SPEECH UNDERSTANDING DRIVEN BY CONCEPTUAL PROCESSING

Masao YOKOTA, Mariko ODA † and Seio ODA
Fukuoka Institute of Technology, Fukuoka, Japan
† Kurume Institute of Technology, Kurume, Japan

ABSTRACT

The authors have been constructing a speech understanding system IMAGES-S that can infer the conceptual information which the speaker would transmit. The processing for this purpose belongs no longer to wave signal processing but to natural language understanding, especially to conceptual processing with background knowledge such as commonsense, world-specific knowledge, etc. And moreover, understanding incompletely perceived speech is nearly equal to estimating the concepts of the words omitted in texts.

MODEL OF SPEECH UNDERSTANDING

Assume that one person "M_1" transmits his conceptual information "c" to the other person "M_2" acoustically in a language. The acoustic expression "r" of "c" which M_1 selects among the various paraphrases that he/she could generate is probably perceived by M_2 as a set of acoustic expressions "R_2" because of M_1's misstating or M_2's mishearing, or the noises during its propagation. Futhermore, each element of R_2 is interpreted as a set of conceptual information which in turn is merged into the total set "C_2", that is , the interpretation of R_2. These can be formalized as (1)-(3) below:

$$r \in \Phi_1(c) = R_1 = \{r_{11}, ..., r_{1l}\} \quad (1)$$

$$R_2 = \Delta_{12}(r) = \{r_{21}, ..., r_{2m}\} \quad (2)$$

$$C_2 = \cup_{i=1}\Phi_2^{-1}(r_{2i}) = \{c_{21}, ..., c_{2n}\} \quad (3)$$

where

Φ_i : M_i's acoustic verbalization process of conceptual information,

Φ_i^{-1} : M_i's interpretation process of acoustic expression,

and

Δ_{ij} : the deformation process of acoustic expression in the environment of M_i and M_j.

The ideal speech recognition in M_2 will easily find "r" in R_2 because even the case $R_2 = \{r\}$ may happen. However, this is very difficult or almost impossible when the environment of the speaker M_1 and the hearer M_2 is not perfect, where "perfect" means "free from either mistakes or noises". Therefore, actually, M_2 is to select some "c'" among C_2 as would be "c" using background knowledge.

IMAGES-S simulates this process instead of the hearer M_2. That is, if the conceptual content "c'" resulting from understanding is reasonable, or not inconsistent with background knowledge, the system deems it as what the speaker would mean, and moreover, "r'", one of its verbalization "$\Phi_2(c')$", as what he would speak, where of course "r'" is not always equal to "r".

For an extreme example, IMAGES-S may transform such a dialogue between two persons as {"Where ?" " Bath."} into a more sophisticated one {"Where are you going ?" "I'm going to the public bath."}.

IMAGES-S consits of three modules: 1) Speech recognition (SRM), 2) Language understanding (LUM), and 3) Task realization (TRM). SRM transforms acoustic signal waves into word-lattices. LUM analyzes them syntactically and semantically and generates meaning representations, employing background knowledge. Finally, TRM realizes the tasks required by the speaker. Here is assumed that the task is limited to dictation.

CONCEPTUAL PROCESSING

LUM, utilizing the background knowledge K_B, estimates the concepts of the words unrecognized in SRM and such an inference process can be formalized as (4).

$$I(P[x_1, ..., x_n] \wedge K_B \vdash I(P[p_1, ..., p_n]) \quad (4)$$

where

$P[.]$: incompletely recognized speech,

x_i : word-sequence not recognized or recognized with a very low likelihood,

and

p_i : estimated word-sequence.

The inference process succeeds when $I(P[.])$ is unified with background knowledge K_B, which superficially, results in substitutions θ_s in (5).

$$I(p) \wedge K_B \vdash I(P\theta), \theta = \{x_1/p_1, ..., x_n/p_n\} \quad (5)$$

The total process is formalized as (6)-(8) below:

$$h(P, K_B) = H \quad (6)$$

$$H = \{P[x_1, ..., x_n]\theta | \textit{hypothrtical restored word-sequence}\} = \{H_1, ..., H_m\} \quad (7)$$

$$e(H) = H' \quad (8)$$

where

h : hypothsis generating function,

H : a set of hypotheses,

e : adequacy evaluating function,

and

H' : ordered H according to a certain preference.

At present, the prefernce order is determined according to the hypothesis as follows : "What is most easily understandable is the best understanding result.". This determination is realized by calculating the complexity of understanding. The representation of knowledge or speech contents in our system is based on the first-order predicate logic and the complexity is deemed as the total cost (C_t) of translation from a surface structure(i.e. sentence) into a conceptual structure(i.e. logical formula). The authors have found C_t given nearly by the equation (9) which approximates the total times of variable unification, predicate insertion, etc. occurring through the translation process.

$$C_t = 2N_0 + W + E \quad (9)$$

where

N_0 : the number of the words recognized in SRM ,

W : the total number of the words inferred in LUM,

and

E : the number of the words representing objects or events inferred in LUM.

