A SPOKEN LANGUAGE TRANSLATION SYSTEM : SL-TRANS2

Tsuyoshi MORIMOTO, Masami SUZUKI, Toshiyuki TAKEZAWA,
Gen'ichiro KIKUI, Masaaki NAGATA, Mutsuko TOMOKIYO

ATR Interpreting Telephony Research Laboratories
Seika-cho, Souraku-gun, Kyoto, 619-02, JAPAN

1. Introduction

An automatic telephone interpreting system will undoubtedly be used to overcome communication barriers between people speaking different languages. Recently, great interest has been growing in this area [Saitoh-88], [Waibel-91], [Kitano-91], [Roe-92]. SL-TRANS2*1) is an experimental system developed at ATR, which translates Japanese speech to English speech. It is composed of three basic components: speech recognition, translation and speech synthesis. This paper introduces the system with emphasis on the translation component. The discourse domain is a dialogue concerning an international conference registration. The distinctive features of the system are as follows.

(1) Japanese continuous speech input can be recognized with high accuracy. Moreover, speaker independent recognition using speaker adaptation technique has been developed.

(2) Various expressions peculiar to spoken language can be accepted and translated properly into the target language. In Japanese, the style of spoken sentences is generally quite different from that of written texts. Spoken utterances are fragmentary and include the speaker's intention directly or indirectly. The system extracts the intention and then transcribes it to a proper expression in the target language.

(3) Linguistic knowledge sources necessary for the translation are defined declaratively to the extent. Such definition improves high modularity, readability and easy maintenance of knowledge description.

In the next section, the system is overviewed and a brief description of the speech recognition mechanism is given. In the following three sections, distinctive technical features of each component of the translation system are described. Experiment results are shown in section 6.

*1) SL-TRANS2 is the extended version of SL-TRANS [Morimoto-90]

2. System Overview

A block diagram of the system is shown in Fig.1. Using a speaker adaptation technique, the

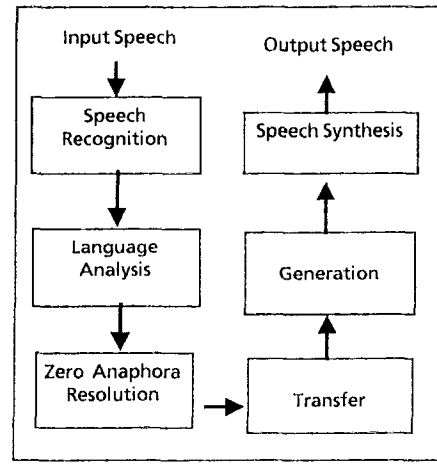

Fig.1 SL-TRANS2

speech recognizer can accept any speaker's speech. Phone level Hidden Markov Models (HMM) and syntactic rules for Japanese are defined in the recognizer [Kita-90]. By referring to these rules, the recognizer predicts the next possible phones and verifies their existence by matching them with corresponding HMMs. The process is invoked for many hypotheses in parallel until the end of the utterance is reached. Finally, the n-best sentential hypotheses are output, with their respective recognition scores. The output hypotheses are tagged with word information such as a part-of-speech label, then

the speech recognizer works as a kind of morphological analyzer for the following analysis component. These hypotheses are all well-formed syntactically, but not always semantically, pragmatically or contextually.

The next analysis component checks the validity of each hypothesis and selects the most plausible one*2). After analysis, some zero anaphora (noun phrase ellipses) are supplemented using pragmatics such as honorific expressions in Japanese. Then, the semantics of the sentence is output in the form of a feature structure. This feature structure is generally composed of two parts: an intentional content and a propositional content. The former indicates the speaker's intention and is expressed in terms of language-independent concepts. The latter is expressed in terms of language-dependent concepts. The subsequent transfer system transfers only the propositional content to their target language concepts. During the generation process, the two components are merged and a final surface expression in the target language is generated. Finally, synthesized English speech is output from the synthesizer. Currently, a commercial English language speech synthesizer is used in the system.

