03－情－９

ヒューマンインタフェース技術に関する
調査報告書

平成15年４月

社団法人 電子情報技術産業協会

KEIRIN

この事業は、競輪の補助金を受けて実施したものです。

序文

わが国の電子工業は、急速に進展する社会・産業の情報化を推進する基幹産業であり、一層高度な研究開発と多様な技術展開の推進が極めて重要である。

このため、当協会では、日本自転車振興会から機械工業振興資金の補助を受け、「情報通信産業の技術戦略等に関する調査研究」を実施し、電子情報通信産業の発展の基礎となるネットワーク技術、高度コンピューティング技術、ヒューマンインタフェース技術、ソフトウェア技術、情報デバイス技術、電子材料・デバイス技術およびセンシング技術について内外の研究開発動向の調査を行った。このうちヒューマンインタフェース技術については、「知識情報処理技術委員会」を設けて「ヒューマンインタフェース技術に関する調査研究」を実施した。

本報告書は、ヒューマンインタフェース技術に関する調査研究に関する調査の成果をとりまとめたものである。調査に当たっては、構成委員からの意見を集約しつつ、また関連学識経験者、関連国際会議出席者を招聘し、講演を聞き、討論を重ね、かつ最新の文献を調査してまとめたものである。この報告書が各方面に広く利用され、わが国の電子工業の今後の研究開発、技術展開に寄与することを念願する次第である。

平成 15 年 3 月

社団法人　電子情報技術産業協会

会　長　　谷　口　一　郎

知識情報処理技術委員会委員名簿

（敬称略・順不同）

委員長	辻井潤一	東京大学
幹事	小原永	日本電信電話㈱
監事	鈴木克志	三菱電機㈱
委員	石崎俊	慶應義塾大学
〃	井佐原均	(独)通信総合研究所
〃	橋田浩一	(独)産業技術総合研究所
〃	徳永健伸	東京工業大学
〃	村田稔樹	沖電気工業㈱
〃	福持陽士	シャープ㈱
〃	住田一男	㈱東芝
〃	奥村明俊	日本電気㈱
〃	松田純一	㈱日立製作所
〃	松井くにお	㈱富士通研究所
〃	佐藤良治	マイクロソフトプロダクトディベロップメントリミテッド
〃	清野正樹	松下電器産業㈱
〃	亀田雅之	㈱リコー
経済産業省	矢島秀浩	経済産業省
事務局	中瀬真	(社)電子情報技術産業協会
〃	鈴木尋士	(社)電子情報技術産業協会

対話コンテンツ技術専門委員会委員名簿

（敬称略・順不同）

委員長	橋田浩一	(独)産業技術総合研究所
委員	石崎俊	慶應義塾大学
〃	加藤恒昭	東京大学
〃	斎藤博昭	慶應義塾大学
〃	大森久美子	日本電信電話㈱
〃	森本由加	㈱東芝
〃	吉田和永	日本電気㈱
〃	小泉敦子	㈱日立製作所
〃	星合忠	㈱富士通研究所
〃	野本昌子	松下電器産業㈱
〃	渡邉圭輔	三菱電機㈱
〃	野本忠司	文部科学省
オブザーバ	藤田慎一	慶應義塾大学
〃	松井洋	慶應義塾大学
〃	伊藤一成	慶應義塾大学
〃	吉田篤弘	慶應義塾大学
〃	深津加代子	慶應義塾大学
事務局	中瀬真	(社)電子情報技術産業協会
〃	鈴木尋士	(社)電子情報技術産業協会

言語資源専門委員会委員名簿

（敬称略・順不同）

委員長	井佐原　均	(独)通信総合研究所
幹事	黒橋禎夫	東京大学
委員	白井清昭	北陸先端科学技術大学院大学
〃	松尾義博	日本電信電話㈱
〃	佐々木美樹	沖電気工業㈱
〃	佐田いち子	シャープ㈱
〃	知野哲朗	㈱東芝
〃	石川　開	日本電気㈱
〃	岩山　真	㈱日立製作所
〃	山下達雄	㈱富士通研究所
〃	佐藤良治	マイクロソフトプロダクトディベロップメントリミテッド
〃	福重貴雄	松下電器産業㈱
〃	相川勇之	三菱電機㈱
〃	望主雅子	㈱リコー
〃	太田浩子	科学技術振興事業団
オブザーバ	荻野紫穂	日本アイ・ビー・エム㈱
〃	森本秀樹	富士通㈱
事務局	中瀬　真	(社)電子情報技術産業協会
〃	鈴木尋士	(社)電子情報技術産業協会

Web 情報アクセス技術専門委員会委員名簿

（敬称略・順不同）

委員長	徳永健伸	東京工業大学
幹事	新納浩幸	茨城大学
委員	長谷川隆明	日本電信電話㈱
〃	池野篤司	沖電気工業㈱
〃	奥西稔幸	シャープ㈱
〃	村上知子	㈱東芝
〃	福島俊一	日本電気㈱
〃	藤尾正和	㈱日立製作所
〃	渡部勇	㈱富士通研究所
〃	佐藤光弘	松下電器産業㈱
〃	増塩智宏	三菱電機㈱
事務局	中瀬真	(社)電子情報技術産業協会
〃	鈴木尋士	(社)電子情報技術産業協会

目　　次

要　約

計算機システムに蓄積された膨大なテキスト情報の中から、真に必要で有効な情報 を取り出すことは、それほど簡単なことではない。特定の個人にとって必要な情報は、必ずしも一般的に役にたつ情報と一致するとは限らない。情報内容に立ち入らず、Webのリンク情報だけから情報の価値を判断する枠組みには、明らかな限界がある。

個人の必要に応じて有効な情報を取り出すためには、個人の持つ背景情報、その時点での特定の興味を示す検索要求（質問）、システム中の情報の相互関係とその内容とを計算機処理の対象としなければならない。また、情報が有効に消費されるためには、同定された情報をどのように提示するかも、蓄積と検索の技術と同様に重要である。

このように、テキスト情報を中心とした情報の蓄積、検索、提示の技術は、知識処理，言語処理，マンマシンインターフェース，マルティメディア技術など、現在の最先端技術を統合する情報技術の集約点となっている。

本委員会では、このような視点から、テキスト情報の検索と提示，知識処理とテキスト情報処理，マルティメディア環境下でのテキスト処理，Webの多言語化とそれへの対処などを対象にして、技術の状況を把握しその将来を展望すること、分野の健全な発展のために必要な資源を共同で構築していくこと、また、急速に活発化しつつある国際的な標準化運動に日本として貢献することを目的に活動を行ってきた。

本委員会の実質的な活動は、井佐原均（通信総研），橋田浩一（産総研），徳永健伸（東工大）の３名の先生方に委員長をお願いしている専門委員会によって行われた。

本年度は、それぞれの委員会は次のような活動を行った。

経済・社会活動において多品種少量生産と個別的サービスの比重が増し、マーケティングを初めとして産業における対話の重要性が高まっている。こうして人間同士の対話に関する科学的・工学的研究が求められているが、それには意味的・対話的構造を明示した対話のデータが必要である。また、インターネットのブロードバンド化に伴って、MM（マルチモーダル;マルチメディア)コンテンツが大量に流通することが予測されるが、これらのコンテンツは、大量のコンテンツの中から検索され、利用者の興味等に応じて対話的・動的に提示されることになるだろう。そうした高度利用のためには、やはりコンテンツの意味的・対話的構造を明示する必要がある。

対話コンテンツ技術専門委員会は、以上のような認識に立ち、対話に関する研究用データと対話的に高度利用可能なコンテンツに共通するMM対話コンテンツ技術に関する調査・検討を行い、関連技術の普及を図るとともに、関連する国際標準の提言を目指して活動している。平成14年度には、映画等のデータに基づくMM対話コーパスへのアノテーション、MM対話コーパスの検索・閲覧用ソフトウェアツールの拡張等を行った。MM対話コンテンツ技術の動向調査においては、eラーニング等への応用を範囲に加えた。

言語資源専門委員会では、昨年度の３ワーキンググループ体制から、1)言語ポータルグループ，2)言語処理応用グループ，3)著作権調査グループ，4)コーパスグループの４グループ体制で、広く言語資源に関わる調査を行った。

言語ポータルグループは、昨年度に行った言語資源およびイニシアティブに関する調査結果を基に、会議案内や新製品紹介、用語集などを含む、我が国初の網羅的な自然言語のポータルサイトを立ち上げた。

言語処理応用グループは、自然言語処理技術を利用したシステムに対するユーザニーズの把握のため、CEATEC2002の来場者に対するアンケート調査を行い、その結果をいくつかの観点から分析した。

著作権グループは、言語資源の利用に関する著作権上の問題点について、ユーザ側の有識者を講師としたヒアリング調査を行った。予め作成した標準質問に回答してもらうことにより、講師間の判断の異なりを明確化した。

コーパスグループは、昨年度作成した日韓（英）コーパスへの対応付け作業の結果を検討した。また、日本語から英語へ、英語から日本語へという翻訳方向の違いが、対応付け作業にどのように影響するかについて、実作業により比較検討した。

Web情報アクセス技術専門委員会では、昨年度に引き続きWeb関連の技術調査を行った。Webはもはや専門家だけのものではなく、一般ユーザにとっても日常生活における様々な問題解決のための有力なツールとなっている。また、Web上に発信されている情報は、企業・組織にとっても迅速な意志決定やマーケッティングのために活用されている。

現時点では、ほとんどのユーザがGoo，Yahoo，Googleといった汎用検索エンジンを用いて必要な情報を得ているのが実状であるが、Web上の情報はますます増加の一途を辿っており、もはや汎用検索エンジンだけでは必要な情報に十分にアクセスできない事態が生じている。実際、これを補うために、最近では特定の問題を解決するための専門分野Web検索サイトも登場してきている。たとえば、旅行計画の立案に関連する情報を集めて整理したサイトや商品の価格比較や評判検索によって商品購入を支援するようなサイトがこれにあたる。

このような背景を踏まえ、本年度は汎用検索エンジンから特定の専門分野に特化した検索サイトに調査対象を移し、専門分野Web検索サイトの事例、およびそこで用いられている要素技術を中心に調査を行った。

１．はじめに

１．はじめに

計算機システムに蓄積された膨大なテキスト情報の中から、真に必要で有効な情報 を取り出すことは、それほど簡単なことではない。特定の個人にとって必要な情報は、必ずしも一般的に役にたつ情報と一致するとは限らない。情報内容に立ち入らず、Webのリンク情報だけから情報の価値を判断する枠組みには、明らかな限界がある。

個人の必要に応じて有効な情報を取り出すためには、個人の持つ背景情報、その時点での特定の興味を示す検索要求（質問）、システム中の情報の相互関係とその内容とを計算機処理の対象としなければならない。また、情報が有効に消費されるためには、同定された情報をどのように提示するかも、蓄積と検索の技術と同様に重要である。

このように、テキスト情報を中心とした情報の蓄積、検索、提示の技術は、知識処理，言語処理，マンマシンインターフェース，マルティメディア技術など、現在の最先端技術を統合する情報技術の集約点となっている。

本委員会では、このような視点から、テキスト情報の検索と提示，知識処理とテキスト情報処理，マルティメディア環境下でのテキスト処理，Webの多言語化とそれへの対処などを対象にして、技術の状況を把握しその将来を展望すること、分野の健全な発展のために必要な資源を共同で構築していくこと、また、急速に活発化しつつある国際的な標準化運動に日本として貢献することを目的に活動を行ってきた。

本委員会の実質的な活動は、井佐原均（通信総研），橋田浩一（産総研），徳永健伸（東工大）の３名の先生方に委員長をお願いしている専門委員会によって行われた。

２．対話コンテンツ技術専門委員会活動報告

2. 対話コンテンツ技術専門委員会活動報告

2.1 はじめに

対話コンテンツ技術専門委員会は、対話コンテンツの処理技術の調査・発展に資することを目的として活動している。対話コンテンツとして主に二話者間の目的志向対話と映画といった映像コンテンツを取り上げ、自然言語処理を用いた言語解析はもちろんのこと、動作や表情といった運動モダリティについても分析を行う。このようなさまざまな解析結果をアノテーションとして付与することで、映像データが言語を用いて構造化され、構造を用いた検索や要約といったより知的で包括的な対話コンテンツの利用が可能となる。

本委員会ではこれまでに二種類の対話課題を収録したマルチモーダル対話コーパスを製作し、それに音声韻律タグ、形態素タグ、構文・意味情報を明示する GDA タグ、発話の役割を示す談話タグなどさまざまなアノテーションを付与し、順次公開してきた。タグの詳細は 2.2 節で述べるが、このような言語情報に関するアノテーションだけでなく、身振り、手振り、表情に対するアノテーションについても検討し作業を進めている。各種タグデータを統合し(2.4.6 節で詳述)、モダリティ間の関係を見ることにより、人間が行っているコミュニケーションの仕組みの解明につながると考えている。4.1 節に見るように、各種アノテーションが施されたマルチメディア対話コーパスは世界的にもまだ数が少なく、貴重な研究用言語資源を提供できたと自負している。

アノテーション付与作業と並行して、平成 12 年からは、対話映像を再生しつつ、リアルタイムで各種タグ情報を表示するマルチモーダル検索/再生ツールを開発してきた。今年度はモジュール化を容易にするためにプラグイン方式に設計し直し、実装を進めた。これについては 2.3 節で詳述するが、プラグイン化がなされた対話的検索エンジンについては 2.5.2 節で述べる。なお、当ツールはアノテーション作業を効率良く進めるためのプラットホームにもなっており、ツールのさらなる改良とともに公開に向けての作業を進めている。

以上のような対話データはマルチメディアコンテンツの一部にすぎず、いわゆる映像データはほとんどこの範疇に入ると考えられ、インターネットや DVD 等を媒体としてデータ量が爆発的に増えている今日、これを十分に活用することが社会的にも重要課題となっている。本委員会では対話データ解析で培った技術を応用し、障碍者向けへの情報提示のあり方や映像要約の方策を探ってきたが、今年度からは e ラーニングへの応用について調査・検討を始めた。これら応用については、MPEG-7 などの標準化動向とともに 2.4 節で述べる。

なお、作成したアノテーションデータおよびツールとともに当委員会のこれまでの活動をまとめたものを JEITA ホームページで公開する準備を進めている。

2.2 アノテーション

本節では、映画「男はつらいよ」シリーズ、「柴又より愛を込めて」に対して行った映像転記作業及び、本委員会が平成 11 年度に公開したマルチモーダル対話コーパスへのアノテーションについて記す。

2.2.1 映像転記

本年度、視覚障害者向け映像転記を踏まえ、実際に映画「男はつらいよ」シリーズ「柴又より愛を込めて」に対して、映像転記作業を進め、映像転記マニュアルの改訂を行った。実際の作業手順について記し、改定後の映像転記マニュアルを以下に転載する。

2.2.2 映像転記作業手順

本委員会で開発されているマルチモーダル対話コーパス検索/再生ツールのプラグインの一貫として、アノテーション支援プラグインが開発された。映像転記作業は、これを利用して行う。ツールの詳細は次節を参照されたい。

実際の作業手順は以下の通りである（準備として、録音機器が必要である)。

(1) 映像を見ながら、映像転記として必要だと思われることを喋り、録音する。特に、主音声からは連想できないような映像情報に注意する。たとえば、場面、時刻などの他に役者の目の動きなどが挙げられる。

(2)(1)の書き起こしを作成する。

(3)主音声を書き起こしたテキストと、(2)で作成したものを時系列で組み合わせる。主音声と映像転記の情報が混在しないよう、留意する。

(4)再度映像を見ながら、発話の開始時と終了時、動作の開始時と終了時にタイムスタンプを付与する。

以上の作業によって映像転記テキストが作成される。視覚障害者向けの映像転記テキストを副音声として利用する場合、次のような手順での作業が加わる。

(5)音声から、無音区間を探す。台詞と台詞の合間、シーンの移り変わりなどがこれにあたる。

(6)その無音区間にあわせ、映像転記を挿入する。

(7)再度映像を見ながら、発話の開始時と終了時、動作の開始時と終了時にタイムスタンプを付与する。

映像転記テキストの作成と視覚障害者向け映像転記テキストの作成は大きく異なる。一般的な映像転記テキストを作成する場合には、実際の映像情報を正確に記述していくことが必要であるが、視覚障害者向け映像転記テキストの場合は、その映像転記テキストの時刻が主音声の時刻と重複してはならない。視覚障害者向け映像転記テキストをいわゆる副音声として使用する場合、主音声と副音声がかぶってはならないためである。

2001 年度ヒューマンインターフェース技術に関する調査報告書において、映像転記マニュアル第 0.1 版を公開した。本年度の映像転記作業に伴い、必要なタグなどを設定した映像転記マニュアル第

1.0 版を作成したので、以下に転載する。

2.2.3 映像転記マニュアル第 1.0 版

(a)　はじめに

本マニュアルは TV ドラマや映画などのストーリー性を持つマルチメディアコンテンツの映像情報を電子媒体に転記する手法について解説する。転記とはあるメディア情報をテキスト情報に変換することである。音声情報をテキスト情報に変換することを音声転記といい、映像情報をテキスト情報に変換することを映像転記という。また、転記により作成されたテキストデータを転記テキストという。音声転記により作成されたテキストデータを音声転記テキスト、映像転記により作成されたテキストデータを映像転記テキストという。音声転記テキストは聴覚障害者のためのクローズドキャプション情報とほぼ同等の情報を含み、映像転記テキストは視覚障害者のための副音声解説情報とほぼ同等の情報を含む。従って、音声転記テキストと映像転記テキストの関係は相互補完的であり、両者を揃えることによりはじめてオリジナルコンテンツのストーリーとほぼ同等の情報を転記したことになる。本マニュアルでは映像転記について、(b)準備、(c)転記の原則、(d)転記の基本形、(e)ファイル形式、の順序で解説する。また、(f)映像転記テキスト例に実際のＴＶドラマの映像転記テキストを掲載したので、適宜参照されたい。

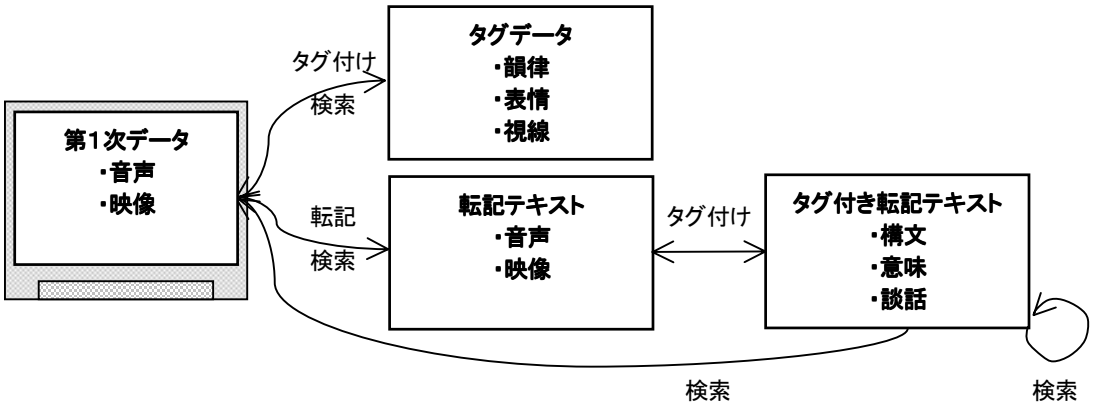


図 2.2.3-1　転記とタグ付け

(b)　準備

前述したように、音声転記テキストと映像転記テキストは相互補完的な関係にあるものの、映像転記情報の取捨選択は音声情報に依存する度合いが強い。これは(c)原則２Ｂ「主音声情報から容易に推測できる情報は記述しない」による。従って、映像転記を行う前に補助資料としての音声転記テキスト若しくは番組シナリオを用意しておくことが望ましい。次に、転記者は、これらの補助資料を参考にしつつも、必ずオリジナルコンテンツを視聴しながら転記を行わなければならない。最後に転記者の心の準備として、転記情報を取捨選択するときの基準は「仮に音声を聴くことしかできない場合に

どの情報を付与すればオリジナルコンテンツのストーリーを過不足なく理解し得るか」にあるということを、心に留め置かなければならない。

(c) 転記の原則

転記に際して転記者は以下の３つの原則を全て一貫して守り通さねばならない。

＜原則１：リアリティーのある表現を重視する。＞

A．客観的に表現する。

B．心理描写的な表現をしない。

＜原則２：コンパクトな表現を重視する。＞

A．重複する情報は記述しない。

B．主音声情報から容易に推測できる情報は記述しない。

＜原則３：映像／音声と映像転記の調和を重視する。＞

A．オリジナルコンテンツの視聴による転記を前提とする。

B．時刻情報（タイムスタンプ）を記述する。

例えば原則１について解説すると、例１～３のように、リアリティーのある表現は対話状況の細部にまでわたる想像を喚起するので、ストーリーを身近に（実際にありそうに）感じることができる。逆に、リアリティーのない表現は想像を掻き立てないので、ストーリーをうそっぽく（稚拙な作り物のように）感じさせることになる。つまり、リアリティーのある表現とはコンパクトで、かつ、対話状況の細部にまで渡る想像を掻き立てる力を持った表現のことである。

例１：　×　__a 部屋に布団を敷く、花子。

　　○　__a 畳の上に布団を敷く、花子。

例２：　×　__a さらさらと雪が降っている。

　　○　__a 粉雪が降っている。

例３：　×　__a あちこちを見回す

　　○　__a[どういう順序でどこを見たのか記述する]

また、表現にリアリティーを持たせるために番組シナリオが役立つこともある。例えば例４のように、番組からは「湖のほとり」という情報しか得られない場合に、シナリオをチェックして「諏訪湖のほとり」と記述するような場合である。

例４：　×　__s 湖のほとり。

○　__s 諏訪湖のほとり。

また、表現のリアリティーを損なうという理由で、例５～７のような心理描写も禁止である。

例５：　×　怒った顔で
例６：　×　うなだれて
例７：　×　おずおずと

(d)　転記の基本形

映像転記で記述すべき情報は表 2.2.3-1 のように 1.場面、2.動作、3.指示対象、4.メタテキスト、5.オブジェクトテキストに分類できる。映像転記テキストはこれらの転記項目を組合せて構成するのが基本である。ストーリーを理解する目的ではこの 5 項目で充分である。また、転記項目の記述には表2.2.3-2の転記記号を使用する。本節ではこれら5つの転記項目の転記手法について順に説明する。

表 2.2.3-1　転記項目

1.場面転記	場面転換が生じた場合に必要に応じて場面を記述する
2.動作転記	ストーリー理解でポイントとなる動作を記述する
3.指示対象転記	セリフ中の指示詞の参照物を記述する
4.メタテキスト転記	キャスト等、映像とは別のレイヤーに表れるものを記述する
5.オブジェクトテキスト転記	T シャツや看板に書かれた文字等を記述する

表 2.2.3-2　転記記号と意味

転記記号	意　味
#	ヘッダ
__	空白
s	場面転記
a	動作転記
開始[:/ 終了]:	メタテキスト転記
o	オブジェクトテキスト転記
←	改行

・転記項目１「場面転記」

場面転換が生じた場合には必要に応じて場面を記述する。場面転記には基本形１を適用する。

＜基本形１：__s 時、場所１、場所２。＞

使用状況１：通常の場合の場面記述。場所２は場所１の構成要素。直前の場面解説との差分のみ記述する（原則２Ａ「重複する情報は記述しない」の適用により）。

L07 __s タイトル、「虹色定期便」－グレープの入院。
L08 __s 教室。
L09 __s 浩市の家、ぶどう園の作業場。
L12 __s タイトル、キャスト。
L13 __s 病院、病室。
L15 __s 夜、佐久間家、ダイニング。
L16 __s 朝、通学路。
L17 __s 小学校、教室。
L18 __s ぶどう園の売店。
L19 __sa 作業場で箱を組み立てる、恭平たち。[1]
L21 __s 病院、病室。
L26 __sa 通りを行く、乙彦と恭平たち。
L27 __s 病院の表。
L28 __s 廊下。
L33 __s 朝、通学路。
L34 __s 解説は茶風林でした。

・転記項目２「動作転記」

ストーリー理解でポイントとなる動作を記述する。動作転記には基本形２か基本形３を適用する。

＜基本形２：__a 何々をする、某。＞

使用状況２：通常の場合の動作記述。
使用状況３：前後にセリフが詰まっている場合の動作記述。

---

[1] 転記記号 sa は場面転記と動作転記の融合形を表す。
融合形はどうしても必要な場合以外で使用してはならない。

L10 __a イスを踏み台にし、棚から箱をおろす、すみ江。

L11 __a 手伝う、楓。

L14 __a 出て行く、静香。

L19 __sa 作業場で箱を組み立てる、恭平たち。

L20 __a ぶどうを箱の中に放り込む、恭平。

L21 __a 湯のみを洗い、部屋に運んでくる、七。

L21 __a 洗濯物を紙袋に入れる、圭子。

L22 __a ベッドの下から尿瓶をとり出す、七。

L24 __a 頭から布団をかぶる、すみ江。

L26 __sa 通りを行く、乙彦と恭平たち。

L29 __a ラッピングした鉢植えを見る、すみ江。

L30 __a ラッピングを手で払う、すみ江。

L31 __a 顔を見合わせる、乙彦とすみ江。

L32 __a 鉢植えを受けとり、乙彦にほほえむ、すみ江。

＜基本形３：__a 某が何々をする。＞

使用状況４：雰囲気や余韻が必要な場合の動作記述。例えば以下の状況では例８よりも例９がふさわしい。

例８：× __a ピアノを弾いている、花子。
(→主音声ピアノ音楽の開始)

例９：○ __a 花子がピアノを弾いている。
(→主音声ピアノ音楽の開始)

L09 __a 浩太郎と静香が箱にぶどうを詰めている。

L15 __a 恭平たちが食事をしている。

L16 __a 恭平とピーチが来る。

L18 __a 浩布と楓が客にお茶を出している。

L23 __a 他の患者が笑っている。

L28 __a すみ江が乗った車イスを楓が押している。

・転記項目３「指示対象転記」

セリフで「これ」「それ」「あれ」等の指示詞が発話され、しかもそれが言語表現の外にある実体を指している場合には、必要に応じて指示対象を記述する。「これとそれとあれ」等のように文中に指示詞が連続して出現する場合には特徴的な指示対象を１つ記述する。指示対象転記には動作転記の基本形２若しくは基本形３を適用する。

＜基本形２：__a 何々をする、某。＞

使用状況５：通常の場合の指示対象記述。

例１０：（花子「これを食べると賢くなるそうよ。」）
__a 太郎にりんごを差し出す、花子。

例１１：（花子「御姉様、私あの方が愛しゅうございますの。」）
__a 振り向き、太郎を見る、花子。

・転記項目４「メタテキスト転記」

脚本、音楽、出演（役、演者）等を記述する。これらは映像とは別のレイヤーに表れるものである。メタテキスト転記には基本形４を適用する。

＜基本形４：__c 担当、担当者１、…、担当者N。＞

使用状況６：通常の場合のキャスト記述。

L12 __c 脚本、相原かさね。
L12 __c 音楽、白石哲也。
L12 __c 出演。
L12 __c 佐久間恭平、佐保祐樹。
L12 __c 宮本桃、山岸絵里奈。
L12 __c 市川武、水野樹希。
L12 __c 亀田七、蓮沼藍。
L12 __c 秋山浩布、西洋亮。
L12 __c 菊島圭子、鈴木麻世。
L12 __c 佐久間家の人々、伊藤隆、三田寛之、藍田みちる、安達俊行。
L12 __c 秋山家の人々、鈴木幸枝、松原誠、大和なでしこ、谷満里奈。

L12 __c 亀田一美、奥井奈緒子。

L12 __c、東京放送児童劇団。[2]

L12 __c、山梨県甲府立富士川小学校のみなさん。

L12 __c、甲府市民のみなさん。

・転記項目５「オブジェクトテキスト転記」

Ｔシャツにかかれた文字や、看板などから読み取れる文字など、映像に現れるオブジェクトテキストを記述する。オブジェクトテキスト転記には基本形５を適用する。

＜基本形５：__o オブジェクトテキスト転記＞

使用状況７：登場人物の来ている宇宙服の図柄の説明。

L29 __o 宇宙服には、アメリカ国旗、「ＮＡＳＡ」のロゴ、「寅」のロゴ

(e)　ファイル形式

文字コードは**ＥＵＣコード**を使用すること。時刻情報と転記記号の記述は全て**半角英数小文字**で行うこと。また、時刻情報と転記記号以外の記述は全て**全角文字**で行うこと。ヘッダ部では以下の例のように、ファイル種目、番組名、タイトル、放送日（封切り日）、転記者などを記述する。

L01 #種目　　映像転記←

L02 #番組名　虹色定期便←

L03 #タイトル　　第 13 便「グレープの入院」←

L04 #放送日　平成 12 年 11 月 15 日←

L05 #転記者　Mr. x←

L06

ボディー部では各行毎に以下の記述形式を守ること。**空白は列区切り**の意味を持つので転記記号の直前以外では使用してはならない[3]。

＜記述形式：時刻情報__基本形１__…__基本形N←＞

L07 0:00:02__s タイトル、「虹色定期便」－グレープの入院。←

---

[2] 読点「、」で始まるキャスト転記は「担当」の記述がないことを表す。

[3] 作成した転記テキストが後に行単位で計算機処理されることを想定している。

L08 0:00:06__s 教室。←

L09 0:00:28__s 浩市の家、ぶどう園の作業場。__a 浩太郎と静香が箱にぶどうを詰めている。←

L10 0:00:37__a イスを踏み台にし、棚から箱をおろす、すみ江。←

L11 0:00:45__a 手伝う、楓。←

時刻情報は、映像検索を目的として転記を行う場合には各行の最初の基本形の開始時刻のみ記述すればよい。もし転記の目的に応じて必要があれば、各行の最後の基本形の終了時刻を記述してもよい。その場合には例１２のように開始時刻と終了時刻を”-“（ハイフン）で接続すること。

例１２：0:15:28-0:15:33__a りんごをかじる、花子。

また、音声による映像解説などを目的として転記を行う場合には、時刻情報の記述にもう少し慎重にならねばならない。場面転換、主音声音楽、主音声音響効果、主音声解説などとの関係を考慮して、主音声に解説をぶつける（主音声の終了直後に間を置かずに解説を挿入する）のか、かぶせる（主音声に解説を重ねる）のか等を決定し、適当な時刻情報を記述する必要がある。解説目的の転記では、原則３「映像／音声と映像転記の調和を重視する」の適用にさらに注意深くあらねばならない。

(f)　映像転記テキスト例

L01[4] #種目　　映像転記←

L02 #番組名　　虹色定期便←

L03 #タイトル　　第 13 便「グレープの入院」←

L04 #放送日　　平成 12 年 11 月 15 日←

L05 #転記者　　Mr.X←

L06

L07 0:00:02__s タイトル、「虹色定期便」－グレープの入院。←

L08 0:00:06__s 教室。←

L09 0:00:28__s 浩市の家、ぶどう園の作業場。__a 浩太郎と静香が箱にぶどうを詰めている。←

L10 0:00:37__a イスを踏み台にし、棚から箱をおろす、すみ江。←

L11 0:00:45__a 手伝う、楓。←

L12 0:01:24__s タイトル、キャスト。__c 脚本、相原かさね。__c 音楽、白石哲也。__c 出演。__c 佐久間恭平、佐保祐樹。__c 宮本桃、山岸絵里奈。__c 市川武、水野樹希。__c 亀田七、蓮沼藍。__

[4] 行番号は本マニュアルの解説のために便宜的に記述したものであり、
実際の転記ではこれを付与しないこと。

c 秋山浩布、西洋亮。__c 菊島圭子、鈴木麻世。__c 佐久間家の人々、伊藤隆、三田寛之、藍田みちる、安達俊行。__c 秋山家の人々、鈴木幸枝、松原誠、大和なでしこ、谷満里奈。__c 亀田一美、奥井奈緒子。__c、東京放送児童劇団。__c、山梨県甲府立富士川小学校のみなさん。__c、甲府市民のみなさん。←

L13 0:02:10__s 病院、病室。←

L14 0:03:48__a 出ていく、静香。←

L15 0:05:15__s 夜、佐久間家、ダイニング。__a 恭平たちが食事をしている。←

L16 0:06:45__s 朝、通学路。__a 恭平とピーチが来る。←

L17 0:08:05__s 小学校、教室。←

L18 0:08:38__s ぶどう園の売店。__a 浩布と楓が客にお茶を出している。←

L19 0:09:08__sa 作業場で箱を組み立てる、恭平たち。←

L20 0:09:55__a ぶどうを箱の中に放り込む、恭平。←

L21 0:10:12__s 病院、病室。__a 湯のみを洗い、部屋に運んでくる、七。__a 洗濯物を紙袋に入れる、圭子。←

L22 0:10:52__a ベッドの下から尿瓶をとり出す、七。←

L23 0:11:02__a 他の患者が笑っている。←

L24 0:11:08__a 頭から布団をかぶる、すみ江。←

L25 0:11:15__s 小学校、教室。←

L26 0:11:54__sa 通りを行く、乙彦と恭平たち。←

L27 0:12:06__s 病院の表。←

L28 0:12:13__s 廊下。__a すみ江が乗った車いすを楓が押している。←

L29 0:13:08__a ラッピングした鉢植えを見る、すみ江。←

L30 0:13:36__a ラッピングを手で払う、すみ江。←

L31 0:13:45__a 顔を見合わせる、乙彦とすみ江。←

L32 0:13:55__a 鉢植えを受け取り、乙彦にほほえむ、すみ江。←

L33 0:14:10__s 朝、通学路。←

L34 0:14:53__s 解説は茶風林でした。←

2.2.4 マルチモーダル対話コーパス

本委員会は平成 11 年 6 月にマルチモーダル対話コーパスを公開した。映像、音声及びさまざまなアノテーションを施した書き起こしテキストからなり、再生／検索を行うツールも開発している。これまでに、音声タグである J_ToBI、形態素タグ、GDA タグ、談話タグのアノテーションが行われてきたが、昨年度、非言語音なども正確に書き起こし、アノテーションの元になるテキストを改定した。それに伴って、昨年度からアノテーションを再度行っている。中でも、今年度は GDA について作業を行っている。最新のアノテーションデータ及びツールは下記 URL からダウンロード可能である。

http://it.jeita.or.jp/eltech/committee/knowledge/mmc/index.html
また、今後、これらのタグを統合的に扱えるよう、改良を加えていく予定である。

2.2.5 GDA とは

GDA とは Global Document Annotation の略称であり、電子化文書の意味的・語用論的な構造を明示する XML のタグセットを作り、普及させようというプロジェクトである。本委員会から公開しているマルチモーダル対話コーパスの書き起こしテキストに対し、GDA のアノテーションを行っている。GDA タグは基本的に意味、特に主題役割、修辞関係、照応に関する情報を記述している。

2.2.6 サンプル

ここでは、実際の GDA アノテーションを施したサンプルを転載する。元になるテキストは顔課題 FM3 である。実際のテキストは
「顔顔は四角い感じ　で　そうね　優しそうな顔をしてるかな」　というものである。
現状では、既に形態素タグとの統合が行われている。それぞれの発話単位は<q>タグで囲まれており、各形態素には音声表記(タグでは tobi)、読み表記(同 yomi)、発音表記(pron)、見出し語表記基本形(fund)が付与されており、<tst val>は発声時刻、<nil val>はポーズを表している。また、照応がある形態素には id が付与されており、その関係を明らかにしている。

```
<gda>
    <su>TASK:FM3</su>
    <np id="A">MNN(E)</np>
    <np id="B">MYS(A)</np>
    <np id="X" />
  <q who="A">
      <tst val="0.315" />
    <su>
      <adp>
        <n syn="r">
            <n id="ID000000000" arg="X" tobi="kao" yomi="カオ"
              pron="カオ" fund="顔">顔</n>
            <nil val="0.059" />
            <n eq="ID000000000" tobi="kao" yomi="カオ" pron="カ
              オ" fund="顔">顔</n>
          </n>
          <ad opr="topic.fit.aen" tobi="wa" yomi="ハ" pron="ワ" fund="
```

```
は">は</ad>
<aj tobi="shikakui" yomi="シカクイ" pron="シカクイ" fund="四角
い">四角い</aj>
<n tobi="kaNji" yomi="カンジ" pron="カンジ" fund="感じ">感じ
</n>
<breath />
<breath />
<nil val="0.957" />
<tst val="2.336" />
<tst val="2.719" />
<ad tobi="de" yomi="デ" pron="デ" fund="で">で</ad>
</adp>
<nil val="0.496" />
<tst val="2.849" />
<tst val="3.344" />
<ij tobi="so'-ne" yomi="ソウネ" pron="ソーネ" fund="そうね">そうね
</ij>
<nil val="0.569" />
<tst val="3.696" />
<tst val="4.265" />
<vp>
<np>
<aj aen="mgn" tobi="yasashi" yomi="ヤサシ" pron="ヤサ
シ" fund="優しい">優し</aj>
<n tobi="so'-" yomi="ソウ" pron="ソー" fund="そう">そう
</n>
<v tobi="na" yomi="ナ" pron="ナ" fund="だ">な</v>
<n tobi="kao" yomi="カオ" pron="カオ" fund="顔">顔</n>
</np>
<ad opr="obj" tobi="o" yomi="ヲ" pron="ヲ" fund="を">を
</ad>
<v obj="ID000000000" aen="X" tobi="shi" yomi="シ" pron="
シ" fund="する">し</v>
<ad tobi="te" yomi="テ" pron="テ" fund="て">て</ad>
<v tobi="ru'" yomi="ル" pron="ル" fund="る">る</v>
```

```
    </vp>
    <ad tobi="ka" yomi="カ" pron="カ" fund="か">か</ad>
    <v tobi="na" yomi="ナ" pron="ナ" fund="な">な</v>
  </su>
  <nil val="0.693" />
```

GDA の詳細については http://www.i-content.org/gda/tagman.html から知ることができる。

2.3 マルチモーダル対話コーパス　検索/再生ツール

動画コーパスとそれを書き起こしたGDAタグ付きコーパスを用いて、意味や構文情報に基づく検索を行い、該当する文節に対応する動画像をユーザに対し提示するツールを一昨年より公開している。今年度様々な機能拡張を行った。

2.3.1　タイムスタンプ付きGDAコーパス

GDAとは、主に主題役割、修辞関係、及び照応に関する情報を記述するためのXML形式のタグセットである[1][2]。特に動画ファイルのアノテーションデータとして利用する場合、動画データの対応する時刻範囲をテキストで埋め込む方法が主に用いられる。同様の方式でGDAファイルと動画ファイルを関連付けるために次の2通りのタグ仕様を定義している。

1. btime(begin time), etime(end time)属性。任意のタグに付加することが可能。タグで囲まれている発話文に対して、btime属性値が対象開始時刻を、etime属性値が対象終了時刻を表す。記述例を以下に示す。

   <su btime=”時刻（秒）” etime=”時刻（秒）”>text</su>

   この表記方法は、無声区間や複数話者の発話の重なりが多い場合に有効である。

2. tstタグ(タイムスタンプタグ)の定義。tstタグはそれ自体要素を持たない、空要素タグであり、以下の様に記述される。

   <tst val="時刻(秒)"/>

tstタグは各文節に付与される。tstタグがテキストエレメントの前に記述される場合はval属性値を対象開始時刻とし、後ろに記述される場合はval属性値を対象終了時刻と定義する。またtstタグがbtime,etime属性を持つことも可能である。この表記方式は任意の場所にタグを埋め込むことが出来るのが特徴である。上記の時刻情報に関するタグを付加したコーパスをタイムスタンプ付きGDAコーパスと呼ぶ。例を図2.3.1-1に示す。GDAのタグの詳しい説明はここでは省略する。

```
<su syn="fc" btime="59.572"  etime="62.386">
  <ad>で<tst val="60.453870"/></ad>
  <adp rel="obj">
    <n btime="60.815086" etime="61.034672">耳</n>
    <ad>は<tst val="61.100256"/></ad>
 </adp>
  <v>かくれ<tst val="61.907619"/></v>
  <ad>て<tst val="62.192223"/></ad>
  <v>る<tst val="62.383523"/></v>
```

図2.3.1-1 タイムスタンプ付きGDAコーパス

2.3.2　ツール概要

本項では一昨年より開発を進めているツールの機能について概説する[3]-[5]。

本プログラムの実行画面はメインウィンド及び複数の内部ウィンドで構成されている。図 2.3.2-1 に画面表示例を示す。一般的なメディアプレイヤ（図 2.3.2-1 の右上）及び、GDA ファイルを解析して表形式で表示するウィンド(図 2.3.2-1 の左上)が実装されている。これらのウィンドは連携して動作する。例えば、テーブルの各行を選択すると対応範囲の動画が再生される。プレイヤーの再生部分に該当するテキストが表中でハイライト表示される等である。

図 2.3.2-1 マルチモーダル対話コーパス検索/再生ツール標準画面

また本年度はプラグイン機能を追加した。これにより、本ツールを利用者が用途に応じて自由にカスタマイズ可能となった。次項で開発した各種プラグインについて概説する。

2.3.3 プラグイン

（1）XQL 検索プラグイン

XQL とは XML ファイルの内部検索のためのクエリー言語である[6]。XQL 検索プラグインを組み込むことにより、GDA ファイルに対して XQL による検索が実行できる。画面を図 2.3.3-1 に示す。図 2.3.3-1 において、左下のウィンドがプラグインによって実装されたインタフェースである。ウィンド上部のテーブルが検索結果表示テーブル、ウィンド下部に検索式入力フィールドが表示される。単語そのものを、入力フィールドに入力することにより、単純キーワードマッチングによる検索も行うことが可能である。

XQL の入力式の例を示す。

//q　　すべての文を抽出

//n[.=’はい’]　「はい」という名詞句を抽出

//q[@who=’A’]　発話者が A である発話文を抽出

図 2.3.3-1 に、XQL プラグインを組み込んでタイムスタンプ付き GDA ファイルを検索した画面例を示す。

図 2.3.3-1 XQL プラグインロード時の画面例

（２）　アノテーション支援プラグイン

アノテーション支援プラグインはデータの編集作業だけでなく、メディアの操作にも用いられる、ツールの基本的なインタフェースである。動画データに対して、「発話情報」、「映像情報」に関する情報を簡易な操作で付与していくことが出来、これらをタイムスタンプ付き GDA ファイルとして保存できる。プラグイン上で、アノテーションの追加／削除、属性値の編集、およびタイムスタンプの修正を行うことができ、さらにアンドゥ／リドゥをはじめとしたアノテーション作業を助ける便利な機能を提供する。また、メディアの再生位置も容易にシークすることが可能である。図 2.3.3-2 にプラグインを組み込んでアノテーション作業を行った場合の画面例を示す。

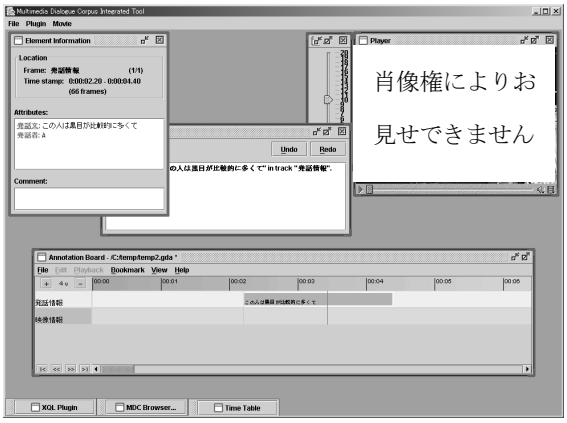


図 2.3.3-2 アノテーション支援プラグイン
ロード時の画面例

（３）　意味･構造に基づく検索プラグイン

GDA ファイルに対して、XML グラフマッチングをベースにした意味･構造に基づく検索処理を行うシステムが開発されている。このシステムは、検索処理の結果を HTML ベースで表示する。本プラグインはこのシステムを外部から利用する機能、及びシステムと連携した動画像再生等の機能を付加する。一般的なブラウザ機能を実装しており、GDA ファイルへのリンクをクリックすると、該当 GDA ファイルを本プラグインが読み込んで操作･閲覧することが可能である。図 2.3.3-3 に本プラグインを組み込んだ場合の検索画面例を示す。

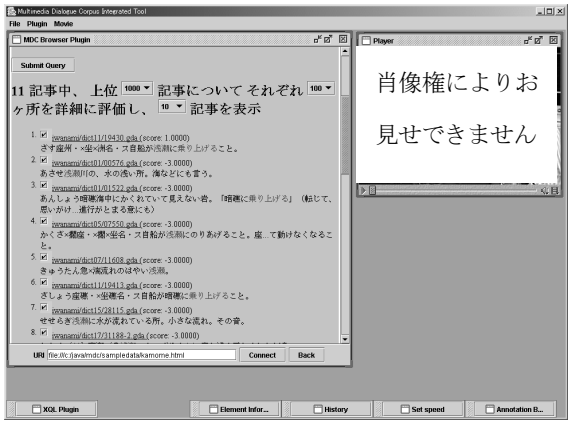


図 2.3.3-3 意味･構造に基づく
検索プラグイン

2.3.4　今後の予定

現状のツールで扱っているデータは GDA コーパスに限定されている。本委員会[7]では GDA に限らず、J-ToBI、談話情報等について記述したマルチモーダル対話コーパスを提供している[8][9]。この対話コーパスは 1999 年 6 月に、本協会より CDROM 媒体にて配布されている。本委員会では、今後これら複数形式のファイルを統合するためのタグ仕様及びその手法を選定する予定である。現在検討中の段階であるが、Annotation Graphs[10]を用いての統合を考えている。これにより韻律、身ぶり、動作、視線、

場所といった情報も条件に含めて、本ツールで検索することが可能となる。

2.3.5 まとめ

本節では、マルチモーダル対話コーパス検索/再生ツールについて述べた。今後は前項で述べた拡張機能を実装していく予定である。また本ツールはインターネット上で公開しており、ダウンロード可能である[11]。ツールに関する詳細な情報はこちらを参照されたい。

[参考文献]

[1] The GDA Tag Set ホームページ, http://www.i-content.org/gda/tagman.html

[2] 橋田 浩一,“GDA 意味的修飾に基づく多用途の知的コンテンツ”,人工知能学会論文誌,Vol.13,No4,pp.528-535, 1998.

[3] 伊藤一成, 斎藤博昭, ″マルチモーダル対話コーパス検索／再生ツールの実装″, 情報処理学会研究報告, NL142-5 (also FI61-5), 2001.

[4]01-情-11「ヒューマンインタフェース技術に関する調査報告書」、(社) 電子情報技術産業協会、(2001 年 7 月)

[5] 02-情-９「ヒューマンインタフェース技術に関する調査報告書」、(社) 電子情報技術産業協会、(2002 年３月)

[6] XQL ホームページ, http://www.w3.org/TandS/QL/QL98/pp/xql.html

[7](社) 電 子 情 報 技 術 産 業 協 会　対 話 理 解 技 術 専 門 委 員 会 ホ ー ム ペ ー ジ , http://it.jeita.or.jp/jhistory/committee/mmc/mmc.htm

[8] 00-情-７「自然言語処理システムに関する調査報告書」、(社) 日本電子工業振興協会、(2000 年３月)

[9] 金子 拓也, 石崎 俊,“マルチモーダル対話コーパスの構築---マルチモーダルデータのタギングについて---”, 電子情報通信学会　思考と言語研究会, TL99-3, pp.17-23, 1999.

[10] Steven Bird and Mark Liberman,” Annotation graphs as a framework for multidimensional linguistic data analysis”, Towards Standards and Tools for Discourse Tagging -- Proceedings of the Workshop, Somerset, NJ: Association for Computational Linguistics, 1999.

[11] マルチモーダル対話コーパス　検索/再生ツールホームページ, http://mdc.comdent.jp

2.4 調査および検討

本節では、マルチモーダル対話コーパスに関わる調査、応用、標準化動向等を述べる。

2.4.1 マルチモーダル対話コーパス

(1) 国内での取り組み

本節では、国内のマルチモーダル対話コーパスの現状についての調査結果を報告し、それらと比較する形で、「JEITA マルチモーダル対話コーパス」の特徴を述べる。国内の対話コーパスについては、2002 年 5 月の第 16 回人工知能学会全国大会で、SLUD 研究会セッション「ここまできた対話データベース--対話の研究開発への利用方法の紹介とデモ」が開催され、マルチモーダル対話コーパスを含む多くの対話コーパスが報告されたので、その内容を中心にまとめていく。そこで報告のあったものについては、コーパス名の後に発表番号(3C5-XX)を付与している。

国内で制作されたもしくは制作を予定されているマルチモーダル対話コーパスには、以下のものがある。

(a) マルチモーダル会話コーパス (3C5-05)

内容　多人数会話を含む対面状況での時事話題などの自由会話 (視線, 姿勢, 身振りを重視したもの)

注釈　モーションキャプチャによる身体の位置座標記録, 言語情報と非言語情報の統合的記述

規模　不明

制作　学術創成研究「人間同士の自然なコミュニケーションを支援する知能メディア」

　　代表　東京大学　西田豊明,　コーパスについては, 千葉大学　伝康晴

公開状況　計画段階であり, 収録がまだ実施されていない

(b) 日本語地図課題対話コーパス (千葉大学マップタスクコーパス)

内容　対面状況および非対面状況で地図を用いて道案内を行うというタスク指向会話

注釈　音声, 転記, 高次注釈

規模　128 対話　23 時間

制作　千葉大学他

参考　人工知能学会　第 27 回 SIG-SLUD　(1999,10)で以下の 3 件の関連発表がある

　　(5) 音声対話コーパス作成、分析ツールについて

　　(7) 日本語地図課題対話における相手話者発話中の発話開始現象(2)

　　(8) 日本語地図課題対話における主導権についての試論

公開状況　現在, 注釈付けを実施中で, その後, 公開される予定

(c) RWC マルチモーダルデータベース (3C5-07)

内容　否定肯定, 方向や大きさなどの身振り表現の「ものまね」動作

注釈　画像及び音声データのみ

規模　48 名×25 動作×4 回

制作　RWCP（技術研究組合新情報処理開発機構）　産総研　速水悟　他

公開状況　研究目的に限定して，実費にて公開中

参考　電子情報通信学会 研究会資料 PRMU97-95（1997-09）

人工知能学会誌，Vol.17，No.2，pp.167-170（2002-03）

http://www.rwcp.or.jp/wswg/rwcdb/mm/index.html　http://www.rwcp.or.jp/wswg/rwcdb/

(d) RWCP 会議音声データベース 2001（3C5-09）

内容　4-6 名の対面状況での会議　テーマ（ツアーの企画立案など）のみ与える自由度の高い対話

注釈　画像及び音声データ，書き起こし

規模　30 分程度の会議，7 件

制作　RWCP 知的資源 WG　産総研 田中和世　他

公開状況　研究目的に限り，実費で公開中

参考　情報処理学会研究報告　SLP-37-15，pp.85-90，2001-7

(e) RWCP 関連マルチモーダル対話データベース

内容　座位対面状況での訪問客と受付係との身振り等を含む対話

注釈　モーションキャプチャによる 3 次元位置データ，音声データ

制作　RWCP　シャープ　綿貫 啓子　他

公開状況　公開の方向で検討中であるが，具体的な日程等については未定

参考　人工知能学会　第 27 回 SIG-SLUD　(1999,10)

(9) 発話時の人間の振舞い -マルチモーダル対話データの解析-

http://www.rwcp.or.jp/outline/gyouseki/copyrightH13.html

また，身振り手振りなどの視覚情報を含まない音声対話もしくは講演会発表など自発的発話のコーパスとしては，以下が報告されている．

(f) 自然発話音声・言語データベース（ATR 音声データベース）（3C5-01）

内容　海外旅行状況で状況のホテルフロントとの通訳を介した会話の模擬発話

注釈　書き起こし，日英形態素情報，日本語構文情報，音声波形，時刻情報付き音素書き起こし

規模　618 会話，延べ 30 万形態素

制作　ATR 音声言語コミュニケーション研究所

公開状況　研究目的に限り，有償で販売中

参考　http://www.red.atr.co.jp/detabase.html

(g) 様々な応用研究に向けたタグ付き対話コーパスとタグ付け支援環境 (3C5-03)

内容　スケジュール調整，道案内等の模擬音声対話，クロスワードパズルなどのタスク音声対話

注釈　転記，韻律，形態素，スラッシュ単位，談話行為，談話セグメント

規模　28 対話，120 分，　最終目標は 40 対話，200 分

制作　談話・対話研究におけるコーパス利用研究グループ

公開状況　作成中と報告されており，公開に関する情報は見あたらない

参考　人工知能学会　研究会資料，SIG-SLUD-9903-4,pp.19-24，2000

(h) 実走行車内音声対話データベース (3C5-04)

内容　実走行車内における仮想ナビゲータシステムとの WOZ 方式による音声対話

注釈　時刻情報付き書き起こし

規模　4 対話，26 分，　収録は 140 時間，500 名，1000 対話

制作　名古屋大学統合音響情報研究拠点(CIAIR)

公開状況　公開中

備考　画像や機器操作などマルチモーダル情報も収録しているとのこと

(i) 日本音響学会研究用連続音声データベース (3C5-06)

内容　観光・旅行案内の主に対面での模擬対話とその読み上げ (収録は音声のみ)

注釈　読み上げについては書き起こし

規模　37 対話 (一対話は 50-200 文)，読み上げは 1027 文

制作　日本音響学会連続音声データベース調査委員会

　　産総研　速水悟，筑波大学　板橋秀一，早稲田大学　小林哲則，ATR　竹澤壽幸　他

公開状況　公開中 http://www.milab.is.tsukuba.ac.jp/corpus/asj_move.html

(j) 電総研道案内対話音声コーパス(1998) (3C5-08)

内容　レストラン，デパートなどへの道案内を対象とした WOZ 方式による音声対話

注釈　書き起こし，意味表現，ピッチパタン，発話単位 (ポーズ長による)

規模　40 名による 197 対話，　1300 分

制作　産業技術総合研究所

公開状況　公開中　http://akiba.media-interaction.jp/ETLSDG/

参考　日本音響学会講演論文集，1998.9，pp. 37-38,

(k) JST/CREST 電話対話データベース (3C5-11)

内容　自由な電話音声対話 (日本語を中心に，日英，日中を含む)

注釈　音声及び書き起こし，発話様式 (談話レベルを含む) に関するタグ

規模　120 対話　3600 分
制作　JST/CREST 発話様式プロジェクト
公開状況　公開を準備中

(l) 日本語話し言葉コーパス（モニター版 2002）
内容　学会講演や模擬講義などの自発音声モノローグ
注釈　代表形，発音形を区別した書き起こし，形態素情報
規模　700 時間を目標，現在の公開は 86 時間分
制作　科学技術振興調整費開放的融合研究
　　国立国語研究所　前川，通信総合研究所　井佐原，東京工業大学　古井
公開状況　研究目的に限定して，モニター版を公開中
参考　音声研究,4-2，51-61，2000 「日本語話し言葉コーパスの設計」
　　http://www.kokken.go.jp/public/monitor_kokai001.html

(m) 重点領域研究　音声対話コーパス
内容　秘書システム，スケジュール管理，クロスワードパズ，地図課題等のタスク指向会話
注釈　平仮名書き起こしなど幾つかの書き起こしテキスト
規模　93 対話　450 分
制作　京都大学，大阪大学，筑波大学，早稲田大学　他
公開状況　研究目的に限り，手数料のみで公開中
参考　http://winnie.kuis.kyoto-u.ac.jp/taiwa-corpus/
備考　音声データの再生ツール，コーパス(テキストと音声)のブラウジングツールあり

以上のマルチモーダル対話コーパス、音声対話コーパスと並べる形で、「JEITA マルチモーダル対話コーパス（以下、JEITA コーパス)」の特徴をまとめると以下のようになる。

JEITA マルチモーダル対話コーパス（3C5-10）
内容　旅行課題，顔課題のタスク指向会話，対面及び非対面状況
注釈　転記，音声転記，形態素，構文意味(GDA)，スラッシュ単位，談話タグ，時刻情報付き音素書き起こし
規模　9 対話　80 分
制作　電子情報技術産業委員会対話コンテンツ技術専門委員会
公開状況　研究目的に限り，無償で公開中
備考　検索・再生ツールあり

国内の現状では、自発的なマルチモーダル対話について、視覚情報を含めて収録したもので、かつ公開されているコーパスは極めて少ないことがわかる。多人数対話であることを特徴とする(d)のRWCP会議コーパスと、２人対話のJEITAコーパスである。課題内容は、共に広い意味では同じタスク指向対話に分類されるが、前者が、比較的自然で緩い対話条件の設定になっているのに対し、後者はやや統制の強い課題設定となっている。注釈付けについては、JEITA コーパスでは、スラッシュ単位等、談話レベルまで行われているが、(d)RWCP会議コーパスは、書き起こしにとどまっている。

音声対話のコーパスは現状でも様々なものがあるが、注釈付けという観点で見ると、統語レベル、談話レベルの注釈付けを行っているものは、かなり限られている。公開されているものでは、(f)ATR音声データベースのみである。この点でも、JEITA コーパスの注釈付けの実例は貴重なものであるといえるし、その経験に基づいて設計された検索・再生ツールも有用であろう。

一方、JEITAコーパスの一番の問題点は、規模の小ささである。2種類のタスクを設定し、対面及び非対面という２状況で、合計の対話数が９対話では、このコーパスだけを使って対話の特徴を比較することは困難であろうし、80分という時間（対面状況はその内の半分程度）では、統計的な処理にかけうるだけの特徴が得られるかも疑わしい。規模と複雑な注釈付けを両立させることが困難であるという本質的な問題であるが、小規模ながら、複雑な注釈付けを行った経験を活かして、今後の展開に繋げていくことが必要であろう。

(2) 海外での取り組み

[研究動向]

マルチモーダルコーパスは、利用目的により、モーダルの種類（音声、表情、視線、ゼスチャなど）、タグ付けの対象や仕様が多様であり、比較したり議論したりしにくい状況にあったが、少しずつ変化の兆しがみられる。言語資源に関する最も大きな国際会議である LREC2002(Third International CONFERENCE on Language Resources and Evaluation) [1] においても、前回よりもマルチモーダル関係の発表が大幅に増えただけでなく、マルチモーダルコーパスのロードマップを作ろうというワークショップの開催に象徴されるように、現状や課題についての共通の認識に基づく議論が活発化している。LREC2002では、マルチモーダルコーパスに関連して以下の４つのワークショップが開催された。

- International Workshop on Resources and Tools in Field Linguistics
- Question Answering: Strategy and Resources
- Multimodal Resources and Multimodal Systems Evaluation
- Towards a Roadmap for Multimodal Language Resources and Evaluation

マルチモーダルコーパスの開発目的には、大きく分けて、次の３つがある。

・　マルチモーダル対話システム等の応用システムに関する研究（ゼスチャ認識、表情認識など）
・　マルチモーダルコミュニケーション・言語獲得・言語教育などに関する言語研究
・　絶滅しつつある言語など言語資源の保存（E-MELD project[2)]、DOBES project[3)] など）

マルチモーダルシステムに関する研究コミュニティーでは、アノテーションスキームに関する議論が活発化している。言語研究や言語資源保存の研究コミュニティーでは、アノテーションも重要だが、それ以前にデータの収集やアーカイブの構築が重要であり、メタデータ（対象言語、話者、作成者、作成時期などコーパスそのものに関する情報）の標準化への関心が高い。具体的な開発事例が増えるにつれ、メタデータやアノテーションスキームについて、かなり具体的に比較・議論されるようになってきた。

また、LREC 本会議において、肖像権やプライバシー保護の観点から、倫理的な問題を議論するセッションが設けられたことは、マルチモーダルコーパスの開発・利用が本格化してきたあらわれと考えられる。音声コーパスの題材のひとつとしてよく利用される自宅への「帰るコール」は個人的な話題が多いため、話者のプライバシーだけでなく話題に出てくる第３者のプライバシーにも配慮する必要がある等の具体的な問題点の提起、コーパス利用にライセンスを設ける（研究者等、審査に合格した人だけがアクセスできる）といった提案等について論じられた。

マルチモーダルコーパスに関する最近の新しい流れとしては、以下のものがある。

(a) Rich Transcription

近年、対話とは少し異なるが映像コンテンツに対して言語、音声、画像に関するタグを付与する Rich Transcription という分野の研究が盛んに行われている。Rich Transcription Evaluation という NIST 主催の団体にて収録されている。2002 年は、ニュース原稿、電話での会話、会議会話の 3 つをタスクとしたコンテンツの収録が行われた。その詳細が以下の URL に公開されている。

http://www.nist.gov/speech/tests/rt/rt2002

2003 年は、収録したコンテンツに対して各種タグの付与に取り組む計画が発表されており、現在、成果はまだ公開されていない。

(b) Advancded QA

質問応答システムの 1 つの拡張として音声対話システムを捉える考え方もある。以下の URL に一覧があるが、研究途上であり、完成し公開されているものは残念ながら存在しない。

http://www.ic-arda.org/InfoExploit/aquaint/index.html

(c) PDA 上のマルチモーダル対話コーパス

AT&T では、Michael Johnston, Marilyn Walker らが PDA 上のマルチモーダル対話コーパス

(Multimodal Human Computer Interaction) の作成に取り組んでいる. ACL2002[4] に MATCH システムとして技術的取り組みについて発表がある。AT&T の研究所の成果であることから公開はされていない。

［コーパスの開発］

コーパスの開発については、ISLE(International Standards for Language Engineering)の Natural Interactivity and Multimodality (NIMM) Working Group が詳細な調査報告を公開している[5]。この調査では、公開されているもの、アノテーションがされているもの、特徴のあるものという観点から６４のコーパス（３６が顔、２８がゼスチャ）を選び、以下の点について調査している。

・データのモーダル（動的な顔／静的な顔、ゼスチャ、視線、音）
・タグ付けの対象（視線、表情、ゼスチャ、音声、テキスト）
・コーディングスキーム(FACS, MPEG-4, MPI GesturePhone, IIME(Multimodal extension of DAMSL), SmartKom Coding scheme など)
・使用しているアノテーションツール
・開発目的・利用しうる分野（開発目的は、唇や表情の認識、音声認識、表情や音声の感情分析、キャラクター合成、音声／ゼスチャ認識システムの教師データ、マルチモーダル対話システムの開発など)

この調査報告で取り上げられているゼスチャーコーパスの概要を以下に示す。

(a) ATR Multimodal human-human interaction database
内容　顔見知り同士の２人のビデオモニターを通しての対話。音声、映像（上半身・体全体)、身体１８箇所の動作データを採取。
言語　日本語
注釈　音声の書き起こし、ゼスチャ（手・腕・上腕・頭・体の位置と方向)
規模　１４対話（２８人。ひとりあたり１０分程度)
制作　ATR、シャープ、産総研
公開状況　不明
参考　Nakamura, S. et al: Multimodal Corpora for Human-Machine Interaction Research. Hayamizu, S., Hasegawa, O., Itou, K., Sakaue, K., Tanaka, K., Nagaya, S., Nakazawa, M., Endoh, T., Togawa, F., Sakamoto, K. and Yamamoto, K.: RWC Multimodal Database for Interactions for Integration of Spoken Language and Visual Information. Proceedings of ICSLP, pp. 2171-2174, 1996.

