A MASSIVELY PARALLEL MODEL OF SPEECH-TO-SPEECH
DIALOG TRANSLATION:
A STEP TOWARD INTERPRETING TELEPHONY

Hiroaki Kitano, Hiroyuki Tomabechi, Teruko Mitamura, and Hitoshi Iida
Center for Machine Translation
Carnegie Mellon University
Pittsburgh, Pennsylvania, 15213 U.S.A.

Abstract
This paper describes the overall picture of \( \Phi \)DMDialog. \( \Phi \)DMDialog is a real-time Japanese-English speech-to-speech dialog translation system that accepts speaker-independent continuous speech inputs. The scientific focus of the project is to model the cognitive process of simultaneous interpreters. As a result, the architecture of the system is very different from machine translation systems. Our model assumes hybridized parallelism as a basic computation mechanism, and the process of translation is highly interactive due to the dynamic participation of knowledge from morphophonetic-level to discourse-level. An almost concurrent parsing and generation scheme provides a simultaneous interpretation capability which is essential to interpreting telephony. \( \Phi \)DMDialog has been publicly demonstrated at the Center for Machine Translation at Carnegie Mellon University.

1 Introduction
We propose a radically new architecture for a speech-to-speech translation system. Traditional approaches for speech-to-speech translation systems (e.g. [17]) have been put together modular components: a speech recognition device, a parser, a generator and a speech synthesizer. The idea behind them has been to extend machine translation (MT) systems to handle speech input and output devices. We claim that the simple extension of MT systems for a speech translation system will not suffice the requirements that emerge when translating spoken dialogues. As an alternative to the traditional model, we propose a model which intends to simulate simultaneous interpretation. Major technical features are: (1) a parallel algorithm of parsing and generation, (2) almost concurrent incremental parsing and generation, and (3) highly interactive processing based on dynamic participation of knowledge from morphophonetics to discourse, and (4) cost-based ambiguity resolution. The model takes into account the major psycholinguistic and cognitive theories[1] [2] [3] [4] [12] [14] [15] [18] [19]. In this paper, we will describe the basic organization and high light features of the system[1].

2 A Basic Organization
Natural language processing can be viewed from two aspects: information-based processing and hypothesis selection. The information-based process is a mapping process between a surface string and a memory network. Due to the ambiguity of the natural language, multiple hypotheses for interpretation of utterances are inevitable. The hypothesis selection process chooses the one most plausible hypothesis among multiple candidates. In our model, a parallel marker-passing scheme and a connectionist network are hybridized in order to capture these two aspects of natural language processing. Computations are performed in a memory network which represents knowledge from morphophonetic-level to discourse-level and world knowledge. Each node has labelled links and weighted links. Labelled links have labels such as IS-A and PART-OF, and they are used for passing markers. Weighted links are used for connectionist type spreading activation.

There are three types of nodes:
- Concept Sequence Class (CSC): Composed of a list or a set of elements to capture complex relations between discourse entities and some sequential phenomena. Phonological sequences, concept sequences, discourse plan sequences are represented by CSC.
- Concept Class (CC): Represent a set or a family of concepts such as phonemes, concepts and discourse plans.
- Concept Instance (CI): Represents discourse entities and instances of utterances.

The marker-passing division of our model captures information-based aspects such as the construction of meaning representation, the propugation of features and constraints, and the symbolic operations involved[2]. Types of markers are:
- Activation Markers (A-Markers): Carry a pointer to a discourse entity, cost, and some linguistic and semantic features. Only pass up through IS-A links.
- Prediction Markers (P-Markers): Located on elements of CSC and containing a binding-list of role-instance pairs, costs, and constraints.
- Generation Markers (G-Markers): Similar to A-Markers but carry surface string or phonological realization in addition to other features.
- Verbalization Markers (V-Markers): Located on elements of CSCs of the target language to record what was
verbalized.

Contextual Markers (C-Markers): Represent contextual priming and are used as an alternative to the connectionist type of priming.

Figure-1 illustrates the movement of P-Markers on a CSC. In (a), a P-Marker (initially located on e₀) is hit by an A-Marker and moved to the next element. In (b), two P-Markers are used and moved to e₂ and e₃. This dual prediction is used for phonological processing.

\[
P < e₁ e₂ e₃ \cdots eₙ > \Rightarrow < e₁ e₂ e₃ \cdots eₙ >
\]

(a) Simple Prediction

\[
P < e₁ e₂ e₃ \cdots eₙ > \Rightarrow < e₁ e₂ e₃ \cdots eₙ >
\]

(b) Dual Prediction

Figure 1: Movement of P-Markers

The connectionist network is formed by weighted links in the memory network. A computation is performed using a competitive activation and inhibition process. As a result of the process, one hypothesis out of multiple hypotheses is selected. The connectionist network can be also applied to speech recognition.

