3 Toward an Education and Learning Support Environment

3-1 New Horizons in Computational Linguistics

Hitoshi ISAHARA

This paper provides an overview of the activities of the Computational Linguistics Group of the Communications Research Laboratory of Japan. We will focus on the following topics: research on natural language processing, fundamental research on natural language, development of practical systems, and collaboration with other research organizations. As for research on natural language processing, we are conducting research on natural language processing using learning-mechanism and neural-network models. As for fundamental research on natural language, our research includes lexical semantics, representations of emotion, and a dialogue model with three participants.

Keywords
Natural Language Processing, Knowledge Engineering, Computational Linguistics

1 Introduction

At the Keihanna Human Info-Communications Research Center, our studies are aimed at the development of environments that support human intellectual activities. To achieve this aim, we are conducting research on natural language processing (NLP) and computational linguistics and developing practical systems using NLP technology, partly in cooperation with outside institutes. So far, we have obtained good results in our development of NLP technologies.

The study of NLP is based on learning mechanisms. We applied the maximum entropy method to the development of a morphological analysis, a dependency analysis, named entity extraction, and the determination of the order of words in a single framework. Studies on the learning-based correction of a corpus, on paraphrasing, and on text segmentation are also under way. Another task being undertaken is language analysis using a neural-network model to implement more accurate processes and to develop practical applications. Our study of information retrieval, information extraction, and summarization received a prestigious award in an objective evaluation contest. In the future, we plan to conduct an advanced study of summarization, by paraphrasing rather than by means of simple extraction of important sentences. By applying a post-processing rule to a system based on the learning mechanism, we will attempt to improve the precision with which language analysis can be conducted. Simultaneous processing will be useful in constructing a model that is more similar to human language processing.

The fundamental study of natural language includes a lexical semantics of Japanese adnominal constituents. We proved the theory by utilizing a self-organizing semantic map of Japanese nouns, with the help of a neural-network model. We plan on developing a framework for expressing the meaning of a sentence by stratifying the semantic map. In a study on Japanese reception expressions, we developed a numerical model using honorific expressions based on an experiment on test subjects, and studied historical changes in honorific expressions. We also worked with test subjects in another experiment to study adjectives that express sensitivity to music and other arts. The construction of a three-person dialogue model was also studied.

The development of a practical system includes the development and disclosure of the intelligent network newsreader. We intend to examine the accuracy of every function of such a newsreader in practical use. Another task in this category was to prepare a learners' corpus and examine error tags for the development of an English-learning support system. This corpus was prepared based on interviews with Japanese learners of English, and includes the results for determined skill levels of English usage. The present development is in progress under a technology transfer project of the Telecommunications Advancement Organization of Japan (TAO), “Research and Development of Adaptive Communication Technology.” The development and disclosure of a KWIC (KeyWord In Context) system, an information retrieval system, and a corpus compilation support system are also under way. We have worked with the National Language Research Institute as an outside institute to conduct a research and development project titled, “Spontaneous speech: corpus and processing technology,” under a Science and Technology Agency Priority Program for the Organized Research Combination System. Under this program, a large-scale corpus of spoken language is being developed, and a study on spoken-language processing technology is in progress based on the corpus. Another cooperative study consisted of consigned research (Breakthrough 21) on aphasia conducted in conjunction with Tokyo Metropolitan Institute for Neuroscience.

The above results will serve as a basis for the continuance of our study of natural language. This report will now sketch out the future research activities of the Computational Linguistics Group.

2 Study of natural-language processing

2.1 Development of a system for analyzing spoken-language text

We have developed a learning-based text analysis system for literary expressions contained in newspapers. This system may be difficult to apply with high precision to spoken language, which differs from written language in terms of commonly used expressions, ambiguity of word units, pauses, accents, and metrical information such as stress and speech speed. In the cooperative study, “Spontaneous speech: corpus and processing technology,” the processing of spoken language is being examined. In the near future, we intend to improve the present system to suit an analysis of spoken language by considering its metrical information characteristics. Experimental work may be necessary to determine whether the metrical information will effectively increase the accuracy of analysis. When fully developed, the new spoken-language system will be used to automatically add morphological, structural information to spoken-language text.

