# 8　産学官連携

## *8 Collaboration with Industry, Academia, and Government*

# 8-1　高度言語情報融合フォーラム（ALAGIN）

## *8-1 Advanced Language Information Forum (ALAGIN)*

内元清貴　鳥澤健太郎　隅田英一郎　柏岡秀紀　中村　哲
UCHIMOTO Kiyotaka, TORISAWA Kentaro, SUMITA Eiichiro, KASHIOKA Hideki, and NAKAMURA Satoshi

**要旨**

高度言語情報融合フォーラム（ALAGIN）では、言語の「壁」を感じさせないコミュニケーションを実現する技術の研究成果の展開・普及と研究開発の更なる推進を産学官連携によって一層進めることを目指している。NICT にとっては、会員の企業、大学などに研究成果（音声言語資源、ツールなど）を提供し、トライアルサービスなどによって評価、実証実験を行い、その評価のフィードバックを得ることができ、企業や大学にとっては、最先端技術情報の収集、ビジネスシーズ発見の場として活用できるというメリットがある。

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.

[キーワード]
言語の壁，音声言語処理，自然言語処理，音声言語資源・サービス，フォーラム
Language barrier, Spoken language processing, Natural language processing, Speech and language resources and services, Forum

## 1　まえがき

高度言語情報融合フォーラム（ALAGIN）（http://www.alagin.jp/index.html）は、NICT における音声・言語処理分野の統合的研究開発の研究成果の展開・普及と研究開発の更なる推進を産学官連携によって一層進めるため、関連分野の企業、有識者、総務省のご協力を得て、平成 21 年 3 月 25 日に設立されたものである。平成 23 年度末の時点で民間企業を中心とする正会員は 85 会員、大学の先生などの有識者を中心とする特別会員は 161 会員、合計 246 会員である。NICT にとっては、この ALAGIN を介してフォーラム会員の企業や大学などに音声言語資源、ツールなどの研究成果を提供し、トライアルサービスなどによって評価、実証実験を行い、その評価のフィードバックを得ることができるというメリットがあり、一方、企業にとっては、最先

端技術情報の収集、ビジネスシーズ発見の場として活用できるというメリットがある。ALAGINはビジネスマッチングの場としても好適であり、研究成果の商用事例が生まれる下地ができつつある。

## 2　ALAGINの組織と活動

図1にALAGINの組織図を示す。以下、この組織の中でも特にフォーラム活動の中核となっている企画推進委員会、技術開発部会、産業日本語推進部会の活動について述べる。

### 2.1　企画推進委員会の活動

企画推進委員会はフォーラム活動の一層の普及と活性化を目指して、次のような方針で活動を推進している。

(1) 部会横断的な活動の推進

アセスメント、標準化、フェアユースなど著作権法改正動向に関連する知的財産といった部会横断的なテーマに取り組む。

(2) 音声言語資源、ツールの配信、サービスの提供

音声言語資源、ツールの配信、サービスの提供を進める。NICTの研究成果だけでなく会員各位からの提供データの配信も進める。

(3) 広報活動の促進

セミナーなど部会活動以外にも、展示会への出展など広報活動を積極的に行い、フォーラム活動の啓蒙、普及を一層促進する。

これらの活動の中でNICTの研究成果の展開・普及に最も貢献しているのは(2)の音声言語資源、ツールの配信、サービスの提供である。平成23年度末までにALAGIN会員を対象に提供した言語資源・サービス、音声資源とその契約状況は以下の通りである。言語資源・サービスの詳細については本特集号 **5-5** を参照されたい。

Ⅰ．言語資源・サービス
(http://alaginrc.nict.go.jp/resources.html)

