# A Corpus-based English Language Assistant to Japanese Software Engineers

Masumi Narita
Software Research Center, Ricoh Co., Ltd.
1-1-17 Koishikawa, Bunkyo-ku, Tokyo, Japan
narita@src.ricoh.co.jp

**Abstract**

This paper presents how we developed an English abstract writing tool for Japanese software engineers, the "Abstract Helper," based on our annotated English-to-Japanese parallel corpus of sample abstracts for research papers. The main purpose of this writing tool is to help Japanese software engineers improve the organization of their writing by enabling them to access and 'borrow' good models of English abstracts. To create this kind of language assistance, we built an E-J parallel corpus of 539 sample abstracts and annotated the corpus with textual and linguistic information focusing on the rhetorical structure of the sample abstracts. By using this annotated corpus as the core language resource, the "Abstract Helper" is shown to be effective at providing users with both discourse-level guidance and sentence-level assistance. We also present some of the user feedback we have gathered at preliminary user trials and discuss our outlook for the further development of the "Abstract Helper."

## 1. Introduction

In today's Information Society, Japanese researchers are required to overcome the language barrier between English and Japanese and achieve better communication in the global community. This also means that we need to improve the quality of the information flow while managing English writing time more effectively than ever before.

To help produce documents in a foreign language, various kinds of language support utilities have been proposed so far. These language assistants include MT (Machine Translation) systems, spelling/grammar checkers (Golding and Schabes, 1996; Jones and Martin, 1997) and writer's workbenches using MT technology (Johnson, 1997; Yamabana *et al.*, 1998). Others have taken a more corpus-based approach (Yamamoto and Kitamura, 1999), which uses a bilingual corpus and NLP techniques to provide word-level, phrase-level or sentence-level translation examples relevant to the writer's intended message. Still others (Shibata and Itoh, 1999) provide a real-time, predictive word look-up from an English-to-Japanese dictionary when writing in English.

These conventional writing tools are useful when producing a target sentence, but they have their limitations since they do not provide discource-level assistance. We think that discourse-level assistance is the most indispensable to improving English documents produced by Japanese authors. This becomes clear when we consider our English writing problems identified by foreign readers such as poor organization, unclear logic and focus, as well as poorly constructed sentences. From his experiences in correcting the English papers submitted by Japanese physicists, Leggett (1966) also notes that 'Japanese English' often seems vague and diffuse because the argument does not run in a logical sequence.

Since our writing problems all relate to content, we have developed a computer-assisted English writing tool, the "Abstract Helper," which is aimed at

helping Japanese software engineers improve the content, in particular, the organization of their writing. Among other types of documents, abstracts for research papers were selected as the target document for our tool because they are written in a concise, logical, and coherent sequence, and thus have the type of organization that is crucial to efficiently producing high-quality documents.

Our approach is different from conventional ones in that it focuses on the rhetorical structure of English abstracts to help produce well-organized abstracts, as well as well-formed English sentences. The "Abstract Helper" encourages users to access and 'borrow' a good model or a good outline of their target abstract and then to flesh out the outline in order to present their original ideas. To develop this kind of TL (Target Language)-driven user assistance, we built an English-to-Japanese parallel corpus of sample abstracts. We also annotated this bilingual corpus with some textual and linguistic information after analyzing the sample abstracts in terms of their textual structure and logical sequence. Once this corpus was built as the core language resource, the "Abstract Helper" was easily designed to help users be more productive in writing English.

The rest of the paper is organized as follows. In Section 2, we describe language resources for the "Abstract Helper," in particular, how we built an annotated E-J parallel corpus of sample abstracts as the core language resource. In Section 3, we provide a system overview of the "Abstract Helper." In Section 4, we present some of the user feedback we have gathered at preliminary user trials. Finally, in Section 5 we discuss our outlook for the further development of the "Abstract Helper."

