Constructing a Tagged E-J Parallel Corpus for Assisting Japanese Software Engineers in Writing English Abstracts

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 constructed a tagged E-J parallel corpus of sample abstracts, which is the core language resource for our English abstract writing tool, the "Abstract Helper." This writing tool is aimed at helping Japanese software engineers be more productive in writing by providing them with good models of English abstracts. We collected 539 English abstracts from technical journals/proceedings and prepared their Japanese translations. After analyzing the rhetorical structure of these sample abstracts, we tagged each sample abstract with both an abstract type and an organizational-scheme type. We also tagged each sample sentence with a sentence role and one or more verb complementation patterns. We also show that our tagged E-J parallel corpus of sample abstracts can be effectively used for providing users with both discourse-level guidance and sentence-level assistance. Finally, we discuss the outlook for further development of the "Abstract Helper."

1. Introduction

The advent of the Information Society has changed our way of communication in such a way that we need to manage English reading/writing time more effectively than ever before. Consequently, Japanese people are required to overcome the language barrier between English and Japanese to achieve better communication in the global community.

In terms of English writing, it is frequently said that English documents produced by Japanese authors are often difficult to understand. The main problems foreign readers identify in these documents are: (1) poor organization, (2) unclear logic, (3) unclear focus, and (4) poorly constructed sentences. Moreover, from his experiences in correcting the English papers written by Japanese physicists, Legett (1966) notes that 'Japanese English' often seems vague and diffuse because the argument does not run in a logical sequence.

Since these problems all relate to content, we developed a computer-assisted English writing tool, the "Abstract Helper," which is aimed at helping the writer improve the content. Among other types of documents, paper abstracts for technical journals 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.

Previous research on computer-assisted foreign language production has primarily focused on the development of MT (Machine Translation) systems, spelling/grammar checkers (Golding and Schabes, 1996; Jones and Martin, 1997) and writer's workbenches using the 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, phraselevel 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. Our approach differs from these in that it focuses on the rhetorical structure of English abstracts to help produce well-organized abstracts, as well as well-formed English sentences. With the "Abstract Helper," users are encouraged to search for a good model or a good skeleton of their target abstract from our tagged E-J parallel corpus of sample abstracts by selecting the specific type of organization required. They are also encouraged to flesh out the selected skeleton with sentence-level assistance from the "Abstract Helper." Thus, our writing tool is aimed at helping users help themselves to be more productive in English writing.

The rest of the paper is organized as follows. In Section 2, we describe how we constructed a tagged E-J parallel corpus of sample abstracts, which is the core language resource for the "Abstract Helper." In Section 3, we give a brief overview of the "Abstract Helper." In Section 4, we summarize our approach. Finally, in Section 5 we discuss the outlook for further development of the "Abstract Helper."

2. Constructing a Tagged E-J Parallel Corpus of Sample Abstracts

Since the "Abstract Helper" was designed to encourage users to search for a good model of their target abstract, the construction of a corpus of well-organized sample English abstracts was the key to making our writing tool useful. The three subsequent subsections describe how we collected sample English abstracts, prepared their Japanese equivalents, and tagged this E-J parallel corpus with some linguistic information.

2.1. Collecting Sample English Abstracts

With permission to use them for research purposes, we collected a total of 539 sample English abstracts (2939 sentences) from widely known technical journals or conference proceedings illustrated in Table 1. Among the abstracts collected, those from the ACL proceedings were not available in electronic form, so they were manually typed into an electronic database.

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

Table 1: Structure of our Corpus of Sample Abstracts

Sample abstracts from the "*IEEE Transactions on Pattern Analysis and Machine Intelligence*" consist of 285 paper abstracts published monthly between July 1995 and July 1997 while those from the "*IEEE Multimedia*" consist of 36 paper abstracts published quarterly between 1995 and 1996. As a third source of data, 218 sample abstracts from the ACL proceedings are paper abstracts published annually between 1990 and 1994.

2.2. Aligning English-Japanese Sentence Pairs of the Sample Abstracts

We prepared Japanese translations of sample English abstracts to make it easier for users to search for a good model of both their target abstract and their target component sentence.

Japanese equivalents were voluntarily constructed on a sentence-to-sentence basis by Ricoh's software engineers, thereby aligning English-Japanese sentence pairs of sample abstracts manually. Since the engineers are well informed about the topic areas, they could produce high-quality Japanese equivalents.

2.3. Corpus Tagging

To help users improve the rhetorical structure of their writing, we examined how to tag our corpus by analyzing the sample abstracts we collected in terms of their textual structure and logical sequence.

As described by Narita (1999a, 1999b), we designed our corpus to be marked up and tagged 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)

Fig. 1 shows a fragment of our manually tagged E-J parallel corpus of sample abstracts where the linguistic information used for tagging is bold-faced. A detailed explanation of each set of linguistic information is given in the subsequent subsections.

2.3.1. Tagging Sample Abstracts with Abstract Types and Organizational-Scheme Types

Our analysis of the textual structure of sample English abstracts led us to use two kinds of tagsets for tagging each sample abstract. One of the tagsets consists of 5 abstract types and the other consists of 4 organizationalscheme types.

Abstract types represent what the authors intend to convey in their papers and are classified into 5 categories:

- 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

With the "Abstract Helper," users are encouraged to select the category that is most closely related to their own work so that the search for finding a good model can be narrowed down.