1．文脈類似語データベース（商用利用可）

約100万の見出し語それぞれに対して、Web文書上での出現文脈が最も類似している名詞最大500個を類似度とともに列挙したものである。

2．動詞含意関係データベース（商用利用可）

含意関係が成立している動詞のペア（52,689ペア）と含意関係が成立していない動詞のペア（68,819ペア）の計121,508ペアを列挙し

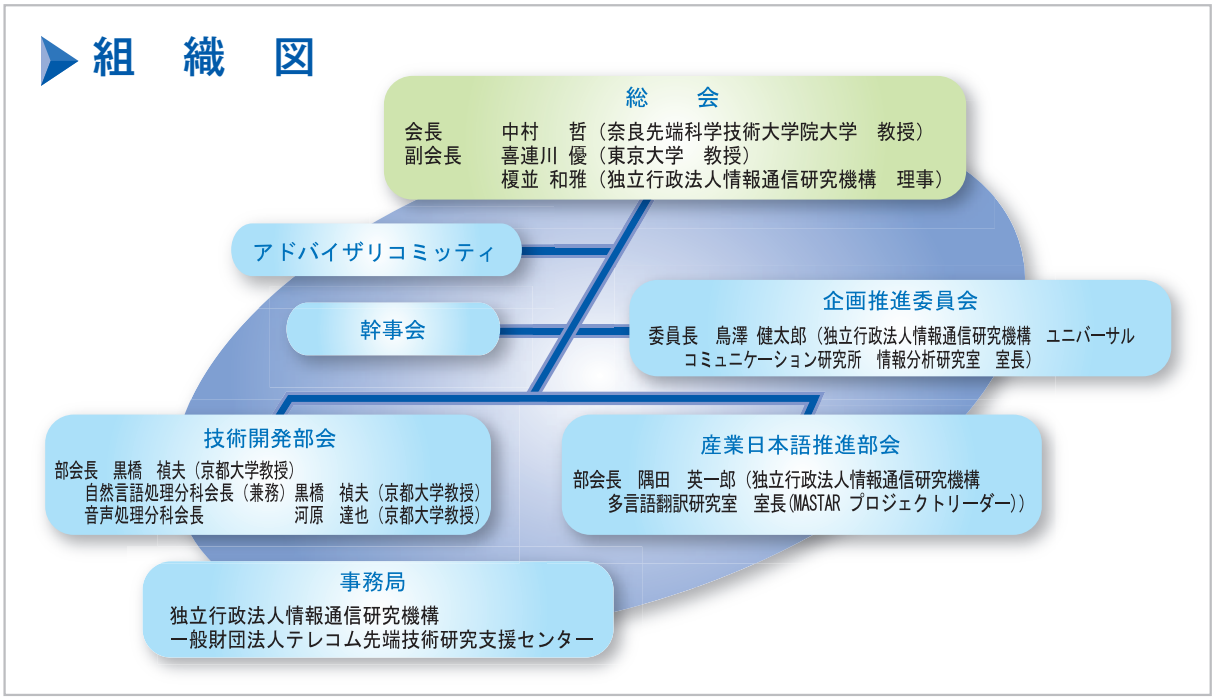


図1　ALAGINの組織図

たものである。

3．負担・トラブル表現リスト（商用利用可）
「災害」「心理的ストレス」「アスベスト汚染」など社会活動に負荷を与えたり、マイナス効果をもたらす問題や障害に関係する表現、20,115件を収録したものである。

4．上位語階層データ（商用利用可）
上位下位関係抽出ツールによって日本語Wikipedia（2007/03/28版）から自動獲得した上位下位関係の上位語を人手で階層化したものであり、合計約69,000名詞句から成る階層的シソーラスである。

5．単語共起頻度データベース（商用利用可）
各単語に対して、それとの意味的関連を表す共起スコアの高い単語を、スコアの高い順に、スコアとともに列挙したものである。

6．日本語パターン言い換えデータベース（商用利用可）
文の係り受け解析の結果を利用して、「AはBが豊富です」のような、一文中で任意の名詞AとBを結ぶパターンに対して、言い換えが可能な別のパターンを収集したものである。

7．日本語異表記対データベース（商用利用可）
文字レベルの編集距離の近い、日本語の語句の異表記対（あるいは「表記揺れの対」）の正例と負例を集めたものである。

8．日本語係り受けデータベース（商用利用可）
大量の日本語文書を係り受け解析した結果から係り受け関係を抽出し、その頻度を収録したものである。

9．基本的意味関係の事例ベース（商用利用可）
約1億ページのWeb文書上において文脈の類似度が高い2語間の意味的関係を人手で分類し、ラベル付けした結果を収録したもので、102,436語対が収録されている。

