Automatic Detection and Correction of Repairs in Human-Computer Dialog*

Elizabeth Shriberg†, *John Bear, John Dowding*

SRI International
Menlo Park, California 94025

ABSTRACT

We have analyzed 607 sentences of spontaneous human-computer speech data containing repairs (drawn from a corpus of 10,718). We present here criteria and techniques for automatically detecting the presence of a repair, its location, and making the appropriate correction. The criteria involve integration of knowledge from several sources: pattern matching, syntactic and semantic analysis, and acoustics.

1. INTRODUCTION

Spontaneous spoken language often includes speech that is not intended by the speaker to be part of the content of the utterance. This speech must be detected and deleted in order to correctly identify the intended meaning. This broad class of disfluencies encompasses a number of phenomena, including word fragments, interjections, filled pauses, restarts, and repairs. We are analyzing the repairs in a large subset (over ten thousand sentences) of spontaneous speech data collected for the DARPA spoken language program. We have categorized these disfluencies as to type and frequency, and are investigating methods for their automatic detection and correction. Here we report promising results on detection and correction of repairs by combining pattern matching, syntactic and semantic analysis, and acoustics.

The problem of disfluent speech for language understanding systems has been noted but has received limited attention. Hindle [5] attempts to delimit and correct repairs in spontaneous human-human dialog, based on transcripts containing an "edit signal," or external and reliable marker at the "expunction point," or point of interruption. Carbonell and Hayes [4] briefly describe recovery strategies for broken-off and restarted utterances in textual input. Ward [13] addresses repairs in spontaneous speech, but does not attempt to identify or correct them. Our approach is most similar to that of Hindle. It differs, however, in that we make no assumption about the existence of an explicit edit signal. As a reliable edit signal has yet to be found, we take it as our problem to find the site of the repair automatically.

It is the case, however, that cues to repair exist over a range of syllables. Research in speech production has shown that repairs tend to be marked prosodically [8] and there is perceptual evidence from work using lowpass-filtered speech that human listeners can detect the occurrence of a repair in the absence of segmental information [9].

In the sections that follow, we describe in detail our corpus of spontaneous speech data and present an analysis of the repair phenomena observed. In addition, we describe ways in which pattern matching, syntactic and semantic analysis, and acoustic analysis can be helpful in detecting and correcting these repairs. We use pattern matching to determine an initial set of possible repairs; we then apply information from syntactic, semantic, and acoustic analyses to distinguish actual repairs from false positives.

2. THE CORPUS

The data we are analyzing were collected at six sites[1] as part of DARPA's Spoken Language Systems project. The corpus contains digitized waveforms and transcriptions of a large number of sessions in which subjects made air travel plans using a computer. In the majority of sessions, data were collected in a Wizard of Oz setting, in which subjects were led to believe they were talking to a computer, but in which a human actually interpreted and responded to queries. In a small portion of the sessions, data were collected using SRI's Spoken Language System [12]), in which no human intervention was in-

*This research was supported by the Defense Advanced Research Projects Agency under Contract ONR N00014-90-C-0085 with the Office of Naval Research. It was also supported by a Grant, NSF IRI-8905249, from the National Science Foundation. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency of the U.S. Government, or of the National Science Foundation.

†Elizabeth Shriberg is also affiliated with the Department of Psychology at the University of California at Berkeley.

[1]The sites were: AT&T, Bolt Beranek and Newman, Carnegie Mellon University, Massachusetts Institute of Technology, SRI International, and Texas Instruments, Inc.

volved. Relevant to the current paper is the fact that although the speech was spontaneous, it was somewhat planned (subjects pressed a button to begin speaking to the system) and the transcribers who produced lexical transcriptions of the sessions were instructed to mark words they inferred were verbally deleted by the speaker with special symbols. For further description of the corpus, see MADCOW [10].

3. CHARACTERISTICS AND DISTRIBUTION OF REPAIRS

Of the ten thousand sentences in our corpus, 607 contained repairs. We found that of sentences longer than nine words, 10% contained repairs. While this is lower than rates reported elsewhere for human-human dialog (Levelt [7] reports a rate of 34%), it is still large enough to be significant. And, as system developers move toward more closely modeling human-human interaction, the percentage is likely to rise.

3.1 Notation

In order to classify these repairs, and to facilitate communication among the authors, it was necessary for us to develop a notational system that would: (1) be relatively simple, (2) capture sufficient detail, and (3) describe the vast majority of repairs observed. The notation is described fully in [2].