3. Analysis

3.1 Grammar Formalization

The grammar formalism adopted was originally based on HPSG (and its Japanese version JPSG) [Kogure-90]. In each lexical entry, syntactic constraints, semantic constraints and even pragmatic constraints are defined as a feature structure (Fig.2).

Parsing is basically accomplished using a unification operation between lexical items or successively derived constituents. This is effective in parsing Japanese spoken sentences which have a variety of expressions. According to the JPSG theory, a few principles (the head feature principle, the subcategorization feature principle, etc.) and one grammar rule (a mother is composed of a daughter and a head) are

```
(deflex-named 持つ-1 持 vstem
(<!m syn head grfs> = =
   [subj ?(subj [[sysn [[head [[pos N][form が]]]]
                        [sem ?subj-sem]])]
   [obj ?(obj [[sysn [[head [[pos N][form を]]]]
                        [sem ?obj-sem]])]
!(sc-s1-2 ?subj ?obj)
(<!m !sem> = = [[reln 持つ-1]
                        [agen ?subj-sem]
                        [obj ?obj-sem]])
(<!m !prag> = = ....)
```

Fig.2 Lexical Entry for Analysis

sufficient to cover all linguistic phenomena. However, naive implementation of the theory as a practical system brings an explosion of processing time and memory consumption, even for a simple sentence. To solve this problem, medium grained context free grammar (CFG) rules are introduced [Nagata-92]. The grammar rules are constructed to maintain declarative description of lexical constraints and also to suppress unnecessary unification execution. For instance, the concatenation conditions between Japanese verbs and auxiliaries are defined explicitly by the rules.

3.2 Parsing Algorithm

Parsing is guided by CFG rules and accomplished by the unification operation as described above. Generally, most of the processing time in a unification-based parsing method is consumed for the unification operation. In this system, besides dividing CFG rules as mentioned above, other efficiency improving technologies have been introduced. For instance, unification execution is delayed until all CFG rules have been applied. Another approach is to improve the unification procedure itself. In our system, an efficient unification mechanism using several techniques such as the quasi-destructive graph unification algorithm [Tomabechi-91] has been implemented. Using these improvements, this system can analyze an input utterance in a fairly short time.

3.3 Zero Anaphora Resolution

*2) The contextual evaluation function is not yet implemented in the current system

Some zero anaphora are resolved and supplemented using pragmatic information in Japanese [Dohsaka-90]. In general, pronouns indicating the participants such as "I" or "You" are seldom explicit in spoken Japanese. On the other hand, Japanese is abundant in honorific expressions, and such information can be used to interpret some zero pronouns. For instance, in the following example, the agent of the predicate "okuru" (send) in the last phrase can be inferred to be the speaker because he (she) is stating a promise and expressing it politely. Then, the indirect object of "okuru" is decided as the hearer, if the dialogue only concerns two participants.

tourokuyoushi-wo
*(a registration form-ACC*4)*
ookuri-itashimasu
(send-Polite/Promise)

Other zero anaphora, especially a subject, in a sentence without such information is interpreted as the speaker.

4. Transfer

Input to the transfer process is a feature structure indicating the semantics of an utterance. The feature structure is represented recursively using relationships. A relationship consists of a relationship name and its case roles. A relationship name is essentially a concept. In the analysis system, the surface illocutionary force type for the utterance is calculated from the surface expression. This is converted to an appropriate illocutionary force type by analyzing and taking into consideration the speaker's plan in that situation. In the current system, however, this conversion is done straight from the surface illocutionary force type because a contextual processing mechanism has not yet been integrated into the system. Typical illocutionary force types established for goal-oriented dialogues, such as those in the target domain, are shown in Table 1.