(b) CHCC OGI Multimodal Real Estate Map
内容　不動産選択を対象とする WOZ 方式によるマルチモーダル対話。ペンによる文字入力を含む。

言語　英語
注釈　音声の書き起こし、ペン入力の内容（手書き文字認識システムで認識）
規模　１８人（ひとりずつ）
制作　CHCC（Center for Human-Computer Communication, Oregon Health & Science University）Sharon Oviatt
公開状況　作者にコンタクト

(c) GRC Multimodal Dialogue during Work Meeting
内容　３人のエンジニアの打ち合わせを題材とするマルチモーダル対話。ワークミーティングにおけるマルチモーダルコミュニケーションの研究を目的に開発。
言語　フランス語
注釈　音声の書き起こし、ゼスチャ（体・頭・手・腕）
規模　不明
制作　GRC group
公開状況　作者にコンタクト

(d) LIMSI-CNRS のコーパス
French National Scientific Research Agency（CNRS）の Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur（LIMSI)が開発したコーパスには以下のものがある。

(d1) LIMSI Multimodal Dialogues between Car Driver and Copilot Corpus
内容　実運転時における運転者とナビゲーターのマルチモーダル対話。カーナビゲーションシステムの研究を目的にルノーのために開発。
注釈　音声の書き起こし、ゼスチャ（手・頭・視線)、ビジュアルな環境
規模　２７対話（５４人。１対話あたり９０分。）
制作　Xavier Briffault and Michel Denis（LIMSI-CNRS）
公開状況　LIMSI library で入手可能

(d2) LIMSI Pointing Gesture Corpus（PoG）
内容　指差しゼスチャのコーパス。地図を指差しながら制限言語を発話するが、指差しには制限なし。指差しゼスチャの認識用に開発。
言語　フランス語
注釈　音声の書き起こし、ゼスチャ（地図への指差し）
規模　１６種類のコマンドについてそれぞれ１２人
制作　Annelies Braffort and Rachid Gherbi　(LIMSI-CNRS)

公開状況　作者にコンタクト

(e) McGill University, School of Communication Sciences & Disorders, Corpus of gesture production during stuttered speech
内容　どもってしゃべったときのゼスチャのコーパス。どもりとゼスチャの関係の分析のために開発。
言語　 英語
注釈　音声の書き起こし（どもりをふくむ)、ゼスチャ、ゼスチャと音声の対応付け
規模　不明
制作　Mayberry, R.I. and Jaques, J. (McGill University)
公開状況　不明

(f) オランダの MPI ではここ数年の間に対象言語・題材・利用目的などの面で多様なマルチモーダルコーパスを作っている。
(f1) MPI Experiments with Partial and Complete Callosotomy Patients Corpus
内容　分割脳患者のマルチモーダル発話コーパス
言語 ケベックフレンチ、アメリカ英語
注釈　ゼスチャ(MediaTagger 使用)
規模　２０人（それぞれの人について２０ほどの MPEG ムービー)
制作　Hedda Lausberg and Sotaro Kita (MPI)
公開状況　作者にコンタクト

(f2) MPI Historical Description of Local Environment Corpus
内容　２～３人のマルチモーダル対話コーパス。スピーチとゼスチャの関係の言語比較研究を目的に開発。
言語　オランダ語、イタリア語、日本語、ラオ語、Arrernte（オーストラリア）
注釈　音声の書き起こし、ゼスチャ(MediaTagger 使用)
規模　１１対話
制作　Sotaro Kita, David Wilkins, Jan Peter de Ruiter, Nick Enfield, Chiara Piccini and Isabella Rega　(MPI)
公開状況　作者にコンタクト

(f3) MPI Living Space Description Corpus
内容　リビングスペースの描写を題材とするマルチモーダル対話コーパス
言語　ドイツ語
注釈　音声の書き起こし、ゼスチャ (MediaTagger 使用)

規模　１１ギガバイト
制作　Mandana Seyfeddinipur and Sotaro Kita (MPI)
公開状況 作者にコンタクト

(f4) MPI Locally-situated Narratives Corpus
内容　オーストラリアとメキシコの物語を題材とするマルチモーダル対話コーパス
言語 Guugu Yimithirr (オーストラリア), Tzeltal z (メキシコ)
注釈　音声の書き起こし、ゼスチャ(MediaTagger 使用)
規模　２対話 (850 メガバイト)
制作　Stephen Levinson　(MPI)
公開状況　作者にコンタクト

(f5) MPI Narrative Elicited by an Animated Cartoon "Canary Row" Corpus 1
内容　アニメーションから誘導したナレーションコーパス (スピーチ、ゼスチャ、手話)
言語　オランダ語、オランダ手話
注釈　ゼスチャ、手話(MediaTagger 使用)
規模 2.5 ギガバイト
制作　Sotaro Kita, Ingeborg van Gijn and Harry van der Hulst　(MPI)
公開状況 作者にコンタクト

(f6) MPI Narrative Elicited by an Animated Cartoon "Canary Row" Corpus 2
内容　アニメーションから誘導したナレーションコーパス
言語　日本語、トルコ語、アメリカ英語
注釈　なし
規模 ４ギガバイト
制作　Sotaro Kita and Asli Ozyurek (MPI)
公開状況 作者にコンタクト

(f7) MPI Narrative Elicited by an Animated Cartoon "Maus" and "Canary Row" Corpus
内容　アニメーションから誘導したナレーションコーパス
言語　オランダ語
注釈　音声の書き起こし、ゼスチャ、目の動き(MediaTagger 使用)
規模 2.5 ギガバイト
制作　Sotaro Kita and Marianne Gullberg (MPI)
公開状況 作者にコンタクト

(f8) MPI Natural Conversation Corpus
内容　３～４人での自然な会話
言語　ラオ語、日本語
注釈　ゼスチャ（ラオ語データ）(MediaTagger 使用)
規模　６４対話（ラオ語)、１対話（日本語)
制作　Nick Enfield and Sotaro Kita(MPI)
公開状況 作者にコンタクト

(f9) MPI Naturalistic Route Description Corpus 1
内容　方向指示のゼスチャを含む道順案内の対話
言語　ガーナ
注釈　音声の書き起こし(MediaTagger 使用)
規模 不明
制作　James Essegbey and Sotaro Kita　(MPI)
公開状況 作者にコンタクト

(g10) MPI Naturalistic Route Description Corpus 2
内容　ゼスチャを含む道順案内の対話。見えない場所での曲がり方の表現を含む点に特徴がある。
言語　日本語
注釈　音声の書き起こし、ゼスチャ(MediaTagger 使用)
規模　１９対話
制作　Sotaro Kita (MPI)
公開状況 作者にコンタクト

(g11) MPI Traditional Mythical Stories Corpus
内容　神話を題材とした２～３人での対話
言語　Yucatec（メキシコ)、Mopan（ブラジル)
注釈　音声の書き起こし、ゼスチャ (MediaTagger 使用)
規模　　６対話
制作　Sotaro Kita, Eve Danziger and Cristel Stolz　(MPI)
公開状況 作者にコンタクト

(g12) MPI Traditional Mythical Stories with Sand Drawings Corpus
内容　神話を題材に砂絵を使ったナレーションおよび対話（１～４人)

言語　Yucatec（メキシコ）Mopan（ブラジル）
注釈　音声の書き起こし、ゼスチャ、砂絵（MediaTagger 使用）
規模　　４対話
制作　David Wilkins（MPI）
公開状況 作者にコンタクト

(h) National Autonomous University of Mexico, DIME multimodal corpus
内容　キッチン設計を題材としたマウスジェスチャを含む WOZ 方式による音声対話
言語　スペイン語
注釈　音声の書き起こし、指示表現（作業中）
規模　　３１対話（合計７時間）
制作　The group of Multimodal Intelligent Systems of the Computer Science Department, IIMAS-UNAM
公開状況 作者にコンタクト

(i) National Center for Sign Language and Gesture Resources
内容　ASL(American Sign Language)の構文を示す例文、特定の語彙のいろいろなコンテクストでの使われ方を集めたもの（画像認識用)、ショートストーリー、２人のネイティブサイナによる２０～２５分の対話（それぞれ上半身・顔の２種類の映像)。正面２種類、側面、顔の４種類の映像（一部、正面１種類のものもある）
言語　American Sign Language
注釈　手指動作（手話単語)、非手指動作（視線、眉、頭部動作など)、手話以外のゼスチャ、英語訳
規模 200 発話（うち３２発話は SignStream によるアノテーションつき）
制作　ボストン大学 C. Neidle を中心とする ASLLRP(American Sign Language Linguistic Research Project)
公開状況　非商用目的に限定して公開（FTP, WWW, CD-ROM)。LDC を通して配布することも検討されている。
備考　The National Center for Sign Language and Gesture Resources (NCSLGR)は NSF サポートにより、ボストン大学とペンシルバニア大学が共同で進めているプロジェクト。NCSLGR の American Sign Language Linguistic Research Project (ASLLRP)では、ASL(American Sign Language)の文法研究の一環として手話コーパスおよびツール群の開発を行っている。現在、アノテーションツール SignStream とそれを使って開発した ASL コーパスを公開するとともに、SignStream を使ってタグ付けしたコーパスの登録も受け付けている。アノテーションツール SignStream は手話の言語研究を目的に開発されたものだが、アノテーションフィールドの定義により、音声言語のゼスチャ分析にも利用可能。現在提供されている手話データは言語分析用のものが中心だが、画像認識用に開発された手話デ

ータもある。
参考：ASLLRP　http://www.bu.edu/asllrp/

(j) RWC Multimodal database of gestures and speech
内容　否定表現、方向や大きさなどの身振り表現の「ものまね」動作
言語　日本語
注釈　音声、ゼスチャ（手の形・位置・方向）
規模　４８名×２５動作×４回
制作　RWCP（技術研究組合新情報処理開発機構）　産総研　速水悟　他
公開状況　研究目的に限定して、実費にて公開中

(k) University of Chicago Origami Multimodal corpus
内容　折り紙タスクを題材とする２人の対話
言語　アメリカ英語
注釈　ゼスチャに対して、発話と同期しているかどうか、コラボラティブかどうかの区別を記述
規模　９対話
制作　Furuyama, N（シカゴ大学）
公開状況　不明

(l)　VISLab Cross-Modal Analysis of Signal and Sense Data and Computational Resources for Gesture, Speech and Gaze Research
内容　リビングスペースを題材とした２人の対話。ゼスチャ、スピーチ、視線に関する言語心理学の研究を目的に開発。
言語　アメリカ英語
注釈　ゼスチャ、スピーチ、視線
規模　３対話
制作　Vislab　(Vision Interfaces and Systems Laboratory, Wright State University)
公開状況　不明

マルチモーダルコーパスについては、上記のほか、以下のWebサイトが参考になる。

Linguistic Annotation
http://www.ldc.upenn.edu/annotation/

Gesture Annotation: Tools and Data

http://www.ldc.upenn.edu/annotation/gesture/

Gesture Annotation では前述した以外に以下のマルチモーダルコーパスが紹介されている。

(m) FORM: A Kinematic Gesture-Annotation Scheme
内容　会話中のビデオに対してゼスチャの動作を記述するためのアノテーションスキーム
注釈　動作情報（上腕・前腕・手・手首・頭・胴体の位置と動き）。Annotation Graph 形式で保存しているため、他の言語情報に関するタグ付けへの拡張が可能。アノテーションツールとして現在は Anvil を使っているが、独自のツールも開発中。
規模　アノテーションスキームの評価のためにコーパスを開発しているが規模は不明
制作　ペンシルベニア大学(Craig Martell)
公開状況　サンプルデータを Web 公開

(n) CHILDES database
内容　CHILDES(Child Language Data Exchange System) project が言語獲得の研究用に開発した大規模な発話データベース。CHILDES の書き起こしフォーマットとツールを使ってたくさんのプロジェクトがデータを作成し、このデータベースに登録している。ビデオデータと書き起こしテキストが独立しているものもあるが、CLAN というアノテーションツールにより、フレーズ単位で対応付けられたものも含まれる。
言語　英語を中心に 30 ヶ国語以上
注釈　CHAT フォーマット、CA(Conversational Analysis)フォーマットによる書き起こし。CHILDES 用に開発された CLAN(Childes Language Analysis)というアノテーションツールを使用。
規模　マルチモーダルコーパスは１４種類

[コーパスの公開]
個別相談に応じて配布されているコーパスはいくつかあるが、ELDA、LDC などの言語資源共有のための機関を通して広く配布されているコーパスは非常に少ない。マルチモーダルコーパスの共有化が進んでいない理由として、目的が特化されているために共有化しにくい、データサイズが大きく公開や配布がしにくい、肖像権やプライバシーの問題、などが考えられる。アメリカでは、マルチモーダル関係の研究およびコーパスの開発はさかんに行われているが、ＬＤＣのカタログ（http://www.ldc.upenn.edu/)には､マルチモーダル関係のコーパスは今のところ記載されていない。ELDA を通して公開されているマルチモーダルコーパスは次の２つである。

(a) M2VTS (http://www.elda.fr/cata/speech/S0021.html)
電話によるサービスにおける認証を目的として収録されたフランス語を中心としたコーパスであ

り、顔の表情と音声が収録されている。37 種の異なった表情と 9 種の声を収録し、顔認証・声認証の研究データとしての使用を目的としている。顔データとしては眼鏡をかけた場合とはずした場合の認証精度の比較などを行うことで、認証精度の強化が行われている。データのみ公開されているが、顔の表情と音声が別に収録されていることから対話コーパスとしての利用には不向きである。

(b) Smarkom (http://www.elda.fr/cata/speech/S0136.html)

1999 年から 2003 年にかけて収録が行われたドイツ語コーパス。45 名の発話者によるものであり、映画館やレストランなどでの顔の表情、上半身を中心に収録されている。現在は、音声のみのデータ、映像のみのデータ、音声と映像両方が含まれたデータが公開されている。

Web で公開されているコーパスとしては、前述の ISLE 調査で取り上げられているもののほか、以下のものがある。

(a) DARPA Communicator

Lyn Walker が取りまとめをしている音声対話コーパスであり、次世代の知的な情報提供対話インタフェースと称されている。音声のみのインタラクションのみでなく身振り、グラフィック、場所等の指定は地図による表示なども含めた情報提供を目標としている。具体的タスクとしては会議のコーディネート、旅行計画、天気予報やフライト予約、ホテルや車のレンタルなどを想定している。現在、発展途上であるが、最新情報とタスクごとの各種コーパスが以下 URL にて公開されている。
http://fofoca.mitre.org/

[参考文献]

[1] LREC 2002(Third International Conference on Language Resources and Evaluation)
http://www.lrec-conf.org/lrec2002/index.html
[2] E-MELD project
http://emeld.org/
[3] DOBES project
http://www.mpi.nl/DOBES/index.html
[4] 40th Annual Meeting of the Association for Computational Linguistics
http://acl.ldc.upenn.edu/acl2002/MAIN/contents.html
[5] ISLE (International Standards for Language Engineering) Natural Interactivity and Multimodality (NIMM) Working Group report D8.1
http://isle.nis.sdu.dk/reports/wp8/index.php?chapter=1

2.4.2 障害者向け応用

本委員会で公開している「マルチモーダル対話コーパス検索/再生ツール」は、視覚障害者が映像メディア（TV、映画など）を楽しむ際に利用する「音声ガイド」の作成にも利用できると考えられる。「音声ガイド」というのは、主音声だけでは内容が理解しにくい場合に付与される、音声による補足説明である。例えば「とらや店先、つっ立っている寅さん」という様に、シーンが変わった際に多く

入る。いつ・どこで・誰が誰といる、などの補足説明を加えることで、視覚障害があっても晴眼者と同じタイミングで笑い、同じように作品を楽しむことが出来るようになる。

2002 年度､対話コンテンツ委員会では視覚障害者に映画鑑賞を楽しんでもらうための環境作りを進めているボランティア団体 CityLights 代表稲葉千穂子氏にヒアリングを行った｡主に視覚障害者が映画館で映画を楽しめるような副音声（音声ガイド）作りを活動の中心にしている団体である。現在、副音声作成作業には、テレビ・ビデオデッキ・パソコン・シナリオ（キネマ旬報などからのコピー）が必要である。リモコンで一時停止ボタンやプレイボタンを頻繁に操作し、目を凝らしてビデオデッキの時刻表示を読み取るなど、必要な時間と労力は大きい。しかしこのツールが副音声作成に使えるようになれば、電源さえ確保できれば、どこでも作業ができるようになる。時間や労力というコストの大きさから、副音声は充分作成されているとは到底いえないが、このツールによって、作成コストが小さくなれば音声ガイド製作者にとっても朗報となるはずだ。

しかし、これまでに本委員会で「映像転記」として考えてきたものと、実際の視覚障害者向けの音声ガイドとの間にはギャップがある。その違いを次の実例から知ることができる。

	A	B	C	D	E
1290	**時間**	**章**	**台詞**	**映像転記**	**音声ガイド**
1291	0:04:30		？:20・19・18・17・……		モニターに映る発射台のロケット
1292	0:04:34		レポーター:あっ、秒読みが始まりました、日本の皆さん聞えるでしょうか、遂に発射の時がやって参りました	マイクを両手で握っているキャスター	
1293	0:04:42		？:15・14・13・……		
1294	0:04:45		:ロケットの中		ロケットの中、宇宙服に身を固めた寅さん、硬く目をつぶる
1295	0:04:49		？:12・11・10・……	ロケットの中宇宙服に身を固めた寅さん、目をつ	顔がこわばってる
1296	0:04:52		寅:よ、ションベン出てるなぁ。よう、ちょっとちょっと、本当にションベンでちゃうからさ	顔こわばってる	
1297	0:04:55		？:9・8・7……		ロケットの中、暗い。赤い光
1298	0:04:57		寅:ちょっと、表出してよ、ね、すぐ……ちょっと出して、何だ出かけんのか？	ロケットの中、暗い。赤い光	
1299	0:04:59		？:6・5・4・3……	シールドに、ロケットの計器類の明かりが映っている	シールドに、ロケットの計器類の明かりが映っている
1300	0:05:02		寅:よう、ちょっとションベン、ションベン出ちゃうよ。ちょっと……、わぁ、わぁーっ		
1301	0:05:04		？:2・1・0……	ロケット発射。凄まじい土煙	ロケット発射。凄まじい土煙
1302	0:05:06		寅:何だ、おい、ほらね、ゆれないのって、本当にゆれてるじゃねえか、	ロケットの中の寅さん	
1303	0:05:09			コントロールセンター寅さんの体のモニターに豆電球が着いていて、その膀胱の部分だけ点滅	コントロールセンター。寅さんの体のモニター、膀胱の部分だけ点滅
1304	0:05:11			上昇していくロケット	上昇していくロケット
1305	0:05:13		寅:ションベン出ちゃう、ションベン出ちゃう、ションベン出ちゃう、ちびっちゃう……ちびっちゃう、ちびっちゃう、ああ	寅さんロケットの中にだんだん日が差してきた	ロケットの中、ヘルメットのシールドに映る計器類の明かり。だんだん日が差してくる

表 2.4.2-1 「男はつらいよ」シリーズ「柴又より愛を込めて」より抜粋

A列は、転記すべき動作の開始時刻、B列は章立て、C列は主音声、D列はこれまで行ってきた映像転記、E列は視覚障害者向けコンテンツとしての映像転記である。D列とE列に注目すると、転記量

の差が一目瞭然である。圧倒的にＥ列の方が少ない。

Ｄ列の映像転記は、作品の映像と映像転記が正確にシンクロしていなくてはならない。但し、タイムスタンプで同期が取れていれば、映像転記自体の文字数などを気にする必要がないので、より詳しい映像転記が可能である。

一方、Ｅ列の視覚障害者向けコンテンツとしての映像転記は、音声で流れることを想定しなくてはならない。つまりテレビの副音声と同様、主音声の他に、人物の動きの説明や、時間・場所などの情報を音声で補足説明する。ここで最も重要なことは、主音声を潰さないことである。視覚障害者は音声だけを頼りに作品を鑑賞するので、役者の台詞と映像転記を音声にしたものが重なると、混乱する可能性があり、作品の雰囲気もぶち壊しになってしまう。これまでの映像転記のように忠実に全てを書き起こすというわけにはいかない。従って、表現方法も限られ、補足説明できる情報も削らざるを得なくなる。最低限の情報だけを、台詞と台詞の合間や場面の転換の際に挿入するしかない。しかも、主音声をたてるために、補足説明と実際の映像のタイミングにずれが生じてしまうが、それによって、Ｄ列とＥ列の転記量に大きな差が出来てしまう。

しかし、Ｅ列に書いてある転記は、凡そＤ列にも含まれる。現段階ではタイムスタンプにずれが生じるため、Ｄ列・Ｅ列と２列になっているが、１ つの転記を様々に応用することが可能ではないだろうか。映像転記のうち、音声ガイドに利用できるものにアノテーションを行う。同時に、アノテーションしたものにだけ、独自のタイムスタンプを付与する。音声ガイド読み上げの際にはそのアノテートされた部分だけを音声にすればよいので、１ つのデータから複数の映像転記としての利用が可能である。

アノテートされたテキストの利用方法として、このような視覚障害者向けコンテンツの開発が考えられる。同様に、聴覚障害者向け、映像ガイドの開発にも利用できる可能性がある。

2.4.3 e ラーニングコンテンツへの適用

当委員会では、マルチモーダル対話データをさまざまな観点から構造化するための記述の枠組を策定している。「マルチモーダル対話」とは、データの内容だけでなく、データの利用の仕方がマルチモーダル的かつインタラクティブであることを意味する。このような構造化されたマルチモーダル対話データが、今後大きく利用されると考えられる分野の一つに e ラーニングがある。今後の e ラーニングコンテンツへ当委員会の提案する記述方式を適用していく可能性を探ることを含めて、e ラーニングの現状について、その利用状況や適用分野、システムやコンテンツの標準化動向について調査を行った。本項では、その調査結果について述べる。

(ア) e ラーニングとは

e ラーニングとはコンピュータやネットワークを利用した学習のことである。少し長くなるが「e ラーニング白書」[1] よりその定義を以下に引用する。

「e ラーニングとは、情報技術によるコミュニケーション・ネットワーク等を使った主体的な学習である。ここでは、コンテンツが学習目的に従い編集されており、学習者とコンテンツ提供者の間にイ

ンタラクティブ性が提供されていることが必要である。ここでいうインタラクティブ性とは、学習者が自らの意志で参加する機会が与えられ、人またはコンピュータから学習を進めていく上での適切なインストラクションが適時与えられることである」。

すなわち、e ラーニングにおいて重要なのはインタラクティブ性である。今後より高度なインタラクティブ性を実現するためには、より詳細に構造化されたコンテンツを用いることが一つの重要なポイントになると考えられる。

(イ) e ラーニング概況

コンピュータによる学習支援システムの歴史は古く、1950 年代後半より CAI(Computer Assisted Instruction)の研究が始まった。その後、人工知能研究の一分野として、知的 CAI あるいは ITS（Intelligent Tutoring System）などの研究が盛んに行われてきた。これらの多くはスタンドアロンのシステムであったが、近年のインターネットやイントラネットなどの普及に伴い、WWW 技術を用いて、サーバ上にある学習コンテンツを学習者側の端末で表示させ学習を行う WBT(Web Based Training)が次第に注目を集めてきている。

WBT は 1990 年代半ばより米国で商用化が始まった。日本でも 1998 年頃より企業内教育として利用されるようになっていきている。そのような中、e ラーニングのシステムやコンテンツの標準化、普及活動を目的に、先進学習基盤協議会 [2]（ALIC: Advanced Learning Infrastructure Consortium）や日本イーラーニングコンソーシアム [3]（eLC: e-Learning Consortium）などといった組織も設立され始め、2000 年は「e ラーニング元年」と呼ばれるほど人々の関心が高まってきている。

① 利用状況

利用が最も進んでいるのは米国であり、主要企業の 40%が何らかの形で e ラーニングを導入し、90%以上が今後導入を検討しているとの報告もある。国内においては導入の初期段階である。e ラーニングを利用する教育現場としては、幼稚園から高等学校までの初等中等教育、大学などの高等教育、その他専修学校などでの教育、企業内教育、生涯学習に分類できる。ここで、学習という用語は、研修やトレーニングを含む広い意味で使用している。通信衛星を利用した遠隔授業などの e ラーニングサービスは 1990 年前半より始まってはいるが、インターネットなどを利用したＷＢＴのようなシステムは、いずれの教育現場においても導入初期段階である。このうち、企業内教育は、一人あたりのパソコン台数（一人一台以上普及している企業が 49.5%）や、インターネットへの接続状況（95.8%）からも最も早く普及すると予測され、ALIC の 2001 年の調査 [1] によると、WBT に興味を持つ会員 33 社のうち半数はすでに WBT を導入している。2010 年の利用状況として、企業内教育・研修が全体の 40%を占めるという予測もある。

② 市場予測

「e ラーニング白書」[1] によると、WBT のマーケット規模は 2005 年に 3,222.5 億円、2007 年に 4,206.0 億円と予測されている。教育分野別で見ると、現在企業内教育ほどは WBT の導入が進んでいない高等教育や専修学校などの教育で、大きく増加すると予測されている（図 2.4.3 - 1）。

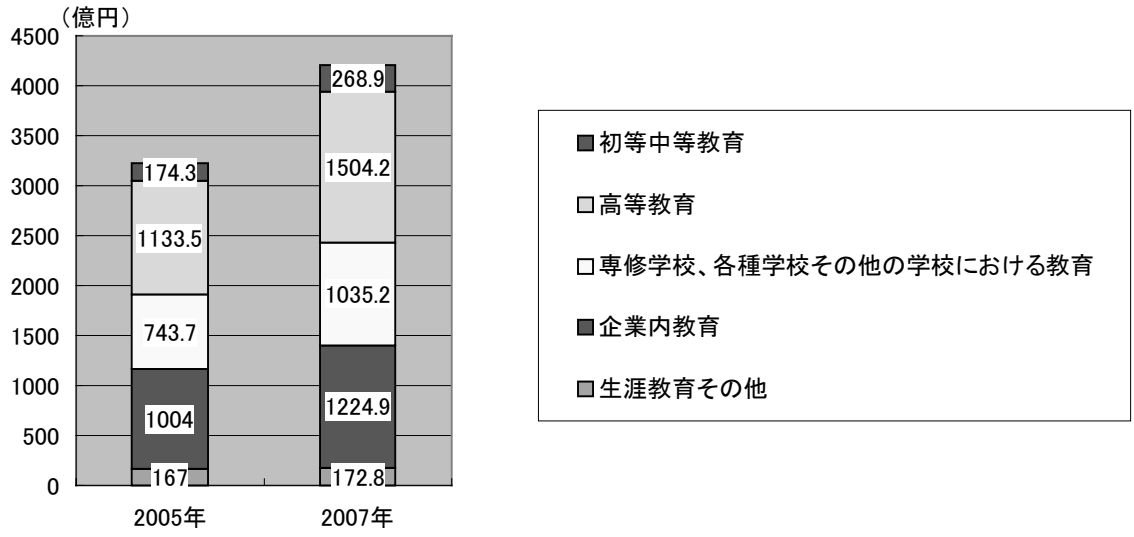


図 2.4.3 - 1　WBT のマーケット規模予測（「e ラーニング白書 2002/2003 年版」[1] より）

(ウ)　コンテンツ

①コンテンツのタイプ

e ラーニングのコンテンツは大きく以下の 2 つに分類できる。

1)　自己学習型

現在最も主流のタイプ。テキストなどの教材をそのまま電子化したいわゆる紙芝居的なものが最も多い。しかし最近では、学習者の演習問題の解答結果に応じてコンテンツを変更する問題駆動型 Web 教材や、実技を伴う学習をシミュレータ上で行うシミュレーション型教材など、より学習効果の高いタイプのコンテンツも現れている。特にシミュレーション型は、ALIC が教育サービスベンダ 14 社に行った調査[1] では、対応しているベンダ数が昨年の 55.6%から 78.6%へと大きく増加している。

2)　協調学習型

電子掲示板、テレビ会議、チャットなどを用いてグループで学習を行うもの。講師に質問したり、学習者同士での討論などが可能。今後の e ラーニングの主流になるとの予測もあり、前述の ALIC の調査では対応ベンダ数が昨年の 16.7%から 28.6%に伸びている。

②コンテンツの分野

日本イーラーニングコンソーシアム（eLC）のホームページ[3] には、会員企業が登録した e ラーニング関連製品およびサービスの検索サイトがある。登録数は表 2.4.3－1 の通りである（2003 年 1 月 1 日現在）。コンテンツの登録数が最も多く、これらはさらに大ジャンル、小ジャンルに分類されている。大ジャンルごとの登録数を図 2.4.3 - 2 に示す。PC リテラシやプログラミングなど IT スキル系のコンテンツが最も多く、全体の 43%近くを占める。ついで、ビジネススキル系のコンテンツが 21%程度となっている。このように、現状のコンテンツはその大半が企業内教育向けのものである。ALIC の調査[1] でも、教育コンテンツベンダ 22 社のうち 80%が今後も企業内教育をメインターゲットとす

ると答えている。また、分野としてはビジネススキル系にシフトする傾向にある。

表 2.4.3－1　eLC の検索サービスに登録されている e ラーニング関連製品・サービス数

プラットフォーム	34
オーサリングツール	20
コンテンツ	1329
ASP・ポータルサイト	26
コンテンツ受託開発	37
コンサルタント	34

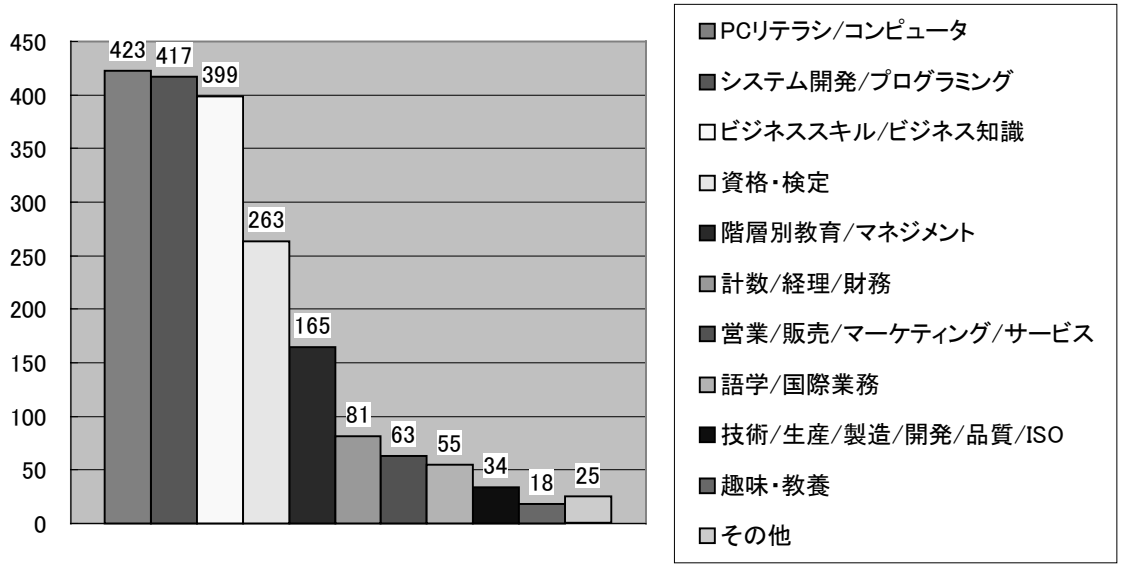


図 2.4.3 - 2　eLC の検索サービスに登録されているコンテンツ分野ごとの登録数

(エ) 標準化動向

①海外の動向

AICC(Aviation Industry CBT Committee)[4]、 IMS(The Instructional Management Systems)、ADL(Advanced Distributed Learning Initiative)[5] 、 IEEE などが米国を中心に標準化を進めている。それらの国際標準化を目指して、ISO と IEC が合同で IT の標準化を行う ISO/IEC JTC1 では、SC36 を組織して多くの国の参加を得て相互に連携しながら、プラットフォームからコンテンツなど広範囲に渡った国際標準化を目指して活動している。SC36 のスコープは学習・教育・訓練のための IT 標準化であり、SC36 の議長や事務局は米国が担当している。

コンテンツの標準規格としては、ADL が中心に進めている SCORM が今後 ISO/IEC の規格として採用される見通しになっている。SCORM は AICC の開発した CMI 規格と、IMS の開発した LOM 規格をベースにしており [6]、2001 年 10 月に SCORM Version 1.2 がリリースされている。

SCORM では、コンテンツは以下の 3 つから構成される（図 2.4.3 - 3)。

1）　コース構造

教材の章立てを定義するもので階層構造をなす。階層の末端はSCO(後述)と一対一に対応する。これはサーバ側（LMS: Learning Management System）で扱われるデータである。LMS はコース構造に基づいてページを選択し、対応する SCO を学習者の WWW クライアントに提示する。

2） SCO (Shareable Content Object)

教材の解説ページ、演習問題ページ、シミュレーションページなど。HTML、JavaScript、JavaApplet、 各種プラグインからなる WWW コンテンツで、演習問題の解答や得点などの学習履歴情報を LMS へ送信する。

3） メタデータ

学習リソースを検索・再利用するためのインデックス情報。LOM 規格では、Dublin Core を拡張した表に示す項目が規定されている。

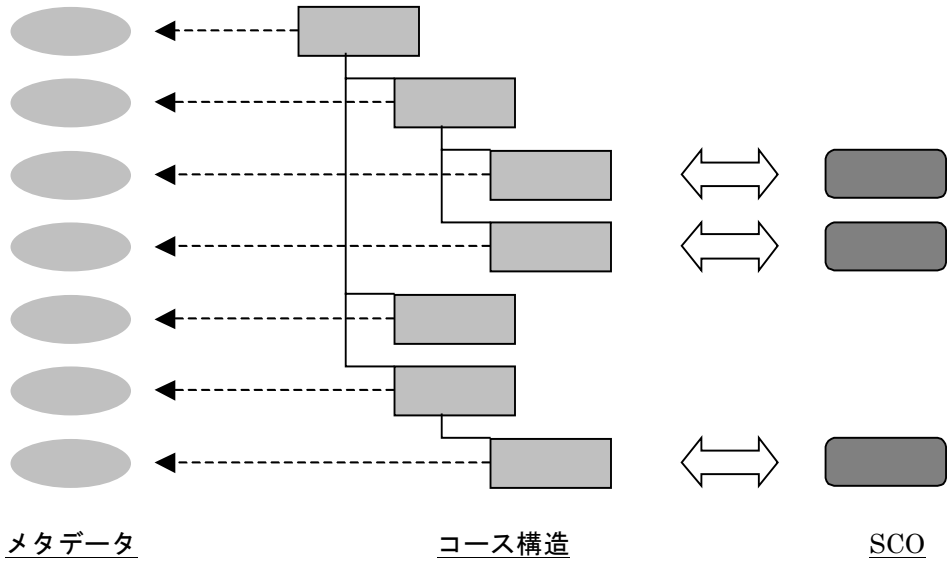


図 2.4.3 - 3　SCORM の教材構造（「e ラーニング白書 2002/2003 年版」[1)] より）

SCORM で規定されているのは、コース構造のデータモデルと XML へのバインディング、SCO と LMS との通信に関する JavaScript API とデータフォーマットである。今後の SCORM(2.x,3.x)では、学習者プロファイル、ナビゲーション、学習者主導などについても標準規格化が進められる予定となっている。

②国内の動向

eLC（日本イーラーニングコンソーシアム）の前身である TBT コンソーシアムが 1996 年より標準化活動を始めている。TBT コンソーシアムは WBT システムベンダの業界団体であったが、2001 年に NPO（特定非営利活動法人）として改組した。2000 年 4 月には、産学官共同の団体として ALIC（先進学習基盤協議会）が設立され、eLC と連携して標準化活動を進めている。日本独自の規格の作

成、CMI 規格のテストベッドシステムの開発、国際規格を使用する際のガイドラインの作成などを行っている。さらに、ALIC の次世代技術研究部会では ISO/SC36 国内委員会と連携して、協調学習に関する標準化も進めており、SC36 における協調学習関係の WG の中心的な活動も行っている。

(オ) まとめ

e ラーニングコンテンツに関する標準化はここ数年精力的に進められている。国際標準となると予想される SCORM では、学習教材の構造に対する記述を規定しているが、それは学習の手順となる教材の骨組み的な構造に関するものである。現時点では、学習者に提示されるコンテンツ(SCO)に関して、その意味内容まで踏み込んだ構造化は議論されていない。しかしながら、より学習効果の高いインタラクティブな e ラーニングシステムの実現には、教材の骨組み的な構造情報だけでなく、学習者に提示されるコンテンツも構造化し、さらにリッチな情報を持つことが必要であろう。そのために、コンテンツ構造化に関する MPEG-7 での標準化と合わせつつ、本委員会で策定している方式を、ALIC や eLC などと連携しながら国際提案していくことも考えられ、今後継続して調査・検討を行う予定である。

[参考文献]

1） 先進学習基盤協議会(ALIC)編著、「e ラーニング白書　2002/2003 年版」、(オーム社、2002)。
2） 先進学習基盤協議会、http://www.alic.gr.jp/。
3） 日本イーラーニングコンソーシアム、http://www.elc.or.jp/。
4） AICC, http://www.aicc.org/.
5） ADL, http://www.adlnet.org/.
6） 仲林、「e-Learning の要素技術と標準化」、情報処理、Vol.43、No.4、pp401-406、(情報処理学会、2002)。

2.4.4 その他の分野におけるマルチモーダルアノテーション技術の応用

(1) エンターティメント分野への応用

エンターティメントの分野においては、人間との対話が前提になる場合が多いので、機械と人間との対話をその場限りのアドホックな実現法ではなく、システマティックに実現していく必要がある。そのためには、場面や状況に応じた発話や動作などのコミュニケーション行動の枠組みを設定し、アノテーションで記述された発話内容やその属性、および、動作内容とその属性等を関連付けながら利用するアプローチが重要になると考えられる。以下、エンターティメント分野の代表であるコンピュータゲームとエンターテインメントロボットへの応用について調査、検討を行った。

①コンピュータゲームへの応用

(a) コンピュータゲームの構造のタイプ

コンピュータゲームには、以下のような種類がある。

a) アドベンチャーゲーム

ゲーム内で実現可能な場面が予め用意され、ユーザは常にその中の１つの場面に位置し、場面ごとに用意された可能な行動の選択肢から次の行動を選び、この行動選択を適切に行うことにより目的を

果たすゲーム。

b) ロールプレイングゲーム

ユーザがゲーム上の登場人物（キャラクタ）を選択できること、キャラクタの種類に応じて、各種の能力が異なり、また、成長し能力を高める機会が設定されているため、アドベンチャーゲームよりも遥かに多くの場面、状況を設定できる。味方のキャラクタと協力して目的を達成するゲーム。多くの場合、物語性が前提となり、「剣と魔法」もののように、戦闘、成長、探索、収集が重要なエピソードとなる場合が多い。

c) シミュレーションゲーム

ロールプレイングゲームに加えて、世界の構造（国家、軍事力と経済力に基づく政治力学、産業、産業の担い手である人民の支持、自然の恵みと災害、など）を導入している点が大きく異なる。さらに、交渉、策略、共同作業などを交えて、実社会と構造的に類似した行動指針を学びながら、歴史上の世界あるいは架空の世界に適応して、生き抜き、成長し、味方を増やすなどして、目的を達成するゲーム。

d) アクションゲーム

操縦、運転、射撃、武術、スポーツなどの動きとリンクさせたコントローラにより、実際と同等の動きを操作（ダイレクトマニピュレーション）して目的を果たすゲーム。

(b) コンピュータゲームとマルチモーダル対話との関係および研究動向

これらの内、a), b), c) はユーザや他のキャラクタとのコミュニケーション行動が重要な要素となっており、母国語を話せるユーザはともかく、状況に合わせて、ゲーム中の各キャラクタや群集のコミュニケーション行動をきめ細かく適切に設計することが重要である。

このために、キャラクタの発話内容や言語行為およびコミュニケーション関連動作（各種の表情や視線移動、お辞儀、握手など各種の動作）の選択可能性や適格性を状況（可能世界や場面の種類、自分のキャラクタの役割・地位、相手との親しさや身分・経済・軍事的優劣関係、行動に関する個人的制約条件など）に合わせて整理、管理、適用する共通の枠組みが望まれる。このための共通の枠組みとして、マルチモーダル情報のアノテーション技術が利用可能と考えられる。すなわち、音声、文法、談話、表情、動作などの各種のレベルのマルチモーダル対話情報、および、それらの関連性（それらの同時生起可能性など）を状況ごとに整理し、格納しておき、状況に応じたコミュニケーション行動の採択を安定的に行うことにより、個々のゲームの違いや、ゲーム設計者、ゲーム会社の違いを越えて安定的なコミュニケーション行動の実現が可能になる。この場合の状況の粒度をどの程度に設定するか、は実際に作成整理するルールの数に影響を与えるので重要であるが、私見では、スピーチアクトのレベルで行うのが、作業量的にも有用性の面からも、また、複数レベルのマルチモーダル情報の関連付けの観点からも妥当と考えられる。

・モノポリーにおける交渉戦略の研究（東工大）[1)]

上記のような市販のコンピュータゲーム以外に、マルチモーダル対話としてのゲームに関する研究もいくつかある。例えば、ボードゲームの一種であるモノポリーに関する研究[1)]では、重要な要素で

ある交渉戦略を研究するため、上級者のゲームをビデオにとり、それを解析することによって、上級者の交渉戦略を検出しようとしている。重要なコミュニケーション行動の一つである交渉の様々な状況におけるマルチモーダル情報の記録およびそれらから得た知見は重要な研究材料になると考えられる。

②エンターテインメントロボットへの応用

自律型の二足歩行ヒューマノイドロボットは、福祉における人手不足の補助のための介護ロボットや、その前提として公共施設や家庭で受け入れられるロボットとしてのエンターテインメントロボットに適した人間との自律的な相互作用が可能で、かつ、相手の人間に大きな違和感を与えない形態を有するロボットとして期待されており、近年、実用化技術の進歩に伴い、社会的脚光を浴びつつある。

(a) 各種のエンターテインメントロボット

a) SONY SDR-4X

2003 年 3 月 19 日ソニーは小型二足歩行エンターテインメントロボット SDR-4X を発表した。[2] SDR-4X は、身長約 580mm、体重約 6.5kg で、技術的には、より高度な運動性能の実現、豊かなコミュニケーション技術、などのコンセプトで開発されており、段差のある路面での歩行や、人とのコミュニケーションが豊富になっている。豊かなコミュニケーション技術という点では、「マルチモーダルヒューマンインタラクション技術」などが導入されている。マルチモーダルヒューマンインタラクション技術は、顔や音声から個人を識別・記憶するといった識別学習技術、無線 LAN と外部 PC による連続音声認識と未知語獲得、記憶に基づく対話と行動制御技術、感情や動作にあわせた音声合成や、楽譜データや歌詞データの入力で歌を歌うことが出来る、といったことが実現されている。

b) 本田技研 新型 ASIMO [3] [4]

本田技研は、世界で始めて人間のように歩くロボット発表し、現在も二足歩行型ヒューマノイドロボット技術をリードしているが、2002 年 12 月 5 日に、人の姿勢やしぐさの意味を理解して自律的に行動できる知能化技術を搭載した新型 ASIMO を発表し、2003 年 1 月から日本国内において順次、公共施設や一般企業にレンタルされている。マルチモーダル対話関連技術としては、認識技術によるコミュニケーションの進化と謳っており、主な特長は以下の通りであり、このような総合的な知能化技術を持ったヒューマノイドロボットは世界初とのことである。

移動体抽出：　人の動きをカメラで追う、人に追従して歩行する、人の接近を検知して挨拶をする。

姿勢と動作の認識：　指差した場所への移動、手を差し出すと握手する、手を振ると振り返す。

環境認識：　移動障害物の前方出現時に停止、静止障害物の発見時に迂回。

音源識別：　名前を呼ばれた方向を見る、相手の顔を見て応答、落下音や衝撃音などの検知。

顔認識：　移動中でも顔認識が可能、登録された人（10 名程度）の顔を認識。

(b) エンターテインメントロボットとマルチモーダル対話との関係および研究動向

・マルチモーダル対話ロボットの研究（早稲田大）

言語コミュニケーションと行動の理解との相互間の役割を明らかにするためにロボットを含めた

グループ会話の研究[5][6]が早稲田大学ヒューマノイド研究所で進められている。ロボットが会話を正確に理解するためには、人間の場合と同様に、発話内容だけでなく、話者が誰を注目しているかや、状況に応じて適切な人を注視することなどのマルチモーダル情報の的確な理解と表現が必要である。

まず、グループ会話の実態を調べるために、コーパス収集ツールを作成し、グループ会話における各会話参加者の発話内容と表情、視線、ジェスチャなどを記録し、対話参加者の顔の向きや発話開始点・終了点などのイベントを半自動的に付与するシステムを作成した[7]。さらに、これに基づき、必要となる状況把握と身体表現機能を整理し、状況に応じた行動制御、発話生成の指針を検討した後、グループ会話を行うロボットを構築した[6]。

・ソフトウェアロボットの研究[8]

実世界ではなく、仮想世界上に自律型のソフトウェアロボット（エージェント）を登場させて、人間との対話を行わせる研究を通じて実際のロボットに利用可能な知見を得ようとするのがソフトウェアロボットの研究である。ロボットの動作の生成や仮想世界における空間指示語の解釈の研究を行っている。現在は、空間的移動に関する動作の枠組みを研究しているが、その後はうなずきや握手などのコミュニケーションのレベルのマルチモーダル情報を扱うようになることが期待される。

上記のように、ヒューマノイド型ロボットは、少量のボキャブラリの会話をこなすだけでなく、顔や音声から個人を識別して行動する能力や、人の移動や身振りに応じた動作など、人間同士のコミュニケーションにとって重要なノンバーバルコミュニケーションの領域に踏み込んできている。生半可なノンバーバルコミュニケーション技術ではかえって誤解を受け反感を買う危険性もあるので、さらに正確で汎用性の広い技術を目指して向上させることが重要である。このためには、場面、状況に応じたマルチモーダル対話行動が重要であり、マルチモーダルアノテーション技術を共通技術として、モダリティのレイヤー（音声、文法、談話、動作など）の切り口や、状況の分類の切り口、モダリティ間の共起性等を少しずつ整理して、マルチモーダル情報のオントロジーを構築する必要がある。また、将来におけるロボットの輸出を考慮する場合、言語知識の入れ替えだけでなく、ノンバーバルコミュニケーション知識を相手国の民族あるいはコミュニティに合わせて切り替える必要があり、このためにも、独立性を保ったアノテーション技術の確立が重要である。

・マルチモーダル情報間の関連分析の必要性について[9]

身振りや指差しなどの動作は本来発話を補完するためのものであり、行為者あるいは対話者の直近の発話や動作とは互いに関連性が強い。従って、マルチモーダル情報は、そのモダリティのレイヤーで個別に記述するだけでなく、同時に観察・分析できるように記述する必要があると考えられる。マルチモーダルコンテンツにおいて、個々のマルチモーダル情報チャネルの相関性を分析しやすいように、多次元のスコア（発話者３次元、粒度の小さい動作、複合的動作、動作解釈、表情、談話、意図、対話状況など）を作成するアプローチも有益と考えられる。例えば、特定の発話の種類と、動作、談話分類、対話状況、発話意図などの間で統計的な共起性が高ければ、それらの相関はある程度の普遍

性をもつと推定でき、類似の状況において同じ相関性を仮定する許容度が高まると考えられる。

[参考文献]
1) 安村, 秋山, 小口, 新田：”モノポリーゲームにおける交渉エージェント”, 情報処理学会論文誌, Vol.43, No.10, pp.3048-3055,(2002).
2) ソニー株式会社：高度な運動性能と豊かなコミュニケーションを実現した小型二足歩行エンターテインメントロボットを開発, http://www.sony.co.jp/SonyInfo/News/Press/200203/02-0319/ , (2002).
3) 本田技研株式会社: HUMANOID ROBOT SITE, http://www.honda.co.jp/robot/ .
4) 本田技研株式会社：知能化技術を搭載した新型『ＡＳＩＭＯ』を発表, http://www.honda.co.jp/news/2002/c021205-asimo.html , (2002).
5) 小林哲則：”会話するロボット”、平成 13 年度科学研究費補助金学術創成研究「言語理解と行動制御」(課題番号 13NP0301) 報告書、pp.177-181, (2002).
6) 松坂要佐, 東條剛史, 小林哲則, “グループ会話に参与する対話ロボットの構築,” 電子情報通信学会論文誌 DII, Vol.J84-D-II, No.6, pp.898-908, (2001).
7) 松坂要佐, 小田勇一郎, 小林哲則, “グループ会話におけるマルチモーダル会話データの収録・解析システムの構築,” 第 15 回人工知能学会全国大会, 3B2-02, (2001).
8) 中嶋正之：”ソフトウェアロボットの行動生成”、平成 13 年度科学研究費補助金学術創成研究「言語理解と行動制御」(課題番号 13NP0301) 報告書、pp.182-185, (2002).
9) 対話理解技術専門委員会：”マルチモーダル情報としての動作タグに関する検討”, 電子情報技術産業協会編「ヒューマンインタフェース技術に関する調査報告書」, pp.74-79, (2001).

(2) 医療

医療分野はもっともマルチメディアサービスへの期待が大きいと言われる [1]。プライバシーやセキュリティ等の問題を解決した上で、膨大な医療データをデジタル化し、検査データ、診療、投薬、看護等の記録を一元管理し、そのデータを複数の医療機関や患者の間で共有することができれば、医学研究上はもちろん、医療の質を向上させる上でも有用であろう。以下では医療データを一種のマルチモーダルコンテンツとみなして、複数の医師間、また医師と患者のコミュニケーションのギャップを埋めるために、アノテーションがどのように利用できるかを検討したい。

(a) チーム医療における情報の共有

日本ではチーム医療はまだほとんど行われていないが、将来そのような体制が整うと、ある患者に対してそのとき最適な治療を各分野の専門医のチームが共同で検討する、ということが行われるようになる。例えば精神的に不安定になっている患者に対し、音楽療法、絵画療法、あるいは言語療法等、様々な療法をそれぞれの専門の療法士が試みたとする。各療法士がそれぞれの治療結果をもちよってこれまでの経過を確認し、今後の方針を話し合うとき、各療法士は他の療法について専門知識を有するとは限らないため、ある療法士がどのような試みをし、それに対して患者がどのような反応を示したかを正確に他の療法士に伝達し、理解させるのは困難であろう。一方、各療法士の試みた具体的な療法やそのときの患者の状態が例えば映像等のマルチモーダルデータとして示され、さらに特定のシーンや患者の示した特定の表情、動作等に対して療法士によるコメント的なアノテーションが加えられていれば、情報の共有化は格段に容易になると思われる。

(b)医師と患者のコミュニケーション

医療データを医師と患者が共有し、双方が特定のデータに対して自由にコメントしたり、質問し

たりする環境が整えば、医師と患者のコミュニケーションはより円滑になり、医療の質が向上することが期待できる。

・在宅健康管理システム

利用者の健康に関するデータ（血圧値、心電図等）を測定した結果を収集･管理し、医師のアドバイス等を送信する在宅健康管理システムは既に利用が進められている。

岩手県釜石市のせいてつ記念病院と釜石ケーブルテレビが共同開発した在宅健康管理システムでは、利用者が毎日定められた測定データを医療端末「うらら」から病院側に送る際に、データと合わせて利用者のメッセージ（例：今日は風邪をひいているので体調が悪いはずだ）も送りたいという要望があるという [2]。現状ではそのような機能がないため、必要に応じて利用者が病院に電話で連絡しているが、利用者が当日の体調について自ら説明したメッセージを測定データとともに病院に送り、医師がそれらの情報が統合された画面を検索して利用者の経過を調べることができれば、より的確な治療が期待できるであろう。

奈良県野迫川村でも、平成１１年度から遠隔医療推進試行的事業により遠隔医療（在宅健康管理）システムを導入した [3]。対象は一人暮らしの高齢者、要観察と指定された住民の世帯（５０世帯）で、「すこやかメイト」という医療端末が各家庭に設置され、血圧、心電図、心拍数、体重、体温、問診などの機能が利用できる。利用者が毎日測定したデータはセンターに蓄積され、保健婦がデータをチェックし、必要に応じ電話や訪問により助言をする。異常があれば診療所に連絡し、患者はテレビ電話により医師の診察を受けることができる。導入により、村内における医療の地域格差の解消、医師や保健婦の医療・保健業務の効率化、通院の負担軽減など患者サービスの向上、在宅介護サービスへの利用、高齢者を中心とした住民の安心感など、はかり知れない効果が上がっている、という。

・電子カルテ

亀田メディカルセンターの報告 [4]によると、患者に対するアンケートで「カルテの内容を見たいですか？」という質問に対し、「見たいです」という人が 87.1%（有効回答人数 2,059 人）であった。また、「カルテを見ることが出来たらどんなことに使いたいですか？」(有効回答人数 1,944 人)という質問に対しては「自分の健康管理に役立てたい」(1,546 人、80%)、「緊急医療時のためにいつも携帯したい」(581 人、30%)、「医師とのコミュニケーションの道具としたい」(371 人、19%)、「旅行や出張の際に携帯したい」（330 人、17%）などとなっており、カルテ中のデータを知り、自己健康管理に役立てたいという要望が多くの人にあり、また、データの携帯や、データを用いて医師とのコミュニケーションを深めることも求められていることがみてとれる。

同センターが取り組んでいる厚生労働省地域診療情報連携推進のモデル事業である医療情報ネットワーク「PLANET (Patient Centered Lifetime Anywhere on the Planet NETworking System)においては、複数の医療機関の連携により 38 万人以上の患者が電子カルテを保有する計画である。患者はカルテを閲覧するだけでなく、自身の PC から「自己記録」として自己の病状記録を書き込むことができる。現在のシステムでは、患者が記入する自己記録は電子カルテ上の特定の

データに対するアノテーションという形にはなっておらず、例えば一日の食事や体調などをフリーフォーマットで自由に記述している。将来、カルテ上の特定のデータを参照して患者が質問を投げかけたり、補足説明を求めたり、それに対して医師がコメントを記入したり、ということが可能になれば、患者は自宅にいながらにして質の高い医療サービスを受けることができるようになるだろう。

以上みてきたように、患者には、医療データを材料として、医師とより深くコミュニケーションしたい、というニーズがあると思われる。医師からの一方的なメッセージの伝達ではなく、マルチモーダルデータを利用した双方向のコミュニケーションが今後の医療システムには求められると考える。

(3) 放送、映画等のコンテンツ制作

放送、映画等のマルチモーダルコンテンツの制作においては、制作の効率化やコスト削減のため、既に作成したマルチモーダルデータの有効利用へのニーズが高いと思われる。マルチモーダルデータとして、例えば、サッカーのシュートシーンを検索する際に、映像中の人物の動作や「歓声が大きい」などの音響情報を用いる方法も研究されている[5]。このようにマルチモーダルデータの特定の部分を、アノテーションを用いて柔軟に検索できるようになれば、コンテンツ制作にも有用であると思われる。

(a)データの再利用

NHK では番組提案、取材、構成、編集を一元管理する番組制作支援システムベアトス(Beatus)を開発した[6]。2001 年５月時点で NHK スペシャル「宇宙・未知への大紀行」「ためしてガッテン」など２０番組が本システムで制作された。パソコンでの「番組提案（企画書)」「番組情報（放送日、出演者、宣伝)」「権利情報」「問い合わせメモ（試聴者サービス)」「放送台本」などの作業結果を一元管理することができる。作成された番組データは NHK アーカイブスに蓄積され、番組素材を検索したり再利用したりすることが可能である。

また、過疎地域である大分県大山町では情報化に向けた取り組みの一つとして CATV 事業に取り組んでおり、町民の日常を題材とした自主番組の制作も行っている[7]。過去に制作した番組の映像とナレーションをすべてデータベース化し、同様に繰り返される放送内容についてナレーションの再利用を行ったことは注目に値する。また、平成８年からは役場の全職員がこのデータベースにアクセスできるようにし、全職員が番組制作スタッフとなったという。素材の再利用により、番組制作が容易になった好例であろう。

(b)参考データの参照

例えば映画の字幕作成のように、作業指針の明確化が難しく、経験が必要とされてきた分野においても、コンテンツ及び作成済の字幕のデータがデジタル化され、一元管理されていれば、アノテーションを手がかりとして類似のシーンを検索し、そのシーンに付与された字幕を参考にすることで、作成が容易になると思われる。

また、複数の人で映画等のコンテンツを制作する際、色や声音、イメージなどことばで表現するの

が難しい内容を指示・伝達する際に、アノテーションを利用してヒントとなるマルチモーダルコンテンツの一部のデータを検索して参照することができれば、有効であると思われる。

例えば映画監督がある画面でどのように雨を降らせてほしいかを指定する場合に、イメージに近い作品の特定場面（例：雨の中、男性が一人ダンスの練習をしている場面）を検索することができれば、「これより少し弱く」などと簡単に指示を出すことができる。

また、例えばラジオドラマの「バーで女主人が客を迎える」という場面で、「いらっしゃい」という台詞をどのような声音で喋ってほしいか声優に指示を出したい場合、同じような場面や台詞をアノテーションを利用して検索することができれば、その例をもとにして「これよりもう少し控えめな感じで」などと指示を出すことができるであろう。

また、バラエティ番組等、客席の反応がそのままデータとして記録可能な場合は、番組終了後にシーン毎の客席の反応を分析し、どのようなシーンで客席の反応が変化したか分析し、次回の番組作成の参考とすることができるだろう。さらに、各視聴者から、番組の特定のシーン（またはその一部）を直接指定してコメントや質問が寄せられるようになると、TV 局側もよりきめ細かな対応ができ、さらにその対応例を一種のアノテーションとして蓄積すると、視聴者の満足度の高い番組作成に役立てることができるようになると思われる。

(4) ショッピング（オーダー）

顧客が既成のカタログを使わずに何かを注文する場合も、コンテンツ制作時と同様、マルチモーダルコンテンツ及びアノテーションを利用することでイメージの伝達が容易になる。

例えば、美容院で、自分の好きな女優の髪型にカットしてもらいたいとき、「女優 A のような髪型」と注文しても、女優 A のいつ頃の髪型なのか特定できず、所望の髪型のイメージが正しく店員に伝わるとは限らない。しかし、「○○という映画で、女優 A が競馬を観戦していた場面」などと特定のコンテンツのシーンをそのシーンのアノテーションを用いて検索できれば、容易にイメージを伝えることができるであろう。

その他、服飾、雑貨、インテリア等も、専門知識のない一般の顧客が自分の注文したいイメージをことばで忠実に伝えるのは難しいと思われるが、例えば「A という映画で主人公が住んでいた家の居間の内装」などと具体的な映像データのシーンを特定し、それを相手と共有することで専門知識の不足をカバーし、満足のいく注文をすることができるであろう。

[参考文献]

1) 「データブック　世界のマルチメディアプロジェクト」、NHK 放送文化研究所編（日本放送出版協会、1996）、頁 416。
2) 小笠原格、在宅健康管理システム、地域情報化―２１世紀へ向けて　第６回２１２情報化セミナー<<ダイジェスト版>>、http://www3.famille.ne.jp/~smhb-ura/rn2-1.html
3) 過疎市町村における情報化施策の先進事例―奈良県野迫川村―、http://www.kaso-net.or.jp/it/nosegawa.htm
4) 亀田メディカルセンター、患者さま中心の医療情報ネットワーク PLANET 事業について、http://planet-med.com/pdf/planet_adreport.pdf
5) 門林他、デジタルコンテンツの高度利用に関する研究、通信総合研究所季報（2001）、第 47 巻、No.3。

http://www2.crl.go.jp/kk/e414/shuppan/kihou-journal/kihou-vol47no3/toku4-1.pdf
6) 番組制作支援システム(Beatus)、
http://www.nhk.or.jp/strl/open2001/tenji/id23/
7) 過疎市町村における情報化施策の先進事例―大分県大山町―、
http://www.kaso-net.or.jp/it/ohyama.htm

2.4.5 ISOにおける国際標準化の動向

ISO（国際標準化機構)における標準化活動のうち、マルチモーダル対話コンテンツに深く関連する案件は 2 つである。ひとつは ISO/IEC JTC1/SC29/WG11 (MPEG)の MPEG-7、もうひとつは ISO/TC37/SC4 である。以下では本年度におけるこれらの進捗に関して概略を紹介する。

(1) MPEG-7

MPEG-7 の概要については昨年度の報告書で述べたので、ここでは今年度追加された内容に関して紹介する。MPEG-7 の中でマルチモーダル対話コンテンツに最も関係が深いのは MDS (Multimedia Description Scheme)であり、今年度は MDS Extension の策定作業が進められた。MDS Extension は言語コンテンツの扱い、分類スキーマ(オントロジー)の拡張、文字列の扱い等の内容を含むが、以下では言語コンテンツの内容記述のための記述ツールである Linguistic DS に関して述べる。Linguistic DS の概要については昨年の報告書を参照されたい。Linguistic DS に関して今年度になされた主な拡張は、外部データとのアラインメントの方法および抽象化(abstraction)と作用域(scope)の記述法に関するものである。

言語データと外部データ(テキスト、ビデオ、オーディオなど)との物理的なアラインメント(音声発話の場合には時間的な対応、書字発話の場合には空間的な対応)を記述するには、<MediaLocator> エレメントと start および length 属性を用いる。たとえば下の例は、transcript.txt というファイルの 120 バイト目から 14 バイト分の部分が文であることを意味する。

```
<Mpeg7>
  <Description xsi:type="ContentEntityType">
    <MultimediaContent xsi:type="LinguisticType">
      <Linguistic>
        <Sentence>
          <MediaLocator xsi:type="StreamLocatorType">
            <MediaUri>http://i-content.org/GDA/transcript.txt</MediaUri>
            <StreamSection unit="byte" start="120" length="14"/>
          </MediaLocator>
        </Sentence>
      </Linguistic>
    </MultimediaContent>
  </Description>
</Mpeg7>
```

抽象化を記述するためのツールとして、copy および substitute 属性がすでに定義されていた。copy は文字通りコピー元のエレメントを指定する属性であるが、たとえばコピー元がコピー先を含む場合などは、コピー元のエレメント全体をコピーするのではなく、一部をコピーの範囲から除く必要がある。今年度新たに導入された noCopy 属性は、コピーの範囲から除くべき部分を指す。下の例では、文全体から expected to be bigger than を除いた Tom lives in a house がコピーされ、その際

substitute 属性によって Tom が Mary に置換されるので、コピーの結果は Mary lives in a house となる。

```
<Mpeg7>
  <Description xsi:type="ContentEntityType">
    <MultimediaContent xsi:type="LinguisticType">
      <Linguistic>
        <Sentence id="TomLivesInAHouse">
          <Phrase id="TOM">Tom </Phrase>
          lives in a house
          <Phrase id="EXPECT">
            expected to be bigger than
            <Phrase copy="#TomLivesInAHouse" noCopy="#EXPECT">
              <Phrase substitute="#TOM">Mary </Phrase>
              does
            </Phrase>
          </Phrase>.
        </Sentence>
      </Linguistic>
    </MultimediaContent>
  </Description>
</Mpeg7>
```

また、このようにコピー元がコピー先を含む場合は、コピーの範囲から除かれた部分を含む最小の最大投射(上の例では a house bigger than Mary does)のコピー先(Mary lives in a house の a house)と共参照する語句によってコピー先が置換される。したがって上の例では原文は Tom lives in a house bigger than X.となり、この X は Mary lives in a house の a house と共参照する。

inScope 属性は、当該のエレメントの指示対象が属する最小の作用域を導入するエレメントを指す。下の例では、Tom loves his wife が So does Bill.にコピーされ、その際に Tom が Bill で置換されるが、inScope 属性により his が Tom loves his wife の抽象化の作用域の中にあるため、his は Bill を指し、したがって、第 2 文の解釈は「Bill も Bill の妻を愛する」(いわゆる sloppy identity)となる。

```
<Mpeg7>
  <Description xsi:type="ContentEntityType">
    <MultimediaContent xsi:type="LinguisticType">
      <Linguistic>
        <Sentence id="TomLovesHisWife">
          <Phrase id="TOM">Tom </Phrase>
          loves
          <Phrase>
            <Phrase inScope="#TomLovesHisWife" equal=”#TOM”>his </Phrase>
            wife
          </Phrase>.
        </Sentence>
        <Sentence copy="#TomLovesHisWife">
          So does
          <Phrase substitute="#TOM">Bill</Phrase>.
        </Sentence>
      </Linguistic>
    </MultimediaContent>
  </Description>
</Mpeg7>
```

Linguistic DS が言語データと関係のない意味記述にも使えることに注意されたい。下の例は、

「Hasida Koiti という人が 2003 年 2 月 13 日に眠る」という意味である。

```
<Mpeg7>
  <Description xsi:type="ContentEntityType">
    <MultimediaContent xsi:type="LinguisticType">
      <Linguistic>
        <Sentence semantics="urn:SomeOntologyOfEvents:sleep">
          <Relation type="urn:mpeg:mpeg7:cs:SemanticRelationCS:2001:time"
                    target="urn:ISO8601:2003-02-13"/>
          <Phrase semantics="urn:SomeOntologyOfObjects:person">
            <Relation type="urn:SomeOntologyOfAttributes:familyName"
                      target="urn:OntologyOfASCIItexts:Hasida"/>
            <Relation type="urn:SomeOntologyOfAttributes:givenName"
                      target="urn:OntologyOfASCIItexts:Koiti"/>
          </Phrase>
        </Sentence>
      </Linguistic>
    </MultimediaContent>
  </Description>
</Mpeg7>
```

前記のような一般的な抽象化の機能も合わせて、このような意味で Linguistic DS は、一般的な意味内容記述のツールである。

ちなみに上の例の ISO8601 は、時刻と日付の表現に関する国際標準であり、無限個の用語からなるオントロジーと見なすことができる。無限であるゆえに MPEG-7 の分類スキーム(classification scheme)によって記述することはできないが、何らかの公開された仕方で登録しておくことにより、上記のように<Relation>エレメントの中から参照できる。

(2) ISO/TC37/SC4

ISO/TC37 ではこれまでターミノロジーの標準化を推進してきた。たとえば言語コード(日本語が ja、英語が en など)は ISO/TC37 で制定された国際標準である。ISO/TC37/SC4 は 2002 年 6 月にスペインのラスパルマスの会合で新設された SC (subcommittee)であり、言語資源とその管理に関する国際標準化を目的とする。言語資源のうち、特にコーパスの書式およびその管理の方法に重点を置く。SC4 には以下のような WG が置かれ、それぞれいくつかの PWI (Potential Work Item)を担当している。正確には、これらのうち正式に発足が認められたものは WG1 のみであり、他の WG は今後順次正式発足していく予定である。

- WG1
  - ⇒ PWI: Terminology of Language Resources
  - ⇒ PWI: Linguistic annotation framework
  - ⇒ PWI: Meta-data for multimodal and multilingual information
- WG2
  - ⇒ PWI: Structural content representation scheme

⇒ PWI: Multimodal content representation scheme

⇒ PWI: Discourse level representation scheme

- WG3

⇒ PWI: Translation Memory, Alignment of parallel corpora

⇒ PWI: Segmentation and counting algorithms (characters, words, sentences etc.)

