8 Collaboration with Industry, Academia, and Government

8-1 Advanced Language Information Forum (ALAGIN)

UCHIMOTO Kiyotaka, TORISAWA Kentaro, SUMITA Eiichiro, KASHIOKA Hideki, and NAKAMURA Satoshi

The goal of the Advanced Language Information Forum (ALAGIN) is to spread and popularize the results of researching technologies that will overcome the language barrier and to promote further research with the collaboration between industry, academia, and government. The merit for NICT contributing to the forum activities is that NICT can provide the research results such as speech and language resources and tools for forum members, conduct evaluation and field tests through trial services, and obtain feedback from them. The other members also have a merit of obtaining the information on the state-of-the-art technologies and finding business seeds.

Keywords
Language barrier, Spoken language processing, Natural language processing, Speech and language resources and services, Forum

1 Introduction

Advanced Language Information Forum (ALAGIN) (http://www.alagin.jp/index.html) was established on March 25, 2009 to spread and popularize the results of integrated research and development of speech language in NICT and to promote further research in collaboration with industry, academia, and government with the support of related fields' companies, influential individuals, and Ministry of Internal Affairs and Communications. There are a total of 246 members; 85 regular members consisting of mainly private companies, and 161 special members consisting of influential individuals such as college teachers as of the end of the fiscal year 2011. The advantages of NICT contributing to the forum activities are that NICT can provide the research results such as speech and language resources and tools for forum members through ALAGIN, conduct evaluation and field tests through trial services, and obtain feedback from them. The other members also have the merits of obtaining information on state-of-the-art technologies and finding business seeds. In addition, ALAGIN is suitable as a business matching opportunity, and the research results are expected to be applied to commercial use.

2 Organization and activities of ALAGIN

The Organization chart of ALAGIN is shown in Fig. 1. Activities of the Planning and Promotion Committee, the Technology Development Division, and the Promotion Division for Industrial Japanese, which are main parts of the forum activities, are described in the next subsections.

