2-6 Development of Language Resources for Natural Language Processing in Deep Level

ZHANG Yujie, KURODA Kow, IZUMI Emi, and NOZAWA Hajime

Techniques for both text analysis and speech transcription are still in an unsatisfactory state even though they cannot be dispensed with Natural Language Processing. To achieve high performance techniques, a large amount of language resources are urgently required. To resolve this problem, Computational Linguistics Group at NICT has been constructing language resources of several sorts, targeting different cases of application. This paper presents some of such resources, including Corpus Annotated for Semantic Frames and their Elements, Japanese Learner's Corpus, and Japanese-Chinese Parallel Corpus.

Keywords
Natural language processing, Multi-planar/Multi-granular Semantic Frame Analysis, Semantic annotation of texts and dialogues, Learner corpus, Error analysis, Parallel corpora, Translation dictionary

1 Introduction

Natural language processing is an area of research that attempts to establish and apply language knowledge through the construction of mathematical models, with the aim of enabling computers to realize some or all of the language-processing capabilities demonstrated by human beings. Language knowledge can be acquired in two ways: by creating rules through human introspection and by using statistical methods to extract this knowledge automatically from corpora. Although corpora provide true language materials, in order to elucidate the language knowledge and linguistic phenomena hidden within the corpora, they need to be annotated with information at a deep level with respect to a variety of aspects. It is recognized in general that the more information with which a corpus is annotated, the more knowledge can be obtained from the corpus. Section **2**, **3**, and **4** below introduce the Multilayered/Multigranularity Semantic Annotation Corpus, the Japanese Learner Corpus, and the Japanese-Chinese Parallel Corpus — all of which are currently under development by the Computational Linguistic Group of the National Institute of Information and Communications Technology (NICT).

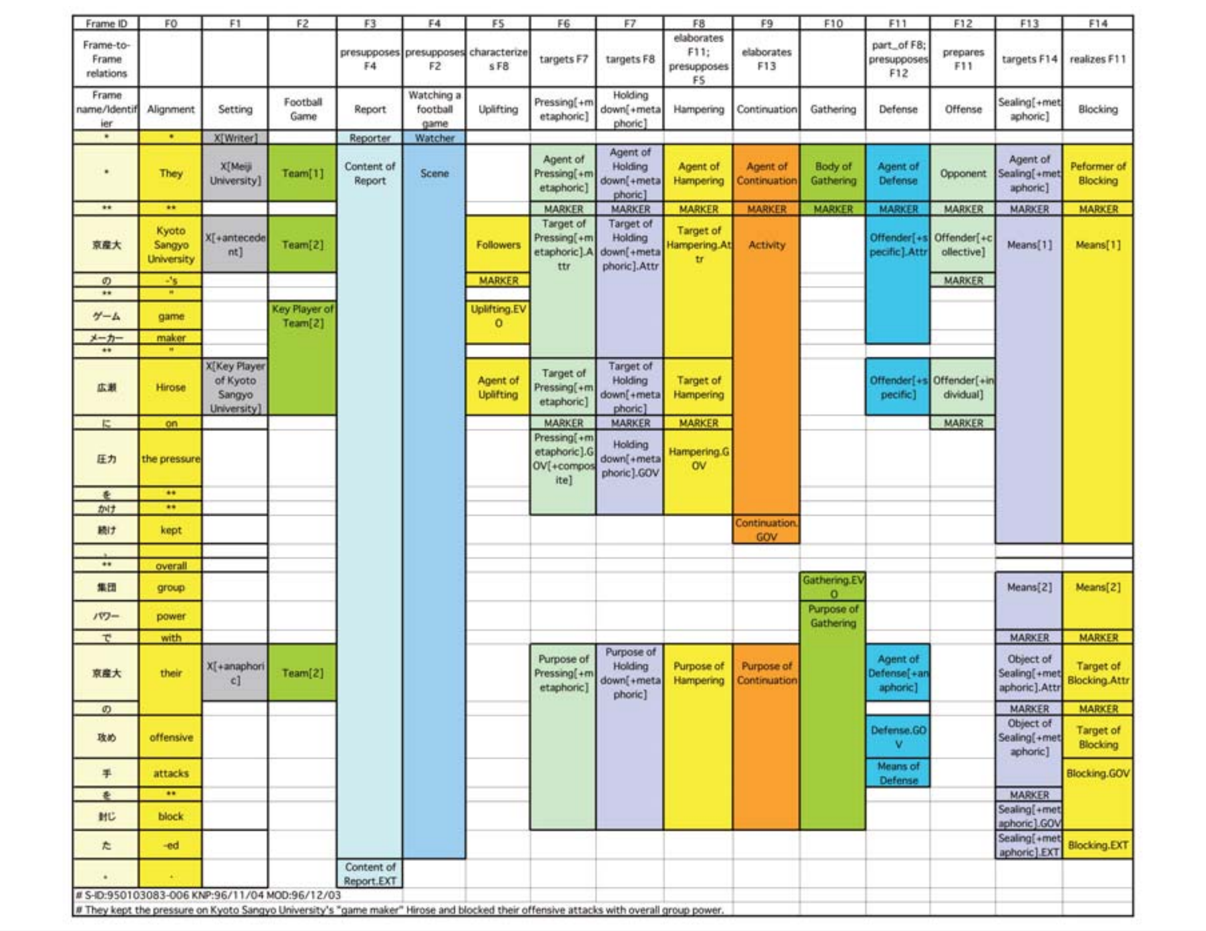
2 Development of semantic resources for written and spoken language, using MSFA and MIFA (Kuroda, Nozawa)

