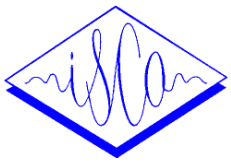

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

The Making of a Speech-to-Speech Translation System: Some Findings from the ΦDMDIALOG Project*

Hiroaki Kitano

Center for Machine Translation
Carnegie Mellon University
Pittsburgh, Pennsylvania, 15213 U.S.A.

NEC Corporation
2-11-5 Shibaura, Minato-ku
Tokyo, 108 Japan

Abstract

In this paper, we discuss some important problems which we encountered in developing the ΦDMDIALOG real-time speech-to-speech translation system. ΦDMDIALOG is a real-time Japanese-English speech-to-speech dialog translation system that accepts speaker-independent continuous speech input and produces audio output of translated sentences. It has been publicly demonstrated at the Center for Machine Translation at Carnegie Mellon University since March 1989. Major problems are seen in the areas of (1) integration of linguistic and statistical nautral language processing, (2) use of discourse knowledge to improve speech recognition rates, and (3) attainment of simultaneity in translation. We chose these problems to discuss here because they are very difficult and unique to speech-to-speech translation systems; they are not encountered in text-based machine translaton systems. We describe our approaches to these problems and describe problems left unresolved, so that this paper can serve as a basis for further discussions on these issues.

1 Introduction

In this paper, we discuss some of the problems which we have encountered in developing ΦDMDIALOG, a real-time speech-to-speech translation system which accepts speaker-independent continuous speech inputs[1]. We focus on problems which we could not resolve, rather than on what was solved, in order to provide a basis for further discussion as to approaches in developing speech-to-speech translation systems. These problems are in the areas of:

1. Integration of Linguistic and Statistical Natural Language Processing,
2. Use of Discourse Knowledge to Improve Speech Recognition Rate,
3. Attainment of Simultaneous Interpretation.

We do not mean that all other problems in other areas which are not mentioned above are resolved. We chose these problems because these are unique to speech-to-speech translation systems. Although, these problems MUST be resolved, convincing solutions are not within sight. Detailed descriptions of each problem are given in individual sections of this paper.

2 The ΦDMDIALOG Project

ΦDMDIALOG is an experimental real-time speech-to-speech dialog translation system which accepts speaker independent continuous speech inputs. ΦDMDIALOG deploys a radically new architecture in that parsing and generation processes are integrated and run concurrently using a hybrid parallel paradigm, which is an integration of a parallel marker-passing scheme and a connectionist network, attaining incremental production of sentences while parsing is in progress. This is an opposite approach to traditional approaches of speech-to-speech translation systems which put together modular components: a speech recognition device, a parser, a generator and a speech synthesizer. One other feature relevant to the topic of this paper is the use of case-based parsing as well as unification-based parsing. Case-based parsing is a method of parsing based on a memory which records past cases of utterance. It locates the nearest parse and applies it to analyze an input sentence. The case-based method is used even for the generation process. These features are incorporated because we intend to simulate simultaneous interpreters at work. Also, the model takes into account major psycholinguistic and cognitive theories of parsing and generation. For details of ΦDMDIALOG refer to [Kitano, 1989]. ΦDMDIALOG was 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. It operates on the ATR's conference registration domain. Current implementation translates Japanese into English, and has been publicly demonstrated since March 1989. On the off-line experiments, the ΦDMDIALOG performs bi-directional translation between Japanese and English. The project is now in a Phase II stage reflecting the problems discussed in this paper. In addition to these problem, the Phase II stage involves implementation on massively parallel machines such as IXM2[Higuchi et. al., 1989], The Connection Machine[Hillis, 1985], and SNAP[Moldovan et. al., 1989]. An implementation on the SNAP is being carried out as a joint project between CMU and USC. In this joint project, SNAP's architecture is designed to meets various requirements of the ΦDMDIALOG, and a version of the ΦDMDIALOG called DMSNAP is up and running at USC on SNAP. Plus, an enhanced discourse model, extension to handle Spanish and French, and interface to HMM-based

*This work is, in part, supported by the National Science Foundation Grant MIP-9009109

[1]The system is one of the three experimental speech-to-speech translation systems currently up and running. Others are SPEECH-TRANS at CMU and SL-TRANS at ATR.

and TDNN-based speech recognition is being developed.