3 Almost Concurrent Parsing and Generation

Simultaneous interpreters start translation even before the end of the sentence. This is especially true when an utterance to be translated is a long sentence with multiple subordinate clauses. We are hypothesizing that parsing and generation are conducted almost concurrently and incrementally in simultaneous interpretation. This assumption is psycholinguistically attractive since the incremental nature of human language processing is confirmed. At parsing, it is intuitively evident, and phenomena such as garden paths support the hypothesis. In generation, studies on spontaneous speech and speech error support the incremental hypothesis[2]. Also, the study of syntactic planning has revealed that the next fragment of the sentence is planned during production of the current fragment, providing independent support for the hypothesis of concurrent and incremental generation.

In parsing with DMDIALOG, A-Markers and P-Markers are used to analyze meanings of the utterance. Phoneme-level nodes are activated by a speech input and trigger A-Marker passing. A-Markers are passed up through IS-A links and each A-Marker hits an element of the phoneme sequence. It, then, collides with a P-Marker and the P-Marker is moved using the dual prediction. When a sequence is accepted, the lexicon is recognized and an A-Marker which carries features of the referred concept is passed up. The A-Marker hits an element of the CSC which captures syntactic-semantic knowledge. Then the simple prediction is performed. Such processing continues up to the discourse-level which captures discourse plan of each speaker. During this process, the memory network is modified according to the meaning and the effects of the utterance.

G-Markers are created whenever A-Markers find the CC which has lexical realization in the target language. In figure-2(a), a word in the source language LEX₁ₐₗ is linked to CC₁ which has a lexical realization in the target language LEX₁ₗ. When LEX₁ₗ do not exist, LEX₂ₗ is used. Figure-2(b) shows that when the concept, expressed as either LEX₁ₐₗ or CSCₐₗ, cannot be expressed in one word in the target language, CSCₗ is deployed to translate using phrases. This scheme guarantees that the most specific expression (either a word or a phrase) in the target language will be selected, even if the word directly corresponding to the word in the source language is not found. The lexical realization or phonological realization is carried by the G-Marker. V-Markers are placed on every first element of CSCs which represent syntactic-semantic knowledge of the target language. When a G-Marker meets a V-Marker and the role of the discourse entity carried by the G-Marker is unambiguous at the moment, verbalization is conducted; otherwise verbalization is halted until the role is made unambiguous. The part of the sentence which is verbalized is recorded in a V-Marker and the V-Marker is moved to the next possible verbalization elements. This will avoid redundant verbalization. Only the element with a V-Marker can be verbalized in order to ensure the consistency of the produced sentence. By this procedure, a concurrent and an incremental parsing and generation is attained. Due to the concurrency of the algorithm the syntactic planning of the next fragment is conducted even during the verbalization of the current fragment, which is consistent with psycholinguistic studies[13]. Apparently, however, this algorithm does not ensure that translation is always made before the whole sentence is parsed. However, the algorithm is effective for utterances composed of multiple subordinate clauses. It is important to notice that the syntactic-semantic-level knowledge and the discourse-level knowledge are integrated, whereas most syntactic theories assume the top-most node as 'S' and the discourse knowledge and syntactic knowledge are isolated from each other. Thus, in translation of long sentences with multiple clauses, knowledge that is on the border of syntactic-semantics and discourse plays an important role.

\[^{3}\]In the case of a short Japanese utterance, verbalization is likely to be halted until the end of the sentence due to the verb-final structure of the Japanese. Even human simultaneous interpreters wait until the end of the sentence on such occasions.

\[^{4}\]Unlike an incremental generation by IPG[6], which assigns a procedure to each syntactic category, our algorithm uses markers to carry information. Also, concepts to be expressed are incrementally determined as parsing progresses. For details, refer [7].
4 Discourse Knowledge

The role of discourse level processing is particularly important in the simultaneous interpretation system because it enables incremental translation of input utterances and helps in resolving ambiguities. Whatever is done to handle these different levels of ambiguities in a uniform manner. This has not been attained in past models of ambiguity resolution. During the parsing process, costs are added when: (1) substitution, deletion and insertion of phemes are performed to activate certain lexical items from noisy speech inputs, (2) new CIs are created, (3) CIs without contextual priming are used for further processing, and (4) the memory network is modified to satisfy constraints. Costs are subtracted when: (1) prediction has been made from discourse knowledge, and (2) CIs with contextual priming have been used.

4.1 Discourse Plan

We use hierarchical discourse plan sequences, represented by CSCs, to recognize and predict possible next utterances. Plan hierarchies are organized for each participant of the dialog in order to capture complex dialog which often takes place in mixed-initiative dialog. This is one of the major differences of our model from other discourse models. Each element of the plan sequence represents a domain-specific instance of a plan or an utterance type (a discourse case) which can be dynamically derived from abstract dialog knowledge and domain knowledge. When an element of the plan sequence is activated at the abstract-level, the rest of the elements of the plan sequence have constraints imposed on them which are derived from the information given to the activated elements. This ensures coherence of the discourse. However, specific plan sequences representing the discourse cases are already indexed in the memory as a result of instantiating abstract knowledge based on past cases of discourse. Therefore, these plan sequences are domain-specific knowledge. When such a plan sequence is activated, it simply predicts the next plan elements because these specific plan sequences are regarded as records of past cases, and, thus, most constraints are already imposed and the sequence is indexed according to the specific constraints.

