[*Proceedings of a Workshop on Machine Translation*, July 1990, UMIST]

Automatic Interpreting Telephony Research at ATR

Tsuyoshi MORIMOTO

ATR Interpreting Telephony Research Laboratories
Seika-cho, Souraku-gun, Kyoto 619-02, Japan
morimoto@atr-la.atr.co.jp

An automatic telephone interpretation system will transform a spoken dialogue from the speaker's language to the listener's automatically and simultaneously. It will undoubtedly be used to overcome language barriers and facilitate communication among the people of the world.

ATR Interpreting Telephony Research project was started in 1986. The objective is to promote basic research for developing an automatic telephone interpreting system. The project period is seven-years.

(l) Continuous Speech Recognition

A grammar-driven new continuous speech recognition mechanism, which is called HMM-LR, has been developed. It combines HMM phone model and generalized LR parsing mechanism.

A generalized LR parser can handle arbitrary context free grammar. Parsing is guided by an LR table which is created from pre-defined grammar, and proceeds left-to-right without backtracking.

Input spoken sentences are those which are uttered phrase by phrase (a phrase means Japanese Bunsetsu). Then, recognition of an utterance is performed for each phrase independently, and the output result is a sequence of candidates corresponding to each phrase (this output is called a "phrase lattice"). To do this, a phrasal grammar for Japanese is defined and converted to the LR table. The lexical items are embedded in the grammar rules, that is, terminal symbols are defined as a sequence of several phonemes.

The system determines which phones should be recognized next by referring the LR table, and then verifies their existence in the input signal by comparison with the corresponding HMM phone model.

If several phones are predicted, it verifies their existence and keeps all possible parsing trees. In this process, the probabilities for each partial parsing tree are calculated and only those with a higher probability are kept using beam search technique. The recognition process proceeds until the end of the phrase, and those which are grammatically accepted and having a higher probability score are output as the final candidates.

(2) Spoken Dialogue Translation

A dialogue sentence is composed of two parts, i.e. a propositional content part and an illocutionary force part. The former corresponds to the content of the sentence and the latter to the speaker's intention or attitude to the hearer such as honorifics. The latter is chiefly expressed in a final predicate phrase by attaching several auxiliary verbs or sentential final particles. The analysis process analyzes an input sentence and extracts semantic expressions for these two parts. The propositional part is described in terms of language dependent concepts, and the illocutionary force part in terms of language independent concepts. The transfer process converts only the propositional part to the target language concepts, and the generation process merges it with the illocutionary force part and then generates a surface expression. In a sense, this method can be considered to be intermediate between the transfer approach and the interlingual approach.

(3)Integration of Speech Recognition and Language Analysis

In spoken language processing, ambiguity of input data itself becomes a crucial problem. To resolve this ambiguity, some linguistic information should be used. On that time, it is necessary to use proper information at appropriate points to avoid an unnecessary increase in processing time. The proposed method is composed of three stages. In the speech recognition stage, syntactical knowledge for phrases is used as described above. Output from this stage is several hypotheses (candidates) for each phrase. In the next stage, Kakariuke dependency between phases is checked to filter out implausible candidates. At the final stage, sentence analysis, the most plausible sentence is selected by checking strict syntactico-semantic and pragmatic appropriateness or by evaluating the preference of sentence structure.

An experimental system called SL-TRANS has been implemented. It recognizes input Japanese speech, translates it to English and outputs synthesized English voice. Some experiment results are reported in the presentation.

Automatic Interpreting Telephony Research at ATR

Tsuyoshi MORIMOTO

ATR Interpreting Telephony Research Laboratories
Seika-cho, Souraku-gun, Kyoto 619-02, Japan
morimoto@atr-la.atr.co.jp

ATR Interpreting Telephony Research Project

(1) ProjectPlan

(2) Requirement for Spoken Language Interpretation

(3) Componential Technologies

- (a) Continuous speech recognition
- (b) Language translation
- (c) Integration of speech recognition and language analysis

(4) Experimental Japanese-English Spoken Language Translation System (SL-TRANS)

(5) Experiment Result and Evaluation

(6) Conclusion and Future Direction

Project Plan

(1) 7 Years Project Period (1986.4~1993.3)

(2)Promote Basic Research for Developing an Automatic Telephone Interpreting System