3 Integration of Speech and Natural Language Processing

To improve the accuracy of a speech recognition system, reduction of search space is essential. This can be attained by using predictions from the language model. In the stochastic model of speech recognition, likelihood of the recognition hypothesis being correct (P) is given by:

$$P = P(y|M)\,P(M) \quad (1)$$

where $P(y|M)$ is a probability measure given from the speech model, and $P(M)$ is an a priori probability given by the language model. If there are two systems which have identical $P(y|M)$ distributions, a system with a language model which provides more accurate $P(M)$ measures would have a better recognition rate.

3.1 Integration of Linguistic and Statistical Parsing

Most languge models deployed in current speech recognition systems are statistical models such as word-pair grammars, N-grams, or finite-state networks which provide $P(M)$ based on the statistics derived from a given corpus. However, this is criticized by NLP researchers because these methods hardly reflect linguistic analysis nor analyze meanings of utterances by themselves. Researchers in the field of natural language parsing claim there needs to be more linguistic methods utilized in developing a language model of a speech recognition system. However, simply using a linguistically based language model does little to improve recognition because only insufficient predictions can be made. This is because most linguistic theories assume pre-terminal symbols (such as N, V, A, P) which are too abstract to make any specific predictions.

We took the approach of integrating linguistic and statistic processing by using stochastic marker-passing in which markers carry a probability measure as well as linguistic and semantic features. This allows our model to assign probability measures to each grammar rule which affects the prediction of possible next words. However, simply assigning probabilistic measures to syntactic knowledge is not good enough, because the real problem in using linguistic analysis has been its weak prediction capability, i.e. pre-terminal symbols are too abstract. Instead of using syntactic knowledge alone, we use cases of utterances which are more specific instances of knowledge of possible utterances. Most specific cases have surface word sequences and are indexed in the memory network in order to map surface word sequence to meaning representation. Generalized cases are located in a middle layer between specific cases and linguistic knowledge since they are of medium level abstraction from cases. Use of cases reduces perplexity because predictions from cases are more specific than predictions from linguitic knowledge, and more constrained than a bi-gram grammar. In an experimental grammar which has a test set perplexity of 41.0 with a bi-gram grammar, the perplexity for the same test set was reduced to 6.4 by using cases of utterances. Probabilities carried by markers from each level of processing are merged when they meet at certain nodes as predicted, and decide a final a priori probability distribution. A word predicted by a specific case has the strongest a priori probability, and words predicted from less specific knowledge have weaker a priori probabilities. Although this method successfully reduces the perplexity measure in the test set used in the experiment, there are some open questions as to its effectiveness when it applied to larger domains.

First, it is not clear that the use of case-based parsing guarantees reduction of perplexity in larger domains, although we are quite sure that use of cases provides more contraints than that of using bi-gram grammars, and it obtains a sufficiently low perplexity measure for the test set we are running at this time. However, it is not clear how this perplexity measure will be worsened as the task domain expands. We need further investigations with larger test sets before we can conclude that the use of cases can be a solution to our task domain.

Second, building a knowledge-base for cases requires far more computational power than building a bi-gram grammar. Each case is indexed into the memory network of the system with assigned semantics and probability measures. Probability measures can be assigned by computing the probability distribution of those cases that can be selected at a specific point of the dialog and so is a relatively straightforward task. However, assigning semantic representation to each case and indexing them into the network is not a trivial task. This is a central topic of the case-based reasoning research community at this moment. Due to this problem, testing the effectiveness of case-based parsing on a large scale domain is a difficult task, since results may vary depending upon how the network is indexed, and how the network should be indexed is as yet not resolved.

Third, a current ambiguity resolution scheme [Kitano et. al., 1989] retains hypotheses unless their probabilities fall below a given threshold. This scheme does not provide sufficient added constraints at the speech processing level. Ambiguities should be resolved as early as possible. However, there are cases where committed choice fails at the end, i.e. garden path sentences. We do not know how to utilize ambiguity resolution without taking the risk of faulty committed choice.

