タイトル	円滑なコミュニケーションのための自然言語処理
著者	渋木，英潔
引用	北海学園大学経営論集，4(2)：55-85
発行日	2006-09-30

円滑なコミュニケーションのための自然言語処理

渋　木　英　潔

目　次

1章　序　　論

現在の「ものあまり」の時代において，消費者の個性化や多様化の傾向は今後も進んでいくと考えられる。それゆえ，マーケット・インや顧客満足といったキーワードにみられるように，消費者や利用者の側に立って顧客のニーズやウォンツを満たすビジネスの構築が重要視されている。マーケット・インとは，全ては市場から始まるという発想のもとで市場のニーズを引き出し，これに基づいて商品企画や生産・販売戦略などを立てることである。そのために，顧客とのコミュニケーションを密接にし，顧客のニーズを収集・分析する努力が望まれている。

顧客のニーズには，顕在的ニーズと潜在的ニーズが存在し，潜在的ニーズは顧客自身にも意識されていない。潜在的ニーズを引き出すために，顧客とのコミュニケーションを通して，顧客の発するあらゆる情報に耳を傾け，内容を吟味することが必要である。

情報技術の発達により，情報収集・分析の手段としてほとんどあらゆる企業でコンピュータが利用されており，利用形態もますます高度化，多機能化している。Web を利用したマーケティング・リサーチや，e-mail で寄せられるレビューなどはその一例である。情報技術を活用することで，従来よりも大規模なデータを，低コスト，かつ，短時間で収集・伝達することが可能となっている。しか

しながら，利用者から収集されるデータは玉石混交であり，収集されたデータからビジネスに有益な情報を抽出するためには，データに対して種々の分析を行う必要がある。

上述した潜在的なニーズを引き出すという観点からは，企業側で用意した項目の中から選択させる形式のデータよりも，顧客が自発的に記述した自由形式のデータの方が有益である(注1)。したがって，収集されるデータには選択された番号などの離散的な記号ではなく，普通の言葉で書かれたデータとなり，言語で記述されたデータをコンピュータ上で処理するための機構がますます重要となってくる。

プログラミング言語などの人工言語と区別する意味で，人間の使う言語を自然言語（Natural Language）といい，自然言語をコンピュータ上で処理する技術を自然言語処理（Natural Language Processing，NLP）という。

自然言語処理の身近な例として，ワープロソフトの「かな漢字変換」，Webにおける「情報検索」や「機械翻訳」などがあげられる。

「かな漢字変換」は，漢字の読みをキーボードから入力し，コンピュータによって適切な漢字に自動変換する文字入力方式である。同音異語など，いくつかの漢字候補があり，その中から選択する場合などでは，どのような順序で変換候補を表示するかによって利便性が左右されるため，一般的な使われやすさや文脈などを考慮する必要がある。

「情報検索」は，蓄積された大量のデータの中から必要なデータを検索することを指す。例として，Webページを検索するGoogleなどのサーチエンジンが挙げられる。今後さらに大量の情報が氾濫することを考慮すると，より高性能の情報検索アルゴリズムの開発が必須課題となっている。

「機械翻訳」は，ある言語を別の言語へコンピュータによって変換する技術をいう。すでに実用化され，翻訳ソフトとして多くのソフトが発売されており，Web上で利用できる例としては，Exciteなどのオンライン翻訳サービスなどがある。しかしながら，これらの翻訳システムでは，翻訳の対象を特定の狭い分野に限定することで精度を上げており，頑健性や網羅性という観点からは課題が残されている。

