Prospects for Advanced Speech Translation

Hitoshi Iida
Speech and Language Information Processing Laboratory
Sony Computer Science Laboratories
Tokyo, Japan
iida@csl.sony.co.jp

Abstract

Speech communication includes many important issues on natural language processing and they are related with desirable advanced speech translation systems. Advanced systems need to be able to handle the interaction for speech communication, pragmatics in speech, and arbitrariness of speech usage. General characteristics of speech communication are discussed. Also the various viewpoints regarding interaction, pragmatics, and arbitrary usage are discussed. Some of the present speech translation approaches are explained and new basic technologies are introduced. In this paper, a synthetic NLP technology such as a composite art form is proposed for speech communication and speech translation.

1 Introduction

When we are doing cross-language communication in daily life, a conventional machine translation system that translates written language texts does not always give the correct translation because the utterances depend on the situation and the real meaning can differ from the literal meaning. Translation results suitable for speech communication cannot be absolutely decided, but relatively dependent on each dialogue situation. Figure 1 shows the state of affairs regarding the differences between a conventional MT world and a speech translation world. In case 1, the source language grammar, target language grammar, and transfer rules might establish an ideal machine translation where the translation result can be determined uniquely and its meaning must be absolutely correct for the readers and the recipients. On the other hand, in case 2, any utterance meaning is situation depended and the translation result is one candidate of potential utterances in the target language. In this case what an MT system requires is just a necessary and sufficient expression suitable for the dialogue situation and the context in the target language.

General characteristics of speech communication are discussed in the next section. Also, various viewpoints regarding interaction, pragmatics, and arbitrary usage are discussed in the section 3, 4, and 5, respectively. Some of the present speech translation approaches are explained in the section 6. And requirements for advanced speech translation and desirable new basic technologies are discussed in sections 7 and 8.

2 Speech Communication

Almost all the current continuous speech recognition systems output word sequences [Lavie 96]. And they generally do not make any spelling errors because wrong phoneme sequences are handled by modifying them into certain word sequences. Thus, a method to eliminate incorrect word sequences including

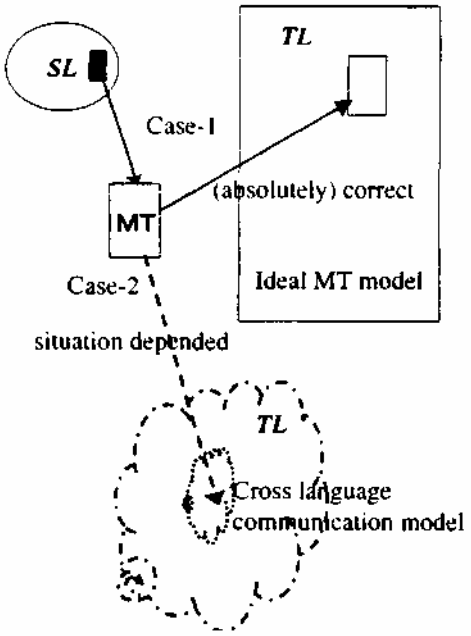

Figure 1. MT for Texts and Speech

unnecessary or unrelated words is different from an error correction method and is expected to avoid errors for speech-to-speech systems. Three types of 'ill-natured' speech recognition outputs can at least be observed.

1) a word sequence that includes an incorrect subsequence that may be analyzed by a kind of syntactic parser,
2) an incorrect subsequence that can be interpreted as a correct meaning by a kind of semantic analysis, and
3) a word sequence that is difficult to divide into sentences or clauses only using syntactic constraints.

In addition, since a general speech recognition system for dialogue utterances does not output a boundary maker between utterances or punctuation in written texts, wrong analysis and translation can easily happen when a system is handling speech inputs that include a lot of fragmental utterances.

So, in order to develop speech translation systems that can produce comprehensible natural conversations between two speakers using different languages, various ideas to recover incorrect outputs from speech recognition systems have been presented. To recover such an error, new technology should be able to handle dialogue contexts, situational information, and world knowledge concerning dialogue tasks and domains. The new technology should also be able to make suitable judgements using common sense. Fortunately, some pieces of information are given through an uttered expression. For example,

- Literal meaning
- Referent objects
- Speaker's interpretation depending on the individual and mutual background between dialogue participants
- Mutual information which forms a context

In this paper, problems regarding speech communications and approaches to extend the current technologies are discussed based on the following three viewpoints.