⇒ PWI: Meta-markup for GIL (Globalization, Internationalization and Localization)

- WG4

⇒ PWI: NLP Lexica

- WG5

⇒ PWI: Validation of language resources

⇒ PWI: Net-based distributed cooperative work for the creation of language resources

2002 年 11 月 21 日から 23 日にかけてフランスのポンタムッソン(Pont-a-Mousson)で WG1 のワークショップが開かれた。WG1 のミッションは、いくつかのアノテーションの枠組を比較してそれらの間の交換を可能にする標準を制定することにある。データモデルとデータ交換に関しては、言語データのアノテーションに用いられるあらゆるデータモデル(たとえばラベル付グラフ)を統一的に表現できる文書形式(ダンプ形式; dump format)を定義して、これをデータ交換用に使おうという方針に関して合意に達し、標準化作業の方向がかなり具体的になってきた。アノテーションの目的等に応じた個別の文書形式(TEI、GDA、MPEG-7 など) が適当なスタイルシートによってこのダンプ形式に変換できるようにする必要がある。ダンプ形式は機械処理用の書式であり、そのまま人間が読んで理解できる必要はない。

誰が考えてもこうなりそうな話ではあるが、とりあえずこのレベルでの合意に達したのはこのワークショップの大きな成果と言えるだろう。今後の標準化作業としては、現存する多くのアノテーションの方式を調査し、それらが表現しうるあらゆる意味をダンプ形式で表現可能にする必要がある。

ただし、個別の形式からダンプ形式への変換は高級言語から機械語へのコンパイルのようなものだが、逆の変換は逆コンパイルのような計算過程であるため、XSLT では処理できない可能性が高いと思われる。また、構造に関するアノテーションについてはこの方法で交換可能性を確立できるだろうが、たとえば品詞体系や主題役割等のオントロジーに関しては別の標準化が必要となる。

2003 年 1 月 14 日に、マルチモーダル意味表現の標準化に関する WG2 の会合が、オランダのティルブルグ大学で開催された。そもそもマルチモーダル意味表現とは何かという議論や、現在開発されているマルチモーダルシステムおよび関連プロジェクト（語彙データベース、RDF/OWL）の紹介がなされた。これからの活動については、現在あるさまざまな提案を調査することやマルチモーダルシステムに必要な情報項目について検討する方針が示されたが、具体的な作業や日程については明確にされなかった。意味論のマルチモーダル情報表現への拡張は、まだまだ研究課題が多く、標準化は時期尚早かも知れない。

TC37 の次回の会合は 2003 年の 8 月にノルウェーのオスロで開かれる。その直前に札幌での ACL

年次大会の際に SC4 の国際会議が札幌コンベンションセンターにて開催される予定である

2.4.6 アノテーションデータの統合

（本稿は 2000 年度対話理解技術専門委員会報告書第 3.5.3 節「アノテーショングラフによるマルチモーダル対話データの記述と統合」をアップデートしたものである。）

(1) 背 景

対話データは音声、音韻、統語、談話、社会的な動作等、多様な側面をもつ。それぞれの側面は固有の分析単位をもち、記述方法も異なる。しかも、過去からの分析データの蓄積がある場合が一般的であり、その多くはアノテーション上、相互の対応付けが困難である。また、将来新たな分析対象が発見され、独自の記述様式が必要になることもありうる。理論の発展、修正に伴い、分析方法の変更も十分考えられる。このような状況を見ると、アノテーション形式の標準化、固定化は非現実的であり、それぞれの記述形式の多様性、独立性を維持したまま統合することを考えるのが自然である。このような統合化問題に対する一つのアプローチとしてアノテーショングラフ[1][2]（以下、AG）が提案されている。AG は以下のような、特筆すべき特徴を備えている。

1) XML、リレーショナル・データベースによる実装が可能

2)（1 の結果による）高効率なデータ管理、スケーラビリティ

3) 柔軟性、独立性の高い、アノテーション、マルチメディアデータの統合

マルチメディアにせよテキストにせよ、およそ言語データに係わるアノテーションであれば、時間を共通の軸として統合できるはずだ、というのが AG の基本的なアイデアである。

本節では、AG の概要と現在当委員会で検討されている各種アノテーションへの適用例を紹介する。

(2) アノテーショングラフ

アノテーショングラフとは、簡単には、ループを含まないラベル付き有向グラフ(directed acyclic graph, DAG)の集合で、以下の条件を満たすものをいう。

それぞれの DAG において、

1) ノードはそれと結線しているノードがすくなくとも一つある。

2) 結線しているノード間の時間線は同一である。

時間線とは適当な要素から成る全順序集合（原典[1]では、実数空間の部分）であり、ノード集合と時間線との間に部分写像 γ を想定する。一方アノテーショングラフは、有向結線を順序としたノード上の半順序集合である。したがって二つのノード A, B について、A から B への結線が存在し、それぞれ時間線への写像が定義されているとき、その順序は γ のもとで保存される、と考えることにしておく。ラベルとは直接連結された二つのノード間のアークに付与される構造化データで、「フィールド付きレコード(fielded record)」と呼ばれる。（以下では、呼びやすさのために FIR と略記する。）構造に関しては、特に制約はないが、原典では、以下のものが提案されている。

type　アノテーションのレベル、単語層、統語層、など。

label　各アノテーションレベル固有のタグ。

class 他の FIR へのインデックス。

例

```
DAMSL
IOS:Commit　A: oh　okay

AG
[1:52.46]　　--- {D/IOS:commit} ---> [3:53.14]
　　+--- {W/oh/} --> [2:] --- {W/okay}->+
```

この例では、"oh okay" という発話に対して、談話層と単語層の二つのレベルでアノテーションを行っている。[a:b] はノードを表わす。a はノード ID、b は発話の開始（終了）点の時刻を表わす。ない場合は未定義。{t/l/c}は FIR で、 t は type、l は label、 c は class を表わす。例では、class は未定義。発話の開始時間は 52.46、終了時間は 53.14 であるが、oh の終了時間、okay の開始時間は未定となっている。

また、AG は表 2.4.6. - 1 のように XML に自然に翻訳することができる。なお、AG ベースの言語データ管理システムの構築を目指している ATLAS プロジェクト[3]では AG の XML への実装フォーマット(DTD)として AIF(ATLAS Interchange format)という規格が検討されている。

以下では、現在当委員会で検討中のいくつかのタグ体系 JEITA－DAMSL（談話機能タグ）GDA（統語タグ)、視線・動作タグ、表情（感情）タグ、J-TOBI（音声、音韻タグ）について、実際に AG による記述例を示し、その適用が可能であることを示す。各タグ体系の詳細については、前年度の本委員会 JEITA 報告書、その他で紹介されているのでそちらを参照されたい。

なお、簡単のためノードの時間情報は省略してある。　また、図中の矢印は時間線、その上の黒丸はノードを表わす。

表 2.4.6-１　アノテーショングラフの XML による記述

```
<annotation>
<arc>
        <begin-node id="1" time="52.46"/>
        <fir type="W" label="oh"/>
        <end-node id="2"/>
</arc>
<arc>
        <begin-node id="2"/>
        <fir type="W" label="okay"/>
        <end-node id="3"    time="53.14"/>
</arc>
<arc>
       <begin-node id="1" time="52.46"/
       <fir type="D" label="IOS:Commit"/>
       <end-node id="3"    time="53.14"/>
</arc>
</annotation>
```

(3)　ＪＥＩＴＡ-ＤＡＭＳＬ

以下の例を考える。その AG 表現は図 2.4.6-1 のようになる。スラッシュは発話単位を表わす。図ではそれぞれの発話に対して三種類の情報（FIR、図中の箱）、話者名、発話の転記、発話タイプが付与されている。転記テキストもアーク上のラベル（FIR）として表現する。たとえば、「まるがおのひと」は T1 時のノードから T2 時のノードにまたがる FIR である。

表 2.4.6-2　対話の例　(DAMSL 形式)

```
A:まるがおのひとです/ で やさしそうなかおです/
%     asr                                  asr
B:
%
A:
%
B:やさしそうなかお /かみのけさらさらですか
%       rr                    ir(yn)
```

図中の A・B は話者名、T1-7 は、FIR の開始・終了時間を表している。

図 2.4.6-1　アノテーショングラフによる表現

図 2.4.6-2 は、リペアなど、いわゆる disfluency のタグを含むアノテーションの AG 表現である。図の中では、F が disfluency（ここでは、言い換え、繰り返し）を表わす。また、AG では発話のオーバラップも厳密に表現できる。

表 2.4.6-3　　対話の例　(DAMSL 形式)

A: [みみは　+　みみが]{U ちょ}ふつう[こう+こう]いうんじゃなくてこうまえにでてる/ %　　asr B:　　はい/　　みみが -/ %　　ack　　un

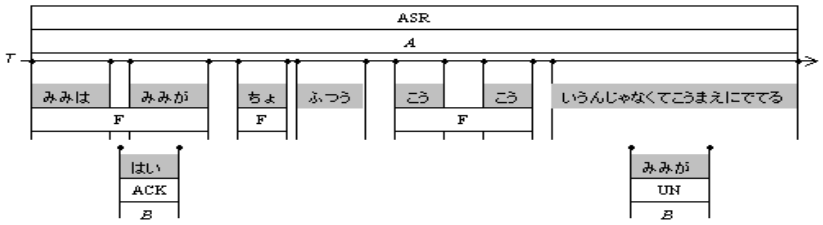


図 2.4.6-2　Disfluency を含むデータの AG 表現

(3) ＧＤＡ

図 2.4.6-3 は、AG による GDA のエンコーディングの例である。Q(uote)、 SU の GDA エレメントの属性が、AG ではそれぞれひとつの FIR に対応する。

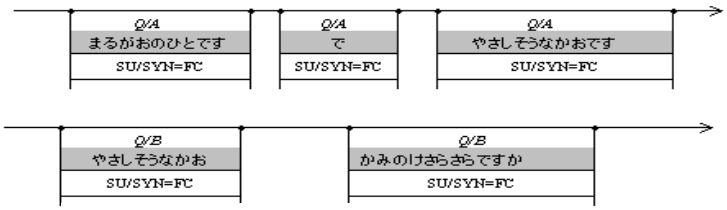


図 2.4.6-3　AG による GDA の記述

図 2.4.6-4 は、図 2.4.6-3 の一部を詳細化したものである。ここでは便宜的に複数の GDA 属性をひとつの FIR にしてある。統語構造の階層性はアークの被覆領域の重なりで（つまり、チャート的に）表現される。

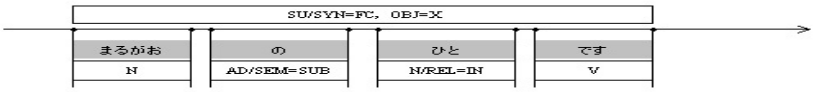


図 2.4.6-4　図 2.4.6-3 の一部の詳細化

(5) ＤＡＭＳＬとＧＤＡの統合

DAMSL と GDA を統合表記すると、図 2.4.6-5 のようになる。特にネーミングに混乱が生じなければ、両者のグラフをマージする（集合和をとる）だけでよい。

(6) 視線・動作

表 2.4.6-4 の例を考えてみる。タグは電子協[4]に基づく。AG 表現は図 2.4.6-6 のようになる。

表 2.4.6-4　視点・動作のタグを含むデータ

A:　アロンさん、こちらオーストラリアのレベッカ・モイヤさんです。 呼びかけ→B 直示→C 紹介→B 視線短　A←B

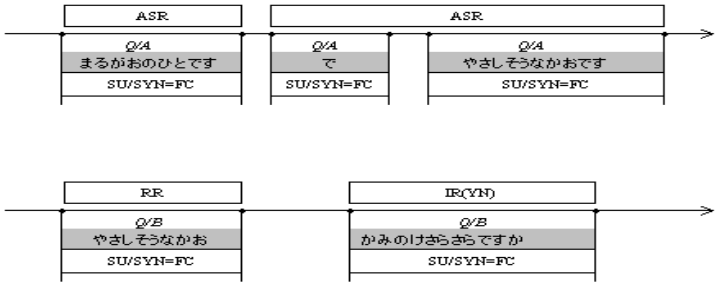


図 2.4.6-5　GDA と DAMSL の統合記述

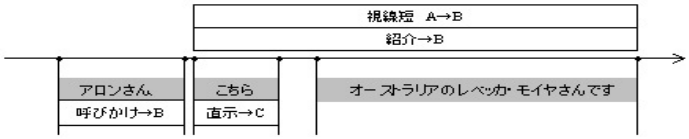


図 2.4.6-6　視点・動作タグの AG 表現

(7) 表情・感情

「はにかみ」、「考察」などの表情・感情タグは電子協[4]に基づく。　表 2.4.6-2 のデータは、AG では図 2.4.6-7 のように記述される。

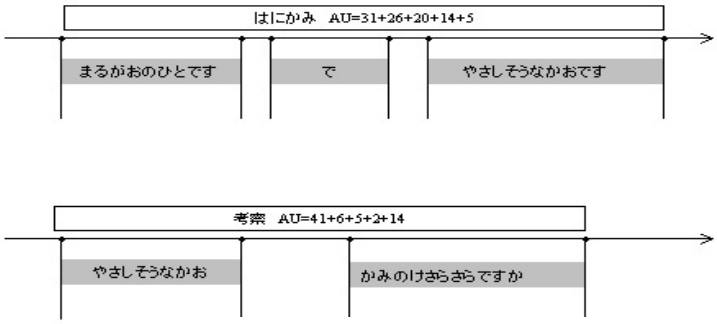


図 2.4.6-7　表情・感情タグの AG 表現

(8)　J-ＴＯＢＩ

以下の例(図 2.4.6-8)は河津[5]のデータに基づく。

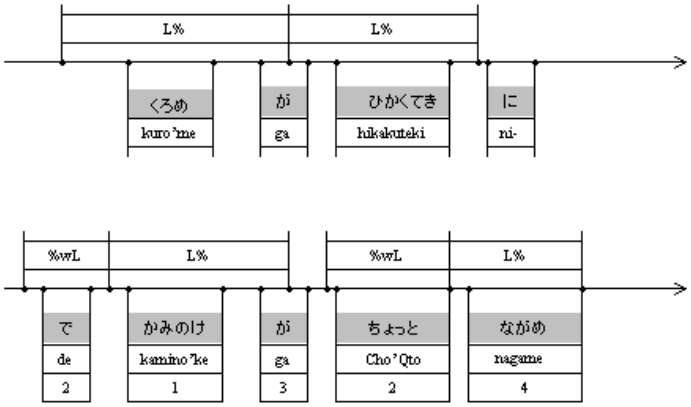


図 2.4.6-8　 AG による J-TOBI データの記述

なお、トーン層は簡単のためアクセント句のみ示している。

(9)　ＡＧの実装

AG の実装にあたっては、データの独立性を保持するため、元データのアノテーション構造、つまり、DTD を一切変更せず、AG へ（から）の変換を一意におこなえるようにするのが望ましい。AG 自体の記述は RDB でも、XML でもよい。以下に変換手法の一例を示す。元データをソース、変換後のデータをターゲットと呼ぶ。また、ソースは XML で記述されているものとする。

まず、AG への変換は、ソースを解析後、XML 構文木(DOM ツリー)上の各ノードについてそれがスパンしている文字列の開始点、終了点を求め、ノードに付随するタグ情報を使って、以下のようなアークを作る。一行がひとつのアークに対応する。

DTD	BEGIN	END	TYPE	ATTRNAME	VALUE	REF
DAMSL001	1	2	slash	UttrType	asr	null
DAMSL001	1	2	pcdata	Text	まるがおのひとです	null
DAMSL001	1	2	slash	Speaker	A	null

ここで DTD は DTD の識別子、BEGIN はアークの開始ノード番号、END はアークの終了ノード番号、ATTRNAME は、ソースタグの属性名、VALUE は、その値、TYPE は、属性が付随しているタグ名（テキストデータは PCDATA とする)、REF は、属性値の参照先（もしあれば）である。ノードはさらに以下のような下位情報をもつ。

node	time
1	4.22
2	6.76

ここでは、node はノード番号、time はデータ開始時からのオフセット時間を表す。
一方､AG から元のアノテーションフォーマットへの変換は､DTD を使ったチャートパーシング(chart parsing)で DOM ツリーを構成し、適当なプリントメソッドで通常の XML フォーマットにすればよい[6]。

(10) アノテーショングラフの課題と最近の動向

AG の重要な課題としては、階層構造の直接的な記述とその上での検索をどうするかという問題がある。AG では、階層構造（たとえば、統語構造）は基本的にチャート式の構造に変換されてしまうので､各階層間の親子関係は明示的に表現されない。この点の解決案として Bird&Liberman[2]では、アークに親ノード（アーク）への参照情報を明示的に導入することが提案されている。さらに、一般的に異なったアノテーション構造間の任意の対応付けを可能にするため、Equivalence Class という概念が導入されている。

また、前節でも述べたように AG の自然な拡張として AG 構造を RDB で表現することが考えられる。この方向での取り組みは文献[7]に詳しい。アノテーションデータの大規模化、共有化、およびデータの保守を考えると、AG の RDB へ実装は当然あるいは必然と言える。AG 構造を RDB にマップするのは前節でも述べたようにそれほど困難ではない。ただ､SQL では検索要求に十分に耐えることができないという問題がある。たとえば、正規表現を用いた検索は言語コーパスを利用する上で極めて重要であるが、いまのところ SQL で実現するのは容易でない。文献[7]では AG の RDB での利用に向けた検索言語の改良が提案されている。

その他最近の動向としては､AG のツールキットがペンシルバニア大学の LDC で開発され､無償配布されている。配布サイトの情報は、www.ldc.upenn.edu にて入手可能である。また、ツリーバンクにみられるようなツリー構造を持ったアノテーションデータの編集を AG でも実現しようという話が文献[8]で紹介されている。文献[9]では、AG ツールキットを用いたアノテーションインターフェイス構築の例が紹介されている。音声動画像との連動も実現されている。

［参考文献］

[1] Steven Bird, and Mark Liberman, Annotation Graphs as a Framework for Multidimensional Linguistic Data Analysis, In Proc. of the workshop “Towards Standards and Tools for Discourse Tagging,” ACL, Maryland, (1999).

[2] Steven Bird, and Mark Liberman, A Formal Framework for Linguistic Annotation, Technical Report MS-CIS-99-01, Department of Computer and Information Science, University of Pennsylvania, (2000).

[3] Steven Bird, David Day, John Garofolo, John Henderson, Christophe Laprun, and Mark Liberman, ATLAS: A Flexible and Extensible Architecture for Linguistic Annotation,

In Proc. of the Second International Conf. On Language Resources and Evaluation, pp 1699-1706,

Paris European Language Resources Assoc.,(2000).

[4] 電子協、自然言語処理システムに関する調査報告書、日本電子工業振興協会、00-情-７、 (2000).

[5] 河津恵、マルチモーダル対話のための音声・談話ラベリング、修士論文、慶応大学大学院　政策・メディア研究科、 (1999).

[6] 丸山、田村、浦本、 XML と JAVA による Web アプリケーション開発、ピアソン・エデユケーション、東京, (1999).

[7] Xiaoyi Ma, Haejoong Lee, Steven Bird and Kazuaki Maeda. Models and Tools for Collaborative Annotation. Proceedings of the Third International Conference on Language Resources and Evaluation, Paris: European Language Resources Association, 2002.

[8] Scott Cotton and Steven Bird. An Integrated Framework for Treebanks and Mltilayer Annotation. Proceedings of the Third International Conference on Language Resources and Evaluation, Paris: European Language Resources Association, 2002.

[9] Steven Bird, Kazuaki Maeda, Xiaoyi Ma, Haejoong Lee, Beth Randall, and Salim Zayat. TableTrans, MultiTrans, InterTrans, and TreeTrans: Diverse Tools Build on the Annotation Graph Toolkit. Proceedings of the Third International Conference on Language Resources and Evaluation, Paris: European Language Resources Association, 2002.

2.5 ヒアリング

ここではコーパスに関連する 2 件のヒアリングの概略を述べる。これらのうち、「言語コーパスと言語知識の統合的管理」は平成 14 年の 3 月に行われたものである。

2.5.1 言語コーパスと言語知識の統合的管理

伝 康晴（千葉大学文学部）

近年、さまざまな機関において、言語・談話コーパスやマルチモーダルコーパスの開発が進んでいる。一般に、これらのコーパスには、さまざまな種類の言語学的情報やその他の情報がタグとして付与されており、これらのタグの仕様の標準化に関する活動も行われている。こうしたタグの仕様は、ある種の知識体系と考えることができる。たとえば、形態論的タグの仕様（語や品詞の定義）は、まさに形態論的知識の体系といえる。

このような言語学的知識の体系は、辞書やシソーラスとして電子化されている。しかし、それらはコーパスとは独立に策定されたものであり、ある言語コーパス中で用いられているタグの仕様を、電子化された言語知識として並行して策定している例はほとんどみられない。あったとしても、コーパス作成後にそこに出現する単語を語彙目録としてまとめるといった形態を取ることがほとんどであり、コーパス開発段階から言語知識と並行して開発を進めるという形態を取ることはごくまれである（EDR コーパスはこれに近い形態を取っている）。

このような方略を欠く際の問題点は、以下のようにまとめることができる。

（1） 哲学的問題

たとえば、語を考えると、コーパス中に出現する語とは、特定の使用文脈を超えた斉一性をもった語タイプの、特定の状況における使用、すなわちトークンとしての語である。ある語のトークンをコーパス中に見出すことができるのは、その語のタイプに関する知識（語形や品詞など）をコーパス設計者が有しているからであり、その意味においてタイプの記述はトークンの記述に先立つ。しかるに、現状のコーパス開発はトークンの記述に終始しており、その過程において暗黙に仮定されているタイプの記述は、作業マニュアルとして文書化されることはあっても、電子化された資源としてコーパスと同列に管理されることはない。

（2） 工学的問題

タイプの記述を、電子化された資源として、コーパス開発時に利用することができないことから、コーパス（トークン）の記述において生じうる不整合を未然に防止するすべをもたない。たとえば、「チ（地）」という語を接尾辞ではなく、普通名詞としてしか認定しない体系にしたがって形態論的情報の付与作業を行っているとき、ある使用文脈で現れた「地」に対して、作業者が誤って接尾辞という品詞を付与してしまうといった類のミスを避けることができない。この種のミスは、タイプの記述（この場合、電子化辞書）をコーパスの記述と並行して開発し、形態素解析システムのようなしかるべき工学的手段によってタグ付け作業を支援するようにすれば、未然に防ぐことができる。

筆者らのグループは、このような問題に対処するために、言語コーパスと言語知識を並行して開発し、統合的に管理する手法について検討している。その理論的な側面については、(伝, 2002) で詳述

している。そこでは、コーパス（トークン）の記述として注釈グラフ(Bird & Liberman, 2001）を採用し、辺ラベルに対する型階層を設定することによって言語知識（タイプ）の記述を行っている。また、(Asahara, et al., 2002）では、リレーショナルデータベースを用いて、コーパスと言語知識を統合的に管理する手法について具体的に述べている。(浅原ほか, 2002）では、上述したようなコーパス開発支援からさらに一歩進んで､複数の相異なるタグの体系の間の関係までも言語知識として記述し、それを利用して体系変換を行うような手法について述べている。

このようなコーパスと言語知識の統合的管理の応用範囲はかなり広い。漢字表記や送り仮名の表記のゆれを統一するといったことも、上記の体系変換の応用として行うことができるし、コーパス開発時には記述していなかったような情報を用いて（言語知識の側に記述しておいて）コーパスを検索するといったこともできる。意味論的オントロジーを記述しておけば、コーパス検索の柔軟さもずいぶんと増すであろう（このような試みはすでに行われているようであるが)。

重要な点は、このようなコーパスと言語知識の統合的管理の恩恵を、コーパス開発のなるべく初期の段階から利用することである。そうすることによって、コーパスの整合性の維持が容易になるだけでなく､作業者に利用可能な情報の範囲も飛躍的に広がる｡作業マニュアルは紙に書くものではなく、電子化された資源としてコーパスと同列に管理すべきものなのである。

[参考文献]

Asahara, M., Yoneda, R., Yamashita, A., Den, Y., & Matsumoto, Y. (2002). Use of XML and relational databases for consistent development and maintenance of lexicons and annotated corpora. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pp. 1372-1378.

浅原 正幸・米田 隆一・山下 亜希子・伝 康晴・松本 裕治. (2002). 語長変換を考慮したコーパス管理システム. 情報処理学会論文誌, 43, 2091-2097.

Bird, S., & Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Communication*, 33, 23-60.

伝 康晴. (2002). 言語学的対象のオントロジー. 人工知能学会研究会資料, SIG-SLUD-A202, 39-44.

2.5.2 意味構造を用いた情報検索

宮田 高志（科学技術振興事業団）

（１） 背景

近年、インターネットの流行を背景として大量の機械可読データとくにテキストが蓄積されてきており、その中から必要な情報をとりだす技術の要求が高まっている。一方、自然言語処理研究の分野では 1980 年代から研究されはじめた統計的アプローチがある程度の成功をおさめ､「書き言葉で比較的短い文」といった限定された条件ながら高精度な形態素解析器・構文解析器が作られるようになってきた([1][3][4])。このような状況の下で、本研究では人間と計算機の協調によって、従来のキーワードに基づく検索では不可能であるような「内容」に基づく検索を実現する。具体的には、検索対象とする文書をあらかじめ形態素解析および構文解析してグラフの形で蓄積しておき、検索質問もグラフで入力するという方法をとる。

（２） 研究のアプローチ

文書を単なる文字や語の列として検索するのではなく、グラフとして検索することで次の三つの利

点が生じる。

① 高精度

入力したキーワードがたまたま同時に出現しているだけで内容としては無関係な文書が排除されるので、より正確な検索を行える。

② より細かいヒントの提示

従来の検索システムにおいてはユーザに提示されるのは文書の列やそれらをクラスタリングしたもので、質問を修正するヒントとしては非常に粒度が粗い。本研究では蓄積してある文書はあらかじめ解析してあるので、その情報をもとに「入力した語と内容的に共起しやすい語」といった、より細かいヒントを提示することができる。

③ ユーザの意図の適切な表現

単純なキーワードの集合では、ユーザがどんな意図をもって検索を行っているのかを推測することはほとんど不可能である。質問をグラフに拡張することは、ユーザが自分の意図を適切に表現するために必要な道具立ての一つである。

検索において言語的な構造を利用するという研究はすでにある[5,6,7,9]が、ほとんどが①の観点からの利用であり、その結果は「キーワードだけを使った場合と大きな差がない」というものであった。本研究では主に②の利点が重要であることを論じる。

また、蓄積する文書を詳細に解析しておくことは、検索に役立つだけでなく解析技術研究のためのデータ蓄積という観点からも有用である。もしこのようなグラフ照合に基づいた検索が一般に広まれば、自分のデータを(自動解析しただけではなく)人手をかけて自動解析の誤りを修正して公開した方が人目に触れやすくなる。現在の自然言語処理研究ではこのような、誤りが少ない文書が多量に必要とされており、研究の進展のためにも有意義である。そのため、各文書および検索質問のグラフ構造(解析結果)は検索だけに特化したものではなく、GDA[2]という一般的な形式で記述する。

GDA による意味構造の記述

Global Document Annotation (GDA)タグ集合 [2]は XML のインスタンスの一つであり、タグの構造とそれらの属性によって、テキストに言語的な情報を付与するための枠組である。例えば図 2.5.2 - 1 は「太郎が買った本を破った」という文に対してタグを付与したものである。

```
<su>
  <adp opr="obj">
    <np>
      <adp opr="agt">太郎が</adp>
      <v obj="mgn">買った</v>
      <n>本</n>
    </np>
    を
  </adp>
  <v agt="hanako">破った</v>
</su>
```

図 2.5.2 - 1:「太郎が買った本を破った」に対するタグ付けの例

「太郎が買った本」が<np>というタグに囲まれていることで、「太郎が」が「破った」ではなく「買った」に係っていること、「太郎が」を囲む<adp>タグの属性 opr に agt という値を指定することで、「太郎」が「買う」というイベントの agent (動作主)であること、などが示されている。また「破った」を囲む<v>タグの属性 agt に hanako という ID を指定することで、この文では明示されていない hanako (という ID 値で参照される者)が「破る」の agent であることも示されている。

図 2.5.2 - 1 のタグ付けされた文を一定の規則に従って変換して得たグラフが図 2.5.2 - 2 である。

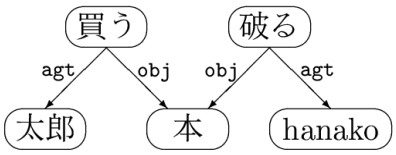


図 2.5.2 - 2: 図 2.5.2 - 1 のタグ付き文を変換して得たグラフ

GDA ではタグ付き文書から変換されるグラフ(意味構造グラフとよぶ)はラベル付きの有向グラフとなるが、本研究では簡単のため辺のラベルと向きを無視し、無向グラフとして扱う。

（3） 情報検索への意味構造グラフの利用

本システムにおける検索は次の手順で行う。

- 検索対象である文書群は(巨大な)意味構造グラフ G_D に変換しておく。
- 問合せはその場で意味構造グラフ G_Q に変換する。
- G_D の部分グラフで G_Q と最も``よく一致する"ものを求める。

上記の問題は graph embedding とよばれ、acyclic なグラフにおいてさえ完全一致解・最大一致解ともグラフのサイズに関して NP 困難であることが知られている[10]。よって本システムでは、まず頂点の組合せだけで第 n 位までの解を求めてから、辺のつながりぐあいを考慮して厳密に比較する、という二段階で近似解を求める。n を大きくすればより正確に最大一致解を求めることができる。現在のところ n は 1000 に固定しているが、頂点の組合せを数え上げるアルゴリズムは[8]をもとに考案したもので、必要に応じて動的に n を増やすことができるので、「スコアの差が一定値以上になるまで」というように動的に変更することも可能である。

図 2.5.2 - 3 に、検索時における問合せ・データベース・シソーラスの関係を示す。

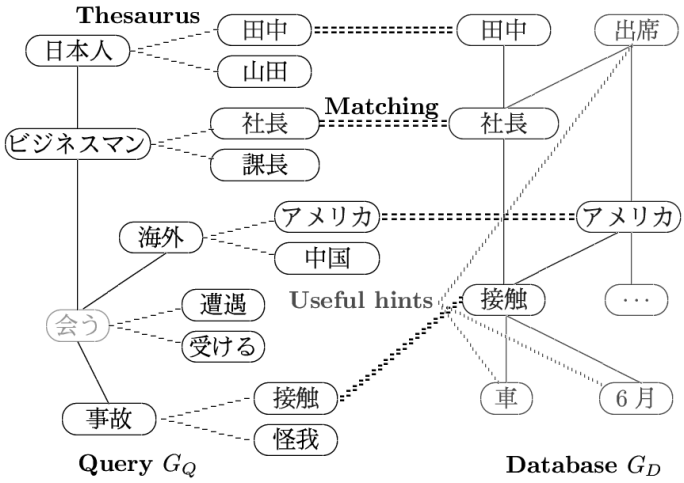


図 2.5.2 - 3: 問合せ・データベース・シソーラスの関係

入力した問合せ「日本人ビジネスマンが海外で事故に会う」を、その意味構造グラフに変換した結果が G_Q である。ユーザは問合せの意味構造グラフ(問合せグラフ) G_Q の各頂点に類義語(と重み)を付与することができ、各類義語がデータベース中のグラフの頂点と照合される。照合した頂点からなる部分グラフ G_D が問合せグラフ G_Q とどれだけ近いかによって G_D のスコアが計算される。さらに問合せグラフ G_Q には含まれていなかったが、G_D に隣接する頂点「出席」「車」「6月」などに関する情報を集計し、ヒントとしてユーザに提示する。

（4） プロトタイプシステム

① 初期画面

ユーザに最初に提示される画面を図 2.5.2 - 4 に示す。

図 2.5.2 - 4: 初期画面

実装したプロトタイプシステムは、Web ブラウザを通じて CGI 経由で検索を行い、ブラウザの履歴管理機能をそのまま利用できるようになっている。初期画面では、文またはキーワードの AND-OR で検索条件を指定する。

② 文書リスト

（4） プロトタイプシステムの頁で説明したグラフ検索は文書単位で行う。すなわち、まず頂点の組合せだけで m 位までの文書に絞り、それぞれ n ヶ所ずつ構造まで含めて照合する (n=1000)。その結果最もスコアの高かった部分グラフのスコアをその文書のスコアとして、文書リストを作成する。図 2.5.2 - 5 に文書リストの様子を示す。

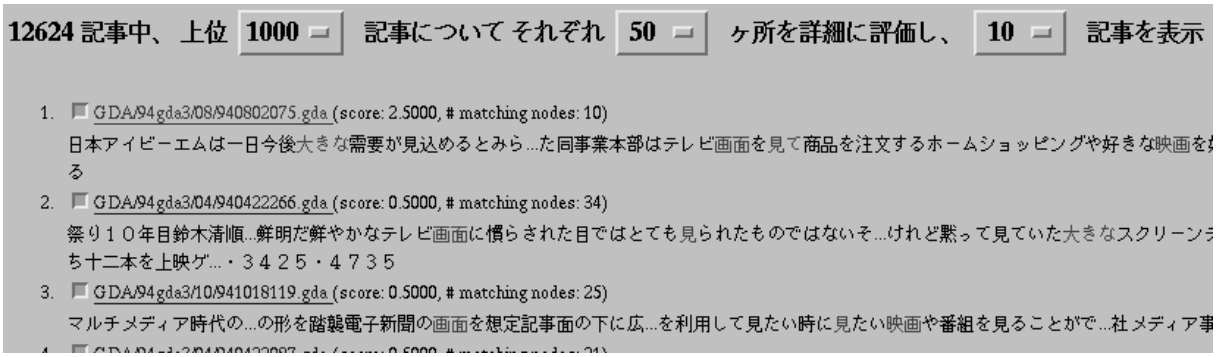


図 2.5.2 - 5: 文書リスト

クリックで文書そのものおよび合致した場所が表示される。

③　意味構造の入力

ユーザは図 2.5.2 - 6 のインタフェースを用いて、問合せグラフを直接編集することができる。

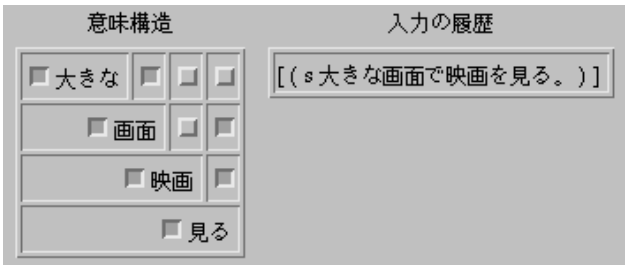


図 2.5.2 - 6: 意味構造編集画面

このインタフェースでは、二つの語の行と列が交差するマスにチェックを入れることで、それらの間に辺があることが表される。また、各語の前のチェックをはずすことで、条件から削除される。

なお、図 2.5.2 - 6 のかわりに二つの語の間の辺を直接描画するインタフェースも用意しており、どちらが使いやすいかを比較する予定である。

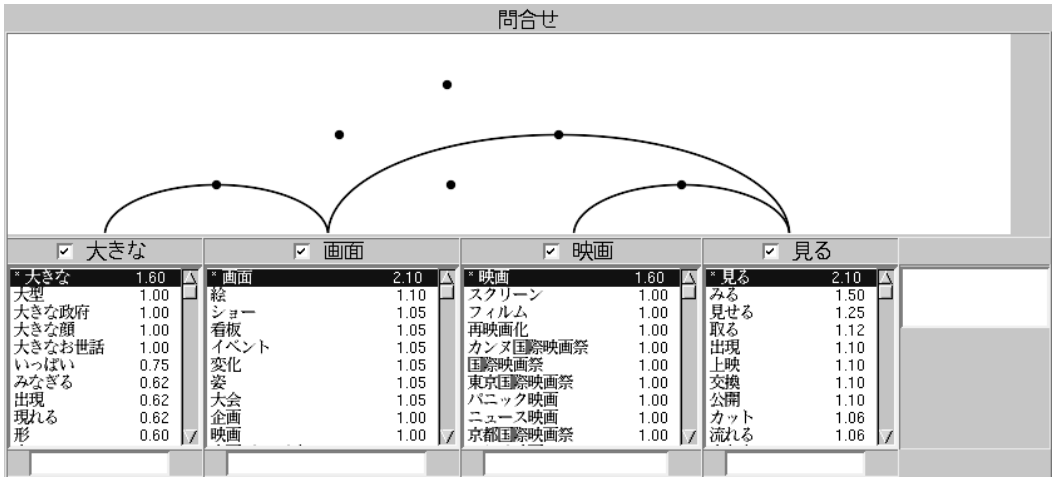


図 2.5.2 - 7: 意味構造編集画面(改良版)

④　類義語の入力

本システムでは、頂点同士の照合を語が(文字列として)同一かどうかで行うので、適切な類義語を指定することが重要となる。現在のところ、グラフ照合がユーザにとってどの程度直観的な動作をするのかが不明なため、類義語を自動的に展開することはしておらず、言語工学研究所(株)のシソーラスを使って「重み付き類似度」の順にソートしたものをユーザに提示するにとどめてある。

図 2.5.2 - 8: 類義語選択画面

一つの頂点に対して複数の類義語を重みを付けて指定することができる。二つの語 x、y のシソーラスにおける類似度を sim(*x*,*y*)とするとき、頂点 *X* に対して各類義語 w_i が重み a_i で指定されている場合、シソーラス中の語 *w* の「重み付き類似度」score(*w*)は、

(式 1) $\mathrm{score}(w) = \Sigma_i a_i \,\mathrm{sim}(w_i,w) / \Sigma_i a_i$

で計算される。さらに、シソーラスの不備を補完するために、入力した類義語と文字列として重なるような語も類義語の候補として提示している。

⑤ 隣接語表

データベース中の文書は全て解析されているので、入力した問合せ中の各類義語とどのような語がデータベース中の意味構造グラフにおいて隣接しているかを、表の形でユーザに提示する。これを隣接語表とよぶ。

図 2.5.2 - 9: 隣接語表

隣接語表の各列 *C* は、頂点 *C* に指定された各類義語と隣接する語が、その語が含まれる文書の最大スコアおよび共起頻度でソートされて並べられたリスト(隣接語リスト)である。なお、頂点 *X* の隣接語リスト中の語は問合せで *X* と隣接している頂点 Y の類義語として利用されることが多いので、図

2.5.2 - 7 の改良したインタフェースでは、類義語選択画面に統合して提示している。問合せ中で頂点 X が頂点 Y に隣接している時、Y の類義語リストに提示される w は、(式 1)で計算した重みに加えて、

a. それが X の隣接語リストにも入っていたら、0.5

b. 同じく w が Y の類義語 y のどれかと文字列として重なっていた時は、0.1

だけ加算する。ただし、ユーザが明示的に重みを指定した時はその値が使用される。

（5） 典型的な例

本システムの特徴を端的に示す例として、「ロボットを使って住宅を安く作る」ということが書かれた文書を検索する例題をとりあげる。

- まずユーザは一つの頂点に一つだけの類義語を指定して、「ロボット」「使う」「住宅」「安い」「作る」を入力する。
  ⇒ この状態では類義語の出現だけで文書のスコアが計算されるので、上位の文書は全て同じスコアであり、どの文書を選択すべきかを考慮する手がかりはない。
- 次にユーザは「ロボット―使う」や「使う―作る」等の辺を追加する。
  ⇒ 辺を追加することで各文書のスコアは変わるが、類義語が適切に指定されていないので、上位にそれらしい文書がみあたらない。
- 検索の結果提示された隣接語表を見ると、(類義語リストには提示されていなかったが)「住宅」に隣接する語として「建設」「建築」を発見、「作る」の類義語に追加する。
  ⇒ 目的の文書に到達

上の例において、「『建設』や『建築』なら、『作る』の類義語として最初から思いつくのではないか」という批判もあり得るが、実際にやってみると適切な類義語を考えつくのは意外に難しいことがわかる。また、「上の例では単にシソーラスに不備があっただけであって、予めシソーラスを作っておけば十分である」という批判もあり得るが、あらゆる分野のあらゆる場面で有効な、実用的規模のシソーラスを予め作っておくのは非現実的である。本研究では「その場で」「文書自身から」作ることが重要であると考える。

（6） 予備実験

1994 年度の毎日新聞約 10 万記事について、KNP[4]で解析したものを意味構造に変換してサンプルデータとした。このデータの統計を表 2.5.2 - 1 に示す。

表 2.5.2 - 1: サンプルデータに関する統計情報

文書中の単語数	13,655,113
意味構造中の頂点数	13,652,694
意味構造中の辺の数	10,928,259
シソーラスの語数	149,270

（6） 予備実験の頁の例のような検索課題 15 題について、著者自身が(正解を知ったうえで)被験者となって検索を行った時の様子を記録し、入力した問合せグラフやどのように絞り込んだかなどを

調べた。被験者は正解を知っているので、問合せの修正には次のような制限を課した:

- 各課題は正解文書のスコアが他の文書のスコアより十分に大きい時に達成されたとみなす。スコアが同じでたまたま最初に提示された場合は課題が達成されたとはみなさない。
- 被験者は類義語リストおよび隣接語表に提示された単語のみを使って問合せを修正することができる。最初の問合せは課題の説明文にある語だけを使う。
- システムは最初に語の出現だけで文書を 1000 個に絞り、それらのスコアだけを詳細に再計算して文書を順位付ける。ユーザにはそれらの上位 10 文書だけが提示される。

各課題に対して平均して約 50 分の時間が費やされた。これは主にシステムの応答時間がまだ遅いからである。

被験者が(課題を読んだあと)最初に入力した問合せグラフに含まれる頂点や類義語・辺の数および、その時に候補に挙がった文書数とその中での正解の順位を表 2.5.2 - 2 に示す。

表 2.5.2 - 2: 初期問合せグラフ

課題	頂点数	類義語数	辺	文書数	順位
1	3	3	0	8773	--
2	3	3	0	8585	1, 10
3	3	3	0	11966	1, 6
4	4	4	3	5256	--
5	3	3	0	573	--
6	2	2	0	1735	1, 10
7	3	3	0	1178	1, 10
8	5	5	4	11407	--
9	3	3	0	3183	--
10	3	3	0	3752	--
11	5	5	4	14680	--
12	3	3	0	7	--
13	3	3	0	6023	--
14	3	3	2	7724	--
15	2	2	0	7731	--
平均	3.2	3.2	0.87	6171.53	--

表 2.5.2 - 2 からは、例えば課題 2 では「頂点が三つで辺をもたないグラフが入力され、各頂点には一つずつ合計三つの類義語が指定された」こと、その時に一つでも類義語を含む文書が 8585 個挙げられ、正解は 1 位～10 位(これらは同じスコアであった)であったこと、が読み取れる。順位が -- の欄は、上位 10 位までに正解が出現しなかったことを表す。

表 2.5.2 - 3 には、課題を達成したもしくは諦めた時点での問合せグラフの情報とそれまでにかかったステップ数を示す。

表 2.5.2 - 3: 最終問合せグラフ

課題	頂点数	類義語数	辺	文書数	順位	ステップ数
1	4	22	3	9604	1, 1	7
2	4	5	5	11044	1, 6	3
3	3	3	0	11966	1, 6	1
4	4	14	4	6811	2, 3	6
5	5	5	2	14958	1, 9	3
6	2	2	1	1735	2, 10	>3
7	5	5	0	3688	1, 5	2
8	5	14	5	17735	1, 1	5
9	5	5	3	11805	--	>6
10	5	12	4	10699	--	>6
11	5	8	7	17026	--	>7
12	4	9	3	8701	--	>6
13	5	11	4	31851	--	>9
14	3	3	2	7728	--	>5
15	5	6	4	56911	--	>6
平均	4. 27	8. 27	3. 13	14817. 27	--	5. 0

表 2.5.2 - 3 において、ステップ数が「>n」とあるのは n ステップまでで課題を諦めたことを示す。15 題の課題のうち約半数が達成され、正解文書の順位も十分上位に提示されていることがわかる。

表 2.5.2 - 4 は最終問合せを得るまでに追加された類義語の数とその内訳である。

表 2.5.2 - 4: 追加した類義語の内訳

課題	類義語数	課題そのもの	シソーラス	隣接語表
1	19	1	18	0
2	2	1	0	1
3	0	0	0	0
4	10	0	9	1
5	2	2	0	0
6	0	0	0	0
7	2	2	0	0
8	9	0	8	1
9	2	1	0	1
10	9	2	7	0
11	3	0	3	0
12	6	0	5	1
13	8	1	5	2
14	0	0	0	0
15	4	2	1	1
平均	5. 07	0. 8	3. 73	0. 53

表 2.5.2 - 4 には類義語はだいたいシソーラスを見て追加されることが多いことが示されているが、隣接語表からも少なからぬ語がとられていることがわかる。

課題 6、9、11、12 については、最終的に達成できなかったが、正解から逆に有効な類義語を推測して追加することで正解まで到達させることができた。もともと達成できた課題と合わせて、それらの最終問合せから辺を除いた場合に正解文書の順位がどれだけ変化するかを見ることで、辺の役割の度合いを示したのが、表 2.5.2 - 5 である。

表 2.5.2 - 5: 辺の有無と順位

	キーワードのみ			辺あり		
課題	頂点数	類義語数	順位	辺の数	順位	文書数
1	4	22	1, 40	3	1, 1	9413
2	5	5	1, 12	4	3, 4	11044
3	3	3	1, 6	0	1, 6	11966
4	5	14	1, 29	4	2, 2	6811
5	5	5	1, 8	2	8, 8	14958
6	2	2	1, 17	1	9, 9	1735
7	5	5	1, 5	3	1, 1	3688
8	5	14	1, 20	4	1, 1	17735
9	5	6	4, 50	3	26, 28	12210
11	5	8	8, 50	4	1, 1	17245
12	3	10	1, 1	2	1, 1	8773
平均	4. 27	8. 55	--	2. 73	--	10507. 1

表 2.5.2 - 5 では、表 2.5.2 - 4 までのように上位 10 文書ではなく、上位 50 文書まで表示して順位の変化を調べている。例えば、課題 1 では 3 本の辺を追加することで、40 位にあった正解文書が、1 位になったことがわかる。課題 9 を除いて、ほとんどの場合、辺を追加することで正解がより上位にスコアリングされることが示されている。

（7） まとめ

本報告書での報告事項をまとめると、次の通りである:

- 現在の自然言語処理技術でも、人間と協調することで、十分役にたつ。
- グラフ照合問題は NP 困難であるが、アルゴリズムを工夫することで十分検索に応用することができる。
- 文書をあらかじめ構造化しておくことで、検索に有用なヒントが提示可能となる。
- インターフェースは洗練する必要がある。
- もう少し規模の大きな評価実験を計画中である。

[参考文献]

[1] Eugene Charniak, A Maximum-Entropy-Inspired Parser, In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP) and the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp.NAACL 132-139, 2000, USA, April-May.

[2] Koiti Hasida, *Global Document Annotation* (http://www.i-content.org/GDA/).

[3] Taku Kudo and Yuji Matsumoto, Japanese Dependency Structure Analysis Based on Support Vector Machines, In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing*

and Very Large Corpora, pp.18-25, 2000, Hong Kong, Oct.
[4] Sadao Kurohashi, *KNP version 2.0 b6 manual* (http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/knp.html).
[5] Kazu Miyakawa, Takenobu Tokunaga, and Hozumi Tanaka, Information Retrieval Using Case Frame, In *Proceedings of the Fourth Annual Meeting of the Association for Natural Language Processing*, pp.112-115, 1998 (in Japanese).
[6] Jose Perez-Carballo and Tomek Strzalkowski, Natural Language Information Retrieval: Progress Report, *Information Processing and Management*, Vol.36, pp.155-178, 2000.
[7] Tomek Strzalkowski, Natural Language Information Retrieval, *Information Processing and Management*, Vol.31, No.3, pp.397-417, 1995.
[8] Henry Thompson, Best-First Enumeration of Paths through a Lattice -- an Active Chart Parsing Solution, *Journal of Computer Speech and Language*, Vol.4, pp.263-274, 1990.
[9] Jason Tsong-Li Wang, Dennis Shasha, George J. S. Chang, Liam Relihan, Kaizhong Zhang, and Girish Patel, Structural Matching and Discovery in Document Databases, In *Proceedings of the ACM SIGMOD*, 560-563, 1997, Tucson, Arizona, May.
[10] Kaizhong Zhang, Jason T. L. Wang, and Dennis Shasha, On the Editing Distance between Undirected Acyclic Graphs and Related Problems, *International Journal of Foundations of Computer Science, Special Issue on Computational Biology*, Vol.7, No.1, March 1996, pp.43-57.

2.6 おわりに

平成 14 年度には、既存のマルチモーダル対話データに関する映像転記や GDA に基づくアノテーションの作業、およびマルチモーダル対話コーパスのブラウジングとアノテーションを支援するソフトウェアの開発を進展させたことに加えて、マルチモーダル対話コンテンツに関する国内外の動向調査に本格的に着手した。後者は、e ラーニング等を含む、マルチモーダル対話コンテンツの産業応用の可能性に関する調査を含む。平成 15 年度には、エンドユーザ向けコンテンツと研究用コーパスとの融合を念頭に置きながら、マルチモーダル対話コンテンツ技術の発展と普及を図るため、次のような活動を行う予定である。

- 国内外におけるマルチモーダルコーパスの作成および利用の動向に関して調査する。特に、コーパスの間の相互運用性や、コーパス管理技術の共有等の可能性を探る。また、ISO/TC37/SC4 における言語資源の標準化作業の進捗状況を見ながら、その作業に対して提案を行いたい。
- 教育やゲームにおけるマルチモーダル対話コンテンツ技術の新たな応用の可能性に関して調査する。これまでの調査の範囲では、実際の応用は、視聴覚障害者向けの映画等のプレゼンテーション等に限定されていた。しかし、映画のようなコンテンツは、前以て定まった順序に従って視聴することを前提に作成されているため、要約や提示順の変更などの加工になじみにくく、マルチモーダル対話コンテンツ技術を応用する素材として必ずしも適切ではない。これに対し、教育やゲームのコンテンツは非線形性が強いので、応用のポテンシャルが高いと考えられる。たとえば e ラーニングの技術に関する議論はこれまでのところ配信のインフラに重点が置かれ、コンテンツそのものの知的な処理についてはほとんど検討されていないので、本委員会の活動に基づいて有意義な提言を行える可能性は高いと考えられる。
- これらの調査に基づいて、マルチモーダル対話コンテンツのさまざまな応用を意識したアノテーションの要件を検討し、これまで作成してきたマルチモーダル対話コーパスのアノテーションの方式を修正・拡張する。マルチモーダル対話コーパスのアクセスツールへの検索機能のプラグインは平成 14 年度にすでに完了しているが、平成 15 年度には教育等における情報提示に必要なアノテーションについて重点的に検討する必要があるだろう。
- マルチモーダル対話コーパスのアクセスツールを拡張して上記の応用のためのアノテーションをサポートするとともに、これを用いてコーパスを増補・修正する。アクセスツールに関しては、上記のような情報提示の応用を想定して。コーパスに関しては、元データを増やすのではなく、これまで扱ってきたデータに対して新たなアノテーションを施す、またはアノテーションを修正する作業に重点を置く。
- これらのコーパス、ソフトウェアツール、API 等を研究開発用に公開する。特に、GSK (言語資源共有機構)を通じて広く配布できるようにする。

３．言語資源専門委員会活動報告

３．言語資源専門委員会活動報告

3.1 はじめに

学習に基づく自然言語処理技術の研究の進展および実際のデータを対象と出来うる頑強なシステムの必要性から、自然言語処理研究用の言語資源に対するニーズはますます高まってきている。言語資源専門委員会では従来より、対訳コーパスを中心とする言語資源に関する調査研究を行ってきた。昨年度は、３つのワーキンググループを設置し、対訳コーパス、自然言語処理応用技術、言語資源およびイニシアティブについての調査を行った。

今年度、言語資源専門委員会では、昨年度の３ワーキンググループ体制から、1)言語ポータルグループ、2)言語処理応用グループ、3)著作権調査グループ、4)コーパスグループの４グループ体制で、広く言語資源に関わる調査を行った。

言語ポータルグループは、昨年度に行った言語資源およびイニシアティブに関する調査結果を基に、会議案内や新製品紹介、言語資源・イニシアティブのリスト、用語集などを含む、我が国初の網羅的な自然言語のポータルサイトを立ち上げた。会議案内や新製品紹介については、頻繁な更新により実用的なサイトとなることを目指した。言語資源等については網羅的なものを目指し、有益な情報をリストできたと自負している。

言語処理応用グループは、自然言語処理技術を利用したシステムに対するユーザニーズの把握のため、CEATEC2002の来場者に対するアンケート調査を行い、その結果を、いくつかの観点から分析した。今回は自然言語処理の応用を広範囲のシナリオで紹介し、どのようなニーズにユーザの興味があるかを調査した。このような調査により、既存の技術の枠を越えた、新しいニーズが発掘できるものと考えている。

著作権グループは、言語資源の利用に関する著作権上の問題点について、ユーザ側の有識者を講師としたヒアリング調査を行った。予め作成した標準質問に回答してもらうことにより、講師間の判断の異なりを明確化した。今年度のヒアリングは、言語資源の提供側ではなく、利用側を対象とするものになったが、それでも、各有識者の間で意見の違いが見られ、著作権に関する判断の難しさをうかがわせた。来年度は提供側の意見のヒアリングも行う予定である。

コーパスグループは、昨年度作成した日韓（英）コーパスへの対応付け作業の結果を検討した。また、日本語から英語へ、英語から日本語へという翻訳方向の違いが、対応付け作業にどのように影響するかについて、実作業により比較検討した。これらの知見により、今後は広くアジア圏言語を対象とできる仕様を作成し、標準化への提案を目指したいと考えている。

言語資源専門委員会では、これらの成果を踏まえて、来年度も引き続き、言語資源に関する調査研究を継続する予定である。

３．２ 言語情報処理ポータル

言語資源専門委員会では、平成13年度、「世界の言語イニシアティブ調査」活動を行い、世界の言語処理／言語資源に関する種々のプロジェクトを調査し、その結果をWWWドキュメントとしてまとめ、公開した。

平成14年度は、この活動をさらに発展させ、言語情報処理に関するさまざまな情報を集約した「言語情報処理ポータル」サイト、http://www.kc.t.u-tokyo.ac.jp/NLP_Portal/の構築を行った(図 ３．２-１)。このサイトに掲載している情報は以下のものである。

- 会議案内
- 製品ニュース
- 言語資源カタログ（アジア・日本（和文・英文））
- 世界の言語イニシアティブ（和文・英文）
- 関連学会・関連機関リンク集
- 用語集

このうち会議案内、製品ニュースについては、週に2, 3回のペースで更新を行っており、最新の情報を掲載する体制をとっている。このように言語情報処理に関する情報を総合的に提供しているサイトは国内ではこのページだけであり、当該分野および周辺分野の研究者、開発者、学生などから好評をえている。

以下、ポータルサイトの内容について紹介する。また、海外における言語処理ポータルサイトの現状についての調査を行ったので、3.2.7で報告する。

３．２．１ 会議案内

言語処理に関連する学会、研究会、国際会議の開催案内や論文募集の情報をまとめ、公開している。主な情報源は言語処理関係のメイリングリストである。速報性を重視し、週に 2,3 回の頻度で更新している。

掲載する情報は、会議の名称、開催日時、場所、論文募集の場合は投稿締切であり、リスト形式にまとめている。また、会議の詳細な情報が容易に得られるように会議の Web ページへリンクをはっている。

図 ３．２-１ 言語情報処理ポータルサイト

３．２．２ 製品ニュース

自然言語処理に関する最新の製品ニュースを収集し、ヘッドラインのリストと元の記事へのハイパーリンクを作成して公開している（図 ３．２．２－１）。製品ニュース記事の収集にあたっては、各メーカーやベンダーがリリースしている製品ニュースの中から自然言語処理関連のものを人手で選別している。ニュースの収集は 2002 年８月から開始され、1 月 28 日現在までで８５件のニュースが集まっている。これらは、カテゴリ別の「電子辞書」、「文書管理ソフト」、「検索システム」、「ナレッジマ

ネジメント」、「機械翻訳」、「音声認識・合成」、「マルチメディア」の７分野および「その他」に分類されている。

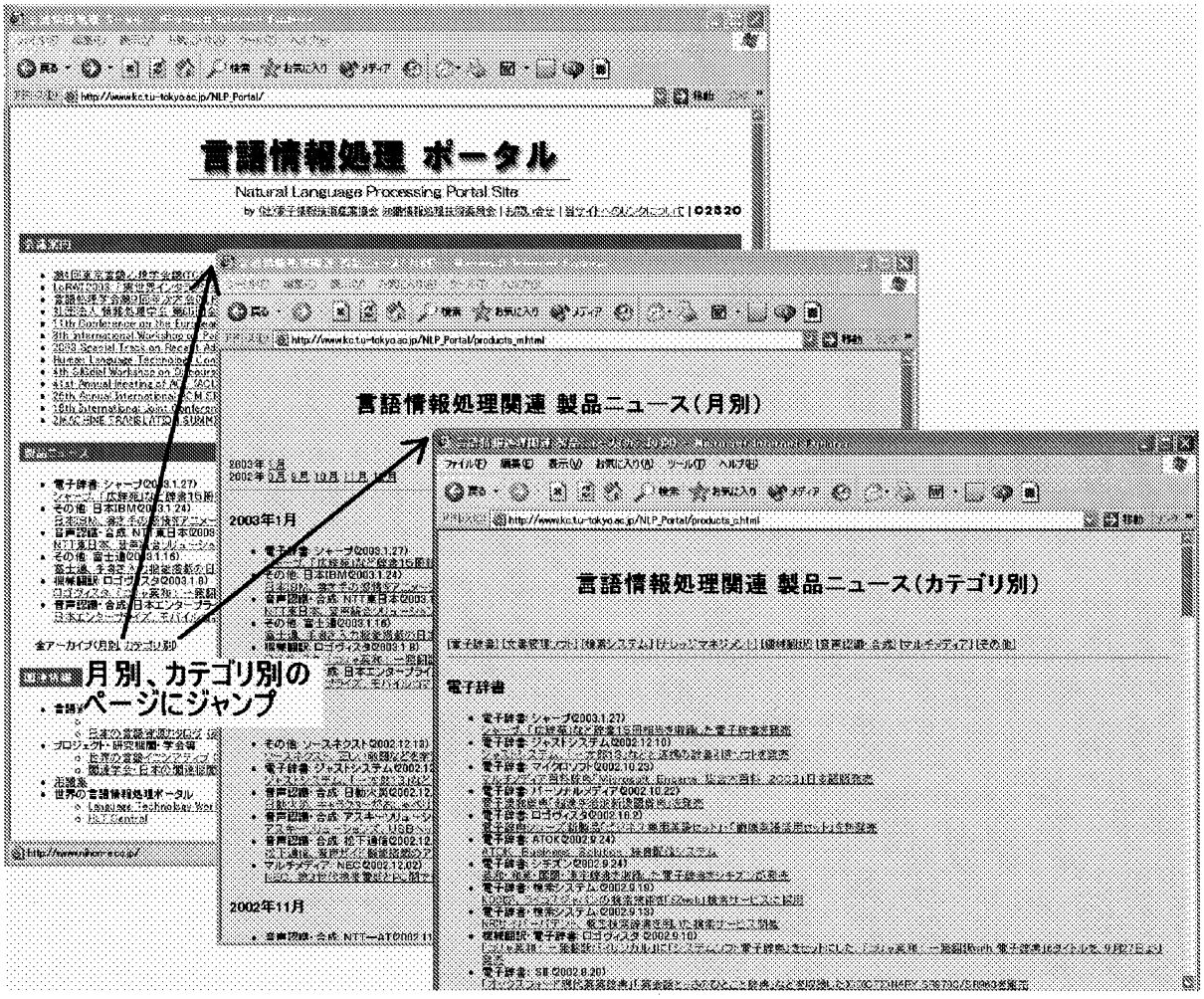
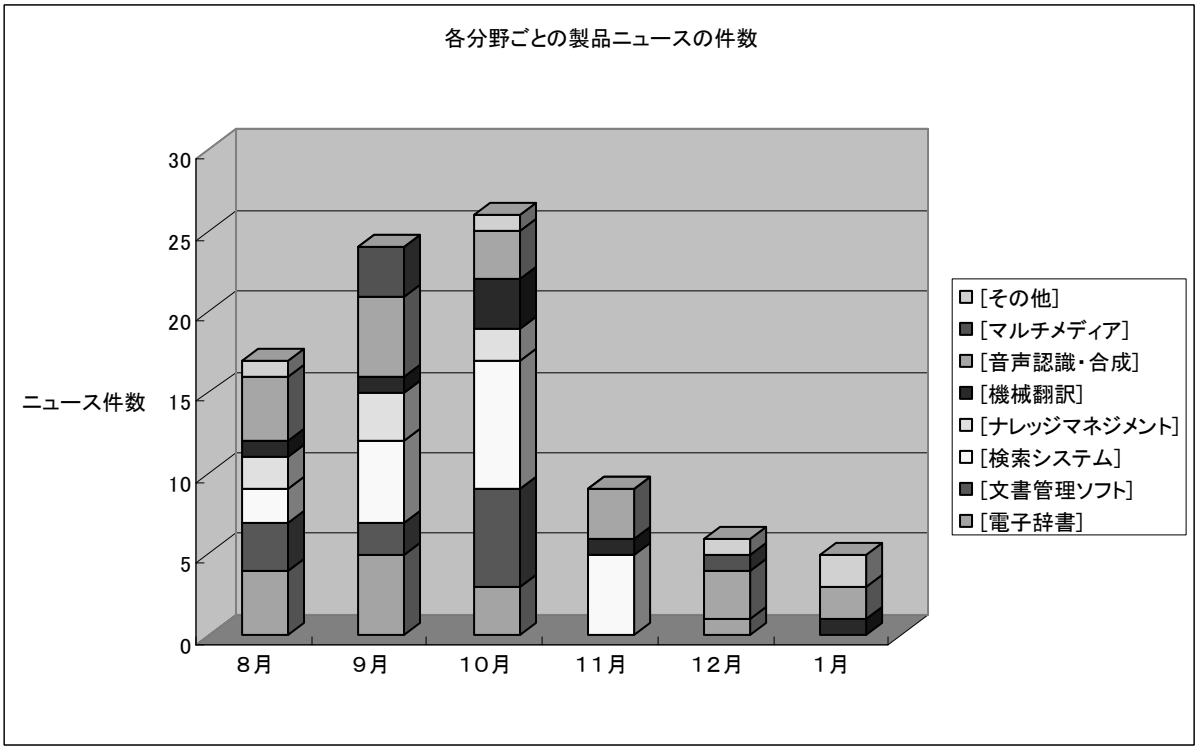


図 ３．２．２－１　製品ニュースのページ

国内で今のところ自然言語関連のニュースを専門に収集しているサイトは存在しない。個別のニュース記事は各関連メーカーやベンダーのニュースリリースや、ＩＴ、コンピュータ関連のニュースサイトに分散している。現状では、最新ニュースをウォッチしたい場合、利用者はこれらの有力な複数のサイトから関連ニュースだけを拾い読みしなければならないが、本ポータルサイトのような最新の関連ニュースを収集提示するサイトが継続的に運用され、情報が蓄積されてくれば、研究者、開発者、製品ユーザなどそれぞれの立場の利用者にとって有用な情報源として利用されるようになることを期待している。

実際に製品ニュースで掲示している製品ニュース 85 件（2002 年 8 月～2003 年 1 月の記事）を分野と月ごとに整理したものを表 ３．２．２－１に示す。

表 ３．２．２－１　製品ニュース８５件の分類



表は各月の製品ニュース件数のヒストグラムを示しており、さらに分野別に色分けによって示している。6 ヶ月で平均すると、毎月ほぼ 10 件以上のニュースがあり、特定の分野に集中するというよりはそれぞれの分野に分散しているように見える。現在、ポータルサイトが立ち上げ完了して間もない段階であり、関連ニュースの選別基準や網羅性などについては今後の検討課題である。今後も最新の関連ニュースを継続的に提供することで、利用者に速報性の高い情報を提供することが主眼であるが、バックナンバーとして提供される一定期間のデータから言語処理関連の分野における技術トレンドや市場規模の変化を読み取るなど、データベースとしての利用も期待される。

３．２．３ 言語資源カタログ

辞書・コーパス・ツールなどの言語資源は、自然言語処理や言語学の分野の研究者にとって貴重な知識源である。そこで、日本国内の言語資源のリストを作成し、言語資源カタログとしてポータルサイトに掲載した。現在、テキストデータ 51 件、音声データ 24 件、ツール 18 件、合計 93 件の言語資源が掲載されている。また、日本語版と英語版のページを用意し、利用者の便宜を図った。日本語版のカタログを 3.2.8 に付録として掲載する。

カタログに記載されている主な項目は以下の通りである。

- Name

  言語資源の名前。

- Type

言語資源の種類。Text(テキストデータ)、Sound(音声データ)、Software(ツール)など。

- Type.linguistics
  言語学的観点から見た言語資源の分類。コーパス、辞書、シソーラスなど。
- Type.functionality
  ツールの機能。形態素解析ツール、構文解析ツールなど。
- Date
  言語資源が作成された日時。
- Creator
  言語資源の作成者
- Contributor
  言語資源の作成に貢献した人。資金提供者など。
- Subject.language
  言語資源が対象としている言語。
- Language
  言語資源の記述言語。
- Description
  言語資源の説明、解説。
- Format
  言語資源の電子フォーマットに関する情報。記録媒体(CD-ROM, テープ)、ファイルサイズ、エントリ数、サンプリング周波数など。
- Format.encoding
  言語資源の文字コード。EUC-JP、ISO-2022-JP、Shift_JIS など。
- Format.markup
  言語資源のマークアップスキーム。SGML、XML など。
- Format.os
  ツールを使用するために必要な OS。
- Format.sourcecode
  ツールのソースコード。
- Relation
  他の言語資源との関連に関する情報。IsPartOf (言語資源 A は言語資源 B の一部である), HasPart(A は B の部分として持つ), Required(A を使うためには B が必要である)などの関係があらかじめ定義されている。
- * Contact person
  言語資源の入手に関する連絡先。
- * URI

(主に作成者が公開している)言語資源の Web ページや FTP サイトの URI。

- * Price
  言語資源の価格。
- * Annotation.corpus
  コーパスの文に付与された付加情報。品詞、構文情報、語義など。
- * Annotation.document
  コーパスの文書に付与された付加情報。キーワード、文書カテゴリなど。

これらの項目は Open Language Archive Community (OLAC)のメタデータセットに準拠している。詳しくは OLAC のサイト http://www.language-archives.org/OLAC/olacms.html を参照していただきたい。特に注意すべきなのは、Subject.Language と Language の違いである。OLAC では、言語資源が対象とする言語を Subject.language、知識を記述する言語を Language として両者を区別する。例えば、英和辞典の場合、これは英単語に関する知識なので Suject.language は英語となり、英単語に関する記述は日本語で書かれるため Language は日本語となる。Creator と Contributor も互いに似ている項目である。前者が言語資源の作成に直接的に関わり、言語資源の管理や配布等にも責任を持つ人や組織であるのに対し、後者は資金の提供など言語資源の作成に間接的に関わった人や組織である。また、OLAC のメタデータセットにない独自の項目もいくつか追加した。上記のリストで * のついている項目がポータルサイト独自の項目である。

また、言語資源に関する情報として、AFNLP で調査・公開しているアジアの言語資源カタログ(http://tanaka-www.cs.titech.ac.jp/ALR/search.html)へリンクを貼っている。

３．２．４ 自然言語処理用語集

自然言語処理に関する専門用語を集め、短い説明を付してリスト形式で提供している。以下の手順で作成した。

1. James Allen による教科書[1]の索引を用語集の一次情報源とした。索引には約 745 語ある。なお、これらの用語はすべて英語である。
2. 1.の索引を辞書順に 4 分割し、4 名の委員に割り当てた。
3. 各委員は各自に割り当てられた用語リストより、10 から 20 個を目安に用語を選んだ。用語の選択基準は特別設けなかった。次に選んだ用語を日本語に訳し、数文程度からなる説明を書いた。図 ３．２．４-１に用語の説明例を挙げる。
4. 東京大学黒橋研究室の学生(博士後期課程)に、3.までで漏れている重要用語(とその説明)を補完してもらった。

以上により現在約 100 個の用語が提供されている。2003 年 1 月現在での全用語を図 ３．２．４-２に示す。

図 ３．２．４-２を見てもわかるように、本用語集は自然言語処理の分野全般を網羅しているとは言いがたい。一つの教科書の索引を情報源としていることが一つの原因である。そこで用語の網羅性

を増すために、閲覧者からの情報を収集する仕組みを設けた。具体的には、用語集に加えて欲しい用語を閲覧者が入力できるようなCGIインターフェイスを組み込んだ。参考までに、2003年1月現在までにサイト閲覧者が入力した用語の中で自然言語処理に関連しているものを一部図 3．2．4-3に示す。用語収集に加えて、閲覧者自身が用語の説明を入力できるようなインターフェイスも用意したが(どの用語に関する説明でも構わない)、残念ながらこれまでに用語説明に関する情報提供はない。

今後は、閲覧者からの用語収集に加え、文献から網羅的に用語を収集することも検討している。近年では、多くの学会誌論文、研究会報告書等が電子化されている。また、個人が Web を介して論文を電子的に公開しているケースも多い。このような専門分野の文献集合から(半)自動的に用語を抽出する試みは興味深い。この手法には以下のような利点がある。

- 網羅性がある。
- 速報性がある。教科書には載らないような新規の用語が即座に発見できる。
- 用語のみならず用語の説明も半自動的に収集できる可能性がある。ただし、自動収集した説明文まで公開する際は、著作権問題を明確にする必要がある.