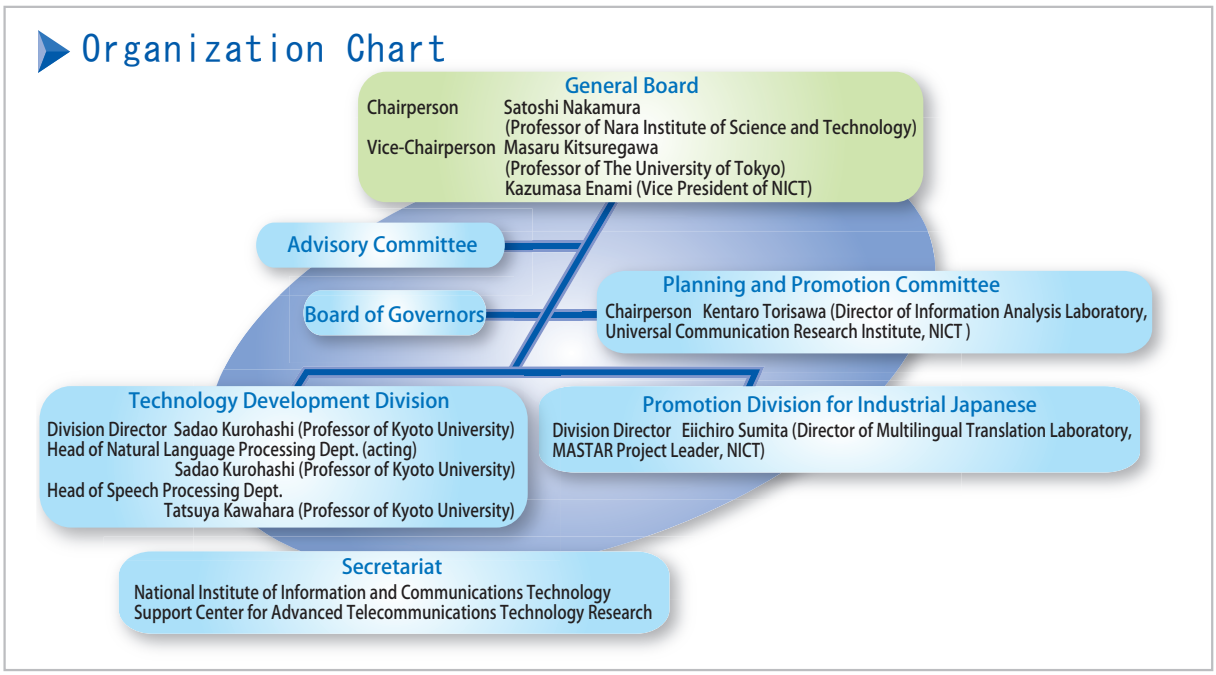

Fig.1 *Organization chart of ALAGIN*

2.1 Activities of Planning and Promotion Committee

The Planning and Promotion Committee is aimed at the further popularization and vitalization of the forum activities by promoting the activities in the following ways:

(1) Promotion of cross-functional activities

Handling intellectual property relating to revision trends of Copyright Law such as assessment, standardization, and fair-use as cross-functional theme.

(2) Provision of Speech and Language Resources, Tools, and Services

Promoting the provision of speech and language resources, tools, and services. Also promoting the distribution of research result data not only provided by NICT, but also by other members.

(3) Promotion of Public Relations

Promoting the enlightenment and popularization of the forum activities by attending outside exhibitions in addition to working group activities such as seminars.

In the above activities, the (2) Provision of Speech and Language Resources, Tools, and Services mostly contributes to the spread and popularization of the NICT research results. Provision and contract statuses of the language resources and services and speech resources to the ALAGIN members by the end of the fiscal year 2011 are as follows. Refer to Chapter **5-5** of this special issue with regards to the details of the language resources and services.

I. Language Resources and Services
(http://alaginrc.nict.go.jp/recources.html)

1. Database of Similar Context Terms (commercially available)

Listing at most 500 nouns according to their similarity in a descending order. Each noun appears in a similar context in the web documents to that of the entry word for each of about one million entry words.

2. Verb Entailment Database (commercially available)

Listing a total of 121,508 verb pairs, with entailment relations (52,689 pairs) and without entailment relations (68,819 pairs)

3. List of Burden and Trouble Expressions (commercially available)

Listing 20,115 expressions related to troubles and obstacles that may be a burden on human activities or have a negative impact, such as “disaster”, “psychological

stress", and "asbestos contamination"

4. Hypernym Hierarchy Database (commercially available)

   Hierarchical thesaurus consisting of about 69,000 nouns which are manually graded hypernyms in hypernym-hyponym pairs automatically acquired from Japanese Wikipedia articles (2007/03/28 version) by using the hyponymy extraction tool

5. Database of Word Co-occurrence (commercially available)

   Listing words with their co-occurring score representing a semantic relationship between the entry word and its co-occurring words in the descending order of the score, for each entry word

6. Database of Japanese Paraphrasing Patterns (commercially available)

   Collecting paraphrasing patterns by using dependency structure analysis results at a sentence level for each pattern that links arbitrary nominals A and B in a sentence such as "A has an abundance of B"

7. Database of Japanese Orthographic Variant Pairs (commercially available)

   Collecting positive and negative examples of Japanese orthographic variant pairs (or "pairs of orthographically inconsistent terms") whose character-level edit distance value is large

8. Japanese Dependency Structure Database (commercially available)

   Collecting dependency structures with their frequencies obtained by syntactically analyzing a huge amount of Japanese documents and extracting dependency structures from the syntactic analysis results

9. Case Base for Basic Semantic Relations (commercially available)

   Collecting 102,436 word pairs that have high context similarities in about 100 million-page web documents and are manually classified and labeled with semantic relations

10. Chinese Parser Model for Chinese Dependency Parser (CNP) (commercially available)

    Chinese parser model parameters for Chinese Dependency Parser (CNP) distributed as an open source software

11. Support Service for Customized Word Set Generation (commercially available)

    Service that allows the users to automatically generate groups of semantically similar words (word classes) by using 10 million words as candidates to be included in the word classes

12. Semantic Relation Acquisition Service (commercially available)

    Web based service that provides the users with word pairs that have a certain relations such as "cause and effect", "trouble and preventive measure(s)", "musicial and song title", "location name and local specialty"

