A PARSER COPING WITH SELF-REPAIRED JAPANESE UTTERANCES AND LARGE CORPUS-BASED EVALUATION

Yuji Sagawa **Noboru Ohnishi** **Noboru Sugie**

Dept. of Information Engineering, Nagoya University, Japan

INTRODUCTION

Self-repair(Levelt 1988) is a repair of utterance by speaker him/herself. A human speaker makes self-repairs very frequently in spontaneous speech. (Blackmer and Mitton 1991) reported that self-repairs are made once every 4.8 seconds in dialogues taken from radio talk shows.

Self-repair is one kind of "permissible ill-formedness", that is a human listener can feel ill-formedness in it but he/she is able to recognize its intended meaning. Thus your partner does not need to interrupt dialogue.

How do you feel if your partner interrupts dialogue every 5 seconds to ask "What do you mean?" or so? You will give up dialogue or choose means of writing. Speaking without self-repair is the most difficult modality of natural language communication.

The goal of our work is to make a dialogue system coping with self-repaired utterances. In this paper we propose a parser called SERUP(SElf-Repaired Utterance Parser), which plays a major part in understanding a self-repaired utterance. That is, because our approach is to translate a self-repaired utterance (Ex.1) into a well-formed version that does not contain self-repair (Ex.2) and parse the well-formed one, we do not need to change the subsequent processes.

[Ex.1] *And from green left to pink, er, from blue left to pink* (from (Levelt 1988))

[Ex.2] *And from blue left to pink*

SERUP uses some linguistic clues to translate utterances, those include a repetition, an unknown word and/or an isolated word. We describe how SERUP uses these clues.

To evaluate SERUP, we analyze a large corpus that contains spontaneous dialogues over telephone. From the result, we estimate that SERUP works well with 88.1 % of 1,082 self-repairs in the corpus.

RELATED WORKS

(Hindle 1983) and (Langer 1990) proposed parsers coping with self-repaired utterances. But they assumed that an interruption point has already been detected. Hindle thought prosodic cues can be used in detection, but it is not clear if they can always succeed. Langer thought editing expressions can be used, but they are not always used in self-repair.

Recently, (Shriberg, Bear, and Dowding 1992) proposed a pattern matching method and used it in GEMINI system(Dowding et al. 1993). This is similar to our method, but the corpus(MADCOW 1992) used is less spontaneous than ours. (Subjects pressed a button to begin speaking to the system)

(Nakatani and Hirschberg 1993) proposed a speech-first method in which prosodic cues are used mainly. We also think prosodic cues are important. But we think people use linguistic cues mainly because they can understand self-repaired utterances in transcripts.

All these works are done on English. (Langer also treats Germany) Because there are many syntactic differences (e.g., left branching v.s. right branching), it is not

clear if their approach is applicable to Japanese.

OUTLINE OF SERUP

Fig.1 shows the outline of SERUP.

Normal Parser is a parser that parses well-formed utterances. When Normal Parser fails to parse an utterance, the utterance is passed to SR-reconstructor that detects a self-repair in it and translates it into well-formed version. The translated utterance is returned to Normal Parser and parsed again.

Because an utterance can contain two or more self-repairs, translation is repeated until Normal Parser succeeds in parsing or translation fails. In the latter case, the utterance has another ill-formedness or self-repair that the SR-reconstructor cannot cope with.

There are two main problems in translation. One is to determine an interruption point, and the other is to determine a reparandum. If these two problems can be solved, then the process of translation is carried out as follows.

1. Remove editing expressions such as *er, no, I mean.*
2. Supersede the reparandum with repair part.

For more detail of SERUP, see (Sagawa, Ohnishi, and Sugie 1993).

CLUES TO TRANSLATION

In this chapter, we will describe a classification of self-repaired utterances. They are classified by clues usable to determine an interruption point and a reparandum.

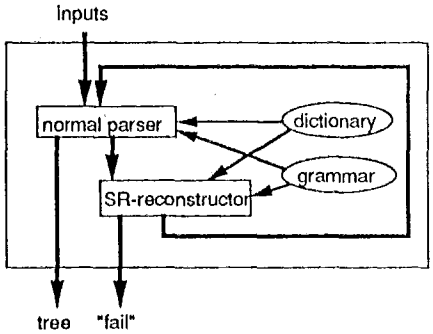

Fig. 1: The outline of SERUP

Table 1 shows the classification. Categories printed in italics have no clue, i.e., SERUP fails to parse utterances in those categories.

with repetition

A self-repair is mostly made in a way to repair a word or a phrase just before an interruption(Levelt 1988). So words or phrases around an interruption are in the same category. For example, in [Ex.1] speaker repairs a prepositional phrase "from green left to pink" to "from blue left to pink", It is rare that he/she just repairs a noun "green" to "blue".