[1] communications style: interaction
[2] communications norm: pragmatics
[3] communications arbitrariness: arbitrary usage

3 Interaction (vs. one-way message)

Interactive communication can be easily seen in various human-computer interfaces, in particular highly resolution and multi-dimensional visual information on multi-media user interfaces. With regard to speech interfaces almost all the current technologies barely recognize grammatical and continuous speech spoken clearly. Therefore, there are some problems like information for users is conveyed subjectively from computers which means that there are few interactions between a user and a computer. With more reliable and accurate spontaneous speech recognition technologies, spoken language translation technologies have to handle a combination of utterances between a user and a computer, an utterance reflected by a real-time reaction, references jointly with dialogue participants or a user and a computer.

3.1 Meaningful communications using a combination of utterances

Previously, many studies on dialogue understanding clearly showed that in a dialogue there are correlations between the participants' utterances, in particular turn taking relations. A goal-oriented dialogue consists of a query, an answer, and a confirmation. Such an information chunk under a turn-taking region satisfies the current information slots with some replied values as fillers, and forms a substructure in a whole dialogue structure. So many goal-oriented dialogues consist of a number of substructures. It is very important to note here that the first utterance in a substructure provides the key information for presuming the topic of the structure and its goal.

A method that unifies two successive sentences such as an unification operation between the main clause and the subordinate one in a complex sentence was created more than ten years ago. This method can resolve zero-anaphora problems and can produce a semantic expression about the whole meanings of two successive sentences. The method can be applied to merging information between two utterances, and also presuming an utterance type of the second one and its possible contents. These processes handle contextual information in such a way that they have a good effect on speech recognition accuracy and spoken language processing capability, especially when handling fragmental utterances. For example, helpful responses and omitted expressions can be handled using the above approach.

3.2 Real-time reaction and confirmation

Speech communication and also speech translation stand up through various real-time reactions including a confirmation not just to physical actions but to speech the other speaker utters. Even if there is no image information during communication, such as during a telephone conversation, communication can be accomplished just by speech. In those dialogues. the pragmatics in a social system play an important role when interpreting the speaker's intentions and conveying correct information since there can be misconceptions. On the way to correctly understand communication undecided situations often happen. This phenomena indicates that it is difficult to decide which possibilities come next on one sub-process of a cascaded hierarchical processing since underdetermined processes have to be considered. Some intermediate processing must be carried out in any stages.

3.3 Promptness

In general, speech inputting and outputting can be considered to be very easy when compared with key-inputting and outputting normal daily life case, such as a face-to-face conversation in a natural sounding environment. The reason is that in speech communication chattering in the form is established by chattering in the form of utterance turn taking without a break. The listener and speaker have the capabilities to understand, manage, and generate the utterances quickly as well as support the realization of such communication. A communication system has to at least be able to manage a cooperative information exchange between a human and a computer. The processes mentioned in subsection 3.1 and section 4 must be carried out in real-time and they must have the above capabilities.

A prompt turn taking of utterances can have another effect on understanding ungrammatical or fragmental ones easily. Promptness of interpretation and reply can be done in a very short time period so that the referred objects and actions are, at least, restricted to a short turn taking scope with few descriptions. When finding out a referred object or an omitted expression in a short turn taking dialogue, a computer system can easily limit its scope to look for the meaning rather than in a sub-domain under some dialogue topic. In other words the promptness helps the listener or system to interpret spontaneous speech because the speech is short and the set of vocabulary is small. The phenomena that identifies omitted phrases is related with identifying the references mentioned as follows.

3.4 References Jointly with Speakers

Cooperative speech communication might be supported by a situation jointly with dialogue participants. Or it might be supported by a virtual situation that would become the situation jointly with them after having introduced it from one participant to the other one. In a situation and context that is made by some sequences of turn-taking utterances, referents including their parts and related actions can be generally restricted, making each participant easy to identify them.

The above referent identification is inside the range of linguistic literal usage. Other referent problems are as follows.

Misconceptions or misunderstandings might be caused by a lack of common sense or by a difference of references between speakers, and it might be caused by differences of both speakers' associative knowledge and their recognized situations and objects. So a computer system that can handle these phenomena has to be able to manage standard and conventional associative knowledge rather than individual specific personal knowledge because a normal spontaneous speech dialogue is carried out under the current daily life pragmatics and the present social systems.