## 2. Language Resources for the "Abstract Helper"

### 2.1 The Core Language Resource: Annotated E-J Parallel Corpus of Sample Abstracts

Building a corpus of well-organized sample English abstracts was the key to developing our TL-driven user assistance. Since the "Abstract Helper" is targeted at Japanese software engineers including our colleagues at Ricoh, we decided to collect sample abstracts from widely known technical journals and conference proceedings in the domain of information engineering.

After receiving permission to use the abstracts for research purposes, we collected a total of 539 sample English abstracts from three sources as shown in Table 1. The abstracts from the ACL conference proceedings were not available in electronic form, so they were manually typed into a computer.

| Source of Sample Abstracts | Form of Samples | No. of Samples |
|---|---|---|
| IEEE[1] Transactions on Pattern Analysis and Machine Intelligence | Electronic | 285 |
| Proceedings of the Annual Meeting of ACL[2] | Paper-printed | 218 |
| IEEE Multimedia | Electronic | 36 |

Table 1: Structure of our Sample Abstracts

We then prepared Japanese translations of sample English abstracts to make it easier for users to search for a good model for their writing. Japanese equivalents were constructed on a sentence-to-sentence basis by Ricoh's software engineers,

thereby manually aligning English-Japanese sentence pairs of sample abstracts. Since these engineers are well informed about the topic areas, they could produce high-quality Japanese equivalents.

As described by Narita (1999a, 1999b), we also designed our corpus to be annotated with the following information in an SGML-conformant way:

(1) Text features – internal organization of each sample abstract
(2) Bibliographic information about each sample abstract
(3) Linguistic information on each sample abstract
  (3-1) Abstract types
  (3-2) Organizational-scheme types
(4) Linguistic information on each sample sentence
  (4-1) Sentence roles
  (4-2) Verb complementation pattern(s)

Figure 1 shows a fragment of our manually tagged E-J parallel corpus of sample abstracts. Tables 2 and 3, respectively, show tagsets of abstract and organizational-scheme types. Abstract types represent what the authors intend to convey in their papers and are classified into 5 categories. Organizational-scheme types represent the location of the topic sentence in an abstract and are classified into 4 categories. Most of our sample abstracts were written in one paragraph and thus typed as S001, S002 or S003. Those abstracts consisting of two or more paragraphs were uniformly tagged as S004, regardless of the position of the topic sentence.

<abs id=A097 type=T001 str=S003> ← **Abstract Type, Organizational-scheme Type**
<issue id=J003>Proc. 30th Annual Meeting of the ACL, 1992.</issue>
<title id=EA097-t>Prosodic Aids to Syntactic and Semantic Analysis of Spoken English</title>
<title id=JA097-t>音声英語の統語的・意味的分析を助ける韻律的情報</title>
<author>Chris Rowles and Xiuming Huang</author>
<keyword>spoken English, lexical ambiguity, structural ambiguity, prosodic information, syntactic analysis, semantic analysis, parsing</keyword>
 ↓ **Sentence Role** ↓ **Verb Complementation Pattern**
<s id=EA097-1.1 role=R010>Prosody can be useful in [resolving@@NP1 resolve NP2@@] certain lexical and structural ambiguities in spoken English. </s>
<s id=JA097-1.1 role=R010>韻律は音声英語の語彙的・構造的曖昧性の解決に役立つ場合がある。</s>
<s id=EA097-1.2 role=R020>In this paper we [present@@NP1 present NP2@@] some results of [employing@@NP1 employ NP2@@] two types of prosodic information, namely pitch and pause, to [assist@@NP1 assist NP2@@] syntactic and semantic analysis during parsing.</s>
<s id=JA097-1.2 role=R020>本論文では、二種類の韻律情報、すなわちピッチとポーズを構文解析時の統語的および意味的解析に役立てるよう取り入れた結果を示す。</s>