13. Kyoto Sightseeing Blogs for Evaluative Information (commercially available)

    Consisting of "Kyoto Sightseeing Blogs" and "Evaluative Information Data" on the blogs. "Kyoto Sightseeing Blogs" is a database containing 1,041 Japanese blog articles (480 Japanese characters per article on the average) exclusively in the Kyoto tourism domain written by 47 authors. "Evaluative Information Data" contains evaluative information (popularity and opinions) manually extracted from Kyoto Sightseeing Blogs and the information on the evaluation holders, expressions used in their evaluation and targets of evaluation.

14. Models for Opinion Extraction Tool (commercially available)

    Consisting of model files for opinion analysis and an evaluative expression dictionary for "opinion extraction tool" distributed as an open source software

15. Bilingual Corpus for Training and Testing Japanese-English Machine Translation Engines (for research use)

    Corpus created based on the basic travel conversation data set which was used in the Japanese-English translation track of the International Workshop on Spoken Language Translation (IWSLT) evaluation campaign in 2005. Consisting of Japanese-English parallel sentences;

20,000 sentences for training and 1,500 sentences for testing machine translation systems

Ⅱ. Speech Resources
(http://alaginrc.nict.go.jp/resources.html)

1. Japanese Aged Persons Speech Database (for research use)
   Reading voices by Japanese native speakers aged 60 years or older
2. Non-native English Speech Database (for research use)
   English reading voices by non-native speakers
3. Chinese Speech Database (for research use)
   Chinese (Putonghua) reading voices and spontaneous speech voices by native Chinese from various locations in China
4. Kyoto Sightseeing Information Dialog Database (for research use)
   Transcript data of recorded face-to-face dialogs between a professional tour guide and a subject acting as a traveler
5. Japanese Elementary School Pupils' Speech Database (for research use)
   Reading voices of travel conversation and phonetically-balanced sentences by first to fourth grade elementary school pupils
6. T^3 Decoder (binary file) (for research use)
   Large vocabulary continuous speech recognition software using Weighted Finite-state Transducer (WFST) which processes 500,000 words accurately in real time
7. Japanese Speech Database (for research use)
   Japanese speech database whose speech contents are phonetically-balanced sentences and finite words developed by ATR, and are recorded by multiple professional narrators
8. Japanese-English and Japanese-Chinese Monologue Speech Database (for research use)
   Speech corpora recorded by Japanese-English or Japanese-Chinese bilingual voice actors or ordinary persons
9. T^3 Decoder (source file) (for research use)
   The source code of the software is also distributed in addition to the executable format module (binary format).

Ⅲ. Number of Subscriptions (gross)

○Language resources and services

	Regular member	Special member	Total
Fiscal year 2000	87	82	169
Fiscal year 2010	156	134	290
Fiscal year 2011	61	128	189
Total	304	344	648

○Speech resources

	Regular member	Special member	Total
Fiscal year 2010	57	54	111
Fiscal year 2011	18	25	43
Total	75	79	154

A separate agreement between the provider and the user is necessary for the use of individual speech and language resources, tools, and services. The above-mentioned number of subscriptions is the total number of agreements for the use of the speech and language resources, tools, and services provided by NICT as its research results. Most of the language resources and services currently distributed via the forum listed in Item I are commercially available. For example, in the agreement of the above-mentioned language resources 1 to 10, users can not only use the speech and language resources, tools, and services ("Research Results") for their use (Fig. 2), but also provide and/or sell products using the Research Results (Fig. 3), and produce a license of reuse of data to forum members when the Research Results are modified more than a certain threshold (Fig. 4), and furthermore impart the reuse licensing to parties who provide services to end users (Fig. 5). As mentioned above, terms of contracts are considered so that users can use it not only for their research but also for their commercial services. In addition, the agreement specifies obligations of reporting and indicating credit so that NICT can understand the usage status of research results. There is only the research use report at the end of fiscal 2011, but commercial use is expected to grow in the future as some parties are considering this option.