次年度は、以上を踏まえて用語集を拡充していくこと、もしくはそのプロセスを確立することを予定している。

あ行

アーリーアルゴリズム（Earley algorithm）
文脈自由文法に基づく構文解析アルゴリズム．ある非終端記号の直後に現われ得る終端記号を事前に予測することによって解析効率を改善している点が特徴.

曖昧性（ambiguity）
自然言語処理では，複数の解析結果が得られることを曖昧性があるという．例えば複数の語義がある場合は語義(選択)に曖昧性があるといい，かかり受け解析において複数の可能性がある場合は，かかり受けに曖昧性があるという．曖昧性は様々な処理レベルで存在し，曖昧性解消(ambiguity resolution, disambiguation)は自然言語処理の真髄とも言われる.

図 3．2．4-1　用語の説明例

アーリーアルゴリズム | 曖昧性 | 曖昧性解消 | アジェンダ | 後入れ先出し | アブダクション | アラインメント | EAGLES | 意味素 | 意味ネットワーク | 意味マーカ | 意味論 | ヴィタビアルゴリズム | 後向き確率 | EDR 辞書 | SGML | XML | n-グラム | LR(1)構文解析法 | 音素 | オントロジー | 格 | 拡張遷移ネットワーク | 格文法 | 確率変数 | 隠れマルコフモデル | 下降型構文解析 | 括弧付きコーパス | ガーテンパス文 | 含意 | 帰納推論 | 共起 | 共参照 | 訓練セット | 形態素解析 | 言語オントロジー | コーパス | コーパスからの知識獲得 | 恒真式 | 構文解析 | 固有名詞 | 語用論 | CYC | 最尤推定法 | CES | C統御 | システミック文法 | 自然言語理解 | シソーラス | 質問応答システム | 修辞構造理論 | 終端記号 | 主辞 | 主辞駆動句構造文法 | 照応 | 深層構造 | 信念 | 情報検索 | 情報抽出 | スクリプト | スロット | 性 | 正規文法 | 選択制限 | 全称限量子 | 相互情報量 | 属性 | 属性値 | 対訳コーパス | 対訳辞書 | 多義性解消 | タグ(tree adjoining grammer) | タグ(tag) | タグ付きコーパス | 談話 | 談話構造 | 談話セグメント | 知識表現 | 知識ベース | チャートパージング | チョムスキーの階層 | ツリーバンク | TEI | 手がかり句 | 電子化辞書 | 統語論 | 統率・束縛理論 | トライグラム | トライ構造 | 人称 | ノード | 発語内行為 | 2-グラム | バックトラック | 評価セット | 品詞タグ付け | BDI モデル | ピッチ | 素性 | 文法 | 文脈依存文法 | 文脈自由文法 | プロローグ | ベイズ規則 | 補部 | 翻訳 | 前向き確率 | 未知語 | EuroWordNet | 様相 | 要約 | ラムダ計算 | リーフ | リスプ | 論理形式 | WordNet

図 3．2．4-2　用語集に含まれる用語(2003年1月現在)

機械翻訳 ｜ ベクトルモデル ｜ Okapi ｜ 正例 ｜ 空範疇 ｜ サンプリング間隔 ｜ ギップの現象 ｜ filler ｜ 標本化 ｜ query ｜ データ分析 ｜ インクリメンタル ｜ SVM ｜ 正規言語 ｜ 関数型プログラミング ｜ ワード ｜ 命題 ｜ modality ｜ Boehm の推移曲線 ｜ ダイナミックプログラミング ｜ Xバー理論 ｜ 線形有界オートマトン ｜ 共起 ｜ トレリス ｜ 意味表現 ｜ 構造化プログラミング ｜ モディフィケーション ｜ 枝分かれ ｜ 文脈 ｜ オペランド ｜ 知識獲得 ｜ ツリー ｜ 状態図 ｜ 袋小路文 ｜ 平文 ｜ dag ｜ フィードバック ｜ 生成文法 ｜ 代名詞

図 ３．２．４-３　ユーザが入力した用語(一部)

３．２．５ 世界の言語イニシアティブ(英文)

昨年度調査報告した世界の言語イニシアティブを英文に翻訳し掲載した。

３．２．６ 関連学会、関連機関へのリンク集

言語処理に関連する学会、研究機関へのリンク集である。学会については、国内・国外あわせて 33 の学会のホームページへのリンクを掲載した。大学の研究機関については、18 の研究室のホームページへのリンクを掲載した。その他、4 つの研究機関と 2 つの協会のホームページへのリンクを掲載した。

３．２．７ 海外の言語処理ポータルサイト

ここでは、海外、とくにヨーロッパにおける言語処理ポータルサイトの現状について紹介する。
以下で、"HLT"とは、"Human Language Technology"の略である。

（1）HLTCentral

HLT および関連トピックスに関するオンライン情報源として、EU の EUROMAP プロジェクトと ELSNET プロジェクトの協力により開設された。前身は LANGLINK プロジェクト。
特にヨーロッパを中心とした、音声処理、テキスト処理、多言語自動翻訳、などの分野におけるニュース、研究開発情報、ビジネス情報を提供する。

表 ３．２．７-1 HLTCentral について

URL	http://www.hltcentral.org/
プロジェクト	EUROMAP および ELSNET
出資母体	EU（EUROMAP および ELSNET）
サイト管理	VDI/VDE-Technologiezentrum Informationstechnik GmbH 社（ドイツの NFP）および Arax Ltd. 社（英国）

以下に、サイト構成を簡単に紹介する。

表 ３．２．７-2　HLTCentral のサイト構成

セクション	内容
Euromap	HLT 技術の実用化促進を目指す EUROMAP プロジェクトのホームページへのリンク
Elsnet	EU における HLT 研究・開発者間の技術交流促進を目指す ELSNET プロジェクトのホームページへのリンク
eContent	ヨーロッパのデジタルコンテンツのグローバルネットワーク上の利用促進を目指す eContent プロジェクトのホームページへのリンク
Events	HLT 関連の会議などのイベントに関する情報、関連 URL。19 件。 過去のイベントに関するアーカイブ（97 件)、うち重要なイベント 3 件を別表示。
News	今年の HLT 関連のニュース（へのリンク)、2000 年以降のアーカイブへのリンク。2002 年の 1 年の蓄積件数は 1,742 件。
What's new	最新のニュース、更新情報
Projects	EC による HLT 関連の 242 プロジェクトに関するデータベース。アルファベット順、年代順のリストあり。うち、現在活動中の 112 プロジェクトについては、カテゴリ別のリストもある。各プロジェクトについては、実施期間、簡単な説明、成果物、連絡先、参加組織、ホームページの URL が記述されている。略称、言語、枠組みプログラム、キーワードによる検索もできる。
Calls　for Proposals	各種 Call For Proposals 現在 2 件、過去分 14 件。応募のための FAQ あり。
Links	HLT 関連のサイトへのリンク集。Web ベースツール、商品紹介、専門組織、言語資源、その他 HLT 関連、EC および政府機関、ニュース・雑誌、パートナー検索、多言語関係、技術移転、ニュースグループおよびメーリングリスト、その他、の 12 分野、22 種類の、のべ 156 件のリンク。
HLT　Overview	EU の HLT に関する概説。新しい EU の枠組みプログラム(FP6)に関する、Call For Proposal などの情報、現行の FP5、過去のプロジェクトに関する資料、市場展望、EC の HLT チーム紹介も。
Repository	HLT 関連のレポート、サーベイ、書籍などのアーカイブ。8 カテゴリ、
Contacts　Database	HLT 関連の研究者の氏名、所属、ホームページ、国のリスト。現在 309 名の登録あり。登録可能。
About HLTCentral	HLTCentral 自身に関する説明および、関連プロジェクト(EUROMAP, ELSNET)へのリンクなど。

数字は 2003 年 1 月 29 日現在の数字。

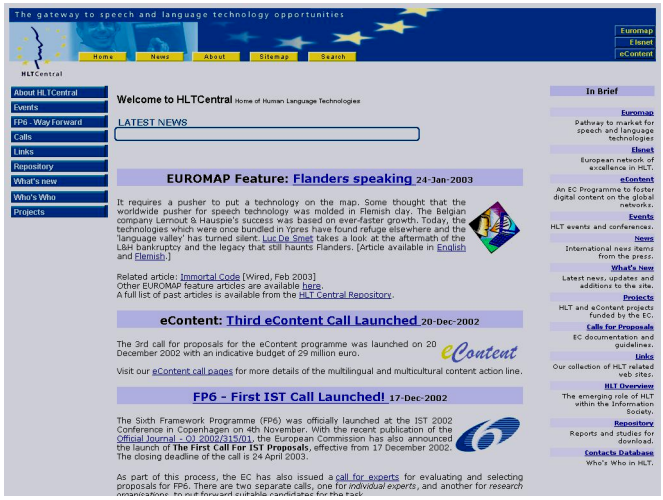


図 3．2．7-1 HLTCentral のトップページ

(2) EUROMAP

ヨーロッパにおける HLT 関連研究開発プロジェクトの成果の、実用のための市場への技術移転促進を主な目的とした、IST プログラムにおける HLT 関連プロジェクトの一つ。プロジェクトは、現在、英、仏、独、伊、西、オーストリア、ベルギー／オランダ、ブルガリア、デンマーク、フィンランド、ギリシャ、の 11 の NFPs（National Focal Points）によって構成されており、各 NFP の活動を基盤に、国際的な活動に高めて行くことを目指す。 ELSNET プロジェクトと共に、 HLTCentral（ http://www.hltcentral.org/ ）というウェブサイトを運営しており、同サイトの中に独自のページを持っている。

表 3．2．7-3 EUROMAP について

URL	http://www.hltcentral.org/euromap/
プロジェクト	EUROMAP
出資母体	EU
サイト管理	VDI/VDE-IT （ドイツの NFP)、および Arax Ltd.社（英国）

以下に、サイト構成を簡単に紹介する。

表 3．2．7-4 EUROMAP のサイト構成

セクション	内容
Articles	EUROMAP のために寄稿された記事 14 記事
Mission	EUROMAP の果たすべき役割について
Team	プロジェクトを構成する各国の NFPs(National Focal Points)へのリンク

Events	HLT 関係のイベントのリスト。各イベントについて、日付、主催、種類、名称、開催都市、ホームページへのリンクを掲載。最新分 4 件と、2001 年からの過去分 50 件。
Success Stories	HLT のさまざまな応用事例に関する紹介記事、36 記事。５つのカテゴリ別、日付順に参照可能。オンラインでの要約版のほかに、フルサイズの報告書もダウンロードできる。各記事は関連する HLT プロジェクトや組織へのリンクつき。
Newsletters	毎月発行される Euromap Language Technologies Newsletter の最新号と、アーカイブ。同 Newsletter には、特集、成功事例紹介、ニュース記事へのリンク、イベント案内、プロジェクト公募情報などが掲載されている。
eBusiness	eBuisiness に関する調査タスクフォースの最終報告（2002 年 5 月）と、関連プロジェクト、機関等へのリンク
Tourism	旅行業務への HLT の応用に関する調査タスクフォースの最終報告書（2002 年 6 月）と、関連プロジェクト、機関、イベント等へのリンク
TechTransfer	HLT 関連の技術移転に関する調査タスクフォースの最終報告書（2002 年 4 月）と、関連プロジェクト、機関、イベント等へのリンク
Reports	EUROMAP の年次報告書(2000, 2001)、EUROMAP が関連したイベントのレポートなど。
Who's who	HLTCentral の Who's who (Contacts Database)へのリンク。HLT 関連の研究者の氏名、所属、ホームページ、国のリスト。現在 309 名の登録あり。登録可能。
HLT Projects	HLTCentral の Projects へのリンク。EC による HLT 関連の 242 プロジェクトに関するデータベース。アルファベット順、年代順のリストあり。うち、現在活動中の 112 プロジェクトについては、カテゴリ別のリストもある。各プロジェクトについては、実施期間、簡単な説明、成果物、連絡先、参加組織、ホームページの URL が記述されている。略称、言語、枠組みプログラム、キーワードによる検索もできる。

数字は 2003 年 1 月 29 日現在の数字。

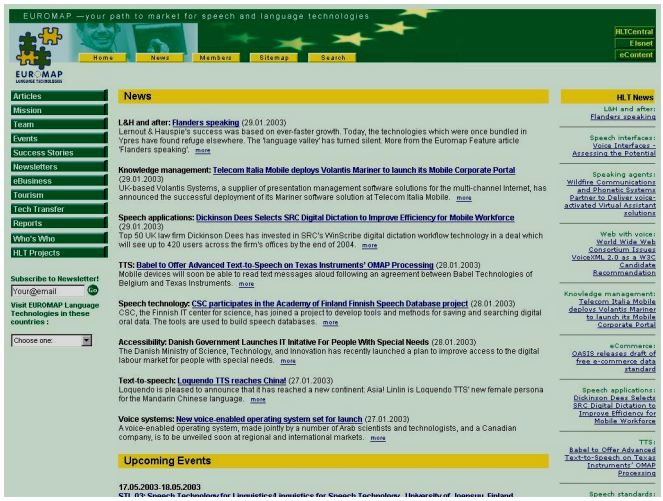


図 3．2．7-2 EUROMAP のトップページ

(3) ELSNET

ヨーロッパにおける広義の HLT 促進を目的とし、言語・音声技術および関連分野の研究・開発・応用に携わる人々の交流を促進するネットワークとして、各種セミナー、ワークショップの開催や、ウェブサイト、メイリングリストの運営を行う。実験用の言語資源の構築・配布も目的の一つ。1991 年に ESPRIT プログラムの下で設立された。現在、約 20 ある IST の Network of Excellence の 一つである。ヨーロッパの 26 ヶ国をカバーしている。メンバーは、言語および音声処理技術の開発・利用を目的とする、公的あるいは民間 の研究機関、企業であり、約 135 のメンバーのうち、60%が大学などの学術的機関であり、40%が産業界からの参加となっている。EUROMAP, eContent プロジェクトと共に HTLCentral というポータルサイトを運営しているが、独自のポータルサイト（下記）も持っている。

表 3．2．7-5 ELSNET について

URL	http://www.elsnet.org/
プロジェクト	ELSNET
出資母体	EU
管理	the Utrecht Institute of Linguistics OTS (ユトレヒト大学, オランダ)

以下に、サイト構成を簡単に紹介する。

表 ３．２．７-6　ELSNET のサイト構成

セクション	内容
Directory of language and speech technology organizations	54 カ国の 2545 の HLT 関連組織について、国名、名称、所在地、説明、連絡先、組織種別、URL のリスト。
Directory of language and speech technology experts	56 カ国、986 名の HLT 関連研究者の、名前、所属組織、連絡先、URL，言語、専門分野、簡単な紹介。
On-line presentations and tutorials in language and speech	HLT に関するオンラインチュートリアルへのリンク。3 カテゴリ８コンテンツが登録されている。
Topics, SIGs and Sites	HLT に関するサーベイやトピックを扱っている、13 カテゴリの 21 のサイトへのリンク。
Paper and Electronic Publications	本、（電子）雑誌、論文、リポートなどの名称、URL、説明、コメントのリスト。9 カテゴリ１６リンク。
Products and Companies in Natural Language and Speech Processing	HLT 関連（商用／非商用）製品、会社の、名称、URL、簡単な説明、コメントのリスト。１７カテゴリ、130 リンク。
Tools for Natural Language and Speech Processing	HLT 関連のツールへのリンク。11 カテゴリ、１４リンク。
Language and Speech Resources	言語資源、関連機関へのリンク。7 カテゴリ、12 リンク。
Newspapers on the Web	Web 上の通常の新聞サイトへ（またはリンク集）へのリンク。19 カテゴリ（地域)、３２リンク。
International Organisations, Infrastructures and Events	HLT 関連の国際組織、学会、会議、メーリングリストなどへのリンク。12 カテゴリ、43 リンク。
National and Regional Organisations, Infrastructures and Events	HLT 関連の各国の組織、学会、会議、メーリングリストなどへのリンク。14 カテゴリ、34 リンク。
EU Funded Projects in Language and Speech Processing	HLTCentral による、EU の HLT プロジェクトのデータベース。
elsnet-list@elsnet.org	ELSNET のメーリングリスト（誰でも参加可能）に関する情報と、アーカイブ。
The Calendar of Events and Deadlines	会議、ワークショップ、サマースクールなど各種イベントの名称、開催日時、カテゴリ、URL のリスト。締切順のリストもあり。
Jobs and CVs	HLT 関連の人材募集広告のアーカイブ。また、個人が求職のため

	の履歴書を登録することもできる。

数字は 2003 年 1 月 29 日現在の数字。

各種情報へのリンクは検索可能。また、オンラインによる（第三者による）登録、変更可能。ただし、組織・個人のリスト以外に関しては、6 ヶ月に登録延長手続きが必要。

図 ３．２．７-3　ELSNET のトップページ

(4) Language Technology World

ドイツ人工知能研究センター(DFKI)により運営されるウェブベースの仮想的な情報センターで、研究開発コミュニティ、言語技術の潜在的なユーザ、学生ほか HLT に関心のある人々向けに、HLT に関する広範囲な技術に関する情報を提供する。

表 ３．２．７-7 LanguageTechnology World について

URL	http://www.lt-world.org/
プロジェクト	COLLATE
出資母体	German Ministry for Education and Research (BMBF)
管理	German Research Center for Artificial Intelligence (DFKI)

以下に、サイト構成を簡単に示す。

表 ３．２．７-8 Language Technology World のサイト構成

セクション	内容
LT WORLD Breaking News	短い本文つきのニュース（3 記事） と、ニュースタイトル（15 記事）。本文は、さまざまなニュースサイト上にある。

Information & Knowledge	Technologies	HLT に関するサーベイ"Language Technology – A Survey of the State of the Art (second edition)" の構成に沿って、17分野 113 の HLT 要素技術に対する、定義、製品、プロジェクト、キープレイヤー、研究組織、参考 URL、論文の紹介。上記サーベイの第 1 版の対応する節（pdf ファイル）も閲覧可能。第 2 版は 2002 年 web 公開、2003 年出版の予定。
	Abbreviations	HLT での主要な略語（251 語）に対する、1 行の簡単な説明。
Players & Teams	Players	HLT 関連研究者（2,204 名）の、名前、ホームページの URL、所属組織、関連プロジェクト。検索・登録・変更可能。
	Projects	144 の HLT 関連技術の、のべ 1,015 のプロジェクトの、名称、テーマ、URL、組織名、責任者名、応用、言語などのリスト。検索・登録・変更可能。
	Organizations	のべ 1,446 の企業、大学／研究機関、学会／ネットワーク、ファンド・プログラムの、国別の、名称、URL、責任者名、連絡窓口、所在地のリスト。検索・登録・変更可能。
Systems & Resources	R&D-Systems	8 分野 64 テーマの、のべ 399 の研究レベルのシステムの、テーマ別の、名称、URL、作者、組織、言語、動作環境、簡単な説明のリスト。検索・登録・変更可能。
	Products	8 分野 66 テーマの、のべ 444 の商品レベルのシステムの、テーマ別の、名称、URL、作者、組織、言語、動作環境、簡単な説明のリスト。検索・登録・変更可能。
	Repositories	メタデータに関するイニシアティブ、言語資源およびツールのアーカイブ、オントロジ関連のアーカイブ、言語データ流通促進機関の、のべ 11 の組織のホームページへのリンク。検索可能。
Communication & Events	News	HLT 関連のニュース記事が各記事の見出し、1～2 行の抜粋と、オリジナルのサイトの本文へのリンク。2002 年については、161 記事が掲載されている。検索可能。
	Conferences	HLT 関連の会議の名称、ホームページの URL、開催日時、主催者、開催場所など。2003 年 2 月から 2004 年 10 月までの６９会議が登録されている。
About LT World		LT World 自身に関する説明

数字は 2003 年 1 月 29 日現在の数字。

図 ３．２．７-4 Language Technology World のトップページ

３．２．８ 付録: 日本の言語資源カタログ

テキスト

・ 分類語彙表

Type	Text
Type.linguistics	lexicon/thesaurus
Description	32,600 語の日本語単語を分類したシソーラス。これらの単語は、4 つの類, 12 のセクション, 798 のグループに分類されている。５０音順に並べられた索引もある。
Creator	国立国語研究所
Contact person	柏野和佳子（waka@kokken.go.jp）
Price	フリー
Subject.language	日本語
Format	3.8 MB.
Format.encoding	Shift_JIS

・ RWC テキストデータベース

Type	Collection
Description	RWCP によって作成されたテキストデータベースのセット。
Creator	Real World Computing Partnership
Contact person	メディアドライブ（ txrwcdb-req@mediadrive.co.jp）
Price	2,000 円
Subject.language	日本語
Language	日本語
Date	1998
Format	381 MB.
Format.encoding	EUC-JP
Relation	HasPart RWC-DB-TEXT-94-1, HasPart RWC-DB-TEXT-94-2, HasPart RWC-DB-TEXT-95-2, HasPart RWC-DB-TEXT-95-3, HasPart RWC-DB-TEXT-96-2, HasPart RWC-DB-TEXT-97-1, HasPart CRL-DB-TEXT-97-1
URI	http://www.rwcp.or.jp/wswg/rwcdb/text/

・ RWC-DB-TEXT-94-1

Type	Text
Type.linguistics	annotation/corpus
Description	通産省の 1993 年から 1995 年の白書を形態素解析したコーパス。人手修正済。
Annotation.corpus	word segmentation, part-of-speech
Creator	Real World Computing Partnership
Subject.language	日本語
Language	日本語
Date	1994
Format	8.1 MB.
Format.encoding	EUC-JP
Relation	IsPartOf RWC テキストデータベース
URI	http://www.rwcp.or.jp/wswg/rwcdb/text/

・ RWC-DB-TEXT-94-2

Type	Text
Type.linguistics	annotation/corpus
Description	日本電子工業振興協会の「自然言語処理の動向に関する調査報告書」を形態素解析したコーパス。人手修正済。
Annotation.corpus	word segmentation, part-of-speech
Creator	Real World Computing Partnership
Subject.language	日本語
Language	日本語
Date	1994
Format	2.1 MB.
Format.encoding	EUC-JP
Relation	IsPartOf RWC テキストデータベース
URI	http://www.rwcp.or.jp/wswg/rwcdb/text/

・ RWC-DB-TEXT-95-2

Type	Text
Type.linguistics	annotation/corpus
Description	毎日新聞の 1994 年の 3000 記事を形態素解析したコーパスの差分データ。人手修正済。
Annotation.corpus	word segmentation, part-of-speech
Creator	Real World Computing Partnership
Subject.language	日本語
Date	1995
Format	1.9 MB.
Relation	IsPartOf RWC テキストデータベース, Requires 毎日新聞 CD-ROM (1994 年)
URI	http://www.rwcp.or.jp/wswg/rwcdb/text/

・ RWC-DB-TEXT-95-3

Type	Text
Type.linguistics	annotation/text categorization
Description	毎日新聞の 1994 年の 30000 記事に対して UDC コードを付与したデータ。
Annotation.document	text category
Creator	Real World Computing Partnership
Subject.language	日本語
Date	1995
Format	1 MB.
Relation	IsPartOf RWC テキストデータベース, Requires 毎日新聞 CD-ROM (1994 年)
URI	http://www.rwcp.or.jp/wswg/rwcdb/text/

・ RWC-DB-TEXT-96-2

Type	Text
Type.linguistics	annotation/corpus
Description	岩波国語辞典(第5版, タグ付き)を形態素解析したデータ。人手修正済。
Annotation.corpus	word segmentation, part-of-speech
Creator	Real World Computing Partnership
Subject.language	日本語
Language	日本語
Date	1996
Format	40.6 MB.
Format.encoding	EUC-JP
Relation	IsPartOf RWC テキストデータベース
URI	http://www.rwcp.or.jp/wswg/rwcdb/text/

・ RWC-DB-TEXT-97-1

Type	Text
Type.linguistics	annotation/corpus
Description	毎日新聞の 1991 年から 1995 年の全記事を自動的に形態素解析したコーパスの差分データ。
Annotation.corpus	word segmentation, part-of-speech
Creator	Real World Computing Partnership
Subject.language	日本語
Date	1997
Rights	research purpose
Format	280.5 MB.
Relation	IsPartOf RWC テキストデータベース, Requires 毎日新聞 CD-ROM (1991 年), Requires 毎日新聞 CD-ROM (1992 年), Requires 毎日新聞CD-ROM (1993年), Requires 毎日新聞CD-ROM (1994年), Requires 毎日新聞CD-ROM (1995 年)
URI	http://www.rwcp.or.jp/wswg/rwcdb/text/

・ CRL-DB-TEXT-97-1

Type	Text
Type.linguistics	annotation/corpus
Description	RWC-DB-TEXT-95-2 のテキストを単文に分割し、係り受け関係を解析したデータ。人手修正済。
Annotation.corpus	syntax
Creator	通信総合研究所
Subject.language	日本語
Language	日本語
Date	1997
Source	jp:rwc95-2
Format	40 MB.
Format.encoding	EUC-JP
Relation	IsPartOf RWC テキストデータベース
URI	http://www.rwcp.or.jp/wswg/rwcdb/text/

・ IPAL 辞書

Type	Text
Type.linguistics	lexicon/subcategorization dictionary
Description	日本語の基本動詞 861, 基本形容詞 136, 基本名詞 1081 語を収録した辞書。語の意味、形態素情報、文法カテゴリ、格フレーム、イディオムなどの情報が記載されている。
Creator	情報処理振興事業協会 (IPA)
Contact person	情報処理振興事業協会 (ipal-info@ipa.go.jp)
Price	フリー
Subject.language	日本語
Language	日本語
Date	1998
Format	11 MB.
Format.encoding	EUC-JP
URI	http://www.ipa.go.jp/STC/NIHONGO/IPAL/ipal.html

・ EDR 辞書

Type	Collection
Description	単語辞書、対訳辞書、概念辞書、共起辞書、専門用語辞書の 5 つからなる電子化辞書。また、共起辞書の付録として EDR コーパスがある。
Creator	日本電子化辞書研究所
Contact person	日本電子化辞書研究所 (thoth@edr.co.jp)
Subject.language	日本語, 英語
Language	日本語, 英語
Format	10 CD-ROM.
Relation	HasPart EDR 日本語単語辞書, HasPart EDR 英語単語辞書, HasPart EDR 日英対訳辞書, HasPart EDR 英日対訳辞書, HasPart EDR 概念辞書, HasPart EDR 日本語共起辞書, HasPart EDR 英語共起辞書, HasPart EDR 専門用語辞書
URI	http://www.iijnet.or.jp/edr/J_index.html

・ EDR 日本語単語辞書

Type	Text
Type.linguistics	lexicon/
Description	単語の意味(概念)や文法属性を記載した辞書。約 260,000 の日本語単語を収録している。
Creator	日本電子化辞書研究所
Contact person	日本電子化辞書研究所 (thoth@edr.co.jp)
Price	150,000 円 (アカデミック), 1,500,000 円 (一般)
Subject.language	日本語
Language	日本語, 英語
Format	103 MB. 260,000 entries.
Format.encoding	EUC-JP
Relation	IsPartOf EDR 辞書
URI	http://www.iijnet.or.jp/edr/J_index.html

・ EDR 英語単語辞書

Type	Text
Type.linguistics	lexicon/
Description	単語の意味(概念)や文法属性を記載した辞書。約 190,000 の英単語を収録している。
Creator	日本電子化辞書研究所
Contact person	日本電子化辞書研究所 (thoth@edr.co.jp)
Price	150,000 円 (アカデミック), 1,300,000 円 (一般)
Subject.language	英語
Language	英語, 日本語
Format	86 MB. 190,000 entries.
Format.encoding	EUC-JP
Relation	IsPartOf EDR 辞書

URI　　http://www.iijnet.or.jp/edr/J_index.html

・ EDR 日英対訳辞書

Type　　Text
Type.linguistics　　lexicon/bilingual lexicon
Description　　約 240,000 の日本語単語について、その対訳となる英単語を記載した辞書。日本語単語は意味によって区別されている。
Creator　　日本電子化辞書研究所
Contact person　　日本電子化辞書研究所 (thoth@edr.co.jp)
Price　　150,000 円 (アカデミック), 1,250,000 円 (一般)
Subject.language　　日本語
Language　　英語, 日本語
Format　　85 MB. 240,000 entries.
Format.encoding　　EUC-JP
Relation　　IsPartOf EDR 辞書
URI　　http://www.iijnet.or.jp/edr/J_index.html

・ EDR 英日対訳辞書

Type　　Text
Type.linguistics　　lexicon/bilingual lexicon
Description　　約 160,000 の英単語について、その対訳となる日本語単語を記載した辞書。英単語は意味によって区別されている。
Creator　　日本電子化辞書研究所
Contact person　　日本電子化辞書研究所 (thoth@edr.co.jp)
Price　　150,000 円 (アカデミック), 1,250,000 円 (一般)
Subject.language　　日本語
Language　　英語, 日本語
Format　　53 MB. 160,000 entries.
Format.encoding　　EUC-JP
Relation　　IsPartOf EDR 辞書
URI　　http://www.iijnet.or.jp/edr/J_index.html

・ EDR 概念辞書

Type　　Text
Type.linguistics　　lexicon/thesaurus
Description　　概念辞書は、単語辞書に含まれる 410,000 の概念に関する情報を記載した辞書で、概念見出し辞書、概念体系辞書、概念記述辞書の 3 つから構成される。概念見出し辞書は概念の定義を記述している。概念体系辞書は、概念間の上位下位関係を記述したシソーラスである。概念記述辞書は、agent, implement, place などの概念間の意味的関係を記述した辞書である。
Creator　　日本電子化辞書研究所
Contact person　　日本電子化辞書研究所 (thoth@edr.co.jp)
Price　　150,000 円 (アカデミック), 1,500,000 円 (一般)
Subject.language　　日本語, 英語
Language　　日本語, 英語
Format　　97 MB. 410,000 entries.
Format.encoding　　EUC-JP
Relation　　IsPartOf EDR 辞書
URI　　http://www.iijnet.or.jp/edr/J_index.html

・ EDR 日本語共起辞書

Type　　Text
Type.linguistics　　lexicon/cooccurrence database
Description　　共起する日本語単語対とそれらの意味的関係を記述した辞書。約 930,000 の単語またはフレーズが記載されている。
Creator　　日本電子化辞書研究所
Contact person　　日本電子化辞書研究所 (thoth@edr.co.jp)
Price　　150,000 円 (アカデミック), 1,400,000 円 (一般)
Subject.language　　日本語
Language　　日本語
Format　　445 MB. 930,000 entries.
Format.encoding　　EUC-JP
Relation　　IsPartOf EDR 辞書, HasPart EDR 日本語コーパス
URI　　http://www.iijnet.or.jp/edr/J_index.html

・ EDR 英語共起辞書

Type　　Text
Type.linguistics　　lexicon/cooccurrence database
Description　　共起する英語単語対とそれらの意味的関係を記述した辞書。約 460,000 の単語またはフレーズが記載されている。
Creator　　日本電子化辞書研究所

Contact person	日本電子化辞書研究所 (thoth@edr.co.jp)
Price	150,000 円 (アカデミック), 1,250,000 円 (一般)
Subject.language	英語
Language	英語, 日本語
Format	242 MB. 460,000 entries.
Format.encoding	EUC-JP
Relation	IsPartOf EDR 辞書, HasPart EDR 英語コーパス
URI	http://www.iijnet.or.jp/edr/J_index.html

・ EDR 専門用語辞書

Type	Text
Type.linguistics	lexicon/technical terminology
Description	情報処理に関する日本語と英語の専門用語を収録した辞書。日本語専門語辞書、英語専門語辞書、日英専門用語対訳辞書、英日専門用語対訳辞書、専門用語概念辞書、日本語専門用語共起辞書、英語専門用語共起辞書から構成されｒている。119,000 に日本語専門用語と 78,000 の英語専門用語が収録されている。
Creator	日本電子化辞書研究所
Contact person	日本電子化辞書研究所 (thoth@edr.co.jp)
Price	150,000 円 (アカデミック), 1,250,000 円 (一般)
Subject.language	日本語, 英語
Language	日本語, 英語
Format	145 MB. 197,000 entries.
Format.encoding	EUC-JP
Relation	IsPartOf EDR 辞書
URI	http://www.iijnet.or.jp/edr/J_index.html

・ EDR 日本語コーパス

Type	Text
Type.linguistics	annotation/corpus
Description	約 200,000 の日本語文に対して、形態素情報、構文情報、意味情報を付加したコーパス。
Annotation.corpus	word segmentation, part-of-speech, syntax, word sense
Creator	日本電子化辞書研究所
Contact person	日本電子化辞書研究所 (thoth@edr.co.jp)
Price	150,000 円 (アカデミック), 1,400,000 円 (一般)
Subject.language	日本語
Language	日本語
Format	355 MB. 200,000 sentences.
Format.encoding	EUC-JP
Relation	IsPartOf EDR 日本語共起辞書
URI	http://www.iijnet.or.jp/edr/J_index.html

・ EDR 英語コーパス

Type	Text
Type.linguistics	annotation/corpus
Description	約 120,000 の英語文に対して、形態素情報、構文情報、意味情報を付加したコーパス。
Annotation.corpus	word segmentation, part-of-speech, syntax, word sense
Creator	日本電子化辞書研究所
Contact person	日本電子化辞書研究所 (thoth@edr.co.jp)
Price	150,000 円 (アカデミック), 1,250,000 円 (一般)
Subject.language	英語
Language	英語, 日本語
Format	218 MB. 120,000 sentences.
Format.encoding	EUC-JP
Relation	IsPartOf EDR 英語コーパス
URI	http://www.iijnet.or.jp/edr/J_index.html

・ 毎日新聞 CD-ROM

Type	Text
Type.linguistics	annotation/corpus
Description	1991 年から 2001 年の毎日新聞の記事を収録した CD-ROM。購入に関する情報は以下の URL を参照。 http://cl.aist-nara.ac.jp/lab/resource/cdrom/Mainichi/MS.html
Annotation.document	keyword
Creator	毎日新聞社
Contact person	日外アソシエーツ
Price	120,000 円 (1 年当たり)
Subject.language	日本語
Date	1991-2001
Format	1 or 2 CD-ROM per year.
Format.encoding	Shift_JIS
Relation	HasPart 毎日新聞 CD-ROM (1991 年), HasPart 毎日新聞 CD-ROM (1992 年), HasPart 毎日新聞 CD-ROM (1993 年),

	HasPart 毎日新聞 CD-ROM (1994 年), HasPart 毎日新聞 CD-ROM (1995 年)
URI	http://cl.aist-nara.ac.jp/lab/resource/cdrom/Mainichi/MS.html

・ 毎日新聞 CD-ROM (1991 年)

Type	Text
Type.linguistics	annotation/corpus
Description	1991 年の毎日新聞の記事を収録した CD-ROM。約 10,000 記事。購入に関する情報は以下の URL を参照。 http://cl.aist-nara.ac.jp/lab/resource/cdrom/Mainichi/MS.html
Annotation.document	keyword
Creator	毎日新聞社
Contact person	日外アソシエーツ
Price	120,000 円
Subject.language	日本語
Date	1991
Format	1 CD-ROM.
Format.encoding	Shift_JIS
Relation	IsPartOf 毎日新聞 CD-ROM
URI	http://cl.aist-nara.ac.jp/lab/resource/cdrom/Mainichi/MS.html

・ 毎日新聞 CD-ROM (1992 年)

Type	Text
Type.linguistics	annotation/corpus
Description	1992 年の毎日新聞の記事を収録した CD-ROM。約 10,000 記事。購入に関する情報は以下の URL を参照。 http://cl.aist-nara.ac.jp/lab/resource/cdrom/Mainichi/MS.html
Annotation.document	keyword
Creator	毎日新聞社
Contact person	日外アソシエーツ
Price	120,000 円
Subject.language	日本語
Date	1992
Format	1 CD-ROM.
Format.encoding	Shift_JIS
Relation	IsPartOf 毎日新聞 CD-ROM
URI	http://cl.aist-nara.ac.jp/lab/resource/cdrom/Mainichi/MS.html

・ 毎日新聞 CD-ROM (1993 年)

Type	Text
Type.linguistics	annotation/corpus
Description	1993 年の毎日新聞の記事を収録した CD-ROM。約 10,000 記事。購入に関する情報は以下の URL を参照。 http://cl.aist-nara.ac.jp/lab/resource/cdrom/Mainichi/MS.html
Annotation.document	keyword
Creator	毎日新聞社
Contact person	日外アソシエーツ
Price	120,000 円
Subject.language	日本語
Date	1993
Format	1 CD-ROM.
Format.encoding	Shift_JIS
Relation	IsPartOf 毎日新聞 CD-ROM
URI	http://cl.aist-nara.ac.jp/lab/resource/cdrom/Mainichi/MS.html

・ 毎日新聞 CD-ROM (1994 年)

Type	Text
Type.linguistics	annotation/corpus
Description	1994 年の毎日新聞の記事を収録した CD-ROM。約 10,000 記事。購入に関する情報は以下の URL を参照。 http://cl.aist-nara.ac.jp/lab/resource/cdrom/Mainichi/MS.html
Annotation.document	keyword
Creator	毎日新聞社
Contact person	日外アソシエーツ
Price	120,000 円
Subject.language	日本語
Date	1994
Format	1 CD-ROM.
Format.encoding	Shift_JIS
Relation	IsPartOf 毎日新聞 CD-ROM
URI	http://cl.aist-nara.ac.jp/lab/resource/cdrom/Mainichi/MS.html

・ 毎日新聞 CD-ROM (1995 年)

Type	Text

Type.linguistics	annotation/corpus
Description	1995 年の毎日新聞の記事を収録した CD-ROM。約 10,000 記事。購入に関する情報は以下の URL を参照。 http://cl.aist-nara.ac.jp/lab/resource/cdrom/Mainichi/MS.html
Annotation.document	keyword
Creator	毎日新聞社
Contact person	日外アソシエーツ
Price	120,000 円
Subject.language	日本語
Date	1995
Format	1 CD-ROM.
Format.encoding	Shift_JIS
Relation	IsPartOf 毎日新聞 CD-ROM
URI	http://cl.aist-nara.ac.jp/lab/resource/cdrom/Mainichi/MS.html

・ 日経新聞 CD-ROM

Type	Text
Type.linguistics	annotation/corpus
Description	1990 年から 2000 年の日経新聞の記事を収録した CD-ROM。購入に関する情報は以下の URL を参照。 http://www.nikkeish.co.jp/gengo/zenbun.htm.
Annotation.document	keyword
Creator	日本経済新聞社
Contact person	日経出版販売 (eizo@nikkeish.co.jp)
Price	78,000 円 (1 年当たり)
Subject.language	日本語
Date	1990-2000
Format	1 CD-ROM per year.
URI	http://www.nikkeish.co.jp/gengo/zenbun.htm

・ 日経産業・金融・流通新聞 CD-ROM

Type	Text
Type.linguistics	annotation/corpus
Description	1994 年から 2000 年の日経産業・金融・流通新聞の記事を収録した CD-ROM。購入に関する情報は以下の URL を参照。 http://www.nikkeish.co.jp/gengo/zenbun.htm.
Annotation.document	keyword
Creator	日本経済新聞社
Contact person	日経出版販売 (eizo@nikkeish.co.jp)
Price	78,000 円 (1 年当たり)
Subject.language	日本語
Date	1994-2000
Format	1 CD-ROM per year.
URI	http://www.nikkeish.co.jp/gengo/zenbun.htm

・ 読売新聞 CD-ROM (邦文記事)

Type	Text
Type.linguistics	annotation/corpus
Description	1987 年から 2001 年の読売新聞の邦文記事を収録した CD-ROM。記事の量は、1987 年から 1997 年までが 1 年あたり 110,000 記事、1998 年から 2000 年までが 230,000 記事、2001 年が 340,000 記事である。購入に関する情報は以下の URL を参照。 http://www.ndk.co.jp/yomiuri/.
Annotation.document	keyword
Creator	読売新聞社
Contact person	日本データベース開発 (yomiuri@ndk.co.jp)
Price	120,000-270,000 円 (1 年当たり, アカデミック), 190,000-490,000 円 (1 年当たり, 一般)
Subject.language	日本語
Date	1987-2001
Format	1 or 2 CD-ROM per year.
Format.encoding	Shift_JIS
URI	http://www.ndk.co.jp/yomiuri/

・ 読売新聞 CD-ROM (英文記事)

Type	Text
Type.linguistics	annotation/corpus
Description	1989 年から 2001 年の読売新聞の英文記事を収録した CD-ROM。記事の量は 1 年あたり約 9,000 記事。購入に関する情報は以下の URL を参照。 http://www.ndk.co.jp/yomiuri/.
Creator	読売新聞社
Contact person	日本データベース開発 (yomiuri@ndk.co.jp)
Price	110,000-170,000 円 (1 年当たり, アカデミック), 190,000-270,000 円 (1 年当たり, 一般)
Subject.language	英語
Date	1989-2001
Format	1 CD-ROM per year.

URI　　http://www.ndk.co.jp/yomiuri/

・ 朝日新聞 CD-ROM

Type　　Text
Type.linguistics　　annotation/corpus
Description　　1985 年から 1997 年までの朝日新聞の記事を収録した CD-ROM。記事の量は 1 年あたり約 100,000 記事。
Creator　　朝日新聞社
Contact person　　紀伊国屋書店 (+81-3-3439-0123)
Price　　120,000 円 (1 年当たり)
Subject.language　　日本語
Date　　1985-1997
Format　　1 CD-ROM per year.

・ 京大コーパス

Type　　Text
Type.linguistics　　annotation/corpus
Description　　毎日新聞の 1995 年の記事の 40,000 文に対して、形態素情報と構文情報を付与したコーパス。人手修正済。毎日新聞の 1995 年の CD-ROM を別途購入する必要がある。
Annotation.corpus　　word segmentation, part-of-speech, syntax
Creator　　京都大学 言語メディア研究室
Contact person　　京都大学言語メディア研究室(corpus@pine.kuee.kyoto-u.ac.jp)
Price　　フリー
Subject.language　　日本語
Language　　日本語
Format　　6 MB.
Format.encoding　　EUC-JP
Relation　　Requires 毎日新聞 CD-ROM (1995 年)
URI　　http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/corpus.html

・ ATR 対話 DB

Type　　Text
Type.linguistics　　transcription/dialogue
Description　　会話の書き起こし文。同じ会話を日本語と英語で収録している。2 種類のトピック(国際会議の予約, 旅行代理店と客の会話)、2 種類の入力方法(電話会話、キーボード会話)の計 4 種類の会話がある。それぞれは 1 枚の CD-ROM に収録されている。
Creator　　国際電気通信基礎技術研究所(ATR)
Contact person　　ATR
Price　　50,000 円 (1 CD-ROM 当たり, 研究用途)
Subject.language　　日本語, 英語
Format　　4 CD-ROM.
URI　　http://www.red.atr.co.jp/database_main.html

・ 英文ビジネスレター文例大辞典 CD-ROM 版

Type　　Text
Type.linguistics　　annotation/corpus
Description　　ビジネスレターを書くための日本語、英語の例文集。
Creator　　日本経済新聞社
Contact person　　日経出版販売(eizo@nikkeish.co.jp)
Price　　70,000 円
Subject.language　　日本語, 英語
Date　　1998
Format　　1 CD-ROM.
Format.encoding　　Shift_JIS
Format.markup　　SGML
URI　　http://www.nikkeish.co.jp/gengo/eibun.htm

・ 中学校・高校教科書の語彙調査

Type　　Text
Type.linguistics　　lexicon/
Description　　1974 年と 1980 年の中学校・高校の教科書の語彙の調査結果のレポート。
Creator　　国立国語研究所
Subject.language　　日本語

・ 勉誠データベース

Type　　Text
Type.linguistics　　annotation/corpus
Description　　古文、和歌、漢文などのテキストデータ。約 50 テキスト。
Creator　　勉誠データセンター
Contact person　　勉誠データセンター(03-5351-3141)

Price 3,000-4,000 円 (1 フロッピーディスク当たり)
Subject.language 日本語
Format 1 floppy disk.

・ 古典対照語彙表
Type Text
Type.linguistics lexicon/
Description 「徒然草」「方丈記」など、14 の古典に現れた約 23,000 の自立語を収録した辞書。語の使用頻度も記載されている。
Creator 笠間書院
Publisher 笠間書院
Contact person 笠間書院(+81-3-3295-1331)
Price 6,695 円
Subject.language 日本語

・ データノベルズ
Type Text
Type.linguistics annotation/corpus
Description 文学作品のテキストデータ。
Creator コンピュータ出版
Publisher コンピュータ出版
Contact person コンピュータ出版(03-5486-9481)
Price 1,800 - 18,000 円
Subject.language 日本語
Format 1 floppy disk.

・ 青空文庫
Type Text
Type.linguistics annotation/corpus
Description インターネットライブラリ。著作権の切れた文学作品など、多数の文学作品を入手することができる。
Publisher http://www.aozora.gr.jp/
Contact person aozora@voyager.co.jp
Price フリー
Subject.language 日本語
URI http://www.aozora.gr.jp/

・ 判例マスター
Type Text
Type.linguistics annotation/corpus
Description 1947 年から 1994 年までの約 95,000 の判例を収録したテキストデータベース。半年に一度更新される。
Creator 新日本法規出版
Publisher 新日本法規出版
Contact person 新日本法規出版(052-211-1525)
Price 267,800 円, 40,000 円(更新)
Subject.language 日本語

・ 特許公報類 CD-ROM
Type Text
Type.linguistics annotation/corpus
Description 1994 年からの特許の公開公報と公告公報の CD-ROM。年間約 150 枚の CD-ROM を発行している。
Creator 日本特許情報機構
Contact person 日本特許情報機構(03-3503-3900)
Price 13,500 - 20,600 円 (1CD-ROM 当たり)
Subject.language 日本語

・ ICOT 形態素辞書
Type Text
Type.linguistics lexicon/
Description 約 120,000 語を収録した形態素解析用辞書。表記、読み、品詞の情報がある。
Creator 新世代コンピュータ技術開発機構(ICOT)
Publisher ftp://ftp.icot.or.jp
Price フリー
Subject.language 日本語
Language 日本語
Format.encoding ISO-2022-JP(JIS コード)
URI ftp://ftp.icot.or.jp/ifs/README.j

・ 講談社和英辞典
Type Text

Type.linguistics　annotation/corpus
Description　講談社和英辞典のテキストコーパス。38,000 文の日英対訳例文を含む。産業技術総合研究所と使用のための誓約書を取り交わす必要がある。
Creator　講談社
Contributor　橋田浩一
Contact person　橋田浩一(hasida.k@aist.go.jp)
Price　フリー
Subject.language　日本語
Language　英語

・ 語の共起関係データ
Type　Text
Description　新聞記事から抽出された「名詞-格助詞-動詞」などの共起関係のデータ。1,160,000 エントリ。
Creator　田中康仁
Contact person　田中康仁 (0794-27-5111)
Price　郵送費のみ
Subject.language　日本語

・ 現代日本語名詞シソーラス
Type　Text
Type.linguistics　lexicon/thesaurus
Description　70,000 語を含む現代日本語名詞のシソーラス。
Creator　荻野綱男
Contact person　荻野綱男 (ogino-tsunao@c.metro-u.ac.jp)
Price　フリー (研究目的)
Subject.language　日本語

・ ZenBase CD-ROM
Type　Text
Type.linguistics　annotation/corpus
Description　禅籍テキストの CD-ROM。
Creator　国際禅学研究所
Contact person　国際禅学研究所(ursapp@mbox.kyoto-inet.or.jp)
Price　1,000 円
Subject.language　日本語
Format.encoding　ISO-2022-JP(JIS コード)
URI　http://www.iijnet.or.jp/iriz/irizhtml/indexj.htm

・ ライフサイエンス辞書
Type　Text
Type.linguistics　lexicon/
Description　ライフサイエンス用語の日本語と英語の辞書。
Creator　ライフサイエンス辞書プロジェクト
Contributor　京都大学薬学部 / 国立遺伝研究所
Publisher　http://lsd.pharm.kyoto-u.ac.jp
Contact person　ライフサイエンス辞書プロジェクト(lsd@lsd.pharm.kyoto-u.ac.jp)
Price　フリー
Subject.language　日本語, 英語
URI　http://lsd.pharm.kyoto-u.ac.jp/index-J.html

・ 英語基本単語リスト
Type　Text
Type.linguistics　lexicon/
Description　Woo Linda さんによって作成された 5,000 語の英語基本単語のリスト。
Creator　Woo, Linda
Contributor　外池俊幸
Publisher　http://www.lang.nagoya-u.ac.jp/~tonoike/linda5000.html
Contact person　外池俊幸(f43633a@nucc.cc.nagoya-u.ac.jp)
Price　フリー
Subject.language　英語
URI　http://www.lang.nagoya-u.ac.jp/~tonoike/linda5000.html

・ 北大英語語彙表
Type　Text
Type.linguistics　lexicon/
Description　北海道大学によって作成された 7,500 語の英語基本語彙表。
Creator　北海道大学
Publisher　http://lexis.ilcs.hokudai.ac.jp/huvl/
Contact person　園田勝英(ksonoda@ilcs.hokudai.ac.jp)

Price フリー
Subject.language 英語

・ テレビ放送の語彙調査
Type Text
Type.linguistics lexicon/
Description 1989 年の 4 月から 6 月のテレビ放送、CM 放送を対象とした語彙調査。26,000 単語。
Creator 国立国語研究所
Contact person 大日本図書 (03-3561-8679)
Price 2,500 円
Subject.language 日本語

音声

・ ATR 音声データベース
Type Collection
Description 6 つのデータセットから構成される音声データベース。
Creator 国際電気通信基礎技術研究所(ATR)
Contact person ATR
Subject.language 日本語, 英語
Relation HasPart ATR 音声 DB(セット A), HasPart ATR 音声 DB(セット B), HasPart ATR 音声 DB(セット C), HasPart ATR 音声 DB(セット D), HasPart ATR 音声 DB(セット E), HasPart ATR 音声 DB(セット F)
URI http://www.red.atr.co.jp/database_main.html

・ ATR 音声 DB(セット A)
Type Sound
Type.linguistics transcription/read speech
Description 日本語読み上げ音声データ。20 話者。8,500 単語。
Creator 国際電気通信基礎技術研究所(ATR)
Contact person ATR
Price 600,000 円 (研究用途)
Subject.language 日本語
Format 1 CD-ROM.
Relation IsPartOf ATR 音声データベース
URI http://www.red.atr.co.jp/database_main.html

・ ATR 音声 DB(セット B)
Type Sound
Type.linguistics transcription/read speech
Description 日本語読み上げ音声データ。10 話者。503 文。
Creator 国際電気通信基礎技術研究所(ATR)
Contact person ATR
Price 350,000 円 (研究用途)
Subject.language 日本語
Format 1 CD-ROM.
Relation IsPartOf ATR 音声データベース
URI http://www.red.atr.co.jp/database_main.html

・ ATR 音声 DB(セット C)
Type Sound
Type.linguistics transcription/read speech
Description 日本語読み上げ音声データ。20 話者。84 タイトル。
Creator 国際電気通信基礎技術研究所(ATR)
Contact person ATR
Price 600,000 円 (研究用途)
Subject.language 日本語
Format 1 CD-ROM.
Relation IsPartOf ATR 音声データベース
URI http://www.red.atr.co.jp/database_main.html

・ ATR 音声 DB(セット D)
Type Sound
Type.linguistics transcription/read speech
Description 日本語読み上げ音声データ。4 話者。400 文書。
Creator 国際電気通信基礎技術研究所(ATR)
Contact person ATR
Price 270,000 円 (研究用途)
Subject.language 日本語
Format 1 CD-ROM.
Relation IsPartOf ATR 音声データベース

URI　http://www.red.atr.co.jp/database_main.html

・ ATR 音声 DB(セット E)
Type　Sound
Type.linguistics　transcription/read speech
Description　英語読み上げ音声データ。4 話者。5,000 単語。
Creator　国際電気通信基礎技術研究所(ATR)
Contact person　ATR
Price　270,000 円 (研究用途)
Subject.language　英語
Format　1 CD-ROM.
Relation　IsPartOf ATR 音声データベース
URI　http://www.red.atr.co.jp/database_main.html

・ ATR 音声 DB(セット F)
Type　Sound
Type.linguistics　transcription/read speech
Description　英語読み上げ音声データ。6 話者。1,100 文。
Creator　国際電気通信基礎技術研究所(ATR)
Contact person　ATR
Price　600,000 円 (研究用途)
Subject.language　英語
Format　1 CD-ROM.
Relation　IsPartOf ATR 音声データベース
URI　http://www.red.atr.co.jp/database_main.html

・ ATR 自然発話・言語 DB
Type　Sound
Type.linguistics　transcription/dialogue
Description　旅行代理店と顧客の模擬対話を収録した音声データ。5 つのセットから成る。日本語での会話が 892、日本語と英語での会話が 618。書き起こし文と形態素情報も付加されている。
Annotation.corpus　word segmentation, part-of-speech
Creator　国際電気通信基礎技術研究所(ATR)
Contact person　ATR
Price　180,000 円 (1 セット当たり, 研究用途)
Subject.language　日本語, 英語
Format　4 CD-ROM.
URI　http://www.red.atr.co.jp/database_main.html

・ ATR 多数話者音声 DB
Type　Collection
Description　多数話者による音声データ。
Creator　国際電気通信基礎技術研究所(ATR)
Contact person　ATR
Subject.language　日本語
Relation　HasPart ATR 多数話者音声 DB(模擬会話), HasPart ATR 多数話者音声 DB(音素バランス文), HasPart ATR 多数話者音声 DB(辞書データ)
URI　http://www.red.atr.co.jp/database_main.html

・ ATR 多数話者音声 DB(模擬会話)
Type　Sound
Type.linguistics　transcription/conversation
Description　多数の話者による音声データベース。3,774 人の話者が会議のスケジューリングに関する模擬対話を行った。4 つのセットから構成される。
Creator　国際電気通信基礎技術研究所(ATR)
Contact person　ATR
Price　180,000 円 (1 セット当たり, (研究用途), 1,000,000 円 (1 セット当たり, 商品化用途)
Subject.language　日本語
Format　3-5 CD-ROM per a set.
Relation　IsPartOf ATR 多数話者音声 DB
URI　http://www.red.atr.co.jp/database_main.html

・ ATR 多数話者音声 DB(音素バランス文)
Type　Sound
Type.linguistics　transcription/read sentence
Description　多数の話者による音声データベース。3,774 人の話者が音素バランス文の読み上げを行った。4 つのセットから構成される。
Creator　国際電気通信基礎技術研究所(ATR)
Contact person　ATR

Price 180,000 円 (1 セット当たり, 研究用途), 1,000,000 円 (1 セット当たり, 商品化用途)
Subject.language 日本語
Format 7-10 CD-ROM per a set.
Relation IsPartOf ATR 多数話者音声 DB
URI http://www.red.atr.co.jp/database_main.html

・ ATR 多数話者音声 DB(辞書データ)
Type Sound
Type.linguistics transcription/read sentence
Description 多数の話者による音声データベース。3,770 人の話者が辞書データの読み上げを行った。
Creator 国際電気通信基礎技術研究所(ATR)
Contact person ATR
Price 180,000 円 (研究用途), 1,000,000 円 (商品化用途)
Subject.language 日本語
Format 5 CD-ROM.
Relation IsPartOf ATR 多数話者音声 DB
URI http://www.red.atr.co.jp/database_main.html

・ 日本音響学会研究用連続音声データベース
Type Sound
Type.linguistics transcription/dialogue
Description 次の 3 つデータから成る音声データベース。(a) ATR 音素バランス 503 文, 64 話者(男性 30 名, 女性 34 名), のべ 9600 文。(b) 案内タスク文, 36 話者(男性 18 名, 女性 18 名), のべ 12474 文。(c) 模擬対話 37 対話, 書き起こしテキスト付き, 37 話者(男性 29 名, 女性 8 名)
Creator 日本音響学会
Contact person 西垣繁雄(〒105 港区芝公園 3-5-8 機会振興会舘内 (財)日本情報処理開発協会 AI ファジー振興センター tel 03-3432-9390, fax 03-3431-4324)
Price 3090 円＋送料
Subject.language 日本語
Format 7 CD-ROM. Sampling: 16kHz, 16bits.

・ 日本音響学会 新聞記事読み上げ音声コーパス(JNAS)
Type Sound
Type.linguistics transcription/dialogue
Description JNAS とは Japanese Newspaper Article Sentences の略。このコーパスは、毎日新聞記事と ATR 音素バランス 503 文を 306 人の話者 (男女そ れぞれ 153 名) が読み上げたデータとそのテキストから構成さ れている。発話はすべて日本語である。
Creator 日本音響学会
Contact person メディアドライブ株式会社 宮井千代子 chiyoko@mediadrive.co.jp
Price 実費
Subject.language 日本語
Format 16 CD-ROM. Sampling: 16kHz, 16bits.
URI http://www.milab.is.tsukuba.ac.jp/jnas/

・ 電総研道案内対話音声コーパス 1998
Type Sound
Type.linguistics transcription/dialogue
Description Wizard of OZ 法によって収録された、道案内に関する機械と人間との 197 個の対話から成る音声対話コーパス。人間の発話の音声データ・ピッチパターン・書き起し・発話の始端と終端・発話の意味表現からなる。
Creator 電子技術総合研究所(現 産業技術総合研究所)
Contact person 電子技術総合研究所(etlsdg@ni.aist.go.jp)
Price 郵送費
Subject.language 日本語
Date 1998
Format 1 CD-ROM.
URI http://akiba.media-interaction.jp/ETLSDG/

・ 電総研音素バランス単語セット WD-I & II
Type Sound
Type.linguistics transcription/word
Description 音素バランス単語セットの単語を男性話者が読み上げた音声データ。WD-I は 492 語、WD-II は 1,542 語から成る。WD-I は WD-II の部分集合である。
Creator 電子技術総合研究所(現 産業技術総合研究所)
Contact person 田中 和世(kaz.tanaka@aist.go.jp)
Price 郵送費
Subject.language 日本語
URI http://unit.aist.go.jp/is/speech/etlwd12a.html

・ 電子協日本語共通音声データ--DAT 版--

Type Sound
Description このコーパスは 110 音節、178 単語、35 個の 4 桁数字、計 323 個の単語を 4 回ずつ読み上げたデータである。録音時間は 120 時間で、76 本の DAT カセットに収められている。それぞれの単語は 20 歳から 60 歳の男女各 75 名ずつによって発音されている。合計のサンプル数は 193,800 である。
Creator 日本電子工業振興協会(現 電子情報技術産業協会)
Contact person 佐々木氏(サンライズミュージック, 〒106 東京都港区六本木 4-11-10 六本木富士ビル 4 階, Tel: 03-3408-6541, Fax: 03-3408-1505)
Subject.language 日本語
Format Sampling: 44kHz, 16bits.

・ 連続音声(文科省 科研費 試験研究)
Type Sound
Type.linguistics transcription/
Description 様々な単音節, 単語, 短文, 文章を 6 名の男女によって読み上げた音声データ。
Creator 筑波大学 板橋研究室
Contact person 板橋秀一(itahashi@milab.is.tsukuba.ac.jp)
Price フリー(CD-ROM 版, 研究者のみ), 70,000 円(DAT 版)
Subject.language 日本語
Format CD-ROM or DAT. Sampling: 16kHz, 16bit.

・ 方言音声データベース
Type Sound
Description 日本語の方言の音声データベース。大学、官公庁研究所に限る。
Creator 田原 広史(大阪樟蔭女子大学), 江川 清(国立国語研究所)
Contributor 文科省 科研費 重点領域 「日本語音声」
Contact person 田原 __萌大阪樟蔭女子大学. Tel. 06-723-8181, Fax. 06-723-8881), 江川 清(国立国語研究所, Tel. 03-3900-3111, Fax. 03-3906-3530)
Subject.language 日本語
Format 19 Audio CD. 3 CD-ROM.

・ 重点領域研究 音声対話コーパス
Type Sound
Type.linguistics transcription/dialogue
Description 93 対話の音声データと書き起こしテキスト。
Creator 堂下修司
Contributor 文科省 科研費 重点領域 「音声・言語・概念の統合的処理による対話の理解と生成に関する研究」
Contact person メディアドライブ株式会社(juten-corpus@mediadrive.co.jp)
Price 10,000 円
Subject.language 日本語
Format 4 CD-ROM.
URI http://winnie.kuis.kyoto-u.ac.jp/taiwa-corpus/

・ RWCP-DB-SPEECH-96-I (RWC 音声対話データベース)
Type Sound
Type.linguistics transcription/dialogue
Description 「海外旅行計画」24 対話、「車の購入」24 対話の音声波形と書き起こしテキスト。
Creator Real World Computing Partnership, Japan
Contact person メディアドライブ株式会社(sprwcdb@mediadrive.co.jp)
Price 2,000 円
Subject.language 日本語
Format 4 CD-ROM.
URI http://www.rwcp.or.jp/wswg/rwcdb/speech/index.html

・ 東北大 -- 松下単語音声データベース
Type Sound
Description 単語音声データベース。大学、官公庁研究所に限る。
Creator 牧野正三, 二矢田勝行, 真船裕雄, 城戸健一
Contact person 牧野 正三(東北大学, Tel. +81-22-262-3469, Fax. +81-22-262-3469)

・ 早大白井研 100 地名単語データベース
Type Sound
Description 100 個の地名の単語の音声データベース。12 人の男性が 2 回ずつ読み上げた。
Creator 早稲田大学白井研究室
Contact person 大平 茂輝 (ohira@shirai.info.waseda.ac.jp)
Subject.language 日本語
Format Sampling: 12.5kHz, 12bit.

・ 京大堂下研 音素バランス単語セット
Type Sound

Description	音素バランス単語セットを男性 28 名、女性 16 名が読み上げたデータ。
Creator	京都大学堂下研究室
Contact person	河原 達也 (kawahara@kuis.kyoto-u.ac.jp)
Format	Sampling: 16kHz, 16bit.

