IS－09－技標－５

知識情報処理技術に関する
調査研究報告書

２００９年3月

社団法人 電子情報技術産業協会
知識情報処理技術専門委員会

知識情報処理技術委員会　委員名簿

（敬称略・順不同）

委員長	橋田浩一	独立行政法人産業技術総合研究所
幹事	奥雅博	日本電信電話㈱
監事	西野文人	㈱富士通研究所
委員	村田敏樹	沖電気工業㈱
〃	福持陽士	シャープ㈱
〃	木下聡	㈱東芝
〃	安藤真一	日本電気㈱
〃	森本康嗣	㈱日立製作所
〃	清野正樹	パナソニック㈱
〃	今村誠	三菱電機㈱
〃	亀田雅之	㈱リコー
〃	井佐原均	独立行政法人情報通信研究機構
〃	石崎俊	慶應義塾大学
〃	石川徹也	東京大学
〃	辻井潤一	東京大学
〃	徳永健伸	東京工業大学
事務局	佐野眞一	(社)電子情報技術産業協会
〃	金丸洋介	(社)電子情報技術産業協会

言語資源分科会　委員名簿

（敬称略・順不同）

委員長	井佐原　均	独立行政法人情報通信研究機構
幹事	白井清昭	北陸先端科学技術大学院大学
委員	下畑さより	沖電気工業㈱
〃	佐田いち子	シャープ㈱
〃	釜谷聡史	㈱東芝
〃	石川　開	日本電気㈱
〃	寺本やえみ	㈱日立製作所
〃	潮田　明	㈱富士通研究所
〃	谷垣宏一	三菱電機㈱
〃	井田裕子	㈱リコー
〃	井上聡子	独立行政法人科学技術振興機構
〃	内元清貴	独立行政法人情報通信研究機構
〃	黒橋禎夫	京都大学
〃	齋藤邦子	日本電信電話㈱
〃	丸山岳彦	独立行政法人国立国語研究所
オブザーバ	石川真奈見	特定非営利活動法人言語資源協会
事務局	佐野眞一	(社)電子情報技術産業協会
〃	金丸洋介	(社)電子情報技術産業協会

マルチモーダルコンテンツ技術分科会　委員名簿

（敬称略・順不同）

委員長	橋田浩一	独立行政法人産業技術総合研究所
幹事	大沼宏行	沖電気工業㈱
委員	長田誠也	日本電気㈱
〃	橋本三奈子	富士通㈱
〃	大森久美子	NTT情報流通基盤総合研究所
〃	伊藤一成	青山学院大学
〃	石崎俊	慶應義塾大学
〃	斎藤博昭	慶應義塾大学
〃	加藤恒昭	東京大学
〃	野本忠司	国文学研究資料館
事務局	佐野眞一	(社)電子情報技術産業協会
〃	金丸洋介	(社)電子情報技術産業協会

情報アクセス技術分科会　委員名簿

（敬称略・順不同）

委員長	徳永健伸	東京工業大学
委員	藤井　敦	筑波大学
〃	池野篤司	沖電気工業㈱
〃	船田純一	日本電気㈱
〃	佐藤祐介	㈱日立製作所
〃	志賀聡子	㈱富士通研究所
〃	續木貴史	パナソニック㈱
〃	相川勇之	三菱電機㈱
〃	長束哲郎	㈱リコー
〃	森田哲之	日本電信電話㈱
事務局	佐野眞一	(社)電子情報技術産業協会
〃	金丸洋介	(社)電子情報技術産業協会

目　　次

エグゼクティブサマリ

知識管理、テキストマイニング、情報検索など、組織や個人の情報管理・発信や知識創造の効率化に資する情報技術は、急速に進展する社会の情報化・知識化において不可欠な基盤となっている。高度な情報通信ネットワーク社会の出現は、人間の生活世界を構造化し、時空を越えた情報や知識の取得と共有と循環を可能にすることで、個人の生活や組織の業務を急速に変えつつある。このような「知識に基づく社会」において個人と組織が有用な知識を容易に取得・創造・発信・共有し拡大再生産するための知識情報処理技術が、本調査研究の対象である。

知識情報処理技術専門委員会においては、自然言語処理や情報アクセス技術等の知識情報処理技術およびそれに関連するデジタルコンテンツやビジネスの動向に関して調査研究を進めている。平成 20 年度は、マルチモーダルコンテンツ技術分科会、言語資源分科会、情報アクセス技術分科会という 3 つの分科会の活動を進めるとともに、知識情報処理技術に関するシンポジウム「先端 Web 技術は企業を変えるか」を開催し、また知識情報処理技術に関する市場調査を行なった。

コンテンツサービス技術分科会は、対話コンテンツをはじめとする様々なコンテンツの処理・利用技術の調査・発展に資することを目的として活動している。平成 20 年度はこのような具体的な作業から得た知見をもとに、これからますます増えると考えられるマルチメディアコンテンツ等の利便性と効率の高い利用のための処理技術の現状と可能性を探った。

言語コンテンツの意味記述における語義の定式化について、構成素分解に関する学術研究の動向を調査した。意味関係による記述と較べて、構成素分解は意味構成素の設定に困難があり、大規模な語彙知識の表現は難しいが、構成素分解に基づく語彙知識は、より深い意味処理の可能性を持っている。

言語コンテンツへのさまざまなアノテーションの方式に関する ISO/TC37/SC4 の活動に参画して国際標準化の動向を調査するとともに、意味内容記述の方式をオントロジーに基づいて統合する方式に関して検討した。

映像コンテンツの構造化オーサリングツールについては、ビデオコンテンツの意味的な構造化とそれに基づく要約に加えて、編集のためのいくつかの付加機能を開発することにより、ユーザビリティの高い編集および利用の方法についての考察を進めた。

これらに関連して 2 件のヒアリング調査を行った。まず、オブジェクト指向の枠組 CLOS (Common Lisp Object System)に基づくセマンティック Web の編集・開発環境である SWCLOS について調査し、CLOS のメタクラス機能を用いた OWL Full を実装について考察した。また、自然言語の文の基

本的な意味構造の定式化であるフレーム意味論に関してヒアリングし、特に、フレーム意味論の枠組みを実際の語義分析と語彙情報資源構築に適用した例としてのフレームネットについて検討した。

言語資源分科会では、言語資源の研究開発利用に関わる問題点を調査することを目的としている。著作権に関する調査、言語資源カタログの整備、また言語資源に限らず広く言語処理に関する情報提供の場としての言語情報処理ポータルの運営を、言語資源協会（GSK）と協調しながら行っている。平成 20 年度は、言語情報処理ポータルの維持管理、言語資源の公開・利用に関する調査、有害サイト関連調査に関する活動を行った。

言語情報処理ポータルの維持管理については、担当委員が運営方針の策定と基本的なコンテンツ作成を行い、特に会議案内、製品ニュース、コラムは定期的に更新を続け、その他のコンテンツについても掲載するべき情報を見つけ次第迅速に対応している。さらに、言語資源の利用事例の公開、アノテーションツールの調査、形態素解析済みコーパスの公開を行った。

言語資源の公開・利用に関する調査としては、言語資源保有者と利用者の間の合意形成や契約締結を円滑に進めるための参考情報を提供することを目的とし、言語資源の公開や利用に関する現状を調査した。また、言語資源保有者が言語資源を提供する際に自分の言語資源がどのように使われるのかを懸念する場合が多いことから、言語資源の使用目的や使用形態といった観点から、言語資源の利用事例を分類した。さらに、主に利用者側の観点から、代表的な言語資源配布団体の言語資源配布に係る契約形態について調査した。

有害サイト関連調査としては、ネット犯罪に関する調査と有害サイトの現状に関する調査を行った。インターネットなどを利用した犯罪に関する相談件数、インターネットを利用したいじめの現状、文部科学省における提言、教育委員会（大阪府）における活動の調査を行った。有害サイトの現状に関する調査としては出会い系サイト、家出サイト等を対象に基礎的な調査を行った。

情報アクセス技術分科会では、場所や利用目的などユーザの利用状況に応じてサービスを最適化する状況依存型サービスの実態とそのための要素技術について動向調査を行った.

状況依存型サービスに関連する技術として、無線 LAN を使って位置情報を取得するための基礎技術、特に、無線 LAN の基地局の位置のデータベースと電波の強度から位置を推定する技術およびこの技術を使ったサービスについて調査した。

企業内の情報検索サービスの事例について調査した。一般に公開された情報の検索と企業内の情報の検索では、異なる観点のサービスが重要となる。特にセキュリティが重要であり、情報内容も他人に見せるというより自分達が使うという観点が重視されている。

個人の情報利用に関しては、Web 検索のパーソナライズと検索結果・検索履歴を備忘録的に使うこ

とを支援する技術、および個人の活動を記録してそれを生活に役立てるといういわゆるライフログの技術について調査した。

現存する状況依存型サービスをインターネット上で調査し、いくつかの特徴素性の観点から分類・整理した。特徴素性のカテゴリとしては、「サービスに使われるデバイス」、「サービスの成熟度」、「サービスに使われる基礎技術」、「サービスの機能・目的」、「サービスが利用する状況情報」、「サービスが分析する対象とその結果」の 6 つのカテゴリで総計 38 個の素性を用いた。サービスと素性の相関関係を見ることにより、現状のサービスと今後現れるであろうサービスの特徴についてまとめた。

また、この分野の研究動向を調べるため、ACM SIGIR 2008 および WWW2008 の内容のうち、関連が深いと思われるものについてまとめた。

1. はじめに

知識情報処理技術専門委員会においては、自然言語処理や情報アクセス技術等の知識情報処理技術およびそれに関連するデジタルコンテンツやビジネスの動向に関して調査研究を進めている。平成 20 年度は、マルチモーダルコンテンツ技術分科会、言語資源分科会、情報アクセス技術分科会という 3 つの分科会の活動を進めるとともに、知識情報処理技術に関するシンポジウム「先端 Web 技術は企業を変えるか」を開催し、また知識情報処理技術に関する市場調査を行なった。

コンテンツサービス技術分科会は、これからますます増えると考えられるマルチメディアコンテンツ等の利便性と効率の高い利用のための処理技術の現状と可能性を探った。

言語コンテンツの意味記述における語義の定式化について、構成素分解に関する学術研究の動向を調査した。また、言語コンテンツへのさまざまなアノテーションの方式に関する ISO/TC37/SC4 における国際標準化の動向を調査し、意味内容記述の方式をオントロジーに基づいて統合する方式に関して検討した。映像コンテンツの構造化については、編集のための付加機能を開発することにより、ユーザビリティの高い編集および利用の方法についての考察を進めた。

これらに関連して、オブジェクト指向の枠組 CLOS に基づくセマンティック Web の編集・開発環境である SWCLOS、および、自然言語の文の基本的な意味構造の定式化であるフレーム意味論に関してヒアリング調査を行った。

言語資源分科会では、言語情報処理ポータルの維持管理、言語資源の公開・利用に関する調査、有害サイト関連調査に関する活動を行った。

言語情報処理ポータルの維持管理については、会議案内、製品ニュース、コラム等を中心にコンテンツを更新した。さらに、言語資源の利用事例の公開、アノテーションツールの調査、形態素解析済みコーパスを公開した。

言語資源の公開・利用に関する調査は、言語資源保有者と利用者の間の合意形成や契約締結を円滑に進めるための参考情報の提供を目的として行った。また、言語資源の使用目的・形態などの観点から言語資源の利用事例を分類し、主な言語資源配布団体の配布契約の形態を調査した。

有害サイト関連調査としては、ネット犯罪に関する調査と有害サイトの現状に関する調査を行った。

情報アクセス技術分科会では、場所や利用目的などユーザの利用状況に応じてサービスを最適化する状況依存型サービスの実態とそのための要素技術について動向調査を行った.

状況依存型サービスに関連する技術として、無線 LAN を使って位置情報を取得するための基礎技

術について調査した。

企業内の情報検索サービスの事例について調査し、一般に公開された情報の検索との違いとして、セキュリティの重要性や情報を社会で使うという観点の重視について検討した。

個人の情報利用に関しては、Web 検索のパーソナライズと検索結果・検索履歴を備忘録的に使うことを支援する技術、および個人の活動を記録して生活に役立てる技術について調査した。

現存する状況依存型サービスをインターネット上で調査し、いくつかの特徴素性の観点から分類・整理した。

また、この分野の研究動向を調べるため、ACM SIGIR 2008 および WWW2008 の内容のうち、関連が深いと思われるものについてまとめた。

２．コンテンツサービス技術

2.1 はじめに

コンテンツサービス技術分科会は、対話コンテンツをはじめとする様々なコンテンツの処理・利用技術の調査・発展に資することを目的として活動している。コンテンツとしては、言語情報が主に入っている映像データを取り上げ、自然言語処理を用いた言語解析はもちろんのこと、韻律、動作、表情といったモダリティについても分析を行なっている。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、映像データが構造化され、その構造を利用して検索や要約といったより知的で包括的なコンテンツの利用が可能になると考える。

本分科会では、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、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てきた。製作したアノテーションデータは使用ツールとともに、言語資源協会から配布している。

今年度はこのような具体的な作業から得た知見をもとに、これからますます増えるであろうマルチメディアコンテンツの利便性と効率の高い利用のための処理技術の現状と可能性を探った。

2.2 動向調査

2.2.1 語の意味知識の表現

(1) はじめに

質問応答や情報抽出等のテキスト処理の高度化を考えた場合、語の意味の扱いが重要となる。構成的意味論による文の意味処理がある程度まで可能であれば、「オズワルドがルビーに射殺された」ことと「ルビーがオズワルトを射殺した」ことが同じ事態を表現していることは理解できるが、「射殺」という語の意味が原子的なもので他の語の意味と関係づけられていないのであれば、この事態と「オズワルドがルビーに撃たれた」ことや「オズワルドが死んだ」こととの含意関係は理解できない。文法規則に基づく意味関係が比較的少数の規則によって記述できるのに対し、語の意味やその間の関係は個別的な知識であるため、その記述は簡単ではない。大量の個別知識をできる限り系統的に表現する仕組みあるいは枠組みが必要になる。

このような枠組みは大きく２つに分類できる[8)]。ひとつは語の意味（語義）をお互いの関係の中で表現する。類義や対義のような関係、更には含意関係等の意味関係で語義どうしを繋ぐことで、例えば、撃つことと射殺すること、射殺することと死ぬこととの関係が明らかになり、それを通じて、「射

殺」の意味が理解される。もうひとつは少数の意味構成素を用意し、それらの組み合わせによって様々な語義を表現するもので、構成素分解、要素分析と呼ばれる。xがyを射殺することの意味を、xがyを撃つという行為がなされ、その結果として、yが生きていない状態に変化することと表現する。yが死ぬことの意味はyが生きていない状態に変化することである。撃つという行為や生きているという状態が意味構成素であるか、更に分解されるものであるのかは別として、このような表現によって上述のものを含め様々な意味処理が可能となる。

一般に、意味関係による記述と較べて、構成素分解は意味構成素の設定に困難があり、大規模な語彙知識の表現は難しいと言われている。実際、現在多くのテキスト処理で用いられているWordNet（http://wordnet.princeton.edu/）は意味関係による記述であり、その有効性が様々な場面で示されているが、規模の面でも利用頻度の面でもそれに匹敵するような構成素分解に基づく語彙知識は存在しない。

一方で、構成素分解に基づく語彙知識は、より深い意味処理への可能性を持っている。それはひとつに語義についての体系的でより精密な処理であり、もうひとつは文レベルの意味の扱いへと縫い目なく繋がるような意味処理である。

本稿では、語義の記述について、構成素分解を中心としたサーベイを報告する。まず、(2)から(4)で、構成素分解の３つの枠組みを解説する。(2)では影山らが提案するLCS（Lexical Conceptual Structure、語彙概念構造）、(3)ではJakendoffのLCSを説明する。これらは共にLCSとして論じられることが多いが、その背景となる哲学と強調点には幾つかの差異があると思われるので、その点を含めて考察してゆく。(4)ではPustejovskyによるGL（Generative Lexicon）について述べる。LCSが特に動詞に注目し、その意味を厳密に記述して統語的な振る舞いとも関係づけているのに対し、GLは動詞だけでなく名詞を含めた多くの語が様々な意味を持つようにみえる現象に注目し、それを説明するためにそれらが生成される仕組みを提案している。(5)では、これら構成素分解による語義の記述と意味関係による記述、特にWordNetによるそれを比較し、そこで表現されている情報の関係を報告する。最後に(6)で全体をまとめる。

(2)　影山のLCS[6)9)]

影山らは、動詞をVendlerの語彙的アスペクト[14)]に沿って分類し、それぞれを以下のようなLCSで表現する。

状態動詞（stative）	y BE-AT z
活動動詞（activity）	x ACT (ON y)
達成動詞（achievement）	BECOME [y BE-AT z]
到達動詞（accomplishment）	[x ACT-ON y] CAUSE [BECOME [y BE-AT z]]

ここでx, yはその動詞を主辞とした文の項となるものの意味が埋め込まれる変項であり、zは多くの場合、語義によって定まっている定項である。
このような表現により、例えば、openの複数の用法について以下のような関係づけが可能となる。

The door is open. y BE-AT open
The door opened. BECOME [y BE-AT open]
John opened the door. [x ACT-ON y] CAUSE [BECOME [y BE-AT open]]

「殺す」ことと「死ぬ」や「生きている」ことの関係も同様に表現される。

また、語義をこのような構造としてとらえることで、"The Sheriff of Nottingham jailed Robin Hood for four years."において、"for four years"が「投獄するという行為を4年間繰り返した」のか「4年間投獄していた（4年間投獄されていた）」かの曖昧性があることを次のように説明できる。つまり、"jail"は、

[x ACT-ON y] CAUSE [BECOME [y BE-AT in jail]]

という意味を持ち、前者の解釈では"for four years"は[x ACT-ON y] を修飾し、後者の解釈では[y BE-AT in jail]を修飾している。「太郎はもう１時間も車を駐車場に入れている」において、入れている行為が1時間かかっている解釈と入った状態が1時間続いている解釈があることも同様に説明される。
更に、語形成における意味の変化にも説明を与える、un-の付加を考えると、形容詞の場合（例えばkind→unkind）、その形容詞が示す状態の否定（kindではない状態）を表すのに対し、動詞の場合（例えばbend→unbend）、その動詞が示す行為の否定というよりも、その行為の結果を取り消すような逆の行為を表すが、このことをun-付加がそれぞれの語の状態を表現する部分を否定すると考えることで説明できる。つまりunkind, unbendの語義は次のようになる。

unkind [y BE-AT NOT kind]
unbend [x ACT-ON y] CAUSE [BECOME [y BE-AT NOT bent]]

このことにより、あわせてpuhやhitのような活動動詞にun-付加が起きないのは、その語義に状態の表現が含まれないためであると説明される。

語形成のもうひとつの例として、名詞や形容詞に-izeが付加されて動詞が形成される場合（例えばunionize, itemize, normalize）も、動詞の意味は基となる名詞や形容詞の語義N/Aから以下のような構成されるとして説明される。

[x ACT-ON y] CAUSE [BECOME [y BE-AT N/A]]

-izeによって構成される動詞の語義はこれだけにとどまるものではない（例えば、carbonize, hospitalize）が、それらについても系統的な分類が可能となる。

このように語義を意味構成素からなる構造的なものとして扱うことで、語の間の関係や、語の振る舞い、更には語形成について、単なる分類や記述にとどまらない議論が可能となる。この点が構成素分解の利点である。一方で影山のLCSは、様々な動詞の語義やその振る舞いを体系的に記述する枠組みたろうとする動機がやや希薄という印象を受ける。もちろん、上述の4種類の動詞だけの議論にとどまっているわけでない。それに加えて移動を表現するMOVEという構成素もある。「移動する」と「歩く」等はMANNERのような構成素を導入すること区別されるし、状態の表現や使役の表現についてもBE-AT, CAUSEのみではなく、幾つかのバリエーションが示されている。「家を建てる」「夕食を作る」のような作成動詞についても

[x ACT] CAUSE [BECOME　y BE-AT in the world]]

のような表現が可能である。

とはいえ、例えば、次に説明するJackendoffのLCSの背景にあるような（特に動詞の）語義に関する包括的な哲学があるかはやや疑問である。また、その記述もある種の分類にとどまっており、原理的に無限の意味が合成的に生じてくるという印象が薄い。例えば、「買う」のようにモノと金銭の両方の所有権が移動するものや、「止める」のように状態変化を起こさないような働きかけ等をどう表現するかというような議論は見当たらない。

(3)　JackendoffのLCS[7)10)15)]

Jackendoffの概念意味論は、有限の構成要素と有限の組み合わせ原理に基づいて意味を表現し理解することを目的としている。このように意味が生成的であることは、列挙による意味の理解が不可能であることから当然であるとされ、更に統語部門が同じように生成的であることから、それとの並行性が主張されている。一方で、心的な実体として意味・概念を扱うということで、自らを形式意味論と差別化している。また、統語部門との並行性ということで、UG（普遍文法）やX'統語論が意識され、様々な観点でカテゴリを横断する一般化が試みられている。

具体的には、Event, Thing, State, Place, Path等のentityを概念の基本要素とし、これが統語的な句（最大投射）に対応するとする。これらの基本要素は関数（ファンクタ）と項からなる構造を持ち、項の単一化（いわゆる単一化文法のそれとは若干異なる）によって語の意味（概念構造）から文全体の意味が合成されていく。例えば、run, intoの語彙項目は図の様に表現される。それぞれの行に、表記、品詞、統語的な下位範疇化情報、語義である概念構造が並んでいる。下位範疇化情報の添字と語義の添字が対応し（添字iは外項（主語）に対応する）、統語構造を参照して意味が合成されていく。

概念構造の開き括弧に添えられているのは、entityの型の情報であり、意味的な制約としても機能する。John ran into the roomの統語構造と概念構造は以下のようになる。前置詞句がPath、文がEvent等、統語構造と概念構造が対応している。

$$\left[\begin{array}{l}\text{run}\\ \text{V}\\ \text{\_\_\_}\ \langle \text{PP}_j \rangle\\ [_{\text{Event}}\ \text{GO}\ ([_{\text{Thing}}\ \ ]_i,\ [_{\text{Path}}\ \ ]_j)]\end{array}\right] \quad \left[\begin{array}{l}\text{into}\\ \text{P}\\ \text{\_\_\_}\ \text{NP}_j\\ [_{\text{Path}}\ \text{TO}\ ([_{\text{Place}}\ \text{IN}\ ([_{\text{Thing}}\ \ ]_j)])]\end{array}\right]$$

$[_{\text{S}}$ John $[_{\text{VP}}$ ran $[_{\text{PP}}$ into $[_{\text{NP}}$ the room$]]]]$

$[_{\text{Event}}$ GO $([_{\text{Thing}}$ JOHN$],\ [_{\text{Path}}$ TO $([_{\text{Place}}$ IN $([_{\text{Thing}}$ ROOM$])])])]$

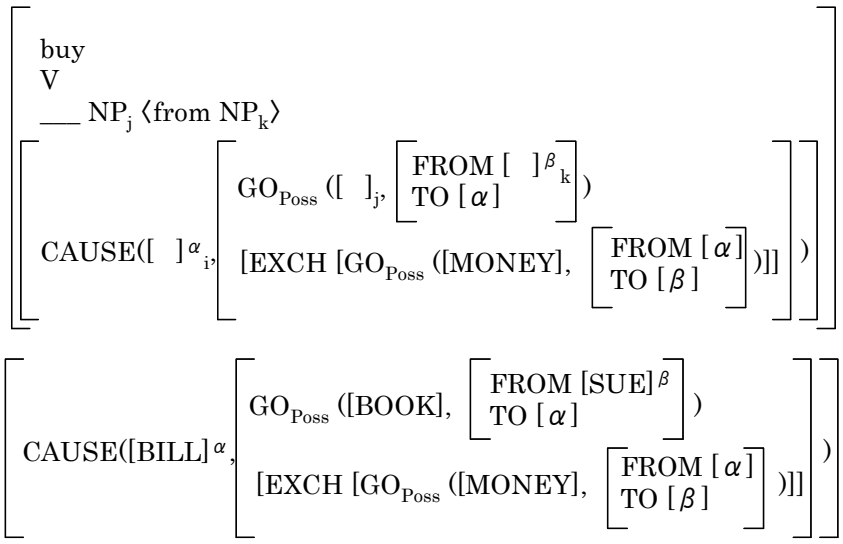
EventにはGO, STAY, CAUSE、 stateにはBE, ORIENT等の基本要素があり。これらは言語に共通な基本的な意味であると共に、意味の場をまたがった共通性を持つ。例えばGOであれば、空間的な移動(The bird went from the tree to the ground)、所有権の移動(The inheritance went to Philip)、属性の変化（The light went from green to red）のすべてに用いられる。このような普遍性・共通性の重視がUGやX'理論を意識したものとなっている。

修飾は、行方向に列挙することで表現される。

John went to the house in the woods quickly.

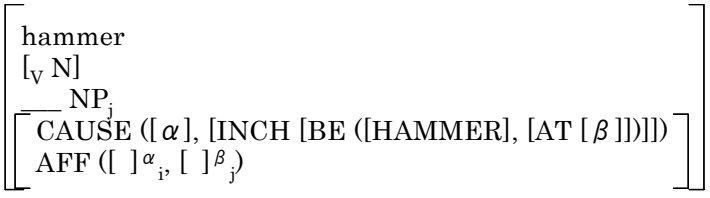
$$\left[\begin{array}{l}\text{GO}\ ([_{\text{Thing}}\ \text{JOHN}],\ [_{\text{Path}}\ \text{TO}\ (\ \underset{\text{Thing}}{\left[\begin{array}{l}\text{HOUSE}\\ [_{\text{Place}}\ \text{IN}\ ([_{\text{Thing}}\ \text{WOODS}])]\end{array}\right]}\ )])\\ \\ \underset{\text{Event}}{[_{\text{Property/Manner}}\ \text{QUICK}]}\end{array}\right]$$

目的や随伴的な事象も同様で、buyの概念構造は、交換事象を表すEXCHを用いて以下のように表現される。”Bill bought a book from Sue.”の概念構造をあわせて示す。ここでα, ßの添字は、統語構造との対応関係を表すi, j, kの添字とは異なり、構造の共有に相当するもので、同じ添字を付与された項は自動的に同じものに単一化されることを意味する。このような仕組みにより、例えば、buyの目的語が代金の移動先であり、品物の移動元であるというように、ひとつの項が動詞に対して複数の意味的な役割を持つことを可能にしている。また、語の意味の内部にMONEYという定項が含まれているが、これは語によって指定されており、この意味クラスに属してそれより特化したもののみと単一化するということで選択制約の表現となる。このような定項の埋め込みはバターを付けるという意味を持つ動詞butter等を表現する場合にも用いられる。

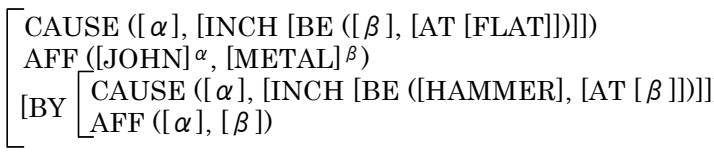


この他にも語の多義性の表現、Agent, Patientを独立した意味の情報（行為層）として扱うこと、統語的な役割と意味的な役割を対応づけるリンキング等の提案がなされており、動詞の語義を分解して表現することで様々な現象の説明が行われている。

人工知能や自然言語処理の研究者にとっては、JackendoffのLCSはSchankのCD[13]を連想させる。実際、意味の表現に関する両者の姿勢はかなり近いと思われる（残念ながらJackendoffの文献にはSchankへの言及は見当たらない）。ただし、JackendoffのLCSでは、統語構造と対応づけられた意味合成の規則が細かく考察されている点が特徴的である。例えば、John hammered the metal flat.のような文は結果構文として知られ、主動詞の語義であるハンマーで叩いたことではなく平らになったことが中心的な意味となり、ハンマーで叩いたことはそれを引き起こした手段としての役割を果たすが、そのような現象を意味合成の規則として定義し、hammerの語義と結果構文という文型から以下のような意味が得られることを説明している。ここで、INCHは状態の変化を表わし、AFFと記述されているのが行為層である。



John hammered the metal flat.



Dorrらは、LCSに基づいた大規模な辞書を作成し、それを中間言語とした機械翻訳システムを提案している。また、同じLCS辞書が外国語学習システムにも利用できることを述べている[1)2)]。

(4) Generative Lexicon[10)11)12)15)]

Pastejovskyは、従来の語義記述のように複数の語義をただ列挙するだけでは多義性（Polysemy）に見られる規則性や語の創造的な使用について適切な扱いができないとし、構成素分解の手法を用いて語義を記述し、意味を構成的に合成する中でこれらを扱うGenerative Lexicon（GL）を提案している。ここでいう創造的な使用とは、例えばa fast boat, a fast typist, a fast gameのように、共通するものを持ちながら、修飾する名詞によって様々にその意味が変化するfastの用法や、wants another cigarette, wants a beer, wants a jobのように目的語によって変化するwantの用法を言う。また、多義性における規則性とは、John baked the potatoes.とJohn baked a cake.のような働きかけと作成の多義が様々な動詞で見られること、名詞においても典型的な多義性のパタン、例えばdoorやwindowの図と地の多義性（John crawled through the window./The window is closed.）やbottleやboxの入れ物と入っている物の多義性（Mary broke the bottle./The baby finished the bottle.）等が存在することをさす。多義の表現をそれぞれの語のレベルで個々の意味を単に列挙することで行うことは、このような一般性を無視していることに他ならない。

GLでは、動詞の語義だけでなく様々な品詞の語義に着目して、言語的な知識と百科事典的な非言語的知識とを区別せずに扱い、それまで動詞の振る舞いと考えられていた現象の源を動詞と結びつく他の語が持つ百科事典的な知識との相互作用に求めている。まず動詞の表現の例として、killの表現を示す。事象構造（Event Structure）がその動詞の意味となる事態を構成する過程（process）や状態（state）とその時間関係を記述する。実際にどのような過程や状態によって構成されているかは特質構造（Qualia）に記述される。これらによって語彙的アスペクトが表現される。この部分の記述は影山のLCSに含まれる情報に近い。項構造（Argument Structure）にはその動詞が項としてとるものの分類とその意味的な制約が記述される。ここに記述される項は統語的に現れるものに限らず、その動詞を含んだ文の意味を構成するために必要なものが記述されている。

$$
\left[\begin{array}{l}
\text{kill} \\
\text{EVENTSTR} = \left[\begin{array}{l} \text{E1} = \text{e1:process} \\ \text{E2} = \text{e2:state} \\ \text{RESTR} = <_{\infty} \\ \text{HEAD} = \text{e1} \end{array}\right] \\
\text{ARGSTR} = \left[\begin{array}{l} \text{ARG1} = [1]\left[\begin{array}{l} \text{ind} \\ \text{FORMAL} = \text{physobj} \end{array}\right] \\ \text{ARG2} = [2]\left[\begin{array}{l} \text{animate\_ind} \\ \text{FORMAL} = \text{physobj} \end{array}\right] \end{array}\right] \\
\text{QUALIA} = \left[\begin{array}{l} \text{create\_lcp} \\ \text{FORMAL} = \text{dead(e2, [2])} \\ \text{AGENTIVE} = \text{kill\_act(e1, [1], [3])} \end{array}\right]
\end{array}\right]
$$

GLに特徴的なのは名詞等の表現での特質構造で、以下の４つの観点から、その名詞の特徴を記述している（動詞が意味する事態の構成要素である過程や状態も特質構造として記述される。その点で動詞とその他の品詞は同じ構造で表現されるが、深いところでの共通性はあるのであろうが、特質構造の意味付けは名詞とそれ以外でだいぶんと異なるように思われる）。

構成役割（constitutive）	材料、成分、部分全体関係等、成り立ちや構成に関する情報
形式役割（formal）	具象物か抽象物か、自然物か人工物か等、それを他から区別する情報
目的役割（telic）	それに意図された目的や機能
主体役割（agentive）	起源や発生等、それを産み出した動作や原因

例えば、小説であれば、構成役割として物語を含むこと、形式役割として本であること、主体役割として書かれるものであること、目的役割として読まれるものであることが記述される。少なくとも後者２つは百科事典的な知識である。

$$\left[\begin{array}{l} \text{novel} \\ \text{ARGSTR} = [\ \text{ARG1} = [1]] \\ \text{QUALIA} = \left[\begin{array}{l} \text{CONST} = \text{hold(e'', [1], narrative)} \\ \text{FORMAL} = \text{book([1])} \\ \text{AGENTIVE} = \text{write(e, z, [1])} \\ \text{TELIC} = \text{read(e', y, [1])} \end{array}\right] \end{array}\right]$$

この特質構造に記述された情報と幾つかの意味合成の仕組みから創造的な言語使用が説明される。そのような仕組みとして、型の強制変換（Type Coercion）、選択的結合（Selective Binding）、共合成（Co-Composition）の３つがあげられている。

型の強制変換では、項となるものの意味的な型が合わない場合、特質構造の情報を使うことで型の辻褄合せがなされる。beginは項として事態を要求するが、そこにa bookのように事態でないものが来た場合、bookの特質構造に含まれる、目的役割の読むという事態、主体役割の書くという事態が使われて型の変換がおこり、本が本を読むあるいは本を書くこととなる。このため、begin a bookは、本を読み始めるあるいは書き始めるという意味となる。同様にwantsも事態を項として要求するので、wants a beerはビールの目的役割が飲まれることであるから、ビールを飲みたいとなり、wants a cigaretteは煙草の目的役割がすわれることであるから、煙草をすいたいと解釈される。

$$\begin{bmatrix} \text{begin} \\ \text{EVENTSTR} = \begin{bmatrix} \text{E1} = \text{e1:transition} \\ \text{E2} = \text{e2:transition} \\ \text{RESTR} = <\text{o}_{\infty} \end{bmatrix} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} = \text{x:human} \\ \text{ARG2} = \text{e2} \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \text{FORMAL} = \text{P(e2, x)} \\ \text{AGENTIVE} = \text{begin\_act(e1, x, e2)} \end{bmatrix} \end{bmatrix} \begin{bmatrix} \text{book} \\ \text{QUALIA} = \begin{bmatrix} \text{AGENT} = \text{write(e, z, [1])} \\ \text{TELIC} = \text{read(e', y, [1])} \end{bmatrix} \end{bmatrix}$$

選択的結合では、修飾語が要求する型と被修飾語の型が一致しない場合、被修飾語の特質構造に含まれる特徴を修飾語が修飾する。例えば、fastは事態を修飾し、typistはその目的役割としてタイプするという事態を含むので、a fast typistにおけるfastは、人物であるtypistそのものではなく、その目的役割であるタイプすることを修飾し、タイプの速いタイピストと解釈される。

共合成は、動詞と項の特質構造に何らかの共通点がある場合に単純な合成にとどまらない変更が生じることをいう。文献では幾つかの例が説明されているものの、形式的に整った定義は見られない。それは問題であるが、考え方として、合成の対象となる項の意味、特に特質構造に記述された百科事典的な情報と交互作用により、様々な意味が産み出されるという点が共通している。例えば、bakeの多義性については、bakeのそもそもの意味はbake_actという目的語となるものを焼くという過程（働きかけ）であり、bakeそれ自体にはそれ以外の意味はない（bake自体に複数の意味があるのではない）とする。bake a cakeが作成の意味を持つのは、cakeがその主体役割としてbake_actを持つからで、この点でbakeとcakeの特質構造が共通することから共合成がおこり、bake a cakeは下記に示すように、焼く（bake_act）という過程の後にその焼かれたものを材料とするcakeが存在するという状態が生じる意味を持つことになる。

$$\begin{bmatrix} \text{bake} \\ \text{EVENTSTR} = \begin{bmatrix} \text{E1} = \text{e1:process} \\ \text{HEAD} = \text{e1} \end{bmatrix} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} = [1] & \begin{bmatrix} \text{animate\_ind} \\ \text{FORMAL} = \text{physobj} \end{bmatrix} \\ \text{ARG2} = [2] & \begin{bmatrix} \text{material} \\ \text{FORMAL} = \text{physobj} \end{bmatrix} \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \text{state\_change\_lcp} \\ \text{AGENTIVE} = \text{bake\_act(e1, [1], [2])} \end{bmatrix} \end{bmatrix} \begin{bmatrix} \text{cake} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} = \text{x:food\_ind} \\ \text{D-ARG1} = \text{y:mass} \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \text{CONST} = \text{y} \\ \text{FORMAL} = \text{x} \\ \text{TELIC} = \text{eat(e2, z, x)} \\ \text{AGENTIVE} = \text{bake\_act(e1, w, y)} \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} \text{bake a cake} \\ \text{EVENTSTR} = \begin{bmatrix} \text{E1} = \text{e1:process} \\ \text{E2} = \text{e2:state} \\ \text{RESTR} = <_{\infty} \\ \text{HEAD} = \text{e1} \end{bmatrix} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} = [1] \begin{bmatrix} \text{animate\_ind} \\ \text{FORMAL} = \text{physobj} \end{bmatrix} \\ \text{ARG2} = [2] \begin{bmatrix} \text{artifact} \\ \text{CONST} = [3] \\ \text{FORMAL} = \text{cake} \end{bmatrix} \\ \text{D-ARG1} = [3] \begin{bmatrix} \text{material} \\ \text{FORMAL} = \text{mass} \end{bmatrix} \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \text{create\_lcp} \\ \text{FORMAL} = \text{exist(e2, [2])} \\ \text{AGENTIVE} = \text{bake\_act(e1, [1], [3])} \end{bmatrix} \end{bmatrix}$$

更に２つの例を挙げる。floatが浮かんでいるという状態を表すのに、float into the caveが浮かんだまま動いてその結果洞窟の中に入っているという過程とその様態としての状態そしてその結果としての状態を表現するということについては、intoの語義として移動するという過程とその結果状態が含まれるとして、それとfloatとが共合成することによって説明される。

$$\begin{bmatrix} \text{float} \\ \text{EVENTSTR} = \begin{bmatrix} \text{E1} = \text{e1:state} \end{bmatrix} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} = [1]\ [\text{physobj}] \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \text{AGENTIVE} = \text{float(e1, [1])} \end{bmatrix} \end{bmatrix} \begin{bmatrix} \text{into the cave} \\ \text{EVENTSTR} = \begin{bmatrix} \text{E1} = \text{e1:process} \\ \text{E2} = \text{e2:state} \\ \text{RESTR} = <_{\infty} \\ \text{HEAD} = \text{e2} \end{bmatrix} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} = [1]\ [\text{physobj}] \\ \text{ARG2} = [2]\ [\text{the\_cave}] \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \text{FORMAL} = \text{at(e2, [1], [2])} \\ \text{AGENTIVE} = \text{move(e1, [1])} \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} \text{float into the cave} \\ \text{EVENTSTR} = \begin{bmatrix} \text{E1} = \text{e1:process} \\ \text{E2} = \text{e2:state} \\ \text{E3} = \text{e3:state} \\ \text{RESTR} = \ \text{e1} <_{\infty} \text{e2, e1} \ o_{\infty}\ \text{e3} \\ \text{HEAD} = \text{e3} \end{bmatrix} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} = [1]\ [\text{physobj}] \\ \text{ARG2} = [2]\ [\text{the\_cave}] \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \text{FORMAL} = \text{at(e2, [1], [2])} \\ \text{AGENTIVE} = \text{float(e3, [1]), move(e1, [1])} \end{bmatrix} \end{bmatrix}$$

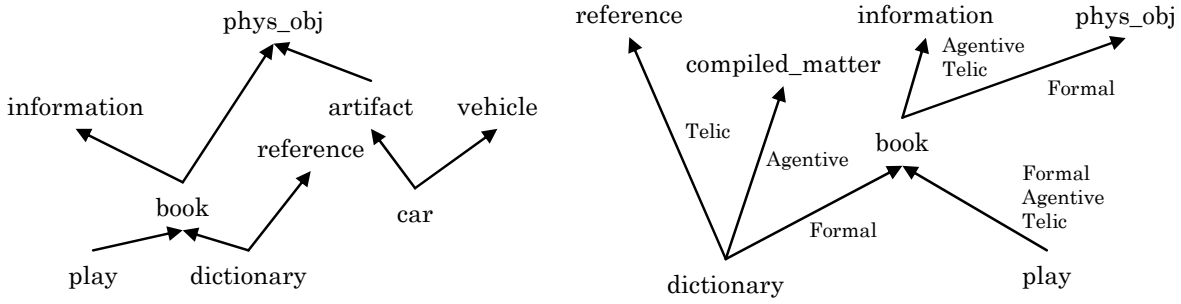
また、hammer the metal flatのような結果構文については、flatの特質構造として何かによってある種の働きかけをされた結果としてflatであるというような情報が記述されていると考えられ、それと共合成することでhammerされた結果flatになるという意味が得られる。なお、形容詞において、ある時点の個体の状況の記述に使われるstage-level predication、例えばsick, hungryと、時間によって変化しない個体そのものの特徴の記述に使われるindividual-level predication、例えばoverweight, intelligentの分類があることに言及し、起源や発生等を記述した特質構造の主体役割の違いがこの分類に対応すると考えて、stage-level predicationとの組み合わせでのみ結果構文が生じることとあわせて、このようなflatの特質構造はひとつの現象を説明するために恣意的に導入されたのではないことが示されている。

$$\begin{bmatrix} \text{flat} \\ \text{EVENTSTR} = \begin{bmatrix} \text{E1} = \text{e1:state} \\ \text{E2} = \text{e2:process} \\ \text{RESTR} = \text{e2} <_{\infty} \text{e1} \\ \text{HEAD} = \text{e1} \end{bmatrix} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} = [1]\ [\text{physobj}] \\ \text{D-ARG1} = [2]\ [\text{T}] \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \text{FORMAL} = \text{flat}(e1, [1]) \\ \text{AGENTIVE} = \text{R}(e2, [2], [1]) \end{bmatrix} \end{bmatrix} \quad \begin{bmatrix} \text{hammer the metal flat} \\ \text{EVENTSTR} = \begin{bmatrix} \text{E1} = \text{e1:state} \\ \text{E2} = \text{e2:process} \\ \text{RESTR} = \text{e2} <_{\infty} \text{e1} \\ \text{HEAD} = \text{e2} \end{bmatrix} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} = [2]\ [\text{physobj}] \\ \text{ARG2} = [1]\ [\text{the\_metal}] \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \text{FORMAL} = \text{flat}(e1, [1]) \\ \text{AGENTIVE} = \text{hammer}(e2, [2], [1]) \end{bmatrix} \end{bmatrix}$$

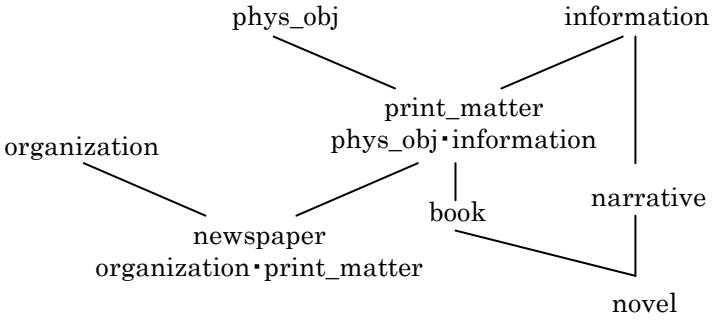
JackendoffのLCSでは、統語的なカテゴリと意味的なカテゴリの対応関係が重視されるので、intoに過程の意味をもたせたり、flatに働きかけの意味を持たせたりすることはそれに反するし、そもそも合成の過程で表に現れてくるような情報が語義の一部として隠れているような記述はできない。そのため、LCSでは、このような現象をある統語的な構成に対応する意味合成の規則の導入で扱うことになる。結果構文については前述の通りである。bakeの多義性については表記法の提案によりそれらをまとめた形で記述することはできるが、それは列挙の別表現としての意味しか持たないように思われる。つまり、多義性をまとめる表記法が非常に強力で任意の多義を記述できてしまうために、その表記法でひとつにまとめられるような多義性だけが存在するというような多義性の規則性を説明するものにはなっていない。

GLのもうひとつの特徴は、名詞の多義性を説明するための語彙の階層構造に関する理論である。図左は一般的な階層的シソーラスであるが、ここで、この階層では共通する特徴を持つはずのplayとdictionaryがどのような動詞の目的語となりうるかを見ると、dictionaryはconsultできるがreadできず、playはその逆であるというような幾つかの違いがあることがわかる。このような現象を説明するために、図右のように語彙の階層構造が特質構造の役割毎に定義されると考える。この例では、playとdictionaryは、形式役割としてbookの下位語として共通点を持つが、その他の役割では異なる特徴

を持つ。例えば、目的役割としてはdictionaryはreferenceの下位語でありplayはinformationの下位語である。consultやreadの目的語としての振る舞いが違うことはこのためと説明される。他にも「食べる」の選択制約を考えてみると、動物、植物、人工物をとり、一般的な階層的シソーラスの中での表現は難しいが、GLの多面的な階層を考えれば、目的役割として食べられるものという情報を持っているものと記述されることになる。



このような多面性を反映して、複合的な型（LCP）という概念が定義される。図のようにbookの上位語である印刷物は物理的実体と情報の2面性を持つ。これをドットによる組み合わせ（phys_obj・information）で表現する。newspaperはさらに「毎朝新聞が人員削減をした」のように新聞社としての語義も持ちうる（換喩とも考えられる）ので、更にそれがドットで繋げられたように表現される。このような複合的な型の存在によって図と地のような典型的な多義性が説明される。また、語義の記述では、複合的な型を構成するそれぞれの側面を項と同じように捉え、以下のように記述される。bookの情報としての側面がx、物理的実体としての側面がyで表現され、その間にholdの関係が存在することが示されている。同様にnewspaperの記述を示す。形式にやや不統一があるが、多義性の性格の違いから来るものと推測される。



$$\begin{bmatrix} \text{book} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} = \text{x:information} \\ \text{ARG2} = \text{y:phys\_obj} \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \text{information・phys\_obj\_lcp} \\ \text{FORMAL} = \text{hold(y, x)} \\ \text{TELIC} = \text{read(e, w, x・y)} \\ \text{AGENTIVE} = \text{write(e', v, x・y)} \end{bmatrix} \end{bmatrix} \quad \begin{bmatrix} \text{newspaper} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} = \text{x:org} \\ \text{ARG2} = \text{y:info・physobj} \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \text{org・info・phys\_obj\_lcp} \\ \text{FORMAL} = \text{y} \\ \text{TELIC} = \text{read(e, w, y)} \\ \text{AGENTIVE} = \text{publish(e', x, y)} \end{bmatrix} \end{bmatrix}$$