</abs>

Fig. 1: A Fragment of our Annotated E-J Parallel Corpus of Sample Abstracts

| Tag | Category |
|---|---|
| T001 | Proposals of new systems/models/algorithms |
| T002 | Technical surveys |
| T003 | Improvements on existing techniques |
| T004 | Reviews of papers |
| T005 | Reports on state-of-the-art technology |

Table 2: Our Tagset of Abstract Types

| Tag | Category |
|---|---|
| S001 | Abstracts starting with the topic sentence |
| S002 | Abstracts with the topic sentence in the middle |
| S003 | Abstracts ending with the topic sentence |
| S004 | Multi-paragraph abstracts |

Table 3: Our Tagset of Organizational-scheme Types

| Tag | Category |
|---|---|
| R010 | Introductory sentence |
| R020 | Topic sentence |
| R030 | Explanatory sentence |
| R031 | Verifying sentence |
| R032 | Supplementary sentence |
| R040 | Concluding sentence |
| R041 | Closing sentence |

Table 4: Our Tagset of Sentence Roles

According to Shinoda's study (1981), the main idea of the abstract is described by the topic sentence and other sentences should have their respective logical relationships with the topic sentence to present ideas in a coherent sequence. Our analysis of the logical relationships between the topic sentence and other component sentences of each sample abstract has led to the classification of sentence roles into 7 categories as shown in Table 4. Note, however, that all of these seven sentence roles are not always linked together in this order within an abstract, although the topic sentence is a necessary component of the abstract.

Narita (1997, 1998) showed that Japanese authors have problems with constructing English sentences and need guidance about the possible grammatical constructions of a given verb. This kind of guidance is also vital for the "Abstract Helper" because users often need information on grammatical constructions even after they have retrieved a model sentence from our corpus. Thus, we annotated each sample sentence with one or more verb complementation patterns based on the COMLEX Syntax V2.2, a computational lexicon which was developed by Grishman *et al.* (1994) at New York University.

### 2.2 Database of Verb Complementation Patterns

Since our corpus of sample abstracts was annotated with verb complementation patterns, we extracted this information to automatically build a separate lexical database of verb complementation patterns with their frequency counts in our corpus. This lexical database is linked to our corpus of sample abstracts so that users can retrieve sample sentences with a specified verb complementation pattern.

### 2.3 Database of English Collocations

Language learners always face the problem of the collocational usage of words in their target language. Among various types of collocates, the "Abstract Helper" was designed to provide three types of collocational information: (1) Adjective + Noun, (2) Noun 1 + Preposition + Noun 2, and (3) Noun + Preposition + V-ing.

We collected collocational information in two steps. First, we parsed our sample English abstracts with the Apple Pie Parser developed by Sekine and Grishman (1995) and automatically extracted candidate collocational patterns which satisfied our pattern matching rules for noun phrases. Second, we manually checked all the candidates and singled out the correct patterns. When users are given a database of English collocations, they can easily find the target word that is likely to co-occur with their input word.

## 3. System Overview of the "Abstract Helper"

We developed a prototype of the "Abstract Helper" on Sun SparcStation 20 using the Mule editor as a user interface. Target users of this writing tool are Japanese software engineers who are intermediate to advanced EFL learners.

The "Abstract Helper" has four major search engines: (1) the Sample Abstract Search engine, (2) the Sample Sentence Search engine, (3) the Sentence Pattern Search engine, and (4) the Collocation Search engine. When in operation, these engines access their respective language resources we constructed as shown in Fig. 2.

| Assistance by the "Abstract Helper" | | Language Resources Used |
|---|---|---|
| (1) Sample Abstract Search | <----------> | Annotated E-J Parallel Corpus of Sample Abstracts |
| (2) Sample Sentence Search | <----------> | Annotated E-J Parallel Corpus of Sample Abstracts |
| (3) Sentence Pattern Search | <----------> | Database of Verb Complementation Patterns |
| (4) Collocation Search | <----------> | Database of English Collocations |

Fig. 2: Search Engines and Language Resources of the "Abstract Helper"

The "Abstract Helper" works as follows. As a first step, when the "Sample Abstract Search" is called for by users on the Mule editor, it prompts them to select one from three sources of sample abstracts. Then a new window is opened to encourage users to specify both an abstract type and an organizational-scheme type from the list on the menu. When both an abstract type and an organizational-scheme type are specified, sample abstracts of specified types are retrieved from our E-J parallel corpus and displayed on the window one at a time. Users can find a good model of their target abstract by scanning each of the retrieved sample abstracts.