(3)Develop a Prototypical System

between Japanese and English
large vocabulary (around 1,000 or more)
goal-oriented task

Requirements

(1) Accurate Speech Recognition and High Quality Language Translation

(a) No existing of pre- or post-editing

(b) Mono-lingual Users

(2) Interpretation of Spoken Dialogue

(a) Ellipsis/anaphora resolution

(b)Interpreting speaker's intention

(3) Integration of Speech Recognition and Language Processing

(a)Ambiguity resolution by using proper linguistic information

(b)Optimum (efficient) interface

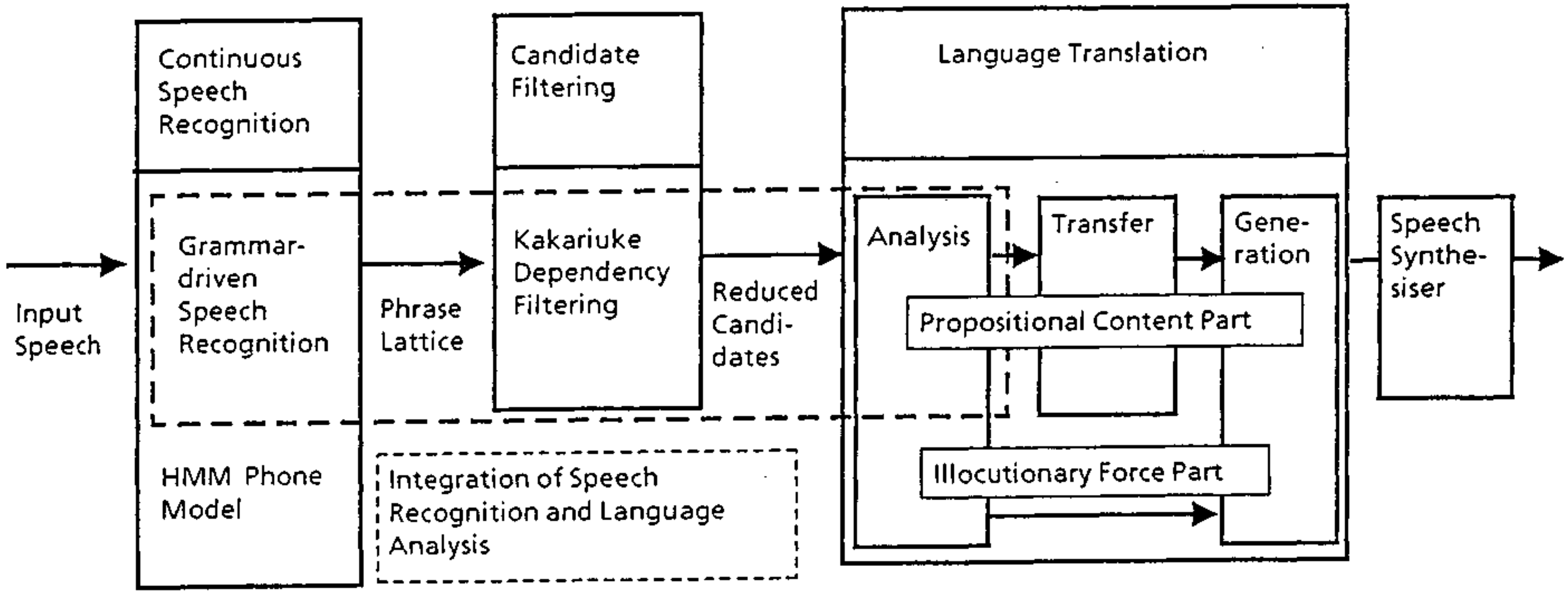


Figure 1 Block Diagram of the Speech Translation System

Continuous Speech Recognition

(1)HMM phone Model

(2)Grammar-Driven Speech Recognition
(a)Generalized LR Parser is used to predict next possible phones

(3) Grammar for Japanese Phrase (Bunsetsu) is Defined

(4) Input is sentence uttered phrase by phrase

Recognition is performed for each phrase.