(5)　意味関係による記述と構成素分解[3)]

意味関係による語義の表現は語を分析の最少要素とする点で構成素分解による表現と大きく異なる。しかし、語義の意味関係が構成素分解によって得られる意味の構造的な関係を示唆することがある。この点で両者の情報には重なりがある。本節では意味関係による語義の表現であるWordNetにおいて、その意味関係が構成素分解（特にLCS）のどの情報に対応するかを考えてみる。このような関係を明らかにすることで、既存の意味関係による記述を参考に構成素分解に基づく語彙知識を構築できる可能性がある。

WordNetにおける多義の表現はsynsetで表現された語義の列挙であり、GLが注目する規則性への関心は全くない。bakeには”cook and make edible by putting in a hot oven”と”prepare with dry heat in an oven”が列挙されており、bookにも”a written work or composition ...”と”physical objects consisting of a number of pages bound together”が並んでいる。様々な意味関係は語でなく語義を関係づけている。上位下位関係（hypernym, hyponym）が重要な関係となっているが、ここにはGLが主張するような役割による区別はなく、bookの２つの語義が異なる上位語（上位語義）を持つことになる。ちなみに前者の上位語はpublication、後者はproduct/productionである。名詞においてはこの他に部分全体関係（meronym）がGLの構成役割と関連する。以下は動詞について考える。

名詞の下位語、上位語の関係に相当する動詞の基本的な関係はtroponymyとhypernymである。V1することはある様態でV2することであると言える時、V1はV2のtroponymyであり、V2はV1のhypernymである。様態については以下の様に様々な観点で整理される。

機会や形式の違い　fightに対するbattle, war, tourney

意図や動機の違い　communicateに対するexamine, confess, preach

メディアの違い　　communicateに対するfax, e-mail, phone

それらに応じて細かい記述は異なると思われるが、この関係は、LCSにおけるmoveとrunの関係

$$\left[\begin{array}{l}\text{move}\\ \text{V}\\ \text{\_\_\_}\ \langle \text{PP}_j \rangle\\ [_{\text{Event}}\ \text{GO}\ ([_{\text{Thing}}\ \ ]_i, [_{\text{Path}}\ \ ]_j)]\end{array}\right] \left[\begin{array}{l}\text{run}\\ \text{V}\\ \text{\_\_\_}\ \langle \text{PP}_j \rangle\\ \left[\begin{array}{l}\text{GO}\ ([_{\text{Thing}}\ \ ]_i, [_{\text{Path}}\ \ ]_j)\\ \ \ [_{\text{Property/Manner}}\ \text{RUNNINGLY}]\end{array}\right]_{\text{Event}}\end{array}\right]$$

等に対応する。この関係は複数の木からなる森の構造として動詞を関連づける。それぞれの木の根（最上位語）は、motion, perception, contact等14種類の行為（action）とそれ以外として状態isの詳細化としてresemble, belong, suffice、助動詞的なもの制御に関するものとしてwant, fail, prevent、アスペクトに関するものとしてfinish等の動詞に対応している。これらの中のどの木に属するかが動詞の一番大きな分類になっている。仮にこれらの最上位語が語彙分解の基本要素の組み合わせで記述でき

るとすれば、様態の表現の問題はあるにせよ、WordNet中の語義に構成素分解による表現の鋳型を与えることができる。

なお、動詞のこの階層は名詞の上位下位関係に較べ、最大4レベルと浅く、多くの動詞が集中する基本レベルの存在が顕著であると報告されている。例えば、最上位のcommunicateの下にはtalk, write等少数の動詞があるだけだが、talkの下にはbabble, mumble, slur, murmur等多数の動詞が存在する。そしてその下のレベルにある動詞は複合語が中心となりその数も少ない。

その他の関係として、Causeがある。これは変化を引き起こす行為（V1）と引き起こされた変化（V2）との関係であり、他動詞と自動詞のbreakの間の関係がその例となる。LCSではそのふたつが

[V1のLCS] = [x ACT-ON y] CAUSE [V2のLCS]

という関係にあることを示唆する。

TroponymyとCauseはより一般的な関係であるEntailmentの下位関係であると考えることができる。EntailmentはV1することがV2することを含意・伴立するという関係であるが、V1が行われている時間とV2が行われている時間に包含（重複）関係が成り立つかで分類され、それぞれが更に2分される。

時間的な包含関係あり

完全に重なるもの（coextensiveness）→Troponymy march-walk, whisper-speak等

一方が他方に完全に含まれるもの（proper inclusion）　walk-step, snore-sleep等

時間的な包含関係あり

後向き前提（backward presupposition）　forget-know, unwrap-wrap等

原因・因果(cause)　show-see, break-break等

TroponymyとCause以外の２つのEntailmentのうち、後向き前提は、典型的にはある状態とそれを打ち消す変化もしくはその変化を引き起こす行為との関係であり、LCSでの一定の関係に対応づけられる。ただし、fail-try等もこの関係にあり、すべてを一律には扱えない。時間的包含関係はあるがTroponymyではないものについては、GLではその事象構造で記述がなされるものの、LCSには一定の対応がないように思われる。ただ、snoreは、その上位語のLCSと組み合わせて、sleepしている時に音を発することのようなLCS記述が行えるかもしれない。

もうひとつの重要な関係は、Opposition（反義・対義）である。これも幾つかに分類される。ひとつはgive/take, buy/sell, lend/borrowのような逆関係で、同じ出来事を異なる視点・参加者から記述したものである。この関係にある動詞は項との対応付けだけが異なる同じLCSを持つと考えられる。Oppositionで中心的なものは、exclude/include, differ/equal, wake/sleep等の相反する状態や、lengthen/shorten, prettify/uglify, tie/untie, appear/disappear等の逆方向の状態変化である。これら

についてもそのLCSには一定の関係が見いだせる。一方、同じ上位語をもつ動詞の様々な対がOppositionの関係とされ、LCSの観点からは一定の関係を見いだしにくい場合がある。例えば、rise/fall, walk/runは同じ移動の下位語であり、前者はその方向で後者はその速度もしくは様態で対立する。これらは共通のLCSを持ち、そのどこか一部に対立する要素を含むという関係とまでしか特定できない。

このように意味関係による記述は構成素分解に様々な示唆を与える。まず、troponymyによる森の構造は、その中に位置づけられた動詞のLCSの鋳型を提供する。そして、Causeに典型的に見られるようにある意味関係はそれによって関係づけられた動詞のLCSの間の一定の関係を示唆する。それらは動詞の意味関係でありながら動詞のLCSの関係であるとも考えられる。

(6) おわりに

本稿では語の構成素分解の手法を概観した。更に関係による記述との対比を行った。語の意味を分解して表現することで様々な現象を説明することができる。このことは少数の規則で多くの意味処理が可能になることを示唆する。語の意味の表現と同じ枠組みで文の意味も表現されていることも重要であり、語の意味を原子としない深い意味処理の基盤となる可能性を持つ。一方で、それぞれに異なる関心に基づき異なる構成素分解の提案がなされており、そこに記述されている情報も異なっている。テキストの意味処理に利用するという前提でどの枠組みがよいかは一概に決定できるものではない。枠組みの問題に加えて、具体的な意味構成素の設計にも問題が多い。カバー率もしくは枠組みの一般性の主張が強いJackendoffのLCSですら、そこに現れる定項の設計には提案はない。何を目的とした意味の表現であるのかを明らかにした設計が必須である。また、既存の資源を利用して作成が可能であるかも重要であろう。逆に他の資源とどう関われるかも重要である。そのような資源のひとつにオントロジーがある。本稿では扱えなかったが、そのような語の意味表現とオントロジーの関係については、文献 4) 5)に詳しい。

参考文献

1) B. J. Dorr Machine Translation Divergences: A Formal Description and Proposed Solution, Computational Linguistics, 20, pp. 597--633, (1994).

2) B. J. Dorr, Large-Scale Dictionary Construction for Foreign Language Tutoring and Interlingual Machine Translation, Machine Translation, 12(4), pp.271--322 (1998).

3) C. Fellbaum. A Semantic Network of English Verbs in C. Fellbaum (eds). WordNet An Electronic Lexical Database, The MIT Press, pp. 69--104 (1998).

4) 林良彦　言語の意味・概念への計算機科学からのアプローチ in 金崎春幸編「言語文化学への招

待」大阪大学出版会 pp. 221--234 (2008)

5) G. Hirst. Ontology and the Lexicon in S. Staab and R. Studer (ed) Handbook on Ontologies, Springer, pp. 209--229 (2004).

6) 伊藤たかね, 杉岡洋子. 「語の仕組みと語形成」 研究社 (2002).

7) R. Jackendoff. Semantic Structures, The MIT Press (1990).

8) D. Jurafsky, & J. H. Martin. Speech and Language Processing, Prentice Hall

9) 影山太郎. 「動詞意味論 -言語と認知の節点-」 くろしお出版 (1996)

10) 松本裕治 「岩波講座 言語の科学 4 意味」 4.3節 語彙意味論 岩波書店 (1998).

11) 小野尚之 「生成語彙意味論」くろしお出版 (2005).

12) J. Pustejovsky. The Generative Lexicon The MIT Press (1996).

13) R.C. Schank. Conceptual Dependency: A Theory of Natural Language Understanding, Cognitive Psychology, (3)4, pp. 532—631 (1972).

14) Z. Vendler. Verbs and Times, Philosophical Review, 56, pp. 143—160 (1957).

15) 由本陽子 語彙意味から見た文法 in 金崎春幸編「言語文化学への招待」大阪大学出版会 pp. 250--264 (2008)

(加藤恒昭)

2.2.2　動画オーサリングツール

インターネット通信環境の向上に伴い、動画共有サービスの利用者が急増している。米国YouTube[1] を筆頭に、動画ファイルを無制限に無料でアップロードできる仕組みが話題性を集めた。動画コンテンツが情報発信を行う表現ツールの一つとして確立したと言える。総務省の発表によると、ここ２、３年ネット上のデータ転送量が急激に増加しており、その大きな要因にもなっている。

国内では2008 年現在、YouTube とニコニコ動画[2]の二つのサイトに多くの利用者が集まっている。そのほかにも代表的なサービスにはソニーのeyeVio[3]、サイバーエージェントの運営するAmebaVision[4]、exciteのDogalog[5]、フジテレビラボによるワッチミーTV [6]などが挙げられる。動画のオーサリングに関しては、従前は編集専用のツールや高い計算処理能力を持つ計算機を必要としたため、一般の利用者にとって敷居が高いとされてきた。しかし現在ではブラウザ上でオーサリング作業が可能なウェブアプリケーションがいくつかサービスを開始している。

ediPa[7] は動画に加え静止画の編集を可能とし、不要なシーンのカット、部分的に隠すマスク機能、シーンの変化をスムーズに見せるトランジションの他、字幕や手書き文字の埋め込みが可能である。ClipCast [8]もカット編集、切れ目をつなぐ効果、エフェクト、文字載せなどに加え、編集結果をYoutubeにアップロードすることもできる。動画だけでなく写真の対応もしており、エフェクトの数は2008年12月現在で3000を超えている。これらのサービスは今も尚改良が進められている。そのほかにもsprasia [9]、Jumpcut [10]、videoegg [11]、PhotoBucket [12]、JayCut[13]、EditorOne[15] 、Movie Masher[14]などがあげられる。スクリーンショットの一例を図2.2.2-1 に示す。

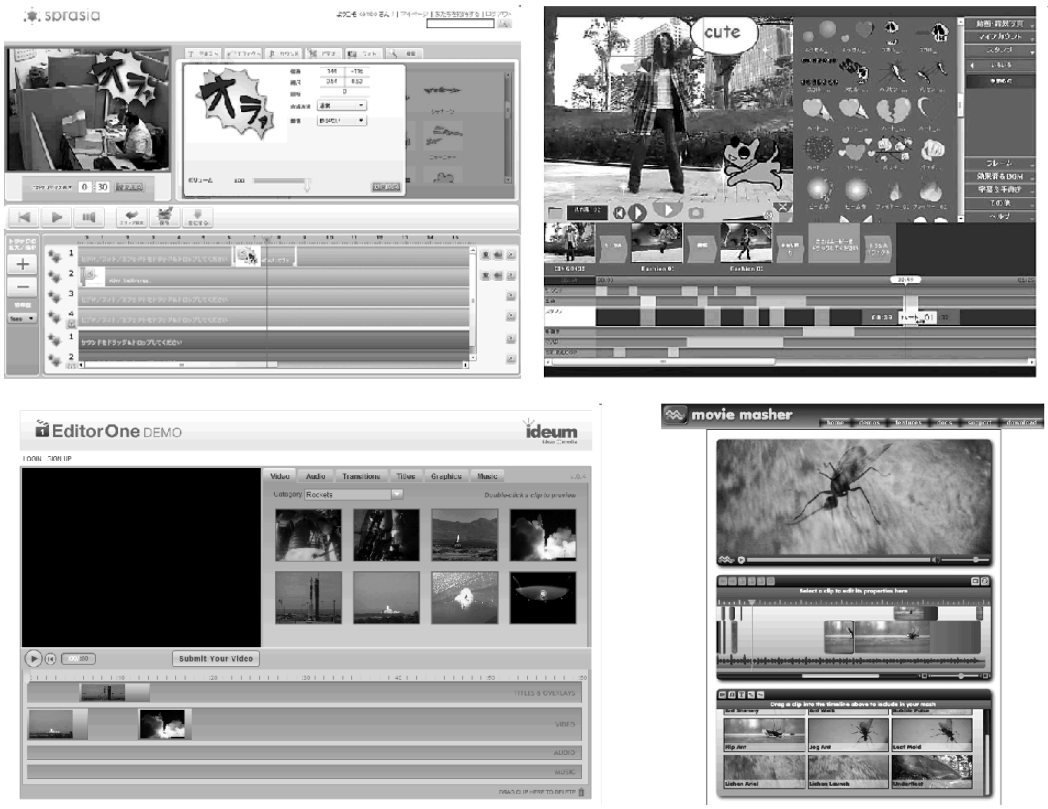


図2.2.2-1：左上：sprasia　右上：ediPa　左下：EditorOne　　右下：movie masher

［参考文献］

[1] YouTube; http://www.youtube.com

[2] ニコニコ動画; http://www.nicovideo.jp

[3] eyevio(アイビオ); http://eyevio.jp

[4] アメブロ; http://www.ameba.jp

[5] ドガログ; http://dogalog.excite.co.jp

[6] ワッチミー！tv; http://www.watchme.tv

[7] ediPa; http://www.edipa.com

[8] Clipcast; http://clipcast.jp

[9] sprasia; http://www.sprasia.com

[10] Jumpcut; http://www.jumpcut.com

[11] VideoEgg; http://www.videoegg.com

[12] Photobucket; http://photobucket.com

[13] Jaycut; http://jaycut.com

[14] EditorOne; http://editorone.ideum.com

[15] Movie masher; http://www.moviemasher.com

（伊藤一成）

2.3　国際標準化

ISO/TC37/SC4 （Language Resources Management; 言語資源管理）においては、さまざまな言語資源に関する言語的構造の記述（アノテーション）の方式、言語資源管理のワークフローなどに関する国際標準化活動を進めているが、国際標準(IS)の作成に直接関わる作業項目（Work Item）だけでなく、そうした作業項目の立ち上げ準備を検討するプロジェクトもあり、その場となるのが TDG（Thematic Domain Group）である。ここではそのうち TDG6（Language Resource Ontology; 言語資源オントロジー）の 2008 年度の活動について報告する。

TDG 6 のミッション

SC4 においては言語的構造の記述の方式に関する国際標準化が進められており、それらを統合するため、LAF （24612: Linguistic Annotation Framework; 言語的アノテーションの枠組）の標準化活動が行なわれている。しかし、LAF そのものには具体的な記述の妥当性（validity）を定義する機能はない。そのような妥当性の定義は FSD（24610-2: Feature Structures – Part 2: Feature System Declaration; 素性システム宣言）によってなされることが想定されているが、実際には、他の具体的な記述方式（対話行為や統語構造の記述の方式）は FSD を使って統一的に表現されているわけではなく、XML の何らかのスキーマ（XML Schema や RELAX NG）や UML によって表現されている。

そこで TDG6 では、多くの記述方法を記述の妥当性を含めてオントロジーに基づいて統一的に表現する方法を検討している。LAF も FSD も何らかのグラフ構造を扱っており、それを RDF（Resource Description Framework）のグラフと見なすことができるので、そのグラフによって具体的な記述の内容を表現し、グラフの妥当性をオントロジーで定義することによって多様な記述方式を統合することができる。これをオントロジー化（ontologization）と呼ぶ。

オントロジー化

オントロジー化とは、オントロジーに基づく再定式化であり、SC4 の場合には、FSD、MAF (24611: Morpho-Syntactic Annotation Framework; 形態統語論的アノテーションの枠組)、LMF（24613: Lexical Markup Framework; 語彙的マークアップの枠組）、WordSeg-1 （24614-1: Word Segmentation -- Part1: Basic concepts and General Principle）、SynAF （24615: Syntactic Annotation Framework; 統語的アノテーションの枠組)、MLIF (24616: Multi-Lingual Information Framework; 多言語情報の枠組)、SemAF-Time (24617-1: Semantic Annotation Framework -- Part 1: Time and Events; 意味的アノテーションの枠組—第 1 部: 時間とイベント）、SemAF-DActs （24617-2: Semantic Annotation Framework – Part 2: Dialogue Acts; 意味的アノテーションの枠組—第 2 部: 対話行為）などの規格が対象である。また、SC3（Systems to Manage Terminology,

Knowledge and Content; ターミノロジー、知識、およびコンテンツを管理するシステム）の DC（12620: Data Categories）もスコープに入ると考えられる。それには、これらの枠組において規定されているアノテーションを RDF グラフに変換する方法と、その RDF グラフの妥当性を定義するオントロジーが必要である。

オントロジー化の最大の目的は相互運用性（interoperability）の確保である。これは、ISO/TC37 の国際標準規格の間での相互運用、それらと他のオントロジーとの相互運用、その他の言語的コンテンツとの相互運用などを含む。

オントジー化の他の目的は、諸規格の仕様をより完全に記述することである。上記のような規格の仕様は XML では表現し尽くせない（XML Schema や RELAX NG では妥当性が定義できない）ので、その大部分は自然言語（ISO の場合には英語およびフランス語）で記述されるが、自然言語による記述はしばしば厳密性に欠ける。また、上述のように UML による記述も用いられているが、UML はほぼオントロジーに相当するので、オントロジー化は UML による記述を徹底することに大体対応する。

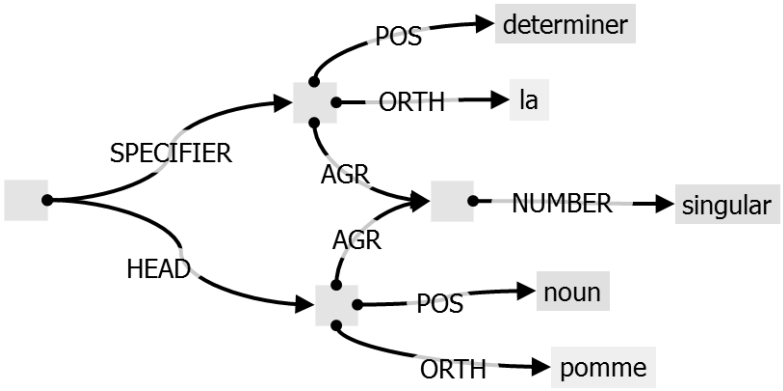
オントロジー化の一環として、DC の仕様の拡張が考えられる。ISO12620 はデータカテゴリを定義する方法を定めているが、各カテゴリが 1 項述語か 2 項関係か、2 項関係の場合には第 1 項と第 2 項のタイプは何か、などを規定できない。ISO12620 をオントロジーで拡張する（または ISO12620 に代わってオントロジーを用いる）ことによって、現在の ISO12620 よりも各データカテゴリの性質を厳密に扱うことができる。

SC4 に限らず多くの標準規格が XML に基づいて定義されているので、オントロジー化は XML から RDF への変換を含む。きわめて大雑把に言って、XML における AnyURI 型および IDREF(S)型の属性（attribute）は RDF の対象属性（object property）に変換され、他の属性はデータ型属性（datatype property）に変換されることになるだろう。また、XML のエレメントの間の埋め込み関係も場合に応じて RDF の属性に変換され、それに対応して XML のエレメントは RDF グラフ中のノード（リソース）またはリテラルに変換されると考えられるが、どのような場合にいずれの形に変換されるかは XML では形式化されていない意味に依存する。

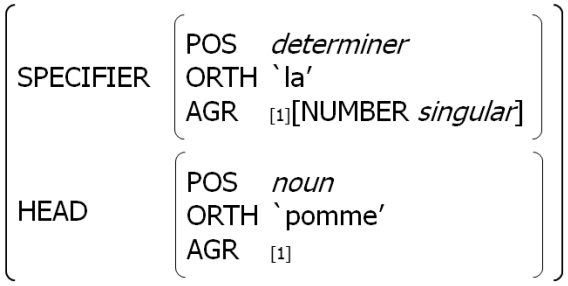
以下では、SC4 のいくつかの規格のオントロジー化に関して検討する。

24610 素性構造

素性構造（feature structure）は主として言語的な構造を記述するために用いられるラベル付グラフ構造であり、RDF とほぼ同じものである。下図に素性構造のグラフ表現を示すが、RDF のグラフ表現も同じものである。



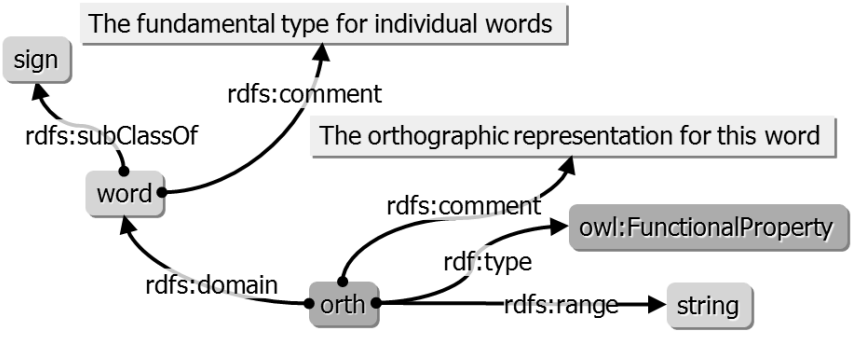
ちなみに下図は素性構造のマトリクス表現である。



下に素性構造のスキーマの一部を示す。これは、word という型の素性構造が orth という素性を持ち、その値が文字列（string）であることを意味する。

```
<fsDecl type="word" baseTypes="sign">
   <fsDescr>The fundamental type for individual words</fsDescr>
   <fDecl name="orth">
      <fDescr>The orthographic representation for this word</fDescr>
      <vRange><string/></vRange>
   </fDecl>
</fsDecl>
```

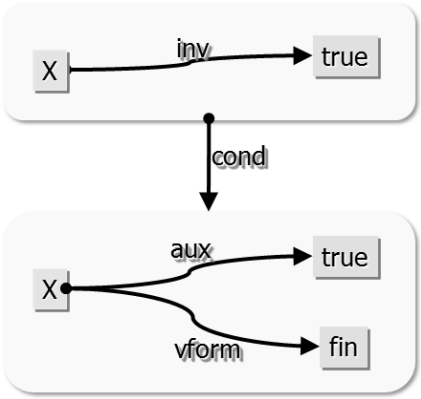
一方、同じ意味のことをオントロジーで表現すると下のようになる。



また、ある素性構造の inv 素性の値が true ならば当該の素性構造の aux 素性の値が true であり vform 素性の値が fin である（つまり、主語と転置された動詞は定型の助動詞である）という意味の制約を FSD に基づいて書くと下のようになる。

```
<cond>
  <fs>
    <f name="inv">
      <binary value="true"/>
    </f>
  </fs>
  <then/>
  <fs>
    <f name="aux">
      <binary value="true"/>
    </f>
    <f name="vform">
      <symbol value="fin"/>
    </f>
  </fs>
</cond>
```

同じ内容の制約を RDF のグラフで表現するには下のような名前付グラフ（named graph）を用いるのが良かろう。ここではこの推育規則の前件と後件をそれぞれ名前付グラフで表現してある。



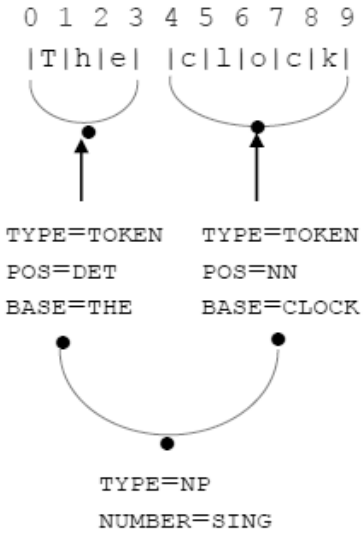
ちなみに、この規則の SWRL (Semantic Web Rule Language) による表現は下記のようになるだろう。

```
inv(?X,true) -> aux(?X,true) & vform(?X,fin)
```

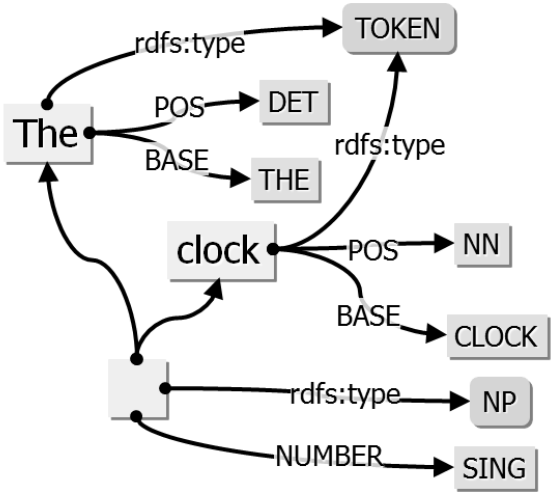
素性構造と RDF はほぼ同じものだが、通常の素性構造のスキーマよりもオントロジーの方が記述力が高い。すなわち、素性構造における素性は部分関数に限られるが、オントロジーにおける属性は部分関数とは限らない。また、素性の間には型階層（type hierarchy）がないのが普通であるが、オントロジーにおける属性の間には型階層がある。このような意味において、オントロジーは素性構造のスキーマを自然に一般化したものと言える。

24612 言語的アノテーションの枠組（LAF）

LAF でも言語的アノテーションを表現するためにたとえば下のようにラベル付グラフを用いる。



TYPE=TOKEN などは FSD で定義される素性であり、この図の下の 3 つの素性の束は FSD に基づく素性構造である。LAF によるこの表現は RDF では下のように表現できるだろう。



ここで、`The'および`clock'という 2 つのノードは、LAF における外付けアノテーション（stand-off annotation）と同様に、それぞれ`The'および`clock'という文字列を何らかの仕方で参照しているものとする。このように LAF も直接的にオントロジー化することができる。

展望

FSD と LAF がオントロジー化可能なので､SC4 の他の規格もオントロジー化が可能と考えられる。それによって複数の規格に基づく言語的構造の記述を統合するには、それら複数の記述の間でどのような対象が共有されるかを確定する必要がある。たとえば、MAF と SynAF とはいずれも統語論的な

構造記述に関する規格なので、それらの構造記述の間で共有される対象は統語論的な対象、つまり語や句や文などの統語範疇（syntactic category）やその素性と考えられる。一方、SemAF-Time や SemAF-DActs などに基づく意味論的な構造記述の間で共有されるのは当然ながら意味論的な対象（個物、事態、時刻、期間など）である。

一方、現状の SC4 の標準規格の範囲では、統語論的な構造記述と意味論的な構造記述の間で共有される対象は存在しない。これは、統語論的な構造と意味論的な構造との関係を記述する規格がないということである。そのような規格があってこそ、言語的な構造の全体を有機的に連関させつつ記述することができ、オントロジーによる相互運用もまた大きな意義を持つことになる。

（橋田浩一）

2.4　作業研究

2.4.1　概要

本分科会では昨年度より、セマンティックオーサリング技術を用いたビデオコンテンツの構造化方式について検討を重ねている。実用を視野にいれたプロトタイプ実装も同時に進めている[1][2]。セマンティックオーサリングでは、伝えたい内容の論理構造を保存するので、人間と計算機の双方にとってコンテンツが分かりやすいのが特徴である。

これまでに情報の要約や検索、アノテーションといった分野を中心に応用研究がなされてきた。これらはテキストコンテンツを対象としているものであったが、動画のショットをメタデータとフレーム画像により表現することで、対象をマルチメディアコンテンツに拡張している。これをセマンティックビデオオーサリングと呼ぶ。本年度は特に、動画のオーサリングツールに関する作業研究を行ったので報告する。

2.4.2　データ構造化

本分科会では、データ構造化方式についての議論を続けている。国立情報学研究所（NII）より提供されている評価用ビデオコンテンツ[3]の中から、日本伝統工芸のドキュメンタリー映像「音色」を構造化対象とした。

当初はショット間の談話関係をリンク付けする構造を想定していた。しかし、グラフ構造を生成するオーサリングは自由度が大きいため、編集を続けるに従って大域的な構造を把握しづらくなる。また関係を示すリンク数も増えるにつれ接続関係もわかりづらくなった。そこで、ショット群を階層的にグルーピングして木構造を構成し、表示位置を固定した上で、必要に応じて談話関係を付加する方式とした。我々の分科会で使用しているセマンティックエディタには、木構造専用の作成モードも搭載されている。図 2.4.2-1 にセマンティックエディタで動画のセグメント群を木構造化した例を示す。左側のウィンドに表示されているのが木構造データである。ショットを構成する代表的なフレーム画像を、木の深さに応じて自動的に配置する。葉以外のノードは、グループを構成する代表ノードのフレーム画像を表示している。葉のノードがショットの実体であり、開始時刻、終了時刻、ショットの説明文などのメタ情報が付加されている。セマンティックエディタの特徴は単なるコンテンツエディタにとどまらず、それに関するオントロジーやスキーマも同様の操作をもって定義できることである。図 2.4.2-1 中の右側に列挙されている内部ウィンド上がそれで、上から談話構造、MPEG-7、 Dublin Core に関する語彙を表現している。これにより属性の可否や属性値の制約に関する支援を受けながらコンテンツを生成できる。

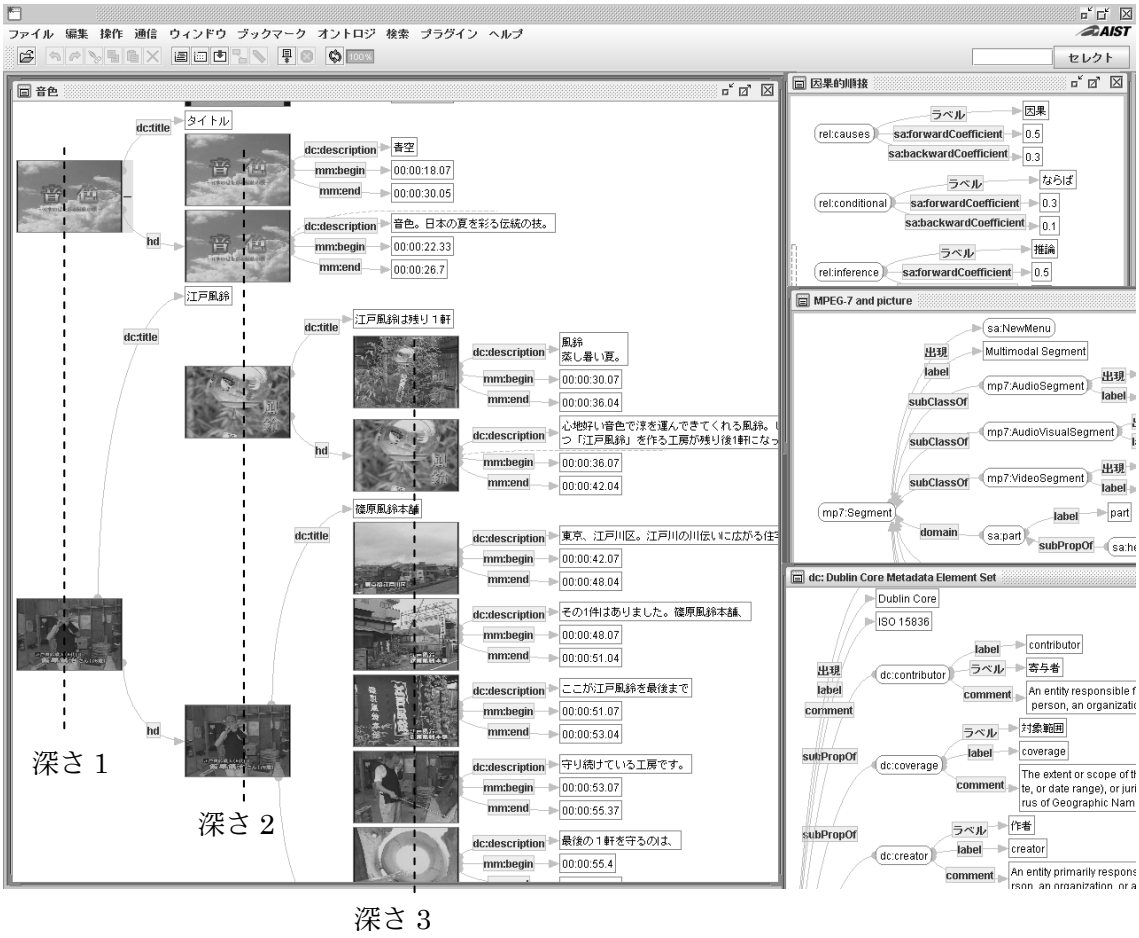


図 2.4.2-1　セマンティックエディタを用いたビデオコンテンツの構造化例

2.4.3　プロトタイプ

今年度は、産業技術総合研究所で開発が進められているセマンティックフレームワークと連携して動作する動画オーサリングアプリケーションを実装した。2.2.5節にて動画編集可能なウェブアプリケーションについて概説したが、高機能ゆえに操作が複雑で習得に時間がかかるものも多い。今回は極力単純かつ直観的な操作性を重視して実装することとした。スクリーンショットを図2.4.3-1 に示す。

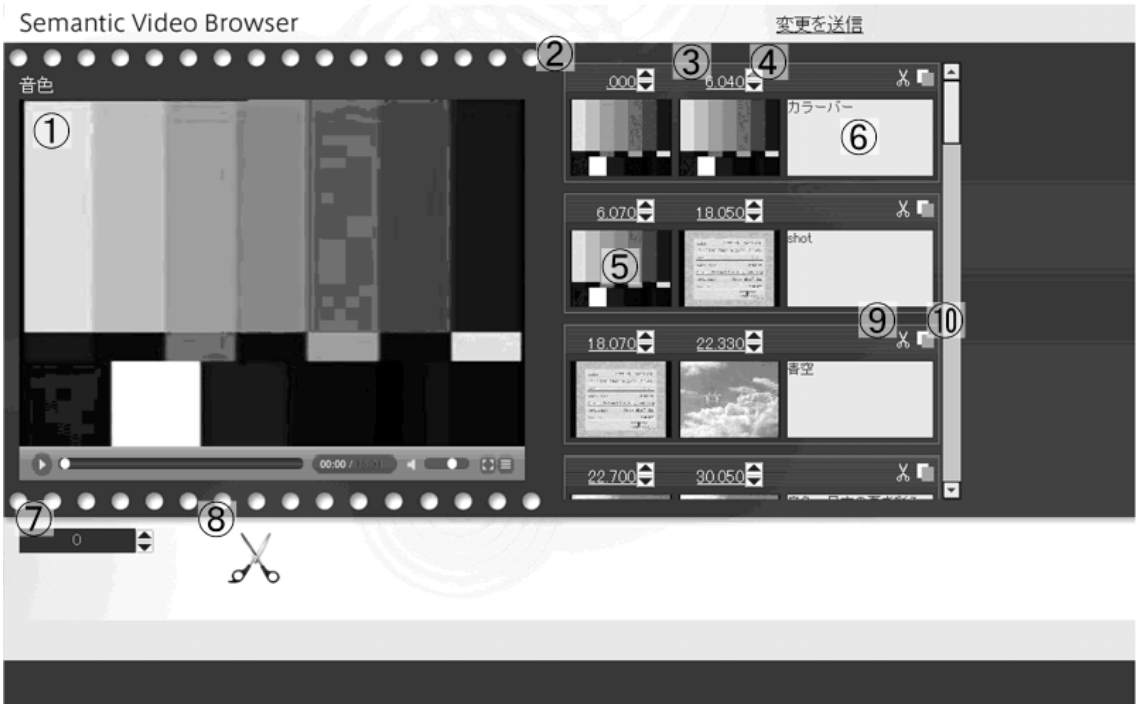


図2.4.3-1 動画オーサリング画面のスクリーンショット

初期状態では、動画全体に相当する単一のセグメントだけが存在し、分割操作を繰り返すことでセグメント数を増やしていく．各セグメントのショットの始点・終点時間は自動的に設定される．基本的に、ユーザが行うべき操作の中心はショットの分割操作となる．また分割されたセグメントを元の状態に戻したい時には、前後のセグメントを結合できる．セグメントの分割と結合のみであるため、複数のセグメントの時間範囲をオーバーラップさせることはできない。図2.4.3-1中に示した構成部品（図中の①から⑩）と可能な操作を説明する。

① 動画プレイヤー

動画の基本操作は映像スクリーン下にあるコントロールバーを用いる。

② セグメントリスト

リスト中のセグメントには、ショットそれぞれの開始時間、終了時間、各時間に対応するサムネイル、そしてそのショットに与えられた説明記述が表示される。

③ 始点・終点時間

各セグメントの始点及び終点の時間を示す。

④　始点・終点時間調節ボタン

始点・終点時間の右端には、上下矢印のボタンを配置している。セグメントの境界時間を変更する。

⑤　サムネイル

該当する時間のショットが表示される。時間の微調整を行うと、調整対象の前後のセグメントの開始時間・終了時間およびサムネイルも連動して修正される。

⑥　説明記述

サムネイルの右側のテキストエリアには、説明記述が記入できる。

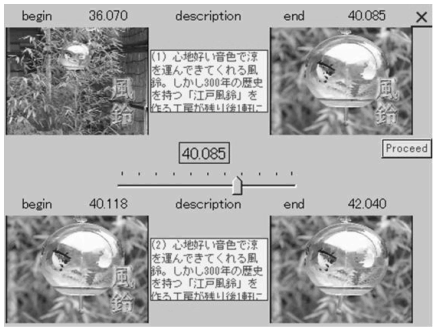
⑦　分割点の時間

動画プレイヤーの現在時間を指す。⑧分割ボタンを押すときに分割点の目安に利用される。また、右側にある上下の矢印をクリックすることでフレーム単位の微調整ができる。

⑧　分割ボタン

動画プレイヤーの中央真下に位置するはさみの画像をクリックすることによって、⑦の分割点の時間を境界としてセグメントを分割する。分割点の時間軸を有するセグメントが分割対象となる。分割されたセグメントは削除される。新規に追加されるセグメントは、一方は削除されるセグメントの開始時間から分割点までの時間を持ち、もう一方は分割点の次のフレームの時刻から削除されるセグメントの終了時刻までの時間を持つ。

⑨　セグメント別分割ボタン



セグメントを個別に分割したい場合に利用する。はさみの形状をしたアイコンをクリックすると新たに右図に示すパネルが表示される。

上半分は分割される前半部分、下半分は後半部分である。中央のスライダーを左右に調節することで、分割点を変更できる。スライダーの上部に表示される、中央部に青く囲われて表示されている値が前半分のセグメントの最終フレームの時間である。次フレームの時刻が後半部分のセグメントの開始時間となる。連動してサムネイルも更新される。分割点を決定すると、分割元のセグメントがセグメントアイテムリストから消去されるとともに、新た

に二つに分割されたセグメントが追加される。

⑩ セグメント結合ボタン

隣接する２つのセグメントを結合する場合に利用する。アイコンをクリックすることで、そのセグメントと、その直下にあるセグメントが結合される。前のセグメントの終点、後のセグメントの始点の時間は破棄される。それぞれの説明記述は結合され、新たに作成されるセグメントの説明記述テキストエリアに入力される。前後の説明記述の間には改行が自動的に加えられる。

2.4.4 まとめ

大量に流通・蓄積されるビデオコンテンツの効果的な活用を目的にした動画コンテンツのオーサリング方式を検討した。また Web アプリケーションのプロトタイプを実装した。

今後は、構造化データを用いた要約、検索方式の提案や、オーサリングアプリケーションのインタフェースに関する検討等について行っていく予定である。

（伊藤一成）

[引用文献]

[1] Hasida, K.：Distributed Semantic Authoring as Foundation of Semantic Computing ,in Notes on From Semantic Web to Semantic World Workshop conjoint with JSAI2003,(2003)

[2] 伊藤一成, 藤原司, 橋田浩一: セマンティックビデオオーサリングによるニュース動画群からのダイジェスト生成, 人工知能学会 第 74 回知識ベースシステム研究会 SIG-KBS-A601,(2006)

[3] 馬場口 登, 栄藤 稔, 佐藤 真一, 安達 淳, 阿久津 明人, 有木 康雄, 越後 富夫, 柴田 正啓, 全 炳東, 中村 裕一, 美濃 導彦, 松山 隆司：映像処理評価用映像データベースについて, 電子情報通信学会技術研究報告, PRMU2002-30, (2002)

2.5 ヒアリング

2008 年度に行ったヒアリングのうち、国立情報学研究所の小出誠二氏のものと、慶應義塾大学の小原京子氏のものを掲載する。

2.5.1 オブジェクト指向に基づく OWL Full 処理系：SWCLOS

WWW が社会的基盤となり、ウェブ検索が人々の生活に必須の技術になりつつあるが、WWW をより知的に組織化する可能性を求めて、セマンティック技術に期待が集まっている．セマンティック技術では、ウェブページやそのコンテンツ中の記述に何らかの説明知識（アノテーション）を付与する

が、そのためには事物の知識を記述するためのオントロジー工学が必要であり、WWW での利用を前提としたオントロジー記述方法として、W3C の勧告にある RDF（S）および OWL が注目を浴びている。RDF はリソースに関する記述を可能にし、RDFS は最低限のクラス分類を可能にし、OWL はより高度な知識記述を可能にするものである。

W3C では、OWL は RDF（S）を前提としてその上に実現されるものとしているが、そのために OWL の OWL Lite、OWL DL、OWL Full の三種類のサブ言語を考案した。このうち完全に RDF 意味論を実現するものが OWL Full である。OWL DL は主に欧州の論理研究者を中心に提案され、一部実現もされているが、記述論理（Description Logic）をベースとしているため、RDF（S）におけるメタクラスの実現において本質的な困難さを抱えていて、これを克服すべく研究が進められている。また詳細仕様の標準化も OWL2 への努力として続けられている。

一方、ANSI Common Lisp におけるオブジェクトシステム Common Lisp Object System（CLOS）では、ベースオブジェクトと同様にクラスもシステム内部でオブジェクトとして実現され、メタクラスを可能とする仕組みも言語機能として備わっている。SWCLOS ではこの CLOS のメタクラス機能を利用して、RDF 意味論に加えて OWL 意味論を実現し、OWL Full を実現した。メタクラス実現には CLOS の作法にしたがったオントロジー構築方法にしたがうことが要求されるが、これを満たすためのいくつかのクライテリアを設けて CLOS CLEAN な方法と称している。メタクラスを考慮したオントロジー構築手法は未開拓な分野であり、CLOS CLEAN 手法はこれに貢献するものである。

本報告では SWCLOS の最大の特徴であるメタクラス機能につき、以下に述べる。SWCLOS 全般については、参考文献やホームページを参照されたい。

RDF 意味論におけるメタクラス

RDF 意味論では RDF グラフ（ノードとエッジを有するラベルつき有向グラフ）をモデルとして意味論を展開する。ある始点ノード／エッジ／終点ノードの組を三つ組みと称して、これを NTriple 構文で記述する。三つ組みにおけるノードの表記を α とし、RDF グラフにおけるそのノードを I(α)とすると、ある三つ組みによる RDF グラフの表示を次のように記述する。

$$\langle\, I(\alpha), I(\beta) \,\rangle \in IEXT(I(\rho))$$

ここで α, β, ρ は三つ組みではそれぞれサブジェクト、オブジェクト、プレディケイトと呼ばれる。I(α)は始点ノード、I(β)は終点ノード、I(ρ)はエッジに相当し、プレディケイトに位置する ρ はその特性から単独でプロパティとも呼ばれる。I(α), I(β), I(ρ)はそれぞれ α, β, ρ の解釈であり、IEXT(I(ρ)) は ρ をプレディケイトとするすべての三つ組みの解釈の集合であり、プロパティ外延と呼ばれる。

RDF 意味論におけるクラス概念はプロパティ rdf:type により次式で導入される。

〈 I(*a*), I(*c*) 〉 ∈ IEXT(I(rdf:type))

このとき、I(*c*)はクラス、I(*a*)はそのインスタンスと呼ばれる。RDFS 意味論ではクラスの外延 ICEXT(I(*c*))を次のように定義する。

I(*a*) ∈ ICEXT(I(*c*)) *iff* 〈 I(*a*), I(*c*) 〉 ∈ IEXT(I(rdf:type))

すなわち、インスタンスはそのクラスのクラス外延の（集合）元であり、そのインスタンス・クラスの 2 項関係は rdf:type プロパティのすべてのプロパティ外延 IEXT(I(rdf:type))の元である。

もう一つの重要な概念はクラスの上位下位関係による包摂概念である。二つの異なるクラスがあって、そのクラス外延において片方 ICEXT(I(*c1*))が他方 ICEXT(I(*c2*))の部分集合のとき、前者 I(*c1*)を後者 I(*c2*)の下位クラスと呼ぶ。つまり、下位クラス I(*c1*)のすべてのインスタンスは上位クラス I(*c2*)のインスタンスでもある。

さて、ここでクラスをインスタンスとするクラス、すなわちメタクラスを考えよう。

I(*c*) ∈ ICEXT(I(*m*)) *iff* 〈 I(*c*), I(*m*) 〉 ∈ IEXT(I(rdf:type))

理論上、このようなクラス・インスタンスの階層はメタクラス、メタ・メタクラス、・・・と無限に考えることができるが、実際には RDFS ではクラス上位下位関係における最上位のクラスを次のように自分自身のインスタンスとすることで、仮想的に無限のクラス・インスタンスの階層を実現する。

I(rdfs:Class) ∈ ICEXT(I(rdfs:Class))

ここで、I(rdfs:Class) ∈ I(rdfs:Class)ではないことに注意されたい。すなわち、RDF 意味論ではクラスオブジェクトとクラス外延を構成するときのクラス役割とを区別する。このようにクラスそのものとそのクラス役割を区別することで、次のような一見不合理な公理も可能となる。

I(rdfs:Resource) ∈ ICEXT(I(rdfs:Class))
I(rdfs:Class) ∈ ICEXT(I(rdfs:Resource))

ここで rdfs:Resource は rdfs:Class も含んですべてのクラスの上位クラスであるが（rdfs:Resource の上位クラスはない）、前述のように rdfs:Class はすべてのクラスのメタクラスであり（1 番目の公理）、rdfs:Class のメタクラス巡回と包摂概念によってクラスとしての rdfs:Class は、メタクラスとしての rdfs:Class が上位クラスである rdfs:Resource に包摂されることにより、rdfs:Resource のインスタンスとなることができる（2 番目の公理）。

RDF 意味論と OWL 意味論の統合

W3C では RDF(S)と OWL の統合について、OWL DL を前提とした RDF 意味論を満足しない統合と RDF 意味論の拡張としての OWL Full な統合について言及しているが、その方法や仕様詳細は明らかになっていない。SWCLOS では CLOS システムの拡張として RDF 意味論を実現した（すなわち CLOS オブジェクトが RDF オブジェクトを包含する）が、同様に RDF オブジェクトが OWL オブジェクトを包含するようにして、 OWL 特有の伴意ルールを実装することで RDF 意味論と OWL 意味論を統合した。繰り返しになるが、このような統合は CLOS のメタクラス機能によって実現されたものである。詳細は参考文献を参照されたい。

CLOS CLEAN なメタモデリング

SWCLOS によれば OWL Full なオントロジー記述と利用が可能となるが、それでも無原則的なメタモデリングが許されるというわけではない。その意味で SWCLOS において可能な、言い換えれば CLOS の意味論を逸脱しない範囲でのメタモデリングを CLOS CLEAN と呼んでいる。大規模オントロジーとして知られている Cyc や、標準的な上位オントロジーと見做されている SUMO は CLOS CLEAN ではない。すなわち、以下のクライテリアを満足しないものである。

クラス *c1* がクラス *c2* の下位クラスであることを $c1 \sqsubseteq c2$ と書くと、以下の関係をクラスとメンバーシップのまっすぐな関係およびねじれた関係ということにする。

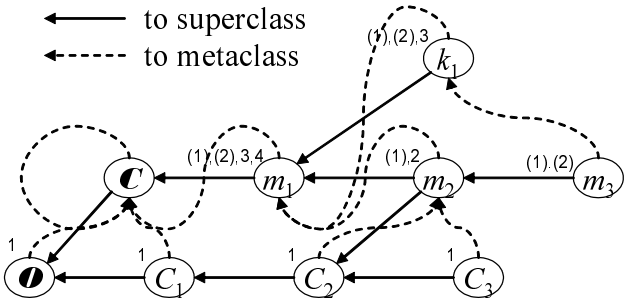
$$I(c1) \in ICEXT(I(c2)) \quad \text{かつ} \quad c1 \sqsubseteq c2$$

$$I(c2) \in ICEXT(I(c1)) \quad \text{かつ} \quad c1 \sqsubseteq c2$$

すると以下の条件を満足するものであれば、SWCLOS においてモデリング可能である。

1. いかなるクラスも上位下位クラス関係において直接あるいは間接にループしてはならない。
2. いかなるクラスもメンバーシップ関係において直接あるいは間接にループしてはならない。
3. すべてのメタクラスは上位クラスとして rdfs:Class(owl:Class)を持たなければならない。
4. クラスとメンバーシップのまっすぐな関係をクラス階層のどこでも置いてよい。これはそのオブジェクトモデルにおいて新しいメタクラス階層を生成するが、あたらしいユニバースを生成はしない。すなわち、クラスの世界を閉じることができない。
5. クラスとメンバーシップのねじれた関係をクラス階層のどこでも置いてよい。これは RDF ユニバース中に OWL ユニバースを生成したように、新しいユニバースを生成する。すなわち、古いクラスの世界と関係なく新しいクラスの中だけに閉じた世界を作ることができる（つくらないこともできる)。

CLOS CLEAN なメタクラスの一例を下図に示す。



ここで、O はrdfs:Resource、C はrdfs:Classである。図中の数字は*Stratification number*と呼ばれるものである。上記クライテリア4は m_1-k_1 関係に見られる。またクライテリア5は C_2-m_2 関係に見られる。Cycではメンバーシップにループがあって上記クライテリア2を満足せず、SUMOはメタクラスではないクラスがクラスをインスタンスとしていて上記クライテリア3を満足しない。

SWCLOS によるメタモデリング例

SWCLOS によるメタモデリングにはすでにいくつかの実例があるが、ここでは簡単に W3C で取り上げている例を以下に示す。この例では最初に OWL クラス Eagle を定義する。Harry はクラス Eagle の個物である。Species はメタクラスであり、EndangeredSpecies は Species の下位クラスである（すなわちメタクラス）。すると Eagle は EndangeredSpecies のインスタンスとしても定義できる。

```
(defConcept Eagle (rdf:type owl:Class))
(defIndividual Harry (rdf:type Eagle))
(defConcept Species (rdf:type owl:Class)
    (rdfs:subClassOf owl:Class))
(defConcept EndangeredSpecies (rdf:type owl:Class)
    (rdfs:subClassOf Species))
(defConcept Eagle (rdf:type Species))
```

EndangeredSpecies をメタクラスとせず、Harry をクラス EndangeredSpecies のインスタンスとしてもよいように思われるかもしれないが、それはオントロジーとして正しくない。絶滅危惧種は鷲であって Harry ではないからである。このように、オントロジー構築においてメタモデリングは必須の機能であり、SWCLOS を用いれば容易にそれを実現できる。

［参考文献］

[1] Hayes, Patrick and Brian McBride. RDF semantics. W3C Recommendation, http://www.w3.org/TR/rdf-mt/, Feb. 2004.