The basic aspects of the notation include marking the interruption point, its extent, and relevant correspondences between words in the region. To mark the site of a repair, corresponding to Hindle's "edit signal" [5], we use a vertical bar (|). To express the notion that words on one side of the repair correspond to words on the other, we use a combination of a letter plus a numerical index. The letter M indicates that two words match exactly. R indicates that the second of the two words was intended by the speaker to replace the first. The two words must be similar, either of the same lexical category, or morphological variants of the same base form (including contraction pairs like I/I'd). Any other word within a repair is notated with X. A hyphen affixed to a symbol indicates a word fragment. In addition, certain cue words, such as "sorry" or "oops" (marked with CR) as well as filled pauses (CF) are also labeled if they occur immediately before the site of a repair.

I	want	fl-		flights	to	boston.
		M_1-	|	M_1		
		what		what	are	the fares
		M_1	|	M_1		
show	me	flights		daily	flights	
		M_1	|	X	M_1	
I want	a	flight		one	way	flight
		M_1	|	X	X	M_1
I want	to	leave		depart	before	...
		R_1	|	R_1		
	what	are		what	are	the fares
	M_1	M_2	|	M_1	M_2	
... fly	to	boston		from	boston	
	R_1	M_1	|	R_1	M_1	
... fly	from	boston		from	denver	
	M_1	R_1	|	M_1	R_1	
	what	are		are	there	any flights
	X	X	|			

Table 1: Examples of Notation

3.2 Distribution

While only 607 sentences contained deletions, some sentences contained more than one, for a total of 646 deletions. Table 2 gives the breakdown of deletions by length, where length is defined as the number of consecutive deleted words or word fragments. Most of the deletions were fairly short. One or two word deletions accounted for 82% of the data. We categorized the length 1 and length 2 repairs according to their transcriptions. The results are summarized in Table 3. For the purpose of simplicity, we have in this table combined cases involving fragments (which always occurred as the second word) with their associated full-word patterns. The overall rate of fragments for the length 2 repairs was 34%.

4. SIMPLE PATTERN MATCHING

We analyzed a subset of 607 sentences containing repairs and concluded that certain simple pattern-matching techniques could successfully detect a number of them.

Deletion Length	Occurrences	Percentage
1	376	59%
2	154	24%
3	52	8%
4	25	4%
5	23	4%
6+	16	3%

Table 2: Distribution of Repairs by Length

Type	Pattern	Frequency
	Length 1 Repairs	
Fragments	$M_1-, R_1-, X-$	61%
Repeats	$M_1 \mid M_1$	16%
Insertions	$M_1 \mid X_1 \ldots X_i M_1$	7%
Replacement	$R_1 \mid R_1$	9%
Other	$X \mid X$	5%
	Length 2 Repairs	
Repeats	$M_1\ M_2 \mid M_1\ M_2$	28%
Replace 2nd	$M_1\ R_1 \mid M_1\ R_1$	27%
Insertions	$M_1 M_2 \mid M_1 X_1 \ldots X_i M_2$	19%
Replace 1st	$R_1\ M_1 \mid R_1\ M_1$	10%
Other	...|...	17%

Table 3: Distribution of Repairs by Type

The pattern matching component reported on here looks for the following kinds of subsequences:

- Simple syntactic anomalies, such as "a the" or "to from".
- Sequences of identical words such as "<I> <would> <like> <a> <book> I would like a flight ..."
- Matching single words surrounding a cue word like "sorry," for example "from" in this case: "I would like to see the flights <from> <philadelphia> <i'm> <sorry> from denver to philadelphia."

Of the 406 sentences with nontrivial repairs in our data (more editing necessary than deleting fragments and filled pauses), the program successfully corrected 177. It found 132 additional sentences with repairs but made the wrong correction. There were 97 sentences that contained repairs which it did not find. In addition, out of the 10,517 sentence corpus ($10,718 - 201$ trivial), it incorrectly hypothesized that an additional 191 contained repairs. Thus of 10,517 sentences of varying lengths, it pulled out 500 as possibly containing a repair and missed 97 sentences actually containing a repair. Of the 500 that it proposed as containing a repair, 62% actually did and 38% did not. Of the 62% that had repairs, it made the appropriate correction for 57%.