2.2 Study of machine translation based on translation memory

We have developed systems for analyzing and generating Japanese sentences to map the text in a syntactic structure showing relationships between phrases. If the Japanese syntactic structure is mapped together with the English syntactic structure, it will be technically possible to translate Japanese text into English text. With this in mind, we intend to map each word, the smallest constituent in a syntactic structure, to develop a system for selecting an equivalent term in translation. This is not an easy task, as a single Japanese word may often correspond not just to a single English word, but to a group of words. Selection of the most accurate English translation depends on gaining an understanding of the context of the Japanese word. We plan to prepare a variety of phrases containing each Japanese word, with an English translation added to each of the phrases. We refer to this preparation as "translation memory." The optimal English translation will be selected from this memory based on the similarity of the Japanese context to each of the prepared phrases. In the future, we will collect tagged data to create a learning-based system for selecting appropriate translations.

2.3 Study of automatic summarization

We have developed a Japanese summarization system based on sentence extraction. Our future plan is to apply this system to other languages, and then to output more natural summaries.

We will first attempt to extend the summarization system so that it can be applied to English documents, and also to separate language-independent and language-dependent parts. A successful separation will be used to create an improved summarization system that may be applicable to other languages. We also intend to summarize a corpus of spoken words to examine how useful the summarization technology based on literary expressions will be for spoken language, and what techniques will be necessary to summarize spoken language.

To output more natural summarizations, we will combine an information extraction technique for extracting specific information with an automatic summarization technique. The information extraction technique is expected to be useful not only in making superficial modifications, but also in determining the meaning of a sentence. Also, introducing methods for summarization to information extraction will be helpful in clearing up the problem of domain dependency in information extraction.

2.4 Study on text separation

Text separation is a task for separating a composition consisting of multiple topics into components each having a single topic. It is an essential process when retrieving bits of information and summarizing a composition. When retrieving information, it is more effective to consider the required topics rather than all topics in the entire document. When dividing a long document into sections by topic, summarization of the entire document may be possible by summarizing every section. Of course, summarization can be performed by selecting important topics only.

When retrieving information from or summarizing various documents in indefinite fields, the procedure for dividing each document should not be limited to a narrow field. Our procedure utilizes the distribution of words in a text in order to divide the text. This makes it possible to divide any text, rather than just specific fields containing training data.

In this procedure, text is divided so as to maximize the probability of division. This is the first approach of a procedure for dividing text without any restriction on its fields. Conventional procedures for dividing text are primarily based on the concept of lexical unity. In some cases, the unity is determined from the active propagation in a significance network, from the cosine of a word distribution, from the repetition of words, or from the order of cosines, rather than considering the cosine itself as a similarity between sentences.

Our procedure was able to divide text with accuracy equal to or greater than that of any of the conventional procedures. This indicates that our procedure is useful in dividing text, and we now intend to examine the general applicability of this procedure.

2.5 Study on a question-answering system

No definite criteria have yet been established for determining whether a computer "understands" a language. One possible criterion is to examine whether the computer gives adequate responses to questions. Turing's test uses a criterion similar to this. In this test, the subject is not told whether he or she is talking with a computer or a human. If the subject cannot tell whether he or she is talking to a computer or a human, it will be judged that the computer is as able to speak like a human. When a human subject is not used, if a computer responds "Tokyo" to the question, "What is the capital of Japan?," it is judged that the computer has the capabilities of such a level. This is referred to as the "question-answering system." We have already completed a preliminary system of this type.

Our question-answering system retrieves promising data from the text files of an encyclopedia and several years' worth of newspaper articles, and collates them with the question before answering. For example, to the question, "The death of cells in what part of the brain is a symptom of Parkinson's disease?," the system finds text such as, "Parkinson's disease is thought to occur when dopamine (formed as a nerve transmission substance within substantia nigra cells) is exhausted by a change in the quality of melanin cells found in the substantia nigra of the midbrain." The system then gives the answer, "the substantia nigra."