[2] Kiczales, Gregor, Jim des Rivi`eres, and Daniel G. Bobrow. *The Art of the Metaobject Protocol*. MIT Press, 1991.

[3] Koide, S. and Takeda, H.: MetaModeling in OOP, MOF, RDFS, and OWL, *2nd International Workshop on Semantic Web Enabled Software Engineering (SWESE 2006) at the 5th International Semantic Web Conference (ISWC 2006)*, Athens, GA, U.S.A., November 2006.

[4] Koide, Seiji and Hideaki Takeda. OWL-Full reasoning from an object oriented perspective. In *Asian Semantic Web Conf., ASWC2006*, pp.263–277. Springer, 2006.

[5] McGuinness, Deborah L. and Frank van Harmelen: OWL Web Ontology Language Overview, W3C Recommendation, February 2004.

[6] Motik, Boris.: On the Properties of Metamodeling in OWL, *ISWC 2005*, Galway, Springer-Verlag, 2005, pp.548-562.

[7] Paepcke, Andreas, editor. *Object-Oriented Programming – The CLOS Perspective*. MIT Press, 1993.

[8] Patel-Schneider, P. F., Hayes, P., and Horrocks, I.: OWL Web Ontology Language Semantics and Abstract Syntax Section 5. RDF-Compatible Model-Theoretic Semantics., W3C Recommendation, http://www.w3.org/TR/owl-semantics/rdfs.html, Feb. 2004.

[9] http://www.w3.org/2007/OWL/wiki/Punning

（小出 誠二）

2.5.2 フレーム意味論

１． はじめに

本発表では、フレーム意味論について、その理論的背景、内容、適用の一例であるフレームネットの観点から述べる。[1]

本発表の構成は次のとおりである。まず、第２節でフレーム意味論の枠組みの出発点となった格文法ついて述べる。第３節では格文法からフレーム意味論への変遷をたどる。第４節でフレーム意味論の内容の詳細について述べた後、第５節ではフレーム意味論の枠組みを実際の語義分析と語彙情報資源構築に適用した例としてフレームネットを紹介する。その際そのコンテンツについて主にオントロジーの観点からみていく。

２． 格文法と格フレーム

格文法（Case Grammar）とは、言語学者 Charles J. Fillmore が 1968 年に “Case for Case”と言う論文で提唱した文法理論をさす[2]。格文法では、述語と述語の項との意味論的な役割関係を深層格（意味役割）と定義し、文を述語とその深層格との組み合わせから構成されるものとした上で、文の構造や意味を分析する。

２.１． 深層格

格文法における深層格とは、言語に依存しない普遍的な意味役割であるとされ、具体的には Agent, Instrument, Dative, Factive, Locative, Objective などが想定された。これらの深層格のいずれを取るかによって、動詞の意味分類が可能となるともされた。動詞の選択する深層格の集合体を動詞の格フレームという。たとえば、格フレーム｛Experiencer Content｝は経験主を主語とする心理動詞群を表す。“I fear the apocalypse”における FEAR などである。格フレーム｛Agent Patient｝は動作主が何かに変化を起こさせるような出来事を表す動詞群、たとえば“I broke the bubble”における BREAK などを指す。｛Stimulus Experiencer｝はある刺激が経験主に引き起こす感情を表す動詞群で、“The lightning scared me”における SCARE などを示す。

このように、Fillmore の格フレームは動詞の意味分類のために動詞のとりうる意味役割を定義するものであり、Chomsky の下位範疇化フレーム（subcategorization frame）が動詞を分類するために動詞の統語論的コンテクストを定義するのとは対照的である。

格フレームには、一つの深層格は一文に一つしか現れないなどの制約がある。また深層格には必須格と随意格がある。必須格を含まない文は非文となるので、たとえば“I gave the money”は非文である。

[1] 以下は、『言語処理学事典』（共立出版から 2009 年刊行）の著者執筆の「5.2.6 格文法」「5.2.8 フレーム意味論」に基づく。

格文法では、主語や目的語などの文法機能（文法格）は深層格に依存して決まると捉える。言い換えると、格文法では文法機能は意味から生成されるものということになる。Fillmore の 1968 年の論文では主語の選択に関する普遍的規則として次の階層が提案された。

動作主（Agent）＞道具（Instrument）＞対象（Objective）

ある動詞の格フレームが動作主を含む場合には、それが能動文の主語となる。動作主を含まない場合には、階層でその次の深層格すなわち道具が主語となる。

２.２． 格文法の影響

格文法は現代言語学に大きな影響をもたらした。統率・束縛理論以降の生成文法における θ 理論など、多くの理論が深層格の考え方を取り入れている。また、格文法の深層格の階層モデルは日本語の格助詞とよく似ている上に、格助詞のもつ統語機能を「格素性」として捉える可能性を示唆したことから、格文法理論は日本語の深層格関係の分析に適した文法理論として期待された。

さらに、格文法は自然言語処理や人工知能の分野にも多大な影響を与えた。たとえば、多くの機械翻訳システムや最近の自動意味役割推定システムは何らかの形で格文法の格フレームから着想を得ている。但し、日本語処理で格フレームという場合には、意味役割ではなく個々の動詞の取りうる表層格（助詞）のパターンをリストアップしたものをさすことが多いので注意を要する。

３． 格文法からフレーム意味論へ

格文法における意味役割の体系としての格フレームは、Fillmore 自身の言葉によれば、後に自らが提唱したフレーム意味論におけるフレームと連続性を持つものであった[7, 10]。と言うのは、格文法における格フレームとは抽象化された「場面」（scene）や「状況」（situation）を特徴づけるものであり、話者が動詞の意味を理解するにはそのような図式化された「場面」について理解している必要があると Fillmore は考えていたからである。

３.１． 格フレームの問題点

しかしながら、格文法における格フレームは、動詞が表われる文の意味構造を完全に記述するには不適切である。たとえば、GIVE と SEND は同じ格フレームをもつにもかかわらず、“give it to John”と“send it to Chicago”では文の意味に違いがある。逆に、ROB と STEAL とは別々の格フレームを持つため、これらの動詞間の意味上の共通点を記述することができない。BUY と SELL、ENJOY と AMUSE の意味上の共通点についても同様である。そこで、語彙の意味分野（domain）ごとに意味役割の集合をきめ細かく記述できるような枠組みを新たに想定する必要性が生じてきた。

また、語や文の意味を理解するには単に意味役割のリストとしての格フレームを参照するのではなく、より包括的な知識構造を参照する必要があることが明らかとなってきた。実際の言語使用における語や文の意味は、概念素性だけでなく、表現主体の視点や表現主体の持つ文化的価値体系、語や文が用いられる文脈などに大きく依存しているからである。

４． フレーム意味論

Fillmore は 1970 年代に、格文法を発展させた理論的枠組みとして言語使用や理解における背景的知識の重要性に着目し、フレーム意味論（Frame Semantics）を提唱した[3]。Fillmore は有機的相互作用をもつ知識体系が言語の発話や理解のための背景知識として必要不可欠であると考え、そのような知識体系を「フレーム」(frame)とよんだ。フレームの概念を中心にすえた、人間の言語の意味理解プロセス解明のための理論的枠組みがフレーム意味論である。その基本的な考え方は、ある語を理解するには、その語についての言語学的知識だけでなく百科事典的知識も必要であるということである[14]。

[6]が定義するフレームとは、言語の発話や理解の際に必要となる、体系的知識構造を指す。テキストや発話によってその知識構造の一部が使われ活性化されると、他のすべての概念集合との有機的繋がりが想起され利用可能となるような知識構造である。

フレームに参与する意味要素をフレーム意味論ではフレーム要素(frame element)という。フレーム要素は個々のフレームにおいて定義づけられる、フレームに依存したカテゴリーである。格文法における意味役割が有限個であり普遍的なものであるとみなされていたのとは対照的である。また、フレーム要素は個々のフレームによって定義づけられることから、意味役割よりも粒度が高い。さらに、フレームを総体的・包括的なものと捉え、フレーム要素をそのようなフレームの種々の側面ととらえることもできる。格文法において意味役割が述語の項を構成するそれぞれユニークな別個の存在と定義されていたのとは大きく異なる。

４.１． 従来の語彙意味論との比較

フレーム意味論は、それまで主流であった意味素性に基づく語彙意味論（checklist theory of semantics）や場の理論（Field Theory）とは異なる[10]。意味素性に基づく語義分析が不可能なためフレームの概念が必要な例として、BACHELOR と言う語を見てみよう。フレームの概念を用いずに意味素性のみを用いて語義分析をする場合、BACHELOR の意味は、[-married, +adult, +man]と定義できる。しかし、ローマ法王はこの三つの意味素性に基づく条件をすべて満たしているにもかかわらず、法王を BACHELOR と表現するのは不適切である。法王は、一般に男性は成人後結婚すると言う、結婚に関する社会文化的通念の該当しない存在だからである。話者が、法王を BACHELOR と呼

ぶことが不適切だとわかるのは、BACHELOR の意味を構築する際に、意味素性を考慮するだけではなく結婚についての社会的文化的通念をも含んだフレームを想起（invoke）しているからである。

場の理論では、意味場ごとに語を分類して、各々の意味場内において一つの語の語義が他の語の語義とどのように異なるかを分析していく。フレーム意味論では、語と語の相関関係を直接記述するのではなく、各々の語をそれらが喚起（evoke）するフレームに即して記述することでそれらの語義の類似点や相違点を分析する点が異なる。

４.２． フレームの例

例として[4,5]の「商取引」に関する動詞群を見てみよう。Fillmore は、BUY、SELL、PAY、SPEND、COST、CHARGE などの動詞が、同一の「商取引」フレームを喚起するという点で意味的に相互に関連していること、しかもそのフレームの喚起の仕方は動詞によって異なることを示した。

この「商取引」フレームには、「買い手」（Buyer）、「売り手」（Seller）、「商品」（Goods）、「代金」（Money）などのフレーム要素が含まれる。動詞 BUY は、「買い手」の「商品」に関する行動に焦点を当て、「売り手」と「代金」を背景化する動詞である。これに対し、SELL は、「売り手」の「商品」に関する行動に焦点を当て、「買い手」と「代金」を背景化している。そして PAY は、「買い手」の「代金」と「売り手」に関連する行動に焦点を当て、「商品」を背景化している。ここで重要なのは、これらの動詞の意味を理解するには、これらを使用する際の背景や動機付けとなる図式化された場面（フレーム）を話者が知っている必要があると言うことである。

４.３． フレームの利点

また、フレームは視点（perspective）を組み込むことができる[9]。視点を組み込んだフレームが必要な例として、GROUND と LAND と言う二つの語を見てみよう。GROUND と LAND はいずれも「大地」をさすが、前者は「空」との対比で後者は「海」との対比で「大地」を捉える語である。たとえば、“a traveler spent a few hours on the ground”と言う発話を聞けば空の旅が中断されたことがわかるし、“a traveler spent a few hours on the land”と聞けば船による海上の旅が中断されたことがわかる。言い換えれば、GROUND と LAND は同じ場面を、異なる視点や背景に基づいて表現している。GROUND が喚起する視点やコンテクストや背景と、LAND が喚起するそれらとを別々のフレームとして定義することにより、人々が GROUND を含む文と LAND を含む文を理解する際の各々のメカニズムを記述することができる。

さらに、フレームはプロトタイプ現象（prototype effects）も扱うことができる。フレームの定義にプロトタイプを組み込んだ例として、BREAKFAST という語を考えてみよう[7]。この語の意味を理解するためには文化的背景や慣習に関する知識が必要である。つまり、人々はある程度決まった時

間帯に一日三回の食事をとり、睡眠から目覚めた後に朝とる食事においては特定の食べ物を食べる、と言うことをあらかじめ知っておかなくてはならない。しかしながら、これらのすべての条件が満たされなければ BREAKFAST という語を適切に使えないわけではない。たとえば、レストランの看板で見かける“breakfast served all day”と言う表現からもわかるように特定の食べ物や料理を指すこともあれば、食べ物や料理の種類を問わず人々が朝の時間帯に食べるものを指すこともある。また、朝に限らず人々が起きてから初めてとる食事を指すこともある。つまり、BREAKFAST と言う語は様々なコンテクストで使われうる言語カテゴリーであり、BREAKFAST が使用可能なコンテクストであるかどうかは個々の具体的なコンテクストをプロトタイプと照らし合わせて決まる。フレームはプロトタイプによって定義されるからである。

４.４． フレームに類似する概念

Fillmore が 1970 年代に言語使用や理解における背景的知識の重要性に着目してフレーム意味論を提唱したのは、ちょうど認知科学一般においてスキーマやフレームの重要性が着目され始めた頃と時期を同じくする。プロトタイプ理論が生まれ、認知研究が単純な素性から複雑なカテゴリーを対象とするようになったのに続き、人間の認知システムがより複雑な構造を持っていることが、Mandler のスキーマ（schema）、Minsky のフレーム（frame）、Schank のスクリプト（script）などの概念を用いて示されるようになった。Fillmore によれば、フレームと言う言葉を使い始めたのは人工知能学者 Minsky や社会学の Ervin Goffman らの研究を知る前だったそうであるが、フレーム意味論における抽象フレーム（abstract frame）は Minsky のフレームや Mandler のスキーマに相通じるものがあるとのことである[10]。

フレーム意味論のフレームの他に、認知、理解、記憶と具体的な言語表現とを結びつけたものに、Langacker の base と profile の概念がある。Langacker は専門用語「斜辺」（hypotenuse）の意味を理解するにはまず「直角三角形」を理解しなければならないと指摘している。フレーム意味論では専門用語だけでなく、全ての語についてこのような背景知識が必要であると考える。

５． フレームネット

1990 年代後半に Fillmore を代表としてバークレーの ICSI（International Computer Science Institute）において、フレームネット（FrameNet）・プロジェクトが始動した[1]。これはフレーム意味論に基づく、大規模なコーパス（主として British National Corpus）を使用した英語の語彙情報資源構築プロジェクトである。あるフレームを喚起する語彙項目ごとに、コーパスから結合価パタンや連語を考慮して用例文を抽出し、それらの用例文にフレーム要素をタグ付けしていっている。

フレームネットが生まれた背景には、1980 年代以降の、コーパスを利用した英語学習者用英語辞典

の相次ぐ出版に象徴されるようなコーパス辞書学、コーパス言語学の発展が挙げられる[8]。これらがフレーム意味論の理論的枠組みと融合することにより、フレームネット・プロジェクトが開始されたといえる。フレームネットでは、言語で表現できる事象全てをフレームで説明するというフレーム意味論の目標を英語に関して実践しようとしている。また、このプロジェクトを通じてフレーム意味論の理論的枠組みや概念の詳細が議論され改訂されてきている。

フレームネットと連動した、同様の枠組みに基づく他言語に関する語彙情報資源構築プロジェクトとして、スペイン語フレームネット、日本語フレームネット[13]、ドイツ語フレームネットがある。

次節では、フレームネット・データベース内のフレームとフレーム要素について、オントロジーの観点から述べる。

5.1. フレームネットにおけるフレーム・フレーム間関係

まず、フレームには二種類の意味タイプを付与することができる[12]。一つ目は、オントロジー意味タイプ（ontological semantic type）と呼ばれるものであり、もう一つはフレームタイプ（framal type）と呼ばれるものである。

オントロジー意味タイプは、そのフレームを喚起する語彙項目全てが満たしていなければならないような意味タイプである。たとえば、「衣類」フレームを喚起する全ての語彙項目は「人工物」という意味タイプをもつ語彙項目でなければならない。

次に、フレームタイプであるが、現在フレームネットでは二種類のフレームが想定されており、一つは言語には関わらないが解釈に必要な抽象フレーム（abstract frame）、もう一つは具体的に言語表現に関わり、語彙的視点を加味した語彙フレーム（lexical frame）である。そしてこれらは、下層フレームは上層フレームから多くの情報を引き継ぐという、継承（inheritance）の概念によって関連付けられている。継承の他にも、サブフレーム（Subframe）、視点（Perspective On）、使用（Using）、使役相（Causative Of）、起動相（Inchoative）、参照（See Also）、先行（Precedes）などのフレーム間関係（fame to frame relations）が想定されている[15]。

5.2. フレームネットにおけるフレーム要素・フレーム要素間関係

以下では、まずフレーム要素のタイプについて述べ、次にフレーム要素間の意味関係について述べる。

フレーム要素には意味タイプを付与することができる。また、全てのフレーム要素には core status と呼ばれる情報が付与されている必要がある。

最初にフレーム要素の意味タイプについてみてみよう。あるフレームにおいてそのいずれかのフレーム要素に意味タイプが付与されている場合、そのフレームを喚起する語彙項目を含む文において、

そのフレーム要素に該当する語句はその意味タイプをもっていなければならない。

たとえば、以下の（1）の主動詞「知る」は「知覚」フレームを喚起する語彙項目である。フレームネット・データベース内のこの「知覚」フレームの定義を見ると、そのフレーム要素である「認識者」にはあらかじめ意味タイプ「知覚者」が付与されている。従って、(1) の[ ]内の「学生は」は知覚者を指すのでこの文は適格であるが、「学生は」の代わりに「部屋は」とすると文は非文となる。

(1)「知覚」フレーム

[学生は 認識者]正解を知らない。

フレームネット上では現在 28 の意味タイプが定義されている。

次に、フレーム要素の coreness status であるが、すべてのフレーム要素には、core、peripheral、extra-thematic のうちのいずれかの coreness status が付与されている必要がある。Core とは、そのフレーム要素が当該フレームに概念的に必要不可欠なフレーム要素であることを示す。Peripheral なフレーム要素とは、どのフレームにおいても出現しうるフレーム要素を指し、例としては「時間」、「場所」、「様態」、「手段」、「程度」などがある。Extra-thematic なフレーム要素とは、それ自体が独自のフレームを喚起するようなフレーム要素を指し、「受益者」、「頻度」などが例として挙げられる。

最後に、フレーム要素間の意味関係について述べる。フレームネットでは現在“Excludes”、“Coreness Set”, “Requires”と言う三つの関係が定義されている。 “Excludes”とは、これらのフレーム要素のうちのいずれかが文中に現れる際には、同じ文中には他方は存在し得ないようなフレーム要素間の関係を指す。たとえば、「殺人」フレームの二つのフレーム要素「殺人者」と「原因」の間の関係がそうである。(2) は「殺人」フレームを喚起する語彙項目“kill”を含む文である。(2a)にはフレーム要素「殺人者」が現れるが「原因」は現れず、(2b)では逆に「原因」のみで「殺人者」は現れない。

(2) a. [A criminal 殺人者] killed two people by shooting.

b. [The earthquake 原因] killed everyone instantly.

“Coreness Set”とは、複数のフレーム要素から成る集合において、それらのフレーム要素のうちの一部が文中で実現されれば文として意味を成すような関係を指す。たとえば、「自律移動」フレームにおける、「起点」、「終点」、「経路」、「方向性」のフレーム要素から成る集合は “Coreness Set”である。これらのフレーム要素のうちどれか一つが、「自律移動」フレームを喚起する語彙項目（例：歩く）を含む文中に現れれば、その文は適格となるからである。

“Requires”とは、複数のフレーム要素から成る集合において、いずれかが文中に実現されれば他のフレーム要素も同じ文中に実現されていなければならない、とするフレーム要素間の関係である。現在フレーム要素間で “Requires”の関係が定義されているフレームは、フレームネット・データベース上に 56 フレーム存在する。

以上、フレームネットにおけるフレーム、フレーム要素に関する様々な意味関係、意味的制約について述べた。これらに基づき、フレームネット上のデータを既存の複数のオントロジーに結びつけ、自然言語の推論関連の様々なアプリケーションに応用する試みが既に始まっている[16]。

６．まとめ

フレーム意味論に関する本発表の要点は、以下の三点に集約できる。

１） 格文法における格フレームが、フレーム意味論における意味フレームへと発展していった。語や文の意味を理解するには、普遍的、有限個の意味役割（深層格）のリストでは不十分であり、表現主体の視点や文化価値体系、語や文が用いられる文脈などを考慮する必要がある。

２） フレーム意味論の理論的枠組みとコーパス言語学の方法論が結びつきフレームネットへと発展していった。フレームネットでは、意味フレームを用いることにより、言語で表現しうる事象すべてを記述することができると言う仮説に基づき、語彙情報資源構築が行われている。

３） フレームネット・データベースに定義されたデータ構造間の意味関係・意味的制約は、既存オントロジーと結びつき自然言語処理アプリケーションに必要不可欠な推論機構に発展する可能性を秘めている。

参考文献

[1] Baker, Collin. “Frame Semantics in Operation: The FrameNet Lexicon as an Implementation of Frame Semantics.” In *The Fourth International Conference on Construction Grammar Plenary Lectures.* pp.34-43. (2006).

[2] Fillmore, Charles J. “The case for case.” Bach, E. and Harms, R. (Eds.). Universals in Linguistic Theory. Holt, Rinehart, and Winston, New York. 1-88. (1968).

[3] Fillmore, Charles J. “Frame semantics and the nature of language.” In Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech. Vol. 280: 20-32. (1976).

[4] Fillmore, Charles J. “The case for case reopened.” Cole, Peter and Sadock, Jerrold (Eds.). Syntax and Semantics. Vol.8: Grammatical Relations. Academic Press, New York. 59-82. (1977a).

[5] Fillmore, Charles J. “Topics in lexical semantics.” Cole, Roger W. (Ed.). Current Issues in Linguistic Theory. Indiana University Press, Bloomington. 76-138. (1977b).

[6] Fillmore, Charles J. “Frame semantics.” The Linguistic Society of Korea (Ed.). Linguistics in the Morning Calm. Seoul: Hanshin. 111-137. (1982).

[7] Fillmore, Charles J. “A private history of the concept ‘frame’.” Dirven, Rene and Radden, Gunter. (Eds). Concepts of Case. Gunter Narr Verlag, Tubingen. 28-36. (1987).

[8] Fillmore, Charles J. and Atkins, B.T.S. 1994. “Starting where the dictionaries stop: The challenge for computational lexicography.” Atkins, B.T.S. and Zampolli, A. (eds.) Computational Approaches to the Lexicon. Oxford: Oxford University Press. 349-393.

[9] 藤井聖子，小原京子「フレーム意味論とフレームネット」，英語青年，Vol.14. No.6，pp.373-376.(2003).

[10] 長谷川葉子、小原京子「Charles J. Fillmore 教授に聞く」、『英語青年』、Vol.152. No.6: pp.354-359. (2006).

[11] Hasegawa, Yoko, Kyoko Ohara, Seiko Fujii, Russell Lee-Goldman, Charles J. Fillmore. “Constructions for measurement and comparison in Japanese and English.” Paper presented at the Fifth International Conference on Construction Grammar (ICCG-5). University of Texas at Austin. September 28th, 2008. (2008).

[12] Lonneker-Rodman, B., C.F.Baker. “The FrameNet Model and its Applications.” ms.

[13] 小原京子「フレーム意味論と日本語フレームネット」、『日本語学』、Vol.25. No.6: pp.40-52. (2006).

[14] 大堀壽夫「語彙記述におけるフレーム意味論」、『日本認知言語学会論文集第 5 巻(JCLA 5)』、617-620. (2005).

[15] Ruppenhofer, Josef, Michael Ellsworth, Miriam R. L. Petruck, Chris R. Johnson, and Jan Scheffczyk. “FrameNet II: Extended Theory and Practice.” (2006).

[16] Scheffczyk, J., C.F. Baker, S. Narayanan. “Reasoning over Natural-language Text by means of FrameNet and Ontologies” In Ontologies and Lexical Resources for Natural Language Processing. Chu-Ren Huang et al. (eds) Cambridge University Press. (2007).

(小原京子)

3．言語資源

3.1　はじめに

言語資源分科会では、平成 20 年度において、(1)言語情報処理ポータルの維持管理、(2)言語資源の公開・利用に関する調査、(3)有害サイト関連調査に関する活動を行った。

言語情報処理ポータルの維持管理においては、言語情報処理に関するさまざまな情報を集約したポータルサイトを 2002 年夏より公開しており、担当委員が運営方針の策定と基本的なコンテンツ作成を行っている。ポータルサイトには多くの情報が掲載されているが、特に会議案内、製品ニュース、コラムは定期的に更新を続けている。その他のコンテンツについても掲載するべき情報を見つけ次第迅速に対応している。当ポータルサイトは言語情報処理に関連する情報を網羅的・総合的に提供しており、当該分野および関連分野の研究者、開発者、学生などから好評を得ている。安定して月平均 2000 程度のアクセスがあり、定期的に閲覧している利用者が多いことが確認できる。今年度はコラム執筆者の交替を行ったほか、新たな活動として、言語資源の利用事例の公開、アノテーションツールの調査、形態素解析済みコーパスの公開を行った。

言語資源の公開・利用に関する調査としては、言語資源保有者と利用者の間の合意形成や契約締結を円滑に進めるための参考情報を提供することを目的とし、言語資源の公開や利用に関する現状を調査した。具体的には、言語資源の利用事例の実態調査として、研究論文を対象に、どのような言語資源を使ってどのような研究開発が行われているかといった言語資源の利用実態を調査した。また、言語資源分科会では非営利団体の言語資源協会（GSK）と協力し、より多種多様な言語資源の流通促進に向け活動しているが、言語資源保有者が言語資源を提供する際に自分の言語資源がどのように使われるのかを懸念する場合が多いことから、言語資源の使用目的や使用形態といった観点から、言語資源の利用事例を分類した。さらに、主に利用者側の観点から、代表的な言語資源配布団体である LDC (Linguistic Data Consortium)と ELDA (Evaluations and Language resource Distribution Agency) について、現在採られている言語資源配布に係る契約形態について調査した。

近年、インターネットを利用した犯罪やいじめが大きな社会問題となっている。言語資源を活用した自然言語処理技術により、これらの問題を抑止する技術開発が可能になると考えられる。たとえば、有害コンテンツの発見・監視に自然言語処理技術を適用することが考えられる。しかしながら、有害サイトを分析研究する心理的負担や、有害サイトの実態に関する公開情報の不足などから、自然言語処理の研究分野において、有害サイト監視等への本格的応用が議論される機会は少なかった。このような状況を踏まえ、新たなテーマとして、有害サイトに関する調査を開始した。今年度は、ネット犯罪に関する調査と有害サイトの現状に関する調査を行った。ネット犯罪に関する調査としては、イン

ターネットを利用した犯罪やいじめを取り巻く現状を調査した。具体的には、インターネットなどを利用した犯罪に関する相談件数、インターネットを利用したいじめの現状、文部科学省における提言、教育委員会（大阪府）における活動の調査を行った。有害サイトの現状に関する調査としては出会い系サイト、家出サイト等を対象に基礎的な調査を行った。

(執筆者 井佐原均)

3.2 言語情報処理ポータルの活動報告

3.2.1 概要

言語資源分科会では、言語情報処理ポータル(http://nlp.kuee.kyoto-u.ac.jp/NLP_Portal/)(図3.1.1-1)の運営を継続して行っている。言語情報処理ポータルは、言語情報処理に関するさまざまな情報を集約したポータルサイトで、2002 年夏より公開している。担当委員が運営方針の策定と基本的なコンテンツ作成を行っている。

2009 年 2 月時点で、次のような内容を掲載している。

- 会議案内
- 製品ニュース
- 新刊案内
- コラム
- 言語情報処理 用語集
- 講義資料リンク集
- 論文データベースリンク集
- 人材募集
- 言語資源カタログ
- プロジェクト・研究機関・学会等
- 世界の言語イニシアティブの紹介
- 主要国際会議の Wiki ページ

上記のうち、会議案内、製品ニュース、コラムは定期的に更新を続けている。その他のコンテンツについても掲載するべき情報を見つけ次第迅速に対応している。このように当サイトは言語情報処理に関連する情報を網羅的・総合的に提供しており、当該分野および関連分野の研究者、開発者、学生などから好評を得ている。安定して月平均 2000 程度のアクセスがあり、定期的に閲覧している利用者が多いことが確認できる。

図 3.1.1-1 言語情報処理ポータル

以下、言語情報処理ポータルのコンテンツ充実に関する本年度の活動について報告する。

- コラム執筆者の交替

  ポータルサイトでは平日に日替わりでコラムを掲載している。これまでは 10 名の執筆者が担当していたが、担当期間が長期に渡ったため、一部の執筆者を除きコラムの担当者を交替した。新しいメンバーは現在の言語資源分科会のメンバーを中心とした 11 名である。また、コラム執筆者の交替は、コラムの内容に新鮮味を持たせる効果も期待している。

- 言語資源の利用事例の公開

  ポータルサイトで掲載している言語資源カタログの新しい掲載項目として、「言語資源の利用事例」を追加した。

- アノテーションツールの調査

  現在公開されているフリーのアノテーションツールを調査し、言語資源カタログに追加した。

- 形態素解析済みコーパスの公開

  フリーで公開されているテキストを形態素解析し、形態素解析済みコーパスとしてポータルサイトにて一般に公開した。

以下、「言語資源の利用事例の公開」の詳細については 3.2.2 で、「アノテーションツールの調査」については 3.2.3 で、「形態素解析済みコーパスの公開」については 3.2.4 で述べる。

(執筆者 白井清昭)

3.2.2 言語資源の利用事例の公開

言語情報処理ポータルでは、日本の言語資源・ツールに関する情報を収集し、カタログとしてウェブ公開している。これまで公開してきた情報は主に、言語資源の仕様、販売元、価格等といった必須の諸元情報であったが、加えてそれらの言語資源が過去、どのような研究事例において利用されているかの情報があれば、ユーザがこれから入手する言語資源を選定する上で有益な情報となるであろう。

そこで本年度、カタログ中の言語資源を利用している論文を自動抽出し、生成した論文リストを言語資源の利用事例としてカタログ情報に追加・公開した。論文テキストからの言語資源名抽出には、抽出対象であるカタログエントリの数が比較的限定されていることから、最も簡易な方法として、異表記を登録した辞書のパタンマッチを用いる方法を用いた。

(1) 予備調査～論文における言語資源名の記載形式調査

まず、論文電子ファイルに記載されている言語資源名がどの程度（再現率視点）自動抽出可能であるかの見当を付けるため、目視による調査を行った。

【調査方法】本分科会のメンバで論文のサーベイを行い、論文中で利用している言語資源（コーパス、辞書）を抽出した（本調査は後述する言語資源の利用目的調査と合わせて実施した。詳細は 3.3.1 節を参照のこと）。抽出した言語資源のうち、カタログに掲載されているもの（表 3.2.2-1）については、更にその言語資源名が記載されている箇所の表記やテキストの状態を調査した。なお、カタログには形態素解析器、パーザ等のツール類も含まれているが、本予備調査では抽出対象としない。

【調査対象】言語処理学会第 14 回年次大会予稿集(NLP2008 CD-ROM)に掲載された論文のうち、講演番号の下 1 桁が奇数の論文 155 件の PDF ファイルを調査対象とした。

表 3.2.2 - 1 言語資源・ツールカタログのエントリ 99 件[1]

毎日新聞 CD-ROM、日経新聞 CD-ROM、日経産業・金融・流通新聞 CD-ROM、読売新聞 CD-ROM (邦文記事)、読売新聞 CD-ROM (英文記事)、朝日新聞 CD-ROM、RWC テキストデータベース、EDR 日本語コーパス、EDR 英語コーパス、京都テキストコーパス、JEITA マルチモーダル対話コーパス、IREX 公開データ・ツール(最終版)、NTCIR テストコレクション、ATR 対話 DB、英文ビジネスレター文例大辞典 CD-ROM 版、勉誠データベース、データノベルズ、青空文庫、判例マスター、特許公報類 CD-ROM、講談社和英辞典、ZenBase CD-ROM、分類語彙表 増補改訂版 (CD-ROM 付き)、分類語彙表 増補改訂

[1] 言語資源が複数のデータセットから構成されており、カタログには階層的なエントリとして複数登録されているときは、上位エントリのみを抽出対象とした。例えば年毎のエントリ「毎日新聞 CD-ROM (1991 年)」などは上位エントリ「毎日新聞 CD-ROM」としてまとめて抽出した。

版 データベース（CD-ROM)、現代日本語名詞シソーラス、日本語語彙大系、BioCaster ontology、IPAL 辞書、EDR 辞書、EDR 日本語単語辞書、EDR 英語単語辞書、EDR 日英対訳辞書、EDR 英日対訳辞書、EDR 概念辞書、EDR 日本語共起辞書、EDR 英語共起辞書、EDR 専門用語辞書、古典対照語彙表、ICOT 形態素辞書、ライフサイエンス辞書、英語基本単語リスト、北大英語語彙表、EDICT、CICC マレーシア語基本語辞書、CICC インドネシア語基本語辞書、CICC 中国語基本語辞書、CICC タイ語基本語辞書、CICC 専門語辞書、MUST1: 日本語複合辞用例データベース v1.0、鳥バンク、中学校・高校教科書の語彙調査、テレビ放送の語彙調査 CD-ROM、語の共起関係データ、女性のことば・職場編、男性のことば・職場編、戦時中の話しことば –ラジオドラマ台本から–、日本語母語話者の雑談における「物語」の研究、科学技術 日英・英日コーパス辞典、Web 日本語 N グラム第 1 版、ATR 音声データベース、ATR 自然発話・言語 DB、ATR 多数話者音声 DB、ATR 多数話者音声 DB(模擬会話)、ATR 多数話者音声 DB(音素バランス文)、ATR 多数話者音声 DB(辞書データ)、日本音響学会研究用連続音声データベース、日本音響学会 新聞記事読み上げ音声コーパス(JNAS)、電総研道案内対話音声コーパス 1998、電総研音素バランス単語セット WD-I & II、電子協日本語共通音声データ--DAT 版--、連続音声(文科省 科研費試験研究)、方言音声データベース、重点領域研究 音声対話コーパス、RWCP-DB-SPEECH-96-I (RWC 音声対話データベース)、東北大 -- 松下単語音声データベース、早大白井研 100 地名単語データベース、京大堂下研 音素バランス単語セット、JUMAN、茶筌、すもも、Breakfast、和布蕪(MeCab)、KNP、MSLR パーザ、SAX、BUP、南瓜(CaboCha)、美寿満（ViJUMAN)、美茶（ViCha)、構文解析過程表示システム（VisIPS)、SUFARY、VisualMorphs、YamCha、TinySVM、日本語スペルチェッカー、Lexical Chainers、テキスト簡易要約器 Posum、DL-MT, Julius

【調査結果】全 155 論文中、37 の論文で、48 のカタログ掲載言語資源が利用されていた。48 件[2]の内訳を図 3.2.2 – 1 に示す。うち、単純な文字列一致により抽出可能なものは 8 件（17%）であった。一方、6 件（13%）は論文の PDF ファイルでフォントがアウトライン化されており、テキストが抽出不能となっている。

残りの 34 件（70%）は論文中の表記がカタログ記載名とは一致しなかった（以下「表記揺れ」と呼ぶ)。表記揺れの例を表 3.2.2 - 2 に示す。最も典型的なケースは、表中の網掛けで示したように、カタログ掲載名が「毎日新聞 CD-ROM」や「分類語彙表 増補改訂版 データベース(CD-ROM 付き)」のように仕様等の説明的な記述を含むため、論文中の表記と一致しないケースであり、表記揺れの約半分（15/34 件=44%）を占めている。

したがって、カタログ掲載名から冗長部分を取り除いた表記や、自明な呼称（「日経新聞」→「日本経済新聞」、「京都テキストコーパス」→「京大コーパス」など）を異表記として追加登録した辞書を用いることで、簡易なパタンマッチでもある程度の抽出率が得られることがわかった。

2 件数は、論文と言語資源の対応（＝利用事例）をカウントしたものである。したがって 1 つの論文中で 2 種類の言語資源が利用されているときは 2 件とカウントする。また、同一の論文中に同一の言語資源名が複数箇所に記載されていても、1 件とカウントする。

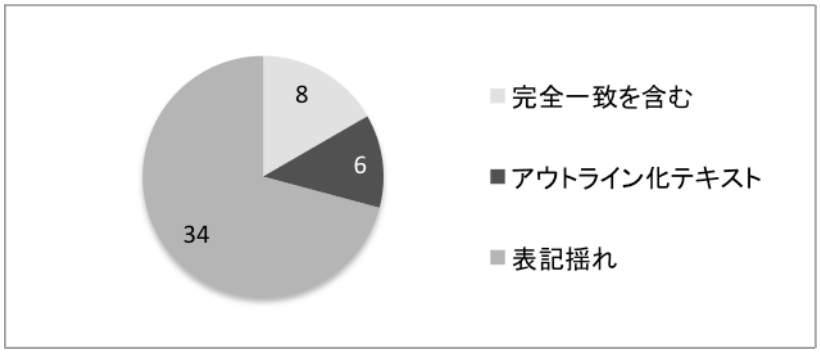


図 3.2.2-1 カタログ掲載名との完全一致による自動抽出可否

表 3.2.2-2　表記揺れの例

#	カタログ掲載名	論文中の記載箇所（抜粋）
1	毎日新聞 CD-ROM	実験として、毎日新聞の 97 年度 [5] の記事を先頭から…
2	毎日新聞 CD-ROM	下記の 3 つのデータに対して実験を行った。... (b) 毎日新聞 91 年から 98 年版を係り受け解析して得られた…
3	毎日新聞 CD-ROM	毎日新聞 1994 年から 2002 年までの 9 年分の記事より、…評価用コーパスを作成した。
4	毎日新聞 CD-ROM	学習データには、毎日新聞 1995 年 1 月 4 日～12 月 31 日の記事を用いた。
5	毎日新聞 CD-ROM	これまでに、毎日新聞 1995 年テキストから選定したテキストから構成される…
6	毎日新聞 CD-ROM	学習文書、評価文書は毎日新聞 [8] の記事を使用し、…
7	毎日新聞 CD-ROM	2006 年度の毎日新聞、読売新聞、日経新聞の三社の朝刊の新聞記事データを利用した。
8	毎日新聞 CD-ROM	新聞は、3 年分の WSJ コーパス（英語）と 5 年分の毎日新聞コーパス（日本語）を用いた。
9	分類語彙表 増補改訂版 データベース (CD-ROM)	…分類語彙データベース（国立国語研究所 2004）を活用して、…
10	分類語彙表 増補改訂版（CD-ROM 付き）	(3)に関しては分類語彙表[12]と日本語語彙大系[13]を利用する。
11	日本音響学会 新聞記事読み上げ音声コーパス(JNAS)	…新聞記事読み上げ音声コーパス（JNAS）より学習した…
12	日経新聞 CD-ROM	日本経済新聞 1998 年のデータ 100 件を用いた。
13	日経新聞 CD-ROM	… 1990 年から 2005 年 の日経新聞から獲得した。
14	日経新聞 CD-ROM	2006 年度の毎日新聞、読売新聞、日経新聞の三社の朝刊の新聞記事データを利用した。
15	日経新聞 CD-ROM	2006 年度の毎日新聞、読売新聞、日経新聞の三社の朝刊の新聞記事データを利用した。
16	特許公報類 CD-ROM	手順 (2) では、公開特許公報 1993～2005 年発行分（文書数は約 456 万件）を対象に検索を行う。
17	朝日新聞 CD-ROM	…利用許諾契約を結んでいる「朝日新聞記事データ集 2002（朝日新聞社 2003)」を活用した。
18	京都テキストコーパス	これを京大コーパスの 1 月 1～8 日分に対し適用した結果は表 1 に示した。
19	京都テキストコーパス	…京大コーパスの文を含むものであり、…

