fairly clear (at least to a human listener). Where a sentence begins, however, is quite difficult to say precisely, due to disfluencies, abandoned utterances, and so forth. This is in contrast to domains where speakers might tend to begin a sentence clearly, such as information retrieval ("Search for books by Kurt Vonnegut").

2. There seem to be two distinct strategies that people can employ: saying a direction all at once ("Put it one inch below the flag") or continuously ("Put it near the flag [pause] but down a bit [pause] a bit more [pause] stop.")

3. Besides a pure All-at-once and Continuous strategy, people sometimes switch between them, employing Both. For example, the director might tell the actor to place an object "right on the flag [pause] down a bit [pause] keep going [pause] stop." We see these as possibilities along a continuum, using the same language mechanisms yet according different emphasis to the strategies.

Our previous findings about these types of language include that continuous-style language uses fewer words per utterance than all-at-once language, and the words themselves are shorter in length as well (reference omitted for review). Furthermore, the use of continuous language increases over the course of the dialogs. Specifically, the relative percentage of continuous language increases over trials. The relative increase in continuous language over time is statistically significant (by logistic regression; style as outcome, subject as categorical, trial as numeric. $B=0.104 \pm 0.037$, $\exp(B) \approx 1.11$, $p < 0.01$). So not only do people engage in dialogue that relies on incremental understanding on the part of the hearer, but such interactions actually becomes more important as the dialogue progresses.

We used these human-human conversations to form the basis for formalizing various aspects of continuous understanding, and for gauging the behavior of the spoken dialog system that we built to operate in this testbed domain. The resulting system is capable of interactions as shown in Figure 3, where the user's utterance is processed as it is received, visual feedback is provided during the course of the utterance, and speech and actions can overlap. As in the human-human interactions, moving an object from one location to another takes time in the working system – that is, the objects are shown moving in a straight line from the beginning point (e.g. the bin at the bottom of the screen) to the end point (the flag in central park.)

4 Related Work

We have previously shown that incremental parsing can be faster and more accurate than non-incremental parsing (references omitted for review.) In addition, we have shown that in this domain the relative percentage of language that is of a more interactive style also increases over time (references omitted.) A number of research efforts have been directed at incremental understanding, adopting a wide variety of techniques including the blackboard architecture, finite state machines (Ait-Mokhtar and Chanod 1997), perceptrons (Collins and Roark 2004), neural networks (Jain and Waibel 1990), categorial grammar (Milward 1992), tree-adjoining grammar (Poller 1994), and chart parsing (Wiren 1989). We compare our work to several such efforts.

Higashinaka et al. (2002) performed a linear regression experiment to find a set of features that predict performance of systems that understand utterances incrementally. The system evaluated by the authors is incremental in that dialogue states are updated as the sentence is processed. However this is a result of incrementally processing the input stream and not the type of continuous understanding we propose. In our approach we allow the parser to make use of information from different layers of processing (i.e. pragmatic constraints from verb-argument constructions, real world knowledge, etc).

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 parsing. 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.

Work by Schuler (2001 and following) has moved away from a pipeline architecture by accessing different sources of knowledge while parsing the sentence. Using real world knowledge about objects improves parsing and can only be achieved by analyzing the sentence from the start. Schuler makes use of potential referents from the environment much the same way that we have also done by the use of model-theoretic interpretations. Thus the system evaluates the logical expressions for all possible potential referents at each node of the tree to know whether they are possible in the current domain. The author provides an example where a PP attachment ambiguity is resolved by knowing a particular fact about the world which rules out one of the two possible attachments. Thus this sort of knowledge comes into play during parsing. Even though the system described in the present paper shares the same goals in using more than just syntactic knowledge for parsing, our parser feedback framework does not require the rewriting of the grammar used for parsing to incorporate environment knowledge. This approach based on probability feedback directly affecting the parser chart is simpler and thus more applicable to and easily incorporated in a wider range of parsers and grammars.

5 Evaluation

We conducted a controlled evaluation comparing incremental understanding to its nonincremental counterpart in our testbed domain. In the nonincremental system, speech and actions alternate; in the incremental system, the actions and speech overlap.

A total of 22 dialogues were collected, each of which consisted of two utterances and the corresponding system responses. Eleven of the dialogues were in the control (nonincremental) condition and eleven of the dialogues were in the experimental (incremental) condition. The utterances were in-domain and understandable by both the nonincremental and incremental versions of the system, they were pre-recorded; and the same utterances were played to each version of the system; this technique allowed us to minimize variance due to extraneous factors such as interspeaker variability, acoustic noise, and so forth, and concentrate specifically on the difference between incremental processing and its nonincremental counterpart. The resulting dialogues were recorded on digital video.

The incremental system was approximately 20% faster than the nonincremental system in terms of time to task completion for each two-utterance dialogue, at 44 seconds per dialogue vs. 52 seconds for the control condition (single-factor ANOVA, F=10.72, df=21, p-value 0.004).