(5) Verify by comparison with HMM

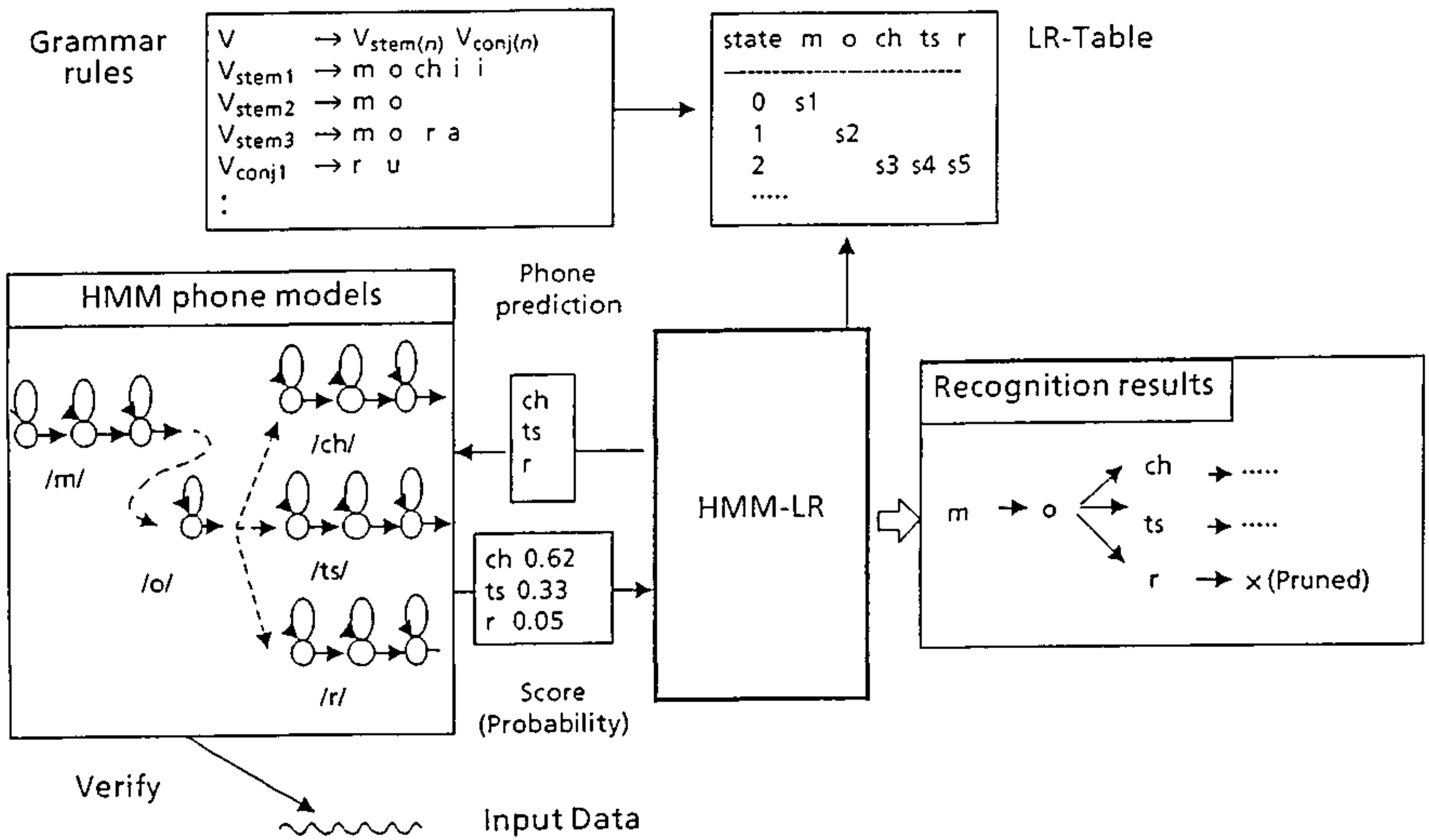


Figure 2 Basic Mechanism of HMM-LR

Language Analysis

(1) Lexicon-based JPSG Grammar and Unification-based Analysis Method

(a) To treat syntactic, semantic and even pragmatic constraints in uniform way

(b) To treat fragmental or complex modal expression in final predicate

intentions, honorifics, indirect-expressions etc.