Development of linguistic resources began with the construction of dictionaries and grew into the development of annotated corpora. Although at present some available corpora are annotated with information on parts of speech and sentence structure, due to technological difficulties almost no corpora are annotated with semantic information. Our goal is to fill this gap. Annotation using Multilayered/Multigranularity Semantic Frame Analysis (MSFA)[2][3] and Multiplanar/Multigranular Interactional Frame Analysis (MIFA)[4], which is an extensive application of the former to analysis on discourse structures, are attempts at semantic annotation. The differences between these two types of analysis reflect differences in the texts studied by each. MSFA specializes more in written language, while MIFA specializes in spoken language. Below, Kuroda comments on activities in the area of written language, using MSFA, and Nozawa comments on activities in the area of spoken language, using MIFA.

2.1 Development of semantic resources for written language, using MSFA (Kuroda)

The goals of activities in the area of written language are to construct linguistic resources (in forms different from those of dictionaries) and steadily release the results to the public as samples. These resources are established based on multidimensional semantic analysis and annotation conducted by linguists on specific samples collected from a corpus. This semantic analysis is characterized by (i) its use of the MSFA method of analysis, which improves on the annotation specifications of Berkeley FrameNet[1], to employ semantic frames and their constituent elements (i.e., semantic roles) as semantic tags; and (ii) its employment of a cycle of updating the semantic tag structure and annotation, without (for the time being) fixing the tag structure. For reference, the (fragmentary) results of semantic annotation using MSFA are shown in Fig. 1.

Semantic annotation of “Using the power



Frame ID	F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14
Frame-to-Frame relations				presupposes F4	presupposes F2	characterizes F8	targets F7	targets F8	elaborates F11; presupposes F5	elaborates F13		part_of F8; presupposes F12	prepares F11	targets F14	realizes F11
Frame name/Identifier	Alignment	Setting	Football Game	Report	Watching a football game	Uplifting	Pressing[+metaphoric]	Holding down[+metaphoric]	Hampering	Continuation	Gathering	Defense	Offense	Sealing[+metaphoric]	Blocking
*	*	X[Writer]		Reporter	Watcher										
*	They	X[Meiji University]	Team[1]	Content of Report	Scene		Agent of Pressing[+metaphoric]	Agent of Holding down[+metaphoric]	Agent of Hampering	Agent of Continuation	Body of Gathering	Agent of Defense	Opponent	Agent of Sealing[+metaphoric]	Peformer of Blocking
**	**						MARKER	MARKER	MARKER	MARKER	MARKER	MARKER	MARKER	MARKER	MARKER
京産大	Kyoto Sangyo University	X[+antecedent]	Team[2]			Followers	Target of Pressing[+metaphoric].Attr	Target of Holding down[+metaphoric].Attr	Target of Hampering.Attr	Activity		Offender[+specific].Attr	Offender[+collective]	Means[1]	Means[1]
の	-'s					MARKER							MARKER		
**	"														
ゲーム	game		Key Player of Team[2]			Uplifting.EVO									
メーカー	maker														
**	"														
広瀬	Hirose	X[Key Player of Kyoto Sangyo University]				Agent of Uplifting	Target of Pressing[+metaphoric]	Target of Holding down[+metaphoric]	Target of Hampering			Offender[+specific]	Offender[+individual]		
に	on						MARKER	MARKER	MARKER				MARKER		
圧力	the pressure						Pressing[+metaphoric].GOV[+composite]	Holding down[+metaphoric].GOV	Hampering.GOV						
を	**														
かけ	**														
続け	kept									Continuation.GOV					
、															
**	overall														
集団	group										Gathering.EVO			Means[2]	Means[2]
パワー	power										Purpose of Gathering				
で	with													MARKER	MARKER
京産大	their	X[+anaphoric]	Team[2]				Purpose of Pressing[+metaphoric]	Purpose of Holding down[+metaphoric]	Purpose of Hampering	Purpose of Continuation		Agent of Defense[+anaphoric]		Object of Sealing[+metaphoric].Attr	Target of Blocking.Attr
の														MARKER	MARKER
攻め	offensive											Defense.GOV		Object of Sealing[+metaphoric]	Target of Blocking
手	attacks											Means of Defense			Blocking.GOV
を	**													MARKER	
封じ	block													Sealing[+metaphoric].GOV	
た	-ed													Sealing[+metaphoric].EXT	Blocking.EXT
。	.			Content of Report.EXT											