20	京都テキストコーパス	…京都大学コーパス 4 を利用する。
21	ライフサイエンス辞書	To supplement this we also used … and the last downloadable version of the Japanese-English Life Science Dictionary Project
22	RWC テキストデータベース	分析対象コーパスとして、『RWCP テキストコーパス』 (RWCP, 1998)、…を用いた。
23	RWC テキストデータベース	本研究では、「SENSEVAL-2 日本語辞書タスク」で配布された RWC コーパスを用いた。
24	NTCIR テストコレクション	…実験データとして、単質問文は NTCIR から、…を使用する。
25	NTCIR テストコレクション	… NTCIR6 QAC4[1] Formal Run のテストセット 100 問を用いて提案システムの性能評価を行なった。
26	NTCIR テストコレクション	本稿では、NTCIR3-WEB で構築されたテストコレクションを利用し、TSUBAKI で用いている各インデックスの効果について述べる。
27	NTCIR テストコレクション	本手法の効果を確かめるために、NTCIR-1 と NTCIR-3 PATENT を使用して評価実験を行った。
28	NTCIR テストコレクション	調査には NTCIR QAC1 [2]テストコレクション 200 問を用いて、…
29	IREX 公開データ・ツール(最終版)	IREX の CRL データを用いた実験により、Wikipeida による辞書、クラスタリングによる辞書ともに、認識精度を向上させることが分かった。
30	EDR 日本語コーパス	本実験では EDR コーパス [8] の一部 (187,022 文) をもとに、…
31	EDR 日本語コーパス	…EDR コーパス(Isahara, 2007)を利用した日本語の頻度辞書を作成した。
32	EDR 日英対訳辞書	To supplement this we also used the EDR Japanese-English lexicon (http://www2.nict.go.jp/r/r312/EDR/index.html)
33	EDR 日英対訳辞書	対訳辞書にはＥＤＲの日英対訳辞書、…を用いた。
34	ATR 対話 DB	実験に用いる対話データとして ATR の対話データベース[Mor94]を用いた。

(2) 辞書パタンマッチによる自動抽出精度評価

前述の予備調査結果に基づき、異表記を登録した言語資源名辞書を作成して論文とのパタンマッチを行い、簡易なパタンマッチでどの程度誤検出が生じるかを調べた（適合率観点)。

【評価対象】(1)の予備調査と同様、NLP2008 の 155 論文の PDF ファイルを用いた。PDF ファイルからのテキスト抽出には pdftotext[1)]を用いた。

【抽出対象】カタログの全エントリ（表 3.2.2 - 1）を抽出対象とする。すなわち、(1)の予備調査で抽出対象とした 48 件の言語資源利用事例（コーパス、辞書）に加え、71 件のツール利用事例（形態素解析器、パーザなど）を合わせた、合計 119 件の言語資源・ツール利用事例を抽出する。

【抽出用辞書】パタンマッチ用に辞書に追加した異表記の一覧を表 3.2.2-2 に示す。カタログエントリ 99 件中、30 件のエントリに対して合計 41 の異表記を追加した。異表記としては、(1)で述べたように「CD-ROM」等の冗長部分をカタログ掲載名から削除した表記や、「京大コーパス」などのよく用いられる名称を登録した。また、(1)の調査で得られた論文中の表記は、辞書エントリとして不自然にならない範囲（原則として文節を跨がない範囲としたが、一部例外あり）で異表記として登録した。

表 3.2.2-2　抽出処理用に登録した言語資源名の異表記

カタログ掲載名	異表記
毎日新聞 CD-ROM	毎日新聞
日経新聞 CD-ROM	日経新聞、日本経済新聞
読売新聞 CD-ROM（邦文記事）	読売新聞
朝日新聞 CD-ROM	朝日新聞記事データ集
RWC テキストデータベース	RWCP テキストコーパス、RWC コーパス
EDR 日本語コーパス	EDR コーパス
京都テキストコーパス	京大コーパス、京都大学コーパス
IREX 公開データ・ツール（最終版）	IREX
NTCIR テストコレクション	NTCIR
ATR 対話 DB	ATR の対話データベース
特許公報類 CD-ROM	特許公報
分類語彙表 増補改訂版 データベース（CD-ROM）	分類語彙表、分類語彙データベース
EDR 辞書	EDR 電子化辞書
EDR 日英対訳辞書	EDR Japanese-English lexicon, EDR の日英対訳辞書
ライフサイエンス辞書	Japanese-English Life Science Dictionary
ATR 音声データベース	ATR 音声 DB
ATR 自然発話・言語 DB	ATR 自然発話・言語データベース
ATR 多数話者音声 DB	ATR 多数話者音声データベース
日本音響学会 新聞記事読み上げ音声コーパス（JNAS）	新聞記事読み上げ音声コーパス、JNAS
電総研道案内対話音声コーパス 1998	電総研道案内対話音声コーパス
電子協日本語共通音声データ--DAT 版--	電子協日本語共通音声データ
RWCP-DB-SPEECH-96-I（RWC 音声対話データベース）	RWC 音声対話データベース
東北大 -- 松下単語音声データベース	東北大・松下単語音声データベース、東北大-松下単語音声データベース
茶筌	ChaSen
和布蕪（MeCab）	和布蕪、MeCab
南瓜（CaboCha）	南瓜、CaboCha
美寿満（ViJUMAN）	美寿満、ViJUMAN
美茶（ViCha）	美茶、ViCha
構文解析過程表示システム（VisIPS）	VisIPS
テキスト簡易要約器 Posum	Posum

【抽出結果】本辞書による抽出精度は、再現率 85%（101/119）、適合率 86%（101/117）であった。誤りの可能性がある旨明記すれば、公開可能な精度が得られていると判断した。なお、再現率に関しては、抽出に失敗した 18 件中の 13 件は PDF ファイルのテキストがアウトライン化されていることが原因なため、OCR 等が必要である。また、適合率に関しては、誤検出の原因は、関連研究などとして言及している箇所にマッチしてしまったケースが殆どであった。簡易なパタンマッチによる識別は難しいと思われる。誤検出箇所の一覧を表 3.2.2-3 に示す。

表 3.2.2-3　論文中の誤検出箇所

#	カタログ掲載名	誤検出箇所
1	EDR 辞書	...[8] 日本電子化辞書研究所:【EDR 電子化辞書】仕様説明書(1993)....
2	JUMAN	...析システムを利用することが不可欠となり、【JUMAN】や ChaSen などのフリーソフトが...
3	京都テキストコーパス	...その 759 文には、【京大コーパス】5 と同じ基準でタグ付けを行っており、これ...
4	IREX 公開データ・ツール（最終版）	...CaboCha による固有表現解析では、【IREX】(Information Retriev...
5	IREX 公開データ・ツール（最終版）	...索質問とコンテキストとの相関固有表現は【IREX】の定義に従えば、人名、地名、組織名、固…
6	NTCIR テストコレクション	...パスとして、MPQA [8]や【NTCIR】－6 における意見分析タスクのコーパス … ...【NTCIR】－6 のタグ付与の仕様は当該文の意見性の…
7	分類語彙表 増補改訂版 データベース（CD-ROM）	...クラスタリングを行っている。シソーラスは【分類語彙表】を用いており、類似度は分類項目の木構造に...
8	EDICT	...ement of the JMdict/【EDICT】 Japanese-English di...
9	京都テキストコーパス	...は、統一的なタグ付け基準を既に定めている【京大コーパス】作成基準 [4] に準拠し、通常の係り受… ...3 タグ付け基準の設計法律文には、【京大コーパス】におけるタグ付け基準をそのまま適用できな… ...態素情報や構文情報は変化しないことから、【京大コーパス】においては括弧書きをあらかじめ除去してタ… ...形態素の品詞、文節の区切り【京大コーパス】においては、等位接続詞は、(4) のよう… ...C 2. A B CP PPP 図 4: 【京大コーパス】基準に基づく並列構造表現 (i) A B ... ...C → (A 並びに (B 及び C))【京大コーパス】においては、並列関係を記号 P のみを用... ...「その他」【京大コーパス】では、「~その他」については「~など」と... ...「その他の」【京大コーパス】では、「~その他の」は、それに後続する適...
10	JUMAN	...態素辞書として、IPADIC[1] や 【JUMAN】 辞書などがある．本稿では、法律文コーパ...
11	KNP	...(ii)自動化された係り受け解析は【KNP】や Cabocha という形で実装され...
12	IREX 公開データ・ツール(最終版)	...よく用いられる MUC の 7 種類や【IREX】の 8 種類の固有表現（人名、地名、組織...
13	京都テキストコーパス	...その後、【京大コーパス】(Kurohashi and Nagao...
14	分類語彙表 増補改訂版 データベース（CD-ROM）	...イル NTT 日本語語彙大系、EDR、【分類語彙表】 概念間関係、単語→概念 単語 (le...
15	EDR 辞書	...れらの修飾要素に対する意味役割の設定は【EDR 電子化辞書】で定義されている．そこで EDR の「事… ...[10] 日本電子化辞書研究所：【EDR 電子化辞書】使用説明書（第 2 版）(1995)…
16	京都テキストコーパス	...ab1 を使用した．この mecab は、【京大コーパス】をはじめ、対象と同一ドメインの対話データ...

(3) 公開データの作成

2005 年～2008 年の 4 年分の言語処理学会年次大会発表論文集を対象に（2）と同様の抽出処理を行うことで、カタログエントリの利用事例論文リストを作成し、言語情報処理ポータル上で公開した（図 3.2.2-2）。利用事例の公開ページは、カタログの各エントリからリンクを辿って閲覧できるようになっている。利用事例として掲載した各論文には、その根拠として、言語資源名が記載されている箇所周辺のテキストを表示するようにした。

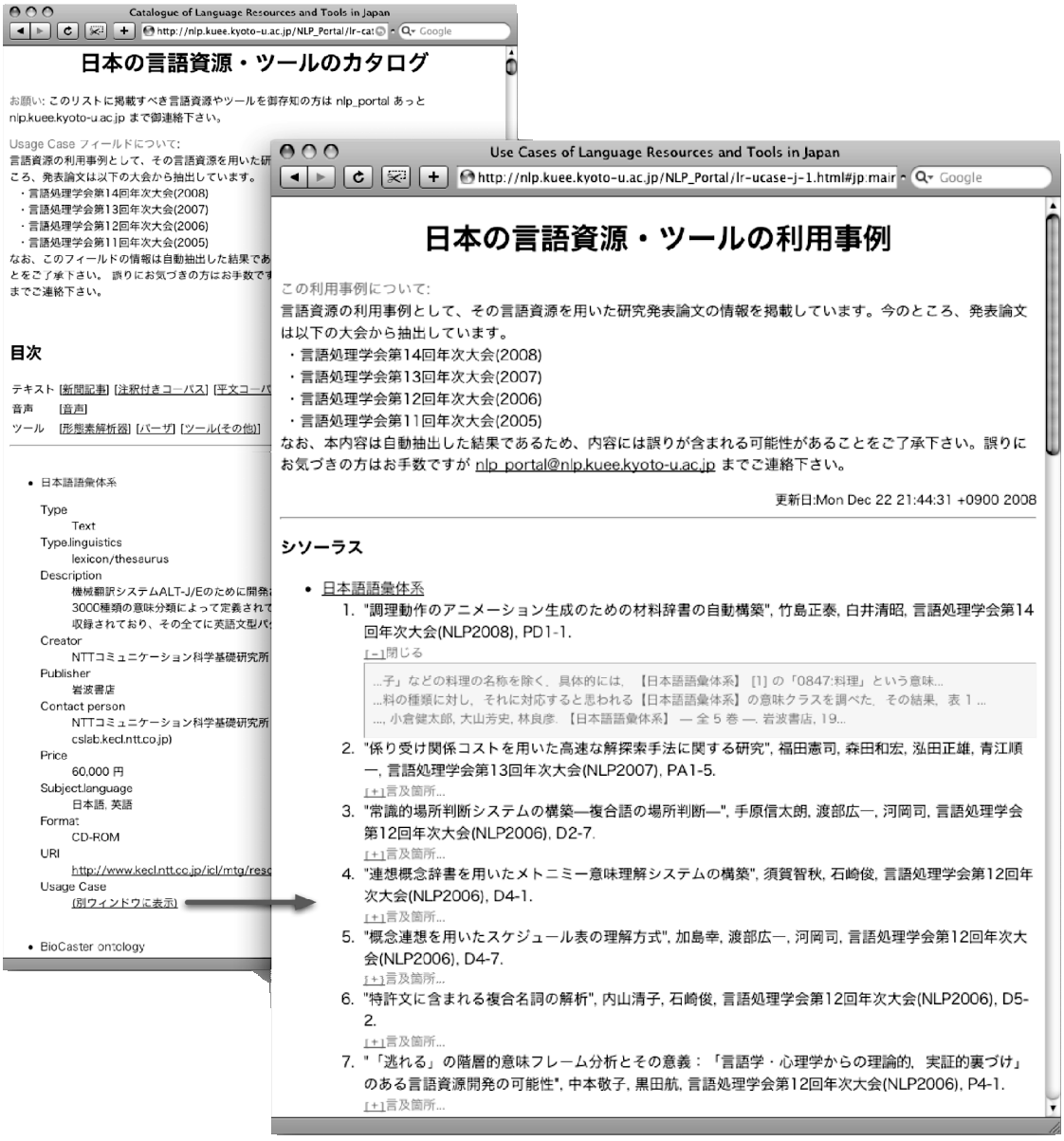


図 3.2.2 - 2　言語資源・ツール利用事例の公開ページ

(4) まとめ

言語情報処理ポータルにおける言語資源カタログの情報充実化の一環として、カタログ中の言語資源を利用している論文を言語処理学会年次大会（2005 年～2008 年）より自動抽出し、利用事例として公開した。本利用事例が言語処理研究者にとって有益な情報となることを願う。

今回公開した情報は個々の言語資源の利用事例（論文）であったが、今後は、言語資源の利用数（論文数）比較や、利用分野（例えばタイトルに含まれる語 [2)]）などのように俯瞰的な情報の公開も検討したい。

参考文献

1) http://www.foolabs.com/xpdf/

2) 村田真樹、一井康二、馬青、白土保、金丸敏幸、井佐原均: “過去 10 年間の言語処理学会論文誌・年次大会発表における研究動向調査”、言語処理学会第 11 回年次大会、pp. 77-80 (2005).

(執筆者 谷垣宏一)

3.2.3 アノテーションツールの調査

言語資源ポータルサイトでは、入手可能な各種言語資源についての情報を広く収集し、掲載している。実際の研究においては、これらの言語資源だけでなく、研究者自身が言語資源を独自に収集・構築するというケースも多くなっている。そこで、注釈付きコーパスを作成する際に活用できるアノテーションツールについて調査した。

調査は、日本語に対応しているもの、および、現在も入手可能なものを掲載すると言う方針で行った。調査方法は、言語処理学会全国大会予稿集でのアノテーションツール開発に関する発表論文のサーベイや、Web を利用した情報収集などである。調査の結果、前述の条件を満たすアノテーションツールが 2 件発見されたので、その情報をまとめ、本ポータルサイトのツールカタログ「ツール（その他）」に掲載した。今回掲載したアノテーションツールの情報は以下の通りである。

Tagrin

【特徴】任意のタグ体系でテキストにタグを付与可能なツール。タグの重複・交差も許される。タグ付きテキストは独自形式 (.tgr) の他、SGML 形式でも import/export 可能。

【開発者】高橋哲朗氏

【対応 OS】windows,linux(Tcl/Tk)

【URL】http://kagonma.org/tagrin/

【参考文献】「アノテーションツール”Tagrin”の紹介」
言語処理学会第 12 回年次大会予稿集 p.228-231

FuuTag

【特徴】初期設定では関根拡張固有表現体系に基づくタグ付与が可能なツール。タグの種類は変更可能。タグ付きテキストは SGML 形式。

【開発者】 関根聡氏

【対応 OS】windows,uinux

【URL】 http://nlp.cs.nyu.edu/ene/

今後も本分科会では、言語資源の普及・発展促進活動の一環として、言語資源構築に活用できるアノテーションツールの情報を継続的に収集し、適宜本ポータルサイトで公開したいと考えている。

（執筆者 齋藤邦子）

3.2.4 形態素解析済みコーパスの公開

本項では、日本語テキストを形態素解析済みコーパスとして公開する活動について報告する。

（1）目的

現在、多くの日本語テキストコーパスが公開され、言語研究に利用されている。これらのコーパスの利用価値を高め、幅広い研究分野で活用してもらうため、コーパスに言語情報を付加して公開することが本活動の主旨である。付加する言語情報は様々考えられるが、本年度の活動では、コーパスに形態素情報を付加し、コーパスの検索や語句分析などを容易にすることを狙いとした。テキストコーパスに形態素情報を付加することで、表現形の違いに捉われず語を検索することや、品詞で語をフィルタリングすること、語の出現の統計をとることなどが可能となる。

（2）コーパスの形式

形態素解析済みコーパスは、日本語テキストを形態素解析して品詞等の形態素情報を付加した XML 形式のデータである。形態素解析済みコーパスの例を図 3.2.4-1 に示す。コーパス公開の際は、利用者の利便性を高めるため、国立国語研究所により公開されている全文検索システム『ひまわり』用の検索インデックスおよび設定ファイルを同梱する。また、形態素解析処理には、２種類の形態素解析器茶筌および、JUMAN を用いることとする。

茶筌：http://chasen.aist-nara.ac.jp/hiki/ChaSen/

JUMAN：http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html

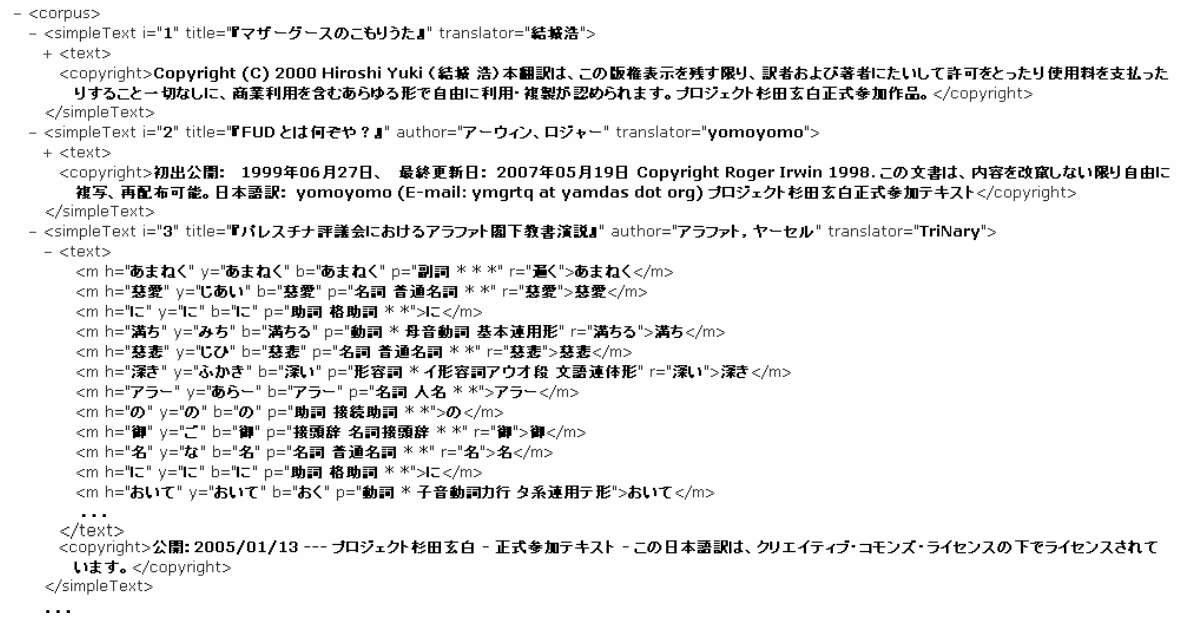


図 3.2.4-1　形態素解析済みコーパス例

(3) コーパスの整備

形態素解析済みコーパスのデータソースには、再利用に著作権上問題のない、オンラインで公開されている日本語テキストコーパスを用いる。対象テキストコーパスとして、下記の２つが候補に挙がった。

1)「青空文庫」(http://www.aozora.gr.jp/)：著作権上問題のない文献を無料で公開するインターネット図書館
2)「プロジェクト杉田玄白」(http://www.genpaku.org/)：著作権上問題のない英語の文献を、日本語に翻訳して無料で公開するプロジェクト。

「青空文庫」に関しては、XHTML 形式で公開されているテキストを形態素解析し『ひまわり』用の検索インデックスを作成するツールが、国立国語研究所により公開されている。よって、本活動では、「プロジェクト杉田玄白」のテキストを処理対象とすることとした。

「プロジェクト杉田玄白」で公開されているテキストには、商用を含めて再利用可能なものと、再利用に制限のあるものがある。本活動では、商用を含めて再利用可能なテキストのみを対象とし、173 件のテキストデータをダウンロードした。さらに、作品によっては利用の際に訳者の著作権表示を義務付けているものもあるため、全テキストの著作権表示を収録テキスト一覧に掲載することとした。このため、テキストデータから著作権に関する記述部分を抜き出す（タグ付けする）作業を人手で行った。なお、この著作権記述は形態素解析済みコーパスにも含まれる。

（4）コーパスの利用方法

本活動では、形態素解析器に茶筌および JUMAN を用い、それぞれ茶筌版、JUMAN 版として２種類の形態素解析済みコーパスを公開することとした。公開するコーパスには、全文検索エンジン『ひまわり』用の検索インデックスを同梱する。

全文検索エンジン『ひまわり』では、形態素情報を持った検索インデックスを用い、下記の４種類の検索が可能である。

1)「単語-表記」モード：検索文字列と表記が一致する単語を検索。

2)「単語-基本形」モード：検索文字列と基本形が一致する単語を検索。 全ての活用形の単語を一度に検索することが可能。

3)「単語-代表表記」モード（JUMAN 版のみ）：検索文字列が代表表記（JUMAN が出力する単語の標準的な表記）と一致する単語を検索。異表記の単語を一度に検索することが可能。

4)「文字列」モード：検索文字列と一致する文字列を全文検索

これらのうち、1)～3)は形態素情報を持つことによって実現される機能である。『ひまわり』の検索結果では、検索文字列に一致した語の、前後の記述（指定文字数長）や、品詞、読み、基本形、表記、書誌情報、などが出力される。また、検索文字列について、品詞、読み、基本形のいずれかを指定して検索結果をフィルタリングすることが可能である。『ひまわり』での検索結果例を、図 3.2.4-2 に示す。

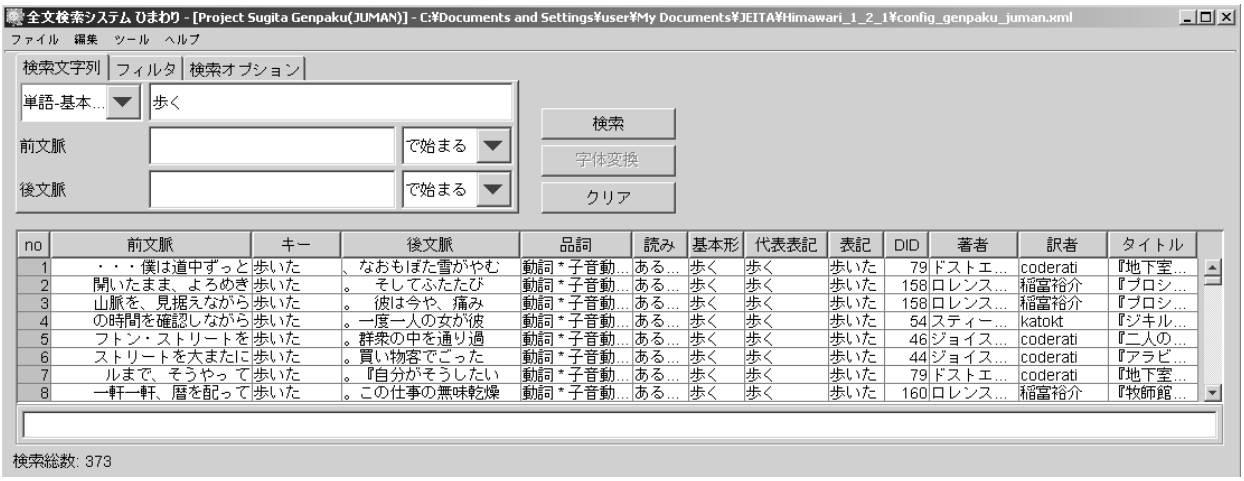


図 3.2.4-2 『ひまわり』検索結果画面（JUMAN 版）

(5) まとめと今後の課題

本年度の活動では、「プロジェクト杉田玄白」で公開されているテキストコーパスに形態素情報を付与し、形態素解析済みコーパスとして公開した。コーパスには、全文検索システム『ひまわり』用の検索インデックスを同梱した。

今後の活動では、「青空文庫」のテキストデータなどを対象にコーパスの量を増やすことや、付与する言語情報を充実させることが課題である。

(執筆者 寺本やえみ)

3.3 言語資源の公開・利用に関する調査報告

言語資源者の保有者が利用者に言語資源を提供する際には、言語資源がどのような目的でどのように使われるかに関して、保有者と利用者とである程度の合意を形成する必要がある。また、両者の間で契約を締結する必要があることも多い。このような背景から、言語資源を公開する際に、言語資源保有者と利用者間の合意形成や契約締結を円滑に進めるための参考情報を提供することを目的とし、言語資源の公開や利用に関する現状を調査した。3.3.1「言語資源の利用事例の実態調査」では、研究論文を対象とした言語資源の利用実態の調査について報告する。3.3.2「言語資源の利用事例の分類」では、言語資源保有者が言語資源を提供する際に自分の言語資源がどのように使われるのかを懸念する場合が多いことから、言語資源の使用目的や使用形態(言語資源をそのまま使うのか、加工して使うのか、など)といった観点から、言語資源の利用事例を分類した調査結果を報告する。3.3.3「言語資源を利用する枠組の調査」では、主に利用者側の観点から、言語資源を利用する際に締結する契約の種別に関する調査について報告する。

3.3.1 言語資源の利用事例の実態調査

近年、自然言語処理技術の研究開発において、ウェブ上のデータやコーパス、シソーラス、辞書といった言語資源の利用が進み、大規模データに基づく音声・言語処理技術を中心に大きな成果を挙げている。分科会では、どのような言語資源を使ってどのような研究開発が行われているかを知ることは意味があると考え、実際の研究における言語資源の利用状況を調査した。

調査対象は、言語処理学会第 14 回年次大会 (NLP2008) 予稿集に収録された論文 159 件である[3]。論文に記載された言語資源を人手により抽出し、集計した。結果を図 3.3.1-1 に示す。全体の 88%にあたる 140 件の論文で言語資源が利用されている[4]。研究領域によるばらつきもなく、自然言語処理

[3] 特定の分野に偏らないよう、予稿集に収録の全論文のうち、論文 ID の下 1 桁が奇数のものを調査対象とした。
[4] 論文中で複数の言語資源を利用している場合も 1 件とした。

のあらゆる分野の研究開発において、言語資源が利用されていることが分かった。

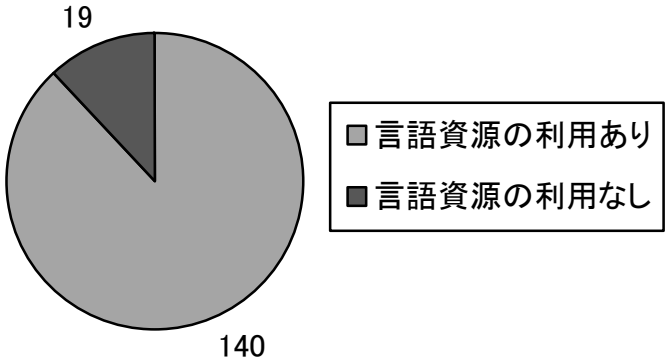


図 3.3.1-1　言語処理学会論文における言語資源の利用状況

次に、どのような言語資源が利用されているかについて調査した。結果を表 3.3.3-1 に示す。利用されている言語資源の種類としては、コーパスが 87 件と最も多く、ついでウェブデータ、辞書・シソーラス、新聞記事[5]と続いている。

コーパス全体では、NTCIR や MUC といったデータコレクションの利用が目立ったほか、EDR コーパス、京都コーパス、Web 5 億文コーパス、青空文庫などの利用があった。コーパスの種類も、話し言葉/書き言葉、生テキスト/タグつきテキスト、単言語/対訳など様々で、入手可能なコーパスの種類が増え、研究目的に適したデータの利用が進んでいる状況が分かる。

ウェブデータとしては、Wikipedia やアマゾンのレビュー記事の利用が多かったが、それ以外にもブログ、レビュー、ニュース記事など多岐に渡った。ウェブデータは公開されているデータのほか、クローリングによって大量のデータが比較的手軽に入手できることもあり、独自に収集、加工している例も見られた。

辞書･シソーラスでは、日本語語彙大系、分類語彙表、WordNet、EDR 辞書の利用が多かった。辞書・シソーラス単独での利用は少なく、他の言語資源とあわせて、語彙の抽出・分類や、結果の判定に利用している。対訳コーパスは日英対訳がほとんどで、中でも NICT 日英対訳コーパスの利用が多数を占めた。その他のテキストデータは、研究目的にあわせて独自に作成したデータで、特に目立った傾向はなかった。

[5] 複数の項目で重複してカウントされる言語資源もある。

表 3.3.1-1　言語処理学会論文における言語資源の利用調査

言語資源分類	論文数
コーパス	87 件
ウェブデータ	56 件
辞書・シソーラス	32 件
新聞記事	22 件
対話･話し言葉コーパス	12 件
対訳コーパス	10 件
その他テキストデータ	14 件

以上のように、言語処理学会年次大会の論文を元に、自然言語処理の研究で利用されている言語資源の状況を調査した。その結果、多くの論文で様々な言語資源が利用されていることが分かった。分科会では、この調査結果を元に、言語資源とそれを利用した研究（利用事例）を対応付け言語資源ポータルで公開している（「3.2.2 言語資源の利用事例の公開」を参照のこと)。

（執筆者 下畑さより）

3.3.2　言語資源の利用事例の分類

(1)　はじめに

「3.3.1 言語資源の利用事例の実態調査」で示す通り、近年、自然言語処理の分野では大規模な言語資源を用いた手法が大きな流れとなっており、言語資源に対する需要が高まっている。こういった需要を背景に、言語資源の商業的な流通も徐々に拡大し、新聞、書籍、辞典等は言語資源データベースとして販売され、学術研究や商用として利用可能な状態となっている。しかし、実用的かつビジネス展開可能な自然言語処理技術の開発には、開発対象に応じた言語資源が商用利用可能であることが望まれている。JEITA 言語資源分科会では非営利団体の言語資源協会（GSK）と協力し、より多種多様な言語資源の流通促進に向け活動しているが、言語資源保有者が言語資源の提供にあたり自分の言語資源がどのように使われるのか懸念する場合が多い。懸念を解消するためには言語資源の利用の仕方を明確にし、言語資源の提供の意義を理解してもらう必要がある。本章では、言語資源の使用目的や使用形態といった観点から言語資源の利用事例を分類した結果を報告する。

(2) 言語資源の利用事例の分類

① 調査方法

学術研究[6]と商用[7]（利用種別）の事例を対象に言語資源がどのような自然言語処理の開発で利用されているか調査した。さらに、各自然言語処理の開発でどのように言語資源が利用されているか使用形態（言語資源をそのまま使うのか、加工して使うのかなど）を調査した。

② 言語資源と自然言語処理の開発

言語資源毎に利用種別と利用技術分野を調査し、言語資源の種別により 6 つに分類した。

✓ 調査項目

調査項目とその分類を示す。

・ 言語資源の種別

－コーパス

対訳コーパスを含むコーパス全般。

－ウェブデータ

Web、メール、携帯で扱うテキスト全般を対象とした。

－辞書・シソーラス

－新聞記事

言語資源データベースとして販売されている言語資源を対象とした。Web やメールによるニュースは「ウェブデータ」に分類。

－対話・話し言葉コーパス

音声のみ、音声とテキストを含む言語資源を対象とした。

－その他テキストデータ

議事録、論文、書籍等。他の分類に含まれない言語資源を対象とした。

・ 利用種別

－学術研究

－商用

・ 利用技術分野

－言語学的解析

文体分析、特徴因子分析、話し言葉らしさ分析など。

---

6 「3.2.2 言語資源の利用事例の公開」の調査結果を基に調査した。

7 Web 上の自然言語処理を利用した製品のホームページを対象に調査した。

－形態素解析

－構文解析

－意味解析

－機械翻訳

－情報検索

－情報抽出

－質問応答

－言語資源開発

　辞書、コーパス開発など。

－文の生成

－校正

－かな漢字変換

－音声認識

－音声合成

－対話システム

－辞書（利用種別が商用の場合のみ適応）

✓　調査結果

言語資源の種別毎に利用事例を示す。

・　コーパス

表 3.3.2‐1　コーパスの利用事例

言語資源	利用種別	利用技術分野
RWC テキストデータベース	学術研究	形態素解析
		情報抽出
EDR 日本語コーパス	学術研究	機械翻訳
		質問応答
		言語資源の開発
		音声認識
		音声合成
京都大学テキストコーパス	学術研究	言語学的解析
		形態素解析

		機械翻訳
NAIST テキストコーパス	学術研究	言語学的解析
		構文解析
田中コーパス	学術研究	構文解析
Web 5 億文コーパス	学術研究	言語学的解析
		情報抽出
		言語資源の開発
現代日本語書き言葉均衡コーパス（BCCWJ）	学術研究	言語学的解析
		形態素解析
		言語資源の開発
太陽コーパス	学術研究	形態素解析
KY コーパス	学術研究	言語資源の開発
The British National Corpus（BNC）	学術研究	言語学的解析
		機械翻訳
MICASE	学術研究	言語学的解析
CREST コーパス	学術研究	音声合成
Kielipankki（フィンランド語コーパス）	学術研究	言語学的解析
Wall Street Journalコーパス(WSJコーパス)	学術研究	言語学的解析
		構文解析
日英対訳文対応付けデータ（EJTAD)	学術研究	機械翻訳
Fry による日英対訳データ	学術研究	機械翻訳
JENNAD（自動対応付けされた日英対訳コーパス）	学術研究	機械翻訳
NICT 日英対訳コーパス	学術研究	機械翻訳
ロイター英日対訳コーパス	学術研究	機械翻訳
IREX CRL データ	学術研究	形態素解析
MUC-7（照応タグ）	学術研究	照応解析
NTCIR CLQA	学術研究	質問応答
NTCIR　QAC 評価データ	学術研究	質問応答
NTCIR-1	学術研究	情報検索
NTCIR-3PATENT	学術研究	情報検索
NTCIR3-WEB	学術研究	情報検索

ACE（照応タグ）	学術研究	照応解析
GENIA コーパス（照応タグ）	学術研究	照応解析
Gideon Mann 06	学術研究	情報抽出
JST 日英抄録コーパス	学術研究	機械翻訳
JST 科学技術用語日英対訳辞書	学術研究	機械翻訳

・ ウェブデータ

表 3.3.2-2　ウェブデータの利用事例

言語資源の例	利用種別	利用技術分野
Web		
tuoitre（ベトナム語オンライン新聞サイト）	学術研究	機械翻訳
Web 新聞記事データ	学術研究	情報抽出
Web ニュース	学術研究	意味解析
		音声合成
Web テキスト（Google ニュース）	学術研究	情報抽出
Web テキスト（goo ニュース）	学術研究	情報抽出
Web テキスト（新聞記事）	学術研究	情報抽出
Amazon 商品レビュー	学術研究	情報抽出
IRC（InternetRelayChat）のチャットログ	学術研究	対話システム
Web から収集したレシピ文書	学術研究	形態素解析
Web5 億文コーパス	学術研究	言語資源の開発
Web コーパス約 1 億文	学術研究	形態素解析
Web サイトの商品レビュー	学術研究	情報抽出
Web テキスト	学術研究	構文解析
		情報抽出
Web テキスト（Amazon レビュー文）	学術研究	文の生成
Web テキスト（at cosme）	学術研究	言語学的解析
Web テキスト（Livedoor レビュー文）	学術研究	文の生成
Web テキスト（Web データ）	学術研究	言語学的解析
Web テキスト（Web ページ）	学術研究	情報抽出
	商用	機械翻訳
Web テキスト（価格.com レビュー記事）	学術研究	言語学的解析

Web 上の質問サイトの質問文	学術研究	質問応答
Yahoo!ショッピング商品データ	商用	情報抽出
Yahoo!検索	学術研究	情報抽出
Yahoo!知恵袋	学術研究	情報抽出
		質問応答
クイズミリオネアの factoid 型質問 100 問（Web）	学術研究	質問応答
サンケイウェブ	学術研究	情報抽出
ブログページ（Yahoo!検索 API）	学術研究	情報検索
国立がんセンターの Web ページ	学術研究	形態素解析
メール		
日経ニュースメール	学術研究	情報抽出
Vine Users ML	学術研究	言語学的解析
		質問応答
携帯		
携帯ブログサイト（aimew）	学術研究	情報抽出

・ 辞書・シソーラス

表 3.3.2-3　辞書の利用事例

言語資源の例	利用種別	利用技術分野
EDR 電子化辞書	学術研究	意味解析
EDR 日英対訳辞書	学術研究	機械翻訳
		言語資源の開発
IPA 品詞体系日本語辞書"IPADIC"	学術研究	文の生成
JMDict（日→英／仏／独）	学術研究	言語資源の開発
JST 科学技術用語日英対訳辞書	学術研究	機械翻訳
MEDLINE	学術研究	情報抽出
Thai-English dictionary（MMT dictionary CICC）	学術研究	言語資源の開発
UniDic	学術研究	形態素解析
Wikipedia	学術研究	言語学的解析
		形態素解析

		機械翻訳
		情報抽出
		言語資源の開発
医学用語辞書	学術研究	情報抽出
岩波国語辞典（形態素タグ、語義タグ）	学術研究	言語学的解析
最新解剖学用語集	学術研究	情報抽出
ライフサイエンス辞書(日→英)	学術研究	言語資源の開発
学研国語大辞典第２版	学術研究	言語資源の開発
情報処理用語辞典	学術研究	情報抽出
DHC 英会話とっさのひとこと辞典	商用	辞書
NTT 出版　パソコン用語知ったか辞典 2002	商用	辞書
Oxford University Press　オックスフォードコンサイス類語辞典	商用	辞書
Oxford University Press　オックスフォード現代英英辞典 第六版	商用	辞書
PHP　経済用語事典	商用	辞書
学研　カタカナ新語実用辞典	商用	辞書
学研　英文手紙用例辞典	商用	辞書
学研　四字熟語早引き辞典	商用	辞書
学研　全国方言一覧辞典	商用	辞書
学研　暮らしのことわざ早引き辞典	商用	辞書
学研監修　漢字字典	商用	辞書
学研監修　世界の料理メニュー辞典	商用	辞書
岩波書店 岩波国語辞典 第六版岩波書店　広辞苑＊第五版	商用	辞書
岩波書店　逆引き広辞苑	商用	辞書
研究社　コンピュータ英語情報辞典	商用	辞書
研究社 リーダーズ英和辞典 第二版	商用	辞書
研究社　新英和中辞典 第六版	商用	辞書
研究社　新和英中辞典 第四版	商用	辞書
研究社監修　カタカナ発音英単語検索辞典	商用	辞書

朝日新聞社監修　知恵蔵 2002	商用	辞書

表 3.3.2-4　シソーラスの利用事例

言語資源	利用種別	利用技術分野
日本語語彙大系	学術研究	形態素解析
		情報抽出
		言語資源の開発
分類語彙表	学術研究	言語学的解析
		言語資源の開発
WordNet	学術研究	情報抽出
	学術研究	言語資源の開発
EuroWordNet	学術研究	言語資源の開発
医学用語シソーラス	学術研究	情報抽出

・　新聞記事

表 3.3.2-5　新聞記事の利用事例

言語資源	利用種別	利用技術分野
毎日新聞	学術研究	言語学的解析
		情報抽出
		構文解析
朝日新聞	学術研究	言語学的解析
		情報抽出
日本経済新聞	学術研究	情報抽出
読売新聞	学術研究	情報抽出
The Daily Yomiuri	学術研究	機械翻訳
APW	学術研究	情報抽出
The New York Times（NYT）	学術研究	情報抽出
XINHA	学術研究	情報抽出
新聞記事	学術研究	情報抽出
	学術研究	音声合成
	商用	機械翻訳

・ 対話・話し言葉コーパス

表 3.3.2-6 対話・話し言葉コーパス

言語資源	利用種別	利用技術分野
ESP_C コーパス（JST/ATR Expressive Speech Corpora サブセット）	学術研究	対話システム
ATR 音声データベース	学術研究	音声認識
		対話システム
新聞記事読み上げ音声コーパス（JNAS）	学術研究	音声合成
日本語話し言葉コーパス（CSJ）	学術研究	言語学的解析
		形態素解析
		情報抽出
		言語資源の開発
		音声認識
		対話システム
三修社監修　ペラペラ イタリア旅行会話	商用	辞書
三修社監修　ペラペラ スペイン旅行会話	商用	辞書
三修社監修　ペラペラ ドイツ旅行会話	商用	辞書
三修社監修　ペラペラ フランス旅行会話	商用	辞書
三修社監修　ペラペラ 英米旅行会話	商用	辞書

・ その他テキストデータ

表 3.3.2-7 その他テキストデータの利用事例

言語資源	利用種別	利用技術分野
議事録		
EU 議事録	商用	機械翻訳
国連議事録	商用	機械翻訳
法律		
国民年金法	学術研究	情報抽出
所得税法	学術研究	情報抽出
法令翻訳データ集	学術研究	言語資源の開発
大日本帝国憲法	学術研究	形態素解析

民法	学術研究	形態素解析
特許		
日英特許データ	学術研究	機械翻訳
特許	商用	機械翻訳
特許明細文	学術研究	構文解析
公開特許公報 1993-2005 年	学術研究	形態素解析
論文		
ACL06	学術研究	情報抽出
ECDL06	学術研究	情報抽出
WEB03	学術研究	情報抽出
生物学論文	学術研究	構文解析
文書		
放射線読影レポート	学術研究	情報抽出
論説文	学術研究	情報抽出
書籍		
青空文庫	学術研究	言語学的解析
		形態素解析
		情報抽出
教育勅語	学術研究	形態素解析
文明論之概略	学術研究	形態素解析
学研「もっとうまい e メールの書き方」	商用	辞書
学研「世界の名言名句」	商用	辞書
講談社インターナショナル「英語で話す「日本」Q&A」	商用	辞書
株式会社アルク「どんな時どう使う日本語表現文型 500 中上級」	学術研究	言語学的解析
その他		
NHK「あすを読む」	学術研究	文の生成
NHK「きょうの健康クローズドキャプション」	学術研究	言語学的解析
NHK 多言語ニュースコーパス	学術研究	機械翻訳
旅行会話文（独自）	商用	機械翻訳

クレーム文	学術研究	情報抽出
小論文課題(原文と要約、その 5 段階評価)	学術研究	情報抽出
単語親密度リスト(英:MRC、日:天野)	学術研究	言語学的解析
英辞郎	学術研究	機械翻訳

③ 言語資源の利用形態

各自然言語処理の開発でどのように言語資源が利用されているか 7 項目について調査した。

✓ 調査項目

調査項目と記述例を示す。記述例は利用技術分野が形態素解析の場合を示す。

- 目的

  言語資源の利用目的。

  例）形態素解析システムを作る

- 使用情報

  言語資源から抽出し使用した情報。

  例）単語、単語の並び

- 成果物

  使用情報を加工して生成されるもの。

  例）単語辞書、単語の連接情報

- 利用者に見えるもの

  言語資源の使用情報が利用者に見える場合の見え方。

  例）入力文の解析結果

- 元言語資源の再現性（再現可能なもの）

  低、高、無の 3 段階で付与。

  例）低（単語）

- 製品形態

  例）開発用コンポーネント（エンジン）、評価システム、開発用リソース（辞書）

✓ 調査結果

自然言語処理技術の開発（技術分野）毎に利用事例を示す。

表 3.3.2-8　自然言語処理の開発と言語資源の利用事例

技術分野	目的	言語資源の使用情報	成果物	利用者に見えるもの	使用した言語資源の再現性	製品形態
言語学的解析	・文体分析 ・特徴因子分析 ・話し言葉らしさ分析 ・照応表現分析 ・単語親密度分析 ・機能表現言い換え ・省略補完（文末表現） ・関係抽出 ・機能表現抽出 ・含意関係獲得 ・因果関係獲得 ・言語表現分析	・辞書の属性情報 ・分類情報 ・文書	・文体分析 ・特徴因子分析 ・話し言葉らしさ分析 ・照応表現分析 ・単語親密度分析 ・機能表現言い換え ・省略補完（文末表現） ・関係抽出 ・機能表現抽出 ・含意関係獲得 ・因果関係獲得 ・言語表現分析	無	無	無
形態素解析	形態素解析システムを作る	・単語 ・単語の並び	・単語辞書 ・単語の連接情報	入力文の解析結果	低（単語）	・開発用コンポーネント（エンジン） ・評価システム ・開発用リソース（辞書）
	形態素解析システムの精度を向上させる	・単語 ・単語の並び ・分野情報	・単語辞書 ・単語の連接情報 ・特定分野の単語出現傾向	入力文の解析結果	低（単語）	
	同表記異音語の解消	・文章 ・語種情報	・評価用データ ・語種情報	無	無	
	単語の収集または辞書の作成（新出単語収集/分野/専門用語/固有表現）	・単語 ・単語の並び ・分野情報	・単語辞書 ・単語の連接情報 ・特定分野の単語出現傾向	見出し語 入力文の解析結果	低（単語）	
	言語資源の開発	・形態論情報	無	無	無	
構文解析	構文解析システムを作る	・文 ・構文情報	・文法規則 ・ツリーバンク	入力文の解析結果	無	・開発用コンポーネント（エンジン） ・評価システム
	構文解析システムの精度を向上させる	・文 ・構文情報	・文法規則	入力文の解析結果	無	
	特定の分野における解析精度を向上させる	・文 ・文章 ・構文情報	・文法規則 ・構文解析モデル	入力文の解析結果	無	