ツール

・ JUMAN

Type	Software
Type.functionality	morphological analyzer
Description	ユーザによる拡張可能な日本語形態素解析ツール。
Creator	京都大学 言語メディア研究室
Contact person	京都大学言語メディア研究室(juman@pine.kuee.kyoto-u.ac.jp)
Price	フリー
Subject.language	日本語
Format	4 MB.
Format.os	unix,MSWindows
Format.sourcecode	C
URI	http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/juman.html

・ 茶筌

Type	Software
Type.functionality	morphological analyzer
Description	茶筌はフリーの日本語形態素解析ツールである。JUMAN に改良を加え、ツールとしての完成度を飛躍的に向上させた。奈良先端科学技術大学院大学 情報科学研究科 計算言語学研究室によって 1997 年 2 月 19 日に ver.1.0 がリリースされた。最新のバージョンは 2002 年 2 月 8 日にリリースされた ver. 2.2.9 である。
Creator	奈良先端科学技術大学院大学
Contact person	奈良先端科学技術大学 松本研究室(chasen@is.aist-nara.ac.jp)
Price	フリー
Subject.language	日本語
Format	3.3MB.
Format.os	unix,MSWindows
Format.sourcecode	C
URI	http://chasen.aist-nara.ac.jp/index.html.ja

・ KNP

Type	Software
Type.functionality	syntactic analyzer
Description	日本語の構文解析ツール。最初に入力文を文節に区切り、次に文節間の係り受け関係を解析する。
Creator	京都大学 言語メディア研究室
Contact person	黒橋禎夫 (kuro@kc.t.u-tokyo.ac.jp)
Price	フリー
Subject.language	日本語
Format	145 KB.
Format.os	unix
Format.sourcecode	C
Relation	Requires JUMAN, Requires 分類語彙表, Requires EDR 日本語単語辞書, Requires IPAL 辞書(optional)
URI	http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/knp.html

・ MSLR パーザ

Type	Software
Type.functionality	morphological and syntactic analyzer
Description	MSLR パーザとそれに関連するツールをまとめたツールキット。MSLR パーザは形態素解析と構文解析を同時に行う LR パーザである。日本語解析のための標準辞書と文法が含まれる。さらに、ユーザは独自の辞書や文法を用いることもできる。
Creator	東京工業大学
Contact person	東京工業大学 田中・徳永研究室 (mslr@cl.cs.titech.ac.jp)
Price	フリー
Subject.language	日本語
Format	1.5 MB.
Format.os	unix
Format.sourcecode	C
URI	http://tanaka-www.cs.titech.ac.jp/pub/mslr/index-j.html

・ すもも

Type	Software
Type.functionality	morphological analyzer
Description	日本語の形態素解析ツール。最適解のみを高速に出力するようにカスタマイズされている。単純な未知語処理も行う。
Creator	NTT コミュニケーション科学研究所

Contact person　鷲坂光一 (wasisaka@nttlabs.com), 山崎憲一 (yamazaki@t.onlab.ntt.co.jp)
Price　フリー
Subject.language　日本語
URI　http://www.t.onlab.ntt.co.jp/sumomo/index.html

・ SAX
Type　Software
Type.functionality　Tool for syntactic analysis
Description　拡張文脈自由文法の一つである DCG(Definite Clause Grammar)に基づいて記述された文法をコンパイルして、上昇型チャート法に基づく構文解析 Prolog プログラムを生成するシステム。SICStus Prolog が必要。
Creator　奈良先端科学技術大学院大学
Contact person　奈良先端科学技術大学院大学 松本研究室 (nlt@is.aist-nara.ac.jp)
Price　フリー
URI　http://chasen.aist-nara.ac.jp/sax.html

・ BUP
Type　Software
Type.functionality　Tool for syntactic analysis
Description　拡張文脈自由文法の一つである DCG(Definite Clause Grammar)に基づいて記述された文法をコンパイルして、左隅構文解析 Prolog プログラムを生成するシステム。SICStus Prolog が必要。
Creator　奈良先端科学技術大学院大学
Contact person　奈良先端科学技術大学院大学 松本研究室 (nlt@is.aist-nara.ac.jp)
Price　フリー
URI　http://chasen.aist-nara.ac.jp/bup.html

・ 美寿満 (ViJUMAN)
Type　Software
Type.functionality　Visualization tool for morphological analyzer
Description　形態素解析ツール「JUMAN」の解析結果を視覚化するツール。
Creator　奈良先端科学技術大学院大学
Contact person　奈良先端科学技術大学院大学 松本研究室 (vijuman-adm@cl.aist-nara.ac.jp)
Price　フリー
Subject.language　日本語
Format.os　unix
Relation　Requires JUMAN
URI　http://chasen.aist-nara.ac.jp/vi4ma.html

・ 美茶 (ViCha)
Type　Software
Type.functionality　Visualization tool for morphological analyzer
Description　形態素解析ツール「茶筌」の解析結果を視覚化するツール。
Creator　奈良先端科学技術大学院大学
Contact person　奈良先端科学技術大学院大学 松本研究室 (vijuman-adm@cl.aist-nara.ac.jp)
Price　フリー
Subject.language　日本語
Format.os　unix
Relation　Requires 茶筌
URI　http://chasen.aist-nara.ac.jp/vi4ma.html

・ 構文解析過程表示システム (VisIPS)
Type　Software
Type.functionality　Visualization tool for syntactic analyzer
Description　構文解析ツールのための視覚化ツール。CKY 表や解析木を図示できる。
Creator　奈良先端科学技術大学院大学
Contact person　奈良先端科学技術大学院大学 松本研究室 (nlt@is.aist-nara.ac.jp)
Price　フリー
Format.os　unix
Relation　Requires SAX
URI　http://chasen.aist-nara.ac.jp/visips.html

・ SUFARY
Type　Software
Type.functionality　Tool for string matching
Description　Suffix array を用いた文字列検索ツール。
Creator　奈良先端科学技術大学院大学
Contact person　奈良先端科学技術大学院大学 松本研究室 (sufary@cl.aist-nara.ac.jp)
Price　フリー
Format.os　unix
Format.sourcecode　C

・ Breakfast
Type Software
Type.functionality morphological analyzer
Description 高速な形態素解析ツール。使用者が形態素文法を自由に記述できる点が特徴。
Creator 富士通研究所
Contact person 颯々野 学 (bf-staff@ling.flab.fujitsu.co.jp)
Price フリー
Subject.language 日本語
Format.os Windows 95, NT 3.51, NT 4.0
URI http://www.labs.fujitsu.com/free/breakfast/index.html

・ VisualMorphs
Type Software
Type.functionality assistant tool for constructing POS-tagged corpora
Description 品詞タグ付きコーパス作成支援ツール。形態素解析システムの出力を表示・修正するための GUI ツール。
Creator 奈良先端科学技術大学院大学
Contact person 奈良先端科学技術大学院大学 松本研究室 (chasen@cl.aist-nara.ac.jp)
Price フリー
Subject.language 日本語
Date 2001
Format.os unix, windows
Format.sourcecode java
URI http://chasen.aist-nara.ac.jp/vm/index.html.ja

・ 南瓜(CaboCha)
Type Software
Type.functionality syntactic analyzer
Description Support Vector Machine に基づく日本語係り受け解析器。
Creator 奈良先端科学技術大学院大学
Contact person 工藤拓(taku-ku@is.aist-nara.ac.jp)
Price フリー
Subject.language 日本語
Date 2001
Format.os unix, windows
Relation Requires 茶筌, Requires YamCha
URI http://cl.aist-nara.ac.jp/~taku-ku/software/cabocha/

・ YamCha
Type Software
Type.functionality chunker
Description 日本語の汎用 chunker。カスタマイズが可能でオープンソース。Support Vectore Machine を利用している。
Creator 奈良先端科学技術大学院大学
Contact person 工藤拓(taku-ku@is.aist-nara.ac.jp)
Price フリー
Subject.language 日本語
Date 2001
Format.os unix
URI http://cl.aist-nara.ac.jp/%7Etaku-ku/software/yamcha/

・ TinySVM
Type Software
Type.functionality machine learning tool
Description 日本語の汎用 chunker。カスタマイズが可能でオープンソース。Support Vectore Machine を利用している。
Creator 奈良先端科学技術大学院大学
Contact person 工藤拓(taku-ku@is.aist-nara.ac.jp)
Price フリー
Date 2001
Format.os unix
URI http://cl.aist-nara.ac.jp/%7Etaku-ku/software/TinySVM/

・ 和布蕪(MeCab)
Type Software
Type.functionality morphological analyzer
Description 形態素解析ツール茶筌の別バージョン。茶筌より 3〜4 倍高速に動作する。
Creator 奈良先端科学技術大学院大学
Contact person 工藤拓(taku-ku@is.aist-nara.ac.jp)
Price フリー
Date 2001

Format.os	unix
URI	http://cl.aist-nara.ac.jp/%7Etaku-ku/software/mecab

・ 日本語スペルチェッカー

Type	Software
Type.functionality	spell checker
Description	大量の平仮名列を用いて辞書を作成し，平仮名列に対してのみスペルチェックを行う。
Creator	京都大学 言語メディア研究室
Contact person	京都大学言語メディア研究室
Price	フリー
Subject.language	日本語
Format.sourcecode	C
URI	http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/spell_checker.html

参考文献

[1] James Allen, “Natural Language Understanding”(2nd Edition), Addison-Wesley, 1995.

３．３ 言語コーパスにおける著作権に関する調査

３．３．１ はじめに

近年、自然言語処理の分野では、言語コーパスを用いた手法が１つの主流となっている。そこでは、実際に使われた言語データがそのまま資料とされる。この言語コーパスに対し、統計的手法などを適応することによって、内省に基づく文法規則ベースの従来の手法では取扱いが困難であった「生の言語現象」を扱うことが可能になった。これにより、自然言語処理技術は言う及ばず、音声認識技術や音声合成技術などが大きく進歩したことは疑う余地がない。つまり、現在の自然言語処理技術の研究・開発においては、良質で大量の言語コーパスの活用が必要不可欠の条件となっていると言えよう。

さらに、近年のインターネットの世界規模の普及に代表されるように、大量の電子化された言語データを容易に入手し利用できるようになったことも、言語コーパスの利用拡大を後押ししている。しかし、全ての技術と同様に、電子データの流通にも功罪の両側面がある。例えば、音楽コンテンツやアプリケーションプログラムの不正コピーがネットワークを経由して無秩序に流通することで、音楽やプログラムの作者らの権利が侵害されるケースも多発し、社会問題化する中、音楽ファイルの無料交換サービスの著作権侵害を東京地裁が認定[8]するなど、知的財産権に関する論議が活発に行なわれている[1,2,3,6,7]。こういった状況を踏まえると、言語コーパスを利用して自然言語処理技術の研究/開発を行なう際にも、著作権への十分な理解と、他者の権利の尊重とが必要となる。そしてこれは、健全で適切な企業活動や学術活動に不可欠な義務である。しかし、著作権法自身は文字通り「法律」であり、専門家以外がその詳細をたやすく理解できるものではない。その上、近年の情報処理技術利用の急速な拡大に伴い、著作権(法)自身の解釈も刻々変化しており、個々の技術者や研究者がこれを随時完全に理解し、把握しておくことは容易ではない。

そこで、本ワーキンググループでは、自然言語処理技術を研究開発において言語コーパスを利用する際に、著作権の観点から留意すべき事柄を再確認することを目的として、「言語コーパスにおける著作権に関する調査」を実施することとした。

本調査では、主たる設定として「企業の自然言語技術の研究/開発者が、自然言語処理技術を研究し開発する為に、自然言語コーパスを利用する状況」を想定して調査を行なった。ここでの自然言語コーパスの入手方法としては、言語コーパスとして整備済みの言語データを契約に基づいて利用する場合から、書籍や新聞に代表される非電子化メディアを参照し利用する場合、あるいは、インターネットからもたらされるデータを利用する場合など、様々なケースを想定して検討した。また、自身の立場としては、上述の主たる設定の他、大学や公立機関などの非営利団体の自然言語技術の研究/開発者を想定した場合についても、補足的に調査した。

以後 第二項では本調査の方法を概説し、続く第三項～第五項で実施したインタビュー結果を報告する。続く、第六項でその分析と考察を行い、第七項に本調査のまとめを示す。

３．３．２ 調査方法

本調査では立場の異なる 3 人の講師を招き、それぞれの立場からコーパス利用時における著作権の捉え方について見解を講演いただいた。講演をお願いするにあたっては、本調査の目的を正解に御理解いただき適切な解説をいただくためと、各講師の見解の比較検討が容易になるように、事前に共通設問集をお渡ししてそれぞれの事例における著作権法の解釈について解説をいただいた。事前に提示した設問は図 3.3-1～2 のとおりである。

設問は大きく分けて 2 種類のケースを想定している。1 つめのケースはウェブページや掲示版、ニュースグループなどの情報から言語規則等に関する知識を獲得する場合を想定している。この場合には、著作権者は極めて多数にわたり、また、掲示版などでは著者の特定すら困難な場合も多いことから、著作権者の許諾を得ての利用は事実上不可能である。こういった場合に著作権法はどのように解釈されるのかを専門家に解説いただいた。

2 つめのケースは、新聞社等の著作権保有者からコーパスの使用権を取得して利用し、この利用によって新たなデータを生成した場合を想定している。どういった性格のデータをどういった手順で生成した場合に著作権侵害となるかを解説いただいた。このケースの場合には著作権の問題というよりは個々の使用許諾契約の解釈の問題ではあるが、使用許諾では単に「研究目的のみに利用できる」と規定している場合も多く、こういった場合にどのように解釈すべきかの調査を意図している。

なお、当然であるが、著作権侵害か否かの判定で個々のケースを精査することなく一般論として断定的な結論を導くことはできない。したがって、各講師の判断が分かれている箇所については、講師の見解の相違とともに、例示された設問から各講師が想定されたであろう具体的事例がそれぞれの講師で異なっている場合もあり得る。また、これも当然であるが、侵害の有無は最終的には裁判所の判断によることであり、たとえ各講師の見解が一致していてもこれをもって確定的な結論というわけではない。

３．３．３ ヒヤリング１

第一回目のヒヤリングでは､(株)東芝研究開発センターの木村和広氏を講師に招き、GSK [4)]の著作権 WG(Working Group)による活動の概要報告を聴講した。これは、EDR、NHK、富士通、NEC、松下、インターG､、ＮＨＫ、電総研(当時) 、および東芝から参加した 10 名のメンバーによって、1999 年 6 月から 2000 年 5 月に渡って実施された活動の総括報告である｡以降本項では､この講演の概要と、前項で設定した共通設問への講師木村氏の見解を紹介する。

実施日：	2002 年 9 月 5 日
実施場所：	電子情報技術産業協会 305 会議室
タイトル：	「GSK 著作権ＷＧ活動報告」
講演者：	株式会社 東芝 研究開発センター 木村和広氏

(1) 講演概要

① 目的：GSK の提供コンテンツの拡充

当時立ち上がり時期にあったGSK活動を軌道に載せる為には、魅力あるコンテンツの提供が必要とされていた。

② 方針：言語コンテンツ提供者と GSK の間のコンセンサスを形成

コンテンツ拡充には、その提供者(プロバイダ)とのコンセンサス形成が必要との見地に立ち、プロバイダの信頼獲得とプロバイダの権利保護に関して検討すべく、以下(ア)～（エ）の活動が行なわれた。

③ 活動内容

(ア) 先行事例の調査： ELRA(European Language Resources Association)[5] の契約

プロバイダはELRAに言語資源の配布権を付与し対価の保証を得る。一方ELRAは配布目的の範囲で、自己方針での宣伝、コピーなどを行なえる。ユーザの権利は、非独占的、譲渡不可、再配布不可で、商用利用は別契約が必要。データは欠陥込みで提供され品質の保証は一切されない。

この調査結果から、プロバイダの信頼獲得が最重要であるとの知見が導かれた。

(イ) デジタル著作権の基本理解

言語資源の利用形態を、図 3.3-3 の通りにモデル化し、このモデルに含まれる各過程における著作権上の留意点を分析。まず、(a)の言語資源の著作物性の有無に関しては、仲介者たる GSK は関与せず、議論も無用との考えが示された。次に、（ｃ）の二次データが複製物であるか二次著作物であるかは、個別事例ごとに判定が必要であり、そこでは元データの推知性/復元性の有無が重要であるとの知見が示された。

(ウ) ケーススタディ：言語資源のプロバイダと利用者との間の認識のずれを仮定し検証

「言語資源の利用者はその情報内容には興味が無いにもかかわらず、プロバイダは情報内容を重要視しており、その侵害を懸念している」のではないかという仮説をたてての検証が行なわれた。ここでは、音声認識処理や文書検索処理といった事例を取り上げて、そこでの言語資源の活用方法を説明する為の資料としての、読み付きの言語コーパスや、文字化した音韻モデル、あるいは言語モデルといった具体的なデータ例が用意された。これらは、続く意見交換の場で、プロバイダに提示されて説明に用いられた。

(エ) プロバイダとの意見交換

新聞社 1 社、インターネットプロバイダ 2 社、放送局 1 社とコンタクトし、上述の説明資料を用いての言語資源の実際の利用の様子の説明がなされた。これに対して、プロバイダからは、これまでわからなかった言語資源の利用形態が理解できたとの好意的反応が得られた。またプロバイダの懸念事項は、無断掲載が最大のも

ので、以下、個人情報の流出、対価などがあることが判明した。また、プロバイダ側には、言語資源の公開は金銭的メリットがあるはずとの認識と、公開のためのデータ加工コストへの懸念があることも判明した。また、言語資源の共有を専門に扱う非営利団体である GSK の存在は、プロバイダ企業内の社内調整を円滑に進めるためにも有益であるとの意見が得られたという。

④ まとめ

プロバイダからの言語資源提供を促進する為の提言として、(a)ネガティブファクタ払拭と、(b)インセンティブ提示の必要性が指摘された。さらに、これらを実現する為の具体的施策として、前者(a)に関しては、利用形態の明確化、プロバイダの権利保護、公開コスト低減支援などを、また後者(b)に関しては、ELRA などによる市場調査結果の利用や、社会的・技術的貢献意義の啓蒙などが提示された。

⑤ 今後の課題

掲示板や、チャット、ニュース、あるいはメイルなどといった、ネットワーク上の言語資源は、言語資源の利用者からの需要が大きなデータであるが、これらは、複数あるいは不特定の発言者が著作権者となるため、個別の交渉が困難で、その利用が難しい今後の課題であるとの見解が示された。

(2). 共通設問への見解概要

まず、(1)「原著作権者の承諾を得ずに行なってよいことは」に関し、著作権法で認められている保護の例外規定の適用される場合には問題がないとの見解が示された。一方、例外規定が適応されないケースについては、統計情報の抽出(c)や、適切な目的(b)(c)の為のデータの一時的な保存、あるいはQ＆Aの知識源としての利用(m)などが可能であろうが、新聞等無断複製を明示的に禁止している情報の利用(g)は承諾が必要との見解が示された。

また、(2)「研究用ライセンスの上でやってよいことは」に関しては、言語コーパスを利用して作成される、統計データ(n)や単語辞書(o)などの二次データ自身の権利は、その作成者に属するが、元データの改変(p)は認められないとの知見が示された。また (q)に関しては利用条件で判断が変わり、可否の判定が困難であることが指摘された。

３．３．４ ヒヤリング２

第２回目のヒヤリングでは、NTT アドバンステクノロジ株式会社の藤波進氏を講師に招き、著作権問題の研究者としての立場から講演をいただいた。氏は著作権法の専門家であり、マルチメディアコンテンツ利用時の著作権問題に関しての著作がある。

講演では、著作物利用にあたっての必要な措置についての総論に続いて、3.3.2 節で設定した共通設問に対しての講師の見解が示された。なお本講演で示された見解は所属組織とはかかわりなく、講師個人の見解であることを強調された。

実施日：	2002 年 10 月 15 日
実施場所：	電子情報技術産業協会 309 会議室
タイトル：	「不当と違法の狭間で…」
講演者：	NTT アドバンステクノロジ 藤波進氏

(1) 講演概要

① 基本的考え方

一般に法は権利者(本例では著作権者)のためにある。ただし、法、特に財産権法はバランスである。(本講演では財産権のみを対象とし、侵害時に回復不能な人格権は対象外とする)

② 知的財産と法

知的財産には、特許、実用新案、商標、意匠、著作権などがあるが、他人の成果に「フリーライド」することは原則自由であり、模倣は進歩の源である。フリーライドが即違法というわけではない。一方、知的財産(権)法はそのフリーライドを禁止する法であり、許されるフリーライドと許されないフリーライドを明確な線引きであらかじめ法令等で明記したものである。

③ 著作物の利用

著作物は著作権の譲渡、もしくは著作権者からの利用許諾により利用可能となる。ただし、利用許諾が可能であるためには(1)対象物が著作権法上の著作物であり(2)許諾される利用の態様が著作権法に権利として明記されていること,が必要である。

④ 著作物利用の課題と措置

著作物の利用にあたっての課題は「判断に迷う」ということである。例えば、単なる事実の報道なのか著作物なのか、または、許諾が必要な利用態様なのか不要な(著作権が制限される)利用態様なのかが容易に判別つかないケースが多い。さらには、'不当'と'違法'の境界も不明確であり、判断に迷う場合が多い。

措置としては専門家への相談が挙げられるが、専門家でも均一な判断が下せるわけではなく、専門家間で異なる「常識」、異なる「認識」が存在している。

したがって、可能な措置は通常一般人の判断を以って、もしくは、社会通念に照らして「適法な」(と判断できる)方法で利用し、異常な事態に至っても「不当な」態様での利用に留まることが肝要である。

(2) 共通設問への見解概要

① 事例回答の前提

まず、著作権法の著作権制限規定要件(教育目的など)を満たす態様での使用であれば事例はすべてＯＫである。以下は要件を満たさなかった場合についての見解。

② 研究用ライセンスについて

「研究用ライセンス」なる権利は効力のある権利としては存在せず、契約当事者間でのみ有効な概念である。したがって契約内容に依存する問題である。契約で定まっていない事項は法の規定に従うことになる。(共通設問後半(n)～(q)に対する見解; 表 3.3-3 参照)

③ fair use について

米国では、公正使用(米:Copyright Act of 1976.107 条:fair use doctrine) として著作権侵害の除外が認められる(ただし、fair use の定義や許容範囲は条文にはない)。批評、論評／解説、ニュース報道、授業、研究、調査等の目的で公正使用されたときは、著作権侵害を構成しない。さらに条文記載の行為は例示にすぎず、対象行為の制限はない。判断は、使用者に使用を認めないときの使用者の不利益と、使用を認めたときの著作権者の不利益を比較した上でなされる。

一方、日本では fair use の主張は認められていない。

④ 著作権者の許諾を受けない場合についての見解

(共通設問前半(a)～(m)に対する見解; 表 3.3-1～2 参照)

３．３．５ ヒヤリング３

富士通　法務・知的財産本部に在籍しておられる亀井氏に著作権に関する講演をお願いした。亀井氏はＪＥＩＴＡの知的財産関係のWGでも活動されており、メーカーにおける知財活動に精通していらっしゃる方である。GSK の著作権WGで活動された経験をお持ちで、言語処理、言語資産に関わる著作権の問題について明るい方である。ご講演でも判定のポイントを明確に説明された。

実施日：	2002 年 11 月 15 日
実施場所：	電子情報技術産業協会 302 会議室
タイトル：	「言語コーパスの著作権」
講演者：	富士通株式会社 法務・知的財産権本部 亀井正博氏

(1) 講演概要

言語資源の利用に際し、著作権者の許諾を必要とするか否かについて以下の３ステップで基本的な判定ができる。

- 言語資源が著作物であるか
- 処理の過程の行為が著作権に触れる行為であるか
- 行為を適法とする権利制限規定があるか

実際にはこれらは法律上の概念であり、解釈の揺れが生じる可能性があるため、確定的な解を出すのは難しい。

言語資源の利用態様は図 3.3-3 のようになる。前述の３ステップでみた場合、言語資源が著作物

であり、解析処理等の過程で複製等の著作権に触れる行為があり、得られる二次データが元の言語資源の二次的著作物ならば、そのような言語資源の利用は著作権法上の権利制限規定に掲げられた場合に相当しない限り、言語資源の著作権者の許諾を要する行為となる。以下では言語資源、二次データ、利用過程の行為ごとに著作権上の評価と注意すべき点について述べる。

①言語資源が著作物であるかどうかの判定の重要性

著作権法において、著作物は「思想または感情が創作的に表現されたもの」（著作権法２条１項１号）と規定している。著作物である限り、諸権利に触れる行為を無断で行なうことはできない。そのため、言語資源が著作物であるかどうかを判定することが重要になる。

著作物にはデジタル化されていないもの(図 3.3-3 の(a))、デジタル化されたもの(図 3.3-3 の(b))とが存在するが、形態に問わず著作権で保護される。

政府が刊行する白書にも著作権は存在する。ただ、禁転載と明示されていなければ転載は可能である。報道記事は事実の部分は著作物ではないが、記事作成者の推測等の部分は著作物である。

②保護される表現の単位、限界

著作物について保護対象となる表現の単位、著作物性なしの限界は必ずしも明確ではない。言語資源の利用に際しては単語のレベルでは著作物性はないとするのが一般的と考えるが、選挙の立候補予定者名簿に付与した記号に著作物性を認めた裁判例も存在し、判定は難しい。

辞書では単語自体には著作物性は認められないが、単語の選択や配列に創作性があれば編集著作物としての権利が存在する。

③言語処理に際して行なわれる行為

処理の結果として出力される二次データが元の言語資源の複製、翻訳、翻案物であるかどうかで判定できる。処理途中で一時的にメモリ上に格納されることを「複製」とするかどうかについては欧州では著作権法上複製ととらえる考え方があるが、日本においては複製と解釈されていない。現状ではこのようなメモリ上の過渡的、瞬間的な蓄積は複製とはしないと判断してよいだろう。しかし、ＨＤＤへの蓄積については、現状のわが国での法解釈論では、複製としない考え方はないので注意が必要である。

また、行為よりも行為を経て得られたアウトプットの形で判断するべきである。

④デジタルデータ化

言語資源のデジタルデータ化(図 3.3-3 の②)自体は複製行為と考えるのが適当である。何らかの創作行為が介在するならば二次的著作物を作成する行為となるが、通常のデジタルデータへの変形は創作行為とはならない。著作権の権利制限が規定された複製がいくつか認められている（図 3.3-4)。

⑤解析処理等

解析処理等(図 3.3-3 の③)は、処理結果である二次データの著作権上の評価によって決まる。

⑥二次データ

言語資源が著作物であり、二次データ（図 3.3-3 の(c)及び(d)）が元の言語資源の複製物か二

次的著作物かであるならば、その二次データには元の言語資源の著作権者の権利が及ぶ（二次的著作物には二次データを作り出した者の権利も生ずる)。

二次データ自体の評価は、元データが推知、感得できるかどうかが問題とされ、同じ単語が現れているといったことだけで複製物とは判定されない。

⑦二次データの公表

二次データが複製物、二次的著作物である場合には、利用において著作権者の許諾が必要である(図 3.3-3 の④)。著作者の権利を図 3.3-5 にあげる。

権利に抵触する行為であっても、研究論文での引用や学会発表での上映（著作権法 38 条：非営利かつ無償での上映等）等では許諾を要しない。

⑧ウェブページに関する著作権上の問題

ウェブページは著作物と考えてよい。ウェブのリンク先表示自体は著作権の侵害にはあたらないが、フレームでコンテンツ自体を流し込んでしまう場合は、見る者に著作物の一部を構成するものとの誤解を与え、同一性保持権の問題が生ずる可能性がある。あるいは氏名表示権を主張される場合もありうる。

また、サーチエンジン等で出典（リンク）と要約だけをのせるケースでは、要約が厳密に言えば著作権上の権利侵害になりうるが、黙示の許諾とも考えられる（ウェブ上で公開すれば検索されることを前提として考える)。

⑵ 共通設問への見解概要

いずれの設問も講演で述べた３ステップで基本的な判断ができる。ただあくまでも目安であり、確定的な解ではない。

①商用、研究用の別について

「目的が商用、研究用での違い」について設定しているが、著作権上の差異はない。商用、研究用よりも私的利用かどうかが意味を持つ。また企業の研究所は商用と考えるべきだろう。

②単語分割、統計情報取得について

基本的に著作権に抵触しない行為と考えてよいだろう。

③人手とロボットでの別

ロボットを操作しているのは人間であり、人間が行なった行為と本質的に変わらない。

④用語集について

用語集が編集著作物である可能性は否定できない(用語の選択・配列に創作性がある場合)。「辞書に加える」行為が、あるまとまりで選択、配列が推知できる場合は「複製権」、「同一性保持権」の問題となる。単語単位に切り分けて取り込んだ場合は問題ないと考える。

元のデータからスーパーセットを作る場合は元のデータの配列を変えれば問題ないし、新たに集めた部分は問題にならない。サブセットの場合は判断が微妙である。

⑤検索システム、自動要約結果の表示に関して

検索システムにおいて本文を部分的に表示すること､元文書と代替性のある要約では同一性保持

権の問題に抵触する可能性がある。但し、「やむを得ない」改変とも考えられる。

⑥研究用ライセンスに関して

ライセンスで禁止されていないか、著作権侵害とならなければ問題がない。研究用ライセンスを持っている辞書を用いて研究用プログラムを作成し、同プログラムを用いた自動学習で商用辞書を作成することについては、研究プログラムの作成は、元の辞書の「表現」が研究用プログラムに使われていない限り、問題はない。仮に元の辞書の「表現」をそのまま使い、元の辞書の著作権を侵害する研究用プログラム（これ自体が著作権侵害）を用いて、新たな辞書を作ったとしても、元の辞書の著作権は及ばない。

３．３．６ 分析、考察

(1) 個別分析

① ヒヤリング 1(木村氏講演)

まず、第一回目にヒヤリングした木村氏の講演ついて分析する。これは、言語資源の共有を目的とする非営利団体 GKS が、プロバイダからコンテンツをスムーズに提供してもらう為の施策を検討した活動の報告であった。そこでは、プロバイダとの信頼関係の確立がその鍵であると結論づけた上で、その為の具体的な要件が示された。GSK の立場は、あくまで言語コーパスのブローカー(仲介者)に徹するというものであり、今回の調査の設定とは立場の異なるものである。よって、その結論をそのまま今回の調査に適応できるわけではない。しかし、本講演は、本ワーキンググループメンバーの言語コーパスに関する著作権の理解に大きく寄与し、特に、言語コーパスがどのように活用のされるかをプロバイダに説明すことが、相互理解に有益であったとの知見は、本調査にとっても示唆に富む指摘である。また、将来の課題として提示された「不特定の発信者を持つ言語資源の著作権」の問題は､今回の調査と完全に合致する問題意識であった。

次に、共通設問に対する木村氏の見解について分析する。同氏は、原データへの復元性を有しない二次データについては基本的に著作権上の問題は生じないとの考えを示されたが、これは本調査でも追認したい見解である。一方、新聞等からの新語の抽出に関して、同氏は慎重な見解を示された。これは現在の企業を取り巻く環境から、配慮されるべきマージンをとった上での見解であるとも推測できるが、例えば、孤立した単語自身の著作物性有無の観点からは、機械処理による新語抽出などに関しては､もう少し広い許容範囲が設定可能であるようにも感じられた。

② ヒヤリング 2(藤波氏講演)

次に、第二回ヒヤリングの講師である藤波氏の見解についてであるが、氏の基本的な考え方は、法、特に財産権法は両者のバランスを衡量して判断されるという考え方である。したがって不必要に著作権者に遠慮しすぎる必要はなく、「フリーライド」は原則自由であるとの見解である。そして例外的にフリーライドを禁止する規定が知的財産

(権)法であるが、問題は、個々のケースを規定に照らしあわせても明快に判定することが困難な場合が多いことであり、さらには、専門家でも均一な判断が下せるわけではないとの指摘は、著作権問題の難しさを認識させられる。そのため極論すれば多くの事例がグレーとなり常に著作権侵害の可能性を秘めているとも言える。以上から、講師の結論としては、「社会通念に照らし」「適法な」利用であれば、異説はあるとしても、許容されるとしている。

共通設問に対する見解においても、法律家として社会的公共性のある行為を許容する立場からの見解と思われ、「勝てる事例」「抗弁可能」との見解も示された。この立場はある意味で、ヒヤリング１の木村氏と対照的であり、最終的に勝てるにせよ、論争になること自体を嫌うような利用者にとっては取りにくい立場とも言える。リスクを正しく認識して可否判断することが必要であろう。

③ ヒヤリング 3(亀井氏講演)

最後に、第三回目にヒヤリングを行なった亀井氏の見解に関してであるが、言語資源が著作物であり、処理行為の過程で著作権に触れる行為であるか、行為を適法とする権利制限規定があるかの３点で基本的な著作権者許諾の判断ができるという見解であり、著作権に関して明るくないものにとっても判断のしやすい見解が示された。また、「著作物であるかどうか」も言語の単位によって段階があることや、行為自体よりも行為によって得られた結果が問題であり、メモリ上での過渡的複製は抵触しないとみなす点など、実使用時の条件によって抵触の判断が異なることが示された。ウェブという新しいメディアの著作物性、その状況によって著作権上留意する点があることもあわせて指摘された。

共通設問においても氏が示された３点による判定と、３点それぞれにおいて実使用によって解釈が変わる点への補足説明がなされた。著作物であっても、それが単語や統計情報になれば抵触しないこと、人手とロボットでは本質的に変わりがないこと、商用、研究用という違いは著作権上意味がない点などがあげられた。ウェブ上のデータに関しては、厳密にいえば抵触となる可能性のあるケース（要約やフレームでのコンテンツの流し込み）があるが、黙示の許諾とも考えられるという見解が得られた。ただ、これらの新しいメディアやその利用方法については判例が少なく、解釈がまださだまらない状況であることもわかった。

(2) 比較分析

以上の各講師の立場から示された設問への見解(表 3.3-1～3)を見ていくと、共通の判断が得られている箇所が見て取れる。設問(a)～(c)をウェブページから統計情報を取得する一連の計算機動作と考えると、単語分割(b)と統計情報取得(c)に関しては著作権侵害にはあたらないとの見解が得られている。しかし、その前段階の HDD への保存(a)に関しては、木村氏の見解では一時複製ならば適法、藤波氏も他説ありとの注釈付きながら個人見解としては適法としている一方で、亀井氏は非私的利用は侵害

との見解である。さらに、亀井氏は、3.3.5 節の講演本文で「メモリ上の過渡的瞬間的な蓄積は複製とはしない」ものの「HDD への蓄積については、現状のわが国での法解釈論では複製としない考え方はない」と前二氏の見解と大きく異なった見解を提示している。また、藤波氏が一時複製に関して「他説あり」と述べている点、木村氏が「無断複製を禁じているページは対象とせず」と述べている点に注意が必要である。

また、単語抽出に関しての設問(f)～(g)に関しても侵害にはあたらないとの見解が大勢である。ただし、木村氏は「新聞は普通無断複製を禁じているので NG」としており、利用既定に複製禁止の表示のあるページに関しては注意が必要である。また、機械的な抽出過程において HDD に一時保存をする場合には、藤波氏、亀井氏が指摘しているように一時的複製そのものを侵害とする説もある点も注意が必要である。

一方、ウェブから得た情報の表示(j)～(m)に関しては、一部見解に差異が見られる。本文の部分表示(j)、要約(k)に関しては、藤波氏は実体法論からは違法なれど抗弁可能との見解で、亀井氏は侵害の可能性ありとの見解である。一方、情報抽出しての結果抽出(l)、Q&A システムでの利用(m)に関しては、藤波氏は基本的に OK、木村氏は Q&A の検索対象としてなら OK(m)、作るのは OK(l)とある。つまり、表示される結果が事実数値などならば OK で部分表示や要約などの非事実数値ならば侵害の可能性ありとの見解があるものの、「抗弁可能」「やむを得ない改変」との見解もあり、また情報抽出に関しても「グレー」との見解もあるなど、差異が見られる。ただ、三者共通しているのは、侵害の有無は要約や情報抽出の処理過程に依存するのではなく、出来上がった出力内容によって判断されるという観点である。この観点は亀井氏の講演中にも「行為よりも行為を経て得られたアウトプットの形で考えるべきだろう」と述べられている。

研究用ライセンスについては、三講師とも契約内容に依るとの前提ながら統計情報や単語の抽出に関しては OK との見解である。一方(p)(q)の設問に関しては、設問の前提条件があいまいだったためか見解に共通性を見出すことが難しい。許諾を得ずとも OK との見解が示されている統計情報や単語の抽出とは異なり、やはりライセンス契約に依拠した判断になることから、一般論としての判断は困難だったと考えられる。

(3) 結論

以上の各講師の見解の比較結果から、言語資源利用時に著作権上留意すべき点として以下があげられる。

①単語、統計情報のレベルでは侵害にあたらないと考えてよい。文のレベルでも著作物性を否定する有力な学説が存在し、文単独では著作物性は認められない可能性もある。辞書についても、単語自体に著作権はなく、単語の配列、選択に著作権上の権利が存在する。

②複製に関しては、複製行為の結果としての二次データが著作権を侵害しないデータの場合で、メモリ上の一時的、過渡的な複製であれば著作権に抵触しない説が有力である。しかしＨＤＤへの複製は、わが国での法解釈論では複製としない考え方はないという見解と、複製としない説があるという

見解とが存在し、現状では抵触、非抵触の判断はできない。メモリとＨＤＤとの違いの問題も含めて、今後の判例に注目しつつ、専門家への継続的な調査、ヒヤリングが必要である。

③複製をはじめ、著作物である言語資源に対する行為では、行為自体よりも行為によって得られた二次データの性質や表現が抵触か否かの判断となる。また行為自体の労力の大きさ（労力がかかる行為は侵害にあたらない等）は抵触か否かの判断に関係しない。

④ウェブの情報自体には著作権が存在すると考える。ウェブから得た情報の部分表示、要約等においては、著作権侵害の可能性があるものの、「黙示の許諾」あるいは「抗弁可能」という見解があり、著作権上大きな問題にならない可能性もある。ウェブのデータから得られた単語、統計情報は前述の①と同様著作権上問題にならない。

⑤人手とロボットを用いた場合での行為に関しては著作権上の判定においてその本質は変わらない。

⑥商用、研究用の別は著作権法上意味がない。米国でいう「フェアユース」も日本においては無関係である。

⑦言語資源の使用許諾に関するライセンス契約がある場合には契約内容によって行為や二次データに関する使用の範囲が決まる。

著作権に抵触する行為か否かの判断において、亀井氏が提示した３ステップ（言語資源が著作物であるか、行為が抵触しないか、著作権制限規定があるか[9]）が、著作権になじみのない者にも明確でわかりやすい判断方法ではないか。この判断基準に加え、各講師が示された個々の留意点を念頭において、言語資源を利用していくのが望ましいといえる。

今回は講師の方々に言語資源利用時の具体的な利用方法を共通設問として提示し、回答を得る方法を用いた。言語資源利用時の個々の具体的な状況を提示することで、実際の利用に即した回答を得ることができたといえる。たとえば単純に「言語資源を複製すること」自体は違法（著作権制限規定に該当せず、言語資源が著作物である場合）である。しかし、言語資源を用いた研究開発を行なう者が実際に行なっている行為や状況を具体的に提示することで、権利抵触しない範囲での利用方法や留意点を明確にできる可能性があることがわかった。また、設問の状況説明自体が曖昧だった場合には明確な回答を得られないケースがあり、今後調査を継続する際には、具体的な状況を説明することが有意義な回答を得るのに必須であることがわかった。

３．３．７ おわりに

本節では、今年度から新たに開始した「言語コーパスにおける著作権に関する調査」について報告した。ここではまず、論点を明確化するために、言語コーパスの具体的な利用状況を複数設定し、これに基づく想定質問集を作成した。その後、立場の異なる３方の講師を招き、計３回のヒヤリングを実施した。そして、ヒヤリング結果の分析によって、著作権の観点から言語コーパスを適切に利用する為の留意点を検討した。

今後の課題としては、まず、**(1)ヒヤリング対象の拡大**が挙げられる。今回の調査では、基本的には、言語コーパスを利用して、自然言語技術の研究開発を行なうユーザ側の立場にある方からのヒヤリン

グを行なった。今後は、言語コーパスを提供する側あるいは、言語コーパスの元になる言語データの著作権者や、さらには、言語コーパスを利用して研究開発された技術を利用する立場の方へのヒヤリングが必要であろう。さらに、著作権とそれを取り巻く状況は刻一刻と変化している。この変化を適切に捉えるためには、**(2)継続的な調査**が欠かせないことは言うまでもない。また、今回の調査では、上述の知見を得たが、これらを**(3)言語コーパスの利用のための著作権に関するガイドライン**として、体系的にまとめることが必要である。そして、本調査の長期的な目標としては、言語コーパスを収集・作成・配布あるいは利用する際に締結すべき契約のための**(4)標準契約書案の提案**を挙げることができる。

謝辞

今回の調査の趣旨をご理解いただき、適切なレクチャとご教示を頂いた、株式会社東芝の木村和広氏、ならびに、NTT アドバンステクノロジ株式会社の藤波進氏、ならびに富士通株式会社 法務・知的財産権本部 亀井正博氏に、言語資源技術委員会著作権ワーキンググループメンバー一同、心よりの感謝の意を表します。

引用・参考文献

1) NIKKEINET 記事,ネット上の課税・著作権保護でＩＴＵが国際基準提唱へ,日経新聞社,2002 年 11 月 11 日, http://www.nikkei.co.jp/news/keizai/20021114AT1FI015313112002.html

2) Impress 記事, デジタルミレニアム著作権法違憲見解/ACM,2001 年 8 月 14 日,
http://www.watch.impress.co.jp/internet/www/article/2001/0814/acm.htm

3) 著作権の保護期間、20 年延長は憲法に違反せず　米連邦最高裁
http://www.mainichi.co.jp/digital/network/archive/200301/16/11.html

4) GSK(言語資源共有機構),「著作権 WG 活動報告資料」,言語資源共有機構
(GSK,http://tanaka-www.cs.titech.ac.jp/gsk/gsk.htm),2000 年 6 月 23 日.

5) ELRA(European Language Resources Association)
http://www.icp.inpg.fr/ELRA/

6) 裁判所：知的財産権判決速報,
http://courtdomino2.courts.go.jp/chizai.nsf/$About

7) 裁判所：知的財産権判例集
http://courtdomino2.courts.go.jp/chizai.nsf/$Help

8) NIKKEI.NET 記事、音楽ファイルの無竜交換サービス、著作権侵害を認定・東京地裁, 2003.01.30,
http://it.nikkei.co.jp/it/news/index.cfm?i=2003012910272j0

9) 文化庁
http://www.bunka.go.jp/

(1)原著作権者の許諾を得ずにやっていいことは？

(a) ロボットでウェブページを集めて HDD に保存すること

(ロボット: 機械的にウェブページを解析し、次々にリンクを辿ってウェブページをダウンロードするプログラム)

(b) ロボットで集めたウェブページを単語分割すること

(c) 単語分割されたウェブページから統計情報を取得すること

(統計情報: 例えば、ある語の使われている頻度情報)

(d) (a)-(c)の "ロボットで" を "人手で" に変えたら状況は変わるか？

(e) ウェブページに公開されている用語集を、自分の辞書に加えること

(f-1) 紙の新聞を人間が読んでいて、新語を発見->辞書登録

(f-2) web の新聞を人間が読んでいて、新語を発見->辞書登録

(g-1) 紙の新聞を機械が解析し、機械的に新語を発見->辞書登録

(g-2) web の新聞を機械が解析し、機械的に新語を発見->辞書登録

(h) google から使用許諾を得て、google API の検索結果から統計情報抽出

(google API: 検索サイト google が機械的にアクセスできるように公開されている。ここにアクセスすると、例えば、ある語で検索される web ページ数や、検索された web ページの要約などを得ることができる)

(i) google から使用許諾を得て、google API の検索結果から新語抽出->辞書登録

(j) 検索システムにおいて本文を部分的(例えば先頭から 100 文字)に表示すること

(k) 自動要約システムでウェブページの要約を表示すること。

(自動要約システム: 一般にディスクへの保存、単語分割などを行なった上で、重要部分の抜き出しなどをします)

(l) 情報抽出システムでウェブページから得た情報を表示すること。

(情報抽出システム: 例えば､各販売店での PC の販売価格をウェブページから収集し一覧表を作成します。これも、一般にディスクへの保存や単語分割などが行なわれます)

(m) Q&A システムのための知識源として、ウェブ上のデータを用いること

(Q&A システム: 例えば「小泉首相は何代目首相?」といった質問に対し、該当しそうなウェブページから答えを見つけるシステム)

図３．３－１　各講師に提示した共通設問(1/2)

2)研究用ライセンスの上でやっていいことは

(n) 研究用ライセンスを持っているコーパス(例えば新聞)から統計情報を抽出し、商用システムに利用すること

(o) 研究用ライセンスを持っているコーパスから単語を抽出し、商用辞書を作成すること

(p) 研究用ライセンスを持っている辞書を xx%程度改変し、商用システムに利用すること

(q) 研究用ライセンスを持っている辞書を用いて研究用プログラム(このプログラム自体は販売などはしない)を作成し、同プログラムを用いた自動学習(例えば新聞からの単語抽出)で商用辞書を作成すること

図３．３－２　各講師に提示した共通設問(2/2)

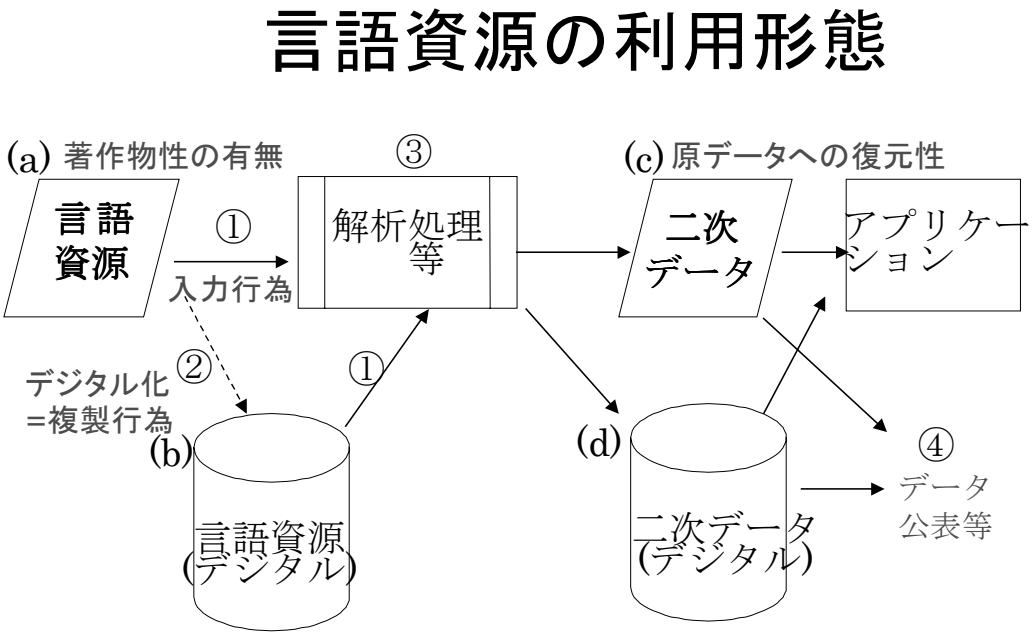


図３．３－３　言語資源の利用形態（GSK 著作権 WG 報告資料より転載）

- 私的使用のための複製（著作権法 30 条：個人的、又は家庭内等での利用目的）
- 図書館等における複製（著作権法 31 条：利用者の求め、保存目的等）
- 引用（著作権法 32 条）
- 教科書用図書等への掲載（著作権法 33 条）
- 学校その他教育機関における複製（著作権法 35 条：授業過程での使用目的）
- 試験問題としての複製（著作権法 36 条）
- 点字による複製（著作権法 37 条）
- 時事問題に関する論説の転載等（著作権法 39 条）
- 政治上の演説等の利用（著作権法 40 条）
- 時事の事件の報道のための利用（著作権法 41 条）
- 裁判手続等における複製（著作権法 42 条）
- 放送事業者による一時固定（著作権法 30 条）

図３．３－４　複製が認められる著作権の権利制限

- 複製権（著作権法 21 条）
- 上映権（著作権法 22 条の 2：公衆に直接見せることを目的とする映写）
- 公衆送信権（著作権法 23 条：公衆に向けての放送、ネットワーク送信）
- 口述権（著作権法 24 条：公衆に直接聞かせることを目的とする口述）
- 譲渡権（著作権法 26 条の 2：譲渡による公衆への提供）
- 貸与権（著作権法 26 条の 3：貸与による公衆への提供）
- 翻訳、翻案権（著作権法 27 条）

図３．３－５　著作権者の権利の例

表３．３－１ 共通設問への各講師の見解(1/3)

				木村氏	藤波氏	亀井氏
1	原著作権者の許諾を得ずにやっていいことは？	a	ロボットでウェブページを集めてHDD に保存すること	NG、複製行為だから。但し、(b)や(c)のために一時的に保存するのは OK	OK。HDD に複製することが直に違法になるわけではありません一時的複製、止むを得ない複製は適法行為です（他説あり）。	1．ウェブページは一応著作物であると考えて差し支えない。2．HDD への保存行為は複製と評価できる。3．「私的使用のための複製」に該当すれば適法。
		b	ロボットで集めたウェブページを単語分割すること	無断複製を禁じているページは対象とせず、かつ、(c)等のための一過程として行うなら OK（＊）但し、原ページ、単語分割結果は速やかに消去する（＊）原データの加工、改変にあたると解釈するとNG になる	OK。自由です。	1．ウェブページは一応著作物であると考えて差し支えない。2．（「分割」という行為が理解しにくいが）単語に分割する行為それ自体は、財産権（著作権）に抵触する行為ではないと思われる。著作者人格権（同一性保持権）にも抵触しないと考えてよいだろう。（元の著作物の推知性が問題となる。すなわち、単語単位に分解すれば、個々の単語から元の著作物を推知することはできないだろう。）
		c	単語分割されたウェブページから統計情報を取得すること	OK。二次著作物でないだろうが基本は、それぞれの許諾条件	OK。自由です。	1．ウェブページは一応著作物であると考えて差し支えない。2．統計情報を取得することは、著作権には抵触しない。
		d	(a)-(c)の ″ロボットで″ を ″人手で″ に変えたら状況は変わるか？	変わらない	特段の事情がある場合を除き、原則変りません。	全く状況は変わらない。ロボットをしてHDDに保存せしめているのは、人間であり、本質的には変わらない。
		e	ウェブページに公開されている用語集を、自分の辞書に加えること	著作権保護の例外にあたるケース(個人ユース、教育ユースなど)なら OK。これは(1)の設問全般についていえる	OK。当該用語が著作物でなければ、自由に使用できます。	1．用語集が編集著作物である可能性は否定できない。（用語の選択・配列に創作性がある場合）2．「辞書に加える」という行為の評価による。自分の辞書の一部に、そのまま（ひとまとまりで、あるいは元の選択・配列が推知できるような程度のまとまりで）取り込む場合、複製権と、同一性保持権（他人の作った編集物の一部に見えることから）の問題となり得る。単語単位に切り分けて取り込んだとすると問題はないと考える。（元の用語集＝編集物の推知性の問題と思われる）3．「私的使用のための複製」に該当すれば適法。
		f-1	紙の新聞を人間が読んでいて、新語を発見->辞書登録	OK	OK	1．単語それ自体は一般に著作物ではない。（造語について、主張はあり得るが保護は困難？）
		f-2	web の新聞を人間が読んでいて、新語を発見->辞書登録		OK	

表３．３－２ 共通設問への各講師の見解(2/3)

		g1	紙の新聞を機械が解析し、機械的に新語を発見->辞書登録	新聞は普通無断複製を禁じているので NG	OK	1．単語それ自体は一般に著作物ではない。（造語について、主張はあり得るが保護は困難？）
		g2	web の新聞を機械が解析し、機械的に新語を発見->辞書登録		OK	
		h	(google 自体の使用許諾は得て)google API の検索結果から、統計情報抽出	？	OK	1．ウェブページは一応著作物であると考えて差し支えない。2．統計情報を取得することは、著作権には抵触しない。
		I	(google 自体の使用許諾は得て)google API の検索結果から、新語抽出->辞書登録	ソフトが行なう行為自身がとがめられることはないと思うが、ソフトの出力結果を著作権者に無断で公表してはいけない	OK	1．単語それ自体は一般に著作物ではない。（造語について、主張はあり得るが保護は困難？）
		j	検索システムにおいて本文を部分的に表示すること	？	実体法論からは違法です。しかし、可罰的違法性論などからの抗弁が可能です。通常は勝てる事例でしょう。	（1．検索対象となった文章が著作物であることが前提。）2．部分的に表示することは、同一性保持権の問題となり得る。3．ただし、「やむを得ない」改変ではないか。
		k	自動要約結果を表示すること	？	実体法論からは違法です。しかし、運用態様によっては可罰的違法性論などからの抗弁が可能です。	（1．要約対象の文章が著作物であることが前提）2．元の文章の代替性がある程度の要約は、翻案権を侵害するものとされる。同時に、同一性保持権の問題ともなる。書誌事項のみであれば権利には触れない。3．翻案権に触れる場合、「私的使用のための複製」に該当すれば適法。
		l	情報抽出をして結果を表示すること	それをしないと意味が無い。作るのは OK、個人利用は OK,サービス提供はグレー＝怪しい	OK。ウェブから得た情報が著作物でない（ex.事実数値）ならば自由に利用できます。但し、サービス態様によっては不競法の検討が必要です。	？？
		m	Q&A システムのための知識源として、ウェブ上のデータを用いること	QA 検索の検索対象が web という意味なら OK。でも QA 検索を使って、例えば FAQ 集みたいなものを作ってしまうと NG	OK。自由です。	（1．利用する「ウェブ上のデータ」が著作物であることが前提）2．「用いる」の内容による。他人の作った文章の全体もしくは一部を、無断で自分の文章の一部に使うことは、少なくとも複製権の侵害となる。3．「私的使用のための複製」に該当すれば適法。

表３．３－３ 共通設問への各講師の見解(3/3)

2	研究用ライセンスの上でやっていいことは	n	研究用ライセンスを持っているコーパスから統計情報を抽出し、商用システムに利用すること	OK.統計情報など原データへの復元性を有さない二次データの権利は、二次データ作成者に帰属すると思うから	OK。特許等に抵触しない限り、アイディアの抽出や利用は自由です。	ライセンス違反でなければよい。(統計情報の抽出については1(h)参照)
		o	研究用ライセンスを持っているコーパスから単語を抽出し、商用辞書を作成すること	たぶん OK	OK。単語は著作物ではありません。因って自由に利用できます。	ライセンス違反でなければよい。(統計情報の抽出については1(h)参照)
		p	研究用ライセンスを持っている辞書を xx%程度改変し、商用システムに利用すること	改変は NG	契約内容に因る。辞書(DB)を構成する著作物とデータベースの著作物についての考察が必要です。	ライセンス違反は別として、1. 辞書は、単語の選択・配列に創作性があれば、編集著作物として保護される。2. 無断で行う改変、および商用システムへの利用は、複製もしくは翻案、各利用権(公衆送信他)の侵害となる。
		q	研究用ライセンスを持っている辞書を用いて研究用プログラムを作成し、同プログラムを用いた自動学習で商用辞書を作成すること	グレー。元の辞書エントリを組み合わせて複合語辞書を作るようなものはいけない気がするが、未知語を抽出するは良い気がする	OK。研究用プログラムの作成は全く自由。単語は著作物ではありませんので自由に利用できます。作成した辞書の著作権は著作者に帰属します(原則)。	研究プログラムの作成は、元の辞書の「表現」が、研究用プログラムに使われていない限り、問題はない。仮に元の辞書の「表現」をそのまま使い、元の辞書の著作権を侵害する研究用プログラム(これ自体が著作権侵害)を用いて、新たな辞書を作ったとしても、元の辞書の著作権は及ばない。

３．４ 自然言語処理の応用に関するユーザ調査

この節では、自然言語処理の応用に関するユーザニーズ把握のために実施したアンケート調査の内容について報告する。３．４．１節で本調査の目的について述べ、３．４．２節で具体的な調査方法を説明する。３．４．３節および３．４．４節で調査結果の分析によって得られた知見について報告する。

３．４．１ 調査の目的

従来の自然言語処理技術関連の研究開発には、音声認識、対話処理など人間に限りなく近い能力を持つコンピュータの開発を目指した長期的視野にたつものが多い。そのため、研究開発の目標が技術シーズからの発想に偏りがちであり、ユーザニーズに即した技術開発がやや軽視されているきらいがある。

しかし、「自然言語処理において重要なユーザニーズは何か」という疑問に答えてくれる調査データを入手することは困難である。自然言語処理関連の製品を開発または販売している企業では、それぞれ独自に市場調査を実施しているはずだが、これらの調査結果は各企業がもつ貴重なノウハウであり、公開・共有される性質のデータではない。また各企業において、検索、かな漢字変換、音声認識、文字認識などの各製品が属する分野に特化した調査はされているとしても、自然言語処理関連の研究開発の方向性を検討するための横断的なユーザニーズ調査はなされていないと思われる。

そこで、本委員会では産学共同での活動という特徴を活かし、企業内の調査では実施しにくいような、将来の応用技術に関する長期的な視点でのユーザニーズ調査を実施することとした。コスト、実現性などにとらわれない自由な発想で設定した設問によるアンケート調査を実施し、特定の製品群や技術に偏らない汎用的な調査データとし、この結果を共有することで自然言語処理関連の研究開発における活用を狙う。

３．４．２ 調査の方法と実施

前節で述べたとおりこの調査は、開発者の立場ではなく、ユーザの観点からみて（ユーザが求める）近い将来実現可能な自然言語処理技術の応用（製品）や、長期的なテーマだが重要な自然言語処理技術を見い出すことを目的としている。従って、単に既存の自然言語処理技術や機能を列挙した設問とはせずに、ユーザが製品やソフトを直感的にイメージできるような「ユーザシナリオ」を用いたアンケートを実施することとした。アンケート対象者には作成したユーザシナリオを評価してもらう。アンケート調査の手順は以下のとおりである。

（１）ユーザシナリオの作成

（２）ユーザシナリオを評価するアンケートの実施

（１）ユーザシナリオの作成

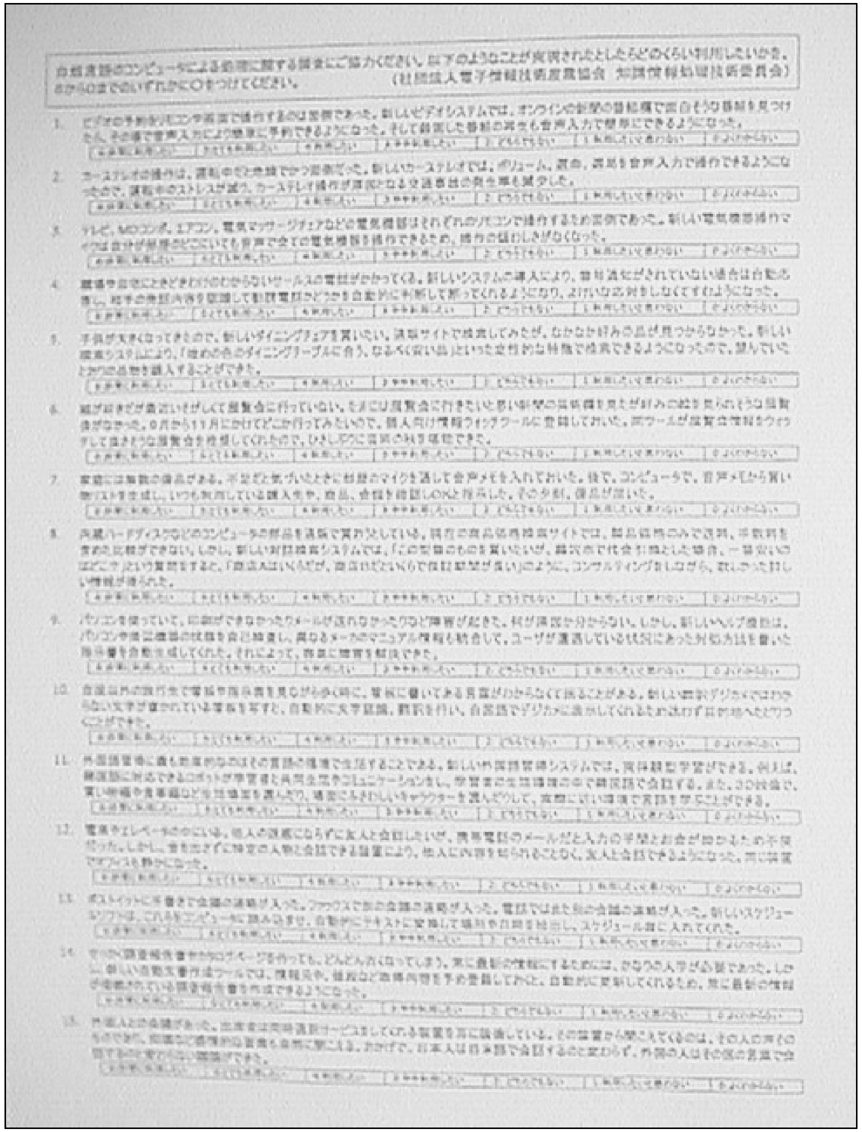
ユーザシナリオとは、「製品やソフトを使うアクター（ユーザ）がいて、具体的な使用シーンが

あり、時系列的なストーリーが書かれたもの」である。まずは、我々が日常生活あるいは業務上において“自然言語処理技術を応用したこんな製品やあんなソフトがあれば、こんなことやあんなことができるだろう！”と思いつくままのシナリオを作成し、以下の点に留意しながらアンケート用のシナリオに修正していった。

- 現時点での実現可能性は重視せず、将来的な展望を含んだシナリオとすること
- 一つのシナリオには部分的であっても必ず自然言語処理技術が応用されていること
- 抽象的ではなく具体的で魅力的なシナリオとなっていること
- 専門用語をわかりやすい表現におきかえること
- シナリオ全体として自然言語処理の技術分野を網羅していること
- シーンが類似しているシナリオはよりわかりやすいシナリオへ統合すること
  （但し処理技術が類似している場合でも、ユーザが遭遇するシーンが異なれば統合しない）
- 知的職業に従事する人のシナリオが多くなりがちなため、より一般的な人を想定したシナリオにコンバートすること

以上の作業の結果、A4サイズ2ページにわたる27問のユーザシナリオとなり、かつフリーインプット欄を設けてユーザからのシナリオも募集するようにした。評価は、個々のシナリオに対し“このユーザシナリオが実現されたらどれくらい利用したいか？”という利用希望度を7段階で示してもらう。また、アンケート対象者のプロフィール設問（性別、年齢、職種）も最後に設けた。以下、ユーザシナリオの分量を把握してもらうためにアンケート用紙のイメージ図を図３．４．２－１に示し、次に実際のユーザシナリオの内容を紹介する。

表



裏

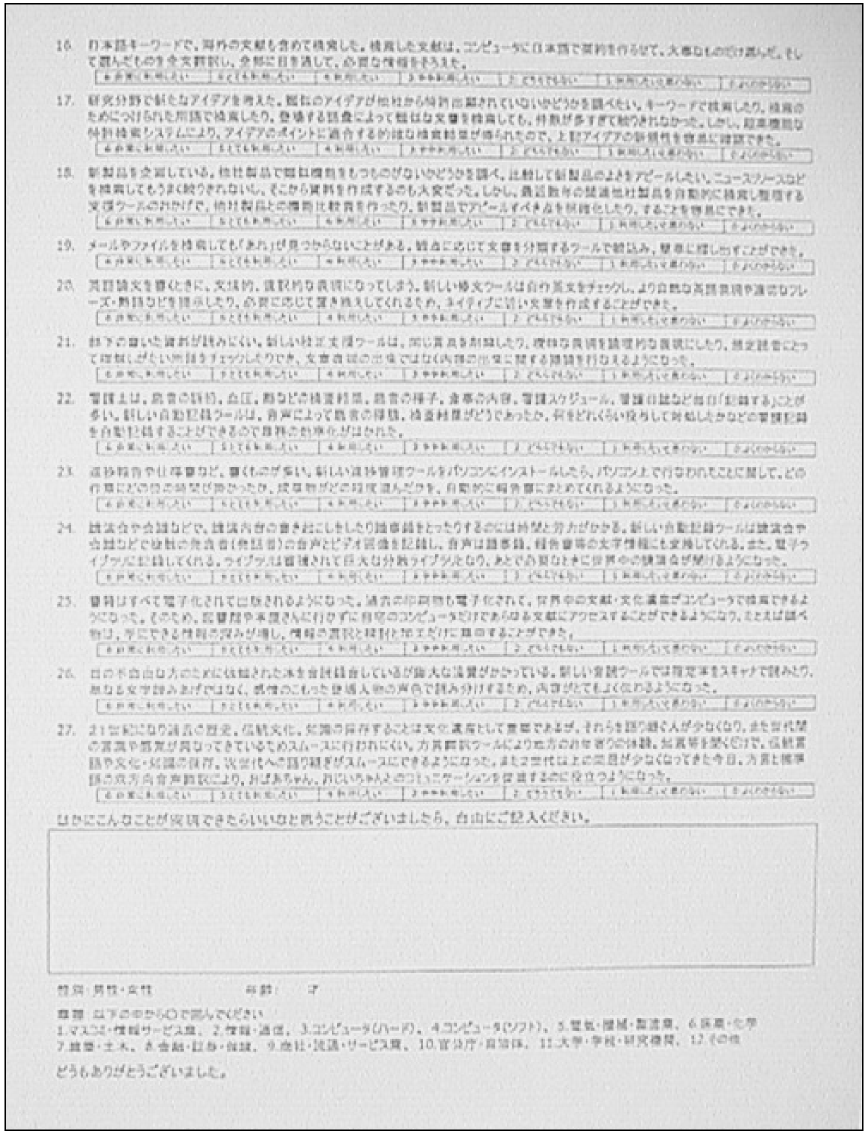


図３．４．２－１　アンケート用紙イメージ

ユーザシナリオの内容　１／３

1. ビデオの予約をリモコンや画面で操作するのは面倒であった。新しいビデオシステムではオンラインの新聞の番組欄で面白そうな番組を見つけたら、その場で音声入力により簡単に予約できるようになった。そして録画した番組の再生も音声入力で簡単にできるようになった。

2. カーステレオの操作は、運転中だと危険でかつ面倒だった。新しいカーステレオでは、ボリューム、選曲、選局を音声入力で操作できるようになったので、運転中のストレスが減り、カーステレオ操作が原因となる交通事故の発生率も減少した。

3. テレビ、MDコンポ、エアコン、電気マッサージチェアなどの電気機器はそれぞれのリモコンで操作するため面倒であった。新しい電気機器操作マイクは自分が部屋のどこにいても音声で全ての電気機器を操作できるため、操作の煩わしさがなくなった。

4. 職場や自宅にときどきわけのわからないセールスの電話がかかってくる。新しいシステムの導入により、番号通知がされていない場合は自動応答し、相手の発話内容を認識して勧誘電話かどうかを自動的に判断して断ってくれるようになり、よけいな応対をしなくてすむようになった。

5. 子供が大きくなってきたので、新しいダイニングチェアを買いたい。通販サイトで検索してみたが、なかなか好みの品が見つからなかった。新しい検索システムにより、「暗めの色のダイニングテーブルに合う、なるべく安い品」といった定性的な特徴で検索できるようになったので、望んでいたとおりの品物を購入することができた。