The transfer system transfers only the feature structure of the propositional content using a feature structure rewriting system [Hasegawa-90]. The rewriting system traverses an input feature structure and rewrites it according to a set of rewriting rules. There are

Table 1 Examples of Illocutionary Force Type

Type	Explanation
PHATIC	Phatic expression such as those to open or close dialogue (Hello, Thank you)
INFORM	Inform a hearer of some facts
REQUEST	Request a hearer to carry out some action (Please tell me···)
QUESTIONIF	Yes/No question
QUESTIONREF	WH question

many kinds of rules such as concept-to-concept transfer rules or structural transfer rules from Japanese-to-English; or even Japanese-to-Japanese paraphrasing rules which make transferring easier. A rule is defined declaratively and composed of three sub-specifications as in Fig.4: an environment condition, an input condition and an output specification. The environment condition is used

```
on  <reln> 持つ in :phase J-E
    in = [[reln 持つ]
          [agen ?agen]
          [obje ?object]
          ?rest]
    out = [[reln have]
           [agen ?agen]
           [obje ?object]
           ?rest]
```

Fig.3 Transfer Rule

to control the application of rules instead of encoding explicit relationships between rules; when some condition is given, only the rules satisfying it are applied. It could also be used to transfer the input properly based on a given context.

Another important problem in the transfer process is how to disambiguate polysemous words and how to choose a proper target concept. In this system, a thesaurus of concepts is defined and used for this purpose. This thesaurus is

implemented as a type system and referred to by related rules.

5. Generation

The basic method used in the generation system is also a unification algorithm. However, unlike the analysis system, each linguistic generation rule is defined for a comparatively large unit. This is because the variety of sentences to be generated is not as great as that in analysis, e.g. an idiomatic expression can be used in certain cases. A generation rule is defined as a phrase definition [Kikui-92]. A phrase definition is basically composed of three sub-specifications as shown in Fig.4: a syntactic phrase structure, syntactic constraints and semantics, and an application environment.

```
structure (S-TOP (S AUX (NP PRON) VP) SIGN)
annotation
    (S-TOP [[syn [[e-cat S-TOP]]]
            [sem [reln REQUEST]
            [agen ?agen *SP*]
            [recp *HR*]
            [obje ?action]]]]
    (S      [[syn [[e-cat S]
                   [e-inv +]
                   [e-slash -]]]]
    (AUX    [[syn [[e-cat AUX]
                   [e-lex "would"]]]]
    (NP     [[syn [[e-cat NP]]]]
    (PRON   [[syn [[e-cat PRON]
                   [e-lex "you"]]]
            [sem *SP*]])
    (VP     [[syn [[e-cat VP]
            [e-vform BSE]
            [e-subj [[sem ?agen]]]]]
            [sem ?action]])
    (SIGN   [[syn [e-cat SIGN]
            [e-lex "?"]]])
environment [ ]
```

Fig.4 Phrase Definition Rule

In principle, a phrase definition is equivalent to a syntactic derivation rule augmented with semantics, other linguistic constraints and environmental constraints. Generation is executed by activating related phrase definitions successively which can subsume the whole semantics of the input feature structure. The validity of a combination of phrase definitions is examined using the unification algorithm. Finally, a set of phrase definitions is determined and their combined syntactic structure is produced as a result. An environment description is not used in the current system, but will be used to generate a more appropriate expression in a given context.

6. Experiment

The SL-TRANS2 system as developed so far can treat utterances from about 600 vocabulary items. It runs on a UNIX-workstation such as SPARC-2. Average run time for one utterance is about one minute, half for speech recognition and half for translation. A preliminary experiment has been carried out on five dialogues, which include 101 sentences. The results are summarized in Table 2. Input speech material are specific speaker's utterances. Abilities of

Table 2 Experiment Result

	As a Component		Total System (SR+TR)
	SR	TR*3)	
Correct Output	86 (85%)	99 (98%)	85 (84%)
Incorrect Output	12 (12%)	2 (2%)	4 (4%)
No Output	3 (3%)	0 (0%)	12 (12%)

Number of sentences
(Percentage)

speech recognition (SR) and translation (TR) as a single component are about 85% and 98% respectively. Correctness of translation is evaluated whether the generated sentences are grammatically correct and semantically understandable (minor errors those involving determiners are ignored). We can see that the translation part has high translation ability as a

*3) The xperiment was carried out on string inputs

single component. Only two sentences indicated below fail to translate properly.

J1: 4-man-yen-desu
(40-thousand-yen-is)

E1: I am 40 thousand yen.