In such self-repairs, a repetition of a word or a phrase often exists. In self-repairs which are intended to correct an error (such as [Ex.1]), words or phrases around the error may be repeated.

In [Ex.1], "from" and "left to pink" are repeated. In self-repairs which are intended to add some information to the item just mentioned, the item may be repeated as in [Ex.3].

> [Ex.3] *I want a fight, one way flight*
> (from (Shriberg, Bear, and Dowding 1992))

In this example a word "flight" is repeated.

A repetition is made with the same constituent or an item in the same category, such as "orange" with "apple".

There are four possible structures around an interruption of a self-repair with a repetition. Fig.2 shows them.

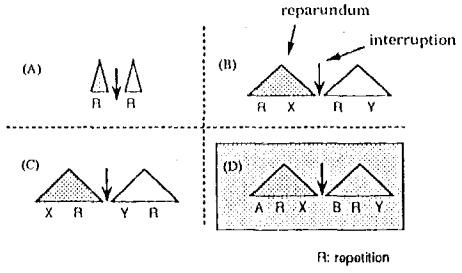

Fig. 2: Possible structures around interruption

A is a case of a simple repetition. B, C and D are cases in which some words exist between repetition. With cases B and C, positions of repetition directory indicate where an interruption occurs and which is a reparandum, but with D case, do not.

SERUP can cope with cases A, B and C.

with syntactic break

A self-repair comes with an interruption of utterance. Because an interruption may occur anywhere in an utterance (even within a word), self-repaired utterance can contain a syntactic break.

If this break can be detected, we can identify an interruption point.

same fragment repetition

When a speaker interrupts an utterance within a word, a fragment of the interrupted word is left. But he/she sometimes starts the repair with a word that begins with the same fragment as in [Ex.4].

[Ex.4] *ten, tenji tanntou no kata to*

This can be treated as A repetition, but to investigate a within-word interruption, we treated it as a separate category.

In this case, an interruption point is just after a repeated fragment. And if within-word interruptions are only made to repair an interrupted word, a reparandum can be identified as the repeated fragment.

with unknown word

Sometimes a fragment left can be detected as an unknown word. For example, if a word "ketueki(blood)" is interrupted and a fragment "ketue" is left, this fragment can be detected because there is no Japanese word "ketue".

In this case, an interruption point is just after an unknown word. And the reparandum can be determined if the same condition as the above case is sufficed.

with isolated word

A fragment left by a within-word interruption is not always detected as the same fragment repetition or an unknown word. For example, a fragment "hon" can be left when "hontou"(real) is interrupted, but this string can be a word meaning "book".

But such a word is always "isolated", that is, both two subtrees in fig3 fail.

In this case, an interruption point is just after an isolated word. And reparandum can be determined if the same condition as the above case is sufficed.

without repetition of a stem

Because Japanese inflectional morphology is complicated, speakers often make inflection errors. To repair such errors a speaker often starts a repair without repetition of a stem as in [Ex.5] not as in [Ex.6].

[Ex.5] *itada i, keru no ka*

[Ex.6] *itada i, itada keru no ka*

In these examples, "itada" is a stem and the speaker first tries to say "itada i ta" or "itada i te" and then changes to "itada keru".

In the case of [Ex.6], a repetition of a stem can be used as a clue. In the case of [Ex.5], existence of an affix without a stem indicates an interruption point and a reparandum.

fresh start

Fresh start is a repair with a completely different utterance. A fragment of utterance before interruption is ignored. SERUP tries the detection of fresh start if all possible clues are not found. It tries to parse the fragment of utterance without a first word of it. It repeats this trial until parsing succeeds.

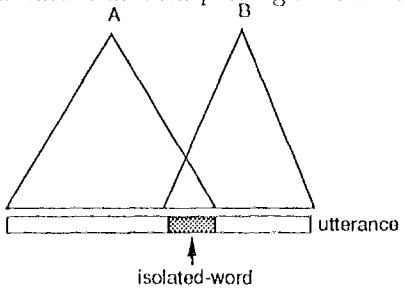

Fig. 3: An isolated word

others

SERUP cannot cope with utterances of all these categories.

changed to well-formed

A self-repaired utterance is occasionally parsed successfully as a well-formed utterance that has a meaning that the speaker does not intend. For example, in [Ex.7], a fragment "kyou" of a word "kyousan "(cosponsorship) is treated as a word "kyou "(today), and parsed successfully but the meaning of it is "cosponsor today".