Organizational-scheme types represent the location of the topic sentence in an abstract. They are classified into 4 categories:

- 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

We focused on the position of the topic sentence which summarizes the central idea of the abstract, as studied by Shinoda (1981). Most of our sample abstracts were oneparagraph abstracts and typed as one between S001 and S003. Those abstracts consisting of two or more paragraphs were uniformly tagged as S004, regardless of the position of the topic sentence.

The tagging of this information enables users to easily search for a good model to use by specifying the position of the topic sentence. Thus, users are expected to learn how to present their ideas in a well-organized sequence.

2.3.2. Tagging Sample Sentences with Sentence Roles

Writing an abstract is a form of paragraph writing because abstracts are mostly written in one paragraph. In paragraph writing, the main idea is described by the topic sentence and other sentences have their respective logical relationships with the topic sentence so that ideas are presented in a logical and coherent sequence.

We analyzed the logical relationships between the topic sentence and other sentences of each sample

¹ Institute of Electrical and Electronics Engineers

² Association for Computational Linguistics



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

abstract in order to design the target which we could use for tagging each sentence of our sample abstracts. Thus, in this tagging, the topic sentence plays an important role in that other sentences are assigned their sentence roles according to their logical relationships with the topic sentence. The sentence roles used for our corpus tagging are:

- R010: introductory sentence
 - e.g. motivation for new research
- R020: topic sentence
- e.g. proposal of a new algorithm
- R030: explanatory sentence
- e.g. description of the algorithm
- R031: verifying sentence
 - e.g. supporting evidence
- R032: supplementary sentence
- e.g. possible applications using the algorithm - R040: concluding sentence
- e.g. conclusion
- R041: closing sentence e.g. future work

Note that all of these seven sentence roles are not always linked together in this order within an abstract, although the topic sentence is the obligatory constituent of the abstract.

2.3.3. Tagging Sample Sentences with Verb Complementation Patterns

Narita (1997, 1998) showed that Japanese authors have problems with sentence construction in their English writing and need guidance as to 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.

We designed to tag 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. An example of this tagging is taken from Fig. 1 and is shown in bold-face below:

The limited capacity of working memory is intrinsic to human sentence processing, and therefore must be **[addressed@@NP1 address NP2@@]** by any theory of human sentence processing.

When this tagging is completed, we can extract only the information on verb complementation patterns to build a separate lexical database with frequency counts of each possible complementation pattern given to a verb entry. This lexical database is linked to our corpus of sample abstracts so that sample sentences of a specified complementation pattern can be retrieved from our corpus at the user's request.

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 ESL



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

learners. The target domain for abstract writing is information engineering.

The "Abstract Helper" has four major 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.

Three types of collocational information were collected 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. With our databases of English collocations, users can easily find out their target word which is likely to co-occur with their input word.

By having access to language resources as shown in Fig. 2, the "Abstract Helper" works as follows. As a first step, when the "Sample Abstract Search" function is requested, users are encouraged to select a journal from three sources of sample abstracts and then to select both an abstract type and an organizational-scheme type from the list on the menu. When an abstract type and an organizational-scheme type are selected, sample abstracts of the specified types are retrieved from our E-J parallel corpus and displayed on the screen one at a time. Users can find a good model of their target abstract by scanning each of sample abstracts retrieved.

Users can start writing by copying and modifying the sample abstract which they have chosen as a good model on their own. Note, however, that this sample abstract functions only as a skeleton of the target abstract, and thus users need to flesh out the skeleton to present their original ideas. When faced with a problem writing a sentence, users can call for the "Sample Sentence Search" function to search for sample sentences which play a specified sentence role. If they need syntactic or lexical information to build up the target sentence, they can call for the "Sentence Pattern Search" or the "Collocation Search" function, respectively.

4. Conclusion

We constructed an E-J parallel corpus of 539 sample abstracts tagged with textual and linguistic information. The linguistic information we employed for our corpus tagging included abstract types, organizational-scheme types, sentence roles, and verb complementation patterns.

This tagged corpus of sample abstracts is effectively used by our newly developed English abstract writing tool, the "Abstract Helper." By providing quick access to this core language resource, 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.

Since the "Abstract Helper" provides both discourselevel and sentence-level assistance in an organized way, we believe that the more Japanese authors become aware of the characteristic organization of an abstract in wellwritten samples, the better they will be able to incorporate intersentential relationships into their own writing.

5. Future Work

We developed a prototype of the "Abstract Helper," which currently runs on the Sun workstation. To make it useful to a much broader community of Japanese software engineers, we will continue to improve the system by gathering substantial feedback on its functionality from our trial user group at Ricoh. We will also work on developing the possibility of semi-automated corpus tagging, based on our experiences in manual tagging, in order to efficiently broaden the coverage of our language resources.

Acknowledgements

This work has been supported by the Ministry of Posts and Telecommunications in Japan. We would like to thank Hitoshi Isahara for his valuable contributions and support.

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. *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. *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. *Technical Report of the Institute of Electronics, Information and Communication Engineers*, pp. 15-20.
- Shinoda, Y., 1981. Technical English. Nanun-Do.
- 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. *Journal of Japanese Society for Information and Systems in Education*, pp. 43-50.