S-ID:950103083-006 KNP:96/11/04 MOD:96/12/03

They kept the pressure on Kyoto Sangyo University's "game maker" Hirose and blocked their offensive attacks with overall group power.

Fig.1 *Text from the Kyoto University Corpus (S-ID: 950103083-006 KNP: 96/11/04 MOD:96/12/03)*

of the team to make Kyoto Sangyo University take the offensive, by continuing to apply pressure to Kyoto Sangyo University's top player Hirose", using MSFA. With each column representing a frame, the information designated (i.e., frame elements) at the points where morpheme columns intersect with frame rows corresponds to semantic tags.

The semantic analysis realized by MSFA as used in Fig. 1 is not a case of assigning word meanings, chosen from an appropriate dictionary, to morphemes. Instead it represents a description (at a granularity more detailed than usual) of the content understood by people when reading or listening to text within a given context. This method is superior in avoiding the excessive particularity about formalization — expected to arise with simplistic implementation on computers — to which traditional semantic description is liable, particularly in regard to description of meanings formed through phrase units (i.e., superlexical units) not described in traditional lexicological resources.[2][3]

In the area of written language, semantic annotation was conducted for two groupings of text: (A) 63 sentences (three articles) from the Kyoto University Text Corpus (administered by the Kurohashi Laboratory at Kyoto University; for one more article, release has not been authorized by the copyright holder Mainichi Shimbun), and (B) 47 sentences (five stories) chosen from a Japanese-English parallel database (developed and administered by Masao Uchiyama of NICT). The results of this semantic annotation are available at the two sites listed below. (Some of the results have not been consolidated and are therefore not currently available.)

Site 1 : http://www.kotonoba.net/~focal/cgi-bin/hiki/hiki.cgi?FrontPage

Site 2 : http://www.kotonoba.net/~focal/cgi-bin/hiki2/hiki.cgi?FrontPage

The released portion of Dataset A on Site 2 features roughly 1,300 recognized frames, while Dataset B on Site 1 has about 400 recognized frames.

2.2 Development of semantic resources for spoken language, using MIFA (Nozawa)

In parallel with the annotation of written language using MSFA, in the area of spoken language an annotated corpus is being constructed using a similar framework, called MIFA. Intended to increase the variety of annotated corpora, this project was launched in the 2006 fiscal year. Simply put, MSFA is an annotation specification suited to written language, and MIFA is an annotation specification suited to spoken language.[4] Also with MIFA, no corpus suited to annotation is available; accordingly, a corpus is being prepared in advance for this framework as well. Under the MIFA framework, instead of using annotation specifications specialized in spoken language — as used in previous studies in language processing — an MSFA-compatible annotation structure is employed that can be shared with written-language corpora.