At present, in the above case, our system roughly collates the question with the text file in the knowledge base before choosing a phrase, "the substantia nigra," that corresponds to the interrogative expression, "what part." This method may be too rough to provide an exact answer in many cases. In the future, we plan to improve the method by collecting equivalent expressions, such as "Change in the quality of cells" = "Death of cells," and "A disease appears due to Cause B" = "A symptom of a disease relates to Cause B." These expressions may allow changes in the expression of the question to make the collation more exact. For example, by applying the latter expression to the text file in the knowledge base, an answer to the above question, in sentence form, may be modified as follows:

"A symptom of Parkinson's disease relates to the death of melanin cells that lie in the substantia nigra of the midbrain." As "that lie in" is equivalent to "in," and "dying" already had to be changed to "death", this is further reduced as follows:

"A symptom of Parkinson's disease relates to the death of melanin cells in the substantia nigra of the midbrain." This sentence may be collated with text files in the knowledge base to produce a more exact answer in the form of the term, "substantia nigra." We are currently working to add other equivalent expressions by identifying other expressions.

3 Fundamental study of natural language

3.1 Study of natural language using neural networks

Neural networks were initially studied for the purpose of imitating the neural structure in the brain. Today, after the overwhelming popularity of the study, research has processed steadily in three areas: theory, brain and cognitive science, and engineering applications. Neural networks have been successfully applied in the area of pattern recognition, and have also made a significant contribution to the study of natural language, including language acquisition, knowledge representation, and disambiguation. Many natural language studies based on neural networks, which are usually called "connectionist models" or "connectionist approaches," are being carried out overseas. In Japan, on the contrary, such approaches are not taken seriously. In 1993, our natural language research group pioneered the study of natural language using neural networks. We have studied associative memories, knowledge representation, semantic maps, morphological analysis, and error detection in a large-scale corpus, all within the single framework of neural networks. In these studies, we aimed to advance our knowledge base while keeping our models at a level equal to or higher than that of various mechanical learning machines.

In the future, we plan to continue these studies and expand our techniques to the processing of natural language, including parsing and identification of name entities. We also intend to develop a learning machine that can acquire the knowledge and information required for various tasks in language processing. This may eventually be made possible through learning with high-accuracy automatic correction of erroneous language data in the real world, but after the above techniques have been merged with other mechanical learning methods.

3.2 Study of lexical semantics

To understand and generate sentences, a computer must properly store the knowledge on the language required for its processing. To understand a language, it is necessary to use this knowledge to determine how to produce the same semantic expression from different structures of sentences, and how to produce a semantic expression from a semantically ambiguous sentence, in accordance with the ambiguity of the context. An object of this study is to construct a dictionary that does not describe words statically or fixedly, but dynamically connects word meanings. We aim to construct a semantic dictionary that aids in the dynamic creation of lexical terms from those coded and stored in the dictionary, and includes a mechanism for systematically associating lexical terms. For this purpose, we must describe a lexical rule to allow a word to indicate its meaning in actual use. We will actively apply both linguistic and engineering knowledge in the continuation of this study.

To use a lexical rule, the “substantiality of the meaning of a word” must be clearly grasped in order to determine how the word functions in actual linguistic use. Without an exact semantic description, the dictionary will be less accurate, the semantic processing will not be performed properly, lexical creation will be excessive, and semantic ambiguities will persist. First, we will determine the substantiality of the meaning of a word through the use of linguistic knowledge.

We plan to conduct analyses from a linguistic standpoint, while at the same time verifying these analyses using the technology of a self-organizing neural-network model. We are currently using this model to construct a semantic map of Japanese nouns. On this map, Japanese nouns are placed near to or far from one another, depending of their degree of intimacy with respect to meaning. Nouns are linked with attributive modifiers (such as adjectives) closely associated with these nouns. Therefore, the entire semantic map contains a network of Japanese lexicons.