6. 絵が好きだが最近いそがしくて展覧会に行っていない。たまには展覧会に行きたいと思い新聞の芸術欄を見たが好みの絵を見られそうな展覧会がなかった。９月から１１月にかけてどこか行ってみたいので、個人向け情報ウォッチツールに登録しておいた。同ツールが展覧会情報をウォッチして良さそうな展覧会を推奨してくれたので、ひさしぶりに芸術の秋を堪能できた。

7. 家庭には無数の備品がある。不足だと気づいたときに部屋のマイクを通して音声メモを入れておいた。後で、コンピュータで、音声メモから買い物リストを生成し、いつも利用している購入先や、商品、金額を確認しＯＫと指示した。その夕刻、備品が届いた。

8. 内蔵ハードディスクなどのコンピュータの部品を通販で買おうとしている。現在の商品価格検索サイトでは、製品価格のみで送料、手数料を含めた比較ができない。しかし、新しい対話検索システムでは、「この型番のものを買いたいが、藤沢市で代金引換とした場合、一番安いのはどこ？」という質問をすると、「商店Ａはいくらだが、商店Ｂだといくらで保証期間が長い」のように、コンサルティングをしながら、欲しかった詳しい情報が得られた。

9. パソコンを使っていて、印刷ができなかったりメールが送れなかったりなど障害が起きた。何が原因か分からない。しかし、新しいヘルプ機能は、パソコンや周辺機器の状態を自己検査し、異なるメーカのマニュアル情報も統合して、ユーザが遭遇している状況にあった対処方法を書いた指示書を自動生成してくれた。それによって、容易に障害を解決できた。

10. 自国以外の旅行先で看板や指示表を見ながら歩く時に、看板に書いてある言葉がわからなくて困ることがある。新しい翻訳デジカメではわからない文字が書かれている看板を写すと、自動的に文字認識、翻訳を行い、自国語でデジカメに表示してくれるため迷わず目的地へたどりつくことができた。

ユーザシナリオの内容　2／3

11. 外国語習得に最も効果的なのはその言語の環境で生活することである。新しい外国語習得システムでは、実体験型学習ができる。例えば、韓国語に対応できるロボットが学習者と共同生活やコミュニケーションをし、学習者の生活環境の中で韓国語で会話する。また、３Ｄ映像で、買い物編や食事編など生活場面を選んだり、場面にふさわしいキャラクターを選んだりして、実際に近い環境で言語を学ぶことができる。

12. 電車やエレベータの中にいる。他人の迷惑にならずに友人と会話したいが、携帯電話のメールだと入力の手間とお金が掛かるため不便だった。しかし、音を出さずに特定の人物と会話できる装置により、他人に内容を知られることなく、友人と会話できるようになった。同じ装置でオフィスも静かになった。

13. ポストイットに手書きで会議の連絡が入った。ファックスで別の会議の連絡が入った。電話ではまた別の会議の連絡が入った。新しいスケジュールソフトは、これらをコンピュータに読み込ませ自動的にテキストに変換して場所や日時を抽出しスケジュール表に入れてくれた。

14. せっかく調査報告書やカタログページを作っても、どんどん古くなってしまう。常に最新の情報にするためには、かなりの人手が必要であった。しかし、新しい自動文書作成ツールでは、情報元や、値段など取得内容を予め登録しておくと、自動的に更新してくれるため、常に最新の情報が掲載されている調査報告書を作成できるようになった。

15. 外国人との会議があった。出席者は同時通訳サービスをしてくれる装置を耳に装備している。その装置から聞こえてくるのは、その人の声そのものであり、抑揚など感情的な要素も自然に聞こえる。おかげで、日本人は日本語で会話するのと変わらず、外国の人はその国の言葉で会話するのと変わらない議論ができた。

16. 日本語キーワードで、海外の文献も含めて検索した。検索した文献は、コンピュータに日本語で要約を作らせて、大事なものだけ選んだ。そして選んだものを全文翻訳し、全部に目を通して、必要な情報をそろえた。

17. 研究分野で新たなアイデアを考えた。類似のアイデアが他社から特許出願されていないかどうかを調べたい。キーワードで検索したり、検索のためにつけられた用語で検索したり、登場する語彙によって類似な文書を検索しても、件数が多すぎて絞りきれなかった。しかし、超高機能な特許検索システムにより、アイデアのポイントに適合する的確な検索結果が得られたので、上記アイデアの新規性を容易に確認できた。

18. 新製品を企画している。他社製品で類似機能をもつものがないかどうかを調べ、比較して新製品のよさをアピールしたい。ニュースリリースなどを検索してもうまく絞りきれないし、そこから資料を作成するのも大変だった。しかし、最近数年の関連他社製品を自動的に検索し整理する支援ツールのおかげで、他社製品との機能比較表を作ったり、新製品でアピールすべき点を明確化したり、することを容易にできた。

19. メールやファイルを検索しても「あれ」が見つからないことがある。観点に応じて文書を分類するツールで絞込み、簡単に探し出すことができた。

20. 英語論文を書くときに、文法的、直訳的な表現になってしまう。新しい修文ツールは自作英文をチェックし、より自然な英語表現や適切なフレーズ・熟語などを提示したり、必要に応じて置き換えしてくれるため、ネイティブに近い文章を作成することができた。

ユーザシナリオの内容　3／3

21. 部下の書いた資料が読みにくい。新しい校正支援ツールは、同じ言及を削除したり、曖昧な表現を論理的な表現にしたり、想定読者にとって理解しがたい用語をチェックしたりでき、文章表現の出来ではなく内容の出来に関する議論を行なえるようになった。

22. 書籍はすべて電子化されて出版されるようになった。過去の印刷物も電子化されて、世界中の文献・文化遺産がコンピュータで検索できるようになった。そのため、図書館や本屋さんに行かずに自宅のコンピュータだけであらゆる文献にアクセスすることができるようになり、たとえば調べ物は、手にできる情報の深みが増し、情報の選択と検討と加工だけに集中することができた。

23. 目の不自由な方のために依頼された本を音読録音しているが膨大な浪費がかかっている。新しい音読ツールでは指定本をスキャナで読みとり、単なる文字読みあげではなく、感情のこもった登場人物の声色で読み分けするため、内容がとてもよく伝わるようになった。

24. ２１世紀になり過去の歴史、伝統文化、知識の保存することは文化遺産として重要であるが、それらを語り継ぐ人が少なくなり、また世代間の言葉や感覚が異なってきているためスムースに行われにくい。方言翻訳ツールにより地方のお年寄りの体験、知恵等を聞くだけで、伝統言語や文化・知識の保存、次世代への語り継ぎがスムースにできるようになった。また２世代以上の同居が少なくなってきた今日、方言と標準語の双方向音声翻訳により、おばあちゃん、おじいちゃんとのコミュニケーションを促進するのに役立つようになった。

25. 書籍はすべて電子化されて出版されるようになった。過去の印刷物も電子化されて、世界中の文献・文化遺産がコンピュータで検索できるようになった。そのため、図書館や本屋さんに行かずに自宅のコンピュータだけであらゆる文献にアクセスすることができるようになり、たとえば調べ物は、手にできる情報の深みが増し、情報の選択と検討と加工だけに集中することができた。

26. 目の不自由な方のために依頼された本を音読録音しているが膨大な浪費がかかっている。新しい音読ツールでは指定本をスキャナで読みとり、単なる文字読みあげではなく、感情のこもった登場人物の声色で読み分けするため、内容がとてもよく伝わるようになった。

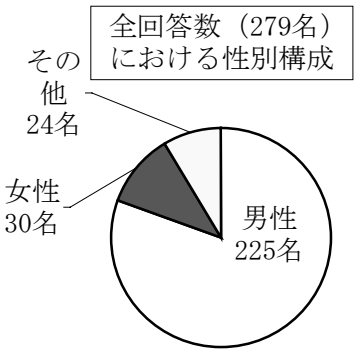
27. ２１世紀になり過去の歴史、伝統文化、知識の保存することは文化遺産として重要であるが、それらを語り継ぐ人が少なくなり、また世代間の言葉や感覚が異なってきているためスムースに行われにくい。方言翻訳ツールにより地方のお年寄りの体験、知恵等を聞くだけで、伝統言語や文化・知識の保存、次世代への語り継ぎがスムースにできるようになった。また２世代以上の同居が少なくなってきた今日、方言と標準語の双方向音声翻訳により、おばあちゃん、おじいちゃんとのコミュニケーションを促進するのに役立つようになった。

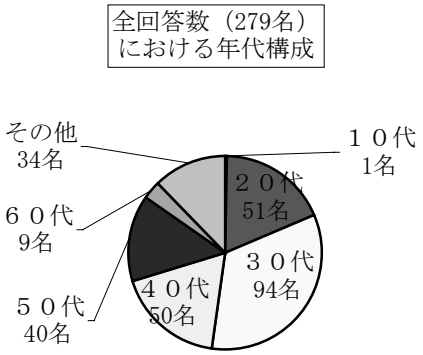
（2）ユーザシナリオを評価するアンケートの実施

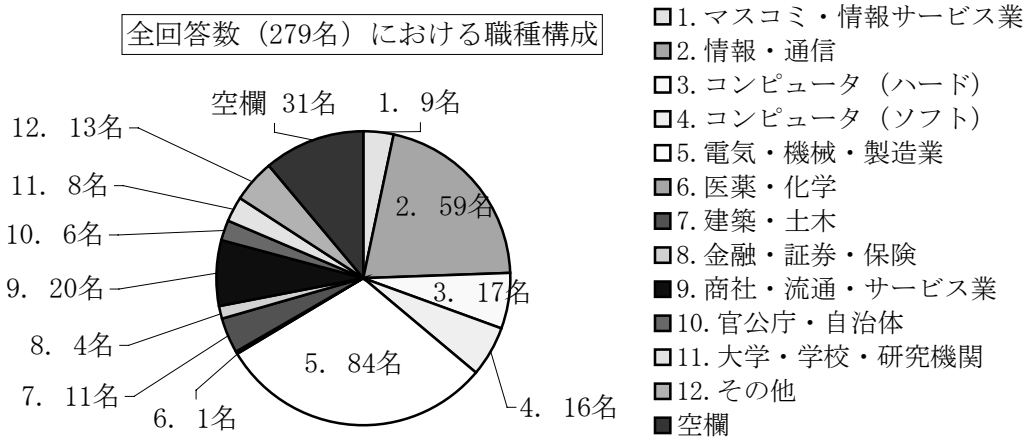
アンケートの実施方法については、「委員会会員の協力による各企業への配布」、「産業イベント等における配布」、「調査会社への発注」などが考えられたが、今後もアンケート調査を継続、拡大していくつもりであるため、今年度は比較的小規模でパイロット的な実施と位置づけ、まずはJEITA関連イベントでアンケートを配布することとした。

平成14年10月1日～5日に幕張メッセで開催されたCEATEC JAPAN 2002（参加者173,021人（プレス，出展関係者等全て含む））において、展示会場ブースでのアンケート配布、及びJEITA関連特別セッションでの配布を行った。なお、できるだけ多く回収するためアンケートに答えていただいた方に「修正テープ」を景品としてお渡しした。

アンケート実施の結果、展示会場ブースにて93回答、JEITA関連特別セッションにて186回答の合計279回答を回収した。以下はアンケート対象者の性別、年代、職種を集計した結果である。







展示会場ブースを回っているお客様の足を止めてアンケートに答えていただくにはユーザシナリオの分量（1シナリオ語数、シナリオ数）が多かったようであり、今後のユーザシナリオ作成、アンケート調査拡大に向けての反省点となった。

次節からの分析は上記アンケート対象者（男性が圧倒的に多く、女性の場合コンパニオンが多かった）を母集団としての分析であるため偏りがあることは念頭におく必要がある。

３．４．３ シナリオの重要度の分析

以下に、集計結果、解釈、応用に関して示唆するであろうことを述べる。

（１） シナリオの重要度ランク

アンケートの７段階評価の上位２つ、すなわち、「非常に利用したい」、「とても利用したい」にチェックをした人の比率を求めた。以下は、その比率でシナリオを順に並べたものである。

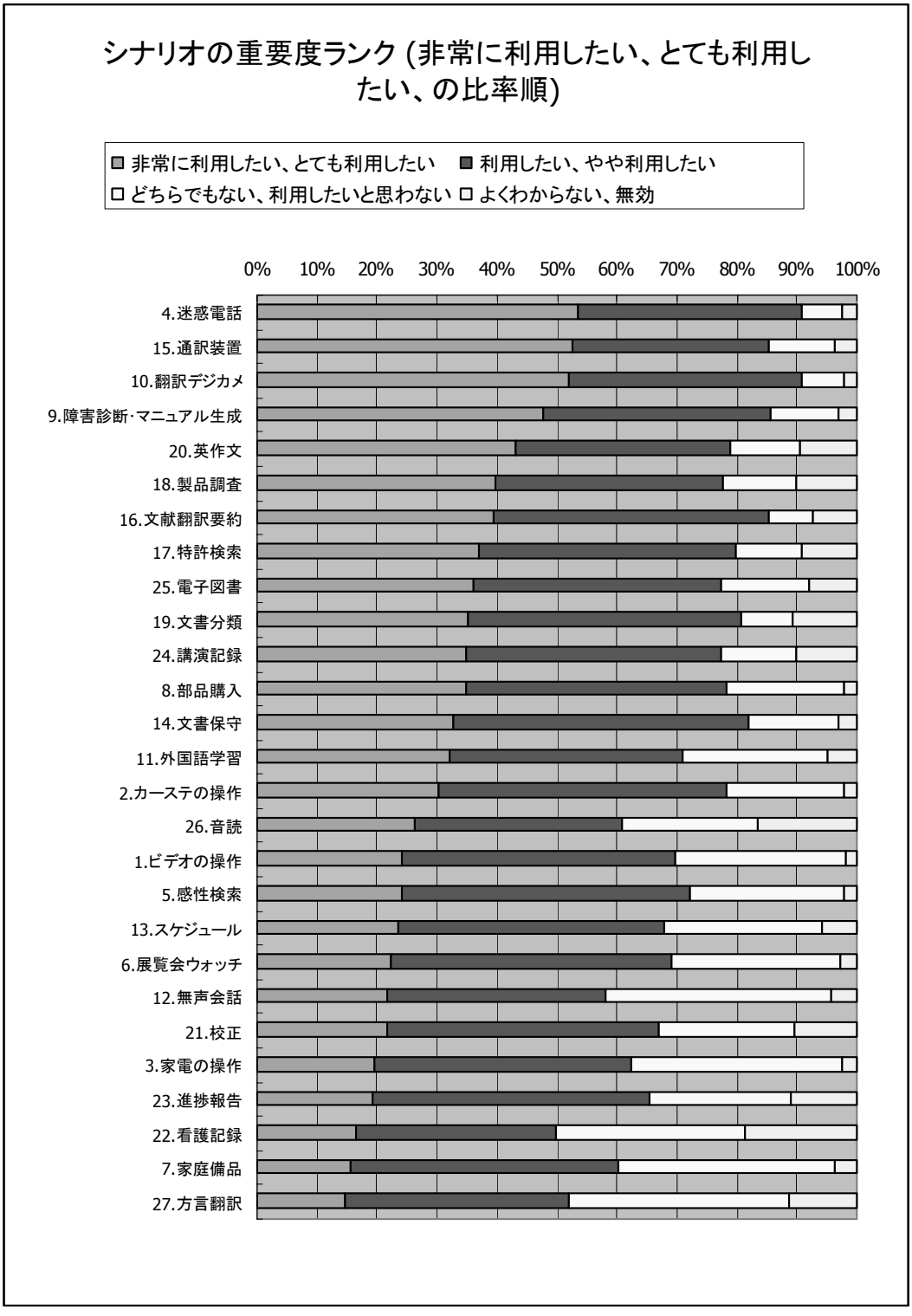


図３．４．３－１　シナリオの重要度ランク

上位は、以下のとおりであった。

１位　4.迷惑電話

２位　15.通訳装置

３位　10.翻訳デジカメ

４位　9.障害診断・マニュアル生成

５位　20.英作文

これらを見ると、ユーザが興味を持つものは、困った状況が手に取るようにわかるようなシナリオであることがわかる。

自然言語処理の応用にユーザが価値を見出して、実際に利用されていくには、ユーザがどういう問題を抱えているかを具体的に絞り、それをいかに解決できるのか、という観点が重要である。

（２） タスクでみた重要シナリオ

シナリオのタスクを、コンピュータと人間の情報の流れで以下の３つに区分する。

１．人間と人間：いわゆるコミュニケーション

２．コンピュータから人間：情報アクセスなど

３．人間からコンピュータ：入力・認識やＡｕｔｈｏｒｉｎｇ

以下は、重要度を縦軸にして、シナリオを３区分に沿ってプロットする。横軸は見やすさのためにずらしただけで数値的な意味はない。

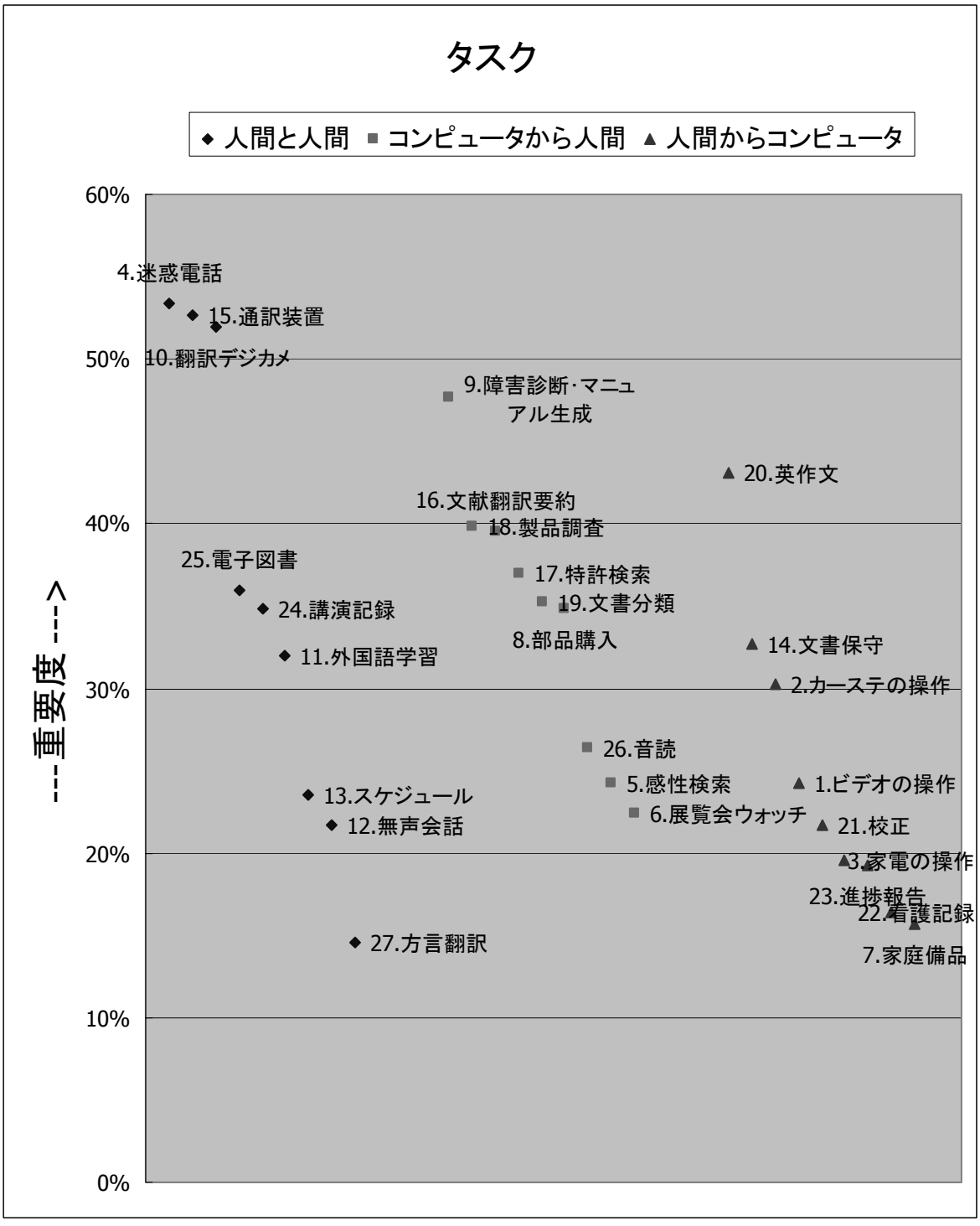


図３．４．３－２　タスクでみた重要シナリオ

これを見ると、コンピュータから人間へ情報を取り出すことが、人間からコンピュータへ認識させることよりも、ユーザの興味を引いたことがわかる。また、特に重視されたシナリオはコミュニケーションに属するものである。

コミュニケーションは言語の本来の機能であり、言語処理の応用タスクとして重要である。また、オンラインで情報が蓄積され、インターネットで情報の流れがよくなっただけに、それらを制御することが重要である。

（3） 処理言語でみた重要シナリオ

以下は、タスクの処理対象を、外国語、自国語、その他、の３つに区分して、シナリオをプロットしたものである。

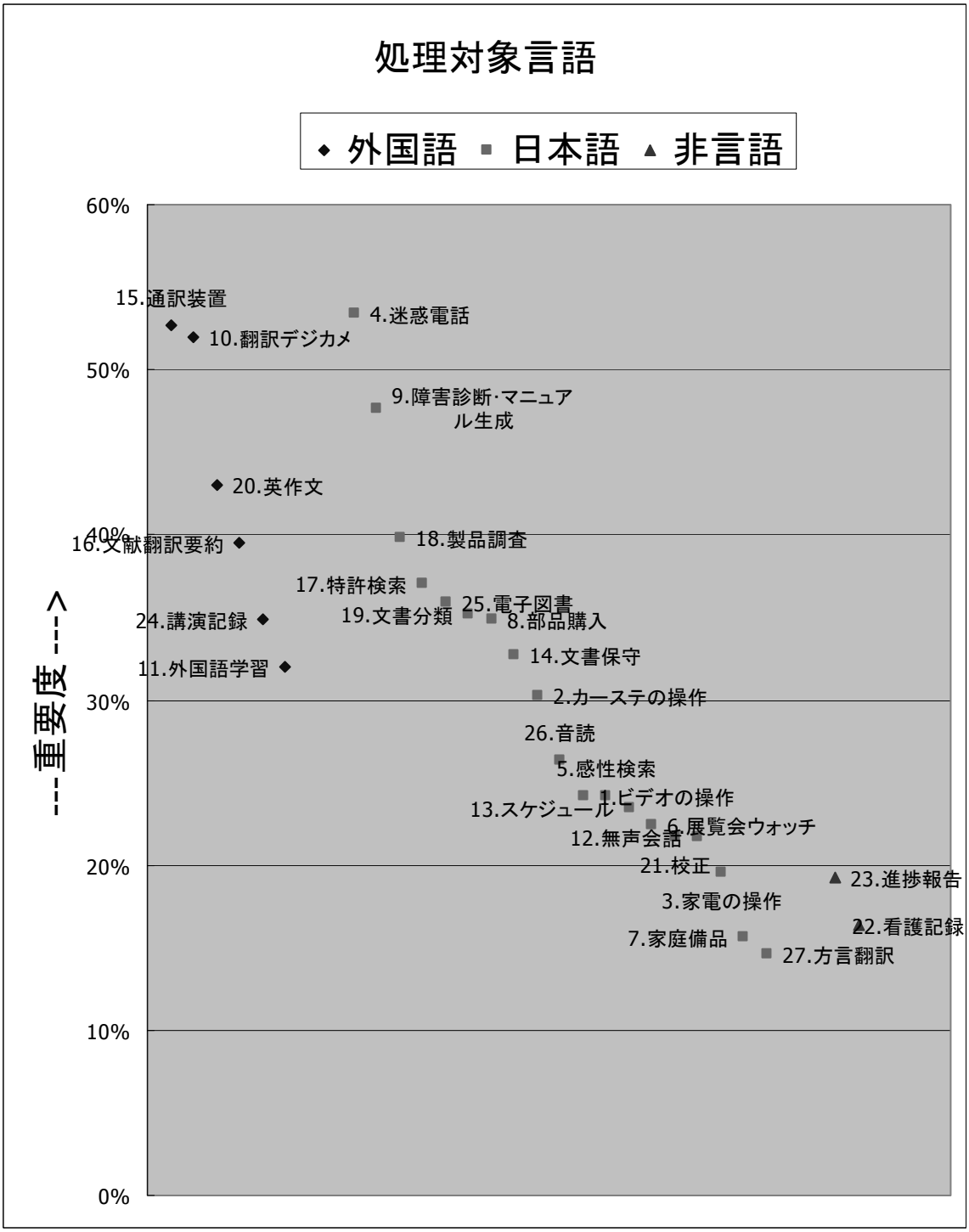


図３．４．３－３　処理対象でみた重要シナリオ

タスクの処理対象が外国語であるようなシナリオは、重要であることがわかる。

日本は歴史的に海外からの貪欲に文化を取り込んできたし、グローバライゼーションの流れもあり、海外の言語の処理に対するニーズが高い。

（4） 処理メディアでみた重要シナリオ

次は、処理メディアで区分したものである。

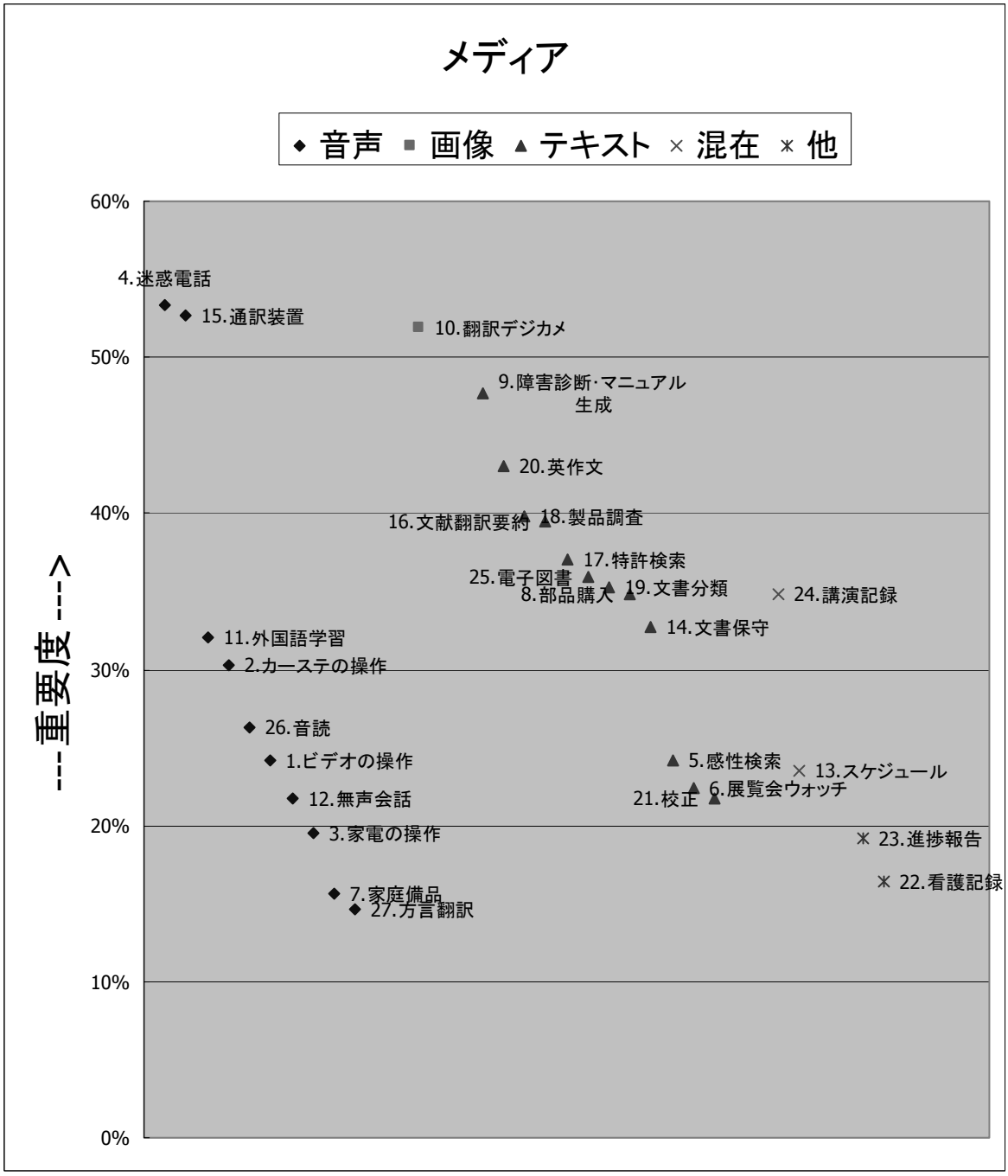


図３．４．３－４

全体的な分布からすると、音声よりも、テキスト処理が重要であることがわかる。

音声はそれだけで付加価値が高いので注目されやすいだろうとの予測に反し、音声だからといって価値を感じるというわけではないようである。そもそも音声というものは、情報の属性で付属的なものである。処理して価値を生じるものは情報そのものである。テキストの情報処理であっても実現が望まれている情報処理の価値は多い。逆にまた、音声等の情報処理に関しては、ユーザの興味を引くような応用開発がまだ不十分であるということを示唆する。

（５） 処理対象でみた重要シナリオ

以下は、処理対象を、会話、文書、ＷＥＢで区分したものである。

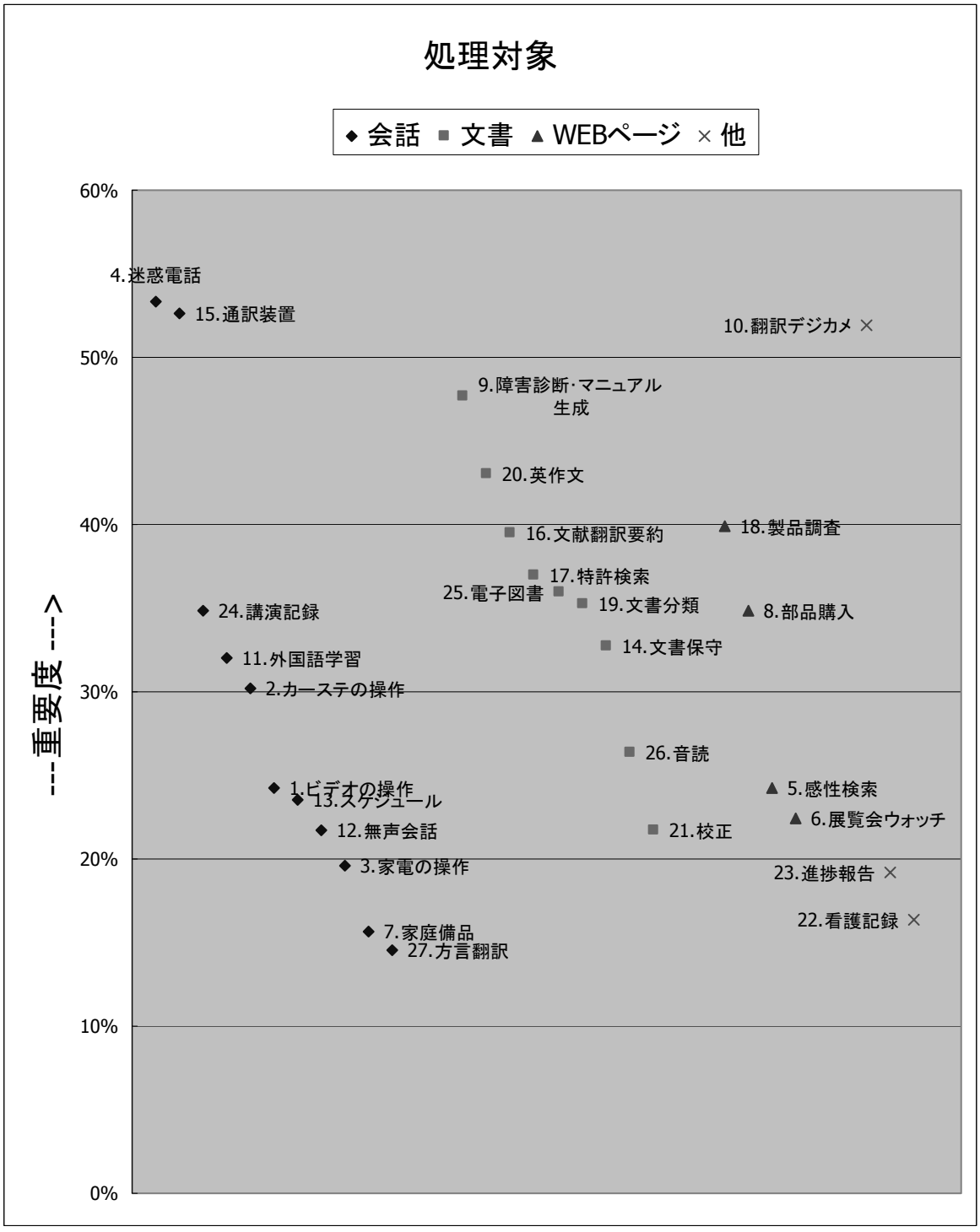


図３．４．３－５

全体的な分布からすると、会話やWEBよりも、文書の言語処理が重視されている。

文書処理の課題の解決が望まれている。また、会話やWEBの情報処理に関しては、ユーザの興味を引くような応用開発がまだ不十分であるということを示唆する。

（6） 年代別でみた重要シナリオ

以下は、年代別に、シナリオの平均点を示したものである。

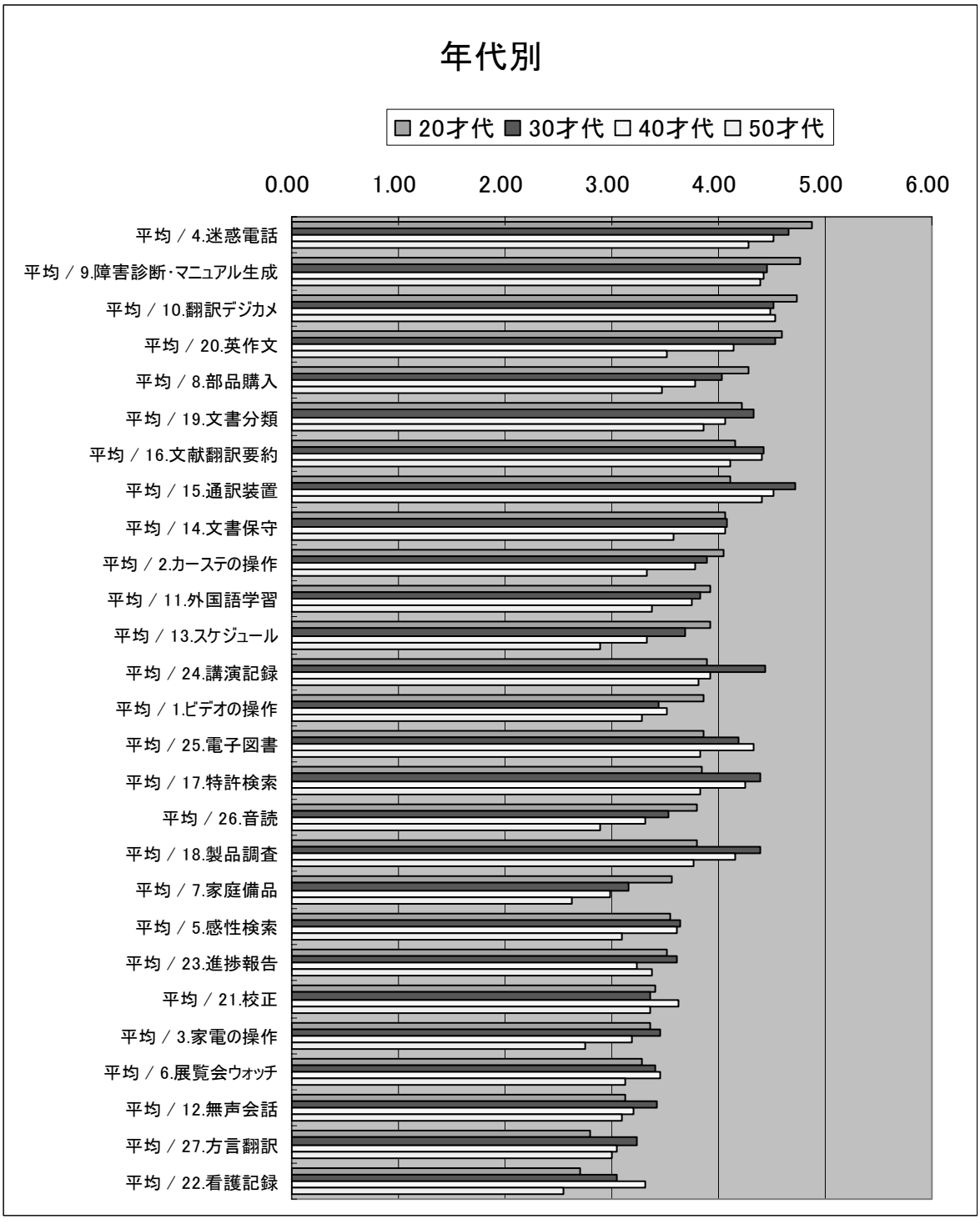


図３．４．３－５

２０歳代は「4.迷惑電話」、「9.障害診断・マニュアル生成」、「10.翻訳デジカメ」、「20.英作文」、「8.部品購入」と、困った状況に直結するシナリオを重視する傾向がある。

３０歳代は、「15.通訳装置」、「16.文献翻訳要約」などのほか、「17.特許検索」、「18.製品調査」など、知識を扱う処理を重視している。

４０歳代、５０歳代になるにつれて、「25.電子図書」や「19.文書分類」など、特定のタスクで必要なことというよりは、より一般的なことやインフラ関係が重視されてくる。

以上から、２０歳代、３０歳代の、現場で具体的な問題と格闘している人たちに注目し、それらの解決方法として自然言語処理の応用を考えていき、次第にインフラなど広範囲の価値につなげていくというアプローチが示唆される。

３．４．４ シナリオ間の関係の分析

（1）シナリオ間の相関係数

それぞれのシナリオについて、他のどのシナリオと相関があるのか調査した。以下の式で、各シナリオ同士の相関係数を求めた。アンケート回答者が回答した各シナリオの 7 段階の評価を 0～6 の数値（点数）として扱う。r_{cw} はシナリオ w とシナリオ c の相関係数を表す。 x_{wi} は i 人目の回答者が回答したシナリオ番号 w の点数を表す。$\bar{x}_w$ は全回答者のシナリオ番号 w への回答の平均点を表す。 n はアンケート回答者数を表す。

$$r_{cw} = \frac{\sum_{i=1}^{n}(x_{wi} - \bar{x}_w)(x_{ci} - \bar{x}_c)}{\sqrt{\sum_{i=1}^{n}(x_{wi} - \bar{x}_w)^2 \sum_{i=1}^{n}(x_{ci} - \bar{x}_c)^2}}$$

この式により計算した相関係数がトップ 30 に入るペアを表３．４．４－１に示す。この相関係数により、ランキングだけでは分からなかったユーザの嗜好が見えてくる。２位のビデオの操作とカーステの操作や、３位の「17. 特許検索」と「18. 製品調査」といった、利用環境的に近いと思われるシナリオ同士の相関係数は高い。「11. 外国語学習」と「14. 文書保守」、「7. 家庭備品」と「13. スケジュール」など、意外な相関関係もある。

表 ３．４．４－１　相関係数トップ 30

順位	シナリオ	シナリオ	相関係数
1	1.ビデオの操作	3.家電の操作	0.608
2	1.ビデオの操作	2.カーステの操作	0.595
3	17.特許検索	18.製品調査	0.561
4	13.スケジュール	14.文書保守	0.538
5	2.カーステの操作	3.家電の操作	0.523
6	16.文献翻訳要約	17.特許検索	0.496
7	8.部品購入	9.障害診断・マニュアル生成	0.491
8	15.通訳装置	16.文献翻訳要約	0.49
9	5.感性検索	6.展覧会ウォッチ	0.489
10	26.音読	27.方言翻訳	0.487
11	22.看護記録	26.音読	0.482
12	16.文献翻訳要約	20.英作文	0.469
13	24.講演記録	25.電子図書	0.468
14	10.翻訳デジカメ	15.通訳装置	0.467
15	3.家電の操作	7.家庭備品	0.464
16	15.通訳装置	20.英作文	0.452
17	7.家庭備品	8.部品購入	0.452
18	11.外国語学習	24.講演記録	0.439
19	11.外国語学習	14.文書保守	0.437
20	20.英作文	21.校正	0.436
21	8.部品購入	14.文書保守	0.426
22	9.障害診断・マニュアル生成	10.翻訳デジカメ	0.425
23	11.外国語学習	13.スケジュール	0.422
24	7.家庭備品	11.外国語学習	0.417
25	7.家庭備品	14.文書保守	0.417
26	7.家庭備品	13.スケジュール	0.416
27	18.製品調査	24.講演記録	0.413
28	5.感性検索	14.文書保守	0.411
29	4.迷惑電話	7.家庭備品	0.402
30	6.展覧会ウォッチ	14.文書保守	0.396

（2）相関係数によるシナリオのカテゴリ分類

シナリオ間相関係数の上位 20 ペアでシナリオをグループ化したものが図３．４．４－１である。シナリオ間の相関の高さは、それらを結ぶ矢印の太さで表した。また、各シナリオの重要度（図３．４．３－１)」の回答率）も、人気が高いほどフォントが大きく濃くなるという方法で表した。相関関係により、シナリオがグループ分けされ、それぞれが特徴を持ったユーザカテゴリを構成することが分かる。ここで図中の各カテゴリについて、説明する。

1. **翻訳機能のある情報分析･作成ツール**:このカテゴリに含まれるシナリオは､「10.翻訳デジカメ」、「15.通訳装置」、「16.文献翻訳要約」、「17.特許検索」、「18.製品調査」、「20.英作文」、「21.校正」である。最大のカテゴリであり、英語に関わることが仕事の一部になっている技術者、といったユーザ像が浮かび上がってくる。情報分析・作成のシナリオに英語に関する要望がここまで共起するとは意外であった。国際化の影響で、海外も含めた調査の需要が高いと考えられる。これらから、技術者向けの英語に関する作業支援といったソリューションなどが考えられる。
2. **電子デバイスの音声操作**：このカテゴリに含まれるシナリオは、「1.ビデオの操作」、「2.カーステの操作」、「3.家電の操作」、「7.家庭備品」である。ビデオ・カーステ・家電の操作に関するシナリオは、全体で見ると重要度はそれほど高くはないのだが、それぞれの相関係数は高い。音声操作はビジネス用途よりも、こうした一般的な用途に対して期待が高いと言える。
3. **情報になっていないものを情報化するツール**：このカテゴリに含まれるシナリオは、「22.看護記録」、「26.音読」、「27.方言翻訳」である。医療・教育・福祉などの意識の高い地域密着型ユーザ像が浮かび上がってくる。
4. **煩雑な業務を解消するツール**：このカテゴリに含まれるシナリオは、「13.スケジュール」、「14.文書保守」である。日々の定型業務をスピーディーにこなすことが求められているユーザが思い描かれる｡現状のIT技術の組み合わせを適用するだけで大きな恩恵を受ける層と言えるだろう。
5. **教育インフラ**：このカテゴリに含まれるシナリオは、「11.外国語学習」、「24.講演記録」、「25.電子図書」である。講演記録や電子図書といった社会インフラ的なものは相関が高いことは予想できたが、それらを求めるユーザ層が他のカテゴリのユーザよりも外国語学習に興味が持っているということは想定外であった。
6. **システムの維持管理**：このカテゴリに含まれるシナリオは、「8.部品購入」、「9.障害診断･マニュアル生成」である。システム管理などを業務とするユーザ層である。
7. **感性がかかわるものを処理**：このカテゴリに含まれるシナリオは、「5.感性検索」、「6.展覧会ウォッチ」である。コンピュータが苦手としてきた定性的な処理を実社会に求めるユーザ層である。

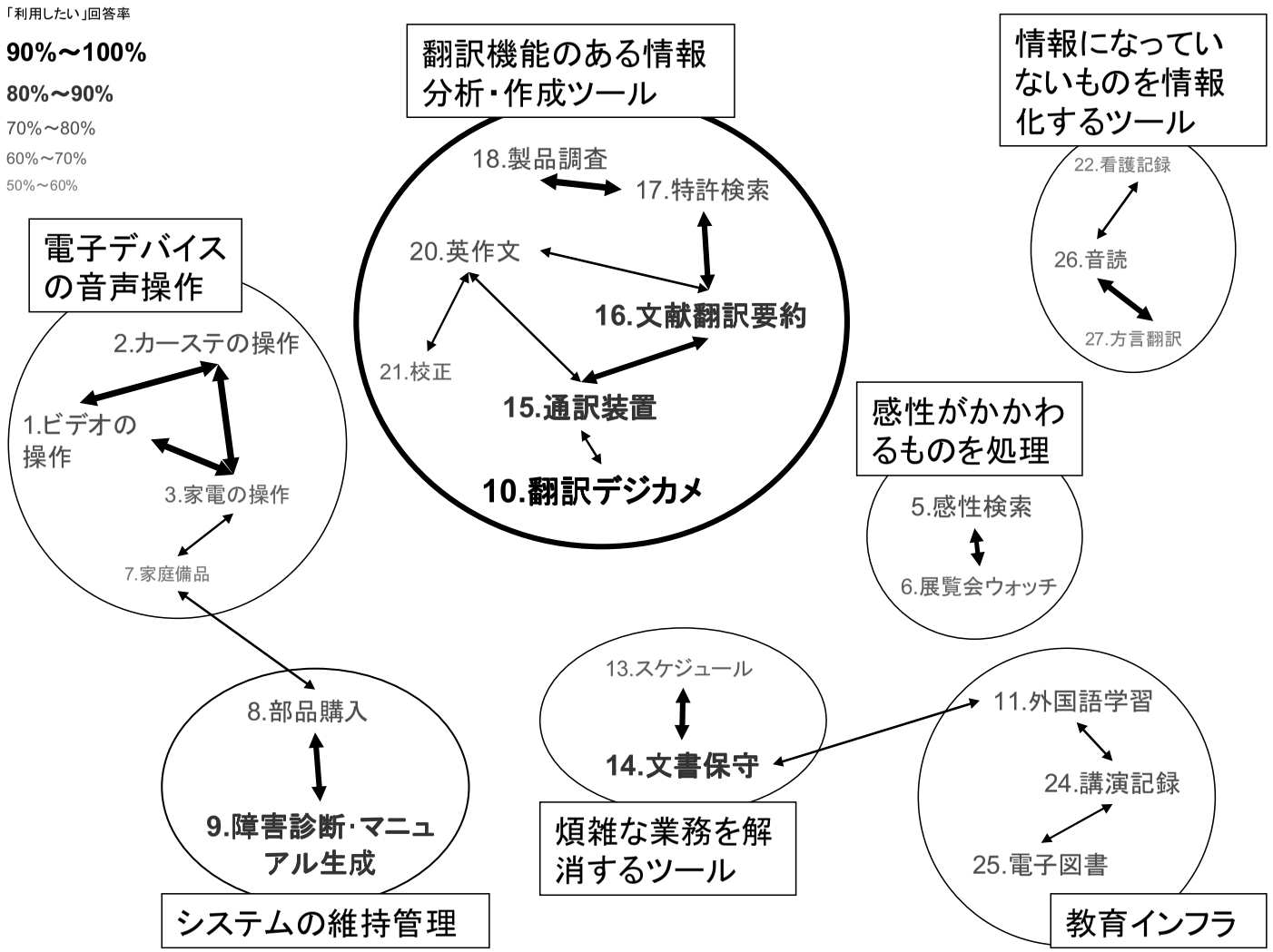


図３．４．４－１　相関係数によるシナリオのグループ化

シナリオ作成時に意図した通りに分類されたカテゴリもあるが、意図していなかったものもある。これらは、開発側ではとらえきれてなかった新たなターゲットカテゴリとして、応用製品開発に活用できると考える。

（3）相関関係ランクが低いナリオの分析

- 重要度ランキングトップの「4.迷惑電話」はようやく 29 位に登場する。トップでありながら相関関係ではこの順位ということは、他のシナリオとは独立なもので、誰からもまんべんなく重要視されているということであろう。
- 同様に、重要度では 10 位と、比較的高順位だった「19.文書分類」は最上位が 41 位と低迷している。「翻訳機能のある情報分析・作成ツール」カテゴリのものと相関していると予想していたが、実際は広く浅かった。

41	6.文献翻訳要約	19.文書分類	0.379
45	4.文書保守	19.文書分類	0.365
57	6.感性検索	19.文書分類	0.350
58	8.製品調査	19.文書分類	0.347
	…		

- 「12.無声会話」はそもそも全体での重要度が低い(21 位)うえ、他のシナリオとの相関も高くない(最上位 37 位)。実現の困難さが敬遠された原因か。

37	7.家庭備品	12.無声会話	0.387

- 「23.進捗報告」ももともと重要度が低い(24 位)。相関ランキングでは 51 位にやっと登場する。業務系のカテゴリのどれかに分類されると予想していたが意外な結果であった。

51	4.文書保守	23.進捗報告	0.360

（4）カテゴリごとの傾向の分析

相関係数は、1 シナリオ同士のものしか求めていない。1 対 1 の相関関係で分類した前記のカテゴリ同士の相関関係を求めることも可能だが、全体のデータ数が少ないためここでは行わなかった。その代わりに、視点を変えて、PrefixSpan というデータマイニングの手法を用いて、あるカテゴリに含まれるシナリオを好む人が、他にどのようなシナリオを好むのかという傾向を大まかに調べた。

PrefixSpan [1]とは、シーケンシャルパターンのマイニング手法であり、大量のデータから高速に頻出パターンを取り出すことができる。分析には、奈良先端大学院大学で公開されているフリーソフトを使用した[2]。

まず、翻訳に興味を持つ人はどういう人なのか、他にどういうことに興味を持っているのかを調べるため、「10 翻訳デジカメ」と「15 通訳装置」に 6 (最高点)をつけた人を 1 グループにして、これと共起する（これらの人が好む）シナリオを分析した。

「10 翻訳デジカメ」と「15 通訳装置」に 6 をつけた人がさらに選んだ (6 をつけた) 1 つのシナリオを以下に挙げる。相関係数とは若干異なる傾向が見られる。

選んだシナリオ	人数
9.障害診断･マニュアル生成	30
20.英作文	29
4.迷惑電話	27
16.文献翻訳要約	26
17.特許検索	25
18.製品調査	23
24.講演記録	23
11.外国語学習	21
19.文書分類	20

「10 翻訳デジカメ」と「15 通訳装置」に 6 をつけた人がさらに選んだ (6 をつけた) 2 つのシナリオを以下に挙げる。

選んだシナリオ	人数
9.障害診断･マニュアル生成、20.英作文	22
16.文献翻訳要約、20.英作文	20
17.特許検索、18.製品調査	20
4.迷惑電話、9.障害診断･マニュアル生成	20
17.特許検索、20.英作文	19
4.迷惑電話、20.英作文	19
9.障害診断･マニュアル生成、16.文献翻訳要約	19
9.障害診断･マニュアル生成、17.特許検索	19
16.文献翻訳要約、17.特許検索	18
18.製品調査、20.英作文	18
18.製品調査、24.講演記録	18
4.迷惑電話、17.特許検索	18

「10 翻訳デジカメ」と「15 通訳装置」に 6 をつけた人がさらに選んだ (6 をつけた) 3 つシナリオを以下に挙げる。相関係数とほぼ同じ傾向になる。

選んだ項目	人数
16.文献翻訳要約、17.特許検索、18.製品調査	16
16.文献翻訳要約、17.特許検索、20.英作文	16
17.特許検索、18.製品調査、20.英作文	16
9.障害診断･マニュアル生成、16.文献翻訳要約、20.英作文	16

- この共起情報から、特許検索や製品調査を業務としている人たちは英語を現場で使う必要性を感じているということが言える。これは先の相関関係分析と同様の結果である。
- 当初の予測と異なり、「11.外国語学習」が上位にない。「11.外国語学習」は重要度ランキングでは中位に位置するのだが、翻訳デジカメと通訳装置にもっと共起しそうなものである。推測になるが、多忙なこのカテゴリのユーザ層は外国語学習の余裕がないのかもしれない。

次に、「教育インフラ」カテゴリに対して、興味を持つ人はどういう人なのか、他にどういうことに興味を持っているのかを調べるため、「11.外国語学習」と「24.講演記録」と「25.電子図書」にそれぞれ 6 と答えた人が好むシナリオを分析した。以下、知見を記す。

- 「翻訳機能のある情報分析・作成ツール」との関係: 教育インフラに 6 をつけた人はこのカテゴリにも興味大。しかし、「17.特許検索」、「21.校正」との共起は弱かった。
- 「電子デバイスの音声操作」との関係: このカテゴリ中の「2.カーステの操作」とだけが比較的高い頻度で共起した。
- 他に比較的高い頻度だったものとして、「システムの維持管理」カテゴリ、「19.文書分類」(どのカテゴリにも含まれていない)、「4.迷惑電話」(どのカテゴリにも含まれていない) が挙げられる。
- 「教育インフラ」カテゴリの各シナリオにそれぞれ 6 と答えた人 (18 人) の性別構成を見ると、女性 4 人、男性 14 人と、女性の比率が全体での比率と比べて若干多い。教育に関する技術に対しての女性の感心の高さがうかがえるかもしれない。しかし、データが少ないのでなんとも言えない。

参考文献

1. J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu, "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth", Proc. 2001 Int. Conf. on Data Engineering (ICDE'01), Heidelberg, Germany, April 2001.
2. 工藤拓, "シーケンシャルパターン マイニングプログラム PrefixSpan+", http://cl.aist-nara.ac.jp/~taku-ku/software/prefixspan/

３．５ 対訳コーパスにおけるタグの妥当性検証の試み

当委員会では、数年前から、言語処理の研究・開発に貢献する対訳コーパスの作成を目指してきた。今年度は、これまで当委員会で模索してきたタグセットやタグ付け単位等の妥当性の検証を試みた。本節では、日英・英日の要素間の対応付けと、英日韓３ヶ国語の句の対応付けについて、述べる。

３．５．１ 日英・英日タグ付き対訳コーパス

（1）「日英・英日タグ付き対訳コーパス」開発の経緯と概要

当委員会では、英語と日本語の対応をとり、句や更に細かいレベルの名詞類等にタグ付けすることによって、言語処理の研究・開発に貢献する対訳コーパスの開発を目指してきた。タグ付け仕様は、コーパスの種類や文体に依存するところも大きく、年々変遷してきているが、基本的にはタグの種類を増やし、より細かい単位での対応をとる方向で、過去の仕様との整合性を図っている。

初年度にあたる1999年度は、経済白書、科学技術白書を題材にタグ付き対訳コーパスを開発した。白書の特徴として、複文や重文が多く一文がかなり長いということが挙げられる。故に、この長文を幾つかに分割するだけでも意義があるとされたため、一つの句は読点単位程度の幾分大まかなものとなった。英語と日本語を対応させるにあたり、表現方法のずれや構文構造の違いが見られる場合には、それを示すタグを設けて通常の句対応タグと区別した。また、用語抽出の目的も視野に入れ、一つの句の中でも「固有名詞」や「複合名詞」には別途細分化タグを付与した。

2000年度は、技術マニュアルを題材とした対訳コーパス開発を行った。当該マニュアルは、原文が英語の文書を日本語に翻訳したものであり、英語としての信頼性が高く、かなり正確に英文・和文の対応が取れていた。このため、ある程度格構造を意識し、白書関連の対訳コーパスよりも更に細かい単位での対応タグ付けを実施したが、この新方式に関しては委員の間でも賛否が分かれた。

そこで2001年度に入り、一旦は細かい単位でタグ付けした対訳コーパスに対し、各要素間の繋がりを示すため、「述部(動詞や動詞句)」とその「目的語(名詞や名詞句)」や「補語」、或いは「名詞句」とその「修飾句」等の対応付けを行うタグを追加した。

2002年度は、京大コーパスとその英訳版、及びPenn Treebankコーパスとその和訳版の各ペアを題材として、2001年度までの仕様を踏襲し、細分化タグと、各要素間を結ぶタグの両方を付与した。

従来仕様では、一方の言語には存在するが他方の言語には存在しない箇所は「対訳なし」として一括りで扱ってきたが、今回はこれを幾つかの現象に分類してみた。新設タグは、主として日本語における省略を扱うもので、英語サイドにのみ付与されることになる。時として同じタグが日本語サイドにのみ付与される場合もあるが、この多くは英語における省略というよりも日本語表現の方が冗長なケースといえよう。具体的には、“my father”や“his leg”のように英語では頻出する所有格の日本語における省略、主語の省略、目的語の省略、副詞句の省略、形容詞の省略、前置詞の省略、動詞の省略であり、個々については次のセクションで例示する。

（2）タグの種類

〈句の対応〉

P:　明確に対応している句対応。

e.g.　P J26:0-4: 路上では

P E26:14-27: in the street

PX:　意味上では完全に対応しているが、構文単位が明確に対応しない句対応。

e.g.　PX J23:2-8: は１勝１敗。

PX E23:6-32: won one game and lost one.

PL:　複数の句同士の対応。「P」や「PX」のような各要素ごとのタグに加え、各要素間の繋がりを分かり易くするため、「PL」というタグを用いる。特に、「述部(動詞や動詞句)」と「目的語(名詞や名詞句)」や「補語」、「名詞句」と「修飾句」などを連続した形で示すことにする。

e.g.　P J19:6-7: は

P J19:11-17: を炊き出し、

P E19:6-13: prepare

P J19:7-11: お節料理

P E19:14-32: New Year's dishes,

PL J19:6-17: はお節料理を炊き出し、

PL E19:6-32: prepare New Year's dishes,

PPRON:　片方の言語で、人称代名詞の所有格が省略されている場合。

日本語では、「ひと」や「もの」の前に所有格を置くのはあまり一般的ではないが、英語ではよく見られる。

e.g.　PPRON J11:0-2: 相手

PPRON E11:15-35: their opponent team.

上記の所有格に加え、文脈上主格の表現に相違がある場合（片方が一般名詞で、他方が代名詞）にも「PPRON」を付与。

e.g.　PPRON J7:8-10: 孫娘

PPRON E7:78-81: she

〈語の対応〉

W:　固有名詞、及び「伏せたい商品名」等を含む固有名詞。

e.g.　W J10:4-10: ギングリッチ

W E10:91-99: Gingrich

WN:　複合名詞、及び「伏せたい商品名」等を含む複合名詞

e.g.　WN J19:7-11: お節料理

WN E19:14-31: New Year's dishes

〈コメント〉

NT:　〈範囲タグ〉〈コメント〉→ 対訳なし。コメントは積極的に省略される。2002年度の新設仕様として、以下に示すように、「NT」を品詞カテゴリごとに細分類して「NT*」とした。

細分類できなかったものや、品詞分類が曖昧なもの等は「NT」のまま。また、「なお、しかし、さらに、それにより」“but, then, therefore”等の接続詞の対応が存在しない場合は、特に「NT」を付与していない。関係代名詞(that, which, who等)も、日本語側に関係代名詞自体の対応は存在しないが、特に「NT」を付与してはいない。

「NT*」の例外として、たとえ品詞ごとに分類できたとしても、一文まるごと対応箇所がない場合や、that節、関係節等のまとまった単位で対応箇所がない場合なども「NT」とした。元々「NT」を細分化した意図は、対応のとりづらい現象を抽出してみる、という試みであったため、全く対応していない大きな単位を品詞ごとに分類する意味は薄いと判断した。

e.g.　NT E265:37-50: I ignored him

「エッ」と思いつつファンデーションをぬっていると、また主人が「もっと、もっと、のばせよ」。

Though I was a little bit surprised, I ignored him and kept putting on foundation; my husband said again, “Go on, go on, spread that foundation thinner!”

NTS:　〈範囲タグ〉→主語（句）が省略されている場合。

e.g.　NTS E245:14-22: his wife

おいのマンションは幼稚園のすぐ隣で、中の様子は手にとるように分かるんです。

My nephew and his wife live next door to a kindergarten, so they can perfectly understand what is going on in the kindergarten.

※「NTS」の例外として、形式主語の “it” には「NTS」を付与しない。

e.g.　カギを開けるのは難しい。

It is such a tough task to unlock the gate to the dream world.

NTO:　〈範囲タグ〉→目的語（句）が省略されている場合。

e.g.　NTO E2:34-42: the gate

カギを開けるのは難しい。

It is such a tough task to unlock the gate to the dream world.

NTADV:　〈範囲タグ〉→副詞（句）が省略されている場合。

e.g.　NTADV E2:43-62: to the dream world.

カギを開けるのは難しい。

It is such a tough task to unlock the gate to the dream world.

※前文に“to the dream world”に相当する部分が表現されている。

cf. ＜前文＞

国境の南の街は夢の国への通用門でもある。

The city to the south of border is also a gate to their dream world.

NTADJ:　〈範囲タグ〉→形容詞（句）が省略されている場合。

e.g.　NTADJ J10:4-15: ニューヨーク株式市場の

三十日のニューヨーク株式市場のダウ工業株三十種平均は、前日比一・〇一ドル高の三八三四・四四ドルで一九九四年の取引を終了した。

The Dow Jones Industrial Average on the 30th, the end of the trading for 1994, was at 3,834.44 dollars, up 1.01 dollars from the previous day.

NTPREP:　〈範囲タグ〉→前置詞が省略されている場合。

e.g.　NTPREP E2:47-56: regarding

当面、行政改革、とりわけ規制緩和、特殊法人の見直し、地方分権など大きな課題がある。

For the time being, there are important issues regarding administrative reform, especially on deregulation, review of government-affiliated corporations, and the decentralization of authority.

NTVB:　〈範囲タグ〉→動詞（分詞を含む）が省略されている場合。

e.g.　NTVB E18:102-108: gained

死亡数は前年を六千人下回り、出生数から死亡数を引いた人口の自然増加数も二十一年ぶりに対前年比で五万二千人アップした。

The number of deaths dropped by 6,000 from the previous year, and the natural increase of population, gained by subtracting the deaths from births, showed an increase of 52,000 over the previous year, for the first time in twenty-one years.

TM1:　〈範囲タグ〉“〈正しい綴り〉” → 綴りが間違っている場合。

TMi:　〈範囲タグ〉“ ” → 必要なスペースが存在しない場合。

ER:　〈範囲タグ〉〈修正案〉→ その他のエラー。

C:　〈範囲タグ〉〈コメント〉→　その他のコメント。

（３）英日対応タグ付け例

===== 10 ======

共和党はギングリッチ新下院議長が作成した選挙公約「米国との契約」の中で「民主主義と市場経済への改革、及びシビリアンコントロール下の軍をもつ国々がＮＡＴＯに加入することを促進する」とうたい、東欧諸国のＮＡＴＯ加入を進めることを明らかにしている。

In its campaign platform set out by the new Chairman of the House of Representatives, Newt Gingrich, and titled, “Contract with America,” the Republican Party promised to “promote the entry into NATO of countries with the will to turn to democracy and a market economy, as well as placing their military forces under civilian control,” thus demonstrating their plan to encourage Eastern European nations to join NATO.

======

P J10:0-3: 共和党

P E10:138-158: the Republican Party

P J10:3-4: は

P J10:89-94: とうたい、

P E10:159-167: promised

W J10:4-10: ギングリッチ

W E10:91-99: Gingrich

PX J10:4-15: ギングリッチ新下院議長

PX E10:36-100: the new Chairman of the House of Representatives, Newt Gingrich,

----

W J10:11-15: 下院議長

W E10:44-84: Chairman of the House of Representatives

P J10:15-20: が作成した

P E10:25-35: set out by

P J10:20-24: 選挙公約

P J10:32-35: の中で

P E10:0-24: In its campaign platform

PPRON J10:20-24: 選挙公約

PPRON E10:3-24: its campaign platform

P J10:24-32: 「米国との契約」

P E10:113-137: “Contract with America,”

P J10:36-50: 民主主義と市場経済への改革、

P E10:227-269: to turn to democracy and a market economy,

P J10:50-52: 及び

P E10:270-280: as well as

P J10:52-65: シビリアンコントロール下の

P E10:311-334: under civilian control,

P J10:65-69: 軍をもつ

P E10:281-310: placing their military forces

PPRON J10:65-66: 軍

PPRON E10:289-310: their military forces

PX J10:69-72: 国々が

PX E10:200-212: of countries

P J10:72-76: ＮＡＴＯ

P E10:195-199: NATO

P J10:76-83: に加入すること

P E10:180-194: the entry into

PL J10:76-88: に加入することを促進する

PL E10:168-194: to ”promote the entry into

P J10:83-88: を促進する

P E10:172-179: promote

P J10:94-98: 東欧諸国

P E10:379-403: Eastern European nations

PX J10:99-105: ＮＡＴＯ加入

PX E10:404-417: to join NATO.