In the first year, the specifications required for handling spoken language were established through experimental annotation of the minutes of the Japanese Diet and the Corpus of Spontaneous Japanese. In 2007, using a research grant from the Hakuho Foundation to study language, culture, and education, annotation of actual commercial conversations started, employing multiple annotators. The goal was to construct an annotated corpus of small scale, but high granularity in both semantics and communicative interaction. We believe that a sufficient volume of highly granular descriptions of communicative interactions, available for the first time with MSFA and MIFA, is an essential element in the development of a properly functioning dialogue system.

3 Toward analysis of language, including errors — The NICT Japanese Learner English (JLE) Corpus (Izumi)

Together with globalization and the spread of information technology, the need to overcome language barriers is increasing. In considering teaching methods that will enable language learners to acquire language skills more effectively, it is vital for language teachers and researchers to ascertain the learner's developmental stages on an overall basis: learner's current proficiency levels and the stages through which they acquire higher abilities in the future. For this purpose, it is necessary to collect actual data of the learner language — to the extent required to measure both the universality and the diversity of learner language — if we are to discuss the subject based on reliable results of analysis. In addition, with advances in computer technology, new learning environments such as computer-aided language learning (CALL) and e-learning are becoming widely adopted. All of the foregoing developments entail requirements for support in communications between language learners and computers.

Traditionally, natural language processing research has essentially targeted processing grammatically and semantically correct semantics. However, in new environments such as CALL and e-learning, where human beings connect with computers, we see the increasing necessity of constructing a framework to analyze the errors included in nonnative speaker's speech to a sufficient degree. To this end, learner corpora consist of sample data of learner language behavior, including learner errors. This could well serve as an effective resource for advancing research into English-language education and second-language acquisition and for developing processing technologies of learner language. This chapter discusses research using the NICT Japanese Learner English (JLE) Corpus[5], a speech database of English-language learners whose mother tongue is Japanese.

3.1 Overview of the NICT JLE Corpus

The NICT JLE Corpus (Fig. 2) consists of text data transcribed from the audio recordings of English-language interview testing (15 minutes per person) of 1,281 learners whose mother tongue is Japanese (for a total of 2 million words). The data for each interview is assigned to one of nine proficiency levels. In addition, some of the data in this corpus (that for 167 speakers) has been annotated by hand with 47 types of error tags covering grammatical and lexical errors. Qualitative and quantitative analysis of learner's errors is one of most effective means of modeling learner's developmental stages. However, since the error tags cover relatively formal aspects of language, such as grammatical and lexical errors, we prepared two sub corpora to enable observation of learner language from a broader perspective. One is a back-translation corpus, consisting of Japanese translations of learner's English, inferring the speaker's intentions as much as possible. The primary goal of this corpus is to gauge the degree of mother-tongue interference in second-language acquisition, through observation using both back-translation and learner language with error tags. The other is a speech corpus of native English speakers. Comparing the speech of native speakers and learners could lead to knowledge that could not otherwise be observed with error tags alone, such as differences in frequency of certain lexical items and in the way speech proceeds.

3.2 Error analysis and applications in development of a learning-support system

We are describing learner language mainly by analyzing learner errors using the NICT JLE Corpus and developing a learning-support system using the resulting knowledge. In this analysis and development, our method is focused on the error analysis procedures that have been employed for many years in second-language acquisition research (Fig. 3).

First, "localizing errors" refers to identifying the errors in learner's speech, such as which words, phrases, grammatical constructs, and word orders are erroneous. "Categorizing errors" refers to categorization of these errors according to the ways their linguistic categories (e.g., morphemes, syntax, lexis) differ from correct usage. Next, "explaining errors"