3.2 Use of Discourse Knowledge for Prediction

Given the fact that syntactic and semantic levels of knowledge are not enough to sufficiently constrain search space, use of pragmantic and discourse knowledge, such as discourse plans [Litman and Allen,] and discourse structure [Grosz and Sidner, 1985], gain attention with the hope that introduction of these higher levels of constraints may help by further reducing perplexity, and thus attain higher recognition rates. Actually, [Young et. al., 1989] reports that perplexity was reduced dramatically by introducing discourse knowledge using a layering prediction method, and that the semantic accuracy of the recognition result was 100%. The introduction of discourse knowledge would be useful for highly goal-oriented and relatively limited domains such as the DARPA resource management domain. We have investigated the effectiveness of using predictions from the discourse level in the ATR conference registration domain, because the ATR domain is a mixed-initiative and less goal-oriented domain.

We assumed plan hierarchies for each participant in the dialog. The questioner and the secretary have their own plan hierarchies. Predictions are made from both hierarchies and nodes predicted from both plans have higher probabilities than those predicted from only one. Discourse plans and goal hierarchies for each speaker interact to produce specific predictions on possible next utterances. We have carried out an experiment using a small corpus of 3 dialogs consisting of 92 utterances. For the test set of perplexity of 19.7 using syntax and semantic constraints, the addition of discourse knowledge reduced the measure to 2.4. This result alone is a significant success, and seems to confirm the effectiveness

of using discourse knowledge. However, the effectiveness of predictions from discourse knowledge largely depend upon the task domain and the coverage of the corpus compared to dialogs in real deployment. There are three basic problems.

First, although, use of discourse knowledge generally helps in reducing perplexity, this assumes that patterns of dialogs, i.e. transition patterns among subdomains, are relatively limited so that discourse level knowledge can further constrain the possible next word choice. We have investigated patterns of subdomain transitions in the ATR corpus in order to examine how this assumption holds in our domain. We took 30 dialogs from the ATR corpus to measure perplexity of subdomain transitions. The total number of utterances in 30 dialogs was 1325. 31 subdomains and 177 transitions were identified. For each subdomain, there were several subdomains within them, but we did not count those details. We simply counted major transitions between relatively abstract subdomains. Each subdomain consisted of utterances ranging from 2 utterances to over 50 utterances. Perplexity of the test set without constraint was 4.95 (note that this is not a word choice perplexity). A test set perplexity of subdomain choices was reduced to 3.44 by using a bi-gram grammar at the subdomain transition level. Yet, on average, the system has to select a domain from 11 hypotheses in order for a new subdomain to transit. Moreover, none of the dialog transitions were equal to any other dialog transitions, which implies that perplexity of the task with a larger corpus of data would be significantly larger than what we encountered in our experiment. It should be noted that a considerable portion of syntactic structures and vocabuluary are shared among subdomains, so that even if the possible next subdomain is reduced to 1/3 of all subdomains, this does not mean possible next words can be reduced to 1/3. Therefore, unfortunately, we must conclude that the use of discourse knowledge which captures transition patterns among subdialog domains, i.e. statistical transition models, bi-grams at the discourse level, and goal calculations, would have only a limited impact in reducing preplexity in mixed-initiative domains with larger topic space.

Second, the nature of mixed-initiative dialog makes accurate prediction even more difficult. Unlike the DARPA resource management domain, the ATR domain is a mixed-initiative dialog where the two participants in the dialog have their own intentions and goals. This is one of the inherent characteristics of the task which the speech-to-speech translation system is expected to process. Here is an example taken from the ATR corpus:

Secretary: Please give me your card name and number.

Questioner: It's American Express, the number is 123-45678-90123. Would the proceedings be published?

Secretary: Yes, it will be published in July. Can I charge a registration fee to your AMEX account?

Questioner: Yes, and send me a registration form please.

Secretary: OK. I need your name and phone number, too.

A domain of this subdialog seems to be a credit card charge, but it has subdialogs of asking if the proceedings will be published and asking for a registration form to be sent out. Although predictions of speech acts may be attainable since more than 80% of the interaction is based on the Request-Inform discourse plan, predictions on which subdomain the dialog may switch into and when it may happen are hopelessly difficult. In the above dialog, how can we predict that the questioner may ask if the proceeding will be published in the middle of a dialog on a credit card? This means that although stronger preferences can be placed on some of the subdomains, the system must be able to expand its search space to nearly the entire domain so that sudden switching of subdomains in such complicated dialog structures can be handled. When this happens, the perplexity measure would drastically increase; in the case of our experimental set, it should fall somewhere in the middle between 2.4 and 19.7. However, obviously expanding search space to entire domains significantly undermines the recognition rate. We still do not have an answer to this problem.