(2) Ellipsis/Anaphora Resolution by Using Pragmatics

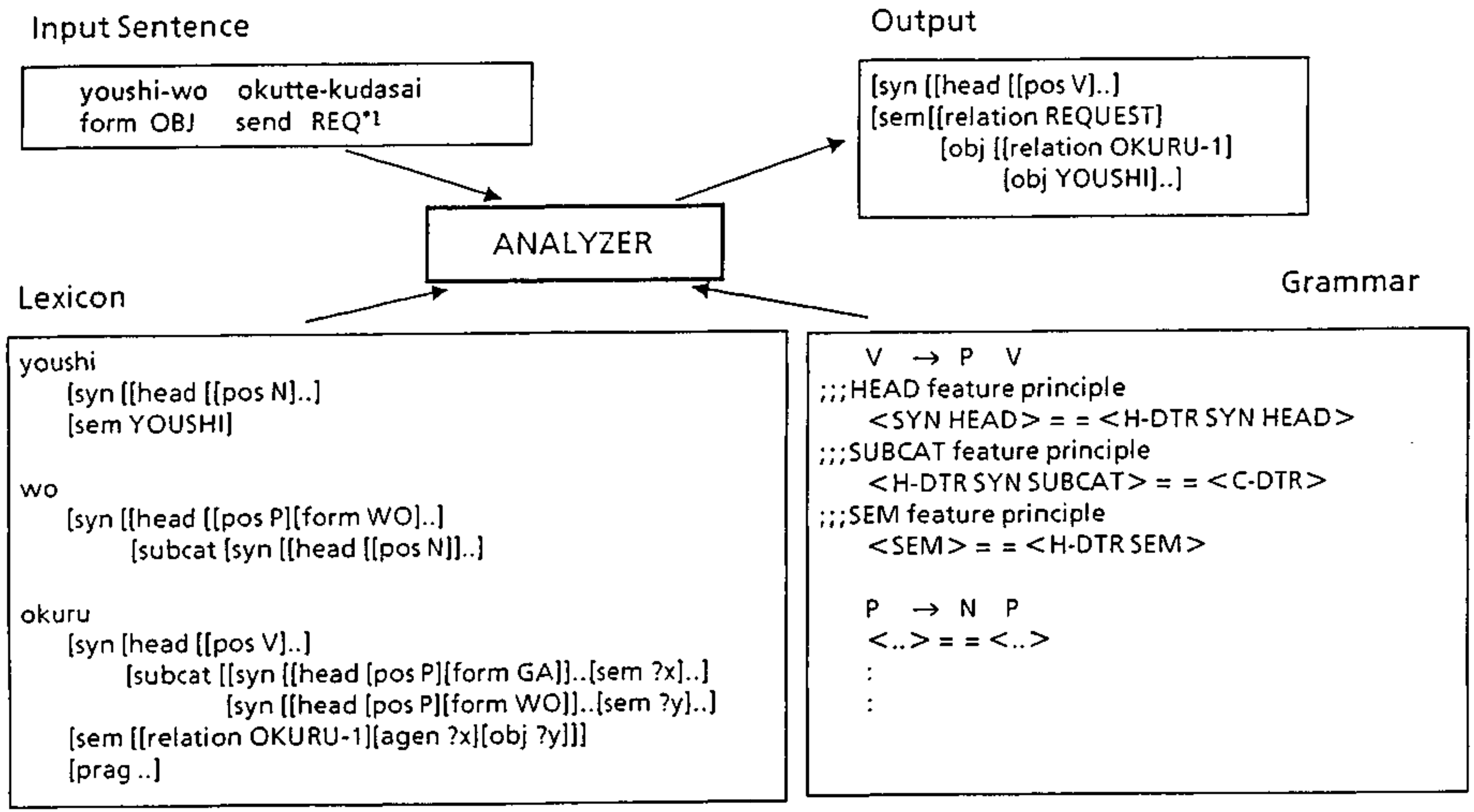


Figure 3 The Outline of the Analysis Process

Resolution of Zero-pronoun on Use of Pragmatics

Ex-1:

o-namae	wo	oshiete	itadake	masu	ka
(your) name		tell	could	POL	Q

(Could <u>you</u> tell <u>me</u> <u>your</u> name, please ?)