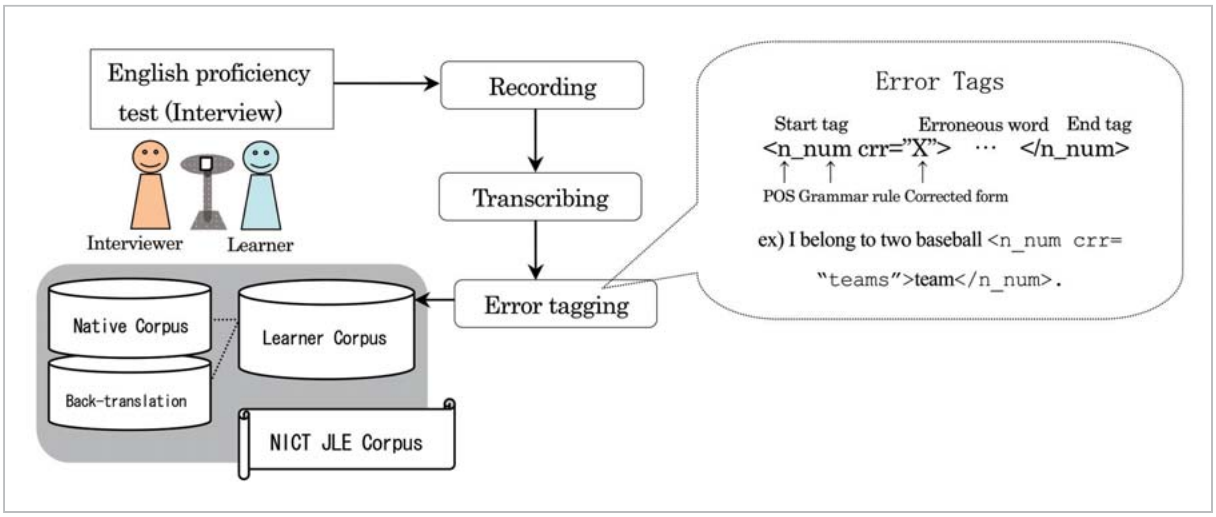


Fig.2 NICT JLE Corpus

refers to identifying the cause of the errors: why did a given error occur? Finally, "evaluating errors" means the evaluation of the ease of understanding sentences containing errors — in other words, identifying "error gravity" by categorizing them into those that hinder communication substantially and those without major effects on communication.

As a task corresponding to localizing and categorizing errors, we conducted various types of analysis of the characteristics of learner errors, based on the 47 types of error tags mentioned above. These included analysis of patterns (forms and contexts) by focusing on individual linguistic items, such as errors in articles or idioms[6], and extraction of acquisition order of key grammatical morphemes (Fig. 4 [7]. Based on the knowledge obtained through this analysis, we conducted experiments in automatic detection of errors using machine learning, leading to the development of the "Eden" (Error Detection in English) automatic error-detection system[8].

To date, analysis and experiments have focused on the accuracy of a certain target language, such as whether any errors were committed. Depending on the purpose of learning languages, precision may be essential; however, in language learning that gives priority to communication — a type of learning that has become mainstream in recent years — it is beneficial first to learn which aspects of language absolutely must be correct and which do not necessarily need to be correct. For this reason, as the second phase of using this corpus we focused on ease of communication

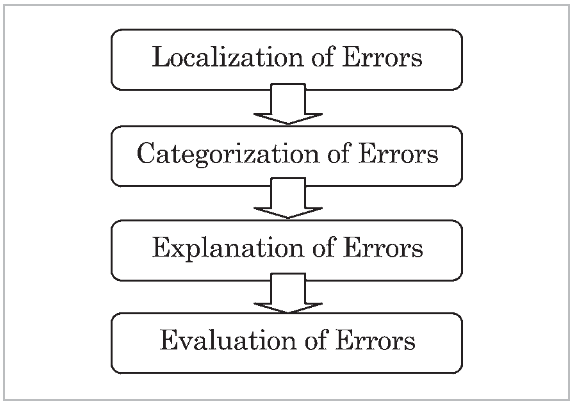
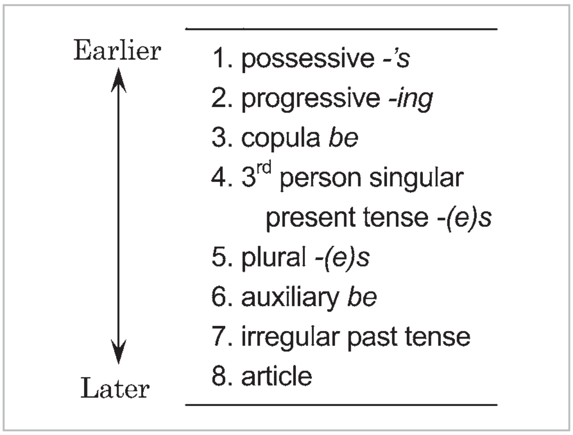


Fig.3 Error-analysis procedures



1. possessive *-'s*
2. progressive *-ing*
3. copula *be*
4. 3rd person singular present tense *-(e)s*
5. plural *-(e)s*
6. auxiliary *be*
7. irregular past tense
8. article