Another problem is supplemental modification of uttered phrases and unfamiliar usage of phrases or words for a hearer. Unfamiliar phrases are promptly repeated, or a speaker explains the phrases in order to make up for miscommunication. Such a modification is temporal in one turn taking chunk. The present speech recognition systems make some misrecognition results. A hearer who listens to the output including errors might misunderstand the original utterance, or. at least, s/he might be not able to understand some phrases in the output.

From these phenomena, monitoring dialogue execution must be required in order to create spontaneous speech interactive functions.

4 Pragmatics in Speech (vs. Grammar in Language)

There are various types of expressions for politeness, modesty, and euphemism. Such expressions are used depending on social roles [Iida 95], [Mima 97]. Conventional communication rules to keep human-to-human information exchanges smooth and friendly work well as pragmatics and are used to create various kinds of social systems.

4.1 Utterances based on a Situation

It is generally well known that there are many linguistic expressions used to present one fact or state of affairs. These expressions are dependent on the speaker's attitudes and his/her situation. In sub-sections 3.1, 3.3. and 3.4. it is said that cooperative speech communication might be executed under a situation jointly with dialogue participants. We can see the main dialogue characteristics in fragmental expressions, omitted expressions, substitute expressions for repeating the same contents, situation-depended expressions, and so on. To understand those expressions, a dialogue environment, the context, and the situation where potential referents might be easily found out must be handled [Levin 95].

A typical Japanese expression that is depended on a situation is shown. The Japanese case particle "wa" is a topic marker and. in general, this marker can be replaced by other case particles. But the following usage cannot be easily replaced by other particles.

a) "Go-yotei-wa o-dekake-wa itsu deshou ka."
('schedule-polite-form-topicalized,' 'departure-polite-form-topicalized,' 'when,' 'be-future-form-question');
the English translation is what is your departure time (on your schedule).

To interpret such utterances, identification of referents in both the dialogue environment and the situation is required, and some prediction on expressions in the next turn helps to interpret the present utterance in comparing with familiar expressions that at least include no irregular sub-phrases. In the above example, it is not necessary to

utter the first topicalized phrase. It is clear to identify the second topic "dekake" under a normal situation.

4.2 Social Norms of Learning (or Education)

There are various types of expressions for politeness, modesty, and euphemisms. Such expressions are used depending on social systems and roles, and are executed satisfying social norms of good attitudes and behaviors in speech (in particular using various types of honorific expressions). The following examples, b) and c) in Japanese, present such expressions using combinations of main verbs, causative verbs, and auxiliary verbs.

b) "Yoyaku-wo kakunin-sasete-itadaki-masu."
('reservation-objective,' 'confirm-modest')

c) "Go-dengon-wo o-tutae-moushiage-masu."
('message-polite-form-objective,' 'inform-honorific')

The propositional contents of both examples are very simple. Each of them can be represented by a combination of the main predicate and some case role arguments. However, those honorific expressions and other various indirect expressions are tough to interpret and generate in satisfying the situation on the spot. In their translation, one of the main problems is how to handle social norms. Those norms help execute cooperative social communication and are related with the pragmatics in an office, a family, an association, and so on. Handling those so-called social manners is required for a speech translation system.

4.3 Expression of Speaker's Attitude

There might be some reasons why many linguistic expressions to present one fact or a state of affairs come out. One reason may be seen from the following utterance.

d) "Shiharai-wa ginkou-furikomi-o o-machi-shite-orimasu."
('payment-topicalized,' 'bank-transfer-objective,' 'wait-for-polite-form-modest')

The phrases of two different sentences are irregularly mixed in one utterance. The first phrase in d) is marked by "wa," and in general goes together with "o-negai-itashimasu" ('ask-for-polite-form-modest') after the second phrase marked by an objective case particle. But the second phrase introduces the speaker to utter the third one that has no correlation with the first one marked by "wa." (If the first one is marked by "o," it agrees with the third one.) In other words, the first phrase works in order to identify a topic scope and in this case the effect introduced by the topicalization is very weak so that the main intention can be expressed by the last two phrases. Those phrases express the speaker's attitude, and in Japanese speech the information he/she wants to convey to a hearer is presented in the last part of the phrases of the utterance as a main verb or a head noun. The first one meets neither a main predicate, nor an obligatory case role in the utterance. The word stream of the utterance is logically twisted.