以上のように，自然言語処理は身近な技術として役立っており，企業経営のための情報処理システムにおいても，自然言語処理に対する重要性が高まっている。例えば，日本信販ではWebサイトを通じて顧客ニーズの把握を目指している。そのため，サイト内検索に入力された膨大なキーワードの分析に着手したが，人手による分析では莫大な時間と手間がかかり，顧客の声をタイムリーに活用できないという問題があった。この問題を解決するために，蓄積された質問ログから顧客のニーズを正確に把握し，Webサイトのコンテンツに反映させることを目標として，自然言語処理を活用したサーチエンジン「ASK NICOS」を導入した。「ASK NICOS」は，集積されたログを詳細に分析することで顧客のニーズを明確化し，コンテンツの充実とWebサイトでの効果的なプロモーションの展開の助けとなっている。しかしながら，その効果は単純な作業で処理できる範囲にとどまっており，未だ改善の余地が残されている。

また，野村證券ではWebサイトを顧客との新たなコミュニケーション手段として位置づけサイト運営を行っていた。しかしながら，サイトの情報量が増加していく過程で，顧客が目的とする情報を的確に紹介するためには，従来のキーワードによる検索では不十分であった。そのことから，サーチエンジンの重要性を感じるようになり，自然文による検索を可能とするために自然言語処理を活用したサーチエンジンを導入した。

このように，自然言語処理は膨大な情報の中から適切な情報を引き出すための核心技術として活用されている。

別の例としては，特許文書の解析技術があげられる。近年の知的財産権を重視する考え方から，企業の特許出願件数は年間40万件以上に増加しており，特許を出願するための出願済み特許の調査業務が重要になってきている。しかしながら，特許文書は曖昧性を排除するため，図1.1の例のように独特の言い回しを多用した回りくどい文章で発明の内容を説明している。そのため，文章を解読し内容を正確に理解するには専門家でも時間がかかるのが実情となっている。

このような背景から，特許文書を理解しやすい表現に変換する技術への要望が高まっている。これを受けて，NTTデータは自然言語処理を活用した特許文書解析技術を製品化に向けて開発しており，調査業務を支援するツールとしての利用が見込まれている。

以上の例に代表されるように，自然言語処理はビジネス情報の理解・伝達を円滑にするツール，あるいは，事業対象である製品として企業と密接に関連している。現在の技術では，経営における意思決定の指標となる情報を抽出するための支援ツールという段階にとどまっているが，将来的には意思決定そのものを補佐する技術としての活用が期待されている。すなわち，単に情報伝達メディアとしての言語を処理する技術ではなく，思考メディアとしての言語を処理するための，真に知的な言語処理を目指すべきである。

LCD 15の液晶画面を縦長画面として使用する第1の状態と上記液晶画面を横長画面として使用する第2の状態とを検出する状態検出部19と，上記縦長画面の表示サイズに対応する第1の画像サイズと該第1の画像サイズより大きな第2の画像サイズの少なくとも2つのサイズから撮影時の画像サイズが選択的に設定されるカメラモジュール17と，上記第1の画像サイズが設定された状態でカメラモジュールを起動した場合に，状態検出部19にて上記第2の状態が検出されると，カメラモジュール17に上記第2の画像サイズで撮影を行わせ，該撮影した画像から上記第1の画像サイズの画像を切り出して保存する制御部10とを有する。

図1.1　特許明細書の例（特許公開番号2005-311851「カメラ付き携帯電話機および画像保存方法」より引用）

しかしながら，自然言語処理の歴史は半世紀程度と比較的短く，ビジネス社会への普及とともに，自然言語処理の技術自体もまだ発展途上にあるといえる(注2)。

自然言語処理は，言語という人間の知的活動において重要な役割を担う情報を扱う技術といえる。したがって，その技術の未成熟さは単純に目的とする処理の問題発生率が高まるだけではなく，自然言語処理そのものに対する拒否反応を引き起こす可能性がある。最悪の場合，そのような拒否反応が自然言語処理の普及を妨げる要因となり，上述した問題の解決を困難なものとし，今後のビジネスにおける発展の機会を逃す恐れがある。

言語という本来人間が扱う情報をコンピュータに処理させることは，単に技術的な問題であるだけではなく，心理的な問題なども考慮に入れる必要があると考える。自然言語処理は工学の一分野であり，これまで工学的な見地から論じられることが多かった。しかしながら，工学的に優れた技術が必ずしも一般に普及するわけではない。それゆえ，より広範，包括的な観点から論じるために，自然言語処理を包含する観点から議論することが必要である。