Fig.4 *Japanese English-language learner's acquisition order of key grammatical morphemes*

rather than accuracy and on analysis of learner language from more of a communicative perspective. This task can be considered to correspond to the error-analysis procedure “evaluating errors” in Fig. 3. First, we had native English speakers judge the data in the NICT JLE Corpus, determining which of three intelligibility levels (intelligible, unclear, unintelligible) best described each sentence. The results showed that the types and frequencies of errors in each sentence affected differences in intelligibility level between the sentences.[9] Using this relationship between error types and intelligibility, we also attempted experiments in automatic judgment of intelligibility level based on machine learning.[10] Although we were able to elicit judgments as to whether a sentence was intelligible on the one hand or unclear or unintelligible on the other at a minimum precision of 90%, judgment as to whether a sentence was “unclear” or “unintelligible” — a judgment that requires contextual-level information — proved extremely difficult.

In this section we have summarized the NICT JLE Corpus, a learner corpus that forms one of NICT’s language resources, one that may prove useful in analysis of sentences including errors — a category not previously covered by existing language-processing techniques. We also discussed analysis of learner language based on this corpus and development of processing technologies. In the future, we plan to continue these development efforts, to apply testing at a fundamental research level to practical-level technologies.

4 Constructing language resources for Japanese-Chinese machine translation (Zhang)

As part of its machine-translation research and development project, NICT is constructing a Japanese-Chinese Parallel Corpus and a Japanese-Chinese translation dictionary.

4.1 The Japanese-Chinese Parallel Corpus[11]

A parallel corpus is a collection of correspondences between different languages, consisting of document, paragraph, or sentence units. Since a parallel corpus can include correspondences between source and target languages at different levels, it is a language resource required for extracting translation knowledge in developing machine-translation systems. NICT began a project of constructing multilingual corpora including Asian languages several years ago, and is now constructing Japanese-English and Japanese-Chinese parallel corpora with standardized specifications. Following is an introduction to the Japanese-Chinese Parallel Corpus.

(1) Japanese-language data

The Japanese-language data consists of approximately 40,000 sentences excerpted from articles published in 1995 in the newspaper Mainichi Shimbun. Based on the specifications of the corpus of spontaneous Japanese, the Japanese-language data are annotated with information such as word segmentation, parts of speech, and syntactic-structure information.

(2) Chinese-language data

The Chinese-language data consists of translations of the above Japanese-language data into Chinese. The translation was performed by professional translators, based on the following standards:

(a) Translation was conducted in units of Japanese sentences.

(b) Priority was given to creating a translation very close in construction to the original.

(c) As needed to communicate the meaning of a sentence, supplementary information was taken from the preceding sentences. In particular, although the subject of a sentence is often omitted in Japanese, in Chinese the subject is necessary. In such a case, the subject would be added to the translation.

(d) As necessary for fluency and ease of reading, word order would be changed, commas inserted, and similar alterations made. To ensure the quality of translation, it would be examined by other translators and by a native Chinese speaker.

(3) Morphological information annotation to the Chinese text

The Chinese translation is annotated with word segmentation and parts of speech. The Chinese part-of-speech tag set defines 39 parts of speech. First, a morphological analysis tool is used for word segmentation and part-of-speech annotation. Next, the results are corrected by hand, with the assistance of a tool we developed to increase the efficiency of these corrections. In addition to editing functions, the tool has the following functions:

(a) It can look up words in the Grammatical Knowledge Base of Contemporary Chinese.

(b) It can look up words in the checked corpus and sort the results of this search according to the context before and after the word.

(c) It can make identical global changes to words with attributes that are the same as those of the most recently changed word. This tool offers a convenient means of ensuring consistency in word segmentation and part-of-speech annotation.

(4) Word alignment in the Japanese-Chinese Parallel Corpus[12]

In a large-scale Japanese-Chinese Parallel Corpus, alignment is required at the word and phrase level for the extraction of translation correspondences. In this research, we propose and evaluate a method of word alignment incorporating the strong points of both the statistics-based approach and the lexical-knowledge based approach.