Table 1 Typical Illocutionary Force Type (IFT)

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

Semantic Representation in Feature Structure

```
Ex-2:   [[relation INFORM]
         [object [[relation MODERATE]
                  [object [[relation DESIRE]
                           [experiencer ?speaker]
                           [object [[relation SURU-1]
                                    [agent ?speaker]
                                    ...]
```

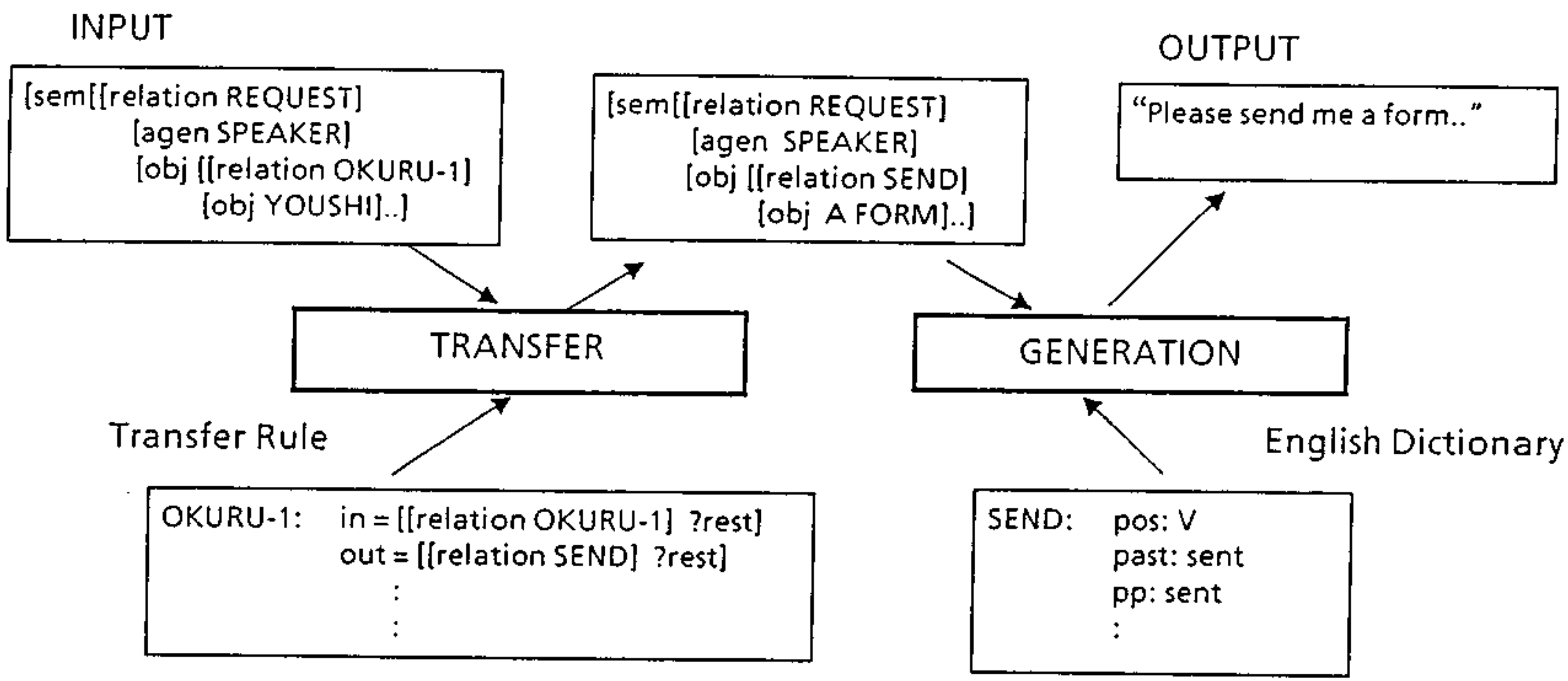


Figure 4 An Outline of the Transfer and Generation Process

Candidate Filtering Using Inter-phrase Kakariuke Dependency

(a) a predicate and its case-filler relationship

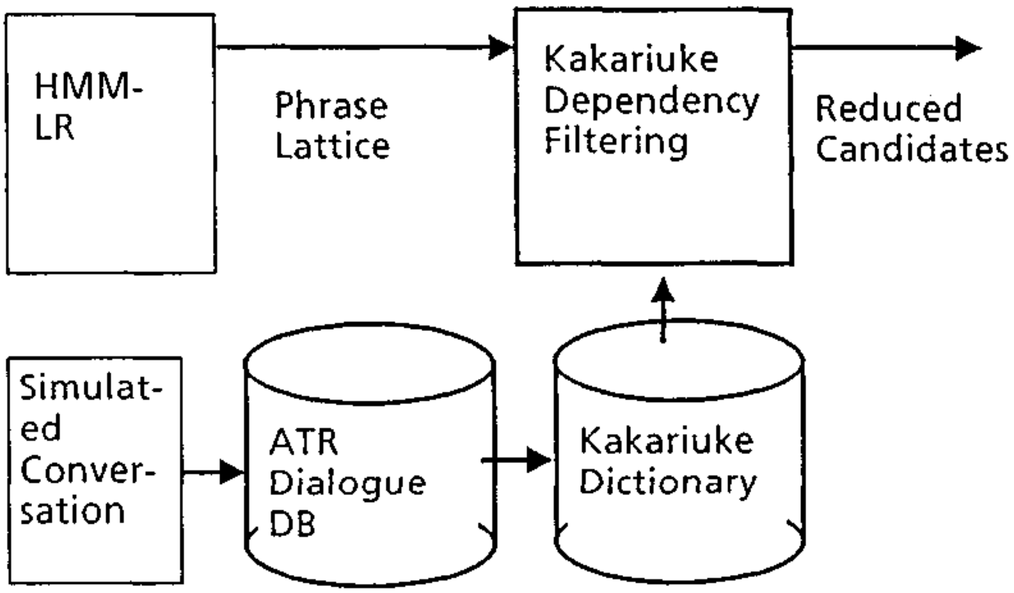
kaigi-ni	**sankasuru**
conference-OBJ2	attend

(b) modification relationship to a noun

kaigi-no	**sankahi**
conference of	registration-fee

(c) modification relationship to a predicate

sikyuu	**okuru**
immediately	send



$$K(X,Y) = F(X,Y) - wl \times D(X,Y) + w2 \times S(Y)$$

where

K(X,Y): matching score between phrase X and Y

F(X,Y): frequency of appearance in the Kakariuke dictionary

D(X,Y): distance between X and Y in the input data

S(Y): speech recognition score

wl,w2: weights determined experimentally

Ex-4:

Original Sentence;

tourokuyoushi-wa sudeni omochi-deshou-ka
registration-form-TOP already have - POL - Q
(Do you already have a registration form?)