	新出表現に対応する	・文 ・構文情報	・文法規則	入力文の解析結果	無	
意味解析	意味解析システムの精度向上	・情報（辞書） ・文	・無 ・文法規則	・無 ・入力文の解析結果	無	・開発用コンポーネント（エンジン） ・評価システム
機械翻訳	訳文検索システムを作る	・対訳コーパス	・対訳用例辞書	入力文の翻訳結果	高（資源そのもの）	・PKG ソフト ・オンラインサービス（Web） ・評価システム
	対訳辞書として利用する	・対訳辞書	無	対訳辞書	高（資源そのもの）	
	新出単語を見つけ出し、訳語を付ける	・新出単語	・対訳辞書	入力文の翻訳結果	低（単語）	
	日本語と英語の文と単語のアライメント	・文章	無	アライメント結果	高（資源そのもの） 低（単語） 無	
	既存の対訳辞書の訳語を調整し、翻訳品質を改善する	・対訳辞書	・対訳辞書	入力文の翻訳結果	高（資源そのもの） 低（単語） 無	
	訳語の曖昧性を解消する	・文	・対訳辞書	入力文の翻訳結果	低（単語） or 無	
	統計手法に基づく機械翻訳システムを作る	・対訳コーパス	・翻訳モデル	入力文の翻訳結果	低（単語）	
	特定分野の文パターンを学習し、翻訳品質を高める	・文章 ・文	・パターン辞書 ・文法規則 ・文モデル	入力文の翻訳結果	無	
	機械翻訳システムを評価する	・対話コーパス	・評価データ	評価結果	無	
情報検索	情報検索システムを評価する	・文書 ・検索クエリ ・検索正解	・評価データ	評価結果	無	・オンラインサービス（Web） ・開発用コンポーネント（エンジン） ・検索システム ・評価システム
情報抽出	略語定義の自動獲得	・文章	・略語と源用語のアライメント	略語定義	低（単語）	・開発用コンポーネント（エンジン） ・評価システム
	語義/人名の曖昧性解消	・文章 ・人名	・分類モデル ・人名	無	無	
	用語抽出（専門用語/話題語など）	・文 ・分類種別 ・シソーラス	・抽出規則 ・分類モデル ・シソーラス	無	低（単語）	

	用語抽出システムを評価する	・見出し語	無	無	無	
	語彙分類	・語彙分類体系 ・文章	・分類モデル	無	無	
	情報の重要度の推定	・文章	・重要度モデル	無	無	
	法令文の論理式への変換	・文章	・文法規則	無	無	
	相関図を作る	・文章	・相関モデル ・無	・無 ・相関図	・無 ・低（単語）	
	要約システムを作る	・文書	・文法規則 ・要約モデル	無	無	
	情報抽出システムを作る	・文書	・抽出規則	情報抽出結果	低（単語）	
	風評分析システムを作る	・文書 ・風評分類	・分類モデル	風評分析結果	無	
	文書分類システムを作る（トピック/ジャンル分類）	・文書 ・分野情報	・分類モデル	分類結果	無	
	情報抽出系システムの評価	・文 ・文章 ・文書	・評価データ	評価結果	高（資源そのもの） or 無（資源からの差分など）	
質問応答	原因表現をマイニングする	・文書	・原因表現のパターン	無	無	・オンラインサービス（Web） ・評価システム
	入力質問の回答を探すデータベースを構築する	・文書	・文書データベース	入力質問に対する回答（候補）	低（単語） ～ 中（文、句）	
	入力質問に対する回答を見つけ出すシステムを構築する	・文書	・質問・回答モデル	入力質問に対する回答（候補）	無	
言語資源の開発	辞書を作成する	・辞書 ・対訳辞書 ・百科事典	・辞書データ ・検索用インデックス ・音声データ	辞書データ	高（資源そのもの	・専用ハード（電子辞書） ・PKG ソフト ・オンラインサービス（Web）
	評価（対訳辞書の評価による再現率評価）	・文書	・単語頻度	無	無	無
	オントロジー/シソーラスを作成する	・辞書 ・対訳辞書	・オントロジー ・シソーラス	見出し語	低（単語）	・開発用リソース（辞書）
	コーパスを作成する	・辞書の属性 ・語義情報 ・文章	・コーパス	・コーパス	高（資源そのもの	・開発用リソース（コーパス）

文の生成	レビュー文の生成	・品詞体系 ・文章	・文生成規則 ・文生成モデル	・レビュー文	低（単語）	・オンラインサービス(Web)
	含意文の生成	・文章		・含意文		
校正	単語入力誤りの検知、修正をする	・単語	・誤りパターン、モデル	修正情報	低（単語）	
	キーワードを基に適切な文例を推薦する	・文	・文	文例辞書	高（資源そのもの）	
かな漢字変換	かな漢字変換ソフトを作る	・単語（＋かな情報） ・単語の並び	・単語辞書読み	かな入力の変換候補	低（単語）	・かな漢字変換ソフト ・IME
音声認識	音声認識システムを作る	・単語（＋読み情報） ・単語の並び(＋情報) ・音声コーパス	・単語辞書 ・単語の並び ・言語モデル ・音響モデル	入力音声の認識結果	低（単語）	・開発用コンポーネント（エンジン） ・評価システム ・音声検索システム
	特定の分野に適用する	・単語（＋読み情報） ・単語の並び(＋読み情報) ・音声コーパス	・単語辞書 ・単語の並び ・言語モデル ・音響モデル	入力音声の認識結果	低（単語）	
	未知語候補と読みの自動獲得	・単語（＋読み情報）	・文字と読みのアライメント	無	低（単語）	
音声合成	音声合成システムを作る(A)	・読み上げ音声 ・文の読み情報	・音素辞書 ・単語辞書 ・読み辞書	入力テキストの読み上げ音声	高（音声）	・評価システム ・開発用コンポーネント（エンジン）
対話システム	対話システムを作る	・対話コーパス	・発話境界モデル ・対話モデル	無	無	無
	対話システムを評価する	・対話コーパス	・評価用データ	無	高（資源そのもの）	評価用データ

（3）おわりに

言語資源がどのような自然言語処理技術の開発に利用されているか、またそれらの開発ではどのように言語資源が利用されているか調査した。言語資源の権利処理においてこのような言語資源の利用事例を言語資源保有者に示すことにより、言語資源の利用目的や利用方法をより具体的に知ることができるため、言語資源の提供への懸念を解消するのに有用であると思われる。

今後は、さらに言語資源提供者が安心して言語資源を提供できる環境の整備と言語資源の提供により自然言語処理技術の向上を通して社会貢献につながることを認識してもらえるよう、言語資源を利用して開発された各自然言語処理技術やその実用化事例の説明資料の作成などを行う予定である。ま

た、本調査では学術研究の事例が主となったが商用の事例についても拡充していきたい。

参考文献

1）言語処理学会第 14 回年次大会予稿集(NLP2008 CD-ROM)

(執筆者 井田裕子)

3.3.3 言語資源を利用する枠組の調査

(1) はじめに

代表的な言語資源配布団体である LDC (Linguistic Data Consortium [1)]), ELDA（Evaluations and Language resource Distribution Agency [2)]）について、現在採られている言語資源配布に係る契約形態について調査した。

(2) 具体例

[LDC]

LDC は、大学、会社および政府研究所が連携して設立したコンソーシアムであり、研究開発目的のための音声、テキストのデータベース、辞書、その他の資源の開発、収集、配布を行っている。

LDC では、会員を所属団体毎に次の 3 つに分けて扱っている。

- 非営利団体　　Not-For-Profit organizations
- 営利団体　　For-Profit organizations
- （米国）政府機関　U.S. Government Entities

ここで、非営利団体に関しては、研究開発の用途が主となり、また、その契約内容（特に言語資源の使用方法）については、営利団体のそれに包含されるところが大きいため、割愛する。また、政府機関に関しても、やや特殊な事例であるため、割愛する。

営利団体向け契約書のテンプレートは、LDC の Web ページに於いて公開されている。LDC とメンバ契約を結んだ者は、LDC から提供されている全てのデータベースを用いた研究開発を認められる。加えて、著作権法を含む法律によって許された範囲、及び、契約時の別途規定に反しない限り、商業を目的とする成果物に LDC データベースの一部を組み入れることを許可されている。ただし、明示的な規定がない限り、データベースそのもの、乃至は、その一部であっても、複製、再配布、送信、公開することはできない。

ここにまとめた LDC 会員として条項に加えて、個別の言語資源に関して、版権所有者によって特

に課された制限により、資源提供の際に個別ライセンスが必要となる場合がある。現在、LDC のカタログには、472 の言語資源が登録されているが、内 87 の資源については、別途書面への署名が求められている。特別なライセンス条項が設けられた言語資源は、一部の言語資源を除いて、使用を研究開発のみに制限し、商用でのデータ組み込みを禁じる条項が設けられている。

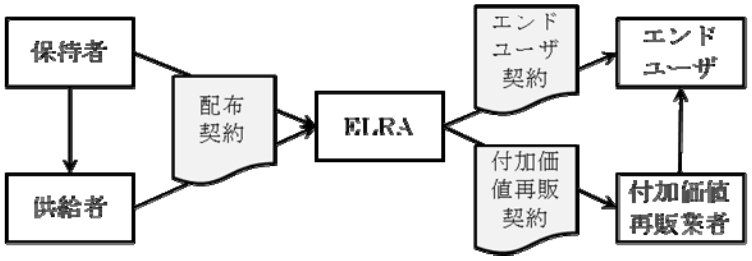
[ELDA]

ELDA は、ELRA(European Language Resources Association [3])の実務機関として設立された団体で、言語資源の開発、配布等を行っている。

ELDA では、会員を所属団体毎に分け、更に言語資源の使用に際しては、その利用目的に応じた契約が結べるようにしている。ELDA の会員／利用目的毎の契約分類をまとめると次の表のようになる。

- 学術（研究）利用　academic use (research)
- 営利団体による学術利用　research use by a commercial organization
- 商業利用　commercial use
- 学術（研究）機関における評価用途　evaluation use by a research organization
- 営利団体における評価用途　evaluation use by a commercial organization

また、ELDA では言語資源の保持者（提供者）とユーザ間の関係を単純化し、契約毎の変化を最小化するために、契約を次図に示すモデルに従って結ぶ。



出版事業者を除いて、多くの言語資源保持者は内部ニーズに沿う資源を開発／獲得することが目的であり、配布などは考えていないのが実情である。そこで、そのような言語資源の、他ユーザへの提供を促進するために、ELRA では、両当事者の責任および義務を規定する包括的な契約を上図のようにモデル化している。

エンドユーザ契約（ライセンス）では、言語資源の使用、及び、合意されたデータアクセスおよび使用に際して必要な翻訳、改編、改作、修正することを認め、契約組織内の自然言語処理研究の目的

に於いて言語資源を再加工し構築することを認めている。しかしながら、提供された言語資源の全体、乃至は、一部に基づく成果物、あるいは、サービスの市場への提供は認められていない。更に、言語資源そのもの、乃至は、その一部であっても、複製、再配布、送信、公開することはできない。

付加価値再販契約では、付加価値再販業者が ELDA から提供されたデータに基づいて制作した成果物、あるいはサービスを非独占的に配布する権利を与えること更に認めている。ただし、原言語資源を改造することを可能にするあらゆる形式で製品またはサービスを提供することは禁止されている。以上の点でエンドユーザ契約と異なるが、それ以外の使用に係る制限事項は同等である。

(3) 議論

今日まで運用されてきた言語資源利用の枠組は、どちらかといえば、言語資源提供者の保護に重点が置かれている。これは、提供者（保持者、著作者）の権利を守る上で、最も重視されるべき事である。また、このことが、多くの言語資源が広く公開され、言語情報処理技術の発展に大きく寄与してきた。

一方で、商用利用に於いて制限があるなど、成果物としての言語情報処理技術が、新しい市場の開拓していく面では、制約が多いことも事実である。一口に商用と言っても、その利用形態は様々であり、個々には、綿密な交渉の上で、解決されてきた事例も見られるものの、一般に適用するには今もって課題が多い。この様な事実は、ここで改めて触れる必要もない程に、長年議論されてきており[4]、根本的な解は与えられていない。

大きな障壁として、多くの言語資源の保持者にとって、自身の資源がどのように使われ、流通していくのかを、必要十分に説明する枠組みが準備されていなかったことが考えられる。例えば、提供した言語資源が検索技術の開発に利用される、という情報があっても、具体的にどのような扱われ方をするかを想起することは難しい。

このような背景から、前節にまとめたように、その用途を権利上課題となる観点から整理し、提供者側にわかりやすい形で提供することを相互理解の行くことを第一歩とし、各事例を蓄積していくことで利用の可能性を広げていくことが今後の課題である。

参考文献

1) LDC, http://www.ldc.upenn.edu/

2) ELDA, http://www.elda.org/

3) ELRA, http://www.elra.info/

4) Jeff Allen, “Language Resources: an Issue of Information Management”, Keynote talk on Resources presented at the Conference on Co-operation in the Field of Terminology in Europe

sponsored by the European Association For Terminology (EAFT) and the Union Latine. 17-19 May 1999.

(執筆者 釜谷聡史)

3.4 有害サイト関連調査報告

3.4.1 ネット犯罪に関する調査報告

(1) はじめに

言語資源の活用法の一つとして、インターネットを利用した犯罪やいじめなどの抑止技術開発への利用が挙げられる。そこで、インターネットを利用した犯罪やいじめを取り巻く現状を調査した。まず、インターネットなどを利用した犯罪（以下、インターネット犯罪）において、どのような犯罪が多いかを確認するために、警察庁に寄せられた相談件数を調べた。その結果、詐欺などの金品を騙し取ることを目的とした犯罪に次ぎ、名誉毀損・誹謗中傷・迷惑メールといったネットいじめの様態として知られる犯罪が多いことがわかった。次に、インターネットを利用したいじめの実態について、日本を始め北米など諸外国における状況を調査した。その結果、様々な国や地域でインターネットの普及に伴い、その問題が拡大していることがわかった。さらに、日本におけるネットいじめへの対応として、文部科学省における対応策を調べた。その結果、学校や家庭に加え、企業に対しても社会的責任の認識や、技術的、人的支援が要請されていることがわかった。最後に、ネットいじめなどに対する教育の現場での取り組みを調べた。その結果、教育委員会では情報モラルの指導が重要であると位置づけ、教員への情報モラル指導資料を作成していることがわかった。また、情報モラルの必要性が低年齢化していることも伺えた。

(2) インターネットなどを利用した犯罪に関する相談件数

インターネット犯罪に対して、警察に寄せられた相談件数を調べると、ネットいじめに繋がる名誉毀損・誹謗中傷や迷惑メールの件数が多いことがわかる。詐欺・悪質商法やインターネットオークションなど詐欺に関する件数が最も多く、全相談件数のうち 5 割以上を占める。次いで、名誉毀損・誹謗中傷や迷惑メールといった犯罪が 2 割にのぼっている。

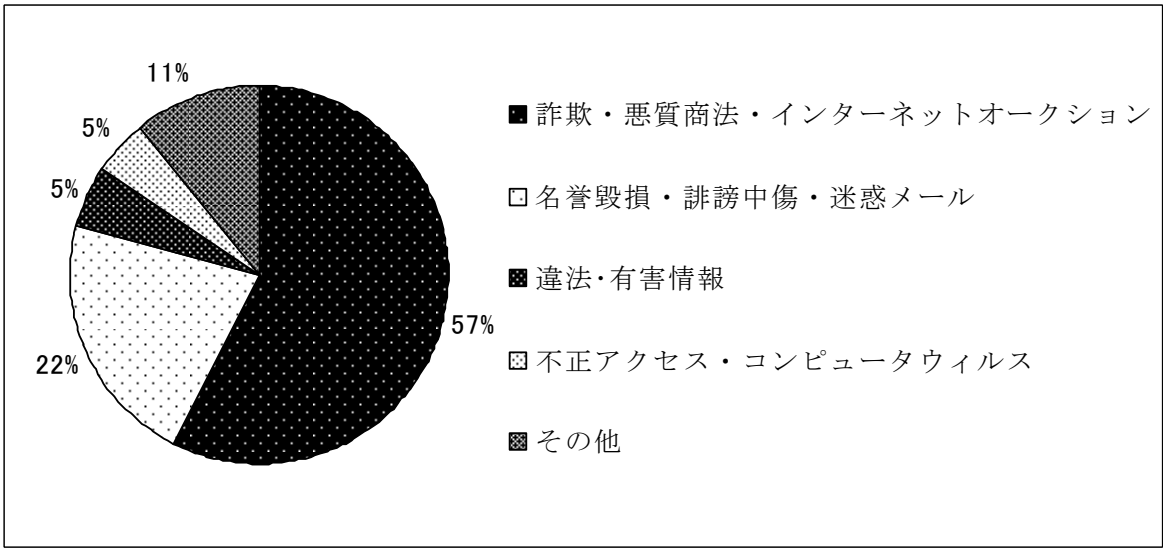


図 3.4.1-1 相談区分の割合（平成 20 年度）

また、インターネット犯罪に関する相談受理件数の平成 15 年から 19 年までの推移をみると、詐欺に関する相談が 60%程度の増加であるのに対し、迷惑メールなどに関する相談は 170%以上も増加していることがわかる。

表 3.4.1-1 相談受理件数の推移（単位：件）

	H15	H16	H17	H18	H19
詐欺・悪質商法・インターネットオークション	26,737	48,864	58,931	35,925	45,531
名誉毀損、誹謗中傷・迷惑メール	4,948	7,631	9,757	10,967	13,516
違法・有害情報	4,225	4,157	5,317	4,335	3,497
不正アクセス、コンピュータ・ウィルス	1,147	2,160	3,965	3,323	3,005
その他	4,697	7,802	6,203	6,917	7,644
合計	41,754	70,614	84,173	61,467	73,193

（3）インターネットを利用したいじめ（ネットいじめ、“cyberbullying”）の現状

アメリカ合衆国における 2000 人の中学生対象に行なわれたネットいじめに関する調査（2007）により、ネットいじめを受けた経験がある学生、ネットいじめをした経験がある学生が、いずれも 17%程度にのぼることがわかった。また、ネットいじめに及ばないまでも、嫌がらせメールなど何らかの被害を被っている学生が 4 割を超えることも明らかになった。

カナダやイギリスにおいても 2001 年の調査により、10 代の 25%程度が嫌がらせメールの被害にあ

っていると思われる。また、カナダの中学生の 74%が何らかのいじめの被害者であり、そのうち 27%がネットによるいじめであったことが明らかになった。

その他、先行研究 [5]のネットいじめ調査対象地域として日本を始め、韓国、シンガポール、中国、タイ、インド、オーストラリア、ニュージーランドが挙がっていることから、ネットいじめの被害はインターネットの普及と共に全世界的な問題を呈していることが伺える。

全世界的なネットいじめの取り組みとして、International Cyber-Bullying Project（国際ネットいじめ対策プロジェクト）が発足し、2008 年 7 月にニュージーランド、クィーンズタウンにおいて NetSafe Conference08 を開催するなど、様々な活動が今後も期待される。

日本におけるいじめに関する調査（文部科学省、平成 18 年度児童生徒の問題行動等生徒指導上の諸問題に関する調査）において、全国の小・中・高等学校及び特殊教育諸学校におけるいじめの認知件数は、12 万件を超えることがわかった。このうち、いじめの様態として「パソコンや携帯電話等で、誹謗中傷や嫌なことをされる」様態が新たに加えられ、いじめにおける新たな問題の浮上を伺わせる。

携帯電話を利用してのネットいじめは日本独自の問題といえる。欧米諸国や日本を除くアジア地域と比較した場合、日本における携帯電話利用者において小学生や中学生を始めとする低年齢層の占める割合が多い。常時接続が可能な携帯電話であるため、被害者は常にいじめの恐怖にさらされることになる。さらに、加害者も常にいじめを行なうことが可能となるなど、パソコン利用によるネットいじめとは異なる問題を呈する。今後、携帯電話の利用制限やフィルタリング機能の拡充など、ネットいじめに対する措置を講じる必要がある。

(4) 文部科学省における提言

平成 19 年 9 月、文部科学省はインターネットを利用したいじめの防止に向けて「子どもを守り育てる体制づくりの有識者会議」を開催し、平成 20 年 6 月に「子どもを守り育てる体制づくりのための有識者会議まとめ（第 2 次）」を作成した。

この会議の背景には、子供たちのインターネットにおけるいじめ問題の深刻化がある。インターネットを利用したいじめには次の特徴が挙げられる。

1) 不特定多数の者から、特定の子どもに対する誹謗・中傷が絶え間なく集中的に行われ、また、誰により書き込まれたかを特定することが困難な場合が多いことから、被害が短期間で極めて深刻なものとなること。
2) ネットが持つ匿名性から安易に書き込みが行われている結果、子どもが簡単に被害者にも加害者にもなってしまうこと。
3) 子どもたちが利用する学校非公式サイト（いわゆる「学校裏サイト」）を用いて、情報の収

集や加工が容易にできることから、子どもたちの個人情報や画像がネット上に流出し、それらが悪用されていること。

4) 保護者や教師など身近な大人が、子どもたちの携帯電話やインターネットの利用の実態を十分に把握しておらず、また、『ネット上のいじめ』の発見が困難であるため、その実態を把握し、効果的な対策を講じることが容易でないこと。

この会議を通じ、いじめ問題に関する提案として、（1）理解促進・実態把握、（2）情報モラル教育の充実とルールの徹底、（3）未然防止・早期発見・早期対応、（4）いじめられた子ども等へのケアの 4 項目が提出された。また、これらの提案の実現に向けて、行政、学校、家庭、関連企業における役割とその課題が提案された。その中で、関連企業に対しては次の役割と課題が下記のとおり提出された。今後、いじめ問題に対して、関連企業としての役割と課題に真摯に取り組む必要がある。

- 社会的責任を認識し、子どもたちの携帯電話やインターネットの利用実態に目を向け、フィルタリングの設定をはじめとした適切な措置を講じていくことが期待されます。
- 悪質な書き込み等に関するチェック体制の整備や日常的な巡回活動、削除要請に対する迅速な対応、ネット上のパトロールを行う人材養成への協力などが期待されます。

⑸ 教育委員会（大阪府）における活動

教育委員会では、情報モラルの重要性を鑑み、情報モラルの指導方法の研究を行い、各学校で教職員が生徒を指導する際の資料として「情報モラル指導資料」を作成、配付を行なっている。

平成 14 年の初版から 5 年における情報機器の発達による情報環境の急速な変化により、児童生徒のコミュニケーションのあり方や情報の受発信等に大きな変化がみられるようになったことから、改訂版が作成された。

最近の児童生徒は、携帯電話のメール、オンラインゲーム等の恩恵を被るようになった。その一方で、出会い系サイト、架空請求詐欺等のトラブル、ウェブサイト上での誹謗中傷、個人情報の流出によるネットいじめ、著作権侵害など、児童生徒自身のモラルが問われる事例が多くなった。このような高度情報通信社会の発展を背景に、様々なメディアから発信される内容について、児童生徒が主体的に判断し、コミュニケーションを創造していく能力、および情報社会で適正な活動を行なうための基盤となる考え方や態度など、いわゆるメディアリテラシーや情報モラルを育成する必要があると教育委員会は考えている。児童生徒のメディアリテラシーの定着を図るためには、児童生徒が安全な環境で適切に情報を活用する能力を身につけられるよう教職員が指導することが肝要であると考えられている。

増補改訂において、児童生徒が学校段階の早い時期に様々な情報を主体的に判断し、自分の責任で行動することができる力を身につけられるよう、小学校、中学校、高等学校におけるメディアリテラシー教育の指導事例が新たに盛り込まれた。

「情報モラル指導資料」が当初、高等学校の教職員向けの資料と作成され、小・中学校の教職員に対しては参考資料程度の位置づけであったのに対し、増補改訂版では、小・中学校での指導事例が盛り込まれていることからも、情報モラルの指導の必要性の低年齢化が伺える。

(6) おわりに

文部科学省の提言や教育委員会の活動を踏まえると、今後、ネットいじめを始めとするインターネット犯罪の抑止に情報処理技術や言語処理技術を役立てることが課題といえる。これまで、情報処理・言語処理技術は、商用システムや法科学的手法の開発において利用されてきた。例えば、情報フィルタリングやプロファイリングといった手法における研究が挙げられる[1-4]。これらの手法を基盤として、ネットいじめの対策ツール開発に必要な自然言語資源の構築といった技術的問題のみならず、そのツールの実装や利用における法律的問題も検討すべきである（フィルタリングにおける法律的問題に関しては先行研究[6]を参照)。

参考文献

1) S. Argamon, S. Dhawle and J. W. Pennebaker: “Lexical Predictors of Personality Type”, Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America, (2005)

2) S. Argamon, M. Koppel, J. Pennebaker and J. Schler:”Automatically Profiling the Author of an Anonymous Text”, Communications of the ACM archive, Vol.52, No.2, pp.119-123 (2009).

3) J. Bachenko, E. Fitzpatrick and M. Schonwetter: “Verification and Implementation of Language-based Deception Indicators in Civil and Criminal Narratives”, Proceedings of the 22nd Coling, pp.41-48 (2008).

4) F. Iqbal, R. Hadjidj, B.C.M. Fung and M. Debbabi: “A Novel Approach of Mining Write-prints for Authorship Attribution in E-mail Forensics”, Digital Investigation, Vol.5, pp.42-51 (2008).

5) S. Shariff, Cyber-Bullying: Issues and Solutions for the School, the Classroom and the Home, Routledge, New York, (2008)

6) 新保史生: “フィルタリングと法”、情報の科学と技術 (特集 情報のフィルタリング)、Vol.56, No.10, pp.475-481 (2006).

(執筆者 小谷克則)

3.4.2　有害サイトの現状に関する調査報告

近年、インターネット上の有害サイトが、様々な犯罪行為を助長・誘発する温床となり、深刻な社会問題となっている。前項で述べたような、警察庁や文科省、教育委員会など、官を主導とした有害サイトの監視や取り締まり、指導、法整備などが進められており、また、携帯電話キャリアやプロバイダによるフィルタリングサービスや、PC 上のフィルタリングツール、ブログや SNS 等のコミュニティサイトにおける発信情報の管理、第三者機関による有害サイトの URL リスト作成など、民間を主導とした対策も進められている。しかしながら、日々インターネット上で更新される大量のテキスト情報を対象に、人手で漏れなく有害サイトの監視を行うのには限界があり、自動的かつ網羅的に有害サイトを発見・監視するための支援技術の確立が急務となっている。

当該技術を包含する自然言語処理の分野では、近年、インターネット上に氾濫する CGM（ブログ、QA、評価評判等）を対象とする分析技術の研究が活発に進められている。有害コンテンツの発見・監視も、ネット上の発信情報を対象とする分析技術の応用という点ではこれらの研究と類似する点が多いものの、有害サイトを分析研究する心理的負担や、有害サイトの実態に関する公開情報の不足などから、自然言語処理の研究分野において、有害サイト監視等への本格的応用が議論される機会は少なかった。

そこで、今回、インターネット上における有害サイトの実態調査を実施した。今回調査対象としたのは、表 3.4.2-1 に示すようなサイトである。「関連犯罪等」は、各サイトで犯罪を助長・誘発すると考えられる代表的犯罪行為を示し、「有害コンテンツ」は、その具体的内容を示した。「存在」は、今回の実態調査において、その存在が確認できたかどうかを示している。直接存在を確認したものは「○」、間接情報から確認したものは「△」で示した。また、「隠語の例」は、コンテンツ中に特有の隠語が見られる場合、その一例を示した。

表 3.4.2-1　有害サイトの実態調査（2009 年 3 月時点）

サイト種別	関連犯罪等	有害コンテンツ	存在	隠語の例（ある場合）
出会い系サイト	買春	買い手の募集	△	＊り切り、ホ別
家出サイト	児童買春	泊まり先の募集	△	神様、神待ち
薬物売買サイト	薬物	薬物の販売	△	
裏情報サイト	危険物	爆発物の作成方法	○	
	薬物	大麻の栽培方法	○	
	偽造	文書・口座の偽造方法	○	

	復讐・殺人	復讐・殺人の方法	○	
自殺サイト	自殺	自殺方法の説明	○	サソポール、ノックダウン
	集団自殺	自殺パートナーの募集	○	氏ぬ
裏バイト	振り込め詐欺	実行犯の募集	△	
	薬物	大麻栽培者の募集	△	
	口座売買	売り手の募集	△	
復讐サイト	殺人	代行者の募集	○	
	いやがらせ	代行者の募集	○	
ブログ	いじめ	特定人物の誹謗・中傷	○	氏ね、SN
学校裏サイト	いじめ	特定人物の誹謗・中傷	○	氏ね、SN

各サイトごとの特徴と調査によって判明した実態を下記にまとめる。

1) 出会い系サイト

会員制。登録されている会員の中から、友人や交際相手を募集し、相手を見つけるサイト。サイトの会員数が多いと、サイト運営側が優良であるかどうかに関わらず、売春などの犯罪につながる有害な書き込みが含まれる傾向にある。サイト単位で有害無害の判断を行うことは難しく、サイト内の各書き込みや会員ごとに、内容ベースで判断する必要がある。売春の買い手募集の書き込みは、犯罪行為の状況証拠を隠蔽する目的で、隠語が用いられる。

2) 家出サイト

会員制。家出中もしくは家出を希望している女性が、会員登録をしている男性に対して泊まり先の募集を行うサイト。出会い系サイトと異なり、その趣旨から、基本的にすべてが有害サイトであると考えられる。

3) 薬物売買サイト

薬物の購入を希望する利用者と、薬物の販売を行う利用者とが、互いに電子掲示板を利用して接触を図るサイト。今回の調査では、実態を把握することができなかった。

4) 裏情報サイト

爆発物の作成方法、大麻の栽培方法、文書や銀行口座の偽造方法、復讐・殺人の方法などの、一

般には公開されていない、特殊な情報を収集して提供するサイト。今回調査したサイトの情報の中には、実施自体が違法行為となるものも多く含まれていた。有料会員制のサイトでは、一部の情報を無料で公開することで会員を集めている。

5) 自殺サイト

自殺を希望する利用者が、自殺の遂行時期や方法などを提示し、集団自殺のパートナーを募集するサイト。募集の書き込み自体が自殺幇助と見なし得るものもある。今回調査した掲示板利用者の中には、自殺希望者に扮して警察に通報している利用者も紛れていた。

6) 裏バイト

違法行為のアルバイトを募集するサイト。今回の調査では、実態を把握することができなかった。周辺情報によれば、振り込め詐欺の実行要員や、大麻の栽培、銀行口座の売買などの違法行為を、高額アルバイトとして募集する掲示板などが存在すると推測される。サイトよりも携帯電話へのDMが主なチャネルである可能性もある。

7) 復讐サイト

特定人物への復讐を希望する利用者が、依頼内容を掲示板に書き込むサイト。今回調査した掲示板では、依頼内容の書き込みは見られたが、実際に復讐代行者とのやり取りの実態は把握できなかった。

8) ブログ

登録制。利用者が自身の日々の体験や思想などを記載した記事を掲載し、利用者が相互に閲覧、感想などを書き込みあうサイト。サイトの会員数が多いと、サイト運営側が優良であるかどうかに関わらず、いじめなどの有害な書き込みが含まれる傾向にある。今回の調査では、恨みを持つ個人の氏名を挙げて中傷する記事が見られた。

9) 学校裏サイト

各小中高校の学生が、自分の在学校ごとに、学校関係者には内緒で作成利用している掲示板。在学学生同士のやり取りが主であるが、その中に特定の学生や教師に対する誹謗中傷などの書き込みが含まれる。「学校裏サイトチェッカー」などの、学校裏サイトへの監視の取り組みが強化される一方で、監視の目が行き届いていない未発見のサイトも多数ある模様。今回調査した中でも、学校裏サイトチェッカーに登録されていないサイトが見られ、問題となる書き込みも確認された。別のサ

イトでは、パスワード認証を導入しており、監視が困難になっている状況が伺えた。

以上、今回調査を行った有害サイトの状況について報告した。

今回の調査を通じて、改めてインターネット上における有害サイトが社会問題化している深刻な状況を認識した。また同時に、同課題における自然言語処理技術の本格的適用が、問題解決の有効な手段となり得るという感触を持った。有害サイトの監視・発見に適用するための応用研究を促進する上で、有害コンテンツを言語資源として収集蓄積し、広く技術者が分析や評価に利用可能とすることが重要であると考える。

（執筆者 石川開）

４．情報アクセス技術

4.1　はじめに

モバイル技術の進展により、場所を問わずにネットワークを介したさまざまなサービスが提供されるようになってきた。しかし、一方で現在のモバイル端末には、通信速度、インタフェース（画面サイズ、入力方式など）などの点において制限があり、固定端末に比べると劣る面もある。今後のネットワークサービスとして、これらの異なるタイプの端末をシームレスに組み合わせて、場面に応じて適切なサービスを提供するようなシステムが考えられる。このような背景をふまえ、今年度は、場所や利用目的などユーザの利用状況に応じてサービスを最適化する状況依存型サービスの実態とそのための要素技術について動向調査をおこなった。

4.2 節では、状況依存型サービスに関連する技術として 4 件のヒアリングをおこなった。4.2.1 は無線 LAN を使って位置情報を取得するための基礎技術に関するものである。位置情報の取得は GPS を用いるものが一般的であるが、GPS では屋内では使えないという問題がある。無線 LAN の基地局の位置のデータベースと電波の強度から位置を推定する技術について紹介している。この技術を使ったサービスはすでにおこなわれており、状況依存型サービスの基盤をささえる技術である。

4.2.2 は企業内の情報検索サービスの例である。一般に公開された情報の検索と企業内の情報の検索では異なる観点のサービスが重要となる。特にセキュリティの観点は重要である。情報内容も他人に見せるというより自分達が使うという観点が重視されている。

4.2.3、4.2.4 は個人の情報利用に関する技術である。4.2.3 は Web 検索のパーソナライズと検索結果・検索履歴を備忘録的に使うことを支援する技術、4.2.4 は、個人の活動を記録して、それを生活に役立てるといういわゆるライフログの技術に関するものである。

4.3 節では、現存する状況依存型サービスをインターネット上で調査し、いくつかの特徴素性の観点から分類・整理したのでその結果について述べる。特徴素性のカテゴリとしては、「サービスに使われるデバイス」、「サービスの成熟度」、「サービスに使われる基礎技術」、「サービスの機能・目的」、「サービスが利用する状況情報」、「サービスが分析する対象とその結果」の 6 カテゴリ、総計 38 素性を用いた。サービスと素性の相関関係を見ることによって、現状のサービスおよび、今後、現れるであろうサービスの特徴についてまとめた。

4.4 節では、この分野の研究動向を調べるために ACM SIGIR 2008 で発表された論文の中でパーソナライゼーション、ユーザーフィドバックなど、状況依存性に関係ありそうな 7 件の論文の概要について紹介している。

最後に 4.5 節では委員のひとりが WWW2008 に参加したので、会議の様子について報告している。

（徳永健伸）

4. 2　状況依存型サービスに関する技術動向

4. 2. 1　無線 LAN を用いたロケーションウェア PlaceEngine の技術とその応用

クウジット株式会社の塩野崎敦氏へのヒアリングに基づき、位置推定技術に関連して、無線 LAN を用いたロケーションウェア PlaceEngine について紹介する。なお、以下に記載した内容はヒアリングが行われた平成 20 年 6 月当時の情報に基づくものであるため、情報が現在と異なっている場合があることを了解いただきたい。

(1)　PlaceEngine とは？

近傍の無線 LAN 電波を利用して位置を推定する基盤技術およびサービスを提供するものである。周囲の無線 LAN 電波をセンサのように感知する技術を用いる。

www.placeengine.com において、周囲の Wi-Fi 情報を電測する PC 用クライアントソフトを無償配布している。ソフトをインストール・起動した状態で、www.placeengine.com/map にアクセスし、このサイトに用意されている機能を利用すると現在位置を地図上に表示できる。サイトにはフィードバックフォームも用意されていて、ユーザはこれを通じて無線 LAN の電測情報を登録することができる。

(2)　位置推定の方法

観測点で感知できたアクセスポイント（AP）の推定位置の情報から、三点測量に類した計算手法を用いて観測点の位置を推定する。基本出力は緯度経度となる。

ビーコン信号には MAC address と ESS ID が入っているので、観測された各々の AP を特定することはできるが、AP の位置（緯度経度）の情報は基本的には手に入らないので、位置推定のためのデータベースを構築して、一つ一つの AP の推定位置を常に更新している。アクセスポイントの推定位置はユーザの位置を特定するための計算のみに用いられ、公開はされていない。

計算のベースとなる各地点の電波の情報は、電測係による計測の他、「位置を教える」ボタンによるユーザからの明示的なフィードバックによって入手する。間違ったデータがフィードバックされても、全体の処理で自然とそのデータの重みが落ちていくので、特別な対処はしていない。

フィードバックによって位置を明示的に登録してもらえなくても、位置問い合わせ（現在位置の取得）をしてもらうだけでもシステムにとって有用な情報が入手できる。

例えば、AP の位置補正については、観測点の位置が既知の場合（例えば明示的に登録してもらえた場合）には、現在の AP の推定位置が、電波が届く範囲よりも遠くであれば少し観測点に近づけるように補正する。一方で、観測点の位置が未知の場合（明示的に登録してもらえなかった場合）でも、

見えている AP 同士の間隔が、電波が届く範囲（の直径）よりも遠ければお互いに少し近づけるように補正する。

緯度経度以外にも、例えばビルのフロアを出力することも可能である。高度を測定しているわけではなく、明示的な登録（人が登録時に「5Ｆ」というように入力する）による情報を提示している。

(3) GPS との比較

PlaceEngine でも GPS と同程度の性能は出せることを実験で確認した。GPS はビルの間などでは大きくずれる場合もある。また、室内や地下街では PlaceEngine の方が強い。GPS はコールドスタートした場合、電源を入れてから位置測位できるまで時間がかかる。5 秒～数十秒かかることもある。

(4) 現在のサービス展開

・PSP ソフト　… みんなの地図 2 と 3、ニッポンのあそこで（地図エンタテインメント）

・web サービスとの連携　…　mapion などのポータル

・アプリケーションとの連携　…　goo スティック、x-Radar、など

・WiFi モバイル機器との連携　…　デジカメ DSC-G1（位置情報を EXIF に保存）など

・コミュニケーションアプリとの連携　…　おともだちレーダーBuddy Radar（Skype 拡張機能用）

(5) その他の話題

・PlaceEngine クライアント API をホームページで公開中。

・ライフタグ

創業者の一人でもある東大の暦本教授が大学で開発中の「WiFi 位置認識に基づくライフログデバイス」。AP 情報を（時間と対比して）常時記録しておき、あとで位置情報に変換することが可能。

・Where 2.0 conference（2008 年 5 月、O'Reilly 主催）にて、iPod 上で PlaceEngine と内蔵加速度センサの組合せで位置推定するものを発表。

(6) 主な質疑応答

Q：データベースに抱えている AP の数は？

A：国内 70 万個、うち 60％が東京。政令都市の主要エリアはカバーしている。

Q：AP 位置データの更新頻度は？

A：インターネットでアクセスの場合はそのときの最新の位置データで推定する。ローカル DB 版もあり、例えば PSP の場合はメモリスティックで DB のスナップショットを持つことになる。更新版（1 週間に 1 回）のダウンロード可能。容量は工夫して小さくしたので 10M で収まっている。

中身を直接見ることはできないようにしてある。ローカルDB版にはフロア情報は入っておらず、緯度経度の情報のみ。

Q：ユーザの情報は何件ぐらい集まったか？

A：AP の推定位置のうち 15％ぐらいがユーザから寄せられた情報によるもの。

Q：利用ユーザ数は？

A：ユーザ ID を取っていないので不明。PSP のタイトルとしては数万個ぐらい出ている。ソフトのダウンロードは月数千、累積では万単位にはなっている。

Q：PlaceEngine ではユーザの移動履歴情報を継続的に捕捉する手段は用意されているか？

A：PlaceEngine では、個人情報はいっさい管理していないため、特定の個人の情報は収集しておらず、ユーザの移動履歴を継続的に捕捉する機構は用意していない。しかし、PlaceEngine を使ったサービスやアプリケーションのレイヤでは、ユーザ情報と位置情報の対応付けをすれば、継続的に移動履歴を記録することはできる。また、PlaceEngine のクライアントを使えば、ユーザは自分の行動履歴を継続的に記録することはできる。

Q：ユーザの位置情報が意図しないサービスやアプリケーションにとられる可能性はないか？

A：PlaceEngine に対応したアプリケーション・サービスは、PlaceEngine の API キーを取得し、その情報を元に PlaceEngine クライアントとやりとりを行う。したがって、ユーザは未登録の（不正な）アプリケーションからのアクセスを識別できる。さらに任意の登録されたアプリケーションが位置情報を取得する場合も、警告メッセージが表示されるので、ユーザはそのサービス・アプリケーションに位置情報を渡すことを自分で制御できる。

Q：AP の位置推定により住所がわかってしまったり、自分の AP の情報を勝手に利用されてしまったり、ということについて、AP を設置している側が問題視することはないか？

A：セキュリティ面など技術的に可能な対応策は日々検討し実施している。また、AP 設置側に関する同様のご意見については、PlaceEngine は、ブロードキャストされている無線 LAN のビーコン信号のみを用いて、ユーザー参加型の位置情報基盤を実現することができる社会性のあるものであると認識いただけるよう努力していきたい。

Q：ビジネスモデルは？

A：今のところ PSP のタイトル販売と、（ポータル）サイトとの連携を中心に実施している。

Q：日本の同業者は？

A：名古屋大の河口先生が少し前から研究されている。

Q：ビジネスとしてやっているところは？

A：日本にはない。アメリカでは skyhook と navizon がある。 iPod Touch に搭載されているのは skyhook 製のエンジン。ゲーム機器にソフトとして入ったのは PlaceEngine がはじめて。

（池野篤司）

4.2.2 情報融合ソリューション －検索基盤構築事例の紹介－

企業内の情報検索サービスの一例として、株式会社エクサの山尾茂氏、安藤寿氏、徳広淳一郎氏へのヒアリングに基づいて、企業内検索システム IDOL（Autonomy 社）および検索ソフトウェアプラットホーム FAST ESP（Enterprise Search Platform：マイクロソフト社）について紹介する。

同社では、ポータル＆コンテンツマネジメントソリューションを提供しており、このうち企業内検索ソリューションとして、IDOL および FAST ESP を活用している。

(1) Autonomy IDOL

- 入力していないキーワードでの絞込み検索 独自アルゴリズム（情報理論、ベイズ推論）に基づき、検索結果から次の候補となる検索キーワードを提示（クエリガイダンス）したり、検索結果をクラスタリングしてビジュアルに表示したりすることが可能
- 豊富なコネクタとフォーマットに対応（Notes, Exchange, DB2, Documentum, ODBC, Oracle, SAP, Siebel, SharePoint, POP3, HTTP, FileSystem など）
- セキュリティモデルに対応（マップドセキュリティによるパフォーマンス確保）
- コラボレーション機能（エキスパートの発見、コミュニティの形成）

(2) FAST ESP

- ホワイトボックス・アーキテクチャー（チューニングが容易）
- 入力していないキーワードでの絞込み検索

クラスタリング：検索結果をその内容に基づいて分類し、階層構造でナビゲート可能に

エンティティ抽出：文章中のエンティティ（人、場所、会社、日付等）を抽出し、キーワードに関連性の高い情報を抽出

- ランキング技術（ランキングのカスタマイズ）
  - ランクプロファイル
  - ブースト＆ブロック：表示直前での制御が可能（例：ＥＣサイトで売切品をブロック、在庫品をブースト）
- セキュリティ対応
- 管理機能
  - クエリレポート
  - ０ヒットクエリとシノニム（同義語）：検索にヒットしなかったクエリ（＝０ヒットクエリ）をもとに同義語登録等の対策を容易にするＧＵＩあり
  - ログ活用によるフレームワーク

(3) 情報融合ソリューション

- 従来の企業内検索の発展形（検索を意識させない検索）
- キーワード入力部をできる限り使わずに業務利用

(4) 適用事例

- 精密機器メーカー
  - 設計時の部品検索で活用（部品データの絞込み検索、関連文書閲覧）
  - PLM システム、生産管理システム、技術情報システムなどと連携
  - データベース検索では難しい高速検索を実現、また手間のかかるスキーマ変更が不要のため柔軟性もあり
- プラント大手
  - ドキュメントの寿命が長い、全社サーバ上にある作成途中のドキュメントも含めて検索したい
  - 既存システム資産を活かしつつ、プロジェクトに関する様々な文書を仮想横断的に検索
  - フォルダパス、ファイルタイプ等のメタデータを自動抽出することで文書を分類
- 機械メーカー
  - 高価な機械製品を販売しており、保守のため部品単位でシリアル番号を管理
  - 問い合わせ対応で横串検索
- 化学系メーカー
  - 技術情報（特許など）の著者をキーにした Know Who 検索で活用

(5) 主な質疑応答

【Autonomy IDOL について】

Q：索引生成時と検索時点とでアクセス権が異なる場合の対処は？

A：ファイルサーバ等は定期的にポーリングしており、アクセス権はほぼリアルタイムで更新可能。

Q：検索ログのセキュリティはどうなっているのか？

A：内部プロファイルの仕組みにより収集時の運用で制御可能。

Q：クエリガイダンスは静的なガイダンス、動的なガイダンスいずれか？

A：動的なガイダンスである。検索結果の文書から抽出したキーワードをガイダンスとして出力している。

Q：ガイダンスで階層分類する仕組みは？辞書なしで可能？

A：単語共起ベースのダイナミッククラスタリングで処理している。

Q：コラボレーション機能におけるエキスパートの意味は？

A：２種類の定義あり。１つは自己申告のプロファイルに基づく定義。もう１つは、ある分野に興味をもつ（その分野に関係する文書を検索している）人の集まり。

Q：エキスパートの検索履歴はクエリレベルか？

A：検索クエリのキーワード＋閲覧ページ内のキーワード

Q：クラスタリングの処理時間は問題にならないのか？

A：あまり問題にはならない。

【FAST ESP について】

Q：エンティティ抽出は、社内文書を対象としても十分な精度が得られるのか？（カスタマイズするためのツールがあるのか）

A：人、場所、日付等は共通の抽出処理。部品名、部品番号等には、応用ごとのカスタマイズあり。固有名詞抽出には、辞書を用いる抽出と文脈からの類推による抽出の２種類ある。