These numbers show that although pattern matching is useful in identifying possible repairs, it is less successful at making appropriate corrections. This problem stems largely from the overlap of related patterns. Many sentences contain a subsequence of words that match not one but several patterns. For example the phrase "FLIGHT <word> FLIGHT" matches three different patterns:

show the	FLIGHT		earliest	FLIGHT			
	M_1	|	X	M_1			
show the	FLIGHT	time		FLIGHT	date		
	M_1	R_1	|	M_1	R_1		
show the	delta	FLIGHT		united	FLIGHT		
	R_1	M_1	|	R_1	M_1		

Each of these sentences is a false positive for the other two patterns. Despite these problems of overlap, pattern matching is useful in reducing the set of candidate sentences to be processed for repairs. Instead of applying detailed and possibly time-intensive analysis techniques to 10,000 sentences, we can increase efficiency by limiting ourselves to the 500 sentences selected by the pattern matcher, which has (at least on one measure) a 75% recall rate. The repair sites hypothesized by the pattern matcher constitute useful input for further processing based on other sources of information.

5. NATURAL LANGUAGE CONSTRAINTS

Here we describe experiments conducted to measure the effectiveness of a natural language processing system in distinguishing repairs from false positives. A false positive is a repair pattern that incorrectly matches a sentence or part of a sentence. We conducted the experiments using the syntactic and semantic components of the Gemini natural language processing system. Gemini is an extensive reimplementation of the Core Language Engine [1]. It includes modular syntactic and semantic components, integrated into an efficient all-paths bottom-up parser [11]). Gemini was trained on a 2,200 sentence subset of the full 10,718-sentence corpus (only those annotated as class A or D). Since this subset excluded the unanswerable (class X) sentences, Gemini's coverage on the full corpus is only an estimated 70% for syntax, and 50% for semantics.[2] Nonetheless, the results reported here are promising, and should improve as syntactic and semantic coverage increase.

We tested Gemini on a subset of the data that the pat-

[2]Gemini's syntactic coverage of the 2,200 sentence dataset it was trained on (the set of annotated and answerable MADCOW queries) is approximately 91%, while its semantic coverage is approximately 77%. On a fair test of the February 1992 test set, Gemini's syntactic coverage was 87% and semantic coverage was 71%.

Syntax Only		
	Marked as Repair	Marked as False Positive
Repairs	68 (96%)	56 (30%)
False Positives	3 (4%)	131 (70%)

Syntax and Semantics		
	Marked as Repair	Marked as False Positive
Repairs	64 (85%)	23 (20%)
False Positives	11 (15%)	90 (80%)

Table 4: Syntax and Semantics Results

tern matcher returned as likely to contain a repair. We excluded all sentences that contained fragments, resulting in a dataset of 335 sentences, of which 179 contained repairs and 176 contained false positives. The approach was as follows: for each sentence, parsing was attempted. If parsing succeeded, the sentence was marked as a false positive. If parsing did not succeed, then pattern matching was used to detect possible repairs, and the edits associated with the repairs were made. Parsing was then reattempted. If parsing succeeded at this point, the sentence was marked as a repair. Otherwise, it was marked as NO OPINION.

Since multiple repairs and false positives can occur in the same sentence, the pattern matching process is constrained to prefer fewer repairs to more repairs, and shorter repairs to longer repairs. This is done to favor an analysis that deletes the fewest words from a sentence. It is often the case that more drastic repairs would result in a syntactically and semantically well-formed sentence, but not the sentence that the speaker intended. For instance, the sentence "show me <flights> daily flights to boston" could be repaired by deleting the words "flights daily", and would then yield a grammatical sentence, but in this case the speaker intended to delete only "flights."

Table 4 shows the results of these experiments. We ran them two ways: once using syntactic constraints alone and again using both syntactic and semantic constraints. As can be seen, Gemini is quite accurate at detecting a repair, although somewhat less accurate at detecting a false positive. Furthermore, in cases where Gemini detected a repair, it produced the intended correction in 62 out of 68 cases for syntax alone, and in 60 out of 64 cases using combined syntax and semantics. In both cases, a large number of sentences (29% for syntax, 50% for semantics) received a NO OPINION evaluation. The NO OPINION cases were evenly split between repairs and false positives in both tests.

The main points to be noted from Table 4 are that with syntax alone, the system is quite accurate in detecting repairs, and with syntax and semantics working together, it is accurate at detecting false positives. However, since the coverage of syntax and semantics will always be lower than the coverage of syntax alone, we cannot compare these rates directly.