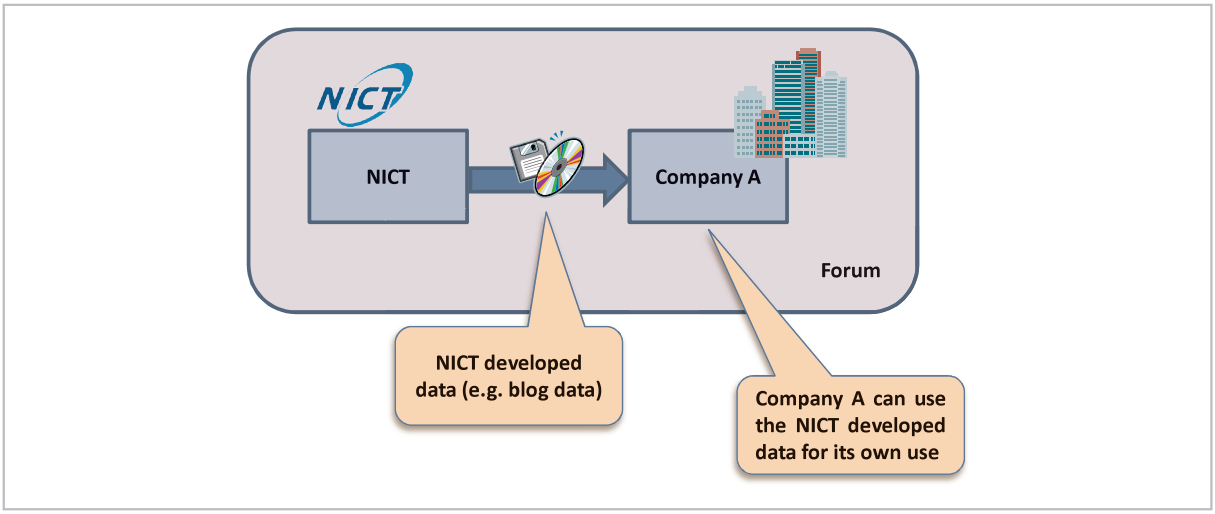

Fig.2 *Self use by users*

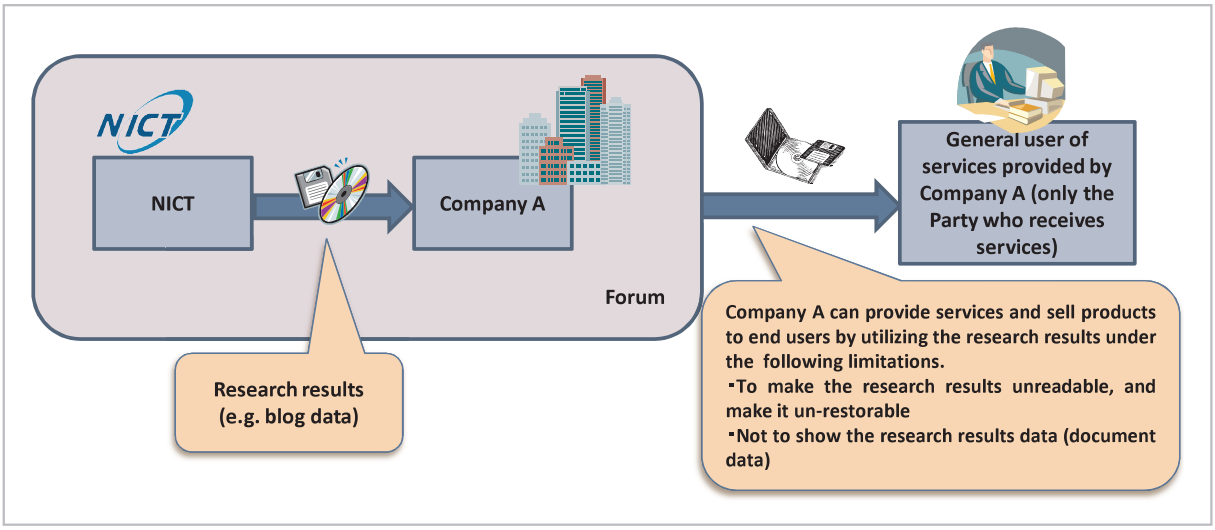

Fig.3 *User services use by users*

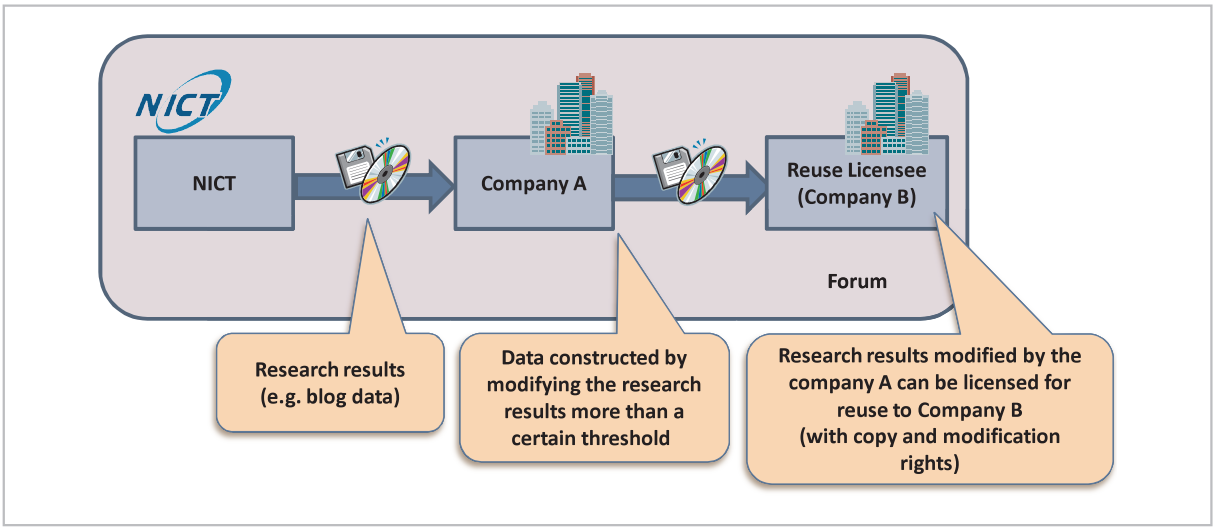

Fig.4 *Third party's use of modified data by users*