Q：エンティティ抽出の言語依存性は？

A：辞書／正規表現での抽出も可能であり、言語依存性はない。

Q：検索エンジンのカスタマイズは、システム管理者等が行うもので、エンドユーザが行うものではない？

A：基本的にはその通りである。場合によっては SNS 的な運用もありえる。

Q：どのような運用が良いのか。草の根的な運用か管理者をきちんと置く運用か。

A：草の根的な運用はまだない。企業内検索ではブースト＆ブロックのユーザが少ない。

【情報融合ソリューションについて】

Q：ポータルでの情報融合表示は難しいのではないか。画面は事前にカスタマイズするのか。

A：BI や DWH を文書単位に展開するイメージ。コンサルティングもあわせ、より業務に密着した検索基盤を構築する。

【適用事例について】

Q：導入時の評価はどのようにされているのか。

A：アンケートやログによる利用頻度等により実施。

Q：導入決定者への説得材料は何か？

A：これまでの例では、熱意のある人が多くトップダウンに導入できた。

Q：自動抽出できなかったメタデータは？

A：検索をしやすくするための補助情報などは自動抽出困難。

Q：Know Who 検索のユーザは知財部門か技術者か？

A：技術者がユーザである。技術が細分化しており、複数部門での開発内容重複による無駄を省きたいというトップの発想があり。

【その他全般について】

Q：インターネット検索とイントラネット検索の違いは何か？

A：セキュリティ（アクセス権）とコンテンツの雑多性（内容ではなくファイル形式など）である。インターネット上の文書は外部に見てもらうための文書だが、イントラネット上の文書は内部メモ（自分が参照する備忘録的なもの）であることも多いという点も異なる。

Q：非公開データは共有しにくいのではないか。

A：「共有できる文書はない」と言われるケースもあり。いっぽう、企業内でブログや SNS を活用して情報共有をはかろうという動きもある。

Q：パーソナライズはどの程度使用されているか。

A：ブックマーク的なパーソナライズは使用事例あり。プロファイルベースのパーソナライズも可能だが使用事例はなし。

Q：プロファイルベースのパーソナライズとはどのようなものか。

A：業務ごとにアプリケーションが異なる。部門ごとに異なる検索結果が優先されるというもの。

Q：知識共有とコンプライアンストレーサビリティとは、検索として異なるものか。

A：異なると考える。前者はマーケティング等がターゲットで試行、気づきを与えるもので、後者はよりピンポイントの検索が要求される。

（相川勇之）

4.2.3　PC ブラウザ履歴活用・検索技術

PC 上でのブラウザ閲覧履歴を、ユーザの過去の状況として活用する技術として、森田哲之委員からの説明に基づき、NTT および NTT レゾナントが公開実験を行った PC ブラウザ履歴活用・検索技術メモリ・リトリーバ、および現在（2008.10）実験中の MyBoom！実験について紹介する。

(1)　メモリ・リトリーバとは？

PC 上の詳細なブラウザ閲覧履歴を利用して、過去に閲覧して有用だった複数の Web ページ、それら Web ページを発見した経緯、Web ページから得た知識を再発見することを助けるツールである。

ブラウザ履歴は、タスクトレイに常駐するソフトウエアで自動的に蓄積する。専用の閲覧ソフトウ

エアを用いることで、任意のキーワードに関連した履歴を、簡単でビジュアルに探し出すことができる。

メモリ・リトリーバは、goo ラボ（http://labs.goo.ne.jp）で 2007 年 1 月～6 月まで公開実験していたものである。

(2) 履歴の蓄積方法

ブラウザ履歴は、タスクトレイに常駐するソフトウエアで自動的に蓄積する。ブラウザに組み込んだアドオンと、OS のウィンドウメッセージを取得することにより、ユーザがいつも何気なく行っている操作だけで、自動的に履歴を蓄積することができる。蓄積する履歴の種類は、URL、個々の Web ページをアクティブに閲覧していた時間の長さ、マウス操作、コピー、印刷、インターネットの検索エンジンに入力した検索語、クリックしたアンカーテキスト、マウスで選択した文字列、Web ページのタイトルなどである。

(3) 履歴の閲覧方法

蓄積した履歴を利用して、さまざまな履歴検索ができる。最も特徴的な機能のひとつは、任意のキーワードに対して、そのキーワードに関係した Web ページを集中して見ていた時間を探し出すことができることである。たとえば、昼休みに「携帯」について調べて良い Web ページをいくつも見つけていたとする。後日、閲覧ソフトウエアに、「携帯」というキーワードを入力して検索すれば、その昼休みの「時間」（12:20～12:37 など）を見つけだすことができる。

期間スコアというスコア付けによって、「時間」はランキングされており、上位に表示された「時間」は重要度が高いと推定されたものである。

表示された「時間」をクリックすると、その「時間」に閲覧したいくつもの Web ページが、サムネイルで時系列に表示される。閲覧時間が長いなど、重要と推測された Web ページは、赤やオレンジの枠で強調表示される。インターネット検索エンジンに入れたキーワードを表示することで、なぜこれらの Web ページを閲覧したのかといった自分の行動の背景を理解することができる。よいと思った Web ページ中の文章をマウスで選択しておけば、閲覧ソフトウエアで表示した時にその文章がハイライトされるので、Web ページ自体を再発見するだけでなく、その Web ページから得た知識（例えば、自動明るさ補正機能が付いた携帯である。）を容易に思い出すことができる。

(4) MyBoom！サービスへの展開

現在（2008.10）MyBoom！という新しいサービスを goo ラボにおいて公開実験している。

メモリ・リトリーバは、自動的に保存した履歴を活用して、有用な Web ページ、閲覧の背景、獲得

した知識などの思い出しを支援するソフトウエアであった。

MyBoom！は、履歴を解析し、履歴から推測される興味語を抽出して新たな情報の獲得を支援するサービスである。

このサービスでいう興味語とは、ユーザが興味を持っていると思われる意味のある文字列である。文字列は、WikiPedia の見出し語を候補としている。詳細な履歴やリンク構造の解析等から興味語の確からしさを推定する。

興味語は随時更新される。PC での閲覧履歴から抽出した興味語を、ブログ検索や画像検索へのリンクと共に携帯上に表示するので、シームレスに興味語に関する検索を続けることが容易になる。

また、自分が興味を持ったキーワードをビジュアルにブログ投稿できる機能もあり、自分の興味を公開して楽しむこともできる。

(5) 主な質疑応答

Q：アンカーテキストとは何か？

A：WEB ページにアクセスするためにクリックしたテキストである。そのユーザにとっての自動的に抽出できるタグと考えられる。

Q：注目した部分にアンダラインが引かれるが、どうやって抽出しているのか？

A：マウスでの文字選択を記録して、注目部分とする

Q：テキストの入力以外に、検索方法はないのか？

A：時間を指定して、閲覧履歴を表示することが可能

Q：キーワードを思いつかない Web ページを検索するときに有効ではないか？　たとえば、画像として覚えているページなど。

A：そうである。

画像ならイメージできる、名称の思い出せないホテルだが、（観光地名など）あいまいなキーワードなら覚えているなどの正確なキーワードが指定できない場合に有効だと考えている。

Q：保存したくない Web ページもあると思うが、すべて記録されてしまうのか？

A：初期設定では、プライバシーの観点から、HTTPS のページは保存しないように設定されている。また、URL のブラックリスト、ホワイトリストを設定して、記録する履歴をユーザがコントロールすることができる。

Q：「時間」を検索するというが、どのように「時間」がランキングされるのか？

A：入力した検索語と適合度が高い Web ページに対して、長い時間見ていた、印刷したなどの行動を起こしているかを判定し、積算することによって Web ページごとのスコアを計算する。次に、それら Web ページの連続性を考慮して、「時間」の開始時刻と終了時刻を決定し、「時間」内に見た

Web ページのスコアを積算して、ランキングしている。

Q：サムネイルや Web ページが表示されるが、履歴はどこに保存されているのか？

A：メモリ･リトリーバでは、プライバシーの観点から、すべてローカル PC 上に保存している。

Q：閲覧時間を抽出しているが、大画面で複数の Web ページを並べて閲覧していた場合は、閲覧時間を抽出できるのか?

A：本ツールでは、アクティブになっているブラウザに最前面表示されている時間を抽出している。よって、非アクティブだが画面の最前面に表示されている Web ページは閲覧しているとは判定していない。視線認識技術などを用いることで正確な判定ができると考えられるが、そのために装置を設定する必要があり、現実的なサービスとしては難しいと考えている。

（森田哲之）

4.2.4 ライフログ

東京大学 相澤清晴先生へのヒアリングに基づき、ライフログの研究動向、事例について紹介する。

(1) ライフログ研究の歴史

現在、我々の身の回りには、様々な情報機器があふれている。この流れに伴い、「人の生活」の情報処理が注目されるようになった。ライフログとは、生活の有り様の記録である。ライフログの研究は、90 年代半ば頃から始まり、特に 2000 年頃から関連研究が散見され始める。これらの研究には、

1. 生活の中で操作した情報の蓄積：e-mail、Web、TV、機器操作履歴など
2. 生活の有り様のデジタル化：体験のデジタル化、体験の情報処理、体験の共有

の 2 つのアプローチが混在している。

これまでに開催されたライフログ関連の国際ワークショップとしては、以下がある。

- Int. Workshop on Memory and Sharing of Experiences 2004 (Pervasive 2004 併催)
- ACM Workshop on Continuous Archival of Personal Experiences (CARPE 2004,2005,2006) (ACM MM 併催)
- IEEE Workshop on Electronic Chronicles e-CHRONICLES 2006 (ICDR 併催)
- ACM Workshop on Human Centered Computing 2007 (ACM MM 併催)
- ACM Conference Multimedia Information Retrieval (MIR2008) (ACM MM 併催)

(2) ライフログの事例

ライフログの事例を表 4.2.4-1 に挙げる。生活の有り様は見えにくく、記録して、見えるようにするだけでも生活の大きな改善につながる。表 4.2.4-1 の事例のように、身近なログ取りが始まっている。

表 4.2.4-1　ライフログの事例

名称	記録する情報	説明
CIMX	電力使用量	工場・店舗の省エネのため、電力をモニタリングし、可視化して、記録解析を行う。http://www.cimx.co.jp/
食事ログ	食事	ダイエット目的で、手帳に食べたものを記録。岡田斗司夫著「いつまでもデブと思うなよ」(2007 年 8 月，新潮新書)
Wii Fit	体重等	体測定、トレーニングと運動貯金の記録、家族データの可視化。
ブログ、Twitter	日記、つぶやき	Twitter は、「今何してる？」の問いに対して、個々のユーザがつぶやきを投稿するミニブログ。http://twitter.com/
キセキ	位置情報	携帯の GPS を使い、10 分以上滞在したところを自動的にメモし、地図付き日記ができる、NTT の Web サービス。 http://lifelog.machi.goo.ne.jp/
ねむログ	睡眠時間	睡眠時間を記録する Web サービス。 http://www.nemulog.jp/
Nike+iPod	運動	Nike の靴と iPod を使ってランニングすると、靴のセンサデータを iPod に送信し、記録を Web で管理できる。他のランナーの状況もわかる。http://www.apple.com/jp/ipod/nike/
Run&Walk		au ケータイを使ってランニングデータを記録し Web で管理できる。http://run.auone.jp/

ライフログは、

- ずっととっておきたい出来事を「保存」
- 意識していなかったことを「発見」
- 忘れかけていることを「追想」
- 行動パターンを「分析」
- 生活の改善のために「予測」

に役立つと期待され、省エネ、健康管理・医療、高齢者のケア、子供の見守り、探し物、サーベイランス、マルチメディア日記、マルチメディア記録の共有等の応用が見込まれる。初期の頃は、「なんでもログをとってみよう」という汎用指向だったものが、最近は特定の応用を指向する流れになってきている。

(3)　相澤研究室の研究紹介

以下に、相澤研究室でこれまで行われた研究を 4 つ紹介する。

①　ウエアラブルイメージングシステムによるライフログの取得と処理

初期の研究（99 年～）。 Sensecam（Microsoft）ベースのシステムで、学生が機器を装着して旅行し、実験を行った。

- 映像（カメラ、ディスプレイ)、音響、GPS、センサ、ジャイロを装着して行動
- 携帯からメールを送ると、地図上のおおよその場所にタグがふられる

- 「Sensecam」は、センサ（加速度センサ、温度計、照度センサ、赤外線センサ）の値をトリガにしてシャッターを切るカメラで、10 時間で約 4000 枚を撮影
- 行動をイベント化し、自動でラベル付与
- 移動度：停留、低速（歩き)、中速（自転車)、高速（交通）
- 停留～低速時の行動：歩行、デスクワーク、食事、静止、就寝など
- 1 名の 40 時間のログデータを取得
- 人手で付与した正解と比較して評価。Recall 80％、Precision 85%の結果

② ユビキタスホームでのライフログ取得とその検索

NICT のユビキタスホーム（04 年築）での実験。完全な 1 軒の家になっており、カメラ、マイク、センサを多用した「生活を記録してくれる家」。古典的ライフログの例である。

- 家の中に、カメラ 17 台、マイク 36 台、その他のセンサ（床圧力センサ、ドアセンサ、RFID 等）を持つ。取得ログは、200GB/day もの大量のマルチメディアデータ。
- 床圧力センサ（18cm 間隔の 2 値データ）が最大の武器。データ量が小さい。
- 被験者家族は、ホームに 2 週間住み、半年後にユーザスタディを行った。
- 被験者家族からのフィードバックコメント：
  - 半年たって忘れていたが、映像を見て鮮明に思い出した。
  - 孤独なシーンは後から探さない。家族と一緒に何かをやっているシーンを見たい。
  - 活用したいシーン：祖母の介護のための行動チェック、なくしたものの発見、なかなか取れない普段の記録、子どもや自分の生活習慣のチェック等
  - 生活全部を連続的に取っておきたい。ただし、記録する場所は限定したい。（リビング、ダイニング、キッチンなどの家の公共スペースに限る）
  - 意外な発見の例：
    - ✧ 子供の目覚めた様子を見ることができて嬉しい
    - ✧ なぜか子供が 2 回朝ごはんを食べている日があった！
    - ✧ 食事時間がたった 10 分だった
    - ✧ 家で仕事をしていたつもりはなかったが、実は仕事をしていた

③ フードログ

1)カメラで食事の画像を撮影しアップロード→2)写真の中から、食事の画像だけ分類→3)食事バランスを推定→4)可視化する Web サービス。衣食住のうち、食は多くの人が興味を持っており、健康の側面からも重要である。特定応用志向のライフログの例である。

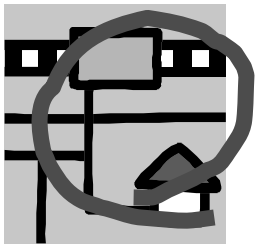
- 1 皿ごとに、フードピラミッドに自動分類
  - 主食、副菜、主催、果物、乳製品の 5 分類
  - 分類は 80%ぐらいの精度で可能。肉、魚といった細かい認識はしないが、ブロック内のグローバルな特徴をつかむ（色のヒストグラム等。料理によらないもの）。機械学習を利用。
- 統計と次の食事の提案を行うアプリができる。運動ログとの連携も考えられる。
- 近日中にサービス公開予定（foodlog.jp）

④　GPS トラックログの検索インタフェース

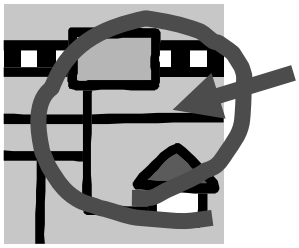
取得した GPS トラックログを、のちに検索するにあたり、「どこからどこまで動いた」というクエリをどう表現するか？ 地図をもとにしたクエリ表現方法を提案。

- 5 種類のスケッチのパターンを用意

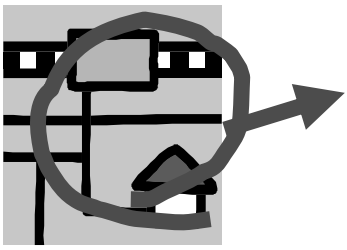
1)囲む
（このエリアにいたとき）



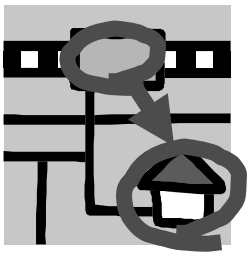
2)囲んで中に矢印
（このエリアに入ったとき出発点は不問）



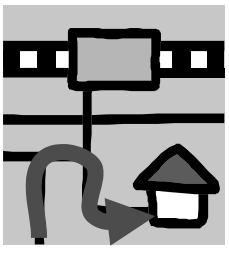
3)囲んで外に矢印
（このエリアからどこかに行くとき）



4)出発点と終着点指定



5)経路指定



- 研究室の学生が 6 ヶ月間、GPS ログを収集
- 検索のバリエーション例：
  - ✧　東京から新潟に行ったログ。いつ行ったのか？
  - ✧　1 月 10 日のサマリを見たい。どこに長く滞在したか。
  - ✧　東京から外に出た今年のログを全部見たい
- 最近は、特定アプリ指向になってきており、UI まわりの研究もいろいろ行っている。

(4) ライフログの今後

ライフログが実際に人の役に立つためには、受容性の高いライフログシステムである必要がある。

- 多様なログの集約（好きなログだけ取ればよく、ログの種類は個人ごとに異なる。また既存サービスとの連携が望ましい）
- 可視化のためのログ処理、コンテンツ処理
- 直観的なインタフェース等が、今後の課題となる。

(5) 主な質疑応答

Q：ライフログ関係の論文、国際会議での発表について。米国では、産業や軍需に使わない研究はあまりしなさそうなイメージがあるが、どういう研究目的か？

A：研究が始まった当初は,

1)モバイルコンピューティングのアプリの一つとして提案

90 年代半ば、”ForgetMeNot”という研究所内 1 行日記のようなものが作られた。誰とすれちがった、誰とメールした等を記録する。

2) DB の発展系として提案

2000 年頃から、マイクロソフトの「マイライフビッツ」プロジェクトが始まった（ゴードン・ベル氏)。 DB の観点から、自分の見た Web や TV や手紙等、なんでもスキャンし、ひたすらログを取る。かなり大きいデータ規模だった。しかし、最近では、軍の研究にもなっているようだ。デモでは、デジカメ写真を整理して見せる DB 作りのようなものになっていたが、兵士がカメラで撮った映像をログとしてためていくようなものを意図しているらしい。

Q：（何か別のデータと組み合わせることを考えると）ログの基本は、時間と場所か？

A：時間と場所は must.

Q：どういう情報を取り、それからどういうアプリができるかは、マトリックスができて、「これをやりたいユーザは、この情報とこの情報を収集しなさい」みたいな感じにはならないのか？

A：まだ今は、全貌が整理できている状況にはない。アプリケーションドリブン。もしくは、今はこんなログがある、という感じ。

Q：先に目的を明確にしないと、何をログにとったらいいのかわからないのでは？

A：まず、食べたら必ずカメラで撮る、と決める方法もある。なお、ライフログとは違う名前で始まろうとしているのが、多くの人のログを統計処理して何かを発見する”Social Signal Processing”, ”Network of Excellence”の研究。ヨーロッパで大規模プロジェクトが始まる模様。また、MIT の実験で、1 年間、学生の携帯電話の基地局ログをとり、100 人×1 年分のログから "Latent Topic Discovery"を行ったという。しかし、専攻もバラバラな 100 人から平均的にとれて

きたイベントは、家を出て学校にいく等の当たり前のものだった。同じデータを使うにしても、母集団を限っていくと、面白い発見になるかもしれない。

Q：ライフログは、レコメンデーションに使えそう。Amazon のおすすめも、ある意味ライフログ。国際会議では、そういう話はあまり出てこない？

A：出てこない。

Q：可視化方面の話になる？

A：可視化の話もまだ不十分。昔ながらのライフログは、イベント検出とかそういう話が多い。一方、ユビキタスの人たちは、レコメンの話が大好き。

Q：工場・店舗等で応用する研究はある？

A：省エネの事例は、工場・店舗向け。ビジネス利用という意味では、サーベイランスがある。ワンフロア丸々、サーベイランスのデータを取って、絶えず人の動線ログを取る研究はある。人の検出、画像処理、あるいは、ミーティングログ。店舗のお客さんの動線は、言われてはいるが、実際利用されているのを見たことはない。

Q：ライフログは広いというのは、少し違和感がある。パーソナルあるいはファミリー内のクローズドなイメージを持っていたが、世の中的にはそうでもない？

A：それがベースライン。自分が見えないログは NG。ライフログの狭い定義は、個人で全部管理できるものが基本。

Q：大学の研究は学術的な価値を問われると思うが、今後、技術的にはどこを狙っている？　それとも、まずはアプリからと思っている？

A：画像処理は最も重要な課題。現在はどの食べ物かまでは判定できず、たとえば 1 年間にラーメンをどれだけ食べたかは判定できない。その辺を細かく行うのが一つの課題。また、検出結果の統計をどう見せるかも課題。結果のバランスをどう人間にフィードバックするか。たとえば、食べ過ぎがずっと続いているが、運動しているから良しとする等、複数のログの組み合わせ処理の工夫も面白いと思っている。

（志賀 聡子）

4.3　状況依存型サービスの事例調査

本節では、現存する状況依存型サービスをインターネット上で調査し、いくつかの特徴素性の観点から分類・整理したのでその結果について述べる。まず、検索エンジンを用いて「状況依存」、「状況適応」、「ユーザの状況」、「Adaptive」などの検索キーワードを手がかりとして、状況依存型サービスと考えられるものを人手で抽出した。次に、それらのサービスが持つ特徴素性を以下の 6 つの大きなカテゴリに部類し、表 4.3-1 のようなマトリックスとしてまとめた。表 4.3-1 の各行には収集したサ

ービスが、各列にはカテゴリに分類した特徴素性が記してある。この表をさらに分析して、以下のような傾向を洗い出した。特徴素性の後の括弧内の数字はサービス数を表わす。

概要

- サービスが持つ素性数をみると、「GPS（19)」、「携帯電話（20)」、「情報共有（16)」、「ライフログ（14)」、「位置情報（22)」、「地図（15)」、「サービス中（33)」などが突出しており、これらのキーワードからモバイル環境での位置情報を使ったサービスが多いことがうかがえる。

デバイス

- 端末に関しては「携帯端末（5)」に対して「携帯電話（20)」が圧倒的に数が多い。最近では iPhone に代表されるスマートフォンなどのデバイスも普及し始めているようだが、依然として携帯電話を対象としたサービスが優勢であることがうかがえる。
- 「携帯電話」と「GPS」を共に利用したサービスが 16 件もあり、この両者は密接な関係にあることがわかる。これは携帯電話に GPS が標準的に搭載されるようになってきたことが大きな要因であろう。
- 位置取得に関しては「無線 LAN（5)」に対して「GPS（19)」が優勢である。無線 LAN の利用は 5 例とまだまだ数は少ないが、その内訳が研究開発 1 件、実証実験 1 件、サービス中 3 件であり、さらに skyhook や PlaceEngine などのプラットフォームがサービスインされていることや iPhone などの出現により、これから増加してくる可能性がある。
- GPS を利用したサービスとしては、「ナビゲーション（4)」や「情報共有（7)」を提供するものが目立つ。情報共有では、場所に関連した店などの情報を共有するのに使われている。

成熟度

- Web 上から調査したためにサービス中のものが 33 と圧倒的に多いが、実証実験中のものも 7 件ある。
- サービス中のものでは、位置情報に関連する素性が頻度で上位に来ており、それに続いて「ログの可視化」、「フィードバック」などのユーザへのフィードバックに関連する素性が目立つ。ブログやログを解析し、レコメンドなども含め、ユーザになんらかのフィードバックをするサービスが多いようである。
- 状況情報活用のサービスは実証実験が活発に行われている。技術的な基盤（サービスプラットフォームとしての携帯電話・ネットワーク接続環境・決済プラットフォーム・センサなど）が揃ってきて，状況情報を利用してどのようなサービスが実現可能かを模索している段階である。しか

しまだ有効なサービスが見つかっていない状況である。実証実験を通じて今後どのようなシーンでどのようなユーザメリットがあるのかが見出されるかどうかが今後の普及への一つの条件となるであろう。

基礎技術

- 「ログの可視化（5)」と「ブログ解析（3)」が比較的多く利用されている。「ログの可視化」は「ライフログ」や「フィードバック」と共に利用され、ユーザのログを分析した結果をユーザにフィードバックするのに使われる傾向にある。

機能・目的

- 「情報共有（16)」が多いのは世の中の Web 2.0 的な動向を反映している。
- 「タスク管理」、「リマインダー」、「リコメンド」なども位置情報と関連付けて、その場所に行くとやるべきことを教えてくれるなどのサービスが出始めている。
- 「ナビゲーション」は当然のことながら「地図」、「GPS」、「携帯電話」といった素性と一緒に利用されることが多い。

状況情報

- 状況情報としては「位置情報（22)」が圧倒的に多い。
- 「ライフログ（14)」に関するサービスも多い。サービスの中で収集される情報としては「食事ログ（4)」、「ブログ（4)」「位置情報（3)」、「地図（3)」などがある。「フィードバック（6)」、「ログの可視化（5)」などと一緒に利用されていることを考えると、ログの分析をおこなった結果を可視化してユーザにフィードバックするというシナリオが見えてくる。また、数は多くないものの「運動ログ（1)」、「睡眠時間（1)」などのキーワードから健康関連のログの取得・分析もこれから増えてくることが予想できる。
- 分析対象・結果
- 「ブログ」についても位置情報との連携が多い。共に利用される「情報共有（7)」、「位置情報（6)」、「地図（4)」、「GPS（3)」、「携帯電話（3)」などのキーワードから、携帯電話に内蔵された GPS を使って、ブログを更新した場所などの情報を自動的に付与するなどのサービスが見えてくる。

（徳永健伸）

表4.3-1 状況依存型サービスと特徴素性（1/8）

			デバイス									成熟度			基礎技術					機能・目的									状況情報					分析対象・結果							
	サービス名	ディスクリプション	GPS	Webカメラ	センサー	ロボット	無線LAN	携帯端末	携帯電話	人体通信	skyhook	研究開発	実証実験	サービス中	サービス終了	タスク管理	ブログ解析	ログの可視化	経験マイニング	環境知能化	タスク管理	リマインダー	ナビゲーション	レコメンド	行動予測	検索	感性制御	情報共有	ライフログ	運動ログ	食事ログ	睡眠時間	位置情報	ブログ	写真	ニュース	TODO	地図	年表	フィードバック
1	山手線乗車時の位置情報を自動表示するiPhone/iPod touchアプリ	山手線乗車中，電車の進行方向や路線上の現在位置を自動的に表示したり，最寄駅の駅構内情報を確認できる．降りる駅を設定することで，乗り過ごし防止機能を使うことができる． http://japan.cnet.com/news/media/story/0,2000056023,20373031,00.htm					1	1				1										1	1		1								1					1			
2	ダイエット応援サイトカロリヲ	体重，体脂肪率，ボディサイズを入力するだけで，ユーザのダイエット状況をグラフ化する． http://www.calorio.jp/f_a/							1					1				1										1	1		1									1
3	いつまでもデブと思うなよ	食事の記録を付けることで食生活を自己管理する「レコーディングダイエット」をNINTENDO DSで実行するソフトウェアである． http://www.nintendo.co.jp/ds/software/cebj/												1				1											1		1									1
4	睡眠時間を記録するサイトねむログ	起床時間と就寝時間を入力すると，ユーザの睡眠時間を記録し，統計情報をグラフ化する． http://www.nemulog.jp/							1					1				1										1	1			1								1
5	アップル － Nike + iPod	Nike+ランニングシューズとNike + iPod Sport Kitまたはセンサーがあれば，走り終わった後に，iPod nanoやiPod touchをMacかWindows PCにつなぐだけで，ワークアウトデータがnikeplus.comに自動的にシンクされる． http://www.apple.com/jp/ipod/nike/						1						1				1										1	1	1										1
6	The flash diet	自分が食べた物の写真を撮っておく人は減量するという研究報告である． http://www.dailymail.co.uk/health/article-1052234/The-flash-diet-Taking-photos-meals-helps-slimmers-lose-weight.html																										1		1										1
7	あすけん-あすの夢を健康でささえる研究所	食事・運動を記録すると，自動的にアドバイスが表示される食事改善プログラムである． https://www.asken.jp/wsp/												1				1							1				1		1										1
8	“街なか”のお奨め情報配信サービス『Magitti』－大日本印刷	ユーザの携帯端末から得られるコンテキスト情報を利用して“街なか”における「食べる」「買う」「見る」「遊ぶ」といった行動からそのユーザが行う可能性が高い行動を予測し，個々のユーザに最適な情報（店舗，施設，イベントなどの情報）を携帯端末に配信するサービス．コンテキスト情報として場所，時間，天気といったユーザの状況に関する情報と，ユーザが事前に登録した嗜好，閲覧履歴を利用する． http://www.dnp.co.jp/jis/news/2008/080425.html	1					1	1				1												1	1							1								

表4.3-1 状況依存型サービスと特徴素性（2/8）

			デバイス									成熟度			基礎技術					機能・目的								状況情報					分析対象・結果						
	サービス名	ディスクリプション	GPS	Webカメラ	センサー	ロボット	無線LAN	携帯端末	携帯電話	人体通信	skyhook	研究開発	実証実験	サービス中	サービス終了	タスク管理	ブログ解析	ログの可視化	経験マイニング	環境知能化	タスク管理	リマインダー	ナビゲーション	レコメンド	行動予測	検索	感性制御	情報共有	ライフログ	運動ログ	食事ログ	睡眠時間	位置情報	ブログ	写真	ニュース	TODO	地図	年表	フィードバック
9	空間ロボット RoomRender ー 日本SGI	会議室や住宅の部屋などの空間をロボットと捉えて，さまざまな電子機器を一括制御することで，人が心地よいと感じる空間演出を提供するシステム． 感性制御技術（ST - Sensibility Technology）により人の感情や感性をコンピュータが認識し，それに応じて空間内の機器やセンサを連動させることで心地よさを演出する． http://www.sgi.co.jp/robot/roomrender/index.html				1							1						1								1												
10	Life-X（ライフ・エックス）－ソニーマーケティング	ブログや画像，動画，ブックマーク，メモ等のウェブ上のコンテンツを一元管理するオンラインサービス．アイコンがスナップ写真のように並べられる「タイムラインビュー」や地図上に表示する「マップビュー」があり，友人とコンテンツを共有できる機能もある． http://life-x.jp/											1														1	1				1	1	1		1	1	
11	ログピ｜ログをピッてするライフログサービス － paperboy&co.	日々の出来事やメモを簡単に記録するライフログサービス．同じテーマをもとにユーザ同士でコミュニケーションする機能や，自分のメモ用に非公開にする機能を持つ． http://logpi.jp/											1														1	1					1					
12	携帯ライフログカプセル ”たむたま”｜エフスパイラルコーポレーション有限会社	携帯から送信した写真やコメントに日付をつけ”玉（カプセル）”に保存しておき，利用者が設定した未来の日付になったときに送信してくれるサービス． http://www.fspiral.co.jp/jp/biz-sns/sns50.php							1				1									1						1											
13	5つ星ミニブログ・ライフログ［staaaaar!（スター！）］	テキストボックスと5つ星だけのシンプルなライフログツール．携帯GPS対応で，訪れた場所に対するコメントと5つ星評価をつけることができる． http://staaaaar.jp/	1						1				1														1	1				1	1			1		
14	キセキ ver.0 ーケータイGPSで簡単日記自動作成！ー goo	携帯電話のGPS機能を利用して，訪れた場所の“住所”と“時間”を自動的に記録し，その記録に対して「食事に行った」「買い物をした」などのコメントをつけることで，簡単に行動日記を作成することができるサービス． http://lifelog.machi.goo.ne.jp	1						1				1															1				1				1		
15	AllofMe｜Timeline yourself	自動的にPC内のデジタルコンテンツ（写真，動画，ブログ，文書等）をスキャンして，個人の年表を作成してくれるサービス． http://www.allofme.com/											1															1										1	
16	NetTansor Web －バンダイ	Webカメラを搭載したホームロボット．自宅のLANに接続しておくと，留守中の自宅の様子を撮影して，ロボットが自動でブログに書き込みをしたり，携帯電話に写真メールを送信する． http://www.roboken.channel.or.jp/		1		1							1														1	1					1					

表4.3-1 状況依存型サービスと特徴素性（3/8）

			デバイス									成熟度			基礎技術					機能・目的									状況情報					分析対象・結果							
	サービス名	ディスクリプション	GPS	Webカメラ	センサー	ロボット	無線LAN	携帯端末	携帯電話	人体通信	skyhook	研究開発	実証実験	サービス中	サービス終了	タスク管理	ブログ解析	ログの可視化	経験マイニング	環境知能化	タスク管理	リマインダー	ナビゲーション	レコメンド	行動予測	検索	感性制御	情報共有	ライフログ	運動ログ	食事ログ	睡眠時間	位置情報	ブログ	写真	ニュース	TODO	地図	年表	フィードバック
17	位置情報プラットフォーム「Fire Eagle」	ユーザーが自分の位置情報をWebで活用できるオープンプラットフォーム．米Yahoo!の開発．ユーザーは自分の位置情報をWeb上に保存して管理できる．情報の公開相手（人，アプリケーション）や，公開レベル（例えば，国，州，通り，など）も自由に設定できる．開発者はその位置情報を活用したサービスやアプリケーションを開発できる．												1														1					1							
18	位置情報確認サービス「今どこ？」	携帯電話のGPS機能を利用して，友人・知人がどこにいるかを知ることができるサービス．相手の携帯アドレスにメールを送信し，相手が位置を知らせることを了承して指定のURLにアクセスすると位置情報を取得できる．Yahoo! JAPAN のYahoo! 地図 Web サービスを利用している．2007年の Yahoo! Japan Web API コンテストで Web+DB Press 賞を受賞．http://api-contest.yahoo.co.jp/prize/10.html	1						1					1														1					1				1			
19	位置表現抽出・管理サービス「LocoSticker」	ブログ中の位置表現を頼りに位置情報を自動的に付与することで，地図や位置キーワードから付近のブログを簡単に探すことができる．												1			1									1		1					1	1						
20	場所にメッセージを残す，携帯用すれ違い系サービス「loc8r」	コメントと現在地の位置情報をあわせて投稿できるコミュニケーションサービス．今いる場所の感想や写真などを投稿すると，携帯電話のGPSや無線LANによる簡易位置情報から取得した現在地の情報を付加する．コメントが書かれた場所が距離的に近いトモダチを近い順に参照することもできる．PlaceEngineを開発したクウジットと，So-netが共同で実施している．	1						1					1														1					1	1						
21	モバイル機器の位置情報を蓄積・統合したホットスポット表示「CitySense」	ユーザーの行動履歴(GPSによる位置情報)を蓄積・統合して，どのエリアが今賑わっているかを表示するSense Network社のサービス．自社開発の，モバイル機器からリアルタイムの位置データを分析利用するソフトウェアプラットフォーム Macrosense を利用している．http://www.100shiki.com/archives/2008/07/citysense.html	1						1					1														1					1							
22	ニュースを地図にマッピング「MappedUp.com」	ニュースサイトのRSSを常時確認して世界のニュースを地図にマッピングする一種のフィーダーとしてのサービス．会員にはパーソナライズや共有の機能も提供している．http://www.100shiki.com/archives/2006/07/_mappedupcom.html																															1			1		1		
23	携帯の使用情報からユーザの現在状況を把握「Jaiku.com」	現在地(GPS利用)や，携帯の電源オン・オフ，携帯操作の有無と間隔などの情報を，特段の入力操作なしに公開することができる．ユーザの現在の状況をおおよそ把握することができるため，友人とのコミュニケーションに有用であるとしている．現在はGoogle傘下となっている．http://www.100shiki.com/archives/2006/07/_jaikucom.html	1						1					1														1					1							

表4.3-1 状況依存型サービスと特徴素性（4/8）

			デバイス									成熟度			基礎技術					機能・目的									状況情報					分析対象・結果						
	サービス名	ディスクリプション	GPS	Webカメラ	センサー	ロボット	無線LAN	携帯端末	携帯電話	人体通信	skyhook	研究開発	実証実験	サービス中	サービス終了	タスク管理	ブログ解析	ログの可視化	経験マイニング	環境知能化	タスク管理	リマインダー	ナビゲーション	レコメンド	行動予測	検索	感性制御	情報共有	ライフログ	運動ログ	食事ログ	睡眠時間	位置情報	ブログ	写真	ニュース	TODO	地図	年表	フィードバック
24	無線LANを利用した位置情報とインターネットとの連携「Loki.com」	無線LAN(Wi-Fi)による位置情報と，webの連携を行うサービスである．Skyhook Wirelessがサービスしている．ブラウザのツールバーをダウンロードするだけで利用できる．利用方法としては，ウェブ検索時に距離が近くのものを優先的に表示させたり，近所の映画や天気の情報を表示させたり，現在地を友人にメールする，などが可能である． (紹介URL http://www.100shiki.com/archives/2006/05/_lokicom.html)					1				1			1													1							1							
25	場所を手がかりとしたリマインダー「Places ToDo.com」	ある場所に行ったら何かをしようという"Places ToDo"の登録と共有ができる．また，携帯の位置情報サービスと連携して，実際にその場所の近くに来たらアラートを出すサービスもある． (紹介URL http://www.100shiki.com/archives/2006/05/_places_todocom.html)	1						1					1								1														1					
26	ブログからのグルメマップ自動作成「30min.ブログマップ」	ブログに書いたグルメ記事を元に，自分専用のグルメマップを自動作成できる機能である．ブログ記事内に含まれる店名や住所，電話番号などの情報を元に，各ブログ記事の該当店舗を特定し，Googleマップ上にマッピングして提供する．												1			1																	1	1				1		
27	名古屋市営地下鉄 無線LAN実証実験	GPSが利用できない地下鉄や地下街において無線LANを用いてユーザの位置を特定することを目的として，名古屋市営地下鉄に設置された無線LAN基地局を用いてユーザの位置特定に関する実用性と応用可能性を確認する実験が行われた．http://locky.jp/pr/nagoyaUG/					1	1					1										1											1							
28	Loopt	GPSの位置情報を利用して友人の位置と現在の状況(通話可能，取り込み中など)を地図上とリスト形式で表示するモバイル位置情報サービス．ユーザーは友人が特定の距離以内に近づいたときにアラートを出すように設定できる．また特定の距離以内にいる友人のグループ全員にメッセージを送ることもできる．（紹介URL: http://jp.techcrunch.com/archives/loopt-to-make-mobile-presence-usable/)	1						1					1															1					1						
29	みんなでつくる国内旅行の旅行記サイト-旅行日和:@nifty	旅行記(写真や記事)投稿や，旅行の思い出・観光スポットの情報を共有するサービス．携帯電話からの投稿では，GPS(位置情報)機能を利用することで現在地の位置情報を自動的に取得し，その周辺の観光スポットの候補一覧が画面に表示されるため，地図から探す手間がなく簡単に旅行記を投稿することができる．（紹介URL: http://www.nifty.co.jp/cs/08shimo/detail/081110003430/1.htm)	1						1					1															1					1	1	1			1		

表4.3-1 状況依存型サービスと特徴素性（5/8）

			デバイス									成熟度			基礎技術					機能・目的								状況情報					分析対象・結果						
	サービス名	ディスクリプション	GPS	Webカメラ	センサー	ロボット	無線LAN	携帯端末	携帯電話	人体通信	skyhook	研究開発	実証実験	サービス中	サービス終了	タスク管理	ブログ解析	ログの可視化	経験マイニング	環境知能化	タスク管理	リマインダー	ナビゲーション	レコメンド	行動予測	検索	感性制御	情報共有	ライフログ	運動ログ	食事ログ	睡眠時間	位置情報	ブログ	写真	ニュース	TODO	地図	年表	フィードバック
30	Remember The Milk	iPhoneのGPS機能を使ってタスクを表示する機能がある．各タスクに場所を設定することで，設定されている場所を訪れた際に場所に対応するタスクを表示できる．http://www.itmedia.co.jp/bizid/articles/0811/07/news077.html	1					1						1							1	1										1							
31	COOLPIX P6000 – コンパクトデジタルカメラ	GPSを内蔵したカメラ．GPS機能を使用して撮影場所の緯度・経度を写真と共に記録する．	1																													1		1					
32	トップーiタウンページ　ハイパーサーチ（β版）	iタウンページの情報を検索する地域情報検索サービス．携帯電話のGPSでユーザ位置情報を取得し，ユーザの現在位置と現在位置を中心にお店や施設の情報を検索し，地図上に表示する．現在位置のエリアにおいて人気のある（他のユーザからのお薦めの）お店も提示する．また，エリアに連動した広告も提示する．	1						1					1										1		1		1					1				1		
33	おでかけ道案内 ｜サービス案内｜駅探	最寄り駅から目的地までの道のりを検索するサービス「おでかけ道案内」駅探がサービス提供．出発地と目的地を入力すると，電車の乗り換え手順と最寄り駅から目的地までの道順を地図で表示する．http://ekitan.com/info/help/odekakemap.html	1						1					1									1														1		
34	au ezナビウォーク	GPS携帯電話で道案内をするサービス「au ezナビウォーク」．KDDI，沖縄セルラー，ナビタイムジャパンが協業でサービス提供．目的地を設定するとGPSによる現在位置からのルートを算出・表示する．乗り換え案内や道路情報とも連携．auの携帯電話の多くにプレインストールされている．2003年サービス開始．	1						1					1									1														1		
35	「Global NAVITIME」	PCや携帯電話から海外の地図検索・目的地までの道案内ができるサービス「Global NAVITIME」．ナビタイムジャパンがサービス提供．28ヶ国でルート案内が可能．単なるルート案内だけでなく，海外で日本語が使える病院の検索などのサービスを組み合わせている．http://global.navitime.co.jp/	1						1					1									1														1		
36	米Ask.com，音声入力に対応した道案内サービス	携帯電話向け検索サイトで音声入力に対応した道案内サービス"Click to Speak"．Ask.com（アメリカ）がサービス提供．音声で現在位置・目的地の入力を行うと道案内情報がテキストメッセージで送られてくる．音声認識は米Dial Directions社の技術を利用．2008年1月3日サービス開始．	1						1					1									1														1		

表4.3-1 状況依存型サービスと特徴素性（6/8）

			デバイス									成熟度			基礎技術					機能・目的									状況情報					分析対象・結果						
	サービス名	ディスクリプション	GPS	Webカメラ	センサー	ロボット	無線LAN	携帯端末	携帯電話	人体通信	skyhook	研究開発	実証実験	サービス中	サービス終了	タスク管理	ブログ解析	ログの可視化	経験マイニング	環境知能化	タスク管理	リマインダー	ナビゲーション	レコメンド	行動予測	検索	感性制御	情報共有	ライフログ	運動ログ	食事ログ	睡眠時間	位置情報	ブログ	写真	ニュース	TODO	地図	年表	フィードバック
37	街を移動する映像と現在地を示す地図を同期した道案内サービス	街を移動する映像を地図を同期して表示するナビゲーションサービス「道ナビ1.0」．ジオバンクがサービス提供．対応エリア内のルートを指定すると，車載カメラで撮影した映像とその位置を示す地図を同時表示するシステムを作成．不動産などの物件情報サイトでの導入を見込んだサービス．2007年5月28日サービス開始． http://www.geobank.co.jp/products/donavi/index.html											1										1															1		
38	街の移動映像とスクロール地図を同期，gooラボで実験	街角の移動映像を地図で検索できる「ウォークスルービデオ」の実証実験．NTTレゾナントがシステム開発および実験運用を行っている．公開Webページで地図からその場所の移動映像を検索する機能を持つ．2007年4月10日から開始． http://map.labs.goo.ne.jp/walkthrough/4/index.php?area=0											1										1															1		
39	カーナビ前にGoogleマップ iPhoneでも利用可能に － ITmedia News	地図サービス「Googleマップ」にドライブルート検索機能を実現．Googleがシステム開発・運用を行っている．ルート上の実際の映像をGoogleストリートビューと連携して表示することが可能．PCだけでなくiPhoneにも対応している．また検索したルート情報を日産のカーナビ「カーウイングス」へ送信する機能も備える．2008年10月29日よりサービス開始． http://maps.google.co.jp/maps	1											1																										
40	報道発表資料：「位置情報を活用したターゲティング情報配信」に関する実証実験を開始｜お知らせ｜NTTドコモ	位置情報を活用したターゲッティング情報発信の実証実験．NTTドコモとJTBパブリッシングがシステム開発および実験運用を行っている．GPS携帯の位置情報や利用者の属性情報を元にユーザの位置・行動履歴・属性にマッチした観光・レジャー・グルメ情報などのコンテンツを提供する．沖縄県と京都府で2008年1月～3月に実証実験． http://www.nttdocomo.co.jp/info/news_release/page/071120_00.html	1						1				1												1	1															

表4.3-1 状況依存型サービスと特徴素性（7/8）

			デバイス									成熟度			基礎技術					機能・目的									状況情報					分析対象・結果							
	サービス名	ディスクリプション	GPS	Webカメラ	センサー	ロボット	無線LAN	携帯端末	携帯電話	人体通信	skyhook	研究開発	実証実験	サービス中	サービス終了	タスク管理	ブログ解析	ログの可視化	経験マイニング	環境知能化	タスク管理	リマインダー	ナビゲーション	レコメンド	行動予測	検索	感性制御	情報共有	ライフログ	運動ログ	食事ログ	睡眠時間	位置情報	ブログ	写真	ニュース	TODO	地図	年表	フィードバック
41	CEATEC JAPAN 2008：目指すのは「空気を読めるケータイ」―ドコモら10社が実証実験 － ITmedia +D モバイル	ユーザの状況に応じて携帯電話が情報提供するサービスの実証実験．経済産業省の情報大航海プロジェクトの活動のひとつでNTTドコモが中心となり10社(NEC，関心空間，東京大学，明治大学，NTTレゾナント，東京急行電鉄，東急エージェンシー，Tカード＆マーケティング，東京データネットワーク)が参加する「マイ・ライフ・アシストサービス」．GPS携帯を通じて10分おきに取得した参加者の位置から参加者の行動パターンを推定し，参加者の行動を先読みすることで適切な情報を提供する仕組みを構築し実証実験を行っている．具体的にはプレビューチャンネル(携帯を開くと自分の行動や欲求にマッチした画面が表示される機能)，Kanshin Airpot(自分が今いるエリアの地域限定情報(口コミ・お勧め等)のうち利用者に合った情報を表示する機能)，ミツケ半島の大冒険(次の行動を誘引するコンテンツを配信することでユーザに行動を起こさせる機能)などの実証実験を実施．							1			1											1	1																	
42	経験マイニングを応用した経験情報検索サービス － みんなの経験	Web上に公開されている個人の経験を整理，分類するための言語処理技術「経験マイニング」を応用し，商品やサービスについての購入経験，利用経験などが書かれたブログ記事を検索できるサービス「みんなの経験」をリリース												1			1		1					1		1								1							
43	センサーチップを搭載した薬による身体反応モニターシステム	センサーチップを搭載した薬剤と，体外で使用する絆創膏に似たパッチにより，患者が薬を服用した場合の生理学的データを自動的に記録できる			1								1															1	1											
44	無線LANを利用した位置情報の推定「PlaceEngine」	PlaceEngineとは，ノートPC やスマートフォンなどの無線LAN(Wi-Fi) 搭載機器で，GPSがなくても現在位置を推定することができる技術である．クウジットの開発・サービスを行っている．地域情報などと連携することで周辺情報を探しやすくすることができる．					1							1																			1							
45	RedTacton	体の表面を伝送経路とし，体の表面に発生する電界を利用する通信技術．人の体の一部が送受信機に触れることで伝送経路が形成され，通信を開始し，離れることで終了する．また，指，腕，足，胴体など体の表面や，靴・衣服を通しても通信が可能である． http://www.redtacton.com/jp/info/index.html			1					1																													