J2: touroku-youshi-wo
*(registration-form-ACC*4)*
okurimasu-node,
(send- CAU)
sore-wo goran-kudasai
(it-ACC see-please)

E2: Please see it since I send an announcement of the conference

In J1, the subject of the sentence is not uttered explicitly, because it is easily inferable as "registration-fee" from the context. The system, however, erroneously supplements "I" since no honorific expression appears in the sentence.

In Japanese sentence J2, a main clause appears later and a pronoun "sore" (it) referring to "touroku-youshi" (a registration-form) in the first clause is used. The system does not see this referential relationship, and so it fails to generate a proper English sentence.

As a total system, about 84% of the sample sentences are recognized and translated into English correctly. Some examples are shown below.

J3: 9-gatsu 27-nichi-ikou-no torikeshi-ni-taisuru haraimodoshi-wa dekimasen

E3: The refund to the cancellation after September 27th is not possible.

J4: dewa dareka-ga watashi-no kawari-ni sankasuru-koto-wa dekimasu-ka?

E4: Then, can someone attend instead of me?

Generally speaking, it is desirable that a translation system be able to detect erroneous speech recognition output. In our system, most of such failures are filtered out, but two sentences are translated into English. These undesirable outcomes are due to inadequacy of selectional restrictions used in the translation component, as indicated below.

[Input] J5:
kouzabangou-ni
(the bank-account-TP)
furikonde -kudasai
(transfer-please)

[Output from SR] J5':
kouzabangou-wo
(the bank-account-ACC)
furikonde -kudasai
(transfer-please)

[Output from TR] E5':
Please transfer the bank account.

*4)Meanings of symbols used here are;
ACC:Accusative, CAU:Cause, TP:To -Place

7. Conclusion

The main features of the translation component of SL-TRANS2 are described. The preliminary experiment has shown promising results. We are currently extending not only the vocabulary size from 600 up to 1500, but also the functionality of the system by improving several functions and introducing a contextual processing mechanism.

Reference

[Saitoh-88] Saito, H., Tomita, M. : "Parsing Noisy Sentences", Proc. of COLING-88, 1988

[Waibel-91] Waibel, A. et al. : "JANUS: a Speech-to-speech Translation System Using Connectionist and Symbolic Processing Strategies", Proc. of ICASSP-91, 1991

[Kitano-91] Kitano,H. : "ΦDM-Dialog", Computer, June,1991

[Roe-92] Roe, D.B. et al. : "Efficient Grammar Processing for a Spoken Language Translation System", Proc. of ICASSP-92, 1992

[Morimoto-90] Morimoto,T., Iida,H., Kurematsu,A., Shikano,K., Aizawa,T. : "Spoken Language Translation - Toward Realizing an Automatic Telephone Interpretation System", Proc. of Info Japan-90, Tokyo, 1990

[Kogure-90] Kogure,K., Hasegawa,T., Ogura,K. : "NADINE - An Experimental Dialogue Translation System from Japanese to English", Proc. of Info Japan-90, Tokyo, 1990

[Kita-90] Kita,K., Takezawa,T., Hosaka,J., Ehara,T., Morimoto,T. : "Continuous Speech Recognition Using Two-level LR Parsing", Proc. of ICSLP-90, 1990

[Hasegawa-90] Hasegawa,T. : "Rule Application Control Method in a Lexicon-driven Transfer Model of a Dialogue Translation System", Proc. of ECAI-90, 1990

[Dohsaka-90] Dohsaka,K. : "Identifying the Referents of Zero-Pronouns in Japanese Based on Pragmatic Constraint Interpretation", Proc. of ECAI-90, 1990

[Tomabechi-91] Tomabechi, H. : "Quasi-destructive Graph Unification", Proc. of ACL-91, 1991

[Nagata-92] Nagata,M. : "An Empirical Study on Rule Granularity and Unification Interleaving Toward an Efficient Unification-based Parsing System", Submitted to COLING-92

[Kikui-92] Kikui,G. : "Feature Structure based Semantic Head Driven Generation", Submitted to COLING-92