(a) The lexical-knowledge based approach works in two steps: alignment based on lexical information and alignment based on dislocation. For each pair of any Japanese morpheme and any Chinese word, we attempt to estimate the possibility of each pair having translation correspondence and then decide on the most plausible choices. This estimation of possibility was performed using three types of lexical information: translation dictionaries, relationships between Japanese kanji and Chinese characters, and relationships between simplified and traditional Chinese characters. For Japanese morphemes and Chinese words that can not be aligned based on lexical information, dislocation-based alignment is employed. In general, it is often observed that when two words belonging to a grammatical component in the original text are translated, the corresponding translated words will also belong to the same grammatical component in the translation. Instead of grammatical structures, previously obtained alignments are used. In estimating the possibility of a translation correspondence between a Japanese morpheme j and Chinese word c, four alignments are used: the closest alignments coming before j and c and the closest alignments following j and c. The correspondence possibility is estimated by measuring the conditions of each distortion and distance between the prospective alignments between j and c and these four alignments.

(b) The existing GIZA++ tool was used for the statistics-based approach.

(c) Integrated approach

We examined an approach of integrating both the lexical-knowledge based approach and the statistics-based approach. Specifically, we use a majority decision on three groups of alignment results which are produced by the lexical-knowledge based approach, GIZA++ application in the Chinese-to-Japanese direction and Japanese-to-Chinese direction. To evaluate this approach, we extracted 1,127 sentence correspondences from the Japanese-Chinese Parallel Corpus and annotated these with word-to-word correspondences by hand. Experimental results show that the recall rate increased to 63% and precision to 79%. These results demonstrate the effectiveness of using the lexical-knowledge based approach and the statistics-based approach simultaneously. Results obtained automatically are then corrected by hand.

4.2 Japanese-Chinese Translation Dictionary[13]

A translation dictionary is a language resource essential to machine translation and cross-language information retrieval. We are constructing Japanese-Chinese translation dictionary based on the EDR Japanese-English Dictionary[14]. As the achievement of this work, a trilingual Japanese-English-Chinese translation dictionary will be obtained. Each record in the EDR Japanese-English Dictionary is to be annotated with the following information:

(a) A Chinese translation based on each record's <semantic information>, (b) part-of-speech or grammatical-category information on each Chinese translation, (c) register information for each Chinese translation, (d) a Chinese translation of the <Japanese conceptual headword>, (e) a Chinese translation of the <Japanese conceptual description>, and (f) variations of the Chinese translation.

Resister information includes the following content:

(1) Type (e.g., word-to-word translation, explanation, or transliteration), (2) style (i.e., archaic, colloquial, literary, or slang), and (3) formality (i.e., honorific, humble, or belittling).

In annotation of part-of-speech and grammatical-category, we will first determine whether the Chinese translation is a word or a phrase, and then annotate the translation with the corresponding category. We have defined 18 types of parts of speech and 11 types of phrase grammatical categories.

The work began in the 2005 fiscal year. To achieve high quality and reduce costs, we are depending on the experts within China, including skilled translators and experts in Chinese language processing. At present, we have completed approximately 140,000 records including high-frequency Japanese words (i.e., those overlapping with the Japanese word list of JUMAN). We plan to complete the project and release the results by the end of the 2007 fiscal year.

5 Conclusions

With the spread of the Internet, natural language processing technologies are becoming ever more important. Applications of natural language processing include machine translation and assistance in language learning. Corpora deeply annotated with information, such as those introduced in this paper, promise to contribute to the establishment of natural language processing to a similarly deep degree, with corresponding improvements in the performance of natural language processing applications.

References

1 Johnson, C. R. and Fillmore, C. J, "The FrameNet tagset for frame-semantic and syntactic coding of predicate-argument structure", In Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL 2000), pp.56-62, 2000.

2 Kuroda, K. and Utiyama, M. and Isahara, H., "Getting deeper semantics than Berkeley FrameNet with MSFA", In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-06), P26-EW, 2006.

3 K. Kuroda and H. Isahara, "Linking Natural Language to Semantic Knowledge using Multilayered Semantic Frame Analysis", Technical Report of IEICE, 104 (416), pp.65-70, 2004.

4 H. Nozawa, K. Kuroda and H. Isahara, “Meaning and Discourse Function in Dialogue Texts — An Approach from Multi-layered Semantic Frame Analysis —”, Technical Report of IEICE, 106 (299), pp.33-38, 2006.

5 Emi Izumi, Kiyotaka Uchimoto and Hitoshi Isahara, Nihonjin 1200-ninn no Eigo Speaking Corpus (Spoken Corpus of 1200 Japanese Learners of English), ALC Press, 2004 (in Japanese).