表4.3-1 状況依存型サービスと特徴素性（8/8）

			デバイス									成熟度				基礎技術					機能・目的									状況情報					分析対象・結果							
	サービス名	ディスクリプション	GPS	Webカメラ	センサー	ロボット	無線LAN	携帯端末	携帯電話	人体通信	skyhook	研究開発	実証実験	サービス中	サービス終了	タスク管理	ブログ解析	ログの可視化	経験マイニング	環境知能化	タスク管理	リマインダー	ナビゲーション	レコメンド	行動予測	検索	感性制御	情報共有	ライフログ	運動ログ	食事ログ	睡眠時間	位置情報	ブログ	写真	ニュース	TODO	地図	年表	フィードバック
46	Eye-Fi Explore	Wi-Fiチップを内蔵しており，SDカード対応のデジタルカメラから，無線LAN経由で直接PCや写真共有サービスなどに写真を転送できる無線LAN内蔵SDメモリーカード．撮影した写真に自動的に撮影した場所の情報を追加する「ジオタグ」機能を搭載している．ジオタグ機能には，Skyhook WirelessのWi-Fi Positioning System（WPS）を利用している． http://www.itmedia.co.jp/news/articles/0805/13/news031.html					1				1																						1		1						
		合計	19	1	2	2	5	5	20	1	2	1	7	33	0	0	3	5	1	1	1	4	8	6	4	4	1	16	14	1	4	1	22	9	4	1	1	15	2	6

4. 4　状況依存型サービス関連技術の研究動向

本節では、SIGIR 2008 で発表された論文の中でパーソナライゼーション、ユーザーフィドバックなど、状況依存性に関係ありそうな 7 件の論文の概要について紹介する。

4. 4. 1　Retrieval and Feedback Models for Blog Feed Search

Jonathan L. Elsas, Jaime Arguello, Jamie Callan, Jamie G. Carbonell (pp. 347-354)

(1)　はじめに

本研究の目的

Large Document Model（ブログ単位）と Small Document Model（ブログ記事単位）を比較する。ウィキペディアのリンク構造を用いたクエリ拡張の手法を提案する。

用語の定義

・ブログ（blog)： ブログ記事の集まり。フィード（feed）と同義

・ブログ記事（blog post)： 記事 URL（permalink document） 、エントリ(feed entry)と同義

ブログ検索

・ある話題に長期的な関心を持つユーザが日常的に行う検索行為である。

・以下の(a) 、(b) 、(c)において従来の随時検索（ad hoc retrieval）と異なる。

(a)検索単位

ブログ検索では、文書集合（ブログ）または文書（ブログ記事）のどちらかが情報の単位となる。この点において複数サイトの横断検索（federated search）に似ている。

(b)言語

ブログ記事は新聞のように編集されておらず、会話に近い。コメントによって文書が長くなる傾向にある

(c)脆弱性

ブログは、広告をブログ記事やコメントとして自動配信するスパム（splog）に弱い。

(2)　フィード検索モデル

ブログ検索に似ている 3 種類の検索タスク

(a)エキスパート検索（expert finding)

- エキスパート（ブログ） ＝ メールの集合（ブログ記事の集合）
- クエリに合致するメール集合に対応するエキスパートが検索される

・ **(Hannah et al., 2007)**

・エキスパート検索のモデルをブログ検索に応用

・クエリと独立の結束性（ブログの重心と各ブログ記事の平均距離）を考慮

(b)クラスタ検索（cluster-based retrieval）

・ クラスタ ＝ 文書の集合

・ クラスタを順位付け、上位のクラスタに属する文書を順位付ける

・ **(Seo & Croft, 2007)**: ブログ ＝ クラスタ

各クラスタに属する上位 K 文書の尤度を幾何平均する

(c)分散情報検索（distributed information retrieval） ＝ 横断検索（federated search）

・ 資源選択（resource selection）は分散情報検索の 1 タスク

・ ReDDE **(Arguello et al., 2008)(Elsas et al., 2007)**

文書集合内に属する適合文書数を推定して、文書集合を順位付ける

(2-1) Large Document Model

・ Large Document (LD) Model: ブログを 1 つの文書と見なす

・ TREC 2007 Feed Distillation task で最も成績が良かった

事後確率 $P_{LD}(F|Q)$ に基づいてフィード F を順位付ける

・ P(F): フィードの事前確率（LD と SD に共通）

・ $P_{LD}(Q|F)$: クエリの尤度

$P_{LD}(Q|F)$の計算方法

1. クエリ Q を素性 Ψ_i に分解して、$P_{LD}(Q|F)$を $P_{LD}(\Psi_i|F)$の積で計算する。
2. $P_{LD}(\Psi_i|F)$に重み w_i を付ける。
3. $P_{LD}(\Psi_i|F)$の計算には、Dirichlet-smoothed maximum likelihood estimate を用いる。

Indri 検索エンジンの形式でクエリを表現する

（例）クエリ ＝ "embryonic stem cells" **(Metzler et al., 2005)**

```
#weight(0.8 #combine(embryonic stem cells)          ← unigram query
        0.1 #combine(#1(stem cells)                 ← ordered window query
                     #1(embryonic stem)
                     #1(embryonic stem cells))
        0.1 #combine(#uw8(stem cells)               ← unordered window query
                     #uw8(embryonic cells)
                     #uw8(embryonic stem)
                     #uw12(embryonic stem cells)))
```

・ #M: ターム間に(M-1)のタームが入ってもよい。#1(stem cells)は、stem と cells が連続していることを要求する。
・ #uwN: N タームからなるウィンドウ内に対象のタームが任意の順番で現れることを要求する。#uw12(embryonic stem cells)は、連続する 12 ターム内に、embryonic、stem、cells が現れることを要求する。

(2-2) Small Document Model

SD の利点

・ フィード（F）とエントリ記事（E）の関係をモデル化することができる。
・ フィード全体の言語に近いエントリ記事を優先することができる。

$P_{SD}(F|Q)$を以下の要素に分解する。

・ P(F): F の事前分布 → Feed Prior
・ P(Q|E): クエリの尤度 → Query Likelihood
・ P(E|F): エントリ中心性 → Entry Centrality

以下でそれぞれについて説明する

Feed Prior

先行研究における文書の事前分布

・ homepage finding: 短い URL を持つ文書を優先
・ web search: PageRank
・ SN 比
  - 文書長に対する索引語数の割合
  - 索引語として抽出されない情報を多く含むほど小さくなる

本研究におけるフィードの事前分布 P(F): 2 通り（LD と SD に共通）

・ $\log(N_F)$ → 予備実験の結果、N_F では検索精度が落ちた。
・ 定数

Query Likelihood

・ Jelinek-Mercer の平滑化を使う。
・ パラメタは訓練データから学習する。

Entry Centrality

P(E|F)の効果

・ フィード全体の言語に近い記事を優先する。
・ フィードの長さを正規化する。P(E|F)を考慮しないと長いフィードほど大きくなる。

P(E|F)の計算方法を 2 通り試した

- F 中の全 E に同じ値を与える。
- クエリに依存した中心性を計算する。

(2-3) Retrieval Model Experiments

検索タスクの目的

クエリに合致するブログを検索する（ブログ記事の検索ではない）

使用したテストコレクション

- TREC 2007 Feed Distillation task on BLOG06、検索課題（topic）45 件
- ブログのホームページ ＋ ブログ記事 ＋ XML フィード
- XML フィードだけを索引付けした
- 訓練データからパラメタを学習するため、5-fold の交差検定を行った

実験結果

- 評価尺度: MAP, P@10
- SD, $P(F)=\log(N_F)$ 、P(E|F)＝式(5) の組み合わせが最も良かった
- MAP の差は、Wilcoxon テスト有意水準 0.05 で有意だった

(3) フィード検索とクエリ拡張

(3-1) Pseudo-Relevance Feedback

比較した手法

(a) QE なし

- NO_EXP

(b) 単純な PRF

- PRF.FEED: 上位 N 件のフィードからタームを抽出
- PRF.ENTRY: 上位 N 件の記事からタームを抽出

(c) ウィキペディアの利用

- PRF.WIKI: 上位 N 件のウィキペディア記事からタームを抽出
- PRF.WIKI.P: 上位 N 件のウィキペディア記事パッセージからタームを抽出

  パッセージの長さ 220 単語（= BLOG06 の平均記事長）

(d) 人手による適合性フィードバック（擬似でない）のシミュレーション

- PF.TTR: 上位 10 件の適合フィードからタームを抽出

・ PR.RTT: 上位 10 件のフィードに含まれる適合文書からタームを抽出

上位 10 件のフィードに適合文書がなければ QE なし

実験結果

・ LD の場合に、ウィキペディアの利用（PRF.WIKI, PRF.WIKI.P）が有効だった。
・ フィードは長すぎる？

RF.TTR, RF.RTT が有効だったことからフィードの長さは問題ではない。

・ 記事は短すぎる？

PRF.WIKI.P が有効だったことから記事の長さは問題ではない。

・ 単純な PRF はスパムに弱い。

フィードバックタームにポルノ関連のタームが多い。

(3-2) Wikipedia Link-based Query Expansion

主旨

・ アンカーテキストを辿ると異表記を検出できる。

（例）US → United States of America

・ 初期検索結果 W のうち上位 R 件の文書に多くリンクしているアンカーテキスト（ウィキペディアの記事タイトル）をフィードバックタームとして使用する

実際には、フィードバックフレーズ

・ アンカーテキストだけを調べるので効率が良い
・ パラメタの値： R=100, W=1000, T=20

実験結果

・ WIKI.LINK はフィード検索に有効である。
・ TB（テラバイトコレクション）の検索には有効ではない

理由

・ TB は.gov から収集したページなのでスパムと無関係である。
・ WIKI.LINK は複数のウィキペディアの記事タイトルからフィードバックタームを集めるので、観点が広くなる傾向にある。

(3-3) Investigating Query Generality

仮説

・ ブログ検索に使われるクエリは他の検索に使われるクエリよりも一般的である。

3 種類の実験

- クエリの長さ： 長いクエリほど要求の特定性が高い。
- Open Directory における深さ： 一般的なクエリで ODP を検索すると階層の上位に位置する文書が検索されやすい。
- ウィキペディアにおける文書の結束性： 省略（詳細の一部が不明）

実験結果

全ての実験結果は、ブログ検索に使われるクエリがテラバイト検索のクエリに比べて一般性が高いことを示した。

参考文献

- J. Arguello, J. L. Elsas, J. Callan, and J. G. Carbonell. Document representation and query expansion models for blog recommendation. In Proc. of the 2nd Intl. Conf. on Weblogs and Social Media, 2008.
- J. Elsas, J. Arguello, J. Callan, and J. Carbonell. Retrieval and feedback models for blog distillation. In Proc. of the 2007 Text Retrieval Conf., 2007.
- D. Hannah, C. Macdonald, J. Peng, B. He, and I. Ounis. University of Glasgow at TREC 2007: Experiments with blog and enterprise tracks with terrier. In Proc. of the 2007 Text Retrieval Conf., 2007.
- D. Metzler, T. Strohman, Y. Zhou, and W. Croft. Indri at TREC 2005: Terabyte track. In Proc. of the 2005 Text Retrieval Conf., 2005.
- J. Seo and W. B. Croft. Umass at trec 2007 blog distillation task. In Proc. of the 2007 Text Retrieval Conf., 2007.

（藤井 敦）

4. 4. 2　Selecting Good Expansion Terms for Pseudo-Relevance Feedback

Guihong Cao, Jian-Yun Nie, Jianfeng Gao, Stephen Robertson (pp. 243-250)

(1)　はじめに.

擬似適合性フィードバックは、擬似適合文書内で最も頻度の大きいタームが検索に有用であると仮定する。本研究ではこの仮定を見直し、検索結果に良い影響を与える有用な拡張タームを予測するためのターム分類手法を提案する。

通常、ユーザのクエリは必要な情報を正確に記述するには短すぎる。多くの重要なタームがクエリ

に含まれておらず、関連する文書を十分にカバーできない。この課題を解決するため、クエリ拡張が広く用いられている[1][2][3][4]。これらのアプローチの中で、検索結果を利用する擬似適合性フィードバック（PRF：pseudo-relevance feedback）で最も有効な結果が得られている。PRF の基本的な仮定は、最初の検索結果上位の文書内に、関連文書を弁別するのに有用な多くのタームが含まれているというものである。一般に、拡張タームはフィードバック文書からタームの分布によって、または、フィードバック文書と全文書との分布の違いをもとに抽出される。別の基準もいくつかの提案されている。たとえばベクトル空間モデルの idf が広く使用されている[2]。文献[5]では拡張タームにクエリの長さを考慮した重み付けをする。文献[6]では言語的な特徴が試されている。

しかし、既存の方法により抽出された拡張タームが、実際に検索に役立っているかどうかを直接調査した研究は少ない。通常、拡張タームの集合による全体的な検索精度向上のみ検討されているが、抽出された個々の拡張タームが本当にクエリと関係があるかどうか、情報検索に有用であるかどうかという基本的な疑問が見落とされがちである。

本研究では、タームの選択に教師あり学習を用いる手法を提案する。タームの選択はターム分類の問題として考えられる。潜在的な検索精度向上への寄与によって、直接的に良い拡張タームを他の拡張タームから分離することを試みる。本手法は、教師なし学習を用いる従来手法とは異なり、タームの分布だけではなくタームの類似性のような追加の基準を用いる。

本アプローチは少なくとも以下の利点をもつ。

- 拡張タームは、単に分布や検索精度向上と間接的に関係するほかの基準に基づいて選択されるのではない。選択されたタームは直接的に検索精度に強い影響をもつことが期待できる。
- ターム分類プロセスは、多様な基準を自然に統合できるので、異なる種類の証拠を統合するためのフレームワークを提供できる。

本手法を TREC の 3 種類のコレクションで評価し、従来手法と比較した。実験の結果、ターム分類を統合する本手法において、検索精度の有意な向上を示すことができた。擬似適合性検証における個々の拡張タームに対する検索精度向上の影響についての研究は、われわれの知る限り初の試みである。

(2) PRF 仮説の再検証

PRF の背後にある一般的な仮説は以下のように形式化できる。

> 【PRF 仮説】擬似適合性フィードバック文書中で最も頻出する、または、最も特徴的なタームが有用であり、クエリにこれらを加えたときに検索精度を向上し得る。

この仮説を検証するため、混合モデルによりフィードバック文書から抽出した拡張タームのすべてについて考える。これらのタームすべてについて順次検索精度への影響を調べる。拡張ターム e を含めたときのスコア関数は以下である。

$$Score(d,q) = \sum_{w \in V} P(t \mid \theta_O) \log P(t \mid \theta_d) + w \log P(e \mid \theta_d)$$

t：クエリターム

e：拡張クエリターム

w：重み

$P(t \mid \theta_o)$：オリジナルクエリモデル（基本ランキング関数）

単純化のため以下を仮定する。

・拡張タームは、クエリに対して他の拡張タームとは独立に働く

・各拡張タームはクエリに対して等しい重み w（0.01 または-0.01）で加えられる

上記の単純化をしても拡張タームのおもな特徴は反映できる。良い拡張タームとは、重みを 0.01 としたときに検索精度が向上し、重みを-0.01 としたときに検索精度が悪化するタームである。悪い拡張タームは上記と逆の効果をもつタームである。中立のタームとは、重みの正負に関わらず、検索精度に同様の影響を及ぼすタームである。下式の chg(e)の絶対値が 0.005 より大きい場合に、それぞれ「良い」「悪い」タームであると定義した。

$$chg(e) = \left[MAP(q \cup e) - MAP(q) \right] / MAP(q)$$

また、上記に加え、フィードバック文書中での頻度が 3 未満のタームは重要タームではないという仮定した。この仮定はノイズのフィルタリングに役立つ。TREC のコレクション AP,WSJ,Disc4&5 を用いて、各コレクションの 150 クエリに対して、スコアが大きい 80 の拡張タームを抽出して実験した結果、以下のことがわかった。

・ 良いタームの比率（17.5%前後）が悪いタームの比率（26%～37%）より小さい。

・ 検索結果への影響が小さい中立タームの比率（48%～56%）が大きい。

中立タームが多いと、検索速度への悪影響（ランキングの評価時間増）があるので、中立タームも削除できることが望ましい。上記の解析結果は、従来の混合モデルによる拡張タームの抽出が不適切であることを示している。また、拡張タームの出現確率の分布を調査すると、中立タームはある程度分離されている（右下に中立語）が良いタームと悪いタームは混在しており、タームの出現確率分布単独では分離が困難であることがわかった。

(3) 拡張タームの分類

分類器として C-SVM を用いる。カーネル関数には RBF（Radial Based Function：動径基底関数）を用いる。定義は下式。

$$K(x_i, x_j) = \exp(-\left\| x_i - x_j \right\| / 2\sigma^2)$$

パラメータのσおよびCは5foldのクロスバリデーションで訓練データの分類精度が最大となるものを推定した。

フィードバック文書中のターム分布や全文書内でのターム分布など、分類に有用な素性はすでに従来の研究で使用されているが、これらの素性は必ずしも適切でないことが分析の結果わかったので、下記素性の追加を検討する。

- 拡張タームと元のクエリタームとの共起
- クエリタームと拡張タームとの近接性

以下、使用した素性をいくつかのグループごとに説明する。

- ターム分布

  最初の素性は擬似関連文書中のタームの分布である。この素性はフィードバック文書Fに対して下式で定義される。

$$f_1(e) = \log \frac{\sum_{D \in F} tf(e, D)}{\sum_t \sum_{D \in F} tf(t, D)}$$

  同様に全文書に対して$f_2(e)$が定義される。これらの素性は従来研究でも使用されている。

- 単一クエリタームとの共起

  多くの研究から、クエリタームと共起するタームはクエリと関係するということがわかっている[7]。そこで、この事実をとらえるための素性を下式で定義する。

$$f_3(e) = \log \frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{D \in F} C(t_i, e \mid D)}{\sum_t \sum_{D \in F} tf(t, D)}$$

  ここで$C(t_i, e \mid D)$は、クエリタームt_iと拡張タームeが文書D中のテキストウインドウ内で共起する頻度である。ウインドウサイズは実験的に12語としている。

- クエリタームペアとの共起

  拡張タームに対するより強い共起関係は、２つのクエリタームと一緒に使用されるということである。文献[7]には、この種の共起関係がクエリの文脈を考慮するので、前項の共起よりも良いことが示されてている。ここではテキストウインドウのサイズは15語とする。可能なタームペアΩに対して、下記素性を定義する。

$$f_5(e) = \log \frac{1}{|\Omega|} \sum_{(t_i, t_j) \in \Omega} \frac{\sum_{D \in F} C(t_i, t_j, e \mid D)}{\sum_t \sum_{D \in F} tf(t, D)}$$

- 重みつきターム近接性

  タームの近接性を使用するというアイデアはいくつかの既存研究にもある[8]。ここでも、2つのタームがより小さな距離で共起するほうがより関係すると仮定する。ある文書集合内で2つのタ

ームの近接性を定義するにはいくつかの方法がある。ここでは 2 つのターム間に現れる最小単語数で定義する。ここで文書集合 B における t_i と t_j の間の距離を $dist(t_i,t_j \mid B)$ として定義する。複数単語からなるクエリに対して、全クエリタームと拡張タームとの距離を集計しなくてはならない。もっとも単純な方法は[8]と同様に平均距離を用いる方法である。しかし、実験では、重みつき平均を用いるほうが良い結果が得られることがわかった。下式のように共起頻度で重み付けして平均距離を計算する。

$$f_7(e) = \log \frac{\sum_{i=1}^{n} C(t_i, e) dist(t_i, e \mid F)}{\sum_{i=1}^{n} C(t_i, e)}$$

ここで $C(t_i,e)$ は t_i と e がコレクション中のテキストウインドウ内で共起する頻度である。ウインドウサイズは前記と同様 12 語とする。

・ クエリタームと拡張タームの共起ドキュメント頻度

このグループの素性は、拡張タームがクエリタームと共起する文書の数である。

$$f_9 = \log \left\lfloor \sum_{D \in F} I((\wedge_{t \in q} t \in D) \wedge e \in D) + 0.5 \right\rfloor$$

ここで I(x)は indicator function であり、論理表現 x が真であるときに 1、偽であるときに 0 を返す。定数 0.5 はスムージング因子で 0 値を避けるためのものである。

素性の値が大きな値や小さな値に偏るのを防ぐため、[0,1]の範囲となるよう下式（query-by-query 方式）により正規化する。

$$f_i(e) = (f_i(e) - \min_i) / (\max_i - \min_i)$$

われわれの実験では、上記の素性のみを用いた。しかし、これらに限定されない一般的な手法である。他の素性を追加することも可能である。拡張ターム選択のために自由な素性を統合可能であることは、本手法の利点である。

(4) ターム分類実験

TREC の 3 種類のコレクションで 150 クエリを用いて実験した。50 クエリずつの 3 グループに分け、クロスバリデーションを行った。3 節で述べた方法で拡張タームの候補（良い、悪い、中立を含む）を抽出し、5.2 節で述べた素性により拡張ターム候補を分類した。フィードバック文書中の頻度が 3 以上のものを対象とした。実験の結果、良いタームの割合は約 1/3 である。SVM 分類器による分類精度は約 69%であり、必ずしも高くはないが、適合率は比較的高く（約 60%）、検索結果に良い

影響を与えると思われる。

(5) タターム分類による拡張タームの再重み付け

分類プロセスでは、レレバンスモデルや混合モデルで提案された拡張タームの選択に対して、さらなる選択を行う。選択された拡張タームは異なる 2 つの手法により従来モデルに統合できる。1 つはハードフィルタリングでもうひとつはソフトフィルタリングである。実験では後者の性能が良かった。ここでは、2 番目のソフトフィルタリングについて説明する。$P(w|\theta_R)$は混合モデルに対して再定義したモデルで、$P(w|\theta_T)$は混合モデルに対して再定義したモデルである。

$$P(w|\theta_R)^{new} = (P(e|\theta_R)^{old}(1+\alpha P(+1|e)))/Z$$
$$P(w|\theta_T)^{new} = (P(e|\theta_T)^{old}(1+\alpha P(+1|e)))/Z$$

ここで Z は正規化因子である。αは係数で、実験データ中の開発用データから line search[9]を用いて推定する。拡張タームの再重み付けがすめば、確率の大きい上位 80 タームを選択できる。80 は、レレバンスモデル、混合モデルとフェアな比較をするためである。

(6) 情報検索実験

- TREC コレクションの 3 種類（AP88-90：24918 文書、WSJ87-92：173252 文書、disk 4&5：556077 文書）の全文書を用いて評価した。各データセットを 3 分割し、それぞれを SVM の訓練用データ、パラメータαを推定するための開発データ、評価データとした。クエリとして TREC のトピックを使用した。対象文書もクエリも Porter stemmer により語幹処理を行い、ストップワードを除去。
- 主な評価指標は上位 1000 文書の Mean Average Precision(MAP)である。いくつかの先行研究では PRF により再現率は向上しているが適合率が悪化しているので、上位 30 位と上位 100 位の適合率も P@30 と P@100 として示した。補足として再現率も示した。query-by-query 分析と t 検定を行い、MAP の向上に有意差があるかどうかを調べた。
- Indri 2.6 検索エンジン[10]がわれわれの基本的な検索システムである。レレバンスモデルは Indri に実装されているが混合モデルは実装されていないので、混合モデルについては文献[4]にしたがって実装した。
- 実験では以下の手法を比較した。

  LM： オリジナルのクエリでの KL-divergence 検索モデル

  REL： レレバンスモデル

  REL+SVM： レレバンスモデル＋ターム分類

  MIX： 混合モデル

MIX+SVM： 混合モデル＋ターム分類

- これらのモデルは、ドキュメントモデルのスムージングのための Dirichlet の事前パラメータなど、いくつかのパラメータを必要とするが、本研究ではパラメータの最適値を求めることが目的ではないので、全モデルに対して同一のパラメータを使用した。
- 表 4.4.2-1～表 4.4.2-3 に 3 種類のコレクションに対する各モデルの結果を示す。表中の Imp は LM モデルに対する精度向上を示しており、統計的に P<0.05 で有意なものは*、P<0.01 で有意なものは**で示している。上付き文字の"R"と"M"は、それぞれレレバンスモデル、混合モデルに対して P<0.05 で統計的に有意な向上があることを示している。

表 4.4.2-1 検索実験結果[AP88-90]

モデル（手法）	P@30	P@100	MAP	Imp	再現率
LM	0.3967	0.3156	0.2407	——	0.4389
REL	0.4380	0.3628	0.2752	14.33%**	0.4932
REL+SVM	0.4513	0.3680	0.2959^{R}	22.93%**	0.5042
MIX	0.4493	0.3676	0.2846	18.24%**	0.5163
MIX+SVM	0.4567	0.3784	$0.3090^{M,R}$	28.36%**	0.5275

表 4.4.2-2 検索実験結果[WSJ87-92]

モデル（手法）	P@30	P@100	MAP	Imp	再現率
LM	0.3900	0.2936	0.2644	——	0.6516
REL	0.4087	0.3078	0.2843	7.53%**	0.6797
REL+SVM	0.4167	0.3120	0.2943	11.30%**	0.6933
MIX	0.4147	0.3144	0.2938	11.11%**	0.7052
MIX+SVM	0.4200	0.3160	0.3036^{R}	14.82%**	0.7110

表 4.4.2-3　検索実験結果[disk 4&5]

モデル(手法)	P@30	P@100	MAP	Imp	再現率
LM	0.2900	0.1734	0.1753	——	0.4857
REL	0.2973	0.1844	0.1860	6.10%*	0.5158
REL+SVM	0.2833	0.1990	0.2002^{R}	14.20%**	0.5689
MIX	0.3027	0.1998	0.2005	14.37%**	0.5526
MIX+SVM	0.3053	0.2068	$0.2208^{M,R}$	25.96%**	0.6025

- 上記の表からは、レレバンスモデル、混合モデルとも LM より良い結果が得られることがわかる。これは従来研究の結果とも矛盾しない。レレバンスモデルと混合モデルを比較すると後者で良い結果となっている。混合モデルではフィードバック文書と全文書を比較して、フィードバック文書に特徴的なタームを選択するので、悪いタームや中立タームがフィルタリングされるためと思われる。
- ターム分類を用いる手法が、3 種類のコレクションいずれでも良い結果となっている。分類モデルを PRF モデルと併用した場合は、常に精度の向上が見られる。AP と Disk4&5 では 7.5%以上向上しており、統計的に有意な差となっている。WSJ ではやや小さい向上（約 3.5%）で、統計的には有意な差になっていない。
- 上位文書の適合率についても、ターム分類により適合率の向上が見られる。Disk4&5 の REL に対して SVM を追加した場合以外では、いずれも適合率が向上している。これは、ほとんどの場合で拡張タームが悪影響を及ぼさず、精度を向上していることを示している。
- クエリ”machine translation”に対する拡張タームとして、混合モデルでは、”50”、”people”、”year”、”work”など有用ではない一般的な拡張タームも含まれるのに対し、提案手法では、”compute”、”soviet”、”company”、”english”など有用な拡張タームのみ得られる。
- 従来のレレバンスモデル、混合モデルに対して２つの変更を加えた。一つは教師あり学習を用いた点、もう一つは素性を追加した点である。どちらの変更が検索精度向上に大きく寄与しているかを調べるため、教師なし学習で素性を追加した場合について調べた（表 4.4.2-4）。

表 4.4.2-4　教師あり学習と教師なし学習の比較

モデル（手法）	AP	WSJ	Disk4&5
MIX	0.2846	0.2938	0.2005
Log-linear	0.2878	0.2964	0.2020
MIX+SVM	$0.309^{M,L}$	0.3036	$0.2208^{M,L}$

- 素性を増やした Log-linear では大きな精度向上は見られず、教師あり学習による効果が大きいことがわかった。
- クエリ拡張には速度性能の問題がある。上記の実験では 80 の拡張タームを使用していたが、実際には多すぎる。上位 10 の拡張タームを選択する実験を行い精度を比較した。予想通り、拡張ターム 10 個では、拡張ターム 80 個よりも低い結果となる。しかし、まだ従来の言語モデル LM よりは高く、統計的にも有意な差があることがわかった。

参考文献

[1] Metzler, D. and Croft, B. Latent Concept Expansion Using Markov Random Fields. In the Proceedings of SIGIR'2007, pp.311-318.

[2] Rocchio, J. Relevance feedback in information retrieval. In The SMART Retrieval System: Experiments in Automatic Document Processing, pages 313-323, 1971

[3] Xu, J. and Croft, B. Query expansion using local and global document analysis. In the Proceedings of SIGIR'2006, pp.4-11, 1996.

[4] Zhai, C. and Lafferty, J. Model-based feedback in the KLdivergence retrieval model. In CIKM, pp.403-410, 2001a.

[5] Smeaton, A. F. and Van Rijsbergen, C. J. The retrieval effects of query expansion on a feedback document retrieval system.Computer Journal, 26(3): 239-246. 1983.

[6] Bai, J. Nie, J., Bouchard, H. and Cao, G. Using query contexts in information retrieval. In the Proceedings of SIGIR'2007, Armsterdam, Netherlands, 2007.

[7] Tao, T. and Zhai, C. An exploration of proximity measures information retrieval. In the Proceedings of SIGIR'2007, pp.295-302, 2007.

[8] Gao, J., Qi, H., Xia, X., and Nie, J. Linear discriminant model for information retrieval. In the Proceedings of SIGIR'2005, pp. 290-297, 2005.

[9] Strohman, T., Metzler, D. and Turtle, H., and Croft, B. (2004). Indri: A Language Model-based Search Engine for Complex Queries. In Proceedings of the International conference on Intelligence Analysis.

（相川 勇之）

4. 4. 3　A Study of Methods for Negative Relevance Feedback

Xuanhui Wang, Hui Fang & ChengXiang Zhai (pp. 219-226)

(1)　概要

Negative Relevance Feedback の各手法の性能について体系的に評価した。

ベクトル空間モデル or 言語モデル

1. Single Query Strategy
2. Score Combination
   i. Single Negative Model
      A) Heuristic 1(Local Neighborhood)
      B) Heuristic 2(Global Neighborhood)
   ii. Multiple Negative Models
      A) Heuristic 1(Local Neighborhood)
      B) Heuristic 2(Global Neighborhood)

上記の手法に対して、TREC コレクションに基づいて Negative Relevance Feedback 用に文書コレクションを作成して精度評価を行った。言語モデルをベースとした Multiple Negative Models が良い結果となった。

(2)　Introduction

Negative Relevance Feedback は Relevance Feedback の特殊なケースであり、検索結果内に適切な正のフィードバック文書が見つからない場合に利用する手法である。著者らは以前に言語モデルに基づく Negative Feedback 方式の提案について報告しているが、以下の理由から評価が不十分であった。

i.　ベクトル空間モデルに適用した場合の評価をしていない
ii.　1 つのテストコレクションでしか評価していない
iii.　体系的な実験設計をしておらず他手法との差異がわからない

これらを踏まえて、本論文では Negative Relevance Feedback に関する手法を体系的に評価した。また、Negative Relevance Feedback を評価する上で必要となるテストコレクションを既存のデータから作成する方法についても提案する。

(3) 評価に用いた Negative Feedback 方式の説明

① 記号の説明

→表 4.4.3-1 を参照

② Single Query の基本方式

クエリに単語を追加し関連文書に多く出現する単語の重みを大きくして、非関連文書に多く現れる単語の重みを小さくする。

③ Score Combination の基本方式

Score Combination の基本式。

$$S_{combined}(Q, D) = S(Q, D) - \beta \times S(Q_{neg}, D) \qquad (1)$$

表 4.4.3-1 本論文で用いる記号の説明

記 号	説 明
Q	クエリ
C	文書コレクション
L	文書のランキングリスト
L_i	L 内の i 番目の文書
f	クエリ Q により得られた上位 f 件の文書は全て non-relevant と仮定する
$N = \{L_1, \ldots, L_f\}$	Negative Feedback に使用した文書集合
$U = \{L_{f+1}, \ldots, L_{f+r}\}$	再ランキングする r 件の文書集合
$S(Q, D)$	文書 D とクエリ Q との関連度スコア
$c(w, D)$	文書 D 内での単語 w の出現数
$c(w, Q)$	文書 Q 内での単語 w の出現数
$\lvert C \rvert$	文書コレクション C 内の文書数
$df(w)$	単語 w の文書頻度
$\lvert D \rvert$	文書 D の文書長
$avdl$	平均文書長

※ 本論文では $f = 10,\ r = 1000$ としている。

Score Combination によりスコア修正を行う文書を以下の 2 つのヒューリスティクスにより選ぶ。

・Heuristic 1 (Local Neighborhood)：Negative query により U 内の文書全てをランク付けし、上位 ρ 件をスコア修正対象文書とする。

・Heuristic 2 (Global Neighborhood)：Negative query により C 内の文書全てをランク付けし、このリストの上位 ρ 件の中から文書集合 U に含まれる文書をスコア修正対象とする。

④　ベクトル空間モデル

(a)　Single Query Strategy (SingleQuery)

N の重心でクエリを拡張する：$\vec{Q}_{new} = \vec{Q} - \gamma \times \frac{1}{|N|}\sum_{D \in N} \vec{D}.$

(b)　Score Combination

1．Single Negative Model (*SingleNeg*)

$$S_{combined}(Q, D) = \vec{Q} \cdot \vec{D} - \beta \times \vec{Q}_{neg} \cdot \vec{D} \quad (\because \vec{Q}_{neg} = \frac{1}{|N|}\sum_{D \in N} \vec{D}).$$

2．Multiple Negative Models (*MultiNeg*)

$$S_{combined}(Q, D) = \vec{Q} \cdot \vec{D} - \beta \times \max(\bigcup_{\vec{Q}' \in N} \{\vec{Q}' \cdot \vec{D}\}).$$

⑤　言語モデル

$$S(Q, D) = -D(\theta_Q \| \theta_D) = -\sum p(w|\theta_Q) \log \frac{p(w|\theta_Q)}{p(w|\theta_D)} \quad (\because p(w|\theta_D) = \frac{c(w, D) + \mu p(w|C)}{|D| + \mu})$$

※ $p(w|C)$ はコレクションの言語モデル、 μ はスムージングパラメータ

(a)　Single Query Strategy (*SingleQuery*)

$$S(Q, D) = -D(\theta_Q \| \theta_D) + \beta \cdot D(\theta_N \| \theta_D)$$
$$\underline{\underline{rank}} \sum_{w \in V} [p(w|\theta_Q) - \gamma \cdot p(w|\theta_N)] \log p(w|\theta_D)$$

(b)　Score Combination

1.　Single Negative Model (*SingleNeg*)

$$S(Q, D) = -D(\theta_Q \| \theta_D) + \beta \cdot D(\theta_N \| \theta_D).$$

$$L(N|\theta_N) = \sum_{D \in N} \sum_{w \in D} c(w, D) \log[(1-\lambda) p(w|\theta_N) + \lambda\, p(w|C)] \quad (\because p(W|C) = \frac{c(w, C)}{\sum_w c(w, C)})$$

※ $\lambda = 0.9$ として、EM アルゴリズムにより $p(w|\theta_N)$ を推定してスコアを計算する。

2. Multiple Negative Models (*MultiNeg*)

$$S_{combined}(Q, D) = -D(\theta_Q \,||\, \theta_D) + \beta \cdot \min(\bigcup_{i=1}^{f} \{D(\theta_i \,||\, \theta_D)\})$$

(4) サンプリングによる擬似的な Difficult Query の作成

ある検索モデルにより得られた上位 10 件に関連文書が無い場合、そのクエリは難しいクエリと見做す。したがって、テストコレクション中のクエリに対する関連文書をサンプリングする以下の 2 つの方法を提案する。

・Minimum Deletion Method：上位 10 件に関連文書が無くなるまで、上位から順に関連文書を削除する

・Random Deletion Method：上位 10 件に関連文書が無くなるまで、ランダムにコレクション中から関連文書を削除する。

(5) 実　験

TREC データセットの ROBUST track と Web track の GOV を使用（表 4.4.3-2）。

表 4.4.3-2　使用したデータセット

		ROBUST		GOV	
Base set	文書数	258,000		1,247,753	
	平均単語数	467		1,094	
	クエリ数	249		125	
		LM	VSM	LM	VSM
パラメータの推定値		$\mu = 2000$	$k_1 = 1.0, b = 0.3$	$\mu = 100$	$k_1 = 4.2, b = 0.8$
上位 10 件に関連文書が無いクエリ数 (Naturally Difficult Queries)		26	25	54	63

※ パラメータは Base set を使用して推定
※ 推定したパラメータを使った検索モデルにより「上位 10 件に関連文書が無いクエリ」を見つける

① Retrieval Effectiveness

(a)　検索精度に関して

(i)　言語モデルの方がベクトル空間モデルよりも良い結果となった。

(ii)　言語モデルにおいて、MultiNeg が SingleQuery や SingleNeg よりも良い結果となった。

(iii)　ベクトル空間モデルにおいて MultiNeg が良い結果を得られなかったのは、高い IDF 値をもつ希な単語を有する特定の文書に従って Negative Query Vector を更新してしまっていたことが原因と考えられる。

(b)　擬似的に作成した Difficult Query を使用した検索結果に関して

5 種の方式間での相対的な性能差は Table 3 での関係と同様の傾向が見られた。また、言語モデルと MultiNeg の有効性も同様に確認できた。

② Effectiveness of Sampling Methods

Difficult Query 作成方法の評価には Kendall の順位相関係数を使用。Minimum, Random いずれにおいても正の相関を示したので有効性が示せた。

③ Parameter Sensitivity Study

・γ（SingleQuery）

γ の値に影響を受けやすく、大きい値になるほど精度が落ちる

・β（Score Combination）

感度はどの手法においてもほぼ同じ

・ρ（Heuristic 1, 2）

SingleNeg1, MultiNeg1 は ρ についての感度が高く、大きくなるほど精度が落ちる。一方で、SingleNeg2, MultiNeg2 は影響をほとんど受けない。Heuristic 2 による罰則対象文書の選択方式が安定した精度を得られる。

・f（Negative Feedback Documents）

多ければ多いほど良い

関連研究

- クエリがなぜ難しいのかを理解するための研究
  - [2] C. Buckley 2004
  - [3] D. Carmel, E. Yom-Tov, A. Darlow, & D. Pelleg 2006
  - [6] D. Harman & J. O. Pedersen 2004）
- クエリが難しいことを特定する研究
  - [17] E. M. Voorhees 2005
- クエリのパフォーマンスを予測する研究
  - [21] E. Yom-Tov, S. Fine, & C. Zhai 2007
- フィードバック技術による検索精度改善に関する研究
  - [1] R. Attar & A. S. Fraenkel 1977

- [5] D. A. Evans & R. G. Lefferts 1994
- [10] S. Robertson & K. Sparck Jones 1976
- [12] J. J. Rocchio 1971
- [13] G. Salton & Buckley 1990
- [14] X. Shen, B. Tan, & C. Zhai 2005
- [20] J. Xu & W. Croft 1996
- [22] C. Zhai & J. Lafferty 2001

● 非関連文書を効果的に探索する手法（Document routing タスクにおいて）の研究

- [16] A. Singhal, M. Mitra, & C. Buckley 1997

（佐藤 祐介）

4. 4. 4 To Personalize or Not to Personalize: Modeling Queries with Variation in User Intent

Jaime Teevan Susan T. Dumais Daniel J. Liebling (pp. 163-170)

(1) はじめに

従来のパーソナライズ検索は、すべてのクエリに対して同じ方法でパーソナライズを行う。しかし、この論文では、クエリ曖昧性（同じ検索クエリでもユーザが検索したい対象にばらつきがあるか）を判別することが、パーソナライズ検索にとって重要な意味を持つと述べている。なぜなら、クエリによって、(i)すべてのユーザが同じ解釈を持つクエリ(ex. “Microsoft Earth”)、(ii)ユーザが意図する検索対象が複数存在し得るクエリ(ex. “street map”)、の 2 種類があり、(i)の場合にパーソナライズを行うと、検索結果が悪くなることがあるからである。クエリ曖昧性の有無に応じて、パーソナライズを行うか行わないか、検索方法を切り替えることができれば、より効果的な検索が行える。

クエリ曖昧性を判定する方法として、この論文では、まず、人手でクエリに対し曖昧性の評価を行う。次に、ユーザが検索結果を「クリックした」という情報が、人手による判定方法の代替になるかを、人手の評価とクリックログの分析結果を比較することで、検証する。

一方で、クリックベースの手法は、クリック履歴を持たない新規のクエリに対しては適用できない。そのため、履歴を使用しなくても、入力されたクエリの特徴（文字数など）を使ってクエリ曖昧性を予測するモデルを検討する。検索時に表れる特徴素（素性）のうち、クリックベースの曖昧性指標と相関のある素性を調査し、実際に機械学習を用いて曖昧性予測モデルを構築する。

(2) クエリ曖昧性の測定方法

① 使用したデータ

・ 検索ログ（implicit データ）

Live Search 2007.10.4-10.11 のログ。言語の違いを抜くため、US の英語圏のエリア分のみを抽出。クエリ数 44002（異なり数。全体では 240 万クエリ、ユーザ数は 153 万人。1 クエリ最低 10 人以上のもの）

・ 人手による評価データ（explicit データ）

ログ中の異なる 12 クエリに対し、128 人から評価を取得。全部で 292 の評価付きセットで、1 クエリあたり 4－81 人の評価者が、検索結果のトップ 50 に対し、「すごく関係がある」「関係がある」「関係がない」の 3 評価をつけたもの。

② クエリ曖昧性のメジャー

表 4.4.4-1 に、人手評価による曖昧性メジャー（explicit メジャー）と、今回評価したい、クリックログから自動獲得可能な曖昧性メジャー（implicit メジャー）を示す。

③ クエリ曖昧性予測に用いる素性

表 4.4.4-2 に、曖昧性予測に用いる素性を示す。クエリと検索結果のどちらを利用するか×履歴利用の有無で分類している。

(3) クエリ曖昧性の分析

① explicit/implicit なクエリ曖昧性メジャーの比較

クリックベースの implicit な曖昧性メジャーが、人手評価による explicit なメジャーを近似できるかを調査するため、メジャー間の相関をみた。結果として、以下のように相関関係があり、クリックベースの implicit なメジャーが有効なことがわかった。

・ クリックエントロピーが高くなるほど、explicit なメジャーの値は小さくなっている
・ implicit なパーソナライゼーションポテンシャルカーブは、クリックエントロピーよりもさらに（explicit メジャーとの）強い相関がある

② 曖昧性メジャーと、曖昧性予測に使う素性との相関関係

クリックエントロピーおよび implicit なパーソナライゼーションポテンシャルカーブ(グループ人数 10)と、表 4.4.4-2 で示した各素性との相関関係を調査した。結果から、曖昧性メジャーと相関の強い素性として、以下が明らかになった。

・ クエリ単体の素性「クエリ長」、「クエリが URL の断片を含むか」、検索結果単体の素性「検索結果内のホストの異なり数」と相関関係がみられる

・ 検索結果の履歴を使った素性「結果エントロピー」との相関が強い
・ 検索結果の履歴を使った素性「クリックした検索結果の平均ランク」「1 ユーザあたりの平均クリックコンテンツ数」は特に相関が強い
・ 検索結果の履歴を使った素性は強い相関が見られる傾向があるが、そのうち「最初の検索結果をクリックするまでの時間」はあまり相関がない

表 4.4.4-1 クエリ曖昧性のメジャー

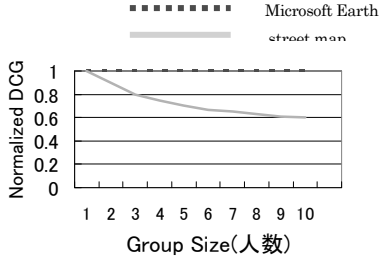
	人手評価による曖昧性メジャー （explicit メジャー）	クリックログから自動獲得可能な 曖昧性メジャー(implicit メジャー)
1	評価者内信頼性 Fleiss の κ [1] $\kappa = \frac{P - P_e}{1 - P_e}$ 全ての評価者がランダムに評価を行なった際の確率(Pe)から、評価者の一致度Ｐがどれだけずれがあるかのメジャー	クリックエントロピー[3] $Click\ Entropy(q) = -\sum_{URL\ u} p(C_u \mid q) * \log_2(p(C_u \mid q))$ ※$P(C_u \mid q)$：URL u がクエリ q でクリックされた確率 クリックエントロピーが大きいと、いろいろなコンテンツがクリックされた（クエリ曖昧性がある）ということ
2	パーソナライゼーションポテンシャルカーブ[2] グループの人数別に、 評価者の一致度を示す値 nDCG (normalized Discounted Cumulative Gain) を計算したもの 	クリックされた回数を、 人が「クエリと検索結果が関係あり」と判断したものとして近似した、 パーソナライゼーションポテンシャルカーブ

表 4.4.4-2　クエリ曖昧性予測に用いる素性

		履歴利用	
		No	YES
情報	クエリ	・ クエリの長さ（文字数、 word 数） ・ クエリが場所を含むか ・ クエリが URL の断片を含むか ・ クエリが advanced operator を含むか ・ クエリ入力された時間帯 ・ クエリ入力が勤務時間内か ・ 検索結果の数 ・ クエリサジェスチョンの数 ・ 表示された広告数 ・ definitive な結果を持っているか	・ Reformulation の確率 ・ そのクエリが何回入力されたか ・ そのクエリを入力したユニークユーザ数 ・ 1 日の平均クエリ数 ・ 勤務時間内の平均クエリ数 ・ 結果の平均数 ・ クエリサジェスチョンの平均数 ・ 広告の平均数

	検索結果	・ クエリの clarity （検索結果のクオリティのメジャー、[4]） $Clarity(q) = -\sum_{Terms\ t} p(t \mid q) * \log_2\left(\frac{p(t \mid q)}{p(t)}\right)$ ※p(t|q)：クエリに対して返ってきた検索結果内にタームが出てくる確率 p(t)：タームが index 内に出てくる確率 ・ 結果のページが持つ ODP（オープンディレクトリプロジェクト）のカテゴリのエントロピー. ・ ODP のカテゴリ数 ・ ODP のカテゴリの異なり数 ・ ODP にマッチした URL の数 ・ HTML でない結果の割合 ・ “.com”もしくは”.edu”の割合 ・ 検索結果中のドメインの異なり数	・ 結果エントロピー（どれぐらいの頻度で検索結果が変わったか） Result $entropy(q) = -\sum_{URL\ u} p(u \mid q) * \log_2(p(u \mid q))$ ※p(u|q)：URL u が、クエリ q の検索結果のトップ 10 入りした回数 ・ ユーザがクリックした結果の平均ランク位置 ・ クエリ入力後、結果をクリックするまでにかかった平均時間 ・ 1 ユーザあたりがクリックするコンテンツ数の平均値 ・ クリックエントロピー ・ パーソナライゼーションポテンシャル

③ クエリ曖昧性メジャーに影響する要因

②で示した曖昧性メジャーと相関の強い素性についてより詳細に分析を行い、クエリに対してユーザが意図している検索対象が複数あるという要因以外に、クリックベースの曖昧性メジャーに影響を与えている要因として、以下の(A) 、(B) 、(C)を挙げている。

（A）検索結果の変化

同じクエリによる検索であっても、検索結果は定期的に変わることがわかっている[5] 。当然、提示される検索結果が異なれば、ユーザがクリックするコンテンツにはばらつきが出る。検索結果の変化指標である結果エントロピーと、クリックエントロピーとの相関をより詳細に見ると、結果エントロピーの値が 2 より大きいクエリは相関が強く、2 以下だと相関が見られない。よって、この論文では分析の際、結果エントロピーが 2 以下の場合を分けて調査を行っている。

（B）ユーザの検索タスク

ユーザのタスクが、特定サイトへのナビゲーションの場合（ex. クエリ”CNN”,”Google”）は、だいたい 1 クリックなのに対し、タスクが情報収集の場合（ex.”cancer”）ならば、いろいろなページを多数クリックする。たとえば、半分のユーザがコンテンツ A を、残りの半分のユーザがコンテンツ B をクリックした場合と、全員がコンテンツ A、B の両方をクリックした場合とでは、クリックエントロピーの値は同じになる。しかし、個人の間の差は前者の方が大きい。ユーザのタスクの違いは、パーソナライゼーションポテンシャルカーブの形に表れる。クリックエントロピーの値は同じでも、1 ユーザがクリックする結果の数が少ないクエリほど、ポテンシャルカーブで大人数になったときに差が開く（カーブがきつい形になる）。

（C）検索結果のクオリティの低さ

クリックされたページのランク位置（平均）が、曖昧性メジャーとの相関が強い。人々は検索結果

の 1 個目をクリックする傾向があることがわかっている[6,7] 。にもかかわらず、クリックされたランク位置が低いということは、検索結果のクオリティが低い（1 個目の結果に満足していない）ということである。

(4) クエリ曖昧性の予測モデルの作成

クリックログを使った曖昧性メジャーが有効であることがわかったが、「このクエリはパーソナライズするべきか？」を判定する際、すべてのクエリに対して履歴情報があるわけではない。ログ中に既出でない新規クエリも多いので、そのような場合でもクエリ曖昧性を予測できるモデルを構築したい。そこで、以下のタスクに対し、学習アルゴリズム Bayesian dependency networks で、予測モデルを学習した。学習結果のモデルは、図 4.4.4-1 のような確率付き決定木の形である。学習には、表 4.4.4-2 に示した素性を利用し、5 分割の交差検定を実施した。また、素性の組み合わせも調査した。

＜タスク＞

- クリックエントロピー、ポテンシャルカーブ（5 人）、ポテンシャルカーブ（10 人） の値の予測問題。値の範囲を示すカテゴリ（4 種類）を予測する。
- ポテンシャルクラスタの予測問題。ポテンシャルカーブの形、値の類似で、15 種類のパターンにクラスタリングし、どのクラスタに属するかを予測する。

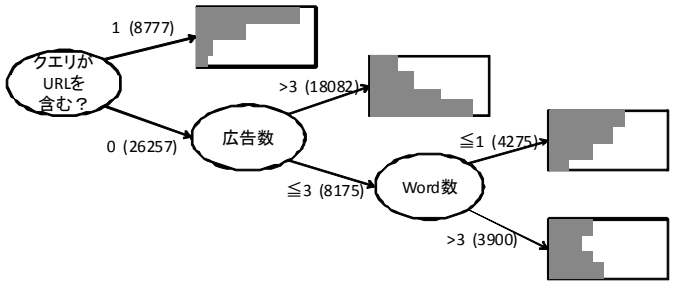


図 4.4.4-1 クリックエントロピーの予測用に学習した木の例（クエリ単体素性のみ利用）

作成したモデルの精度評価の結果、クリックエントロピー、ポテンシャルカーブ（5 人）、ポテンシャルカーブ（10 人）の予測については、検索結果や履歴情報を使わずに、クエリ単体から得られる素性を使うだけでも、効果があることがわかった。 accuracy が 38～40％となり、ベースライン（25～26％）から 50～57％の改善率である。（なお、結果エントロピーの値が低いものだけに絞って評価した場合は、38～39％の改善率であった。）さらに、検索結果単体の素性、またはクエリ履歴に関する素性を追加して評価してみたが、accuracy の改善にはほとんど影響しないこともわかった。最後に、すべての素性を使った場合には、クリックエントロピー、ポテンシャルカーブ（5 人）、ポテンシャルカーブ（10 人）の accuracy は飛躍的に高くなり、80％となった。

(5) まとめ

検索結果をパーソナライズする価値があるクエリを見極めるため、検索のクリックログを利用する方法を調査した。人手で評価を付けた結果と、膨大なログ内のユーザのクリック行動の分析結果から、クリックベースのクエリ曖昧性指標（クリックエントロピーとパーソナライゼーションポテンシャルカーブ）が有効な指標であることを確かめた。また、クリックエントロピーとパーソナライゼーションポテンシャルカーブの値に影響を与える要素として、検索結果のゆれやユーザの検索タスク、結果のクオリティ等も関係することがわかった。一方、クリックベースの指標は、クリック履歴がないと使用できないため、クエリの素性を使ってどのぐらいうまく曖昧性を予測できるかも調査した。その結果、クエリ文字長など、クエリ単体が持つ情報を使うだけでも、曖昧性予測に効果があることがわかった。

参考文献

[1] Fleiss, J. L. Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5): pp.378-382(1971).