10. A Chinese Dependency Parser（CNP）用中国語解析モデル（商用利用可）
オープンソースソフトウェアとして配布している係り受け解析器（A Chinese Dependency Parser、略称CNP）のための中国語解析用モデルパラメータである。

11. カスタム単語集合作成サポートサービス（商用利用可）
1,000万語の単語を対象に意味的に類似する単語の集合（単語クラス）を自動作成するサービスである。

12. 意味的関係抽出サービス（商用利用可）
「原因–結果」関係、「トラブル–予防策」関係、「音楽家–曲名」関係、「地名–名物」関係などの何らかの意味的関係を持つ単語対を抽出するWebサービスである。

13. 京都観光ブログの評価情報付与データ（商用利用可）
「京都観光ブログ」と京都観光ブログの「評価情報付与データ」から構成される。「京都観光ブログ」は、日本語ブログ記事のデータベースである。京都観光を中心とした内容で、執筆者は47名、合計1,041記事（平均約480字）から構成される。「評価情報付与データ」は「京都観光ブログ」に対して評価情報（評判・意見）が人手で抽出され、評価保持者、評価表現、評価対象などが付与されたデータである。

14. 意見（評価表現）抽出ツール用モデル（商用利用可）
オープンソースソフトウェアとして配布されている「意見（評価表現）抽出ツール」のための意見解析用モデルファイルと評価表現辞書から構成される。

15. 日英翻訳エンジン学習・評価用対訳コーパス（研究利用）
International Workshop on Spoken Language Translation（略称IWSLT）の2005年評価キャンペーンの日英翻訳で使用された基本旅行会話データセットに基づいて作られたコーパスである。翻訳機学習用データ20,000文、評価用データ1,500文（日英対訳文）から構成されている。

Ⅱ．音声資源
（http://alaginrc.nict.go.jp/resources.html）

1．日本語高齢者音声データベース（研究利用）
日本語を母国語とする60歳以上の話者の読み上げ音声

2．ノンネイティブ英語音声データベース（研究利用）
非母語話者の英語読み上げ音声

3．中国語音声データベース（研究利用）

中国各地域出身の母国語話者による中国語（普通話）読み上げ音声および自由発話音声

4．京都観光案内対話データベース（研究利用）
プロの観光ガイドと、旅行者を模した被験者の2名による対面対話を収録し、書き起こしたデータ

5．日本語小学生音声データベース（研究利用）
小学校1年生から4年生までの話者が読み上げた旅行会話及び音素バランス文章

6．$T^3$ デコーダ（バイナリファイル形式）（研究利用）
単語数50万語彙を実時間で高精度に処理可能な「重み付き有限状態トランスデューサ（Weighted Finite-state Transducer: WFST）」を用いた大語彙連続音声認識ソフトウェア

7．日本語音声データベース（研究利用）
ATRにて開発された、音素バランス文などの文や定形単語を発話内容とする、プロナレータによる多数話者日本語音声データベース

8．日英・日中バイリンガル独話音声データベース（研究利用）
日英または日中のバイリンガルである声優または一般人が発声した音声コーパス

9．$T^3$ デコーダ（ソースファイル形式）（研究利用）
これまで実行形式モジュール（バイナリ）にて配信していたものに、ソースを追加して配信している。

Ⅲ．契約者数（延べ数）

○言語資源・サービス

| | 正会員 | 特別会員 | 合計 |
|---|---|---|---|
| 平成21年度 | 87 | 82 | 169 |
| 平成22年度 | 156 | 134 | 290 |
| 平成23年度 | 61 | 128 | 189 |
| 計 | 304 | 344 | 648 |

○音声資源

| | 正会員 | 特別会員 | 合計 |
|---|---|---|---|
| 平成22年度 | 57 | 54 | 111 |
| 平成23年度 | 18 | 25 | 43 |
| 計 | 75 | 79 | 154 |