既に述べたように，今後の企業における自然言語処理の重要性がますます高まっており，その重要性は，情報の通信・伝達手段としての技術よりも，得られる情報を効果的に分析・加工する技術の方に大きいという状況を迎えている。情報を分析・加工するには専門的な知識が必要であり，そうした本来人間がもっている知識をコンピュータに移転して処理を行わせることを一般に知識処理(注3)とい

う。

自然言語処理は，知識処理の対象を言語に限定したものとみなすことができるため，知識処理における問題点や対策などは自然言語処理にも当てはまることが多い。そこで，人間がもつ知識をコンピュータに移転させることによる問題点を考察するために，知識処理を人間のもつ知識を移転・活用したコンピュータ処理と定義し，知識処理という観点から自然言語処理の問題に言及していくこととする。

今後の知識処理は，現在よりも効果的な分析・加工を行うために，さらに知的な領域に踏み込む，すなわち，人工知能との結びつきを強める必要があると考えられる。しかしながら，コンピュータが知的な領域を扱う場合には，そうでない場合と比較して以下のような問題を抱えることとなる。1点目は，変換ミスや翻訳ミスなど，コンピュータの誤りに対してユーザが過剰な拒否反応を起こす可能性があること，2点目は，知識処理システムを導入した場合に要求されるコストの大きさである。

1点目の過剰反応の原因を更に分類すると，コンピュータが「知」を扱うことへの抵抗感によるものと，コンピュータに対する過剰な期待が裏切られたことによる失望感によるものとに分離することができるであろう。前者の抵抗感は，機械が人間の領域を侵犯することが原因であると考えられているが，この問題は非常に深い問題であり，ユーザへの啓蒙活動をはじめとして，様々な面から解決の手段を講じていく必要があると考えられる。後者の失望感は，コンピュータの誤りが人間では考えられないような誤りであることへの苛立ちなどが原因であると考えられる。この問題も決して浅い問題ではないが，処理技術の見直しによって解決できる部分が大きいと考えられる。

2点目のコストの問題に関しては，知識処理システムを導入する時点で支払うコストと，利用している間，支払い続ける導入後のコストに分類することができる。

情報は高い鮮度が求められるため，知識処理に用いられる知識も新鮮な情報に対応できるようにする必要がある。また，処理に必要とされる，システムがもつべき知識は利用者ごとに異なる場合が多く，有効に活用するためには利用者ごとにシステムの知識をカスタマイズする必要がある。このようなことから，知識処理システムを利用するために，導入後に支払うコストは，ルーチンワークを代行する類のシステムと比較して大きいと考えられる。

また，支払うコストの種類の面からも，金銭的なコストだけではなく，導入したシステムを使いこなすための技能的なコストを考慮する必要がある。上で述べたように，有効に活用するためには知識を利用者ごとにカスタマイズする必要がある。利用者ごとのカスタマイズを行うためには，何らかの形で利用者側の作業が必要であるが，この作業に要求される技能があまりに煩雑なものであった場合にも，一般利用者への普及を妨げる要因となるであろう。

したがって，利用者が求める理想的なシステムとは，利用者がそのシステムを使用しているうちに，システムの方で利用者に必要な知識を推測し，自動的に知識のカスタマイズを行うような形式でありと考えられる。

以上の問題を解決し，理想を実現するための手段として，コンピュータに学習を行わせること，すなわち，機械学習の利用が考えられる。機械学習は，人工知能における研究課題の一つで，人間が自然に行っている学習能力と同様の機能をコンピュータで実現させるための技術・手法のことである。機械学習を組み込んだシステムであれば，知識処理システムを保守するために必要なコストの内，最も不明瞭で大きいと考えられる知識整備に必