[2] Teevan, J., Dumais, S. T., and Horvitz, E. Potential for Personalization. Under submission.(2008).

[3] Dou, Z., Song, R., and Wen, J.R. A large-scale evaluation and analysis of personalized search strategies. In Proceedings of WWW .07, pp.581-590 (2007).

[4] Cronen-Townsend, S., Zhou, Y., and Croft, W. B. Predicting query performance. In Proceedings of SIGIR .02, pp.299-306(2002).

[5] Selberg, E. and Etzinoi, O. On the instability of Web search engines. In Proceedings of RIAO 2000 (2000).

[6] Guan, Z. and Cutrell, E. An eye tracking study of the effects of target rank on Web search. In Proceedings of CHI 2007, pp.417-420(2007).

[7] Joachims, T., Granka, L., Pan, B., Hembrooke, H. and Gay, G. Accurately interpreting clickthrough data as implicit feedback. In Proceedings of SIGIR 2005, pp.154-161(2005).

（志賀 聡子）

4. 4. 5　User adaptation: good results from poor systems

Catherine L. Smith & Paul B. Kantor (pp. 147-154)

(1)　はじめに

検索システムの性能を改善するために、システムはユーザからクエリ以外の追加情報を得る必要がある。たとえば、検索文脈、個人の検索履歴などの情報を用いたり、適合性フィードバックによりユーザから直接情報を得ることもある。しかし、これらの方策が必ずしも効果的であるとは限らず、現実には、ユーザ（人間）はシステムよりはるかに柔軟であり、ユーザがシステムを使いこなしているというのが実状である。本論文では、システムの挙動（性能）の違いによりユーザがどう検索行動（戦略）を変え、最良の検索結果を引き出そうとするかを被験者実験を通して明かにすることを目的としている。

(2)　関連研究

このようなユーザ研究は過去にいくつか報告されている。

(Turpin & Scholer 2006)

標準システムと意図的に性能を落した劣化システムで 1 つの適合文書を見つける時間に差はあるかを調査し、両者に差がないという結果を得ている。

(Allan, Carterette & Lewis 2005)

ユーザの生産性はシステム性能の両極端な場合（非常に良い場合と悪い場合)にシステムの影響を受けるという結果を得ている。

(Turpin & Hersh 2001)

ユーザは標準システムでも性能をよくした拡張システムでも同じような検索結果を得る。ただ、標準システムでは 3 倍のクエリを必要とするなど、効率が悪いという結果を得ている。

要するに、これらの研究からわかることは、ユーザはシステムの性能に応じて適応的に検索行動を変えているということである。

(3)　実験設計

36 名に学生を被験者として用い、与えられた各トピックに対するよい情報源（good information sources）を検索システムを用いて見つけるタスクを実行させた。被験者には$15 の報酬を与えた。また、見つけた情報源に全体で一番よい情報源が入っており、かつ、見つけた情報源の中の悪い情報源

数が一番少ない者には$40 のボーナスが与えられることとした。よい情報源とは、そのトピックに関して自分が使うことができそうなものであると定義されている。36 名の被験者を 12 名ずつ、3 グループに分割。また、トピックは 12 トピックで、1 被験者のセッションは 4 トピックずつ 3 つのブロックに分割されている。このブロック分割については被験者には知らされない。 最初と最後のブロックは全グループ同じで標準システムを使う。真中のブロックは使うシステムをグループで変える。検索順序はブロックでバランスするように調整する。練習セッションの後、実験手順で実験をおこなう。検索の時間制限は設けない。

・トピックを記述した文章を提示する

・事前アンケートに答える

・検索（システムの出力の各項目に対して GIS (Good Information Sources)フラグを付ける）

・事後アンケートに答える

実験システムでは、Google の出力から広告、Cached, Similar pages などのリンクを削除し、ランキングリストのうち 20 項目を提示する。この 20 をどう選ぶかで以下の 3 つのシステムを作る。

・Controlled：Google の出力上位 20 個をそのまま出力する

・Consistently low ranking (CLR)：常に Google の 300-320 位を出力する

・Inconsistently low ranking (ILR)：Google の出力中のあらかじめ決めた 20 項目のブロックをトピックの順序に応じて選択して出力する

システムのディスプレイはシステムインタフェース用とページ内容の表示用に 2 つある。

(4) 採取データ

以下のデータを 1 次データとして採取する。

・検索開始時刻

・クエリ（タイムスタンプ付き）

・被験者に表示される項目とそのランキングリスト

・GIS 情報

・検索終了時刻

・demographic 情報、検索経験（アンケートで採取）

(5) 正解判定

GIS タグの付いた URL について実験条件を隠して以下の基準で第一著者が判定をおこなう。

・good: 必要な情報を（1）含むページ、(2) 含むページへの直接リンクがあるページ、(3) 検索できるページ

・marginal: トピックの一部しか網羅しないもの

・bad: トピックに関係ないもの

・missing: リンク切れ

複数の被験者が見つけた GIS タグ付きのページの分布は、good: 51.8%, marginal: 19.3%, bad: 24.4%,, missing: 4.5%であった。

(6) 分析

各検索ごとに以下の変数を調査した。

・GID: good と判定された表示項目の数

・AID: すべての表示項目の数

・GUI: GID から重複を除いたもの

・AUI: AID から重複を除いたもの

・GTID: GID のうち GIS タグのついた項目の数

・ATID: AID のうち GIS タグのついた項目の数

・BTUI: GIS タグのついた重複のない bad 項目の数

・MTUI: GIS タグのついた重複のない marginal 項目の数

・GTUI: GIS タグのついた重複のない good 項目の数

・ATUI: GIS タグのついた重複のない項目の数

・TGI: コレクション中の good 項目の数

システムの評価尺度としては以下の 2 つを考える

・GPrec = GID/AID：精度

・GRec = GID/TGI：再現率

ユーザの評価尺度として以下のものを考える。

・GTUI: GIS タグのついた重複のない good 項目の数

・BTUI: GIS タグのついた重複のない bad 項目の数

・MTUI: GIS タグのついた重複のない marginal 項目の数

・ETTime: 検索開始から終了までの時間（分）

・Good item ratio = GTUI/ATUI

・Marginal item ratio = MTUI/ATUI

・Bad item ratio = BTUI/ATUI

・Searcher selectivity = 1 - (ATID/AID)：ユーザが GIS を付けない割合、GIS を付けにくいと高くなる

・Searcher sensitivity = (GTID/GID)：ユーザが good 項目に GIS を付ける割合

ユーザの行動とシステムの応答の評価尺度

・Query rate = 入力クエリ数/ETTime：単位時間のクエリ投入数

・Average list length = AID/入力クエリ数：クエリに対する平均表示項目数

・Unique items per query = AUI/入力クエリ数：上のユニークなもののみ

・Item display repetitions = (AID - AUI)/AID：項目の重複の度合

まず、アンケート調査、demographic 情報の ANOVA 分析からブロック間で被験者の偏りはないことを確認した。

データに影響をおよぼす要因として、検索トピック、ユーザの検索スキル・領域知識、ユーザの学習効果などが考えられる。純粋にシステム要因だけを抽出するために、計測値を、ユーザ要因+システム要因+トピック要因+検索順序要因+誤差という線型結合で表現するモデルを導入し(Wacholder et al. 2007))、システム要因 (+誤差)を分離抽出する。また、本論文では、最初の 2 ブロックのみを分析する。

計画対比 (planned contrasts)により、第 1 ブロックと第 2 ブロックの差 (1 次差分 (d_v))と 1 次差分のグループ間の差 (2 次差分 (Δ_v))を算出し、これらの統計的な有意差を検定する。

(7) 結果

システムの評価尺度

実験条件では統制条件に比べ、ブロック 2 で GPrec, GRec が共に劣化している。

ユーザの評価尺度

有意な差があったのは CLR の Bad item ratio と Searcher Sensitivity だけであった。

ユーザの行動とシステムの応答

・CLR では Query rate が上昇し、Average list length が短かくなる

・ILR では Unique items per query が上昇する

・CLR, ILR の両方で Item display repetition が低下する

(8) 議論

実験条件によりシステム性能は確かに低下しているが (GPrec, GRec), 検索の総合的な性能は落ち

ていない（ユーザの評価尺度）。この結果は過去の研究とも一致する。ユーザはどうやってるのか? まず、やみくもに GIS フラグをつけているのではないことは Selectivity に差がないことからわかる。CLR の Bad item ratio に有意差があったことから、CLR のユーザは自分の要求レベルを下げて、本来なら marginal や bad のものまで GIS フラグをつけている可能性がある。劣化システム（CLR. ILR）のユーザは自分の検索行動をシステムに適用させて、つまり、システムの劣る部分を補うために、より高い適合文書の認識力を発揮している。また、単位時間のクエリ投入数は、システムが単調で性能が悪いと多くなる傾向にある。CLR のユーザは標準システムに比べ、短いリストを得る傾向にあるため、単位時間のクエリ投入数が多くなっている。逆に ILR のユーザは長いリストを得る傾向にあり、単位時間のクエリ投入数は増えていない。

(9) 結論

本論文では、システムの性能の違いによってユーザがどのように振舞いを変えるかを分析し、いくつかの知見が得られた。

今後、さらに他の要因も考慮する必要がある。たとえば、標準システムで検出率が十分高くないのはなぜだろうか? システムがそこそこ性能がよいとそれに満足してしまってまじめに探すことをやめてしまうのか? また、あらかじめ期待する項目数を尋ねておくと、その数に達したらやめてしまうようだ。さらに、よい結果の中の順序付けは怠る傾向にある。逆に劣化システムはユーザにやる気を起させるのかもしれない。

参考文献

Allan J., Carterette B. and Lewis J. When will information retrieval be 'Good Enough'? In 28th annual international ACM SIGIR conference on Research and development in information retrieval. (Salvador, Brazil). ACM, New York, NY, USA, 2005, 433-440.

Turpin A. H. and Hersh W. Why batch and user evaluations do not give the same results. In SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. (New Orleans, Louisiana, United States). ACM, New York, NY, USA, 2001, 225-231.

Turpin A. and Scholer F. User performance versus precision measures for simple search tasks. In SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. (Seattle, Washington, USA). ACM, New York, NY, USA, 2006, 11-18.

Wacholder N., Kelly D., Kantor P., Rittman R., Sun Y., Bai B., Small S., Yamrom B. and Strzalkowski T. A model for quantitative evaluation of an end-to-end question-answering system. J. Am. Soc. Inf. Sci. Technol., 58, 8 (2007), 1082-1099.

（徳永健伸）

4. 4. 6　Query Expansion Using Gaze-Based Feedback on the Subdocument

Georg Buscher, Andreas Dengel, and Ludger van Elst (pp. 387-394)

(1)　概要

本論文では、情報検索プロセスのパーソナライゼーションにおけるユーザの視線データを利用したフィードバック手法の評価を行っている。視線追尾装置を利用することでユーザが読んでいる文書の部分を特定することができ、それをクエリ拡張や検索結果の再ランキングのための暗黙的フィードバックとして利用する。

視線データから得られる文書の部分レベルの情報を利用する 3 種類のフィードバック手法を評価し、文書全体のコンテキストを利用するフィードバック手法との比較を行った。その結果、ユーザの文書を読む振る舞いをフィードバックとして利用することで、一般的なケースにおいて検索結果の精度が 32%もの大きな向上をもたらすことがわかった。ただし、精度の向上は閲覧する文書の内部構造や必要とする情報の種類によって異なることも確認された。

(2)　はじめに

情報検索におけるパーソナライゼーションへの取り組みは大きく 2 つの方法がある。1 つはロングタームモデリングで、これはユーザの持続的な興味をモデル化するものであり、ユーザの個人的な文書コレクションの解析やユーザの生成したクエリや訪れた Web サイトなどの行動の解析に基づくものである。もう 1 つはショートタームモデリングで、これはユーザの今現在の情報ニーズをモデル化するものであり、ユーザが今現在見ている文書やスクロール、クエリなどのアクションを利用する。

本論文ではクエリ拡張、再ランキングによって検索の精度を高めるために、目の動きをユーザのショートタームのコンテキストとして利用する方法について述べている。今日、視線追尾の技術は十分に実用化レベルに達しており、視線追尾のデータを利用することは、他の一般的な開いた文書や閲覧時間などの文書レベルでの情報に比べて、文書の部分レベルでユーザがどこを見ているのか、どのように見ているのかがわかる利点がある。そのため視線追尾情報はより正確な暗黙的フィードバックとして非常に有用な情報である。

(3)　関連研究

・ランキング精度向上のための暗黙的フィードバックに関する研究

表示時間、クリックスルー、マウス移動、スクロールなどのユーザのアクションと関連フィードバックの関連性を調べる研究は多く行われている。しかし、実験レベルでは暗黙的なアクションデータと関連フィードバックの関連性は認められているが、実世界においてはこれらのアクションデータは

解釈することが難しいことがわかっている[Radlinski and Joachims, Qiu and Cho][Agichtein et al][Shen et al.][Sugiyama et al.]。

ユーザが蓄積した文書集合全体をユーザのコンテキストと捉えて利用する手法も研究されている[Chirita et al.][K. J¨arvelin and J. Kek¨al¨ainen][J. Teevan, S. T. Dumais, and E. Horvitz]。

一方、文書の部分レベルまでを扱う情報検索のパーソナライゼーションの研究は少なく、[Golovchinsky et al.]らは、ユーザによるアノテーションを興味のマーキングとして利用して、単語、文、段落レベルでパーソナライズされたクエリを生成することで検索結果が向上することを明らかにしている。しかし、現実世界ではこうしたアノテーションがつけられることはまれである。また明示的にユーザにフィードバックを入力させる手法もあるが、これはユーザに非常に大きな認知的負荷を強いることがわかっている。視線データを利用することは、少なすぎるフィードバックデータの問題と多すぎる認知的負荷の問題の両方を解決することができる。

・情報検索における視線追尾データを利用した研究

視線追尾装置を用いた情報検索の改善という研究は少ないが、最近は目の動きを暗黙的フィードバックとして利用する研究が出てきている[Granka et al.][P. Brooks et al.] [J. Salojarvi et al.][T. Miller and S. Agne.]

(4) 視線データを利用したアノテーション抽出

本研究では、視線データから Reading と Skimming を判定し、アノテーションとして文書のメタデータを記録する視線アノテーション抽出アルゴリズムを開発し、利用している。Reading の場合は目の動きは「凝視」と「サッケード」の組み合わせになり、Skimming しながらの Reading では目は文章上を独特な動きで凝視することがわかっている。

(5) クエリ拡張のための単語抽出

クエリを生成する直前に見ていた文書の部分の情報を利用することで正確なクエリ拡張のための単語抽出を行うことができる。本研究では、視線フィードバックを利用したクエリ拡張単語抽出方法を 3 種類と基準手法として視線フィードバックを利用しないクエリ拡張単語抽出手法を 1 種類実装して比較する。

基準手法

文書全体での TF･IDF を計算し、最も高い値の単語を抽出する。

Gaze-Filter 法

視線アノテーションがつけられている文書部分においてのみ TF･IDF を計算し、単語を抽出する。

Gaze-Length-Filter 法

視線アノテーションがつけられている文書部分において、単語 t の興味度を計算する。

短い領域部分での凝視はユーザの興味を示さないという仮説。

$$\mathrm{interest}(t)=\frac{\mathrm{LA}(t)}{\mathrm{LA}(t)+\mathrm{SA}(t)} \qquad (式1)$$

SA(ｔ)：単語 t が出現する 230 文字未満の視線アノテーション領域の数

LA(ｔ)：単語 t が出現する 230 文字以上の視線アノテーション領域の数

Reading-Speed 法

視線アノテーションにおける Reading の比率を算出し、Gaze-Filter 法と組み合わせて単語の興味度を計算する。

より注意深く読んだ部分がユーザの興味を表しているという仮説。

$$\mathrm{r}(t)=\frac{1}{|\mathrm{At}|}\sum_{a\in \mathrm{At}}\mathrm{r}(a) \qquad (式2)$$

単語 t を含むすべてのアノテーション At における Reading スコア r （ｔ）を算出

(6) 実験手法

・被験者タスク

被験者は新聞記事を書くジャーナリストという設定で実験を行い、メールで書くべき記事のトピックを指示される。メールには一行のトピックの説明と 4 つの関連文書が添付される。被験者は 15 分程度で 4 つの関連文書を読み、より詳細な情報を得るために 3 種類の異なるクエリの検索を行う。

1）各トピックについて書く材料を探すためのクエリ検索

2）添付文書にあるサブトピックに関するクエリ検索

3）添付文書に含まれていない関連トピックに関する材料を探すためのクエリ検索

検索実施後、被験者は結果の上位 20 文書について明示的な関連性評価を行う。1 つ目のトピックが終わったら、2 つ目のトピックについて作業を行う。被験者は大学生 21 人で、60 分～80 分の作業に対して 10 ユーロの報酬を与える。

・検索システム

関連添付文書からは視線データによるアノテーションが抽出される(被験者にはアノテーションは見えない)。被験者がひとつのトピックに関するすべての添付文書に目を通すと、システムは自動的に (5) に示した 4 種類の方法を用いてクエリ拡張単語を抽出し、4 種類のクエリを生成する。クエリは

ユーザが生成したもともとのクエリ単語と、自動抽出されたスコア上位 50 個の単語を含んでいる。単語には重みをつけることができ、60%は自動生成された単語に、40%はユーザが生成した単語に付与される。生成された 4 つのクエリを用いて以下の検索プロセスが実行される。

1）検索結果集合は拡張されていないオリジナルのユーザクエリにより生成する。

2）拡張クエリを用いて検索結果の再ランキングを行う（4 つのランキング結果)。

3）4 つのランキング結果をマージして 1 つのユーザ提示用のランキングを生成する。

・評価指標

以下の 3 種類の評価指標を用いて評価している。

Precision at K

あるクエリ q の検索結果上位 K 文書内に関連文書がある割合。

MAP：Mean Average Precision

1 つの単語抽出手法について 1 つの評価値を返す。

$$MAP(Q)=\frac{1}{|Q|}\sum_{q\in Q}\frac{1}{K}\sum_{K=1}^{K}Prec@k(q) \qquad (式3)$$

Q はクエリー集合、K は 5 または 10 で計算

DCG at K：The Discounted Cumulative Gain

あるクエリ q について最初の K 個の結果のランキングを考慮して算出。

$$DCG@K(q)=\sum_{j=1}^{K}(2^{r(j)}-1)/\log(1+j) \qquad (式4)$$

r（j）はクエリ q に対する検索結果順位 j 番目に対する関連度のレーティングを表す整数値

(7) 結果

・全体評価

3 種類の評価指標とも、視線データを利用しない基準手法に対して、視線データを利用したクエリ拡張手法の方が大きな向上（MAP で 21.9%～31.7%の向上）が見られ、文書全体レベルでのフィードバックよりも、文書の部分レベルでの視線データによるフィードバックの方が検索精度を大きく改善する効果があることが確認された。

視線データを利用した手法間の比較では、Reading-Speed 法は Gaze-Filter 法、Gaze-Length-Filter 法に比べて少し効果が落ちること、また Gaze-Length-Filter 法は Gaze-Filter 法に比べて少し良いが大きな差ではないことがわかった。

・詳細分析

文書が強い構造を持つ場合は、基準手法に対して視線データを利用したフィードバックは大きな改善効果が現れているが、弱い構造の文書（関連情報が文書内に広く分散している文書）の場合は改善は見られない。

三種類の検索タスク（メイントピックに関する検索、サブトピックに関する検索、関連トピックに関する検索）において、基準手法ではメイントピックからサブトピック、関連トピックとなるほど評価指標の値が下がっている。それに対して Gaze-Filter 法、Gaze-Length-Filter 法では評価指標地の下がり幅が小さい。これはユーザの情報ニーズがコンテキストの中心から離れるほど、結果の関連性が低くなるからであると考えられる。

(8)　考察および今後の課題

本論文では、視線データを用いた文書部分レベルでのフィードバックが情報検索におけるクエリ拡張や検索結果の再ランキングに大きな効果があることを示した。今回の実験は 1 種類のタスクのみを対象にした実験であり、今後視線データによるフィードバックがどのような種類のタスクに最も効果的であるかについて調べる余地がある。

他の文書部分レベルを利用する関連フィードバック手法と比較しても、視線データを用いるフィードバック手法は十分に正確、かつ十分な量のデータが利用でき、ユーザへの認知的負荷も少ないという特長がある。

視線データを用いるフィードバックは､他の手法を組み合わせることによるさらなる精度を向上や､クエリ単語拡張以外の情報検索問題への応用など、今後も研究の余地が多くある領域である。

（長束 哲郎）

4. 4. 7　Crosslingual Location Search

Tanuja Joshi, Joseph Joy, Tobias Kellner, Udayan Khurana, A Kumaran, Vibhuti Sengar (pp. 211-218)

(1)　概要

本論文では、クエリとは異なる言語で用意された地図データを検索する問題（Crosslingual Location Search：言語横断位置検索)を「空間的な一貫性を保ちつつ、クエリとデータの字面の類似性を最大にするような地理的範囲を探すこと」と定式化した。翻訳や翻字（transliteration）の曖昧性の問題を解決するために空間的制約を導入した。

英語の地図データに対してアラビア語、ヒンズー語、および日本語で検索する実験を行った結果、

現在利用されている単言語の地図データ検索システムと transliteration を結合したものよりも、本論文で提案したシステムの方が良い結果を示すことがわかった。

(2) はじめに

Location Search（位置検索）では、構造化された情報と、構造化されていない情報との 2 種類の情報を扱う。構造化された情報を扱うものとしては、住所を地図上の位置にマッピングする “Address geocoding” がある。それに対して、構造化されていない情報とは、例えば、“Corner of Grand Ave and Walnut St in Everett”のような表現を指す。構造化されていない情報から位置を探すことができるだけでも大変有用であるが、本論文ではさらに、言語を横断して検索する方法について述べる。Google Maps, Yahoo Maps, Windows Live Local などの地図検索サービスがすでにあるが、単言語のみの取り扱いである。

本論文の手法は、翻訳や transliteration と空間的制約とを統合した点が新しい。本論文では主に以下の内容について説明する。

・Crosslingual location search を入力から出力までひとつながりの問題として定式化

・空間的制約の導入

・比較評価

(3) 背景

著者らは、本論文に先立って、単言語での位置検索のシステムの提案と実験を行っている [1)]。単言語での位置検索での課題としては以下の点などがある。

・構造化されていないクエリが様々な形で入ってくる。

・クエリ中の言葉（句）の区切りがあいまいである。

・スペルミスや矛盾する内容を含むクエリがある。

・よくある名称（例：”Main St”）やよく似た名称（例："Auckland” の別表記多数）を持つ位置がある。

言語横断位置検索の場合は、上記の課題に加えて以下の問題点にも対処しなければならない。

① 固有名詞は翻字（transliteration）しなければならない。

地理上のものを示す名称は固有名詞であることが多く、一般的には翻訳では処理しきれない。同種の文字を使う言語間でもスペルが異なることはあるし、文字が違う言語間では transliteration をする必要がある。Transliteration とは、原言語での発音を、目的言語での発音ルールに従いながら移し変える処理である。双方の言語において、発音とスペルの対応は、必ずしも 1 対 1 ではないので、transliteration の処理にはあいまい性がつきまとう。

② 一般的な名詞は transliteration と翻訳とのいずれを行うべきか確定できない。

例えば ”Hospital” を transliteration により「ホスピタル」と変換すべきか、翻訳して「病院」とすべきかは、それぞれのケースで異なってくる可能性があるので、常に両方のパターンを考慮しなければならない。

③ 言語や住所表記方法によって語順が違う可能性がある。

これらの crosslingual の場合の課題を考慮すれば、単純に transliteration または翻訳を 単言語の位置検索につなげただけでは十分な性能は出ないであろうことが想像できる。

(4) アプローチとアルゴリズム

本論文でのアプローチは、上記の課題から生じる様々なあいまい性を解消するために空間的制約を導入することが鍵である。

原言語での位置クエリは、いくつかの部分に分けられ（分け方にもバリエーションがある）、その各部分ごとにいくつかの transliteration または翻訳の結果が用意され、組み合わせたものが目的言語でのクエリ解釈候補となる。このクエリ解釈候補のうち、各部分が空間的に重なり合っている(spatially overlap)ものが正しい候補であると考える。同時に、重なり合った位置が位置検索の結果となる。

① 問題定義

クエリ Q =(q_1, q_2, q_3, …, q_n) は、トークン q_i （単語、数字、など）のリストである。Q 中の各語が集まって、いずれかの空間領域を示す。クエリ・サブシーケンス $q_{i\text{-}j}$=(q_i, q_{i+1}, q_{i+2}, …, q_j) は、Q 中の隣接するトークンのリストである。

サブシーケンスのうちのいくつかは、一定の形状の空間領域の固有表現(named entity)となっているはずである。

位置あるいは物体 エンティティ e=(g,{n_1,n_2,…,n_n}) は、幾何学的形状 g と、名称 n のペアで表される。ただし、あるエンティティは複数の呼び名を持つことが可能なので、n はリストとなっている。空間リポジトリ（位置情報のデータを収納した場所）S は、このエンティティの集合であるとする。

このとき、空間リポジトリ S 対するクエリ Q の解釈 I={($q_{i1\text{-}j1}$,e_1), ($q_{i2\text{-}j2}$,e_2), …, ($q_{im\text{-}jm}$,e_m)} とは、「Q の重なりがないサブシーケンス群を、S 中のエンティティに 1 対 1 で対応づけること」である。もちろん数多くの解釈がありうるが、位置検索の目的は、「最も矛盾がない解釈を見つけること」であると言うことができる。

「矛盾がない解釈」については、以下の 2 つの性質を考慮する必要がある。

(a) 文字列の親和性（類似性）

クエリ中のサブシーケンスと、その解釈中のエンティティの名称（の発音）が似ている必要がある。意味的に似通っていることも含めて考える（例： “Main Street” と “Main Road”）。

(b) 空間的一貫性

解釈中のすべてのエンティティが空間的に一貫性を保っている必要がある。これはつまり、すべてのエンティティが空間上で重なりあっている（部分がある）ということである。エンティティを外側に境界幅 β だけ広げた領域がすべて重なりあう部分が存在すれば、それらのエンティティは空間的に一貫性があるとする。重なり合った領域のことは、解釈の幾何学的範囲(geometric scope)と呼ぶことにする。図 4.4.7-1 では、エンティティ e_1 と e_2 は空間的に一貫性があるが、 e_3 はいずれのエンティティとも一貫性を持たない。エンティティ e_1 と e_2 が示す幾何学的範囲は図で影になっている部分である。

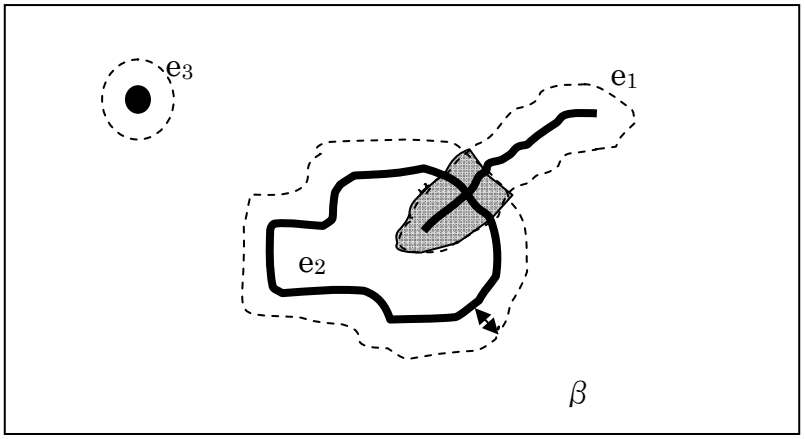


図 4.4.7-1 空間的一貫性

上記の定義により、言語横断の位置検索は「空間的一貫性を維持しながら文字列の親和性を最大にする解釈を探すこと、そして、その解釈の幾何学的範囲を返すこと」と定義できる。

② 解釈の計算アルゴリズム

解釈のランクつきリストを返すアルゴリズム XL-QUERY を提案している。

まず MC(match candidate)=($q_{i\text{-}j}$, name) のリストを作成する。MC とは、クエリ・サブシーケンス $q_{i\text{-}j}$ を、空間リポジトリ S 中の name に対応付けたものである。手順はおおよそ以下の通りである。

- クエリ中の各単語に対して transliteration と翻訳（訳語の辞書引き）を行い、さらに transliteration 結果の抽象化（発音に着目した変換）を行う。
- すべてのサブシーケンス（単語が n 個なら、サブシーケンスは n(n+1)/2 個）について、
  - transliteration と翻訳の結果を組み合わせた候補と、抽象化された語を使った候補とを、それぞれランク付けしたリストを作る。

- それぞれのリストを、Fuzzy Text Index FI(S) または $FI_A(S)$ と照合する。FI(S)は、空間リポジトリ S 内のエンティティの名称のインデックスであり、$FI_A(S)$はそれらの名称を抽象化したもののインデックスである。
- 照合後の 2 種類のリストの集合和が MC のリストとなる。

その後、空間的一貫性を保ちながら、深さ優先探索で再帰的に解釈を計算していく。具体的には、以下の手順を経る。

(a) MC のリストから、一つの MC を選び、その MC を現在の解釈に追加する。

(b) その MC と現在の Focus （対象となる空間領域）の交叉空間を新しい Focus とする。

(c) その MC と空間的にも文字列上も矛盾がない MC だけを残した MC のリストを作り、(a) に戻る（MC リストが空になったら解釈が一つ成立したとしてその枝の探索は終了)。

と同時に、今対象としていた MC のリストから、対象にしていた MC を取り除いたリストを新たに MC のリストとして、(a) からの処理を行う。

最後に、発見された解釈をランク付けして出力する。

③ Machine Transliteration

Transliteration は統計的に学習させたシステムを用いた。Grapheme-to-grapheme のアプローチ（原言語の綴りから、対象言語を綴りに直接置き換える方法）を取った。学習の分類器には最大エントロピー法を用いた場合の結果が今回の実験には最も良好であった。

(5) 実装

ある言語で位置検索のクエリを受け付けて、その解釈を出力するシステムを実装した。Geocoding（緯度経度の出力）部分はシステムの深部に組み込まれており、言語や対象領域に依存しないようになっている。そのため、新しい言語での検索や、新しい地域を対象にすることも容易である。現在は、英語で用意された地図データに対して、アラビア語、英語、ヒンズー語、日本語での検索をサポートしている。

Machine transliteration では、ユニットの対応付けには Viterbi training を用いた。原言語の文字を、前後 3 文字ずつの情報を使って、対象言語の文字列に transliteration できる確率を求める。15000 件のデータで学習した。

Transliteration の実験（他の三言語から英語への変換）結果では、上位 4 件にかなり正解に近い結果が含まれることがわかった。

位置検索のシステムは、イギリス・アメリカ・インドの大都市を一つずつ選び、24 万エンティティ（通りの名前、地名、郵便番号）をインデックスしてある。

(6) 実験結果

ベースラインとして、商用の単言語の位置検索の前に transliteration を通したものを用意した。

テストクエリは 275 で、うち 155 は構造化された住所（地域の慣行に従った記述法に基づく）にで、残りが構造化されていない記述である。正解位置の 1km 以内を示したかどうかで結果を確認した。

構造化された住所での実験結果では、単言語のアメリカの地名に関する検索の場合は、ベースラインのシステムが上回ったが、特に言語横断の検索結果は本論文のシステムがかなり差をつけて上回った。

構造化されていないデータの場合には、ベースラインのシステムはほとんどできていなかったのに対し、本論文のシステムは 70-90%代の高い性能を示した。

以上の結果から、入力から出力までひとつながりにしたシステムに transliteration を組み込むことにより、言語横断での位置検索でも、単言語での位置検索に匹敵する性能を上げることができることがわかった。

(7) 今後

すでに取り組み中の試みとして、「～の近くの病院」といった “what” に相当するものも扱えるように拡張して、ローカルサーチをサポートする予定である。また、分散システム化してデータを 2 ケタ増にすることと、コーパスを探して翻訳できる言語対を増やすことを考えている。

参考文献

1) Vibhuti Sengar, Tanuja Joshi, Joseph Joy, Samarth Prakash and Kentaro Toyama: “Robust Location Search from Text Queries”, In Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems (2007).

[池野 篤司]

4.5 WWW2008 参加報告

4.5.1 はじめに

2008 年 4 月 21～25 日に中国の北京で開催された「第 17 回 World Wide Web 会議（以下、WWW2008)」について報告する。インターネットの普及と関連技術の発展によって、ワールドワイドウェブに関する技術の重要性が高まっている。こうした状況の中、World Wide Web 会議は回を重ねるごとに規模を拡大している。WWW2008 本会議の口頭発表（「Refereed Papers Track」呼ばれる）

には 880 件という過去最多の投稿があり、その中から 103 件の投稿が採択された。12%という採択率は情報処理関連の国際会議としては狭き門である。参加者は約 800 名だった。プログラムは以下の通りであった。

・ 4 月 21 日: チュートリアル（午前・午後 3 件、午前のみ 6 件、午後のみ 6 件）
・ 4 月 22 日: ワークショップ（午前・午後 8 件、午前のみ 2 件、午後のみ 2 件）
・ 4 月 23～25 日: 本会議

本会議では、基調講演、口頭発表、ポスター発表、展示等があった。以下、チュートリアル、ワークショップ、本会議の口頭発表について報告する。

4.5.2 チュートリアル

筆者は、以下に示す 2 つのチュートリアルに参加した。

・ Learning to rank for information retrieval (Tie-Yan Liu, Microsoft Research Asia, China)

Learning to rank (L2R)は、情報検索研究において数年前から台頭した技術動向である。機械学習手法を応用して文書を順位付けるためのモデルや関数を学習する。本チュートリアルでは L2R の技術動向や利用可能な資源等について紹介された. チュートリアルの資料は以下の URL より入手可能である。

http://research.microsoft.com/users/tyliu/

・ Online advertising: business models, technologies and issues (James Shanahan, Independent Consultant, San Francisco, USA)

Web 上のビジネスモデルとして急速に発展している「Web 広告」について、最先端の研究動向が解説された。WWW2008 の本会議や SIGIR 等の情報検索に関する国際会議においても Web 広告に関する研究発表が増えている。

4.5.3 ワークショップ

筆者は、以下に示す 2 つのワークショップに参加した。

・ Question Answering on the Web

質問応答に関するワークショップであり、研究発表というよりも議論や啓発を意識した構成と内容であった。前半と後半に分かれており、各パートの冒頭で Lead talk があり、その後で短めの発表が数件行われた。前半の Lead talk は、Web から質問と回答のペアを自動抽出し、さらに抽出した質問・回答ペアから質問応答システムを学習する内容だった。後半の Lead talk は、Yahoo! Answer（日本の Yahoo!知恵袋）に蓄積された質問と回答を情報源として利用する質問応答に関する種々の報告があった。日本では国立情報学研究所を通して Yahoo!知恵袋の質問・回答

集合が研究利用で配布されていることに対して、Yahoo! Answer のデータは非公開である。本ワークショップの情報は以下の URL より入手可能である。http://www.buyans.com/qaweb2008/

・ NLP Challenges in the Information Explosion Era

文科省科研費特定領域「情報爆発時代に向けた新しい IT 基盤技術の研究（通称、情報爆発)」に関わっている研究者のうち、自然言語処理や情報検索の研究者が有志で企画した。情報爆発時代における自然言語処理の役割やあり方等について検討することを目的とした、研究発表を中心としたワークショップである。Web 上の人物検索における名寄せ、批評や意見の分析、対話、分類などに関する発表が合計 6 件行われた。本ワークショップの情報は以下の URL より入手可能である。

http://www.slis.tsukuba.ac.jp/~fujii/NLPIX2008/

4. 5. 4　本会議

本会議の口頭発表（Refereed Papers Track）では 4 つのセッションが並行して行われた。以下にセッション名を示す。また、セッションの件数を括弧内に示す。

Search (5), Data Mining (4), Social Networks (3), Semantic Web (3), Internet Monetization (3), XML (2), Security (2), Web Service Deployment (1), Web Service Composition (1), Web Engineering (1), Technology for Developing Regions (1), Rich Media (1), Performance and Scalability (1), Mobility (1), Browsers and UI (1)

これらのうち、Social Networks, Internet Monetization, Rich Media は WWW2008 でできた新しいセッションである。セッション数が最も多いのは Search に関するセッションであり、当該セッションは採択率も最も厳しかった。Web 関連技術において情報検索に関する研究開発が活発であることが分かる。

筆者は、「Search」と「Data Mining」に関するセッションだけを聴講した。12%の採択率を勝ち抜いた研究だけあって、聞き応えのある発表が多かった。ACL や SIGIR といった国際会議では予稿論文のページ数が 8 ページであるのに対して、WWW2008 では最大 10 ページと長い。そのため、複数の要素を含んだ大きめの研究テーマであるか、もしくは技術の細部に渡って入念に工夫が施された研究内容が多かった。相応の研究体制と年月をかけて成熟した内容が発表されているという印象を受けた。Search と Data Mining のセッションを聴講して印象に残ったキーワードは、「検索ログ」、「Community QA」、「トピックモデル」の 3 つだった。

検索ログとは、ユーザが Web 検索システムに入力した検索質問と、検索結果のうちクリックされたページ（クリックスルー）に関する履歴情報である。検索ログを分析することで、「キーワードとページの関係」、「キーワードどうしの関係」、「ページどうしの関係」など種々の関係を発見し、検索に応

用することが可能になる。ただし、分析に値する大規模な検索ログは検索サービス大手の企業でない限り入手が難しい。

Community QA とは、Yahoo!知恵袋や OKWave のような Web サイトの投稿機能を介して、人間どうしが行う質問応答である。Community QA に蓄積された質問と回答の集合を用いて質問応答のモデルやシステムを学習する手法がいくつか提案された。

トピックモデルに関する発表では、PLSA（Probabilistic Latent Semantic Analysis）や LDA（Latent Dirichlet Allocation）などの確率的なトピックモデルを拡張して、あるトピックを表現するための「語」に関する分布を高精度で学習するための手法が提案された。

また、本分科会の調査テーマである「状況依存型サービス」に関連する発表について、以下、発表タイトルと概要を示す。

- Modeling Anchor Text and Classifying Queries to Enhance Web Document Retrieval
  ユーザの検索意図に基づいて検索モデルを使い分ける。
- Spatial Variation in Search Engine Queries
  検索質問の地理的依存性を特定する。
- Mining the Search Trails of Surfing Crowds: Identifying Relevant Websites From User Activity
  Web 検索においてユーザの閲覧履歴を利用する。
- Using the Wisdom of the Crowds for Keyword Generation
- Contextual Advertising by Combining Relevance with Click Feedback
  検索連動型広告を高度化する。
- Ranking Refinement and Its Application to Information Retrieval
- Learning to Rank Relational Objects and Its Application to Web Search
- Personalized Interactive Faceted Search
  ユーザからのフィードバックによって検索精度を向上させる。

4.5.5 おわりに

本節は、WWW2008 のチュートリアル、ワークショップ、本会議の口頭発表に関する参加報告を行った。2009 年の大会は 4 月 20～24 日にスペインのマドリードで開催される予定である。

- WWW2008 のホームページ　http://www2008.org/
- WWW2009 のホームページ　http://www2009.org/

（藤井 敦）

知識情報処理技術に関する調査研究報告書

発　行　日　平成22年３月
編集・発行　社団法人電子情報技術産業協会
　知識情報処理技術専門委員会
　インダストリ・システム部　企画グループ
〒100-0004
東京都千代田区大手町１丁目１番３号
大手センタービル
TEL　(03) 5218-1057

印　　刷　三協印刷株式会社