個々の音声言語資源、ツールやサービスの利用に際しては提供者と利用者との間で個別に利用契約書を取り交わす必要がある。前述の契約者数はNICTの研究成果である音声言語資源やツール、サービスごとの契約数を合計したものである。前述Ⅰに挙げた言語資源・サービスについては現在フォーラムで配信しているもののほとんどが商用利用可能である。例えば、上記言語資源1～10の利用契約では、利用者が音声言語資源やツール、サービス（以下、「対象成果物」と呼ぶ）を自己利用できる（図2）だけでなく、利用者が対象成果物を利用してエンドユーザ向けにサービス提供、商品販売をすることを可能としたり（図3）、対象成果物に一定以上改変を加えたデータはフォーラム会員に再利用許諾可能としたり（図4）、さらに再利用許諾先がエンドユーザ向けサービスを提供することを可能とする（図5）など、利用者が研究だけでなく実サービスにも利用しやすいように契約条件を工夫している。この他、契約書では、NICTとして対象成果物の利用状況を把握できるように報告義務やクレジット表示義務を課している。平成23年度末の時点までは研究利用の報告のみであるが、商用利用について検討されているところもあり、今後、商用事例が増えることが期待される。

### 2.2 技術開発部会の活動

技術開発部会の主な活動方針は次の通りである。

- 音声・言語処理、Web関連技術の分野における研究開発の方向性や目標について議論を行い、新しい市場の開拓などを行うことに重点を置く。具体的には、音声・言語処理、音声・言語処理と関連の深いWeb関連技術のニーズやシステムイメージ、未開拓分野のアプリケーションを検討する。
- 目標とするシステムの要件や、実現に必要な要素技術、基盤技術、音声・言語資源、性能の評価基準、標準化項目などを明確にする。これにより、研究者、開発技術者が目指すべき方向や目標、課題、マイルストーンなどの研究指針を提示する。
- 会員や国の求めに応じて、政策立案への提言、産学官連携の研究プロジェクトの立ち上げや予算確保、研究遂行、報告・発表・デモ等の