[Ex.7] *kyou, kyousan suru*

Some of these utterances can be detected as an error in semantic interpreter. And we think prosodic cues can be used effectively, because a fragment "kyou" and a word "kyou " is pronounced differently. So far, SERUP cannot cope with such utterances, because it uses well-formed first method.

dividing word

In [Ex.8] the speaker starts repair within word.

[Ex.8] *junji, bi ni desu ne*

The speaker tries to say "junbi ni desu ne", but makes a lexical error "junji". He starts the repair with a fragment "bi" of "junbi", instead of a complete word "junbi". This is a very rare case.

repetition with different category

Speakers occasionally repair with different category of words. A human listener can draw some inference and find relation between words, but automatic detection is difficult.

ambiguous repair

In [Ex.9], it is ambiguous what kind of self-repair is made.

[Ex.9] *apointo wo, ni, er, suuzitu tyuu ni*

The speaker may repair a particle "wo" with "ni", or repair a fragment "ni" of a word "nisanniti" that has the same meaning of "suuzitu"(some days). We cannot solve this ambiguity automatically.

LARGE CORPUS-BASED ANALYSIS

To investigate effectiveness of SERUP we analyzed a large corpus called ADD(Ehara et al. 1990). ADD contains one million words of dialogues about registration to an international conference over telephone. ADD is created at ATR Interpreting Telephony Laboratories.

There are 1,082 self-repairs in the corpus. With these self-repairs, we investigate the categories they belong to. Table 1 shows the result.

DISCUSSION

In sum, SERUP seems to cope with 953(88.1%) of self-repairs. We think SERUP is effective to Japanese self-repaired utterances.

Most of utterances that SERUP cannot cope with are in the category "Changed to well-formed". As we mentioned, these utterances might be processed successfully with semantic constraints or prosodic cues. If we could implement them, SERUP would cope with 1,064(98.3%) self-repairs.

CONCLUDING REMARKS

We proposed SERUP, a parser coping with self-repaired Japanese utterances. SERUP uses some linguistic clues and translates a self-repaired utterance into well-formed version and parses it. The result of large corpus-based analysis suggests that 88.1% of 1,082 self-repairs can be processed by SERUP.

Our future directions are to test the system with large grammar and lexicon and to incorporate prosodic processing.

Table 1: The result of ananlsys

With repetition	A repetition	same constituent repetition	141(13.0%)
		same category repetition	108(10.0%)
	B repetition	same constituent repetition	96(8.9%)
		same category repetition	2(0.2%)
	C repetition	same constituent repetition	136(12.6%)
		same category repetition	3(0.3%)
	D repetition	*same constituent repetition*	4(0.4%)
		same category repetition	0(0%)
With syntactic break		Same fragment repetition	105(9.7%)
		With unknown word	98(9.1%)
		With isolated word	235(21.7%)
		Without repetition of a stem	23(2.1%)
		Fresh restart	6(0.6%)
Others		*Changed to well-formed*	111(10.3%)
		Dividing word	4(0.4%)
		Repetition with different category	5(0.5%)
		Ambiguous repair	4(0.4%)
Total of successable			953(88.1%)
Total			1,082

References

Blackmer, E. R. and J. L. Mitton (1991). Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition 39*, 173–194.

Dowding, J. el al. (1993). Gemini: A natural language system for spoken-language understanding. In *Proceedings of the 31st Annual Meeting of ACL.*

Ehara, E. et al. (1990). Contents of the ATR dialogue database. Technical Report TR-I-0186, ATR Interpreting Telephony Research Laboratories.

Hindle, D. (1983). Deterministic parsing of syntactic non-fluencies. In *Proceedings of the 21st Annual Conference of the ACL*, pp. 123–128.

Langer, H. (1990). Syntactic normalization of spontaneous speech. In *COLING 90*, pp. 180–183.

Levelt, W. J. M. (1988). *Speaking: From Intention to Articulation*, Chapter 12, pp. 458–499. Cambridge, MA: The MIT Press.

MADCOW (1992). Multi-site data collection for a spoken language corpus. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pp. 7–14.

Nakatani, K. and J. Hirschberg (1993). A speech-first model for repair detection and correction. In *Proceedings of the 31st Annual Meeting of ACL*, pp. 46–53.

Sagawa, Y., N. Ohnishi, and N. Sugie (1993). Repairing self-repairs in Japanese. In *Proceedings of Natural Language Processing Pacific Rim Symposium (NLPRS '93)*, Fukuoka, pp. 191–198.

Shriberg, E., J. Bear, and J. Dowding (1992). Automatic detection and correction of repairs in human-computer dialog. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pp. 419–424.