Third, prediction failures run the risk of undermining recognition rates by pruning out a correct hypothesis in favor of incorrect but predicted hypotheses. Chances of making wrong predictions depend upon the coverage of the corpus collected from real dialogs. If the corpus covers a sufficient portion of possible dialog transitions, the chances of making wrong predictions would be much lower. In the ATR's conference registration domain which involves various topics such as sightseeing, dinner, and hotel reservations, covering all possible subdomains and transitions is nearly impossible. Actually, one dialog of the corpus involves how to spend time with geishya girls at Kyoto! While covering all possible transitions is not feasible, the problem remains of how to avoid selecting wrong but predicted hypotheses when an unexpected utterance is made. We believe that higher level knowledge can help only a little to with this problem, and that it can even be harmful in some cases. The only solution we suggest is to improve speech recognition at lower levels.

In summary, the language model cannot be 100% correct in providing a priori probability to the speech processing level. Use of discourse knowledge is effective only with a task of a relatively limited domain, and it would be less effective in mixed-initiative and wide domains with which we intend to deal. Given the fact that a highly accurate prediction of what may be told next is not feasible, we still need to improve the speech recognition system's accuracy without depending on higher levels of knowledge sources such as discourse knowledge.

4 Simultaneous Interpretation

In real dialogs, the length of each utterance can be considerably long. Utterances of sentences where each took 10-15 seconds are frequently observed. This imposes critical problems in deploying sequential parse-and-generation type architectures. Supposing one utterance 15 seconds in length, the hearer would need to wait more than 15 seconds to start hearing the translation of her/his dialog partner's utterance. Then, assuming that she/he responds with an utterance of 15 seconds in length, the first speaker would have to wait at least 30 seconds to start hearing her/his dialog partner's response. We believe that unless speech-to-speech translation system overcome this problem, practical deployments are hopeless.

The only approach we have today, and we believe this is the approach we should take, is to simulate actual simultaneous interpreters at work. The important point is that the simultaneous interpreter starts translation even before the end of a sentence. This is especially true when an utterance to be translated is a long sentence with multiple subordinate clauses. It should be also noted that the one sentence in Japanese can be translated as several English sentences, and vise versa. We believe that this incremental and simultaneous parsing and generation is the key to practical speech-to-speech translation system.

The Simultaneous Interpretation algorithm is the solution

which we are proposing and implemented in the ΦDMDIALOG system. In the algorithm, incremental parsing and incremental generation processes [2] are coupled so that generation of a part of a sentence can start even before the entire sentence is heard by the system. The incremental parsing process provides multiple hypotheses of fragments of sentences so that the incremental generation process can start building up sentence fragments in the target language. This process is inherently parallel due to the ambiguities involved in both the parsing and generation processes. Whenever ambiguity is resolved, either completely or with high certaintity, the parser provides the generator with a meaning representation of the part of the utterance which is disambiguated. The generator picks a hypothesis which meets this meaning representation, and produces a part of the sentence in the target language in an incremental manner.

Obviously, the most significant issue is how to resolve ambiguities of the parsing process as early as possible, so that the final translation hypothesis can be determined as early as possible. A conventional method of pipe-lining the syntactic parser and the semantic ambiguity resolver is not an appropriate solution. The ambiguity resolver needs to be embedded in the syntactic parsing stage so that feedback from semantic and even discourse levels can be obtained immediately after an ambiguity is detected. This is a very difficult task because (1) it undermines the ease of trace and debugging, and (2) ambiguity resolution itself is not fully accomplished. However, one hope is that, as can be seen from the transcript, the interpreter does not start translating unless she/he is sure about what the sentence means so that if ambiguity can be resolved at the end of each clause, delay in generation would be minimal. Another problem is how to decide to which hypothesis to commit. If some ambiguities still remain, the generator needs to commit to one of the hypotheses, which may turn out to be false. This would be even complicated when a source language and a target langauge has substantially different linguistic structures. For example, in English, negation comes before a verb, whereas Japanese negation comes after a verb, and the verb comes very end of a sentence. In such case, translation cannot be started until the verb, which comes the end of the sentence, was processed, and existance of negation after the verb is checked. Decision has to be made, for this case, to wait translation until these ambiguities are resolved by encountering a clause which follows the initial clause. Fortunately, most Japanese utterance consist of multiple clauses which makes simultaneous interpretation possible. In order to cope with these ambiguities, a simultaneous interpretation system should be capable of (1) anticipating possiblity of negation at the end, (2) incorporating some heuristics which recover false translation to correct one, and (3) making a decision on when to start or wait translations. Theories of commitment in ambiguity resolution and generation are not established yet, thus they are a subject of further investigations.