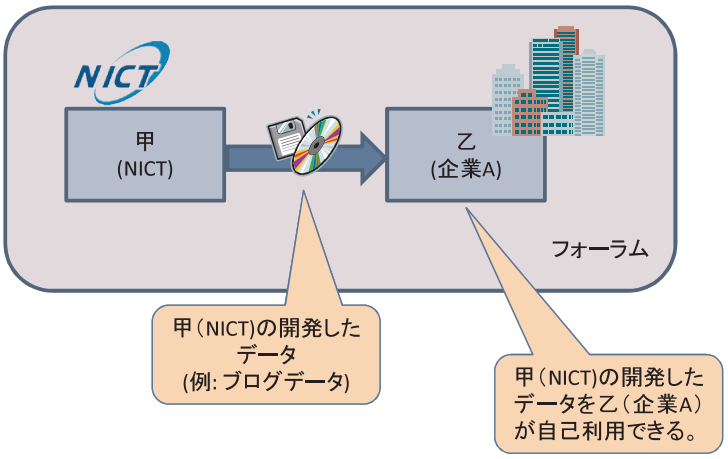


図２ 利用者（乙）の自己利用

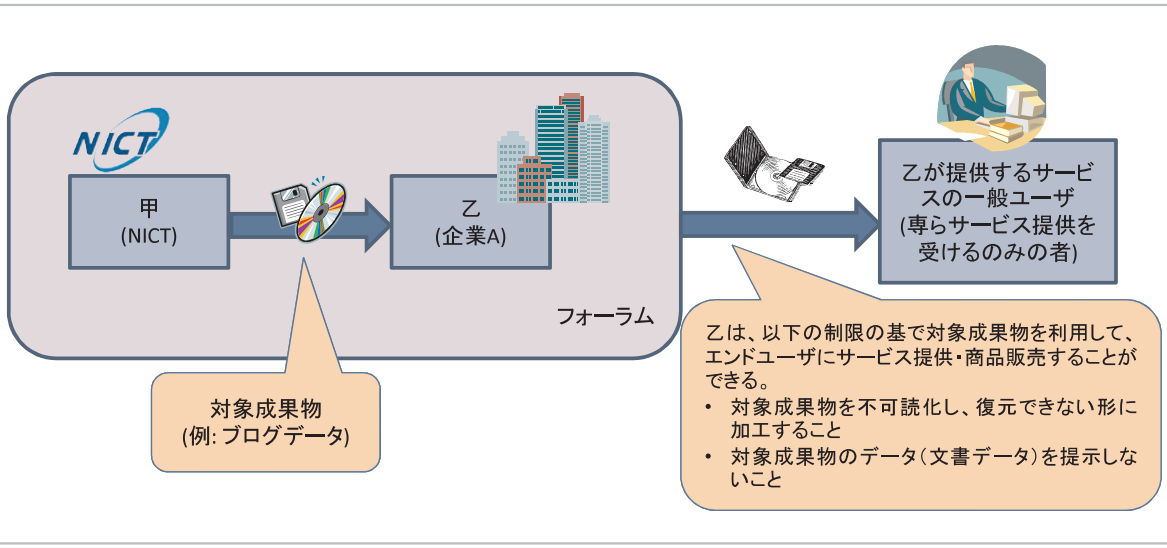


図３ 利用者（乙）のユーザサービスでの利用

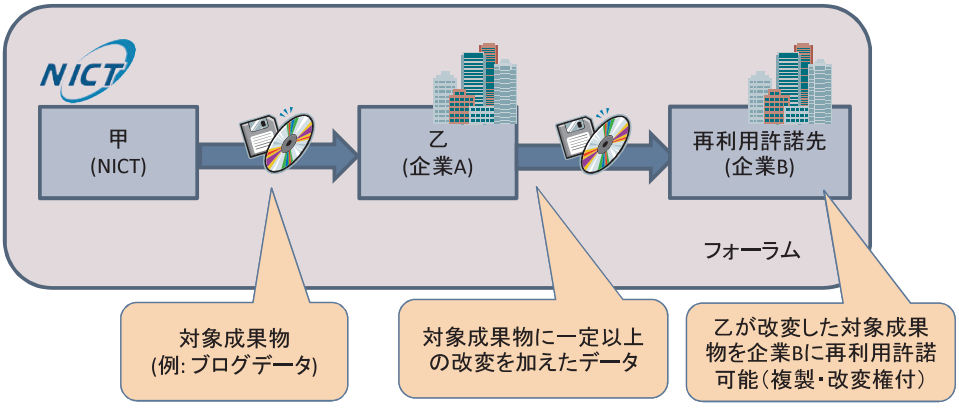


図４ 利用者（乙）の改変データの第三者の利用

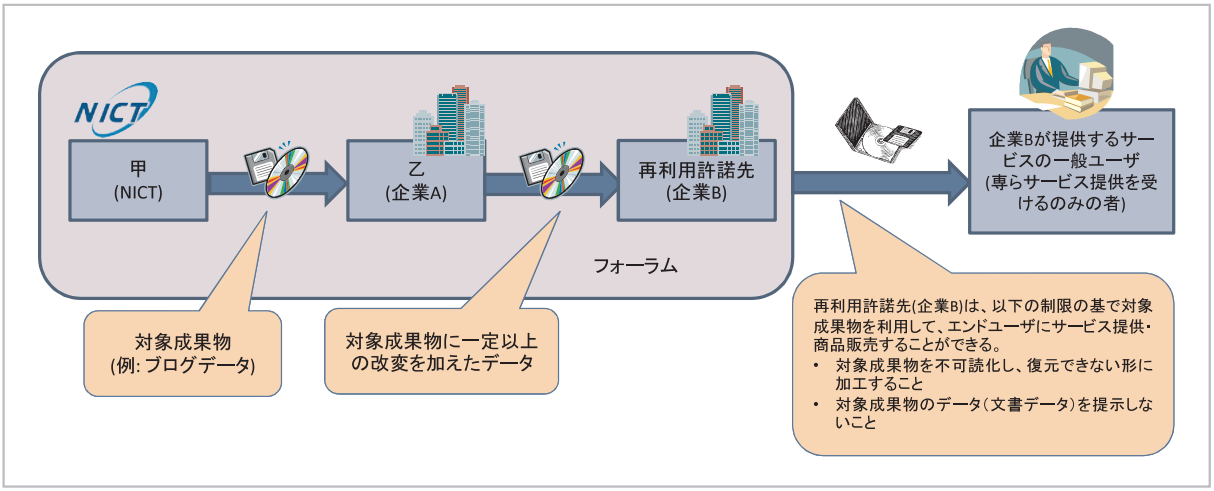


図5 再利用許諾先のユーザサービスでの利用

支援を行う。

●音声・言語処理、Web 関連技術の分野の底上げを意識した各種講演会、講習会、セミナーを開催する。

これまでの活動の中で特に好評なのは音声・言語処理技術分野の底上げを意識した講習会、セミナーである。具体的には、音声認識・音声対話技術の基礎理論の講義からシステムを作成する演習までを 4 日間連続で集中的に行う講習会や、機械学習や自然言語処理の入門から応用、最新の研究動向まで演習を含めて分かりやすく講義するセミナーなどである。これらの講習会、セミナーを通して日本の学術界、産業界の技術レベルが底上げされ、NICT の研究成果がより理解、応用されやすくなり、研究成果の展開・普及が加速されることが期待される。

### 2.3 産業日本語推進部会の活動

産業日本語推進部会の主な活動方針は次の通りである。

●産業日本語の基本テーマである情報を客観的かつ正確に表現し伝えるための日本語の研究・普及の活動を行う。

●特に、産業活動の様々な局面で作成される文書に着目し、人に理解しやすく、かつ、機械にも処理しやすく記述するための日本語（これを「産業日本語」と呼ぶ）を中心に検討し、他分野への適用も視野に入れて活動を進める。

●活動にあたっては、技術開発部会、言語処理学会等と連携・交流しながら進め、政策立案への提言、産学官連携の研究プロジェクトの立ち上げ、予算確保、報告・発表・デモ等の支援を検討する。