6 Emi Izumi, Toyomi Saiga, Thepchai Supnithi, Kiyotaka Uchimoto and Hitoshi Isahara, “Error tag tsuki nihon-jin eigo gakusyu-sya hatsuwa corpus wo mochiita gakusyusya no kanshi syutoku keiko no bunseki (Analysis of Japanese learner’s acquisition of English article based on the learner corpus)”, In Proceedings of the 9th Annual Meeting of the Association for Natural Language Processing, pp.19-22, 2003 (in Japanese).

7 Emi Izumi, Kiyotaka Uchimoto, and Hitoshi Isahara, “Error tag tsuki nihon-jin eigo gakusyu-sya hatsuwa corpus wo mochiita nihon-jin eigo gakusyu-sya no syuyou bunpou keitaiso no syutoku junjo ni kansuru bunseki (Corpus-based analysis of Japanese learners’ acquisition order of major English grammatical morphemes)”, In Journal of Natural Language Processing, vol.12, no.4, 2005.

8 Izumi, E., Saiga, T., Uchimoto, K., Supnithi, T., and Isahara, H. “Automatic error detection in the Japanese Learner's English spoken data”, In Companion Volume of the Proceedings of the Association of Computational Linguistics (ACL) 03. Japan, pp.145-148, 2003.

9 Emi Izumi, Kiyotaka Uchimoto, and Hitoshi Isahara, “Nihon-jin eigo no tsujiyasusa ni kannsuru kennkyu (Investigation of intelligibility of Japanese learner English)”, In Proceedings of the 12th Annual Meeting of the Association for Natural Language Processing, 2006 (in Japanese).

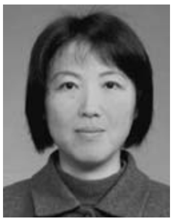
10 Izumi, E., Uchimoto, K., and Isahara, H., “Measuring intelligibility of Japanese learner English, Salakoski, T., Ginter, F., Pyysalo, S., & Pahikkkala, T. (Eds)”. Advances in Natural Language Processing, 5th International Conference on Natural Language Processing (FinTAL), Lecture Note in Artificial Intelligence, Springer, Berlin, pp.476-487, 2006.

11 Zhang, Y., Uchimoto, K., Ma, Q., and Isahara, H., “Building an Annotated Japanese-Chinese Parallel Corpus — A Part of NICT Multi lingual Corpora”, In the Tenth Machine Translation Summit Proceedings, pp.71-78, 2005.

12 Zhang, Y., Liu, Q., Ma, Q., and Isahara, H., “A Multi-aligner for Japanese-Chinese Parallel Corpora”, In the Tenth Machine Translation Summit Proceedings, pp.133-140, 2005.

13 Zhang, Y., Ma, Q., and Isahara, H., “Automatic Construction of Japanese-Chinese Translation Dictionary Using English as Intermediary”, Journal of Natural Language Processing, 12(2), pp.63-85, 2005.

14 NICT (National Institute of Information and Communications Technology), “EDR Electronic Dictionary Version 2.0 Technical Guide”, 2002.



***ZHANG Yujie**, Ph.D.*

Expert Researcher, Computational Linguistic Group, Knowledge Creating Communication Research Center (former: Expert Researcher, Computational Linguistic Group, Keihanna Human Info-Communication Research Center, Information and Communications Department)

Computational Linguistics

***KURODA Kow**, Ph.D.*

Expert Researcher, Computational Linguistic Group, Knowledge Creating Communication Research Center (former: Expert Researcher, Computational Linguistic Group, Keihanna Human Info-Communication Research Center, Information and Communications Department)

Linguistics, Cognitive Science

***IZUMI Emi**, Ph.D.*

Expert Researcher, Computational Linguistic Group, Knowledge Creating Communication Research Center (former:Expert Researcher, Computational Linguistic Group, Keihanna Human Info-Communication Research Center, Information and Communications Department)

Corpus Linguistics, Second Language Acquisition

NOZAWA Hajime

Expert Researcher, Computational Linguistic Group, Knowledge Creating Communication Research Center (former: Expert Researcher, Computational Linguistic Group, Keihanna Human Info-Communication Research Center, Information and Communications Department)

Cognitive Linguistics, Pragmatics