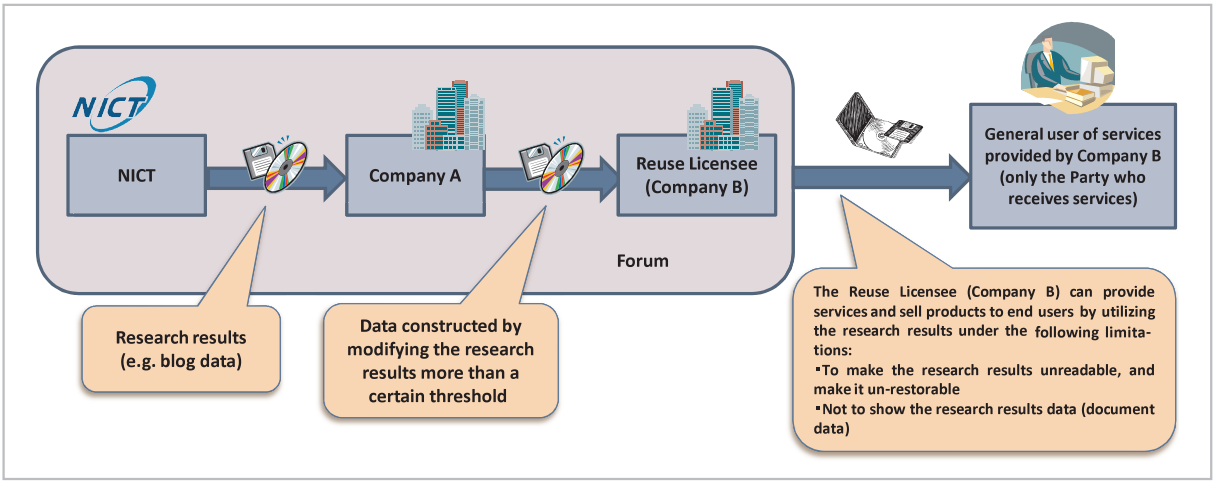

Fig.5 *Reuse license for user services*

2.2 Activities of Technology Development Division

The main activities of Technology Development Division are as follows:

- Discussing the direction and goal of the research and development in speech and language processing, web-related technology areas, and putting an emphasis on exploiting new markets. Specifically, discussing needs and system images of speech and language processing and the web-related technologies which have a deep relationship with speech and language processing, and applications of unexplored areas
- Defining targeted system requirements and clarifying necessary conditions for realizing the system; element technology, fundamental technology, speech and language resources, performance evaluation criteria, standardization items. Then, providing research guidance such as direction, goal, issues, and milestones for researchers and development engineers
- Supporting recommendations to policy planning, launch of research projects by collaboration with industry, academia, and government, budgetary requests, research accomplishments, reporting, presentations, and demonstrations etc. per request from members or the government
- Holding lecture meetings, training sessions, and seminars aiming to raise the level of speech and language processing and web-related technology areas

Training sessions and seminars aiming to raise the level of speech and language processing are well accepted in the activities. In particular, intensive four-day consecutive training sessions from basic theory to system construction of automatic speech recognition and speech dialog technology, and seminars including exercises that feature a lecture from basic to advanced levels and the state-of-the-art research trend of machine learning and natural language processing are well accepted. Through these training sessions and seminars, it is expected that the technology level of Japanese academic and industrial fields will be raised, NICT research results will be more accepted and applied, and the spread and popularization of research results will be accelerated.

2.3 Activities of Promotion Division for Industrial Japanese

The main activities of Promotion Division for Industrial Japanese are as follows:

- Performing research and popularization of Japanese that can transfer information in objective and a precise manner as the basic theme of Industrial Japanese
- Focusing attention especially on documents used in various situations of the industrial activities, mainly reviewing a way of Japanese writing for the ease of human un-

derstanding and ease of machine processing (called a way of "Industrial Japanese" writing), including penetration to other fields

- Collaborating and intercommunicating with Activities of Technology Development Division and the Association for Natural Language Processing, and supporting the recommendation to policy planning, launch of research projects by collaboration with industry, academia, and government, budgetary requests, reporting, presentations, demonstrations, and so on

Specific activities are collaborating with the Association for Natural Language Processing and Japan Patent Information Organization in developing a new interdisciplinary field by holding symposiums yearly, with a theme of Industrial Japanese to expand information-sharing cross-functionally, and promote cross-functional communication. Establishing the standardization of Industrial Japanese is a required approach to improve the automatic business translation accuracy. It is expected in the future that the translation technology of NICT will be accelerated to be spread and popularized within the industrial field.

3 Summary

This paper summarizes the organization and activities of the Advanced Language Information Forum, and describes the merit of supporting these activities by NICT. The forum plays an important role in spreading and popularizing speech and language resources and services as NICT research results, and acquiring the feedback deserving further research and development for industrial and academic fields. The total number of the language resources, services, and speech resources contracted through ALAGIN exceeds 800 at the end of fiscal year 2011, and the NICT research results are utilized in various ways in private companies and universities. It is expected that ALAGIN will become an open organization not only domestically but also internationally, and will serve as a place to recommend the policy planning to the government by studying roadmaps including speech, language, translation, and information analysis in collaboration with industry, academia, and government.

(Accepted June 14, 2012)

***UCHIMOTO Kiyotaka**, Ph.D.*

Research Manager, Planning Office, Universal Communication Research Institute

Natural Language Processing

***TORISAWA Kentaro**, Ph.D.*

Director, Information Analysis Laboratory, Universal Communication Research Institute

Computational Linguistics, Knowledge Acquisition, Web Mining

***SUMITA Eiichiro**, Dr. Eng.*

Director, Multilingual Translation Laboratory, Universal Communication Research Institute

Natural Language Processing, Machine Translation

***KASHIOKA Hideki**, Ph.D.*

Director, Spoken Language Communication Laboratory, Universal Communication Research Institute

Spoken Language Processing, Speech Translation, Spoken Dialogue

***NAKAMURA Satoshi**, Ph.D.*

Professor, Nara Institute of Science and Technology

Spoken Language Processing