The above example, in which the topicalized phrase is unnecessary for a hearer to interpret the utterance, shows some useful information for grasping a speaker's view rather than a necessity of handling irregular or ungrammatical expressions for computer processing. Much of the useful information that appears in such irregular expressions can help to identify the speaker's focus, consciousness, and so on. An interpreting system is required to handle the speaker's thought processes using expression irregularities as clues to his/her attitude.

4.4 Metaphorical Language Generative Competence

In considering metaphors, a relationship between an object and its part must be imagined first. Such a relation can be seen in example e). The literal meaning is that someone parts one's head into some pieces, but the interpretation is far from common sense. The object can be inferred to be the hair of a head that is directly described. The word "head" is used as a metonymical expression, while "hairs on the head" is a direct expression.

e) "Atama-o wakeru."
('a head-objective,' 'part (one's hair in the middie)')

On the other hand, in f) there is a relationship between an object and a related action under the specified topic, 'motorcycling,' i.e., a crash helmet and its supported action 'putting on.' The relations between the nouns and the verb can be settled in a suitable relation under a restricted topic that governs them by the first phrase. This phrase, followed by the topic marker "wa," has many meanings depending on the interpretations of "wa." One of the meanings is that it would be better for a person who motorcycles to put on a crash helmet. At least, the first phrase marked by "wa" cannot be interpreted as the agent role. This is not a metaphorical expression, but a normal interpretation. This fact shows us that we must pay special attention to handling a topic phrase, especially a Japanese topicalized phrase marked by "wa."

f) "Ohtobai-wa herumetto-o kaburoh."
('(When motorcycling) a motorcycle-topicalized,' 'a crash helmet-objective,' 'put-on-suggestion')

It might be judged whether a given expression presents a metaphorical meaning or literal one by some relations between phrases with associative relations under common sense knowledge. In order to handle common sense, a computer system has to pay much attention to noun phrases, in particular complex nominals, word formation, etc.

5 Arbitrary Language Usage (vs. Standard Writing System)

In daily life conversations, grammar restrictions do not work well because various types of arbitrary language usage appear and they are not captured by a standard writing system, i.e. school grammar. Those facts have been mentioned in the previous two sections. Moreover, 'paralinguistic' expressions should be worth discussing since they come from human arbitrary and temporal behavior accompanied with tone of voice, prosody, phrase repetition, and so on. In this section those paralinguistic usages and their problems of handling them on speech translation are discussed.

5.1 Conveying Speaker's Mental State with the Tone of Voice

No matter how simple a speaker's query may be, his/her tone of voice conveys his/her own mental state which might be skeptical, hurried, confident, acceptable, and so on. At present, there is no reliable technology that can recognize voice tones in spontaneous speech. By using a combination of expressions that reflect a speaker's attitude (mentioned in the subsections. 4.1 and 4.3) and arbitrariness of speech generation that reflect human behaviors (mentioned in subsection 5.3), information concerning voice tones ought to help to infer a speaker's mental state even if the technology is poor.

5.2 Prosodic Information

Information patterns and pauses that discriminate between various meanings give very important clues for a speech translation system to utilize paralinguistic information. Such a prosodic information can play a role in eliminating many of the candidates derived from literal language analysis and thus establish candidates suitable for the situation. Various interpretations of linguistic phenomena and their constituents are well known: speech act. modifier scoping, old or new information, discourse structure, response extent. and demonstratives [Iida93]. If speech information is to be fully grasped, mental states related to feelings such as anger or nervousness must also be considered. This would make possible the correct interpretation of errors arising from morphological linguistic expressions, and understanding the intention of an utterance. In spontaneous speech processing, the character string "5.000 en desu ka" (this means "Is it 5,000 yen?") in Japanese is not only the question type but also more mood types, for example skeptical mood and wondering mood. This utterance interpretation by the hearer is made possible by an exchange of information between the mental states related to the utterance intention and the mental states related to the feelings determined by prosodic information. Also, Japanese final particles added to the end of an utterance, such as "ne," and the intonation applied to them can indicate intention such as confirmation, command, or persuasion.

At present, there is also no reliable technology for handling prosodic information. In this case such a prosody handling technology has the same important role as the technology for handling voice tones.