3.3 Study on the misuse of honorific expressions

Confusion regarding honorific expressions and changes thereto have been increasing in recent years. Here, we define the misuse of honorific expressions as the use of non-standard honorific expressions from a linguistic viewpoint. There are many types of such misuse, including simple errors in word form and misuse due to misunderstanding of the function of the expressions. The degree of the sense of incongruity caused by the misuse of an honorific expression may depend on the type of misuse. We plan to investigate and statistically analyze misuses of honorific expressions by type, as well as the extent of the unnaturalness of the misuse.

To be specific, after collecting and arranging misused honorific expressions, we will conduct a psychological experiment by the pair-comparison method and digitize the psychological impressions according to the naturalness or unnaturalness of the misused honorific expressions. We have thus far confirmed in a small-scale test that the degree of naturalness depends on the type of misuse and the attributes of the test subject. We feel that when this tendency is confirmed in a large-scale experiment, it may be applicable to an education system such as a system for learning honorific expressions.

In the future, the results of the large-scale experiment will be analyzed in detail for the above dependency, and will be compared with the reaction to the correct honorific expressions to examine 1) the dependence of learning honorific expressions on sex and age and 2) the degree to which each of the expressions is recognized.

3.4 Study of the communication process

We study the communication process in order to clarify the cognitive mechanism that allows communication to be flexible and robust. We will focus on the function of speech acts within a dialogue, based on a model for advancing a dialogue (the process of creating regularity in the conversation sequence and the mechanism of inference based on relevance in the dialogue), and on a cognitive model for dialogues (specifically, the planning of dialogue participants, and the context sensitivity of recognition and belief).

At present, we are studying 1) annotation of discourse-level tags to a spoken-language corpus and the construction of a dialogue corpus, 2) annotation and reexamination of speech-act tags and relation tags, 3) the natural behavior of a social agent in ongoing conversation, 4) the recording of three-person dialogue data and the construction of a three-person dialogue model, and 5) the fundamental framework for realizing education and learning through communication.

3.5 Study on sensitivity to sounds

Music and language are forms of communication that have different objects. The first object of a language is to precisely communicate thoughts, while one of the main objects of music is to artistically express intense feelings. Some languages involve difficulties in expressing emotions. For example, we cannot tell whether readers are able to relate to poetry and tankas written from the writer's viewpoint. In such a case, it may be necessary to add other information to the poem or tanka. As for spoken language, spoken words disappear immediately after the speaker speaks. Therefore, whether the listener truly understands what the speaker says depends on the skill of the speaker and the attention paid by the listener. To enhance a listener's understanding, the speaker places stresses by raising the level of his or her voice and repeating words. To make a stronger impression, the speaker increases the intensity of his or her speech. The listener observes a speaker's breathing and habits, which can be considered musical expressions in a sense. A musician playing an instrument can create strong feelings and emotions. The audience naturally synchronizes in harmony with the musician, catching the musician's intention.

In spite of the object of precisely communicating through the use of a spoken language, the feelings of the speaker may play a major role in the precision of communication, even if this is not apparent. This is also true of music. Therefore, we feel it may be possible to apply our past experiments and models in which music information was used as a variable to the study of sensitivity to sounds. Our future study is described below.

Spoken language in an academic meeting includes various conjunctions in speeches, to indicate a beginning, a change in topics, a closing, or the like. Some speakers repeat the same conjunction several times. Others have different speech characteristics. We intend to classify these characteristics into groups, since once we identify them, it may be easier to deal with spoken language. If we know in advance that a speaker frequently uses the conjunction "soredewa" at the beginning of a sentence, it may be easier to identify a change in topics by the speaker.

Adjectives are necessary for expressing complex feelings, and are considered a means of describing sounds and music. Today, music is often treated like a document, and hence it is sometimes necessary to access the information that it contains. In such a case, live or recorded music must be described. To aid in such description, we intend to study sensitivities to sounds from the viewpoints of musicology and systems development.