具体的な活動としては、毎年、言語処理学会、日本特許情報機構と共同で産業日本語をテーマにしたシンポジウムを開催し、分野横断的に情報共有を進め、異分野間の交流を深めることにより新たな学際領域を切り開こうとしている。産業日本語の標準を確立することは、産業翻訳などの自動化精度を向上させるために必要な取り組みであり、今後、それが実現できれば、NICT の翻訳技術の産業界への展開・普及も加速されることが期待される。

## 3 むすび

本稿では高度言語情報融合フォーラムの組織と活動について概観し、NICT としてこの活動を支援することのメリットについて述べた。本フォーラムは NICT の研究成果である音声言語資源、ツール、サービスを展開・普及するとともに産業界および学術界から更なる研究開発に資するフィードバックを得る上で重要な役割を果たしている。ALAGIN を介して契約された言語資源・サービス、音声資源の延べ数は平成 23 年度末までに 800 件を超え、多くの民間企業や大学などで NICT の研究成果が様々な形で利用されている。今後、ALAGIN は、国内だけでなく国際的

にも開かれた組織になるとともに、音声・言語、翻訳、情報分析を含めたロードマップを産学官で検討し政府の政策立案などへ提言する場としても機能していくことが期待されている。

（平成 24 年 6 月 14 日 採録）


**内元清貴**（うちもときよたか）
ユニバーサルコミュニケーション研究所
企画室研究マネージャー
博士（情報学）
自然言語処理
uchimoto@nict.go.jp

**鳥澤健太郎**（とりさわけんたろう）
ユニバーサルコミュニケーション研究所
情報分析研究室室長
博士（理学）
自然言語処理、知識獲得、Web マイニング
torisawa@nict.go.jp

**隅田英一郎**（すみたえいいちろう）
ユニバーサルコミュニケーション研究所
多言語翻訳研究室室長
博士（工学）
自然言語処理、機械翻訳
eiichiro.sumita@nict.go.jp

**柏岡秀紀**（かしおかひでき）
ユニバーサルコミュニケーション研究所
音声コミュニケーション研究室室長
博士（工学）
音声言語処理、音声翻訳、音声対話
hideki.kashioka@nict.go.jp

**中村　哲**（なかむら　さとし）
国立大学法人奈良先端科学技術大学院
大学情報科学研究科教授
工学博士
音声言語処理
s-nakamura@is.naist.jp