5.3 Arbitrariness of Human Behavior

When a spoken dialogue translation system handles speech recognition results including misrecogmtion parts, it must consider various pieces of information derived from characteristics of an expression uttered under the present situation. The system handles such ill-formed expressions as well as irregular (or, in other words, fluent) expressions that come from human arbitrary and temporal behaviors, i.e. phrase repetitions, insertions of any additional phrases, inversions and self-corrections. On the other hand, hesitation in speaking or mistakes in enunciation are done in the same utterance. A logically twisted thread in the utterance described in subsection 4.3 is also related with the phenomena of human arbitrariness behaviors. Those expressions are very hard to treat deterministically. The translation system needs to grasp locally correct grammatical meanings and to expect speaker's intention. And. moreover, such a real technology requires an interaction mechanism between dialogue participants so as to resolve misunderstanding, misconception, and miscommunication.

6 Recent Approaches

Right now, it is necessary to establish a method for handling recognition results including misrecognized parts in which an incorrect word or word sequence is included, or unnecessary or unrelated words might appear. Such a method would be different from the string error correction technique used in general text processing. Three types of 'ill-natured' speech recognition outputs can naturally be observed: 1) incorrect word subsequences: 2) incorrect word subsequence types that could be interpreted as correct in another context: 3) a word sequence type that is difficult to divide into some sentences or clauses due to syntactic constraints. A new translation method we call "Maximally Partial Translation (MPT)" makes good effects on handling the above problems of misrecognized speech [Wakita 97].

Few trials of simultaneous interpretation in speech have been reported. One trial using MPT has shown a possibility of making a real system, and moreover, future possibilities of taking in various characteristic problems of speech translation which are mentioned in the previous sections [Furuse 96], [Mima 98].

6.1 Maximally partial translation

A general speech recognition system for dialogue utterances produces neither boundary markers nor written

punctuations between utterance units. If two sentences are concatenated without a sentence boundary in Japanese, the first one might be regarded as an embedded clause to modify the succeeding noun that must be the head of the second one because a Japanese declarative sentence, in particular, requires a verb at the end of a sentence. Even if the speech recognition result is correct, a wrong analysis can easily happen when a system handles speech inputs that include a lot of fragmental utterances. By certain parsing techniques and utilizing memorized phrases as examples of linguistic usage, dialogue participants are able to understanding each other when handling misrecognition results by a speech recognizer [Wakita97].

Through actual experiments, it has been proved that various types of linguistic information and their combinations are very useful for accurately detecting such errors as in:

(1) The connectivity of words;
(2) N-grams on words and characters; and
(3) Statistical information on dependency structures.

After error detection, the correct/plausible parts from the speech recognition results can be extracted so that some examples similar to the input string can be retrieved. By doing this, a translation that embodies the intention of the original utterance can be produced.

So-called Constituent Boundary (CB) parsing has good practical results [Furuse 96], and on the parsing processes the estimation of looking for a suitable example corresponding to the chunk depends on a certain criterion that is judged comparing with examples which are similar to the chunk in the example data. This means that capturing a similar example to a phrase in the input makes it possible to translate, which can then be regarded as a correct speech recognition phrase. Moreover, even if translation fails for an entire sentence, CB parsing can deliver meaningful partial structures. The extraction of correct parts from poor speech recognition results can be achieved using those partial structures. The new method MPT extracts correct parts from a speech recognition result and it translates the extracted parts. During CB parsing, this translation method utilizes the following two factors in order to obtain correct parts:

• A semantic distance between the input expression and an example expression.
• A structure selected by the shortest semantic distance.

A prototype system of speech translation using the following component technologies has already been made by ATR-ITL.

6.2 Simultaneous interpretation

A simultaneous interpretation system should have the capability to incrementally translate with synchronous activation according to speech inputs. To realize such capability the system, at least, has abilities to translate any phrases and any clauses as same as any types of sentences. According to speech inputs a chunk that consists of a meaningful grammatical sequence of words is extracted and simultaneously translated when the chunk can be identified with one grammatical constituent in the target language [Mima98]. Many characteristics on speech communication and speech translation have been explained in the previous sections. Concerning simultaneous interpretation, moreover, patterns of ways how to connect constituents, i.e. phrases, clauses, or sentences, with another ones are captured as constituent example - terns, and those patterns fill roles in determining a how to describe the relationship between the two related constituents. And the determined way can decide real linguistic expressions to connect with the following constituent meanings. Those processes using constituent example patterns show a method to determine how to express the linguistic relation between the present constituent and the dependent constituent that might follow the present one. This expression method looks like one of strategies a human interpreter uses on the real stages. Figure 2 shows a rough process image of a simultaneous interpretation process example for handling Japanese-to-English translation.