Input (Output from HMM-LR);

tourokusi-ta sudeni omoi-mashou-ka
register-PAST already think-POL-INT-Q

tourokuyoushi-ga itsu-ni omochi-deshou-ka
registration-form-SBJ when have - POL - Q

tourokuyoushi-wa sen-ni omochisi-mashou-ka
registration form-TOP thousand-OBJ2 bring - POL-INT-Q

Output;

tourokuyoushi-wa sudeni omoti-deshou-ka

Sentence Preference During Language Analysis

(a) If several sentential candidates remain after Kakariuke filtering, their syntactico-semantical legitimacy is checked by the language analyzer.

(b) Nevertheless, if there are still several candidates, the sentence which has the highest preference score is selected.

$$P(X) = a1 \times S(X) - a2 \times Nt(X) - a3 \times Nu(X)$$

where

$P(X)$: *preference score of sentence* X

$S(X)$: *speech recognition score*

$Nt(X)$: *number of nodes of syntax tree*

$Nu(X)$: *number of unfilled obligatory elements*

$a1, a2, a3$: *weights determined experimentally*

Nt and Nu reflect the heuristics that a simpler sentence is more plausible.

Ex-5:

Input;

(1-1) **soredewa**	(2-1) **saremasu**
then	(someone) do (something) / (something) is done
(1-2) **sureba**	(2-2) **sitsureisimasu**
if (someone) do (something)	goodbye

Evaluation of Preference;

Of these, the combination of (1-1) and (2-2) has the fewest nodes and unfilled obligatory elements and thus, is selected.