The understanding result with the least C_t is determined as the best.

Assume that EX-7 below is one of the word-sequences generated from the word-lattice put out by SRM. LUM ,using the background knowledge in Table 1, understands it and TRM generates such sentences (S1.1)-(S2.5). The underlined parts are the inferred and inserted words.

The preference order among the sentences is shown by P in Table 2, which implies that S1.1 is the best and that S2.3 and S2.4 are the worst.

EX-7 父親·X_1·自動車·X_2·学校·X_3·通勤.
(= Father·X_1·automobile·X_2·school·X_3·commute)

S1.1 父親が自動車学校に通勤.
(= Father commutes to the automobile school.)

S1.2 父親が経営する自動車学校に誰かが通勤.
(= Someone commutes to the automobile school owned by Father.)

S2.1 父親が自動車で学校に通勤.
(= Father commutes to the school by automobile.)

S2.2 父親の所有する自動車で学校に誰かが通勤.
(= Someone commutes to the school by the automobile owned by Father.)

S2.3 父親の経営する自動車について教育する学校に誰かが通勤.
(= Someone commutes to the school for automobile education which is owned by Father.)

S2.4 父親の所有する自動車のある学校に誰かが通勤.
(= Someone commutes to the school where the automobile owned by Father is.)

S2.5 父親が自動車について教育する学校に通勤.
(= Father commutes to the school for automobile education.)

CONCLUSION

The modules LUM and TRM are almost equal to IMAGES-II[1, 2], that is, almost completed. The simulation of these modules has proved the validity of IMAGES-S. In near future, C_t will be improved in order to reflect coherence and cohesion in context. The problem left unsolved is the connection of SRM and LUM. The module SRM will be realized by employing Hidden Markov Models (HMMs).

References

[1] Yokota,M. et al (1984), Language-picture question-answering through common semantic representation and its application to the world of weather report, in (Bolc,L. ed.) Natural Language Communication..., Springer-Verlag

[2] Yokota,M. et al (1991), "On Natural language understanding system, IMAGES-II", IEICE Japan

[3] Flanagan,J.L. (1994), Technologies for multimedia communications, Proc. of IEEE, 82-4

Table 1: A part of background Knowledge (word-meanings)*

word	word-meaning = [concept : unifying operations ;]
通勤	通勤 $(x) \Leftrightarrow L(z, y_0, y_0, y_1, A_p) \wedge \ldots \wedge$ 通勤$^+(x) \wedge$ 人間 $(y_0) \wedge$ 施設 $(y_1) \wedge$ 手段 (z):ARG(dep(が),y_0),ARG(dep(に),y_1),ARG(dep(で),z),...;
学校	学校 $(x) \Leftrightarrow$ 施設 $(x) \wedge$ 教育$^{++}(y, x, \ldots) \wedge \ldots$: ;
父親	父親 $(x) \Leftrightarrow$ 男 $(x) \wedge$ 親 (x): ;
自動車	自動車 $(x) \Leftrightarrow L(\phi, x, p, q, A_p) \sqcap L(x, y, p, q, A_p) \wedge p \neq q \wedge$ 自動車$^+(x) \wedge$ 物 $(y) \wedge$ 所有 $(z_1) \wedge$ 教育$^{++}(z_2, \ldots, x, \ldots) \wedge \ldots$: ;
教育	教育 $(x) \Leftrightarrow$ 教育$^+(x) \wedge$ 教育$^{++}(x, y, z, \ldots) \wedge$ 施設 $(y) \wedge$ 事物 (z) ; ARG(dep(が),y),ARG(dep(について),z),... ;

*$\sqcap$: "simultaneously AND", $L(X, Y, U, V, A_p)$: "Y moves from U to V by X", ϕ: "don't care", 男 : "male", 親 : "parent", 施設 : "institution", 物 : "object", 事物 : "object or event", ARG(X, Y) : "unify X with Y".

Table 2: Evaluation of understanding results *

I.D.	θ	N_0	W	E	C_T	P
S1.1	{X_1/が, X_2/ϵ,X_3/に}	3	2	0	8	1
S1.2	{X_1/が経営する, X_2/ϵ,X_3/に誰かが}	3	5	2	13	3
S2.1	{X_1/が, X_2/で, X_3/に}	4	3	0	11	2
S2.2	{X_1/の所有する, X_2/で, X_3/に誰かが}	4	6	2	16	5
S2.3	{X_1/の経営する, X_2/について教育する, X_3/に誰かが}	4	7	3	18	6
S2.4	{X_1/の所有する, X_2/のある, X_3/に誰かが}	4	7	3	18	6
S2.5	{X_1/が, X_2/について教育する, X_3/に}	4	4	1	13	3

*ϵ means empty.