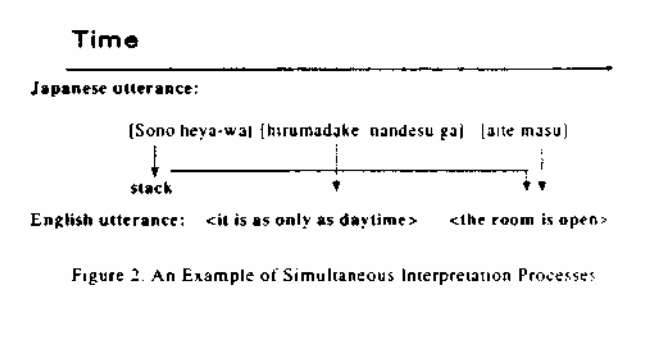

Figure 2. An Example of Simultaneous Interpretation Processes

7 Requirements

An individual process produces less efficiency. A conventional sequential process of natural language processing looks like a sequence of discrete information filters with local decision processes. From such a point of view the following two processes are required.

#1) A total judgement covered a whole process from speech input to speech output using situation-depended preferences and abductive common inferences can gently comprehend dialogues in each process and at each turn-taking subsection between speakers or between a human and a machine.

#2) Monitoring what dialogue participants image in their temporal memories and think in a conversational give-and-take action principle leads a system to mutual comprehension with the participants.

A new approach to establish a synthetic natural language processing technology is necessary for spontaneous

speech communication and translation. In this synthetic technology, language is placed at the center of the conveyed information and has various information elements tied in speech.

8. New Basic Technologies

To develop the synthetic technology much effort devotes to basic research on divergent interpretation and fitness for a real situation. The following three technologies can be enumerated so as to push forward with the basic research.

1) Chart structure under graph connection instead of the structure under linear connection
2) Maximally partial understanding or translation instead of fully understanding
3) Abduction instead of deduction

These research topics are concerned with a main mechanism to carry out total judgement in a whole system. The first one is needed for handling various possible processes in parallel, in particular integrated processes for merged regions to treat speech and language simultaneously. The second one handles partiality of information concerning characteristics of speech communication. The last one is concerned with judgement under partial information and is needed for handling vagueness and arbitrariness of human behavior.

9 Conclusion

Speech communication includes many important issues on natural language processing and they are related with desirable advanced speech translation systems. Advanced systems need to be able to handle the interaction for speech communication, pragmatics in speech, and arbitrariness of speech usage. General characteristics of speech communication are discussed. Also various viewpoints regarding interaction, pragmatics, and arbitrary usage are discussed. Some of the present speech translation approaches, for example maximally partial translation and simultaneous interpretation, are explained and new basic technologies, in particular 'Chart' structure under graph connection, maximally partial understanding and translation, and abduction, are introduced. Concerning speech translation, processing in progress is very. complicated and the result is simple. It is, however, fit to the situation where a dialogue is on going, and produces much effect to the dialogue participants. From this point, it is said that to handle speech communication and speech translation seems equal to treating a synthetic composite art.

References

[Furuse96] Furuse O. and Iida H. (1996). "Incremental Translation Utilizing Constituent Boundary Patterns," in Proc. of COLING'96, pp. 412-417.

[Iida93] Iida H. (1993). "Prospects for Advanced Spoken Dialogue Processing." IEICE TRANS. INF. & SYST.. VOL. E-76-D. No.1, pp. 2-8.

[Iida95] Iida H. (1995). "Spoken dialogue translation technologies and speech translation." in Proc. of MT Summit V. Addendum.

[Lavie96] Lavie A., et al. (1996). "Multilingual Translation of Spontaneously Spoken Language in a Limited Domain." in Proc. of l6th ICCL, pp442-447.

[Levin95] Levin L., et al. (1995). "Using Context in Machine Translation of Spoken Language," in Proc. of TMI-95, pp. 173-187.

[Mima97] Mima H., Furuse O. and Iida H. (1997). "A Situation-based Approach to Spoken Dialog Translation between Different Social Roles," In Proc. of TMI'97. pp. 176-183.

[Mima98] Mima H., Iida H. and Furuse O. (1998). "Simultaneous Interpretation Utilizing Example-based Incremental Transfer." In Proc. of COLING/ACL'98. pp. 855-861.

[Wakita97] Wakita Y, Kawai J. and Iida H. (1997). "Correct parts extraction from speech recognition results using semantic distance calculation, and its application to speech translation," in Proc. of ACL'97 Workshop on Spoken Language Translation, pp. 24-31.