P J10:105-111: を進めること

P E10:366-378: to encourage

NTVB E10:105-112: titled,

PL J10:109-121: ことを明らかにしている。

PL E10:341-365: demonstrating their plan

P J10:111-121: を明らかにしている。

P E10:341-354: demonstrating

NTADV E10:213-226: with the will

PPRON E10:355-365: their plan

PPRON J10:109-111: こと

（4）分析

今回は、この二種類のコーパスを用いることで、「翻訳方向とタグ付け容易性との相関関係」を調査することも目的の一つであった。これまでの対訳コーパス開発の経験上、ある程度予測はできていたが、実際に作業にあたってみると、やはり翻訳方向による明らかな差異が認められた。

日本語から英語に翻訳した京大コーパスでは、英語と日本語の間で明確な対応がとり難い、或いは対応がとれないといったケースが多く挙げられた。特に、スポーツ記事や会話文では、所謂ジャーナリズム特有の簡潔な表現や臨場感に溢れる表現が多いためか、こうした傾向が顕著であった。これは、日本語の言語特性にも大いに依存するものと思われる。改めて言うまでもないが、日本語では、自明の主語や文脈から判断できる目的語等を省略するのが常であり、これを英訳する際には、省略語句を補わねば英文として理解できないものとなる。

勿論、翻訳方向が逆であっても同じ現象は起こるはずだが、Penn Treebankコーパスを英語から日本語に翻訳したものでは、京大コーパスの場合と比べて問題の件数が少なかった。一つの仮説としては、初めから日本語で文章を執筆する場合と、英語を日本語に翻訳する場合とでは、用いられる日本語の構造、延いては自然さが異なるのかも知れないといえよう。尤も、翻訳家が邦訳書を出すような場合の翻訳は別だろうが、通常新聞記事のようなものを翻訳する場合には、原文である英文に忠実に翻訳しようという意識が働き、自ずと英語の構造を多少引きずった和文となるものである。

従って、英語と日本語の対応関係が、より正確に、より多くとれたタグ付き対訳コーパスを目指すなら、原文が英語で、それを日本語に翻訳したものを題材にする方が向いていると考えられる。英文の質の面からも、英語ネィティブの英語、即ち原文が英語であることが望ましい。翻訳品質に関しては、翻訳者個人の素養によるところが大きいが、一般に英語のノンネイティブによる日本語から英語への翻訳は、逆方向に比べてコストもかかるうえ、質も下がるようである。

但し、翻訳品質が保証されているのであれば、日本語から英語に翻訳したものであっても、「言語特性に起因する対応関係のずれ」といった観点から分析を行う等、価値ある対訳コーパスを開発することはできよう。機械翻訳システム等を想定した場合、この対応関係のずれが大きくなるほど処理の難易度は上がると思われる。将来的には、タグ付き対訳コーパスに付与されたタグの種類を手掛かりに、言語処理の難易度に応じた評価コーパスを抽出することも可能だろう。

３．５．２ 英日韓３ヶ国語タグ付き対訳コーパス

（1）「英日韓３ヶ国語タグ付き対訳コーパス」開発の経緯と概要

2001 年度には、これまで当委員会にて模索してきたタグセットやタグ付け単位等の妥当性を検証することを主眼に、前記技術マニュアルの英日対訳タグ付きコーパスをベースとして、韓国語のタグ付けを試みることにした。

まず、英文か和文のいずれかを韓国語へ人手で翻訳する必要があり、翻訳方向については委員会でも議論が分かれたが、元々英語で書かれたマニュアルであるため、日本語を介さずダイレクトに英韓翻訳を行うことになった。また、言語の類似性の観点からも、今回の主目的が日韓対訳タグ付け作業

を通じて現行の英日対訳コーパスを検証するという点にあることを考慮すると、日本語から韓国語への翻訳は敢えて避けた方がよいとの結論に至ったものである。

本調査は、日本語、英語、韓国語の三つの言語を対象に言語間に見られる諸特徴を記述するためにマニュアルの翻訳文を対象に分析を進めた。作業対象は、下記に示す３ヶ国語に対応付けられた文章とそれを翻訳に便利な単位に区切ったものに分けられ、分析に先立って対応付けられた文章の翻訳を検証し、自然な翻訳文になっているかどうかをチェックした。特に、英語から韓国語への翻訳が自然な韓国語になっているかをチェックした。

（2）タグ

日韓対訳タグ付け作業に先立ち、現行の英日対訳コーパスから、各文単位で対訳タグ付けセットを抽出し、各英日対訳ペアを日本語を基準として出現順に並べ替えるツールを作成した。この各文の先頭に英語から韓国語への人手による翻訳文（※１）を追加し、各英日対訳ペア（むしろ「日英対訳ペア」と呼んだ方がよいかも知れない）の日本語フレーズに対応する韓国語のフレーズ（※２）を追加することにした。

韓国語原文１（英語原文１を韓国語に翻訳したもの）※１
日本語原文１
英語原文１
韓国語フレーズ１（「日本語フレーズ１」に対応するもの）※２
日本語フレーズ１
英語フレーズ１
韓国語フレーズ２（「日本語フレーズ２」に対応するもの）※２
日本語フレーズ２
英語フレーズ２
…　　…

（3）例

・3ヶ国語対応文章

(韓国語)

사용자와 다른 관리자들이 XXXXX Server 에 접속할 수 있도록 설정하는 여러 개의 Windows NT 로컬 그룹 중의 한가지

(日本語)

ユーザーおよび別管理者が XXXXX サーバーにアクセスできるように, XXXXX 管理者がセットアップする Windows NT ローカルグループの１つです.

(英　語)

One of several Windows NT local groups the XXXXX administrator sets up so that users and other administrators have access to a XXXXX Server.

・翻訳に便利な単位に区切られたもの

사용자와 다른 관리자들이

P J108:0-12: ユーザーおよび別管理者が

P E108:79-109: users and other administrators

XXXXX Server

W J108:12-21: XXXXX サーバー

W E108:127-139: XXXXX Server

XXXXX Server 에

P J108:12-22: XXXXX サーバーに

P E108:122-140: to a XXXXX Server.

XXXXX Server 에 접속할 수 있도록

PL J108:12-33: XXXXX サーバーにアクセスできるように,

PL E108:71-78: so that

PL E108:110-139: have access to a XXXXX Server

・分類したもの（分類記号の後はコメント）

PO > :영어는 주어에 조사를 붙이지 않음

사용자와 다른 관리자들이

P J108:0-12: ユーザーおよび別管理者が

P E108:79-109: users and other administrators

PM >

XXXXX Server

W J108:12-21: XXXXX サーバー

W E108:127-139: XXXXX Server

PM >

XXXXX Server 에

P J108:12-22: XXXXX サーバーに

P E108:122-140: to a XXXXX Server.

PO > : 영어는 어미 변화가 없음

XXXXX Server에 접속할 수 있도록

PL J108:12-33: XXXXXサーバーにアクセスできるように,

PL E108:71-78: so that

PL E108:110-139: have access to a XXXXX Server

（4）分析

分析作業は、区切られたものを対象にその形式的な面に注目し、次の五つの観点から分類した。

①日英韓が一致するもの　→ＰＭ

②英韓は一致し、日本語のみが異なるもの　→ＰＮ

③日韓は一致し、英語のみが異なるもの　→ＰＯ

④日英は一致し、韓国語のみが異なるもの　→ＰＱ

⑤日英韓がすべて一致しないもの　→ＰＸ

①は、日英韓が完全に一致するもので、次のように品詞構成などにおいてずれが見られないものである(韓国語のハングル表記の後の括弧は日本語の翻訳である)。

액세스 그룹(アクセスグループ)

アクセスグループ

access group

ActiveX 콘트롤(ActiveXコントロール)

ActiveXコントロール

ActiveX control

上記の例でも確認できるように、日英韓で一致するものは名詞、より具体的には専門用語が圧倒的に多い、一方、このような名詞の場合、分析の精度をどこまで設定するかによって結果が異なる。今回は問題にせず、品詞的観点からその対応関係がすなおに対応していれば一致するものと判断することにする。したがって、次のような例は、日本語と英語が共に「形容詞＋名詞」の構成になっているのに対して韓国語の場合は、「名詞＋名詞」の構成になっているため、韓国語のみが異なるものとして分類することになる。

기본 대기열(基本待機列)

プライマリキュー

primary queues

一方、次の場合は、日本語と韓国語が共に「名詞＋名詞」の構成になっているのに対し、英語の場合は、「形容詞＋名詞」の構成になっているため、この場合は英語のみが異なるものとして分類する。

처리 대기열(処理待機列)
處理キュー
processing queue

②は、英語と韓国語が一致し、日本語のみが異なるものである。この分類に属するものとしては、①と同じく、品詞のずれが目立ち、英語と韓国語は、「形容詞＋名詞」の構成になっているのに対し、日本語の場合は「名詞＋名詞」構成になっている場合がある。

실패한 대기열(失敗した待機列)
失敗キュー
failed queue

これと類似した例として、英語と韓国語は共に「名詞＋名詞」構成になっているが、日本語の場合、「手動による」のようになっている例がある。

수동 라우팅(手動ルーティング)
手動によるルーティング
manual routing.

さらに、最もこの分類で一般的に観察されるのは、次のように英語と韓国語は「名詞＋名詞」の構造を持つのに対し、日本語の場合は、「名詞＋の＋名詞」の構造を持つ例である。

폴더 활성화 (フォルダ活性化)
フォルダの有効化
folder enabling

팩스 표지
ファックスの 表紙
fax cover pages

次の例はかなり複雑な例で、英語と韓国語の場合は、「名詞＋動詞」の構造をしているが、日本語の場合は「ユーザーが定義する」のように完全な文の形になっているのが特徴である。さらに、韓国語の場合は受動表現を用いるが日本語の場合は、能動表現であるのも注目される。

사용자 정의된(使用者定義された)
ユーザーが定義する
User-defined

③は、日本語と韓国語が一致し、英語のみが異なるものである。この分類に入るものとしては、関係詞によって表現されるが、日本語や韓国語には関係詞がないため英語のみが特殊な構造を持つ例が多く見られる。

COM(Component Object Model)에 기초하는 기술 셋트로(COM(Component Object Model))に基づく技術セットで)
コンポーネントオブジェクトモデル (COM) に基づく技術セットで,
A set of technologies,
that is based on the Component Object Model (COM)

을 나타내는 (を表す)
を開示する
that exposes

에 유용한(に有用な)
に利用できる
that are useful for

英語の場合は文の中での位置によってその文法的意味を表すが日本語と韓国語は共に助詞を用い文の中での名詞の文法的意味を表示するといった違いがある。

ActiveX 컨트롤을(ActiveX コントロールを)
ActiveX コントロールを
ActiveX controls

英語は前置詞によって表現するところを日本語や韓国語は助詞によってそれを表現する。

와 ---를 일치시킴으로써(と…を一致させることで)
と
を符合させることで,
by matching
with

한 번에(一回に)
一度に
at once

英語は主語明示型であるのに対し、日本語や韓国語は主語を明示しない言語である。

추가할 수 있습니다 (追加できます)
追加できます.
you can add
to

처리가 완료된 문서(処理が完了されたドキュメント)
処理が完了したドキュメント
documents you have finished processing

문서를 넣는 대기열.(文章を入れる待機列)
ドキュメントを配置するキューです.
The queue in which you place documents

선택합니다.
選択します.
You select

なお、日本語と韓国語は語順が比較的自由であり、特に述語の場合は、文末に来るという特徴を共有する。一方、英語の場合は、命令文の場合は文頭に動詞が来るといった特徴がある。

실패한 대기열 및 처리 중인 대기열을 비교해보십시오(失敗した待機列および処理中の待機列を比較してください)
失敗キューおよび處理キューと比較參照してください.
Compare with failed queue and processing queue.

語順と関連して、英語も日本語も韓国語も主要部が後に来るという特徴があるが、次のような場合、日本語と韓国語は主要部が後に来るが、英語の場合は前に来る。

아키텍처 모델(アーキテクチャモデル)
アーキテクチャモデル
model for the architecture

④日本語と英語は一致し、韓国語のみが異なるものはかなり数が少なく、次のように韓国語のみが「名詞＋名詞」の構造をしているものがあった。しかし、これは翻訳に強く影響されるもので韓国語でも日本語のように外来語として翻訳をしていたら日英韓が一致するものとして分類することができる例である。

기본 대기열(基本待機列)
プライマリキュー
primary queues

⑤は、日英韓がすべて一致しないものであり、日本語と韓国語は一致し、英語のみが異なる場合と基本的にはその原因が類似している。数としては、日本語と韓国語が一致しないがためにこの分類に入ったものが多い。その原因を見ると次のように日英韓の言語の表現形式の違いから来るものがある。

를 설정할 때(を設定する時)
の設定時に
when setting

英語の場合は関係詞によって時間を表すが、日本語と韓国語は「名詞(時)＋に」のような表現を用いる。しかし、上記の例でも確認できるように、日本語の場合は“ドキュメントプロパティの設定時に”のように「の」を用いるが韓国語の場合は「를 설정할 때(を設定する時)」のように日本語の「を格」にあたる「 를 」を用いる。もちろん、日本語も“ドキュメントプロパティの設定時に”ではなく、“ドキュメントプロパティを設定する時に”のように表現すれば「を格」を用いることができる。しかし、韓国語の場合、「設定時」のように表現する場合は助詞が省略されることはあっても日本語の「の」

に該当する「의」を用いることはない。

一方、次のように日本語と韓国語の翻訳文体の違いによってこの分類に入ったものもかなりある。

전자 목록(電子目録)
電子的なリストです.
An electronic list

日本語と英語の場合は、基本的に「形容詞＋名詞」という構造を持っているのに対し、韓国語の場合は「名詞＋名詞」の構造を持つといった違いがある。また、日本語の場合は「～です」の形で文が終わっているが、韓国語の場合は、「電子目録」で終わっている。もちろん、韓国語にも日本語の「～です」に該当する「～입니다」という形式があり、先の例を「電子目録입니다」のように翻訳することもできる。これは翻訳における文体の違いということで無視しても良いかもしれない。

以上、分類の基準と幾つかの具体例を示したが、今回調査の対象になった 1,512 件のそれぞれの内訳をまとめると次のようになる。

パターン	出現回数	割合(%)
ＰＭ	920	60.85
ＰＮ	76	5.03
ＰＯ	342	22.61
ＰＱ	13	0.86
ＰＸ	161	10.65
計	1,512	100.00

表３．５．２－１ 分類ごとの内訳

日英韓の全体が一致するものが 920 件で全体の 60.85％を占めている。この数字は、マニュアル類を対象に行ったものであり、翻訳のために区切られた比較的限られたものを対象にした分析であることからかなり高い一致度を見せたものであると考えられる。また、今回の調査が語構成や語種の問題などは考慮に入れなかったことも一つの原因であると考えられる。一方、日英韓の全部が不一致であったものは 161 件で全体の 10.65％を占めている。

英語のみがことなる場合は、342 件 22.61％で最も高く、その次が日本語のみ異なる場合で、76 件の 5.03％であり、韓国語のみが異なる場合は、13 件の 0.86％である。英語を基準にした場合、日本語と韓国語は非常に近く、韓国語の方が日本語より英語に若干近いことが言える。しかし、その原因は、日本語の訳文の場合はかなり日本語として自然な文になっているのに対して韓国語の訳文の場合は、英語の直訳といった感じがするのが最も大きい原因であると考えられる。したがって韓国語の翻

訳をより自然な韓国語に直した場合、日本語と韓国語はほとんど差がないと言えるかも知れない。

以下にまとめを述べる。

- 句の切り出し単位については、英日対訳コーパスの作業結果のタグセットで、韓国語に対しても問題はない。
- 英韓対訳コーパスからの日韓対訳タグ付け結果では、日韓のずれが少ない。英日対訳コーパスの作業結果のタグセットで、英韓対訳タグ付けも英日対訳タグ付けと同じ内容に作業できた、といえる。

３．５．３ 今後の予定

日英・英日双方向の翻訳コーパスで、句対応が一致している文ほど、翻訳過程で特別な構造変換などが必要なく、逐語訳に近い方式での翻訳が適用できそうである。逆に、句構造が一致していない文は、一般的な方式では扱いきれない言い回しや構文を含み、翻訳が困難な文であろう。このような、句対応の一致・不一致の特徴は、難易度段階別の翻訳評価文セットの開発に利用できる可能性が高いため、来年度は、評価文セットの開発を視野に入れて活動する予定である。また、タグセットなど、作業の成果をISOなどの標準化委員会に提案することも検討に入れたい。

４．Web 情報アクセス技術専門委員会活動報告

４．Web 情報アクセス技術専門委員会活動報告

４．１ はじめに

昨年度まで「文書情報技術専門委員会」として活動してきたが、本年度から「Web 情報アクセス技術専門委員会」と改称し、引き続き Web 関連の技術について調査をおこなった。

Web はもはや専門家だけのものではなく、一般ユーザにとっても日常生活におけるさまざまな問題解決のための有力なツールとなっている。また、Web 上に発信されている情報は、企業・組織にとっても迅速な意志決定やマーケッティングのために活用されている。

現時点では、ほとんどのユーザが Goo、Yahoo、Google といった汎用の検索エンジンを用いて必要な情報を得ているのが実状である。そこで、昨年度は汎用の検索エンジンについてのヒアリングを中心とした調査をおこなった。の結果、Web 上の情報はますます増加の一途をだとっており、もはや汎用検索エンジンでは必要な情報にアクセスするのに十分でないという考えに至った。実際、これを補うために最近では特定の問題を解決するための専門分野 Web 検索サイトも登場してきている。たとえば、旅行計画の立案に関連する情報を集めて整理したサイトや商品の価格比較や評判検索によって商品購入を支援するようなサイトがこれにあたる。

そこで、本年度は汎用検索エンジンから特定の専門分野に特化した検索サイトに調査対象を移し、このようなサイトの事例、およびそこで用いられている要素技術を中心に調査をおこなった。

4.2 節では、早稲田大学の山名助教授をお招きして、Web の現状と課題についてヒアリングをおこなった結果をまとめた。4.3 節では今年度の調査対象とした専門分野 Web 検索を定義し、いくつかの観点から分類をおこなった。4.4 節では、専門分野 Web 検索サイトの実例として、2 つの事例についてヒアリングをおこなった。価格.com はユーザの商品購入を支援することを専門としたサイトである。一方、モバイルインフォサーチはユーザの現在地を考慮して、検索をおこなうユニークな検索サイトである。4.5 節、4.6 節では、文献調査を中心に調査をおこなった。特に 4.5 節では、システム技術を中心に調査し、4.6 節では各システムで用いられる要素技術を中心に調査した。4.5 節、4.6 節で調査した論文は以下のとおりである。

- Robert Steele: “Techniques for Specialized Search Engines”, in Proceedings of Internet Computing, 2001. (4.5.1 項)
- Robert B. Doorenbos, Oren Etzioni, Daniel S. Weld: "A Scalable Comparison-Shopping Agent for the World-Wide Web", In Proceedings of International Conference on Autonomous Agent, 1997. (4.5.2 項)
- Panos M.Markopoulos, Jeffrey O. Kephart: "How valuable are shopbots?", In Proceedings of International Conference on Autonomous Agents, 2002. (4.5.2 項)
- E. Brynjolfsson and M. D. Smith: “Frictionless commerce? A comparison of internet and conventional retailers”, Management Science, 46(4), 2000. (4.5.2 項).
- Thorsten Joachims, Dayne Freitag, Tom Mitchell: "WebWatcher: A Tour Guide for the World Wide Web", In Proceedings of International Joint Conference on Artificial

Intelligence, 1997. (4.5.3 項)

- 立石健二, 石黒義英, 福島俊一: “インターネットからの評判情報検索”, 情報処理学会第 62 回全国大会, 4W-5, 2001. (4.5.4 項)
- 立石健二, 石黒義英, 福島俊一: ”評判情報検索システムの試作と評価”, 情報処理学会第 63 回全国大会, 2V-1, 2001. (4.5.4 項)
- 立石健二, 石黒義英, 福島俊一: ”インターネットからの評判情報検索”, 情報処理学会研究報告, NL-144-11, 2001. (4.5.4 項)
- 立石健二, 福島俊一: ”意見分析システムにおける意見抽出方式の検討と評価”, 第 1 回情報科学技術フォーラム, D-1, 2002. (4.5.4 項)
- 立石健二, 森永聡, 山西健司, 福島俊一: ”Ｗｅｂ上の意見分析－情報抽出とテキストマイニングの融合－”, 情報処理学会第 64 回全国大会, 2X-4, 2002. (4.5.4 項)
- Steve Lawrence: "Online or Invisible?", Nature, Vol. 411, No. 6837, 2001. (4.5.5 項)
- Steve Lawrence and C. Lee Giles and Kurt Bollacker: "Digital Libraries and Autonomous Citation Indexing", IEEE Computer, Vol. 32, No. 6 , 1999. (4.5.5 項)
- Steve Lawrence and Kurt Bollacker and C. Lee Giles: "Indexing and Retrieval of Scientific Literature", CIKM'99 - Eighth International Conference on Information and Knowledge Management, 1999. (4.5.5 項)
- Kurt Bollacker and Steve Lawrence and C. Lee Giles: "Discovering Relevant Scientific Literature on the Web", IEEE Intelligent Systems, Vol. 15, No. 2 , 2000. (4.5.5 項)
- Steve Lawrence and C. Lee Giles: "Searching the Web: General and Scientific Information Access", IEEE Communications, Vol. 37, No. 1, 1999. (4.5.5 項)
- Andries Kruger, C. Lee Giles, Frans M. Coetzee, Eric Glover, Gary W. Flake, Steve Lawrence, Christian Omlin: “DEADLINER: Building a New Niche Search Engine”, CIKM 2000 - Ninth International Conference on Information and Knowledge Management, 2000. (4.5.6 項)
- Kushmerick, N.: "Wrapper induction: Efficiency and expressiveness", Artificial Intelligence, Vol. 118, pp.15-68, 2000. (4.6.1 項)
- Line Eikvil: "Information Extraction from World Wide Web - A Survey", Norweigan Computing Center, No.945, 1999. (4.6.1 項)
- Eric J. Glover, Gary W. Flake, Steve Lawrence, William P. Birmingham, Andries Kruger, C. Lee Giles and David M. Pennock: “Improving Category Specific Web Search by Learning Query Modifications”, Symposium on Applications and the Internet, SAINT, 2001. (4.6.2 項)
- Chakrabarti, S., van der Berg, M and Dom, B.: "Focused crawling: a new approach to topic-specific Web resource discovery", *Computer Networks*, Vol.31, No.11-16, pp.1623-1640,

1999. (4.6.3 項)

- Diligenti, M., Coetzee, F.M., Lawrence, S., Giles, C.L. and Gori, M.: "Focused Crawling using Context Graphs", 26th International Conference on Very Large Databases, VLDB 2000, pp.527-534, 2000. (4.6.3 項)
- 松田勝志, 福島俊一: “文書タイプ分類による問題解決向き WWW 検索システムの開発と評価”, 情報処理学会研究報告, FI-53-2, 1999.(4.6.4 項)
- K. Matsuda and T. Fukushima: “Task-Oriented World Wide Web Retrieval by Document Type Classification”, CIKM’99 - 8th International Conference on Information and Knowledge Management, 1999. (4.6.4 項)
- 山田洋志, 福島俊一, 松田勝志: “Web ページからのタイプ別情報抽出・分類方式”, 情報処理学会研究報告, FI-57-19, 2000. (4.6.4 項)
- 有吉勇介, 福島俊一: “目的および個人に特化したサーチエンジンの開発”, 人工知能学会誌、Vol. 16, No. 4, 2001. (4.6.4 項)
- E. J. Glover, G. W. Flake, S. Lawrence, W. P. Birmingham, A. Kruger, C. L. Giles and D. M. Pennock: “Improving Category Specific Web Search by Learning Query Modifications”, SAINT 2001 – Symposium on Applications and the Internet, 2001. (4.6.4 項)

４.２ 検索エンジンの現状

本委員会では、2003 年 6 月 26 日に、早稲田大学理工学部情報学科 山名早人助教授から Web 検索エンジンの最新技術動向に関するヒアリングを行なった。以下はそのヒアリング内容をまとめたものである。

（１） Web ページ数の調査

インターネットの普及の度合いを調べる場合でも、検索エンジンのカバー率を調べる場合でも、現在どの程度の Web ページが存在しているのかという情報がまず必要となる。このように、Web ページ数は Web の現状を把握するための最も基本的なデータであるといえる。Web ページの調査方法としては、調査対象が膨大であり実数調査が困難なため、Web サーバ数といった他の統計値やサンプリング調査の結果を組み合わせることにより統計的に推定する方法が一般的となっている（調査方法自体が研究テーマとなっている）。

Web ページ数推定では Lawrence らによる調査研究が有名である。Lawrence らは、大規模な検索エンジン間のデータの重なりから、1997 年 12 月時点での Web ページ数を 3.2 億ページと推定している[1]。1999 年 2 月には、360 万の IP アドレスに対して 80 番ポート（Web サーバが使用するポート）をチェックし、2500 の Web サーバから実際に Web ページ収集を行なうことにより、8 億ページ、15TB という推定値を得ている（推定サーバ数×１サーバ当たりの平均 Web ページ数・サイズから算出）[2]。なお、Netcraft の調査[3]によれば、2002 年 5 月時点での Web サーバ数（サイト数）は、約 3800 万であり、1 サーバ当たりの平均ページ数を約 190 ページ([1][2]での値)とすると、約 70 億ページとい

う推定値が得られる。

国内の Web ページ数に関しては、来住らが、国内の検索エンジンを利用した Web ページ推定方法により、1999 年 11 月時点で 1 億 2000 万ページ[4]、2000 年 10 月時点で 2 億 5600 万ページ[5]との推定値を得ている。西村らは、jp ドメインからランダムに Web サーバを選択して Web ページを約 5 週間に渡って収集。第二レベルドメイン（ad, ac, co, go,…など）毎に平均 Web ページ数を計算することにより 3 億 1300 万ページとの推定値を得ている。

（2） Web 検索エンジンの研究動向

現在の一般的な Web 検索エンジンは、以下の 3 つのデータ処理・要素技術により構成されている。

- Web ページを収集する「Crawling」
- 収集したページを格納して検索用インデックスを作成する「Storage & Indexing」
- ユーザからのクエリに対して順位付けされた結果を返す「Ranking & Link Analysis」

以下では、それぞれの要素技術についての技術課題と研究動向を整理する。

(a) Crawling

Web ページは、以下のような特徴を持つといわれている。

- データ量が膨大（約 70 億ページ）
- 毎年ほぼ倍増
- 23%のページが Daily Change[7]
- 10 日で半分のページが更新[7]
- リンク構造は蝶ネクタイ構造[8]

データ量が膨大であり、かつ更新が頻繁に行なわれるため、単純に収集していく戦略では収集速度が追いつかず、いかに効率的に収集・更新するかが大きな技術課題となっている。

Web ページの効率的な収集方法としては、重要なページ（あるいは特定の話題に関連するページ）を優先的に収集していく方法がいくつか提案されている[9][10]。これらの方法では、

- アンカー文字列
- バックリンク数などのリンク情報
- URL の形式（index.htm など）

などの情報を用いることにより、これから集めようとするページの重要性・関連性を、ページの内容を見ずに判断（推定）する。この推定がある程度の確からしさで実現できれば、不要なページを飛ばして効率的にページ収集を行なうことが可能となる（実際にそのような効果が得られるものと報告されている）。Web ページは絶えず更新されているので、ただ収集するだけでなく、一度収集したページをいかに効率的に最新の状態に保つかも重要な課題になっている[11]。

(b) Storage & Indexing

収集した Web ページを検索できるようにするためには、データの格納・インデックスの作成を行う必要があるが、ここでは、数十 TB(未圧縮時)のデータをいかに効率よく格納するのか、数十億の Web ページに対していかに高速にインデックスを作成するかが技術課題となる。Melnik らは、イン

デックス作成処理を、メモリにページの内容を読み込む「loading」、単語抽出・ソート処理を行なう「processing」、ソート結果をディスクに書き込む「flushing」の 3 フェーズに分割し、パイプラインで実行する方法を提案している[12]。processing では CPU、flushing では二次記憶といったように、各フェーズで主に使用する計算機リソースが異なるため、パイプライン処理が有効であるとしている。

(c) Ranking & Link Analysis

Web 検索エンジンに入力される検索語は、通常１～２語程度であり、検索語のみでランキングを行なうことは難しい。そのため、Web が持つハイパーリンク構造を用いてランキングに使用するページ重要度を計算する方法が研究されており、Google などの検索エンジンで実用化されている。

Kleinberg による HITS（Hyperlink-Induced Topic Search）は、特定の話題に関する、有用な情報を沢山含む Authority ページ（有用ページ）と、Authority ページへの沢山のリンクを持つ Hub ページ（有用リンク集）を見つけ出すアルゴリズムであり、繰り返し計算により、特定の検索語で検索されたデータセット（数千ページ）についてのページ重要度を計算する[13]。

Google のランキングに使用されている PageRank 法[14][15][16]は、被リンク数が多いページは重要であり、重要なページからリンクされているページは重要であるという考え方に基づく方法である。ページ A からページ B にリンクが存在する場合には、ページ B にページ A が一票投じていると考える。ページ A の重要度によって、一票の重みを変えることにより、ランクを上げるために複数のダミーサイトからリンクを張っているようなケースを排除している。

[参考文献]

[1] S. Lawrence, C. L. Giles: “Searching the World Wide Web”, Science, Vol.280, No.5360, pp.98-100 (1998).

[2] S. Lawrence, C. L. Giles: “Accessibility of Information on the Web”, Nature, Vol.400, pp.107-109 (1999).

[3] Netcraft Home Page, http://www.netcraft.co.uk/.

[4] 来住伸子、大森貴博、笹塚清二、近藤晶子、水谷正大、小川貴英: “統計的推定による日本語 Web の調査”、インターネットコンファレンス’99 論文集、pp.21-28 (1999)。

[5] 来住伸子、大森貴博、水谷正大、小川貴英: “検索エンジンを利用した日本語 Web ページ数の統計的推定”、情報処理学会論文誌、データベース、Vol.42, No.SIG 8、pp.47-55 (2000)。

[6] 西村真幸、山名早人: “ドメイン毎の Web ページ数の偏りを考慮した日本の Web ページ数推定調査”、情報処理学会 第 64 回全国大会、2X-6 (2002)。

[7] J. Cho and H. C. Molina: “Estimating Frequency of Change”, Technical report, Stanford University Computer Science Department (2000).

[8] A. Broder, R. Kumar et.al.: “Graph structure in the web: experiments and models”, Proc. of the 9th WWW Conf. (2000).

[9] A. Arasu, J. Cho, H. C. Molina, A. Paepcke and S. Raghavan: “Searching the Web”, ACM Trans. on Internet Tech., Vol.1, No.1, pp.2-43 (2001).

[10] M. Diligenti, F. M. Coetzee, S. Lawrence, C. L. Giles and M. Gori: “Focused Crawling Using Context Graphs”, Proc. of the 26th International Conf. on Very Large Database, pp.527-534 (2000).
[11] J. Cho and H. G. Molina: “Synchronizing a database to improve freshness”, Proc. of International Conf. on Management of Data, pp.117-128 (2000).
[12] S. Melnik, S. Raghavan et.al.: “Building a Distributed Full-Text Index for the Web”, WWW Conf. 10 (2001).
[13] J. Kleinberg: “Authoritative sources in a hyperlinked environment”, J. of the ACM, Vol.46, No.5 (1999).
[14] L. Page, S. Brin, R. Motwani and T. Winograd: “The PageRank citation ranking: Bringing Order to the Web”, Stanford Digital Library Technologies, Working Paper SIDL-WP-1999-0120 (1998).
[15] M. Henzinger: “Link Analysis in Web Information Retrieval”, IEEE Bull. of the Tech. Committee on Data Engineering, Vol.23, No.3, pp.3-8 (2000).
[16] T. H. Haveliwala: “Efficient Computation of PageRank”, Stanford Digital Library Technologies, Technical Rep. 1999-31 (1999).

４.３ 専門分野 Web 検索の定義と分類

本報告書では、専門分野 Web 検索機能を提供するサイト（専門分野検索サイト）を、次のような 2 つの条件を満たす Web サイトであると定義する。条件 1 は専門分野を対象とすることを意味し、条件 2 は検索サイトであることを意味している。

［条件 1］限定した目的で広くコンテンツを集めていること。

［条件 2］コンテンツの検索・ナビゲーション機能を備えていること。

この定義によれば、Google や Yahoo!のような Web サーチエンジンは、対象コンテンツの目的が限定されておらず汎用であることから、条件 1 を満たさず、専門分野検索サイトには該当しない。また、サイト内検索機能を備えた企業サイトは、目的は限定されているが、対象コンテンツが単一で広く集められていないために、やはり条件 1 を満たさず、専門分野検索サイトには該当しない。

具体的に専門分野検索サイトの例をあげると、まず、ショッピングの分野では、価格.com（http: //kakaku.com/）、MySimon（http: //www.mysimon.com/）、DealTime（http: //www.dealtime.com/）などの価格比較サイトがよく知られている。これらのサイトでは、ジャンル別に分けられたなかから商品カテゴリや商品名を選ぶと、該当する商品がどこのショップサイトでいくらで売られているかを、比較表の形で提示してくれる。国内の価格.com は商品情報・価格情報を人手で集めているが、海外の MySimon や DealTime では、Shopbot（ショッピングエージェント）の技術を用いて、様々なショップサイトから商品情報・価格情報を自動収集・抽出する方式がとられている。

別なタイプの専門分野検索サイトとして、各種学校のホームページを集めた SCHOOL NAVI（http://www.schoolnavi-jp.com/）があげられる。このサイトでは、全国各地の大学・短大、高校、

中学校、小学校などのホームページを、都道府県別に分類して掲載している。

また別なタイプとして、インターネット上に公開されたコンピュータサイエンス分野の論文のサーチエンジンである ResearchIndex（CiteSeer ともいう、http://citeseer.nj.nec.com/cs）がある。このサイトでは、ユーザが入力したキーワードや人名にマッチする論文（HTML や PDF の形式のファイル）が Web から検索される。Google などの一般の Web サーチエンジンがあらゆる種類の Web ページを検索対象としているのに対して、ResearchIndex では、論文のみが検索対象になっている。また、論文に特化することで、引用関係の分析など、対象により特化した検索機能も提供されている。

このような専門分野検索サイトは、そのサイトで提供するデータの作成方法という観点から、いくつかのタイプに分類できる。ここでは、データが自作か、他から集めて加工するか、という度合いに着目して 4 つのタイプに分けた。

［タイプ A］他者のデータに対して選別・スコア付け

［タイプ B］他者のデータに対して分類・メタデータ付与

［タイプ C］他者のデータに対して情報抽出・データベース化

［タイプ D］データを自作あるいは共同作成

タイプ A は、汎用 Web サーチエンジンでいえば Google のようなタイプで、外部の Web ページを収集・インデクシングし、ユーザの検索キーワードに応じて選別・スコア付けして提示する。上述の専門分野検索サイトの例では、ResearchIndex がこのタイプ A に該当する。

タイプ B は、汎用 Web サーチエンジンでいえば Yahoo!のようなタイプで、外部の Web ページをカテゴリ別に分類したり、概要紹介文などを追加したりする。上述の専門分野検索サイトの例では、SCHOOL NAVI がこのタイプ B に該当する。

タイプ C は、外部の Web ページから着目する情報の部分のみを抽出し、自分のデータベース内に登録する。上述の専門分野検索サイトの例では、MySimon や DealTime がこのタイプ C に該当する。

タイプ D は、自前でデータベースを作成する、あるいは、外部からデータの提供を受けてデータベースを作成する。上述の専門分野検索サイトの例では、価格.com がこのタイプ D に該当する。

データの作成という観点では、さらに、上記のような 4 種類の方法が人手で行われるか、自動化されているか、という区別も考えられる。

また、別な観点として、どのような検索・ナビゲーション機能が提供されているかも着目すべき点である。汎用 Web サーチエンジンでは、キーワード検索とディレクトリ（ジャンル別分類）が基本的な機能として提供されているが、専門分野 Web 検索では対象を特定のものに絞り込んでいるため、その対象に特化した検索・ナビゲーション機能を提供することが可能になる。例えば、商品の価格を条件とした絞り込み機能や、学校の所在地による分類機能や、論文の引用関係の参照機能などが、その一例である。

現在、インターネット上には様々な目的に対する多数の専門分野検索サイトが存在し、ここでこれらを列挙し切れるものではないが、下記の表に、それらのいくつかをあげておく。

表 4.3－1　専門分野検索サイトの例

専門分野検索サイト	ジャンル	タイプ	特記事項
価格.com http://kakaku.com/	商品情報・価格比較	タイプ D 人手	口コミ情報の掲示板も提供
MySimon http://www.mysimon.com/	商品情報・価格比較	タイプ C 自動化	海外のサイト
DealTime http://www.dealtime.com/	商品情報・価格比較	タイプ C 自動化	海外のサイト 日本にも進出したが既に撤退
FlipDog http://www.flipdog.com/	求人情報検索	タイプ C 自動化	海外のサイト 登録データ数は 30 万件弱
Ecareer http://www.ecareer.ne.jp/	求人情報検索	タイプ D 人手	登録データ数は 6 千件弱 職種や勤務地などから絞り込み可能
旅の窓口 http://www.mytrip.net	ホテル検索	タイプ D 人手	国内 10465 件、海外 4419 件 条件検索、ホテル予約も可能
ぐるなび http://www.gnavi.co.jp/	レストラン検索	タイプ D 人手	場所、種別、キーワードなどでの条件検索
@グルメぴあ http://g.pia.co.jp/	レストラン検索	タイプ D 人手	場所、種別、キーワードなどでの条件検索
SCHOOL NAVI http://www.schoolnavi-jp.com/	学校検索	タイプ B 人手	大学 737、短大 578、高専 62、高校 4500、中学校 4525、小学校 8641 など
この指とまれ! http://www.yubitoma.or.jp/	コミュニティ検索	タイプ D 人手	ウェブ同窓会
掲示板探し http://pochi.cside2.jp/sagasi/	コミュニティ検索	タイプ B 人手	掲示板をもつサイトのディレクトリ
ResearchIndex（CiteSeer） http://citeseer.nj.nec.com/cs	論文検索	タイプ A 自動化	全文インデックスとサイテーションインデックスを提供
Attayo ケイタイ 携帯から http://attyao.jp/	i モードページ検索	タイプ A 自動化	i モードページ 100 万件以上 地域別分類も自動化
OH!NEW? http://www.ohnew.co.jp/i/	i モードページ検索	タイプ B 人手	i モードサイト 62,290 件

４.４ 専門分野 Web 検索の実例

４.４.１ 価格.com

（１）日時　　平成 14 年 10 月 28 日（月）16:00-17:10

（２）場所　　（株）カカクコム（東京都台東区浅草橋）社内会議室

（３）応対　　（株）カカクコム：穐田（あきた）氏（CEO）、中川氏（システム開発部長）、久米川氏（広報室）

（４）ヒアリング内容

① 設立経緯・現状

6 年近く前、メルコの営業担当だった槙野氏が、営業先の秋葉原や量販店での競合他社製品の価格調査を通して、メーカ、消費者、販売店共通の関心事である「価格」情報の重要性を実感。これをヒントに入社 1 年で同社を退社後、創業。

設立当初は、店頭に直接出向き販売価格を調べて回ったが、すぐにネット上の販売価格調査に切替え、コミ、雑誌紹介で人気が上がり、送客数の増加とともにショップの対応が変わってきた。当初、各商品の価格の安い順に上位 15 店舗を掲載、価格が高い下位 5 店を毎週入れ替えていたが、人気が出てくるにつれて常時掲載してほしいというショップが増えてきた。現在、価格情報は店舗側からＩＤ・パスワード管理で直接登録・更新されている（ブランド、スポーツ等一部商品を除く）。

昨年 1 年でショップ数は 3 倍（1324 店、2003 年 1 月 1 日現在）、ページビューは 4 倍（約 1 億 5600 万 PV/月、約 316 万ユニーク IP アドレス/月、2002 年 12 月現在）。掲示板の書き込みは約 2,500 件／日、約 75,000 件／月（2002 年 12 月現在）。

② ビジネスモデル

収益の柱は 4 つ。

(a) ショップのエントリ料

掲載は基本は無料。一部の有料カテゴリ（パソコン、デジカメ等人気の高いもの）のみ、一定のエントリ料（＝ミニバナー広告出稿）を情報掲載の条件としている。

(b) メーカ直販サイトとの販売提携契約（アフィリエイト）

サイトを通じての購入の何%という形の成功報酬。

(c) 情報提供売り上げ

自動車保険各社商品の一括見積もりなど。個人情報が各保険会社へ提供される。

(d) 一般的なバナー広告

比率はエントリ料 25%、アフィリエイト 25%、情報提供料 25～30%、残りが一般バナー広告料。ただし月によって多少のばらつきはある。昨年まではショップからのエントリ収益がほとんどであったが、アフィリエイトと情報提供部分の成長により収益が延びた。

③ 質疑応答

●対象商品に関して

Q. 対象商品は何か？

A. 創業当初からのメイン商材は秋葉原商品。その後、現金問屋の扱い商品（ブランド品）。さらに、近接商業エリアのお茶の水のスポーツ用品に広がることになった。

Q. 取り扱う型番の選定は？

A. 基本的に新製品は全部取り扱う。定期的にメーカーサイトを見たり、ニュースリリース情報から新製品をチェックしたりしている。

Q. Web での販売に向いているものは何か？

A. 単価が高いもの、見なくて変えるもの、ブランドとして認知されているもの、比較で差がはっきりでるもの。当初は型番で買えるものだけかと思っていたが実際にはそれだけではなかった。

Q. ものとサービスとでは何か違うところがあるか？

A. サービスは実際にものを見なくても比較検討が容易である（クレームが少ない）、という点は違うかもしれない。

Q. 今後の取り扱い範囲の拡大は？

A. 一部中古品を取り扱うサービスも始めた。あとは秋葉原商品の深掘り。中古車査定は（すでに実施しているが）ユーザメリットがあるからか反応がいい。パソコンの買い取り査定は以前から行っている（買い取り希望の法人にメールで連絡する仕組み）。中古市場は大きくなってきているので Web 向きである。

●データ入力に関して

Q. データ入力はショップのみか？

A. 通信（ADSL など）やブランド、スポーツカテゴリの一部商品に関しては担当者が調査して入力しているものもある。

Q. リアルタイム性はどうか？

A. パソコンをはじめとするほとんどの商品はほぼリアルタイムに更新されている。ブロードバンド（ADSL など）は月に 2、3 回。ブランド、スポーツは毎週更新。

Q. ショップの虚偽入力の問題はあるか？

A. ない。現金問屋ではサービス用員が少ないので、不要な問い合わせが来ると手間がかかってしまうのでそういうことはしない。ただし、在庫が少なくてもデータに入れるので、人気が集中する商品などは問い合わせをした段階では既に「ない」という場合はある。

●口コミ情報（掲示板情報）・評価などに関して

Q. 口コミ情報はいつから始めたのか？

A. 最初は価格情報だけで、それから製品情報紹介を加え、その後に掲載を開始。当初ユーザーサービスとして始めたが予想以上に支持された。弊社は実購買率ランキングが高い（1 位は 2001 年 2 月日経流通新聞／日経ネットブレーン「e コマースサイト調査」。2 位は旅の窓口）が、これには口コミ情報の存在も大きいと思われる。

Q. 口コミ情報のチェックは？

A. 1 日 2 回、誹謗中傷は削除。半分匿名（IP アドレスが表示されるので）。

Q. 掲示板情報をマーケティングする考えはあるのか？

A. 野村総研と既に取り組んでいる。さらに広く展開することを考えている。現在は、どういうカテゴリを選んでいるか、トータル検索回数（何を購入したかは不明）などのユーザの検索結果データをショップに販売する形態を取っている。レジ POS のデータは実際に買った商品のデータなのに比べて、上記データはそれより前の（潜在）顧客の情報を提供できる点が特徴である。他に、お知らせメールを利用した「メーカ競合情報」を収集している。お知らせメールは商品を購入したら解除されるので、解除の際に実際にどの商品を購入したかアンケートをしている。勝ち負け表（単独指名か競合か）データが集まる。この情報をメーカに販売するかどうかはまだ検討中である。

Q. 野村総研とテキストマイニングの話をもう少し詳しく伺いたい。

A. 野村総研にはローデータを渡して、抽出された情報を野村総研を介してメーカなどに販売する（その何%かをもらう）というモデルを考えている。そのフィードバックは来ているが、システムの改良を伴うのと、販売相手によって作り込みも変わってくるので、まだ特に対応していない。

Q. 情報は一般にも販売しているということだが、販売価格は？

A. 最低 15 万円程度から。クライアント様のニーズにより価格は異なる。個人情報(IP アドレス)は出さない。同一人物の特定はしていない（ハンドルネームの一致による同定は可能）。解析に取り組んで試験的にでも実施したいと考えている。今年はこれに注力している。

Q. 商品情報ページの「製品評価」の尺度は誰が決めるのか？

A. ユーザの投票。ただし、当初予想していなかったことなのだが、一部熱狂的ユーザが意図的コントロールを試みているので仕様の変更を予定している。

Q. お知らせメールについて

A. 人気が高い商品は 1000 人ぐらいの登録がある。ショップが、商品をいくつ売るのにどれくらい価格下げればいいか、も登録者人数の表を見て類推できる。希望商品の掲示板に新しい書き込みがあれば自動でメールで知らせるというサービスもある。

●ロボット検索型価格比較サイトとの比較・運用コスト・自動化に関して

Q. DealTime が日本市場に参入（その後撤退）したが？

A. 当時は脅威を感じたが、結局先行メリットがものを言った。4 年先行していたため、十分にユーザの意見・クレームを拾い上げてサービス・機能に反映できていたこと、DealTime が網羅できなかった現金問屋の情報を扱っているのでこちらの方が安い価格が出せたこと、などが勝因となったと思う。DealTime の日本でのサービス開始時には、競合としてカカクコムの名前も出たので知名度は一緒に上がった。型番単位で価格情報を掲載している EC サイトは少ないので（カタカナのバイオしか探せないとか）手入力によるデータ整理が重要と考える。また、価格の網羅性、正確性もポイントである。販売店側は型番入力等の手間無く価格を入力するだけなので間違いは少ない。

Q. 規模拡大に伴って手作業ではコストがかかりすぎるのではないか？

A. 規模に比例してコストがかかるわけではない（見る人が倍になっても対応する人数を倍にしなくてもいい）。DB 外注や広告に金をかけるよりも自社スタッフを充実させる方がトータルでは安い。担

当者が商品や業界の仕組み､動向に詳しくなるので手作業に意味はある｡価格の更新が一番大変だが、それはショップが行う。ショップは顧客獲得というモチベーションがあるので積極的に行っている。例えば、他のショップが更新しない時間（夜中など）に変えたりしている。部門別の収益管理を今後は考えたい。

Q. 集中した処理をさばくのには自動化が必要ではないか？

A. 手作業を補足する意味では機械もほしいが、本質的にはあまり必要性を感じていない。

●その他

Q. エントリ有料化の基準は？

A. 販売店との話し合いを通じて決定した。事前にプランがあったわけではない。カテゴリに集まるお客さんの数にもよる。まずは無料で参加してもらい、価格表に多数店の価格情報が掲載できるようになってから有料化を考える、という形を取っている。ショップと一緒にコンテンツを作っている、という姿勢で臨んでいる。ショップの中には年商 100 億円と言われるところもある。

Q. ユーザからのフィードバックにはどのように対応しているか？

A. ひとつひとつ細かな反映はその都度実施している｡商品やカテゴリの追加の際にはユーザの意見が大きな決め手になる。ただし、（デジカメでない一般の銀塩カメラも含めた）カメラのカテゴリ追加のときは売り手（デジカメだけでなく一般のカメラも売っているカメラ屋）のニーズにも後押しされた。ユーザの問合せ・サポート窓口はカテゴリ別。

Q. コンテンツ運営・管理は担当者ベースの方がよいか？

A. 例えばパソコンと ADSL では比較検討の道筋が違うので、担当者ベースの方がよい。新製品ニュースのページも作っているが、それぞれの商材把握や業界動向を知るのに必要。各カテゴリ別に担当者が決まっている。社員は 25 名（派遣も含めて）、コンテンツ維持だけなら 15 人。情報収集者がページの運営管理やショップ対応、ユーザーサポートまで担当する。

Q. ハードは自前か？

A. 自社で購入し、有明のサーバセンターに設置している。管理は自社で行っている。

4.4.2 モバイルインフォサーチ

モバイルインフォサーチ(http://www.kokono.net) [1]とは、NTT 情報流通プラットフォーム研究所が実験プロジェクトとして開発した、実世界の位置情報をキーとした情報検索の草分け的サイトである。ユーザは自分の現在地から半径 500 メートル以内のレストランやお店をインターネットで探すことができるといった、実世界とインターネットを位置情報により結び付けていることが最大の特徴である。以下では、まず研究の背景について述べてから、モバイルインフォサーチの実験のきっかけとなったインターネットタウンページについて紹介した後に、モバイルインフォサーチの特徴である位置指向のインターネット情報検索の仕組みを紹介する。

今日でこそ携帯電話によるエリアナビゲーションやカーナビを用いたテレマティクスが普及しつつあるが、元々インターネットにある情報は実世界における地理情報を条件として検索することはできないという問題があった。例えば、出張や旅行、食事や買い物の計画を立てるときに、現在地、移動

先や観光地、レストランや店舗に関する情報は、個別の特定の情報源を頼りにせざるのを得ないのが実情であった。一方、ユーザの位置情報を特定する技術の進歩により、GPS(Global Positioning System)や携帯電話の基地局 Cell 情報（PHS 等）に代表されるインフラの整備で、ユーザは自分の現在地が簡単にわかるようになった。そこで、インターネットにおいて位置情報を含む Web ページを収集して、ユーザの位置情報を考慮して地理的に近い店舗等の検索を実現することがモバイルインフォサーチのモチベーションとなっていた。

モバイルインフォサーチの出発点となったのは、インターネット上で電話帳情報のサービスを提供しているインターネットタウンページ(http://itp.ne.jp)である。インターネットタウンページには、全国 1,100 万件のタウンページデータが格納され、その業種は約 2,000 種類に上る。ここで特に課題となったのは、電話帳と地図の統合である。つまり、検索された店舗の情報を表示するために地理情報（緯度経度）が必要であるし、逆に検索を行う条件にも必要である。そこで、インターネットタウンページでは、電話帳に記載されている住所と住宅地図に記載されている住所文字列のマッチングを行うことからスタートした。このような経緯を経てサービスが提供されているインターネットタウンページは、日本におけるインターネットでの位置指向検索サービスの最初のひとつであり、現在のモバイル系サービスの布石となった。これをさらに電話帳だけでなくインターネット上の Web ページへと対象を広げたものがモバイルインフォサーチである。

モバイルインフォサーチを実現するための位置指向検索エンジン「ここのサーチ[2]」の一連の仕組みは次のようになっている。

（1） ロボットによる Web ページ収集

インターネットに存在する無数の Web ページからロボットを用いて位置情報を含むページを効率的に収集する。アンカー部分に地名が含まれる場合は、参照先の Web ページも位置情報を含む割合が高いので、アンカー部分の地名の有無を解析することにより、位置情報を含むページを優先して収集している。

（2） 位置情報の抽出

Web ページに存在する位置情報の代表的な例は住所である。住所の抽出には、いくつか問題点がある。

- 同一の単語でも地名にも人名にも使われる（例：千葉県と千葉さん）
- 住所がユニークであることを保証しなければならない（銀座だけでは全国に多数存在）
- 丁目表記の揺れを統一しなければならない（銀座 1-2-3 と銀座一丁目二番三号）

これらの問題に対処するために、まず形態素解析を行い、都道府県から始まる前方完全表記を採用し、市区町村の接尾辞が付いているものを対象とし、丁目等の表記はルールを用いて統一して照合する手法により、住所を高精度に収集している。

（3） Web ページへの緯度経度の付与

Web ページに含まれる住所やランドマーク、駅、郵便番号等から、地図情報を基にした位置データリポジトリを参照して Web ページに緯度経度を付与する。

（4） 地理的検索

緯度経度を利用して、特定地点からの半径を指定することにより検索結果の絞込みを行い、特定地点と検索対象までの距離によりランキングされた結果を表示する。検索結果が 100 件程度になるように検索半径を決める手法を採用している。単純な市町村名での検索よりも高い再現率を実現していて、特に距離的には近いが市町村名が異なる店舗の検索に有効である。

モバイルインフォサーチにおける実験では、同じ位置に対して天気に続いてお店検索を行うといった複数サービスの横断検索が同一ユーザにより短時間に行われることが明らかになり、位置指向の情報統合というニーズがあることが確認されている。また、ロボットが収集した Web ページの緯度経度を日本の白地図にプロットした結果、都市部に Web ページが集中していることが明らかになっている。さらに、都道府県別に見た人口と Web ページ数の関係では、人口が多い県ほど Web ページ数も多いが、北海道や京都などの観光地を持つ都道府県には人口に比べて多くの Web ページが存在することがわかるなど、大変興味深い知見が得られている。

[参考文献]

[1] Takahashi 他: “Mobile Info Search: Information Integration for Location-Aware Computing”, 情処論, Vol.41, No.4 (2000).

[2] 横路他: “位置指向の情報の収集、構造化および検索手法”、情処論、Vol.41、No.7 (2000)。

4.5 専門分野 Web 検索のシステム技術

4.5.1 専門分野 Web 検索システムの概要

本項では、参考文献 [1] を元に、専門分野 Web 検索技術の必要性、分類、展望について述べる。

現在、ほとんどのユーザは Web 上で必要な情報を探すために、Google や Yahoo などに代表される汎用検索エンジンを用いている。しかし、インターネット上には 5,000 億以上の Web ページが存在するという説もあり、これらの汎用検索エンジンがインデックスを持っているページは高々20 億ページ程度にすぎない。さらに、最近では、バックエンドにデータベースシステムを備え、動的にページを生成するようなサイトも増えている。これらのページに関してはあらかじめクローラーによってインデックスを生成しておくことはできない。Web ページの多様化、肥大化によって従来の汎用検索エンジンのアプローチでは、技術的にも経済的にも Web 上の情報を十分に網羅することが不可能になっている。

この問題を解決するためのひとつのアプローチとして検索の対象分野を限定することが考えらえれる。ここでは、これを専門分野 Web 検索と呼ぼう。検索を対象を限定することによって、インデックスの更新の周期を短くし、情報の即報性を高めたり、より深くリンクを辿ることによってより多くの情報を収集できる。また、分野に特化することによって、分野独自のインタフェースの提供や分野の知識を利用した検索支援なども考えられる。

4.3 節では、専門分野 Web 検索を主にユーザの観点から分類したが、ここでは、以下の 2 つの技術的な観点から分類する。ひとつはインデックスが分野に特化しているかどうか、もうひとつはクエリ

の処理が分野特化しているかどうかである。

インデックスの分野特化に関してはさらに 3 つのアプローチに細分化できる。ひとつは Web ページを収集する時点で分野特化をする方法であり、このためには特定分野の情報を含むページだけを収集する Focused Crawling の技術などが用いられる。Focused Crawling の技術については 4.6.3 項を参照されたい。また、このアプローチを採用したシステムの例としては、4.5.5 項の Research Index や 4.5.6 項の DEADLINER などがある。その他にも以下のようなシステムが実現されている。

・HPsearch (hpsearch.uni-trier.de/hp)：個人の Web ページの検索
・Moreover (www.moreover.com)：最新のニュースの検索
・FlipDog (www.flipdog.com)：求人情報の検索

インデックスを生成する際に分野を特化することの利点は、単にページから「単語」を取り出すのではなく、情報抽出の技術を利用してより高精度で必要な「情報」を取り出せるという点である。情報抽出の技術がうまく機能するためには、分野特定がされていることが必要であるが、それに加えて、特定の分野のページの記述方法がある程度定型化しており、ページの構造を利用することによってより高い抽出精度を得ることができる。

また、分野を特定することによってより深いリンクまでたどることができるので、より詳細な情報を収集可能となる。たとえば、Moreover のクローラーは 1,800 のサイトを 1 時間に 4 回程度探索すると言われている。あるいは、掲示板や特定のユースグループを監視して、Web ページへのリンクを含む投稿があったら、そのページをすぐに探索に行くというような技術も使われている。このような手法は即報性が要求される分野では必要不可欠なものである。

インデックスの分野特化に関する 2 番目のアプローチはメタ検索により分野特化をしようとするものである。CompletePlanet の調査によれば、Web 上のデータベースサイトは 200,000 以上あるといわれており、これを単独の検索エンジンでメタ検索するのは現実的ではない。したがって、これらをメタ検索しようとすれば、必然的に分野を限って、その部分集合のサイトに対してメタ検索をおこなうことになるたとえば、4.3 節でも例として紹介されている、MySimon はこのアプローチをとっているサイトである。MySimon は商品比較・購買支援をおこなうサイトである。機能的には 4.4.1 項で紹介している価格.com と同じであるが、アプローチがまったく違う点は興味深い。MySimon や同種の DealTime は検索サイトが自動的に情報を収集に行くのに対して、価格.com は商品の提供者（人間）が情報を提供し、サイト自身はその情報交換の場を提供しているにすぎない。商品提供者にとっては少しでも安い値段を提示して顧客を得るというインセンティブがあり、これをうまく利用したビジネスモデルになっている。DealTime が一時期日本にも上陸したが、ほどなく撤退したという事実がある。これが文化による違いなのかどうか興味深いところである。

メタ検索は既存の情報源をうまく組み合わせて、より豊富な情報を提供できるという利点がある反面、その性能が一次情報提供サイトの中の最低の性能に抑えられてしまうという限界がある。また、一次情報提供サイトのインタフェースが頻繁に変るとシステムの保守にコストがかかるという問題もある。

インデックスの分野特化の最後のアプローチは実時間でクローリングするというものである。これは分野特化していて、探索範囲が狭いからこそ可能になる。

インデックスの分野特化に対して、クエリの処理に分野特有の知識を使うアプローチがある。これらの 2 つのアプローチは直交しており、同時に使用することも可能である。クエリに対する処理としては、クエリ中の語を類義語の集合で置換するクエリ拡張が代表的であるが、一般にクエリ拡張をする際には専門分野のシソーラスを利用するほうが性能が改善されることが知られている。検索対象分野を限定することはこのような観点からも有効である。また、クエリのタイプを分類し、その答えを含みそうな表現にクエリを拡張するアプローチもある。これをつき進めればいわゆる QA システムに近いものになるだろう。

このように Web 上の情報の多様化、肥大化は汎用の Web 検索から専門分野 Web 検索の移行を必然的なものにすると思われる。しかし、ユーザの観点から見ればさまざまな専門分野 Web 検索サイトが存在し、その中から適宜適切なサイトを選んで検索をするよりは、万能(汎用)の検索エンジンがひとつあったほうが使い勝手がよい。したがって、専門分野 Web 検索はいかに普及させるかが現実的な課題となる。そのためのアプローチとしては、Yahoo などのディレクトリのように、専門分野 Web 検索のサイトのディレクトリを用意してユーザの選択を支援する、専門分野 Web 検索のためのメタ検索サイトを用意して、見かけ上はひとつの検索エンジンのようにみせかける、既存の汎用検索エンジンから必要に応じて専門分野 Web 検索を呼びだす、などが考えられる。

[参考文献]

[1] Robert Steele: “Techniques for Specialized Search Engines”, in Proceedings of Internet Computing (2001).

4.5.2 Shoping Agent

論文[1]は、インターネットショッピングを支援するソフトウェアエージェント(ShopBot)の構築についての最初の実例研究(Case Study)である。具体的には、ShopBot は、Web 上のショップの商品について、その価格の比較と商品説明をしてくれるものである。

論文では、まず Web を扱ったエージェントが何を考慮すべきかを述べたのち、ショッピング支援エージェントが何を支援すべきかを考察した後、そのような問題を解決した ShopBot の構築方法を紹介し、その有効性を、ショッピング時間とユーザが見つけることができた商品の最低価格を基に示している。

Web の情報を扱ったエージェントを設計するときの考慮すべき問題として、以下の項目を挙げている。

・Ability：HTML で書かれたコンテンツからどのように情報を取得するか？

・Utility：ユーザに有用な情報をいかにして提供できるのか？

・Scalability：新たに出現する Web サイトの情報を自動的に扱うことができるか？

・Environmental Constraint：エージェントは、自然言語をどれくらい理解できるか。
どれくらいドメインに特化した知識を持つべきなのか？

ShopBot では、Utility を実現するために、価格を比較するとともに、製品情報を提示する。また、Ability や Scalability を実現するために、HTML 文章から商品の価格や説明などの情報を、ドメインに依存しない方法で取り出す方式を採用している。この方式は、特別なドメイン知識や言語処理系を必要とせず、特定のサイトや商品情報に依存しない方式である。

ShopBot は、特定の Web サイト上の HMTL 文書から商品の情報を摘出する仕方を学習するための ShopBot Learner モジュールと、学習した結果を使って、ユーザに商品を提示する ShopBot Buyer モジュールから成る。

ShopBot Learner モジュールは、与えられたサイトのすべての商品リストを取得する方法を学習する部分と、検索した商品からその価格や商品説明を取得する方法を学習する部分とからなる。商品リストを取得するために、Learner には、商品の検索ページの登録と、そのページで指定すべきパラメータを指定する。また、一致する商品が存在しなかった場合のページもあわせて指定する。検索した情報から、個々の商品の情報を取得するためには、HTML のタグとテキストの並びをパターンで指定する。これによって、サイトごとがページのフォーマットが違っていても、正しく商品情報が取得できるように学習可能となる。

ShopBot Buyer モジュールは、ユーザが指定した商品について、あらかじめ登録し、学習しておいたサイトを検索し、その商品情報、価格情報を抜き出し、リストにして表示を行う。この検索のための学習には時間がかかるが、検索自体には、さほど時間はかからない。

本論文では、ShopBot の有用性の評価として、Web 上の実際のソフトウェアショップで、ある商品を検索した時の検索時間と、そのときに見つけた最低価格について調べている。具体的には、ShopBot を使って検索した場合と検索エンジンのみを使った場合、ショップの URL を与えた場合の３つの場合で比較している。その結果、ShopBot を使うと、安い商品を他の方法より早く見つけることができることがわかった。また、実際に最安値のショップを見つけることができることも確認できた。

一方、論文[2]では、商品の売買がすべてエージェント化された世界を想定し、商品を紹介するエージェント(Shopbot)の情報がどれくらいの価値を持っているのかを評価している。実際には、Shopbot が各ショップのサイトから収集してきた情報に対して価値を決定し、それらを購入エージェントに対して売買するエージェント(infoseller)を導入し、infoseller の機能を評価している。論文では、実際のオンライン書店を元に、商品の安さと実際のシェアは一致しないことから、商品の売買は、価格以外にブランドを考慮しないといけないことを考察し、ブランドを考慮した販売者の価値(utility)を扱っている。

この価値のモデル化は、Brynijolfsson と Smith らによる論文[3]で提案されたものである。論文[3]では、ブランドを商品の到着時間の短さとして、その価値を金額に換算して計算し、実際のシェアと価格とのギャップを説明している。しかしながら、ブランドとして考えられる要素には、郵送コストや、過去の売買経歴、宣伝なども考えられ、それらの要素との関係については課題としている。

論文[2]では、まず、購入エージェントに対するショップ情報の価値についてモデル化した後、infoseller がどれくらいの価格で情報を売るべきか(適正価格)をモデル化している。

商品の情報を提供する価値は、具体的には、その商品の期待値(価格)と、購入エージェントがそれまで収集している商品の値段によって決まる。この論文では、まず、購入エージェントが、(a)あるショップ(i)の商品価格を既に知っていたとき、他のひとつのショップ(j)の価格を知る時の価値についてと、(b)あるショップ(i)の製品価格以外すべてのショップの価格を知っていたときにそのショップ(i)の価格を知る価値についてモデル化を行っている。(a)の場合、平均価格よりも安いショップを既に知っていたとき、高いショップを知った場合の価値は、高いショップを既に知っていた場合に安いショップを知るときの価値よりも低い。これは、購入エージェントが持つすべての商品情報の平均 utility 値の変化の度合いでモデル化している。また、(b)の場合、すでに多くのショップの情報を知っていたとき、他のショップの情報の価値は低くなるようにモデル化している。

これらのモデル化に基づいて、購入エージェントが既に知っているショップの情報に対して、infoseller が他のショップの情報をいくらで売ればよいのかの適正価格を考察している。最後に、ショップの数が多くなり、その選択の価値基準にいろいろある場合、商品情報を売買することは有効であろうと結んでいる。

[参考文献]

[1] Robert B. Doorenbos, Oren Etzioni, Daniel S. Weld: "A Scalable Comparison-Shopping Agent for the World-Wide Web", In Proceedings of International Conference on Autonomous Agent (1997).

[2] Panos M.Markopoulos, Jeffrey O. Kephart: "How valuable are shopbots?", In Proceedings of International Conference on Autonomous Agents (2002).

[3] E. Brynjolfsson and M. D. Smith: “Frictionless commerce? A comparison of internet and conventional retailers”, Management Science, 46(4) (2000).

４.５.３ Webwatcher

Web 利用上におけるユーザの興味を理解し、必要な情報をユーザに提供するための、ソフトウェアエージェントによる支援の仕方には、ユーザとの対話を繰り返しながらナビゲーションしたり、Web 上のページのコンテンツに関する知識や、ユーザの利用履歴を利用してユーザの興味をあらかじめ学習しておくなどの様々なアプローチが存在する。

ここでは､両方の特徴をもつ WebWatcher に焦点を当てる｡WebWatcher は常にユーザと共にあり、ユーザの行動を観察、学習し、ユーザの求める情報が掲載されているページに、最短経路で到達できるようにナビゲーションするものである。WebWatcher が、一般的な Web 検索エンジンなど(Google, Lycos etc.)と異なる点は、

1. ユーザの要求を的確に述べるキーワードを探す必要が無い。
2. Web 上のコンテンツ間の関係性を考慮している。

と言う点である。

ユーザの興味にかなうハイパーリンク先を紹介するためには、主に、過去のユーザのクリックストリームからキーワードを蓄積しておく方法と、クリックストリームから集めた情報量が最大化するよ

うなツアーを強化学習する方法があるが、WebWatcher はこの 2 つの方法をマージした手法を提案する。

詳細は以下の通りである。

まず、過去のユーザのクリックストリームからキーワードを蓄積しておく段階においては、過去のクリックストリームに存在するサイトに対して、ハイパーリンク中に登場する語を対象に tf-idf 法で重み付けし、そのうち上位 k(default:5)個をハイパーリンクの代表語としておく。これによって、後のナビゲーションの際、現在の検索先との類似度を距離計算によって評価し、上位のリンク先を関連ページとして推薦できるようになる。

次に、クリックストリームから集めた情報量が最大化するようなツアーを強化学習する段階においては、興味に関連する情報が最大になるような Web 上の検索パスを見つける。この最適パスの発見に関しては、強化学習を採用する。

具体的には、Web 上のあるページ s にアクセスしている際、エージェントは、一定の報酬 $R(s)$を受け取る。また、ある状態から起こされるアクション a は評価関数 $Q(s,a)$で評価され、この値は、エージェントが状態 s においてアクション a を実行し、その次に最適なアクションを選択した際に得られる将来的な報酬の合計に依存し、現在の状態 s の次の状態 s'を導入とすると、、以下の式で求められる。

$$Q_{n+1}(s,a) = R(s') + \gamma \max_{a' \in actions\_in\_s'} [Q_n(s', a')]$$

一度 $Q(s,a)$が決定されれば、エージェントは、現在の状態 s における $Q(s,a)$が最大になるようなアクション a を繰り返し選択し行動するように制御される。

本手法の効果を確かめるため、ランダムにページを推薦する方法(Random)、ユーザの興味に関係なくクリックストリーム中で最もよく参照されているページを推薦する方法(Popularity)、tf-idf 法に基づいて計算された文書ベクトルが近いと判断されたページを推薦する方法(Match)、ハイパーリンク中に登場する語を代表語と決めて置き、近いと判断されたページを推薦する方法(Annotate)、強化学習によりページを推薦する方法(RL)と、今回のシステムで採用した、前記すべてを考慮した方法(Combine)の比較実験を行った。実験は、CMU コンピュータサイエンス学科のフロントページを入り口として、学科内情報にアクセスした、約 7 ヶ月間に及ぶユーザのクリックストリームを対象とし、少なくともリンクを 4 つ以上たどらなければ必要な情報を得ることのできない検索課題を用意し、そのページを模範解答とした。結果は以下のように Combine が最も正解率が高かった。

	Accuracy
Random	31.3%
Popularity	41.9%
Match	40.5%
Annotate	42.2%
RL	44.6%
Combine	48.9%

［参考文献］

[1] Thorsten Joachims, Dayne Freitag, Tom Mitchell: "WebWatcher: A Tour Guide for the World Wide Web", In Proceedings of International Joint Conference on Artificial Intelligence (1997).