Output;

soredewa sitsureisimasu

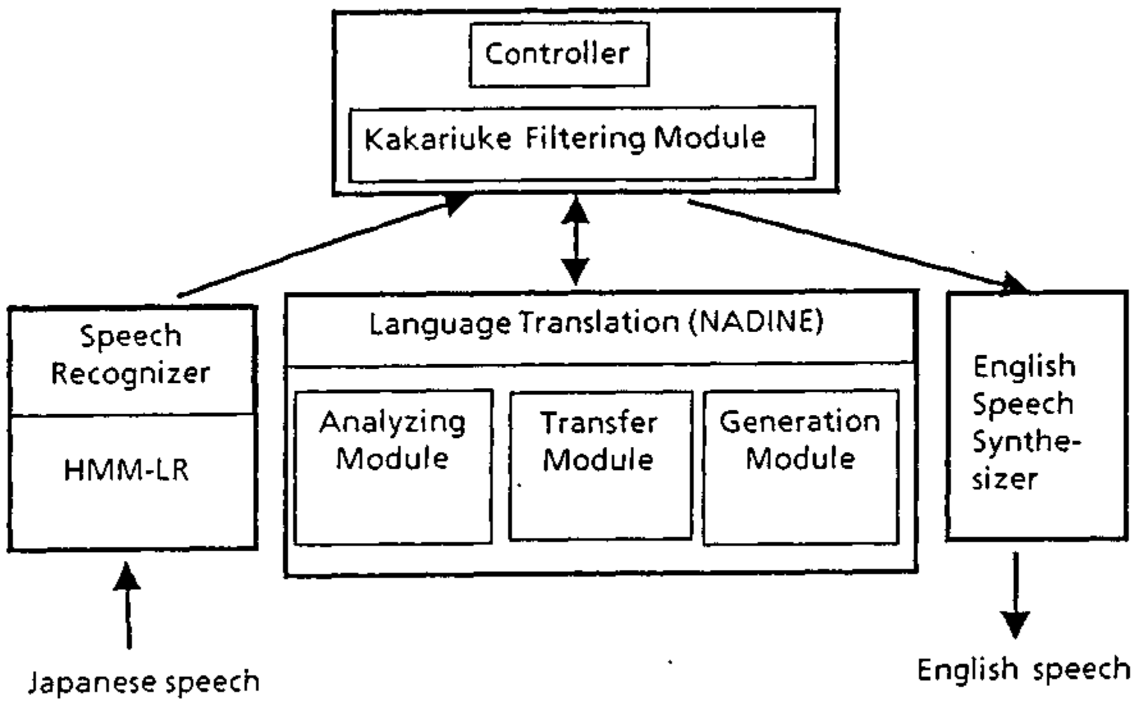


Figure 5 Configuration of SL-TRANS

Table 2 Experiment Results of SL-TRANS

Input	Specific speaker Number of sentences: 37 Number of phrases: 83 Average number of phrases/ sentences: 2.2
HMM-LR	Bunsetsu recognition rate 87% for the 1st rank ($0.87^{2.2}=0.74$) 96% for the top 5 ranks Average number of output candidates/phrases : 4.6 →number of sentential candidates: $4.6^{2.2}=28.7$
Kakariuke Filtering	Average number of selected candidates/ phrases: 1.5 →number of sentential candidates: $1.5^{2.2}=2.4$
Language Translation (System total)	Number of sentences selected correctly: 34 Number of sentences translated correctly: 34

Japanese utterance inputs	Translated English outputs
A: Moshi moshi Sochira wa kaigijimukyoku desu ka	A: Hello. Is that the office for the conference?
B: Hai Soudesu	B: Yes. That is right.
A: (Watashi wa) kaigi ni moushikomi tai nodesu ga	A: I would like to make a registration for the conference.
B: (Anata wa) touroku-youshi wa sudeni o-mochi des- you ka	B: Do you already have a registration form?
A: Iie mada desu	A: No. Not yet.
B: Wakarimasita Soredewa (watashi wa anata ni) touroku-youshi wa o-okuri itasi masu	B: I see. Then, I send you a registration form.
;;; where parentheses are missing phrases	;;; where underlined are supplemented words

Figure 6 Translation Examples

Conclusion and Future Direction

(1) Enhancing the Efficiency of Speech Recognition
  (a) Introducing more precise HMM model
    multiple code-book
    context-dependent model

  (b) Introducing Sentential Grammar in Speech Recognition
  (c) Introducing Stochastic Grammar in Speech Recognition

(2) Integrating Kakariuke dependency filtering (or some semantic based candidate filtering) function with either a speech recognizer or language analyzer rather than making it as an independent process.

(3) Improving the efficiency of the language analyzer.

(4) Extending translation both in vocabulary size and sentence expression variation.

(5) Introducing a contextual processing for the sake of enhancing the ability of:
  (a) Ellipsis/anaphola resolution.
  (b) Speech recognition ambiguity resolution.