Users can thus start writing on the editor by copying and modifying the sample abstract which they have chosen as a good model for their own. Note, however, that this sample abstract functions only as an outline of the target abstract so that users need to flesh out the outline to present their original ideas. In so doing, users can call for the "Sample Sentence Search" to find and 'borrow' sample sentences playing a sentence role as desired. When they also need syntactic or lexical information to build up the target sentence, they can call for the "Sentence Pattern Search" or the "Collocation Search," respectively.

In short, the "Abstract Helper" gives discourse-level assistance, not only by enabling users to access and 'borrow' a sample abstract, which has the organization they consider to be suitable for their target abstract, but also by helping users build up each component sentence in a logical and coherent sequence. It also provides sentence-level assistance when users need information on possible complementation patterns of a given verb or collocational information.

## 4. User Feedback

We asked 3 software engineers to use the "Abstract Helper" for a month in order to gather user feedback. User feedback was obtained by two means: (1) questionnaire measure of user satisfaction and (2) asking users to write their reactions to our tool in their own words. To measure user satisfaction, we designed a questionnaire where we included several factors that affected computer user satisfaction. Our trial users were asked to rate their satisfaction on a five-point scale (-2, -1, 0, +1, +2) so that an individual's feeling can be placed somewhere between a "most positive" reaction and a "most negative" reaction.

Tables 5 and 6 show the average rating for each factor and the users' perceived utility of the software component, respectively.

| Factor | Average Rating |
|---|---|
| General Impression | + 0.33 |
| Accessibility | + 0.67 |
| Format of Output | + 1.33 |
| Effectiveness of Samples | + 1.67 |
| Volume of Samples | + 0.33 |
| Operation Manual | - 0.33 |
| Response Time | + 2.00 |

Table 5: User Satisfaction with the "Abstract Helper"

| Software Component | Average Rating |
|---|---|
| Sample Abstract Search | + 1.50 |
| Sample Sentence Search | + 1.00 |
| Sentence Pattern Search | 0.00 |
| Collocation Search | + 0.33 |
| Templates | + 1.33 |
| Bad Examples | + 1.33 |

Table 6: Users' Perceived Utility of Software Component

The factors "Response Time" and "Effectiveness of Samples" were fairly positively evaluated by our trial users, whereas our "Operation Manual" turned out to be less user-friendly. Among our software components, the "Sample Abstract Search" was most positively evaluated. Unexpectedly, however, the "Sentence Pattern Search" was evaluated as neutral. Our oral interviews with users made it clear that they did not heavily access this function because the "Sample Sentence Search" gave them enough information to construct the target sentence.

In Section 3, we did not explain about the "Templates" and "Bad Examples" functions because these are still under construction. The "Templates" function provides some ready-made templates of English abstracts in a "fill-in-the-blank" fashion in order to help users with low proficiency in English. On the other hand, the "Bad Examples" function gives negative evidence of word usages in English so that users can avoid lexical errors common to Japanese EFL learners. It is interesting to note that these 'underdeveloped' functions were evaluated as fairly positive.

As mentioned earlier, we also asked our trial users to write their reactions to the "Abstract Helper" in their own words. Their comments and requests are summarized as follows:

- The "Sample Abstract Search" was the most helpful and most heavily accessed in writing.
- They want to access corpora from a wider range of information science domains.
- They want to access a larger number of samples within a specific domain.
- They need a Japanese-to-English lexical look-up function and a function to facilitate the transition between sentences in terms of information structure.
- A Windows-based version of the tool should be produced.