4 Development of practical systems

4.1 Study and development of adaptive communication technology

With the goal of developing an English-learning support system, we are compiling a learners' corpus and examining error tags. The learners' corpus was compiled based on interviews with Japanese learners of English. It is characterized by a judgment of the learners' English level. This study is under way as part of a technology transfer project of the Telecommunications Advancement Organization of Japan (TAO) titled, "Research and Development of Adaptive Communication Technology."

As a highly information-oriented society develops, flexible communication for the exchange of information is required between humans and between humans and computers. We must not mistake the slight misuse of proper words or erroneous expressions for the truth. We must not react improperly to such misuse or errors. It is important to infer what the speaker is really trying to say in order to establish a technology for continuous adequate communication, known as "adaptive communication technology." It is an object of our study to apply our developed NLP technology to the development of a technology effective to analyze error-containing English spoken by a Japanese person. Another object is to create and open to the public the database itself.

The techniques required for the adaptive communication technology include gaining an understanding of the intentions of utterances and the generation of summarized sentences. A core technology for these techniques is one that automatically parses a sentence into single words and adds such linguistic information to the sentence as the part of speech and the modifier of a word. Conventional techniques for the automatic addition of linguistic information are not effective as inputs when there are grammatical errors or when there is no superficially expressed intention of the speaker. Thus, it is difficult to apply these techniques to the adaptive communication technology. In this study, in order to transfer techniques to meet the above objects, we intend to do the following:

(1) Prepare and open to the public a database with attached information on error-containing utterance of Japanese speakers of English.
(2) Develop a technology effective in adding linguistic information to error-containing language
(3) Demonstrate that both the database and the linguistic information adding technology are effective

We used error-filled English language spoken by Japanese people as the subject of this study and development. When speakers have poor English-speaking abilities, they are seldom conscious of implied meanings in speech. Thus, their speech may contain many stereotypical errors involving the basic usage of words and grammar. These errors are easy to detect and classify into groups, and are therefore suitable for the quick production of results and the transfer of technology.

In this study and development, we will prepare and open to the public a database with attached information on error-containing utterances of Japanese speakers of English, and will simultaneously develop a technology effective in adding linguistic information to error-containing sentences. In this latter development, it would be time-consuming to manually correct each error in usage and grammar. It is thus necessary to automatically determine the correct grammar and usage of words from data on erroneously used words. The system we have developed is capable of efficiently acquiring information from a small amount of language data, and is able to add language-related information with a high degree of accuracy. In this study and development, we intend to extend and transfer this system.

4.2 Language processing in symbiotic communication

As information technology continues to develop and increase in popularity, the gap is expanding between IT-oriented people and those that cannot use IT. If everyone is to be able to freely use IT, learning support through human communication is required. Such communication differs entirely from the reading of a printed manual.

We aim to realize human-learning support systems with the aid of communication using a language, the non-verbal transfer of information using the eyes and hands, and a communication environment including an objective interface that provides a feeling of human interaction. The NLP technology in this system is directly related to the dialogue-processing and knowledge-processing section.

The dialogue-processing section combines non-verbal information with linguistic information to symbolize user input information in a text file. Response information prepared by the knowledge-processing section will be used to create a response in each medium. It will be necessary to study and develop a new conversation analysis technique using multi-modal information from the interface, in addition to conventional techniques. We intend to present adequate information based on a speaker model, and to effectively use multi-modal information.

The knowledge-processing section uses textbooks, manuals, newspaper articles, and the like as knowledge sources. It extracts and symbolizes information requiring a response by using symbolized information input as a retrieval key. We intend to develop a high-precision technique for extracting information from a large-scale document, and to develop techniques for analyzing text information with added multi-modal information and for producing other text information.

5 Concluding remarks

In this report, we described some of the research and development activities concerning natural languages being conducted by the Computational Linguistics Group. It includes a wide range of activities ranging from fundamental linguistic studies to the development of language-related systems. In the future, we intend to advance our current work in these respective fields.

***Hitoshi ISAHARA**, Dr. Eng.*

Leader, Computational Linguistics Group, Keihanna Human Info-Communications Research Center

Natural Language Processing