To further evaluate the effectiveness of the incremental system, we conducted an onlooker study where 18 subjects, mostly from the University of Rochester community, rated the interactions in the dialogues. First, each subject watched one video clip once and only once; then, the subject filled out written responses to questions about that video clip. Subjects provided responses for each dialogue video clip to four Likert-scaled (1-7, 1=less) questions on speed, accuracy, match-to-intent, and satisfaction:

[FAST] "How fast did the computer respond?"
[ACC] "How accurately did the system understand?"
[ACT] "How well matched were the computer's actions to what the person wanted?"
[SAT] "If you had been the person giving the commands, how satisfied overall would you be with the interaction?"

In order to check that people's responses were objectively correlated with actual system performance, four "wrong" system videos were included in the study, two for each condition (nonincremental control and incremental / experimental condition). That is, the user in the video said one thing, but the system did something else. In this way, we experimentally manipulated the "right/wrong" response of the system to see how people would rate the system's correctness.

We measured speed, accuracy, and match to user intentions with a subjective survey; as it happens, our results are compatible with methods that measure these factors objectively and then relate them to subjectively reported user satisfaction. For example, the PARADISE model (Walker et al. 1997) found that speed, accuracy, and match to user intentions well predicted user satisfaction. Using a linear regression model as in the original PARADISE framework, we confirmed that with our data a linear model with speed (FAST), accuracy (ACC), and match-to-actions (ACT) as input variables predicts well the output variable satisfaction (SAT) (R=.795, R Square=.631, Adj. R Square=.625; df=3, F=91.389, $p<0.001$).

Since the input and output variables are seven-item Likert scale responses it turns out that ordi-

nal regression models are a better match to the experimental setup than the linear regression models. Ordinal regression models are specifically designed for cases where the variables are a set of levels that are ordered (N+1>N) but not necessarily linear (1 to 2 may not be the same as 4 to 5.) We thus adopted ordinal regression models for the remainder of the analyses. In addition, since some of the subjects indicated in written comments that they got used to the behavior of the system over time, we included the dialogue number (NTH; 1=first seen, 22=last seen) as a covariate. And, since individual subjects tend to vary in their responses (some subjects more negative than other subjects), we also included subject (SUBJ) as an input variable.

The model we built to analyze the effects of right/wrong system response (RIGHT) and nonincremental vs. incremental processing (INC) was as follows. We built an ordinal regression model predicting satisfaction (SAT) by right/wrong (RIGHT) and nonincremental/incremental (INC) and subject (SUBJ) with FAST, ACC, and ACT as covariates (Table 1).

The first result we found was that there was a significant effect for RIGHT as a predictor of user satisfaction, in the expected direction: wrong responses predict lower satisfaction (or, equivalently, correct responses predict higher satisfaction.) These results help validate the external reliability of the experimental design.

Next, to evaluate the effects of incremental vs. nonincremental processing, we examined the model coefficient for INC. In this case, nonincremental processing was a significant predictor of lower satisfaction (p=.026) – or, equivalently, incremental processing was a significant predictor of higher satisfaction.

Table 1. Parameters of ordinal regression model predicting satisfaction (SAT).

Variable	Estimate	Std. Error	Sig.
NTH	.188	.058	.001
FAST	.770	.176	.000
ACC	1.411	.341	.000
ACT	.616	.304	.043
RIGHT=0 (0=wrong, 1=right.)	-1.855	.903	.040
INC=0 (0=control 1=incr.)	-2.336	1.051	.026

6 Conclusion

Our results show that – at least for this task – incremental processing predicts higher user satisfaction. Why? The statistical model makes clear that this preference is the case after controlling for factors such as speed, accuracy, and match-to-intent. Explanatory factors that remain include naturalness – that is, the ways in which incremental systems are more like human-human conversation than their nonincremental counterparts. Nonincremental dialogue systems require many artificial restrictions on what the user and the system can say and when they can say it, and therefore exclude many important characteristics of natural human dialogue. Incremental understanding has the potential to remove such obstacles. The work presented here suggests that successful incremental understanding systems will improve both performance and usability

References

Ait-Mokhtar, S. and Chanod, J.-P. Incremental finite-state parsing. ANLP 1997.

Altmann, G. and Kamide, Y. Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition* 73(3):247-264. 1999.

Collins, M. and B. Roark. Incremental parsing with the perceptron algorithm. ACL 2004.

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

Higashinaka, R., Miyazaki N., Nakano, M., & Kiyoaki, A. A method for evaluating incremental utterance understanding in spoken dialogue systems. ICSLP 2002.

Jain, A. & Waibel, A. Incremental parsing by modular recurrent connectionist networks. NIPS 1990.

Milward, D. Dynamics, dependency grammar and incremental interpretation. COLING 1992.

Poller, P. Incremental parsing with LD/TLP-TAGS. Computational Intelligence 10(4). 1994.

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

Schuler, W. Computational properties of environment-based disambiguation. ACL 2001.

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

Traxler, M.J., Bybee, M.D., & Pickering, M.J. Influence of Connectives on Language Comprehension: Eye-tracking Evidence for Incremental Interpretation. The Quarterly Journal of Experimental Psychology: Section A, 50(3), 481-497. 1997.

Wiren, M. Interactive incremental chart parsing. In Proceedings of the 4th Meeting of the European Chapter of the Association for Computational Linguistics. 1989.

Walker, M., Litman, D., Kamm C., and Abella, A. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. ACL 1997.