These statements suggest that we should produce a Windows-based version of the tool which can provide a larger number of samples in a wider range of domains.

## 5. Conclusion and Future Work

We developed a prototype of the "Abstract Helper," which currently runs on the Sun workstation. The "Abstract Helper" provides users with relevant information to help them produce a well-organized abstract, as well as well-formed English component sentences in the abstract. To give users both discourse-level and sentence-level assistance in an organized way, our newly developed E-J parallel corpus of sample abstracts plays an essential role as the core language resource for this writing tool. We believe that the more Japanese authors become aware of the typical organization of an abstract in well-written samples, the better they will be able to incorporate intersentential relationships into their own writing.

To make it useful to a much broader community of Japanese software engineers, we will continue to improve the "Abstract Helper" by gathering more substantial feedback on its functionality from our trial user group at the R&D division of Ricoh. We will also make a plan to produce a Windows-based version of our tool. Moreover, in order to efficiently broaden the coverage of our language resources, we will work on developing the possibility of semi-automated corpus tagging, based on our experiences in manual tagging.

## Notes

[1] Institute of Electrical and Electronics Engineers
[2] Association for Computational Linguistics

## References


Golding, R. and Schabes, Y. (1996), 'Combining Trigram-Based and Feature-Based Methods for Context-Sensitive Spelling Correction', in Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, pp. 71-78.

Grishman, R., Macleod, C. and Mayers, A. (1994), 'Comlex Syntax: Building Computational Lexicon', in Proceedings of COLING-94, pp. 268-272.

Johnson, I. (1997), 'Personal Translation Applications', in Translating and the Computer, pp. 37-50.

Jones, M. P. and Martin, J. H. (1997), 'Contextual Spelling Correction Using Latent Semantic Analysis', in Proceedings of the 5th Conference on Applied Natural Language Processing of the Association for Computational Linguistics, pp. 166-173.

Leggett, A. J. (1966), 'Notes on the Writing of Scientific English for Japanese Physicists', in Journal of the Physical Society of Japan, pp. 790-805.

Narita, M. (1997), 'Error Analysis of English Sentences Composed by Japanese University Students', in Proceedings of the 36th Annual Meeting of the Japan Association of College English Teachers, pp. 69-72.

Narita, M. (1998), 'Language Resources for Writer's Helper', in Proceedings of the 1st International Conference on Language Resources and Evaluation, pp. 269-273.

Narita, M. (1999a), 'Constructing a Tagged E-J Parallel Corpus of Sample Abstracts', in Proceedings of the 5th Annual Meeting of the Association of Natural Language Processing, pp. 173-176.

Narita, M. (1999b), 'Construction of English Abstract Writing Tool', in Grant-in-Aid for COE Research Report (2): Researching and Verifying an Advanced Theory of Human Language, Kanda University of International Studies, pp. 807-819.

Sekine, S. and Grishman, A. (1995), 'A Corpus-based Probabilistic Grammar with Only Two Non-terminals', in Proceedings of the 4th International Workshop on Parsing Technologies, pp. 216-223.

Shibata, M. and Itoh, H. (1999), 'An Editor with a Simple Artificial Brain Agent and a Search Agent for an E-J Dictionary', in Technical Report of the Institute of Electronics, Information and Communication Engineers, pp. 15-20.

Shinoda, Y. (1981), Technical English, Nanun-Do: Tokyo.

Yamabana, K., Kamei, S., Doi, S., and Muraki, K. (1998), 'An Interactive English Writing Support Platform with Translation-aid and Information-Access Functions', in Proceedings of JSPS-HITACHI Workshop on New Challenges in Natural Language Processing and its Application, pp. 128-132.

Yamamoto, H. and Kitamura, M. (1999), 'Corpus Based Natural Language Processing and an Education System Using it', in Journal of Japanese Society for Information and Systems in Education, pp. 43-50.