４.５.４ 評判情報検索システム

評判情報検索システムは、インターネット上に発信された様々な人の意見を自動検索・分類する専門分野 Web 検索エンジンである。NEC 関西研で開発された[1][2][3][4]。Shopbot（価格比較システム）では、表のような項目の繰り返し構造をもつ Web ページからの情報抽出が実現されていたが、評判情報検索システムでは非構造 Web ページからの情報抽出を行う点が特徴である。

ユーザは商品名 と商品カテゴリを入力し､その商品に関する人の意見を検索結果一覧の形で見ることができる。各意見は、これが出現した Web ページ単位にまとめられ、肯定意見／否定意見を区別するマークも付与される。なお、対象物は商品名に限定されず、ブランド名・サービス名・企業名・人名などであってもよい。

1. A 氏のホームページ
   ○ モバイル 777 のほうがよほど役に立つ。
   ○ ここ半年ぐらいモバイル 777 という携帯情報機器を愛用しています。

2. モバイルマシン入門のページ
   × 良くできたモバイルマシンだが、モバイル 777 の欠点としては、カラー液晶になって…
   ○ プレゼンテーションするだけなら、モバイル 777 とディスプレイケーブルを持っていけば充分です。

3. B 氏の日記
   ○ わが家の新メンバーとなったモバイル 777 には、本当に重宝しています。
   ○ モバイル 777 のキーボードはとても打ちやすいです。

図 4.5.4－１　評判情報検索結果（架空の商品名での検索結果イメージ）

評判情報検索の処理の流れは以下のようになる。

まず、Web クローラやメタサーチを用いて、対象とする商品名を含む Web ページをインターネットから収集する。次に、意見抽出部が近接演算処理を実行する。この処理では、商品名と評価表現とが一定の距離以内にあるという条件に該当する箇所を見つけ、両者を含む文字列を意見として抽出する。評価表現は、「良い」や「悪い」等の評価を示す表現のことで、商品カテゴリごとにあらかじめ辞

書として用意している。さらに、抽出した意見について、構文的な意見らしさのスコアを計算する適正値判定処理を実行する。最後に、肯定･否定分類部では、あらかじめ定めた評価表現の基本属性（肯定／否定）に対して、評価表現の近くに「ない」のような否定表現が出現する場合は属性を反転する。

意見抽出の精度に関しては、コンピュータ分野での評価では、意見らしさのスコア上位 10 件の適合率が 72%、再現率 43%のときの適合率が 62%であり、アルコール飲料分野では、上位 10 件の適合率が 84%、再現率 61%のときの適合率が 59%という結果が得られている[4]。

評判情報検索システムの具体的な応用として、インターネットを情報源としたマーケットリサーチが提案されており、テキストマイニングシステム SurveyAnalyzer と組み合わせた Web 上の評判分析システムも開発されている[5]。

[参考文献]

[1] 立石健二、石黒義英、福島俊一: “インターネットからの評判情報検索”、情報処理学会第 62 回全国大会、4W-5 (2001)。

[2] 立石健二、石黒義英、福島俊一: ”評判情報検索システムの試作と評価”、情報処理学会第 63 回全国大会、2V-1 (2001)。

[3] 立石健二、石黒義英、福島俊一: ”インターネットからの評判情報検索”、情報処理学会研究報告、NL-144-11 (2001)。

[4] 立石健二、福島俊一: ”意見分析システムにおける意見抽出方式の検討と評価”、第 1 回情報科学技術フォーラム、D-1 (2002)。

[5] 立石健二、森永聡、山西健司、福島俊一: ”Ｗｅｂ上の意見分析－情報抽出とテキストマイニングの融合－”、情報処理学会第 64 回全国大会、2X-4 (2002)。

４.５.５ ResearchIndex

ResearchIndex（別名：CiteSeer）インターネット上に公開されているコンピュータサイエンス分野の論文を自動収集・インデクシング（Full-text indexing および Citation indexing）する専門分野 Web 検索エンジンである。NEC 北米研で開発され、http://citeseer.nj.nec.com/cs で公開されている。世界最大の無償のコンピュータサイエンス関連論文ライブラリということもできる。50 万件を収集しており、月間 500 万アクセスという実績がある。

以下に、ResearchIndex の機能と技術的な特徴を列挙する。

(1) Locating Scientific Articles

複数のサーチエンジンの検索結果から研究論文を含むページを抽出する。例えば Postscript、PDF、technical report、conference、proceedings などのキーワードを利用。また、メーリングリストを監視し、そこからも投稿論文を収集する。さらに、論文著者が CiteSeer に直接登録することもできる。広域クローリング機能ももつが効率的でないので使っていない。

(2) Full-text Indexing

Postscript や PDF もテキスト形式に変換して全文インデックス作成する。論文著者のイニシャルも精度上重要なのでストップワードは使用しない。

(3) Autonomous Citation Indexing

論文の引用関係のインデックスを作成する。同一の論文に関する引用は 1 つにまとめて表示。何人に引用されたか、自己引用は除外等のカウントも可能。発行年別のヒストグラムも生成。

(4) Information Extraction

論文内の引用論文リストの箇所を抽出する。引用論文リストから個々の引用論文を抽出する。個々の引用論文から著者やタイトル等の情報を抽出する。これには隠れマルコフモデルを利用している。論文本体の著者やタイトルも抽出する。

(5) Context and Query-Sensitive Summaries

引用論文内で検索論文を実際に引用している箇所をサマリーとして表示する。

(6) Related Documents

検索論文に関する引用論文だけでなく、関連論文も表示する。関連論文は引用と単語情報を利用して計算している。同じ著者で同じ組織であるほど優先度を高くする。

(7) Overlapping Documents

検索結果からほぼ同一内容の論文を削除する。共著者同士で書く論文や同一著者のマイナーチェンジ等がこれに該当する。2 つの論文間で同一と認められる文章の割合が多いと同一論文とみなす。

(8) Citation Graph Analysis

自己引用かどうかは論文著者をもとに判断する。リンク情報を利用したランキング（Google 等と方式を若干改良）をしている。Authority は被リンク数の多いページ、Hub は多数の Authority にリンクしているページ、Authority と Hub の各々のランキングを表示する。Hub の論文は研究者が最初に読む論文として適している。新しい論文ほど引用される数が少ないので時間軸で正規化してランキングする。

(9) User Profile

お薦め論文を自動配信する機能もある。

(10) Distributed Error Correction

ユーザがデータベース内容を修正することも可能。

(11) External Links

著作権問題で無償で扱えない論文には外部リンクを張っている。

[参考文献]

[1] Steve Lawrence: "Online or Invisible?", Nature, Vol. 411, No. 6837 (2001).

[2] Steve Lawrence and C. Lee Giles and Kurt Bollacker: "Digital Libraries and Autonomous Citation Indexing", IEEE Computer, Vol. 32, No. 6 (1999).

[3] Steve Lawrence and Kurt Bollacker and C. Lee Giles: "Indexing and Retrieval of Scientific Literature", CIKM'99 - Eighth International Conference on Information and Knowledge Management (1999).

[4] Kurt Bollacker and Steve Lawrence and C. Lee Giles: "Discovering Relevant Scientific

Literature on the Web", IEEE Intelligent Systems, Vol. 15, No. 2 (2000).
[5] Steve Lawrence and C. Lee Giles: "Searching the Web: General and Scientific Information Access", IEEE Communications, Vol. 37, No. 1 (1999).
その他多数（下記 URL に文献リストがある）
http://www.neci.nec.com/~lawrence/pub-ri.html

4.5.6 DEADLINER

DEADLINER は、アカデミックな研究コミュニティ向けにコンファレンスやワークショップ等に関するアナウンスの一覧を作成する専門分野 Web 検索エンジンである。コンファレンスのアナウンスに関する WWW、ニュースグループ、メーリングリストなどを監視し、講演者、開催場所、開催日、論文提出等の期限、トピック（キーワード）、プログラム委員会、要旨、所属等のフィールドを抽出し、データベースに登録する。NEC 北米研で開発された。

DEADLINER は、コンファレンスのアナウンス(Call for Papers)に特化したものとして動いているが、そのアーキテクチャは、人手によるルールのチューニングや自然言語処理を用いるのではなく、単純なモジュールの学習と統合によって構築されており、専門分野 Web 検索エンジンの汎用的なアーキテクチャを提案している。その処理手順は、次の通りである。

(1) 文書の検索・収集： メタサーチ、ニュースグループウォッチ、フォーカストクローリングという 3 通りの方法を用いて、候補となる文書を効率的に収集する。

(2) 不適合文書のふるい落とし：適合／不適合文書セットに 7.5%以上出現する単語・bigram・trigram のうちで、不適合文書中の頻度に対する適合文書中の頻度の比が大きいもの上位 100～300 個を、文書ベクトルに利用し、SVM (Support Vector Machine)によるフィルタリングをかける。

(3) ターゲットフィールドの検出：多数の検出器（正規表現マッチあるいはフォーマットテンプレートマッチ）を統合し、ターゲットフィールドを判定する。訓練データでの検出精度に基づいて検出器（ルール）に優先度付けし、上位ルール何個まで使用するかは、システム管理者がアプリケーションに要求される適合率と再現率のバランスから判断する。

(4) フィールドから値の抽出：複数の検出器による検出結果をマージし、最小共通領域等のヒューリスティックスを用いて、抽出する値を決定する。

Call for Papers を対象として実装した際の精度については、以下のように報告されている。

まず、SVM による不適合文書のふるい落としの精度については、人手で選別した正解例 592 件、不正解例 2269 件（うち 850 件はコンファレンス関係）を準備し、ランダムに選択された訓練セット(正解例 249 件、不正解例 1250 件)とテストセット（正解例 343 件、不正解例 1019 件）で評価した。その結果、ポジティブ精度 95.9%、ネガティブ精度 98.6%が得られた。

ターゲットフィールドの抽出精度については、DBWorld 500 文書（208 件のトピック、338 件のコンファレンスタイトル、906 件の提出期限、197 件のプログラム委員会を含む）を用いて評価した。提出期限（300 件）の抽出精度については、検出成功&抽出成功が 214 件、検出成功&抽出失敗が 2 件、検出失敗が 31 である。プログラム委員会（1455 件）の抽出精度については、検出成功&抽出成

功が 1252、検出成功&抽出失敗が 72 件、検出失敗が 136 件である。

[参考文献]

[1] Andries Kruger, C. Lee Giles, Frans M. Coetzee, Eric Glover, Gary W. Flake, Steve Lawrence, Christian Omlin: “DEADLINER: Building a New Niche Search Engine”, CIKM 2000 - Ninth International Conference on Information and Knowledge Management (2000).

４.６ 専門分野 Web 検索の要素技術

４.６.１ Wrapper Induction

（１）LR Wrapper

専門分野 Web 検索サイトの中には外部の Web ページから注目する情報のみを取り出して、それら情報を自分のデータベースに登録するタイプが存在する。そして Wrapper Induction は、上記のように Web ページから注目する情報を自動的に取り出すための技術の一つである。

ここでは Wrapper Induction の中心的な論文である Kushmerick の論文[1]を紹介する。

インターネットの出現以来、インターネット上の情報を一種のデータベースとして扱おうという研究がある。もちろん、インターネット上の情報は通常のデータベースのように規格化された形で蓄えられてはいないため、所望の情報を取り出すことが困難な場合も多い。そこで、ユーザが目的の情報を容易に取り出せるようにするための１つの手段として、Web 全体に皮（Wrapper）をかぶせることが考えられる。するとこの場合 Wrapper とは、Web 上の文書からある種の情報を抽出するプログラムに対応する。

当初 Wrapper は人手で作成されていた。ただし、その作成コストが高いことや、抽出目的の対象が変化すれば､また一から作成しなおさなくてはならない｡そのため自動構築の技術が望まれており、Wrapper Induction はそのような背景から生まれた技術である。

Wrapper Induction は、概略述べれば、文書とその文書内の抽出対象の対の集合を与えて、文書から目的の情報を抽出する規則、つまり Wrapper を帰納学習する技術である。Kushmerick が対象としたのは主に表タイプの Web ページである。そこから表の項目を抽出する規則を自動構築することを目指した。

例えば、以下は国名とその国の電話の国番号が書かれている Web ページのソースの例である。

```
<HTML><TITLE> 国名と国番号 </TITLE><BODY>
<B>コンゴ</B><I>242</I><BR>
<B>エジプト</B><I>20</I><BR>
<B>ベリーズ </B><I>501</I><BR>
<B>スペイン</B><I>34</I><BR>
</BODY></HTML>
```

このページから以下のような国名と国番号のペアを抽出することを考える。

{<コンゴ,242>,<エジプト,20>,<ベリーズ,501>,<スペイン,34>}

抽出するプログラム（Wrapper）としては図 4. 6. 1-1 のようなものがあればよい。このプログラムのポイントは l_k と r_k の組の部分、すなわち {(<B>,</B>),(<I>,</I>)} である。直感的には、抽出項目（国名や国番号）を挟む左右のタグが l_k と r_k である（ただし Wrapper Induction は HTML 文書だけを対象にしているわけではなく、 l_k や r_k は一般に文字列であり、タグになるとは限らない)。上記のプログラムを自動構築するとは、この組の集合を学習することに他ならない。このような Wrapper を LR Wrapper、そして LR Wrapper により目的とする抽出項目を抽出できる文書の集合を LR Wrapper class と呼んでいる。

```
function   LR_wrapper (page P)  {
   while   there are more occurences in P of <B>
   {
      for each  (l_k, r_k) in {(<B>,</B>),(<I>,</I>)}
       {
          scan in P to next occurence of l_k
          save position as start of kth attribute
          scan in P to next occurence of r_k
          save position as end of kth attribute
       }
   }
   return  extracted  {..., <country,code>,...} pairs
}
```

図 4. 6. 1－1

LR Wrapper class の文書 P と P から抽出すべき項目 L の対 <P,L> の集合を入力として <l_k,r_k> の組を学習するアルゴリズムが Wrapper Induction である。この学習アルゴリズムは l_k と r_k に関するある制約充足問題に変形でき、この問題を解くことで目的の l_k と r_k、つまり LR Wrapper が得られる。概略述べると、解法のポイントは l_i と r_j を独立に扱っている点である。このために効率的な解法となっている。 l_i や r_j の候補の文字列を作成し、ある制約から候補を絞る。そして最終的に残った候補の中で最長の文字列を l_i や r_j に割り当てる。

ここで、現実に対象となる Web ページは都合よく LR Wrapper class に属しているのかどうかという疑問が当然起こる。後述する実験では、実際のサイトの 53% がこのクラスに属していると予想している。このカバー率を上げるために、LR Wrapper class 以外に、HLRT Wrapper class、OCLR Wrapper class、HOCLRT Wrapper class、N-LR Wrapper class 及び N-HLRT Wrapper class の 5 つのクラスとそのクラスに対する Wrapper の学習手法を提案している。ここでは学習手法の説明は省略し、それぞれがどのようなクラスかだけを示す。

1） HLRT Wrapper class

最初に例で出した国名と国番号の HTML のソースが以下のようになっていたとする。

```
<HTML><TITLE>国名と国番号</TITLE><BODY>
<B>国名と国番号</B><P>
<B>コンゴ</B><I>242</I><BR>
<B>エジプト</B><I>20</I><BR>
<B>ベリーズ</B><I>501</I><BR>
<B>スペイン</B><I>34</I><BR>
<HR><B>終わり</B></BODY></HTML>
```

この場合、LR Wrapper ではうまく抽出できない。それは国名を取り出す左右の文字列(<B>,</B>) が、繰り返しの始まり（Head）にある「国名と国番号」や繰り返しの終わり(Tail)にある「終わり」を国名として取り出してしまうからである。この不具合に対処するために、LR の規則の他に Head と Tail の部分を認識するように拡張したのが HLRT Wrapper であり、このような文書のクラスが HLRT Wrapper class である。上記の例の場合、学習される規則は、Head の終わりの文字列 <P> と Tail の始まりの文字列 <HR> が LR に追加された形になり、{<P>,<B>,</B>,<I>,</I>,<HR>} となる。

2) OCLR Wrapper class

最初に例に出した国名と国番号の HTML のソースが以下のようになっていたとする。

```
<HTML><TITLE> 国名と国番号 </TITLE><BODY>
<B>国名と国番号</B><P><UL>
<LI><B>コンゴ</B><I>242</I><BR>
<LI><B>エジプト</B><I>20</I><BR>
<LI><B>ベリーズ</B><I>501</I><BR>
<LI><B>スペイン</B><I>34</I><BR>
</UL><HR><B>終わり</B></BODY></HTML>
```

この場合も、HLRT のときと同様、LR Wrapper ではうまくいかない。ここでは、LR の規則を含む始まり(Open)の文字列と終わり(Close)の文字列を取り出すことで対処する。具体的には、<LI> と
 で LR の部分が挟まれている。上記例の場合、学習される規則は、{<LI>,
,<B>,</B>,<I>,</I>} となる。

3) HOCLRT Wrapper class

最初に例に出した国名と国番号の HTML のソースが以下のようになっていたとする。

```
<HTML><TITLE> 国名と国番号</TITLE><BODY>
<B>国名と国番号</B><P><UL>
<LI><B>コンゴ</B><I>242</I><BR>
<LI><B>エジプト</B><I>20</I><BR>
<LI><B>ベリーズ</B><I>501</I><BR>
<LI><B>スペイン</B><I>34</I><BR>
</UL><HR><B>終わり</B></BODY></HTML>
```

この例は HLRT と OCLR の両方のクラスに属している。このようなタイプの文書に対しては、LR の他に Head、Open、Close、End を学習する。上記例の場合、学習される規則は、{<P>,<LI>,
,<B>,</B>,<I>,</I>,<HR>} となる。

4) N-LR Wrapper class

LR、HLRT、OCLR、HOCLRT は対象として HTML 文書を想定している。 一方、N-LR は通常のテキストを想定している。例えば、以下のようなテキストを考える。

```
name: 鈴木
        address:  東京都
            phone:  123-4567
            phone:  444-5555
name: 田中
        address:  千葉県
            phone:  666-7777
name: 佐藤
name: 山田
        address:  埼玉県
        address:  栃木県
            phone:  888-9999
            phone:  222-1111
```

これは通常のテキストであるが、改行やインデントが視覚的に表を構成しており、表が埋め込まれていると考えられる。この表の項目の値を抽出するのも LR で実現できる。例えば住所に対応する値は、'address:' と '¥n' (改行) で囲まれている。上記のような文書のタイプを N-LR Wrapper class と呼んでいる。LR Wrapper class はほぼ N-LR Wrapper class に含まれている。

5) N-HLRT Wrapper class

LR のクラスに対して、HLRT のクラスがあったように、N-LR のクラスに対しても N-HLRT クラスが存在する。これは N-LR のクラスの例に、Head と Tail が付いたものである。

以上、この論文では 6 つの文書のクラスとそのクラスに対する Wrapper の学習手法を提案しているが、さらにこれらクラスが現実にカバーできる文書の割合と各クラスの学習の効率についても論じ

ている。

カバー率に対しては、以下のような実験を行っている。www.search.com からランダムに 30 サイトの検索エンジンを選ぶ。これらはある分野専用の検索エンジンとなっている。各検索エンジンに対して、適当なクエリを与え、検索されたページを集める。それらのページのうち 10 ページに抽出項目のラベルをつけ各クラスの Wrapper の学習を行い、100%の正解率で目的の情報を抽出できる Wrapper が学習可能かどうかを調べる。もし学習可能であれば、最初の検索エンジンが対象とする分野に関しては Wrapper の学習が可能だとする。この学習可能な分野の割合を調べる。結果、LR は 53%、HLRT は 57%、OCLR は 53%、HOCLRT は 57%、N-LR は 13%、N-HLRT は 50%のカバー率になった。6 つのクラスが学習可能な分野の和集合をとると 70%となった。N-LR や N-HLRT の成績はよくないが、他の 4 つではカバーできない文書の 25% がこれらでカバーされている。また理論的に 6 つの文書クラスの包含関係も与えているが、単純な包含関係になるものはほとんどなく、各文書クラスの関係は複雑である。

次に学習の効率に対しては、各クラスの学習に必要とされる事例の数や学習時間について、実際の実験と理論的解析の二面から述べている。ここでは理論的解析の結果だけを紹介する。そこでは、PAC 学習理論を用いて必要とされる事例数について以下の結果を得ている。

・LR と N-LR については $1/\varepsilon \cdot (2K \ln R - \ln \delta)$

・HLRT,OCLR,N-HLRT については $1/\varepsilon \cdot ((2K+2)\ln R + \ln(R^2 - 2R + 1) - \ln(4\delta))$

・HOCLRT については $1/\varepsilon \cdot ((2K+4)\ln R + \ln(R^4 - 4R^3 + 6R^2 - 4R + 1) - \ln(16\delta))$

ここで K は抽出すべき l_i と r_i の組の数（最初の国名と国番号の例では 2）、R は訓練データのページの大きさの中で最小のものの値を示す。そしてエラー率 ε を確率 $1-\delta$ で達成するのに必要な最低事例数が示した値である。また学習時間に関しては以下の結果を得ている。

・LR については $O(KM^2 |\varepsilon|^2 V^2)$

・HLRT については $O(KM^2 |\varepsilon|^4 V^6)$

・OCLR については $O(KM^4 |\varepsilon|^2 V^6)$

・HOCLRT については $O(KM^4 |\varepsilon|^4 V^{10})$

・N-LR については $O(M^{2K} |\varepsilon|^{2K+1} V^{2K+2})$

・N-HLRT については $O(M^{2K+2} |\varepsilon|^{2K+3} V^{2K+4})$

ここで V は訓練データのページの大きさの最大のものの値を示す。M は訓練データのラベルの大きさの総数を示す。

Wrapper の帰納学習で重要な点は、できるだけカバー率をあげられるクラスとその上での学習手法を考案することであろう。ここでの学習では、あるページで項目が欠損するような場合や、他のタグで表現される場合などに対応できない問題が指摘されている。また訓練データを作成するためにラベルを付けるコストが高いことも問題としてある。また Wrapper は情報抽出の抽出パターンの一種とも捉えられる。そのため、Wrapper Induction は情報抽出における抽出パターンの自動獲得に関する研究とも関連がある。

（2）Wrapper Induction の比較

複数の Web サイトから情報を取得してデータベースに格納する処理には、当初はサイトごとに人手で作成された Wrapper が用いられていた。このような最初の Wrapper 構築フレームワークが TSIMMIS である。これは、複数の情報源にアクセスして情報を取得し、それらに矛盾がないかを検証するためのツールを提供するものである。

一方、情報量が多く、構造が動的に変化する Web サイトにおいては、より効率的な Wrapper の構築が求められている。Database 分野では異種情報の統合方法が中心議題であり、AI 分野では機械学習を Web サイトの学習に利用する方法が中心となった。ここでは論文[2]を元に、後者について述べる。

①では、構造化された Web ページの Wrapper とその生成について述べる。これらのシステムは、おもに Wrapper 生成の分野において提唱されたものである。②では、構造化されていない Web ページに関する技術について述べる。こちらのシステムは、主に IE 分野から提案されたものである。

① Structured and semistructured Web pages

この節では、Wrapper 生成システムとして、ShopBot, WIEN, SoftMealy, STALKER について説明する。これらは、サーバーに送ったクエリへの応答として生成された Web ページからの情報抽出が主な応用である。主に、区切り文字を利用した抽出パタン(delimiter-based extraction patterns)を利用し、文法解析は行っていない。また、構造化されたデータのみに対して動作することができる。

(a) ShopBot

ShopBot は、比較ショッピングエージェントであり、フォームに入力したクエリに基づいて生成された商品一覧画面からの情報抽出を目的としている。抽出処理は、ヒューリスティック検索とパタンマッチ、そして、帰納学習の組み合わせである。

Shopbot の動作は、オフラインの学習フェーズと、オンラインの比較ショッピングフェーズからなる。学習フェーズではヒューリスティックを用いて、クエリ入力フォームと、クエリの与え方を推定し、次に検索結果として得られたページのフォーマットを推定する。また、検索結果の Web ページは header、body、tail から構成され、header 部分と tail 部分は異なる Web ページの間で共通であると仮定する。検索結果 Web ページは、以下のステップにより推定される。

- 存在していない製品キーワードを与えて、検索できなかった場合のテンプレートを学習する。
- 次に、いくつか既存製品のキーワードを与えて、head と tail を推定する。
- 最後に、body 部分のフォーマットを解析する。タグとテキストの列として定義されたフォーマット仮説をいくつか用意する。ボディ部分を論理行に区切り(
等で区切る)、各仮説と比較して、最も一致するものを選択する。

以上の手順でフォーマットを解析する。ただし、「価格」といった情報スロットのラベルは導けないので、「価格」だけは特別にハンドコードされたアルゴリズムで抽出する。

(b) WIEN

WIEN(Wrapper Induction ENvironment)は、wrapper 構築支援ツールである。WIEN は ShopBot

に強く影響されたものであるが、term wrapper induction を初めて導入したものであり、このアルゴリズムは表形式のさまざまな構造化データに対して有効である。

このアルゴリズムでは、Web ページの構成として、Head 区切り文字(delimiter)と、抽出情報両側にある Right 区切り文字および Left 区切り文字、そして Tail 区切り文字からなるものを対象としている(この構成を論文では HLRT と呼んでいる)。まず、スロットの前後を決める区切り文字を探し、次に表部分と周りのテキスト部分とを分ける区切り文字を探す。

WIEN は抽出情報の前と後ろを区切る区切り文字のみを用いる。またいくつかの項目が抜けていたり、項目の出現順序が変則的な文書には適用できない。

一方、マルチスロットルールを用いることができるので、複数の項目を用いて特定しなければならない情報も抽出できる(例： 住所録の表において、特定の個人の住所を抽出する際)。

(c) SoftMealy

WIEN の登場後に、Wrapper Induction を改良したいくつかのシステムが提案された。SoftMealy は、半構造化された Web ページから情報を抽出する Wrapper を、非決定性有限オートマトンとして学習する。

帰納生成アルゴリズムは、訓練例から、文法ルールを生成するのに用いられる。訓練例は、抽出情報とそれらを区切るセパレータからなるリストである。Wrapper induction は、セパレータの場所と抽出情報の出現順序を得るための手がかりとなるラベルが付けられた行を入力として受け取る。このラベルによる場所情報を用いて文脈ルールを生成する。

生成された Wrapper は非決定性有限オートマトンであり、各状態は抽出する情報を表し、状態遷移は抽出情報の間にあるセパレータを定義した文法ルールを表す。情報抽出処理時には、Wrapper は抽出情報の周りにあるセパレータを利用する。

SoftMealy は、ワイルドカードが利用可能であり、抜けている項目(Missing items)にも対応可能である。また、抽出項目の出現順序に依存しない抽出が可能ではあるが、可能な出現順序全てをカバーする訓練例が必要である。SoftMealy の抽出パタンは、WIEN よりも表現力が高い。

(d) STALKER

STALKER は情報抽出ルールの導出のための学習アルゴリズムである。ユーザーは幾つかのサンプルページとそれに関連したデータの組を訓練例としてシステムに与える。システムはページからトークン列とそのインデックス列を作成する。これらのトークン、およびワイルドカードの列は、抽出情報の位置を特定するためのランドマークとして使用される。

STALKER は順列変換アルゴリズムを使用する。はじめに可能な限り多くの正の訓練例を生成することができるリニア・ランドマーク・オートマトンを生成する。次に残りの訓練例を生成できるオートマトンを生成する。すべての正の訓練例がカバーされたら、獲得したオートマトンに相当する単純なランドマーク文法を出力する。

STALKER は文書中の情報をそれぞれ独立に抽出するので、文書中における情報の出現順序に依存しない抽出が可能である。このことにより WIEN よりも柔軟性の高い抽出が可能である。また

SoftMealy とは対照的に、STALKER は抽出項目のすべての出現順を含んだ訓練例を用いなくてよい。

② Semistructured and unstructured Web pages

この節では、より広範囲のテキストに適用可能なシステムとして、RAPIER, SRV, WHISK について述べる。これらのシステムでは、意味論や文法を利用していないが、今後利用することも可能である。

これらのシステムは、IE と WG(Wrapper Generation)の中間にある。手法は帰納論理プログラミングや関係学習に基づいており、SRV と WHISK は帰納アルゴリズム FOIL に、RAPIER は帰納アルゴリズム GOLEM と関係がある。

(a) RAPIER

RAPIER(Robust Automated Production of Information Extraction Rules)は、文書と情報の抽出部分に記述が埋め込まれたテンプレートを受け取り、抽出ルールを学習する。

この学習アルゴリズムは帰納学習であり、はじめにターゲットのスロットと一致する最も特殊(汎用の逆の意味)なルールが適用される。次に 2 つのルールをランダムに選択して、2 つのルールに対して汎化されたルールを生成する。そしてルールに変化がなくなるまで制約を追加する。

抽出パタンは、区切り文字とコンテンツ記述からなる。抽出ルールは、テンプレート名とスロット名でインデクシングされ、次の 3 つの部分から構成される。pre-filler(ターゲットテキストの直前と一致するパタン)、 filler(ターゲットテキストと一致するパタン)、 post-filler(ターゲットテキストの直後に一致するパタン)

各パタンは pattern items の列であり、一つの単語に一致するものか、もしくは複数の語に一致する pattern lists である。パタンとテキストの比較においては、(i)単語、(ii)品詞、(iii)意味クラス、の 3 点が同じ場合に一致とみなす。

このシステムは、シングルスロットとして動作するが、テキストを 3 対上の部分に分割することで、マルチスロットとしての動作の可能性もある。

(b) SRV

SRV(Sequence Rules with Validation)は、トップダウンの関係学習アルゴリズムである。入力は抽出するフィールドに対してラベル付けされた文書集合と、トークンの特徴情報である。出力は、抽出ルールである。

SRV では IE を分類問題として扱う。文書中の全ての抽出フレーズ候補に対して、ターゲット・スロットを満たす文字列の候補としての確信度を表す基準を割り当てる。オリジナルの SRV では、FOIL に類似したトップダウン処理の関係ルール学習を行う分類機を用いている。

SRV に与えるトークンの特徴情報には、文字列の長さ、文字列タイプ、つづり、品詞、語彙意味などを使うことができる。

処理は、まず全ての正例、負例のセットから開始し、以前に学習したルールがカバーしていない範囲の例と比較される。そして、ルールが正例のみをカバーする状態や、これ以上の汎化が無意味であるとなったときに、ルールと一致する全ての正例が取り除かれる。

SRV のルールの表現力は高い。SRV は特定のアイテムを他の関連したアイテムとは独立に抽出できる点がSTALKERやRAPIERに似ている。関係学習ではシングルスロットの抽出のみ可能である。一方、WIEN ではマルチスロット抽出のルールを学習できる。

(c) WHISK

WHISK は、構造化されたテキストや半構造化されたテキスト、そして、ニュース記事などのフリーテキストからの抽出など、全てのタイプの抽出に対応するシステムであり、文法情報と意味情報を用いる。構造化・半構造化テキストからの抽出では、文法情報は必要ない。一方、フリーテキストからの抽出では、入力テキストは文法解析と意味タグが付与されている状態で最も効率的に動作する。

WHISK の処理は、タグ付けされていないインスタンスと、空のタグ付けされたインスタンスがある状態から開始する。インスタンスとは、文書中の小さな単位のことである。半構造化された文書では、HTML などのタグを用いて文書がインスタンスに分割される。フリーテキストでは、文解析により文書を文に分割する。ユーザーにいくつかのタグ付けされてないインスタンスを提示し、ユーザーが抽出部分を囲むようにタグを付与する。文書に埋め込まれたタグは、ルールの生成と評価に用いられる。

WHISK は、トップダウンの分類ルール学習に属している。最初に最も一般的なルールを生成し、エラーがゼロになるか、または事前枝狩り(pre-pruning)基準を満たすまでルールにタームが追加される。追加されるタームの選択では、ラプラシアン期待エラー(laplacian expected error) (e+1)/(n+1)が使われる。e はエラーの発生数であり、n は訓練例から得られた抽出情報の数である。学習は全ての正の抽出が行えるルールが生成されるまで行われる。また、事後枝狩り(post-pruning)が、オーバーフィッティングを防ぐために行われる。

③ Summary

本節では、Wrapper の自動学習システムについて述べた。表 4.6.1-1 は、各システムの特徴をまとめたものである。X 印が、そのシステムが処理できる項目を表している。

上の 4 つのシステムは Wrapper Generation をベースとしたシステムであり、構造化された Web ページからデリミタを用いた抽出規則により抽出処理を行うものである。一方、下の 3 つのシステムは、IE をベースとしたものであり、関係学習を用いている。SRV と WHISK はトップダウン探索を行う。RAPIER は基本的にはボトムアップ探索である。

表 4.6.1－1

	struct	Semi	free	Multislot	Missing items	Permutations
ShopBot	X					
WIEN	X			X		
SoftMeary	X	X			X	X
STALKER	X	X			X	X
RAPIER	X	X			X	X
SRV	X	X			X	X
WHISK	X	X	X	X	X	X

struct：構造化された文書からの情報抽出

semi：半構造化された文書からの情報抽出

free：フリーテキストからの情報抽出

Multislot：マルチスロットの抽出ルールの生成

Missing items：文書中に抽出情報の一部が含まれていない場合の抽出処理

Permutations：文書中に出現する抽出情報の順番が不定の場合の抽出処理

[参考文献]

[1] Kushmerick, N.: "Wrapper induction: Efficiency and expressiveness", Artificial Intelligence, Vol. 118, pp.15-68 (2000).

[2] Line Eikvil: "Information Extraction from World Wide Web - A Survey", Norweigan Computing Center, No.945 (1999).

４.６.２ Specialized Query Modification

汎用的な検索エンジンの限界における問題を解決するために、ユーザが入力するクエリの処理に着目して、クエリの処理時に分野特有の知識を用いるアプローチがある。本項では、ユーザの必要とする分野の情報を含む文書に典型的に出現するような表現をクエリに追加することにより、特定の分野の情報を含む文書を高い精度で検索する論文を紹介する。

Glover らは、Support Vector Machine (SVM) による Web ページのカテゴリ分類とカテゴリ分類に基づくクエリ拡張により、分野特化型の検索エンジンを実現している[1]。なお、特定分野の文書をより多く検索するために、複数の汎用の検索エンジンによるメタ検索が用いられるが、各検索エンジンの特徴に依存して、適切なクエリ拡張が異なるので、本論文では最適なクエリ拡張と検索エンジンの対を求めることが目的である。以下では、本論文で述べられている（１）分類器の学習、（２）クエリ拡張の選択、（３）クエリ拡張と検索エンジンのランク付けについて技術の詳細を紹介する。

（1） 分類器の学習

ある特定のカテゴリの文書を取得するには、はじめに Web ページがその特定のカテゴリに属するかどうかを判定する必要がある。あるカテゴリに属する Web ページ A（正事例）と属さない Web ページ B（負事例）を訓練データとして SVM に与えることによって、分類器を学習する。分類器の学習において、文書は 2 値の素性ベクトルに変換される。素性には、単語、フレーズ、HTML の構造、テキストの位置情報が用いられる。例えば、”my”がタイトルにあるか等が素性になる。

さらに素性ベクトルの次元数について、SVM が学習するのに適切な数に圧縮する。まずある一定の出現頻度に満たない素性について消去し、その後期待エントロピー損失（期待情報利得）で素性をランキングすることにより、出現頻度は高いが分類に有効でない素性を取り除いている。

（2） クエリ拡張の選択

クエリ拡張は、カテゴリ分類の際に用いられた素性のひとつ以上の組み合わせとして表現される。ただし、クエリ拡張は検索エンジンの入力になるので、素性はテキスト中に出現する単語かフレーズに制限されている。クエリ拡張の目的は適合率を上げることであるが、適合率を上げようとしてクエリを拡張すると再現率が下がるので、パラメータにより下限の適合率を指定する方法を採っている。

あるクエリ拡張 qm のすべての素性を含む正事例集合 A_{qm} と qm のすべての素性を含む負事例集合 B_{qm} から次式により計算される予測適合率 P’(qm)を求める。ただし、α は訓練データにおける適合率と実際の検索エンジンの結果から得られる適合率の違いを考慮した定数である。

$$P'(qm) = \frac{\left|A_{qm}\right|}{\left|A_{qm}\right| + \alpha\left|B_{qm}\right|}$$

この値が指定した下限の適合率のパラメータよりも小さい場合はスコアを 0 にする。そうでない場合は、次式によりスコア（予測再現率）を求める。

$$Score(qm) = \frac{\left|A_{qm}\right|}{\left|A\right|}$$

これをすべてのクエリ拡張に対して繰り返し、各クエリ拡張は計算されたスコアによりソートされる。

（3） クエリ拡張と検索エンジンのランク付け

検索エンジンによって望ましいクエリ拡張は異なるので、スコアリングされているクエリ拡張を使って、実際に検索エンジンを使って評価することが必要になる。評価するクエリとソートされている上位のクエリ拡張を用いて、検索エンジンから取得された上位 N 個の URL から Web ページをダウンロードし、そのページが望ましいカテゴリに属する場合にスコアをインクリメントしていく。これを評価するすべてのクエリについて調査し、検索エンジンに対するクエリ拡張のスコアを返す。すべての検索エンジンとクエリ拡張の組み合わせにおいて以上の処理を行い、最

終的にスコアでソートされたクエリ拡張と検索エンジンの対が得られる。

本技術を評価するために、個人のホームページと論文募集の二つのカテゴリについてクエリ拡張なしと単純なクエリ拡張をベースラインとする実験が行われた。個人のホームページのカテゴリでは、327 ページの正事例と 1,500 ページの負事例を学習し、391 ページの正事例と 400 ページの負事例が評価された。実験に用いられたクエリは、"chess", "ballroom dancing", "beagles", "cat"の 4 つである。検索エンジンは AltaVista, FAST, Google の 3 つが使用された。最も予測再現率の高いクエリ拡張は `my + welcome' の 25%で、これを含め予測再現率の上位 8 個のクエリ拡張を用いて、各検索エンジンにより各クエリ 50 ページの 200 ページにおける適合率が評価された。その結果、各検索エンジンでは最適なクエリ拡張が異なり、それぞれ最も適合率が高かったのは、`my + "home page"'と FAST の組で 64.5%、`+ "my name is"'と AltaVista の組で 57.5%、's home'と Google の組で 54.7%であった。クエリ拡張なしでは最も良い AltaVista で 8%、単純なクエリ拡張の`home page'では最も良い Google で 28.5%であった。このように、クエリ拡張はベースラインと比較して高い適合率で個人のホームページの検索が実現されている。

一方、論文募集のカテゴリでは、249 ページの正事例と 1,250 ページの負事例が学習され、183 ページの正事例と 1,019 ページの負事例が評価された。実験に用いられたクエリは、"databases", "natural language processing", "algorithms"の 3 つである。最も予測再現率の高いクエリ拡張は'"for papers" & "will be"'等で約 50%であり、予測再現率の上位 5 個のクエリ拡張を用いて、個人のホームページのカテゴリと同じ 3 つの検索エンジンにより各クエリ 50 ページの 150 ページにおける適合率が評価された。その結果、`call +for papers" "+will +be"'と Google の組で 88%、'"call for papers" "will be"'と FAST の組で 85%となり、検索エンジンの違いにより適合率にそれほど差がなかった。また、クエリ拡張なしでは最も良い Google で 2%だが、単純なクエリ拡張の`call for papers'では最も良い Google で 83%となり、本技術によるクエリ拡張とそれほど差が出なかった。

以上のように、本技術は、特定分野の Web ページだけを検索する際に、個人のホームページのようにカテゴリの幅が広く、どういう単語やフレーズをクエリに補って検索すればよいかが不明なカテゴリに対して、特に有効に機能すると言える。

なお、本技術はメタ検索エンジン Inquirus 2[2]に採用されている。

[参考文献]

[1] Eric J. Glover, Gary W. Flake, Steve Lawrence, William P. Birmingham, Andries Kruger, C. Lee Giles and David M. Pennock: "Improving Category Specific Web Search by Learning Query Modifications", Symposium on Applications and the Internet, SAINT (2001).

[2] Eric J. Glover, Steve Lawrence, William P. Birmingham and C. Lee Giles: "Architecture of a Metasearch Engine that Supports User Information Needs", In Proceedings of the Eighth International Conference on Information and Knowledge Management (CIKM'99), (1999).

4.6.3 Focused Crawling

Web 上の情報量は増加の一途をたどっており、従来の汎用サーチエンジンが利用しているデータ収

集手法（以降、Crawler と呼ぶ）のように、Web 上の情報をできる限り網羅的に集めて索引づけするという方式には、データ容量および収集速度の両面において限界が見え始めている。Crawler の収集方式を改良し、より重要なページを優先的に収集する手法の研究もされている[1] [2]が、これらは汎用サーチエンジンでの利用を前提とした網羅的収集を目的としており、問題の根本的解決には至っていない。

一方で、特定分野に限定した情報の収集を目的とする Focused Crawling 技術が近年注目を集めている。この技術は、本年度の主な調査対象である専門分野 Web 検索サイト構築に必須の要素技術となりうるばかりでなく、ユーザ端末での Web 情報巡回ソフトウェアなどにも応用可能であると考えられる。

汎用 Crawler が、ページに含まれるすべてのリンクをたどって収集を繰り返すのに対し、Focused Crawler は、あらかじめ与えられたトピックに基づき、関連するリンク先 URL を優先的に収集することが特徴である。リンクの探索優先度を決定するにあたり、トピックに対する関連度を確率モデル等に基づいて計算する手法が主流となっている（以降、トピックに対する関連度計算を行うエンジンのことを Classifier と呼ぶ)。また、優先度決定の方法には、大きく分けて Context Independent なモデルと Context Dependent なモデルが知られている。本節では、まず Context Independent なモデルを用いた例として、トピックカテゴリツリーに基づく言語生成モデルを利用した Chakrabarti らの手法[3]について、次に Context Dependent なモデルを用いた例として、Context Graph を利用した Diligenti らの手法[4]について紹介する。

(1) トピックカテゴリツリーを用いた Focused Crawling

(a) アルゴリズム

初期設定

Chakrabarti の手法では、Classifier を設定するために、基準となるカテゴリツリーを最初に用意する。図 4.6.3-1 にカテゴリツリーの具体例を示す。図 4.6.3-1 の左側にフォルダ階層上に示されたものが、トピックの階層を表したものである。

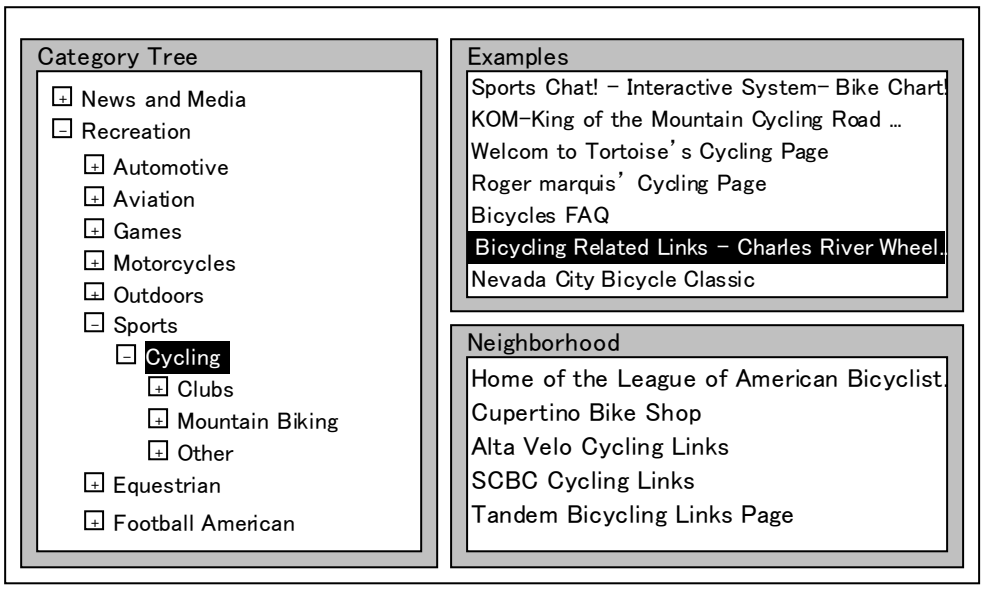


図 4.6.3－1：基準となるトピック分類例

カテゴリツリーとは、様々な粒度のトピックに対応するカテゴリを、Yahoo!のディレクトリ構造のように木構造上に表現したものである。各カテゴリにはそのカテゴリに分類される URL が割り付けられており、図 4.6.3-1 右上の"Examples"の部分に示されている。Examples の各 URL をクリックすると、右下の"Neighborhood"の部分に、関連すると思われる URL が提示される。関連するかどうかは、例えば内容を表す単語を共通に含むかどうかで判断する。

次に、探索対象とするトピックを決める。これは、クエリの形で与えるのではなく、関連すると思われる URL を 20～30 ほど用意する。以降、探索対象トピックを決めるために最初に与える URL の集合を、シーズ URL と呼ぶことにする。

カテゴリの選択と精緻化

システムは、与えられたシーズ URL を関連度の尺度に従い、既出のカテゴリツリー上に配置する。ユーザは、割り付けられたカテゴリの良し悪しを判定し、よいと判断したカテゴリを good としてマークする。マークされたカテゴリは図 4.6.3-1 のカテゴリツリー上でハイライト表示される。

カテゴリツリーは必要に応じて詳細化することが可能であり、また、割り付けられている具体例を別のカテゴリに移動することも可能である。

以上の操作を繰り返し、カテゴリツリーを精緻化することで、Focused Crawling のための初期設定が終了する。

Focused Crawling プロセス

処理プロセスは、各 URL の関連度計算のため毎回実行される Classify プロセスと、有効なリンクをもつページを評価するために断続的に実行される Distillation プロセスからなる。

Classifier は、先に設定したカテゴリツリーに基づく言語生成確率に従い、各 URL の関連度を計算する。即ち、カテゴリ C^*に関する関連度を Rc^*で表すと、任意の URL dについて、関連度 R は、$R_{root}(d) = 1$, および、$R_{parent}(d) = \Sigma_{ci} R_{ci}(d)$ {ci：c0 の子カテゴリ} という条件を満たす。Focused Crawling

における関連度は、$R_{topic}(d)$で表される。ここで topic とは、先の初期設定においてユーザが good と判定したカテゴリのことである。関連度の尺度に従いリンクを辿る方法として、以下の 2 通りの方法がある。

1) Hard focus rule：探索候補の URL に対して、最も関連度の高い（生成確率の最も高い）カテゴリツリー上のリーフノードを選択し、ルートノードからのパス上に good とマークされたカテゴリがあれば、探索候補として残す。

2) Soft focus rule：探索候補の URL に対して、good ノードに関する関連度 $R_{topic}(d)$を計算し、関連度の高いものから優先的に探索する。次のリンクを辿る優先度は、現在のページの関連度とし、当該 URL へのパスが複数ある場合は、最大の関連度を優先度とする。

Distiller は、集められてきた URL 群の中から、各 URL と URL 間のリンク構造に基づき、探索対象となる URL へのリンクを多く持つ Hub ページを見つけ出してくる[5]。その他優先度を決める要因として、特定の URL から同じ URL に対して張られているリンクの回数や、URL を訪れる回数などを用いることができる。

(b) 紹介手法の特徴

Chakrabarti の手法のメリットをまとめると以下のようになる。

1) 負例に対するよりよい確率モデル
   - 話題集合に含まれるか含まれないかという設定では、負例の確率分布がうまく推定できない
2) 負例を多く集めるより、正例を分類階層上に集めて必要に応じて編集する方が利便性が高い
3) 親ノードや兄弟ノードの関連クラスの事例をもとに、必要に応じて探索候補を拡張できる

(c) 手法の評価

評価の方法として、①取得した URL の平均関連度による評価、②シーズ URL の別々の部分集合から出発して得た URL の重なり割合、③実際に結果を人手でみた評価、の 3 つについて行っている。比較した手法は、Hard Focus rule, Soft Focus rule, Unfocus rule である。Unfocus rule とは、ランダムな順序でリンクを辿る方法である。対象分野として、ガーデニング、投資信託、サイクリング、エイズ等の各トピックについて crawl を行った。その結果、①では Hard Focus, Soft Focus 共に、Unfocus に対して顕著な違いが見られたが、Hard Focus と Soft Focus の差はそれ程見られなかった。②についてもサイクリングについては 8 割以上の一致率、投資信託では 6 割以上の一致率を達成している。③については、上位 100 URL を人手でチェックし、関連するページが得られたことを確認している。

(d) 手法の限界

紹介手法では、ターゲットの URL の探索途中に関連度の低いページが必ず含まれると、効率よくターゲットに到達できないという限界がある。

(2) Context Graph を用いた Focused Crawling

Diligenti らは、探索途中に関連度の低いページがある場合の悪影響を低減する方法として、Context

Graph を用いた Focused Crawling 手法を提案している[4]。

(a) アルゴリズム

Context Graph の構築

最初にトピックを表すドキュメントを与える。このドキュメントが含まれる層を Layer 0 とする。次に、このドキュメントにリンクをはっているドキュメントを求める(backward-crawling)。これにはいくつかの手法があるが、例えば Google のリンク逆引き機能を利用して求めることができる。こうして求められたドキュメントを Layer 1 のノードとし、それぞれから Layer 0 のドキュメント（トピックを示すドキュメント）に対して一方向リンクが張られたグラフを構成する。さらに、Layer 1 の各ドキュメントから同様にしてリンク元ドキュメントを求め、Layer 2 を構成する。これを N 回繰り返すことにより、Context Graph が構成される。図 4.6.3-2 に、N=2 の場合の Context Graph 構成例を示す。

なお、トピックを示すドキュメントを複数準備する場合には、各ページ個別に Context Graph を構成した後に単純に合成することで、Merged Context Graph を構成する。Merged Context Graph においてに Layer 0 に位置するドキュメント群が、Chakrabarti らの手法におけるシーズ URL に相当する。

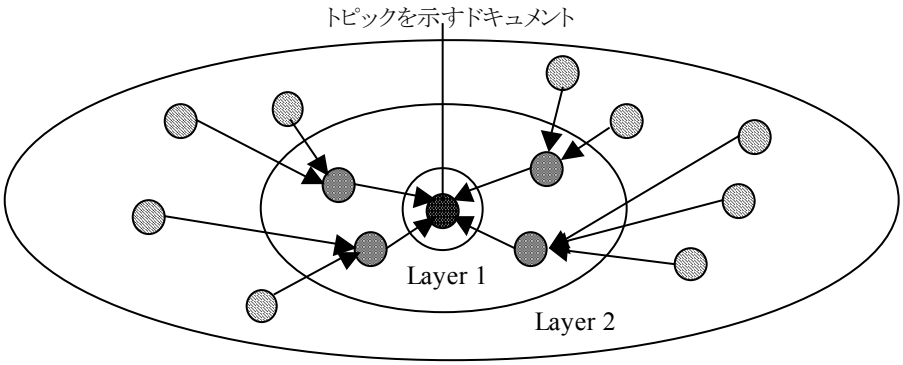


図 4.6.3－2 Context Graph の構成例

Classifier の学習

本手法における Classifier は、Naive Base Classifier の考え方に基づいている。まず、あるドキュメントは、その中に出現する語または句 w_t を次元とするベクトルとして表現される。ベクトルの値は情報検索でよく用いられる TF-IDF により計算される。いま、Context Graph が Layer 0～N の N+1 個のレイヤから構成されるとすると、Classifier の処理は、ベクトルで表現されたドキュメントを、N+1 個のいずれのレイヤに属するか、またはどのレイヤにも属さないかを判別することとなる。

あるドキュメント d_i が Layer j に属する確率 $P(c_j | d_i)$ は、下式により求められる。

$$P(c_j \mid d_i) \propto P(c_j)P(d_i \mid c_j) \propto P(c_j)\sum_{k=1}^{N_{di}} P(w_{di,k} \mid c_j)$$

$w_{di,k}$は、ドキュメント d_iに出現する語または句である。上式により、Context Graph 中のすべてのレイヤに対してドキュメント d_i が属する確率を求め、値が最大となるレイヤ j^*を決定する。ここで、$P(c_{j^*}|d_i)$があらかじめ定められたしきい値より小さい場合、ドキュメント d_iはどのレイヤにも属さないと判定され、そうでない場合にはドキュメント d_iはレイヤ j^*に属すると判定される。

以上のような動作をする Classifier を構築するためには、レイヤごとに、そこに属するドキュメント中の語 w_t の、レイヤにおける出現確率 $P(w_t|c_j)$をあらかじめ計算しておく必要がある。これは Context Graph を構成するすべてのドキュメント中に出現するすべての語について、それがどのレイヤに何回出現したかを求めて正規化することで計算できる。

Crawling プロセス

Context Graph が N+1 個のレイヤから構成されるとき、Crawler は N+2 個の queue を持つ。Crawler は、まず空でない最も番号の小さい queue からドキュメントを取り出し、その中に含まれるリンクをたどって新たなドキュメントを取得する。ドキュメント取得後、Crawler は Classifier によって当該ドキュメントがどのレイヤに属するか（あるいはいずれにも属さないか）を判定し、これにしたがって対応する queue にドキュメントを格納する。ここで、あるドキュメントが Layer 0 に属すると判定され、queue 0 に格納された場合、Crawler はその直接の親ドキュメントを queue 1 に格納するように動作させることもできる。これにより、Context Graph の再構築を行うことなしに、backward-crawling を行って Context Graph を更新したのと同様の効果を得ることができる。

(b) 手法の評価

一般的な幅優先探索手法と、従来の Focused Crawler、および Context Graph を用いた本手法のそれぞれで、"Call for Paper"に関連するドキュメントを収集する実験を行ったところ、本手法により収集されたドキュメントのうち、関連ドキュメントの占める割合は、幅優先探索はもとより、従来の Focused Crawler と比較しても平均で 50～60%高い値となった（ただし、ここでいう従来の Focused Crawler とは Context Graph を Layer 0 のみに制限し、収集順を Naive Base により計算される確率の高い順としているものである）。

(c) 手法の特徴と限界

以上のように、本手法では Context Graph を用いることで、ある程度広い範囲のドキュメントを収集対象としつつ、よりトピックに関連するものから収集するように動作することで、効率良く関連ドキュメントを収集することができる。

本手法においては、以下の改良すべき点があり、今後の検討が必要である。

- トピックとなるドキュメントに対するリンク元ドキュメントが探索できない（トピックドキュメントがサーチエンジンに登録されていない、など）場合に Context Graph を構築する方式
- Classifier において「どのレイヤにも属さない」と判定するしきい値の決定方法
- 収集中に Context Graph を動的に更新する仕組みの実装（現手法では擬似的に実施）

[参考文献]

[1] Cho, J., Garcia-Molina, H., and Page, L.: "Efficient Crawling Through URL Ordering", in

Proceedings of the 7th World-Wide-Web Conference (1998).
[2] Najork, M. and Wiener, J.: "Breadth-first Search Crawling Yields High-quality Pages", in Proceedings of the 10th World-Wide-Web Conference (2001).
[3] Chakrabarti, S., van der Berg, M and Dom, B.: "Focused crawling: a new approach to topic-specific Web resource discovery", *Computer Networks*, Vol.31, No.11-16, pp.1623-1640 (1999).
[4] Diligenti, M., Coetzee, F.M., Lawrence, S., Giles, C.L. and Gori, M.: "Focused Crawling using Context Graphs", 26th International Conference on Very Large Databases, VLDB 2000, pp.527-534 (2000).
[5] Authoritative sources in a hyperlinked environment, in: Proc. ACM-SIAM Symposium on Discrete Algorithms, 1998, also appears as IBM Research Report RJ10076 (91892) and online at http://www.cs.cornell.edu/home/kleinber/auth.ps

４.６.４ ページタイプ判別

ユーザがインターネットを検索する場合、個々の目的に対して、多くの場合、それに適合するWebページのタイプが定められる。例えば、商品購入という目的に対する「商品カタログ」や「オンラインショップ」タイプ、就職・転職という目的に対する「求人情報」タイプなどである。ここでいうページタイプは、テキストの意味内容に深く立ち入らないと判別できないような分類ではなく、むしろ、Webページを一見して得られるような外観的な特徴に基づいて判別できる性格のものである。

ページタイプの判定には、ページ内の特徴的なキーワードに加えて、URL文字列、HTMLタグ構造、リンク数、画像の有無など、スタイル的なファクタにも着目する。あるページタイプに該当する典型的な条件をウェイト付きで列挙したものをルールとして記述しておき、これをどの程度満たしたかによって、そのページタイプらしさのスコアとする[1][2][3][4]。

例えば、「商品カタログ」タイプの判定ルールは次のようなものになる。

- 特定の単語を含む：「商品」「サービス」「製品」「お客さま」「問い合わせ」「価格」「仕様」「特長」
- co.jpドメインで、URLに”product”という文字列を含む。
- <TABLE>タグを使用している。
- ドメイン内へのリンクが多く、ドメイン外へのリンクが少ない。　　など

インターネット上から集めたWebページ群をこのようなページタイプに着目して選別することで、特定目的に特化した情報のみを集めた専門分野Web検索エンジンが構築できる[3][4]。従来、ページ内の単語頻度に着目して、特定のトピック／ジャンルに該当するWebページのみを選別・収集するアプローチがあるが、ここで述べたページタイプによる選別は、トピック／ジャンルとは別な切り口になる。問題解決をドメインとタスクという2面で考えた場合、トピック／ジャンルの指定はドメインを絞り込むことに相当するが、ページタイプの指定はタスクを絞り込むことに相当する[2]。

ページタイプの判別精度に関しては、「求人情報」タイプが適合率98%、再現率66%、「イベント情報」が適合率89%、再現率74%と報告されている[3]。適合率は各タイプ300件のサンプルを判定し

たものであり、再現率はキーワード検索の結果との比較によって擬似的に求めたものである。また、再現率は求められていないが、適合率のみの評価結果として、「商品カタログ」タイプで 89%、「リンク集」タイプで 95%、「調査レポート」タイプで 95%、「プレゼント情報」タイプで 80%なども報告されている[2]。

なお、以上で説明したページタイプ判別方式[1][2][3][4]は、NEC 関西研で開発されたものであるが、着目する様々な特徴別に判別条件をルールの形で人手で記述しておくものであった。一方、NEC 北米研では、着目する特徴はあらかじめ定義しておくものの、それらの重み付けは SVM (Support Vector Machine)によって正例・負例から学習する方式が開発されている[5]。

[参考文献]

[1] 松田勝志、福島俊一: "文書タイプ分類による問題解決向き WWW 検索システムの開発と評価"、情報処理学会研究報告、FI-53-2 (1999)。

[2] K. Matsuda and T. Fukushima: "Task-Oriented World Wide Web Retrieval by Document Type Classification", CIKM'99 - 8th International Conference on Information and Knowledge Management (1999).

[3] 山田洋志、福島俊一、松田勝志: "Web ページからのタイプ別情報抽出・分類方式"、情報処理学会研究報告、FI-57-19 (2000)。

[4] 有吉勇介、福島俊一: "目的および個人に特化したサーチエンジンの開発"、人工知能学会誌、Vol. 16、No. 4 (2001)。

[5] E. J. Glover, G. W. Flake, S. Lawrence, W. P. Birmingham, A. Kruger, C. L. Giles and D. M. Pennock: "Improving Category Specific Web Search by Learning Query Modifications", SAINT 2001 – Symposium on Applications and the Internet (2001).

Proceedings of the 7th World-Wide-Web Conference (1998).
[2] Najork, M. and Wiener, J.: "Breadth-first Search Crawling Yields High-quality Pages", in Proceedings of the 10th World-Wide-Web Conference (2001).
[3] Chakrabarti, S., van der Berg, M and Dom, B.: "Focused crawling: a new approach to topic-specific Web resource discovery", *Computer Networks*, Vol.31, No.11-16, pp.1623-1640 (1999).
[4] Diligenti, M., Coetzee, F.M., Lawrence, S., Giles, C.L. and Gori, M.: "Focused Crawling using Context Graphs", 26th International Conference on Very Large Databases, VLDB 2000, pp.527-534 (2000).
[5] Authoritative sources in a hyperlinked environment, in: Proc. ACM-SIAM Symposium on Discrete Algorithms, 1998, also appears as IBM Research Report RJ10076 (91892) and online at http://www.cs.cornell.edu/home/kleinber/auth.ps

４.６.４ ページタイプ判別

ユーザがインターネットを検索する場合、個々の目的に対して、多くの場合、それに適合するWebページのタイプが定められる。例えば、商品購入という目的に対する「商品カタログ」や「オンラインショップ」タイプ、就職・転職という目的に対する「求人情報」タイプなどである。ここでいうページタイプは、テキストの意味内容に深く立ち入らないと判別できないような分類ではなく、むしろ、Webページを一見して得られるような外観的な特徴に基づいて判別できる性格のものである。

ページタイプの判定には、ページ内の特徴的なキーワードに加えて、URL文字列、HTMLタグ構造、リンク数、画像の有無など、スタイル的なファクタにも着目する。あるページタイプに該当する典型的な条件をウェイト付きで列挙したものをルールとして記述しておき、これをどの程度満たしたかによって、そのページタイプらしさのスコアとする[1][2][3][4]。

例えば、「商品カタログ」タイプの判定ルールは次のようなものになる。

- 特定の単語を含む：「商品」「サービス」「製品」「お客さま」「問い合わせ」「価格」「仕様」「特長」
- co.jpドメインで、URLに”product”という文字列を含む。
- <TABLE>タグを使用している。
- ドメイン内へのリンクが多く、ドメイン外へのリンクが少ない。　　など

インターネット上から集めたWebページ群をこのようなページタイプに着目して選別することで、特定目的に特化した情報のみを集めた専門分野Web検索エンジンが構築できる[3][4]。従来、ページ内の単語頻度に着目して、特定のトピック／ジャンルに該当するWebページのみを選別・収集するアプローチがあるが、ここで述べたページタイプによる選別は、トピック／ジャンルとは別な切り口になる。問題解決をドメインとタスクという2面で考えた場合、トピック／ジャンルの指定はドメインを絞り込むことに相当するが、ページタイプの指定はタスクを絞り込むことに相当する[2]。

ページタイプの判別精度に関しては、「求人情報」タイプが適合率98%、再現率66%、「イベント情報」が適合率89%、再現率74%と報告されている[3]。適合率は各タイプ300件のサンプルを判定し

たものであり、再現率はキーワード検索の結果との比較によって擬似的に求めたものである。また、再現率は求められていないが、適合率のみの評価結果として、「商品カタログ」タイプで 89%、「リンク集」タイプで 95%、「調査レポート」タイプで 95%、「プレゼント情報」タイプで 80%なども報告されている[2]。

なお、以上で説明したページタイプ判別方式[1][2][3][4]は、NEC 関西研で開発されたものであるが、着目する様々な特徴別に判別条件をルールの形で人手で記述しておくものであった。一方、NEC 北米研では、着目する特徴はあらかじめ定義しておくものの、それらの重み付けは SVM (Support Vector Machine)によって正例・負例から学習する方式が開発されている[5]。

[参考文献]

[1] 松田勝志、福島俊一: "文書タイプ分類による問題解決向き WWW 検索システムの開発と評価"、情報処理学会研究報告、FI-53-2 (1999)。

[2] K. Matsuda and T. Fukushima: "Task-Oriented World Wide Web Retrieval by Document Type Classification", CIKM'99 - 8th International Conference on Information and Knowledge Management (1999).

[3] 山田洋志、福島俊一、松田勝志: "Web ページからのタイプ別情報抽出・分類方式"、情報処理学会研究報告、FI-57-19 (2000)。

[4] 有吉勇介、福島俊一: "目的および個人に特化したサーチエンジンの開発"、人工知能学会誌、Vol. 16、No. 4 (2001)。

[5] E. J. Glover, G. W. Flake, S. Lawrence, W. P. Birmingham, A. Kruger, C. L. Giles and D. M. Pennock: "Improving Category Specific Web Search by Learning Query Modifications", SAINT 2001 – Symposium on Applications and the Internet (2001).

ヒューマンインタフェース技術に関する調査報告書

発　行　日　　平成15年４月
編集・発行　　社団法人　電子情報技術産業協会
〒101-0062　東京都千代田区神田駿河台３丁目11番
三井住友海上別館ビル　郵便番号101-0062
TEL　(03)3518-6434
印　　刷　　三協印刷株式会社