6. ACOUSTICS

A third source of information that can be helpful in detecting repairs is acoustics. While acoustics alone cannot tackle the problem of locating repairs, since any prosodic patterns found in repairs will be found in fluent speech, acoustic information can be quite effective when combined with other sources of information, particularly, pattern matching.

Our approach in studying the ways in which acoustics might be helpful was to begin by looking at two patterns conducive to acoustic measurement and comparison. First, we focused on patterns in which there is only one matched word, and in which the two occurrences of that word are either adjacent or separated by only one word. Matched words allow for comparisons of word duration; proximity helps avoid variability due to global intonation contours not associated with the patterns themselves. We present here analyses for the $M_1|M_1$ ("flights for <one> one person") and $M_1|XM_1$ ("<flight> earliest flight") repairs, and their associated false positives ("u s air five one one," "a flight on flight number five one one," respectively).

Second, we have done a preliminary analysis of repairs in which a word such as "no" or "well" was present as an editing expression [6] at the point of interruption ("...flights <between> <boston> <and> <dallas> <no> between oakland and boston"). False positives for these cases are instances in which the cue word functions in its usual lexical sense ("I want to leave boston no later than one p m."). Hirshberg and Litman [3] have shown that cue words that function differently can be distinguished perceptually by listeners on the basis of prosody. Thus, we sought to determine whether acoustic analysis could help in deciding, when such words were present, whether or not they marked the interruption point of a repair.

In both analyses, a number of features were measured to allow for comparisons between the words of interest.

	Pauses before/ after X		F0 of X	
	before X (only)	after X (only)	greater than F0 of 1st M_1	less than F0 of 1st M_1
False Positives (N=24)	.08	.58	.08	.42
Repairs (N=12)	.83	.00	.92	.08

Table 5: Acoustic Characteristics of $M_1|XM_1$ Repairs

	Pauses after X (only) and F0 of X less than F0 of 1st M_1	Pauses before X (only) and F0 of X greater than F0 of 1st M_1
False Positives	.58	.00
Repairs	.00	.92

Table 6: Combining Acoustic Characteristics of $M_1|XM_1$ Repairs

Word onsets and offsets were labeled by inspection of waveforms and parameter files (pitch tracks and spectrograms) obtained using the Entropic Waves software package. Files with questionable pitch tracks were excluded from the analysis. An average F0 value for words of interest was determined by simply averaging, within a labeled word, all 10-ms frame values having a probability of voicing above 0.20.

In examining the $M_1|M_1$ repair pattern, we found that the strongest distinguishing cue between the repairs ($N = 20$) and the false positives ($N = 20$) was the interval between the offset of the first word and the onset of the second. False positives had a mean gap of 42 ms ($s.d. = 55.8$) as opposed to 380 ms ($s.d. = 200.4$) for repairs. A second difference found between the two groups was that, in the case of repairs, there was a statistically reliable reduction in duration for the second occurrence of M_1, with a mean difference of 53.4 ms. However because false positives showed no reliable difference for word duration, this was a much less useful predictor than gap duration. F0 of the matched words was not helpful in separating repairs from false positives; both groups showed a highly significant correlation for, and no significant difference between, the mean F0 of the matched words.

A different set of features was found to be useful in distinguishing repairs from false positives for the $M_1|XM_1$ pattern. These features are shown in Table 5. Cell values are percentages of repairs or false positives that possessed the characteristics indicated in the columns. Despite the small data set, some suggestive trends emerge. For example, for cases in which there was a pause (defined for purposes of this analysis as a silence of greater than 200 ms) on only one side of the inserted word, the pause was never after the insertion (X) for the repairs and rarely before the X in the false positives. Note that values do not add up to 100% because cases of no pauses, or pauses on both sides are not included in the table. A second distinguishing characteristic was the F0 value of X. For repairs, the inserted word was nearly always higher in F0 than the preceding M_1; for false positives, this increase in F0 was rarely observed. Table 6 shows the results of combining the acoustic constraints in Table 5. As can be seen, although acoustic features may be helpful individually, certain combinations of features widen the gap between observed rates of repairs and false positives possessing the relevant set of features.

Finally, in a preliminary study of the cue words "no" and "well," we compared 9 examples of these words at the site of a repair to 15 examples of the same words occurring in fluent speech. We found that these groups were quite distinguishable on the basis of simple prosodic features. Table 7 shows the percentage of repairs versus false positives characterized by a clear rise or fall in F0, lexical stress, and continuity of the speech immediately preceding and following the editing expression ("continuous" means there is no silent pause on either side of the cue word). As can be seen, at least for this limited data set, cue words marking repairs were quite distinguishable from those same words found in fluent strings on the basis of simple prosodic features.