4.2 Discourse Entities and Their Relations

Discourse entities and their relations are established as a result of understanding the utterance in the dialog. In our model, understanding is regarded as the activity of memory network modification. The memory network is modified by creation of a new instance in the network, and by creation and deletion of links between nodes. CIs are created when new discourse entities are introduced and are referred to later when the same discourse entity is referred to in subsequent utterances. One other type of CI represents the meaning of the specific utterance. Such CIs are recorded in the memory network as cases of utterance in the specific discourse context. In whatever is done to handle these different levels of ambiguities in a uniform manner. This has not been attained in past models of ambiguity resolution. During the parsing process, costs are added when: (1) substitution, deletion and insertion of phemes are performed to activate certain lexical items from noisy speech inputs, (2) new CIs are created, (3) CIs without contextual priming are used for further processing, and (4) the memory network is modified to satisfy constraints. Costs are subtracted when: (1) prediction has been made from discourse knowledge, and (2) CIs with contextual priming have been used.

Cost-based Ambiguity Resolution

A cost-based scheme of ambiguity resolution has been adopted in Whatever is done to handle these different levels of ambiguities in a uniform manner. This has not been attained in past models of ambiguity resolution. During the parsing process, costs are added when: (1) substitution, deletion and insertion of phemes are performed to activate certain lexical items from noisy speech inputs, (2) new CIs are created, (3) CIs without contextual priming are used for further processing, and (4) the memory network is modified to satisfy constraints. Costs are subtracted when: (1) prediction has been made from discourse knowledge, and (2) CIs with contextual priming have been used.

Costs for noisy speech inputs are assigned based on the previous course of discourse because previous utterances affect costs assigned to each hypothesis. Major factors include the state of the memory network modified as a result of understanding previous utterances (this modification reflects what the hearer knows and infers about the discourse world from the utterance), contextual priming, and predictions from discourse plans which assign negative cost values to predicted hypotheses as pre/.

Phonological Processing

Costs for noisy speech inputs are assigned based on the phonological distance of substituted phonemes, the cost of deletion and insertion of extra phonemes, and prefer-

---

5 Suppose the correct phoneme sequence is /m a z w/, /m o z w/ has a substitution between /a/ and /o/, /m a z w/ has a deletion of /a/, and /m o z w/ has an extra phoneme /o/.

6 This means that order-strict or order-free constraints apply for determining the order of the plan sequence.

---
ence based on the context of the discourse. (This preference is a factor which decreases costs, whereas other factors increase costs.) The phonological distances are determined by distinct features and actual experiments using the speech recognition device. In order to keep beam width within a certain size, hypotheses whose cost exceeds the threshold will be deactivated.

Reference to the Discourse Entity
Whenever a discourse entity[19] is referred to in the utterance, a CI that represents the discourse entity is searched for. When a lexical node activates any CC node, a CI node found, the reference succeeds and no cost is added. Failure to find the CI (reference failure) will result in added cost with the creation of a new CI representing the discourse entity. This is based on the psycholinguistic study of [1]. Anaphoric reference is handled within this scheme.

Contextual Priming
Effects of contextual priming are well known[14][18]. No cost is added to hypotheses that use contextually primed concepts (either by C-Marker passing or by competitive activation[16] [9]). Hypotheses using concepts without priming will be assigned added cost depending on the level of priming the concepts have. (Of course, less priming equals more cost.)

Constraints
Constraints describe presuppositions that need to be satisfied to accept an interpretation of the utterance. If constraints are already satisfied when parsing the utterance, hypotheses that only require these constraints do not incur costs. On the other hand, costs are added to hypotheses which do not satisfy constraints. Such cost operation reflects the principle of a priori plausibility [1].

Lexical Preference
The lexical preference [3] is incorporated as a bias; term in the equation.

Generation
In the generation process, there are cases where multiple hypotheses are activated for one utterance. On such occasions, the hypothesis with the least cost will be selected. The cost in the generation phase will be added when the hypothesis uses lexical entries without C-Markers. Pragmatic considerations [5] are other factors in the cost calculation.

6 Conclusion
We described a model of speech-to-speech dialog translation. The model is intended to simulate the simultaneous interpreter's cognitive process. Several psycholinguistic studies were taken into account in designing the framework of the model. The almost concurrent parsing and generation scheme is one example of this scheme, which was derived from psycholinguistically realistic processes. The architecture proposed is new and different from that of any other existing model. We believe that our model can serve as a basis for solid discussion about the development of interpreting telephony.

Implementation
The model described has been implemented as DDMIALOG, developed at the Center for Machine Translation at Carnegie Mellon University, using CMU-CommonLisp and the Mach operating system on the IBM RT-PC workstation. Speech recognition and synthesis devices are Matsushita Institute's Japanese speech recognition and synthesis devices and DECTalk.

References