5 Conclusion

In this paper, we discussed some unresolved problems which we have encountered in Phase I of the ΦDMDIALOG Project. In essence, we made the following conclusions:

First, we claim speech recognition rates need to be improved drastically without depending upon higher level knowledge sources such as discourse knowledge, because accurate predictions of possible next subdomains in wide and mixed-initiative domains are less effective than in limited domains where successful use of discourse constraints has been reported. In addition, syntactic and semantic knowledge does not sufficently reduce preplexity to the point where high sentence recognition rates can be obtained in larger domains.

Second, the sequential and modular architectures which most machine translation systems use are inappropriate for practial speech-to-speech translation systems due to the inherent delay in transactions. A new architecture using the simultaneous interpretation algorithm is an approach to the problem, yet it would be premature to say at this point that this is a solution. There needs to be further investigation.

Third, ambiguity resolution schemes which resolve ambiguity at the earliest possible point need to be investigated because (1) they help in reduceing perplexity of next possible word choice, and (2) the elimination of ambiguous parses is critical in starting to generate translation while the rest of the sentence is in the process of being parsed. However, which hypothesis to commit to is a major issue, since faulty decisions degrade an entire system's performance.

In addition to these issues, we need to realize that the computational cost of simulating a parallel process would be enormous when our scheme is applied to large scale domains that contain a few thousand vocabulary items and concepts. The solution to this problem is to implement our scheme on a massively parallel machine such as the connection machine [Hillis, 1985] or SNAP [Moldovan et. al., 1989].

We believe these problems need be resolved in order to implement speech-to-speech translation systems which can actually be deployed, and therefore these problems should be considered as research topics of prime importance in the 1990s.

[2] For details of our incremental generation scheme, refer [Kitano, 1990].

References

[Grosz and Sidner, 1985] Grosz, B and Sidner, C.L., *The Structure of Discourse Structure*, CSLI Report No. CSLI-85-39, 1985.

[Higuchi et. al., 1989] Higuchi, T., Furuya, T., Kusumoto, H., Handa, K., and Kokubu, A., "The Prototype of a Semantic Network Machine IXM," *Proceedings of the International Conference on Parallel Processing*, 1989.

[Hillis, 1985] Hillis, D., *The Connection Machine*, The MIT Press, 1985.

[Kitano, 1990] Kitano, H., "Parallel Incremental Sentence Production for a Model of Simultaneous Interpretation," In *Current Research in Natural Langauge Generation*, (Eds.) Mellish, C. and Dale, R., 1990.

[Kitano, 1989] Kitano, H., *A Massively Parallel Model of Simultaneous Interpretation: The ΦDMDIALOG System*,CMU-CMT-89-116, Center for Machine Translation, Carnegie Mellon University, 1989.

[Kitano et. al., 1989] Kitano, H., Tomabechi, H. and Levin, L., "Ambiguity Resolution in DMTRANS PLUS", In *Proceedings of the Fourth conference of the European Chapter of the Association for Computational Linguistics*, 1989.

[Litman and Allen,] Litman, D. and Allen, J., "A Plan Recognition Model for Subdialogues in Conversation", *Cognitive Science 11* (1987), 163-200.

[Moldovan et. al., 1989] Moldovan, D., Lee, W., and Lin, C., *SNAP: A Marker-Propagation Architecture for Knowledge Processing*, CENG 89-10, Department of Electrical Engineering - Systems, University of Southern California, 1989.

[Young et. al., 1989] Young, S., Ward, W. and Hauptmann, A., "Layering Predictions: Flexible use of Dialog Expectation in Speech Recognition," In *Proceedings of IJCAI-89*, 1989.