	F0 rise	F0 fall	Lexical stress	Continuous speech
Repairs	.00	1.00	.00	.00
False positives	.87	.00	.87	.73

Table 7: Acoustic Characteristics of Cue Words

Although one cannot draw conclusions from such limited

data sets, such results are nevertheless interesting. They illustrate that acoustics can indeed play a role in distinguishing repairs from false positives, but only if each pattern is examined individually, to determine which features to use, and how to combine them. Analysis of additional patterns and access to a larger database of repairs will help us better determine the ways in which acoustics can play a role in detection of repairs.

7. CONCLUSION

In summary, disfluencies occur at high enough rates in human-computer dialog to merit consideration. In contrast to earlier approaches, we have made it our goal to detect and correct repairs automatically, without assuming an explicit edit signal. Without such an edit signal, however, repairs are easily confused both with false positives and with other repairs. Preliminary results show that pattern matching is effective at detecting repairs without excessive overgeneration. Our syntax-only approach is quite accurate at detecting repairs and correcting them. Acoustics is a third source of information that can be tapped to provide corroborating evidence about a hypothesis, given the output of a pattern matcher.

While none of these knowledge sources by itself is sufficient, we propose that by combining them, and possibly others, we can greatly enhance our ability to detect and correct repairs. As a next step, we intend to explore additional aspects of the syntax and semantics of repairs, analyze further acoustic patterns, and examine corpora with higher rates of disfluencies.

ACKNOWLEDGMENTS

We would like to thank Patti Price for her helpful comments on earlier drafts, as well as for her participation in the development of the notational system used. We would also like to thank Robin Lickley for his helpful feedback on the acoustics section.

REFERENCES

1. Alshawi, H, Carter, D., van Eijck, J., Moore, R. C., Moran, D. B., Pereira, F., Pulman, S., and A. Smith (1988) *Research Programme In Natural Language Processing: July 1988 Annual Report*, SRI International Tech Note, Cambridge, England.
2. Bear, J., Dowding, J., Price, P., and E. E. Shriberg (1992) "Labeling Conventions for Notating Grammatical Repairs in Speech," unpublished manuscript, to appear as an SRI Tech Note.
3. Hirschberg, J. and D. Litman (1987) "Now Let's Talk About Now: Identifying Cue Phrases Intonationally," *Proceedings of the ACL*, pp. 163-171.
4. Carbonell, J. and P. Hayes, P., (1983) "Recovery Strategies for Parsing Extragrammatical Language," *American Journal of Computational Linguistics*, Vol. 9, Numbers 3-4, pp. 123-146.
5. Hindle, D. (1983) "Deterministic Parsing of Syntactic Non-fluencies," *Proceedings of the ACL*, pp. 123-128.
6. Hockett, C. (1967) "Where the Tongue Slips, There Slip I," in *To Honor Roman Jakobson: Vol. 2*, The Hague: Mouton.
7. Levelt, W. (1983) "Monitoring and self-repair in speech," *Cognition*, Vol. 14, pp. 41-104.
8. Levelt, W., and A. Cutler (1983) "Prosodic Marking in Speech Repair," *Journal of Semantics*, Vol. 2, pp. 205-217.
9. Lickley, R., R. Shillcock, and E. Bard (1991) "Processing Disfluent Speech: How and when are disfluencies found?" *Proceedings of the Second European Conference on Speech Communication and Technology*, Vol. 3, pp. 1499-1502.
10. MADCOW (1992) "Multi-site Data Collection for a Spoken Language Corpus," *Proceedings of the DARPA Speech and Natural Language Workshop*, February 23-26, 1992.
11. Moore, R. and J. Dowding (1991) "Efficient Bottom-up Parsing," *Proceedings of the DARPA Speech and Natural Language Workshop*, February 19-22, 1991, pp. 200-203.
12. Shriberg, E., Wade, E., and P. Price (1992) "Human-Machine Problem Solving Using Spoken Language Systems (SLS): Factors Affecting Performance and User Satisfaction," *Proceedings of the DARPA Speech and Natural Language Workshop*, February 23-26, 1992.
13. Ward, W. (1991) "Evaluation of the CMU ATIS System," *Proceedings of the DARPA Speech and Natural Language Workshop*, February 19-22, 1991, pp. 101-105.