要なコストを大幅に抑えることができる。

以上から，知識処理システムの普及のためには，ユーザに拒否反応を抱かせないような知識処理技術を開発することと，知識の学習において人手による労力を軽減することが必要であると考えられる。このことは自然言語処理システムの普及においても同様であると考えられる。

自然言語処理において，ユーザに拒否反応を抱かせないためには，処理過程に人間の言語処理過程との親和性をもたせる必要があると考える。それゆえ，今後の自然言語処理の発展のためには，かつて行き詰まりを見せた，文法・意味理論を再び考慮する必要がある。かつての行き詰まりは，人手によるコンピュータへの知識付与というアプローチによる部分が大きいと考え，知識付与の部分に，現在，主流となっているコーパス・統計モデルのアプローチを応用することで，意味に関する知識を学習する段階に進むべきであることを提言する。

そのような段階へ進むための第一歩として，コーパスから学習された知識を用いて，深層格を自動的に推測する手法DCAPRの提案を行う。深層格とは，主体や対象など動作に対する意味的な役割を表す格のことであり，深層格を明らかにすることはコンピュータに意味を扱わせるための基礎となる。今後のグローバル社会における活用を目指して，DCAPRは深層格推測に必要な知識を，深層格選好と一文一格の原理という言語普遍の枠組みの中で学習を行う。具体的には，網羅性とコストの問題に対処するため，一定量のタグ付きコーパスから得られた知識を中核として，タグなしコーパスから規則として推測に用いる知識を学習する。DCAPRは，単語の概念ごとに，それぞれ特定の深層格に解釈される傾向（深層格選好）が存在するという仮定に基づき，一定量のタグ付きコーパスから深層格選好を計算した後，その値を用いてタグなしコーパスから多様な言語表現に対処するための規則を学習する。深層格選好を手がかりの主体とすることにより，DCAPRは言語の違いに依存せずに深層格を推測することが可能となる。言語非依存性を検討するため，雑誌や新聞記事などの文を集めた日本語と英語のコーパスを用いて，一つの動詞と二つの名詞で構成される単文を対象に実験を行い，その結果，クローズドデータにおける日本語の精度81.2%（再現率78.0%），英語の精度78.5%（再現率77.8%），オープンデータにおける日本語の精度69.5%（再現率62.1%），英語の精度73.5%（再現率73.3%）となり，従来手法と同程度の精度で日本語と英語の深層格を推測することができたことを示す。また，今後のビジネスのための応用例の一つとして，自動要約システムへの応用を試みる。

本論文では，以上で述べた，コンピュータ上で言語情報を扱う場合に起こりうる問題点を明らかにし，注意すべき点の指摘を行う。また，指摘した問題点を解決するような試みを同時に行うことが必要であると考え，以上の考察の結果を反映した自然言語処理として深層格推測手法の提案を行い，実験による考察を行う。

なお，本論文で想定するコミュニケーションは言語を媒介とした情報伝達全般を指しており，メールやチャットなどによる特定の相手とのコミュニケーションだけではなく，不特定の人物に向けて情報を発信することや，そのように発信された情報の中から自分に必要な情報を得ることも含まれている。

本論文の構成は以下の通りである。

1章は「序論」であり，研究背景および研究目的について述べる。

2章は「情報化社会における知識処理」である。2.1節では，今日の情報化社会の状況と今後の動向を述べる。2.2節では，知識処理と自然言語処理の関係について論じる。

2.3節では，現在，自然言語処理がビジネスの世界において，どのように活用されており，今後どのように活用されていくか幾つかの事例を示す。2.4節では，利用者側から見た知識処理を活用する際に起こりうる問題点に関して考察を行い，その中で知識処理を普及させるためには，どのような視点が必要なのかを考察する。

3章は「自然言語処理」であり，2章で考察した情報化社会における知識処理という観点から，自然言語処理の問題を分析し，今後の自然言語処理に求められる機能を提言する。3.1節では，自然言語処理の歴史を概観し，自然言語処理が主流がどのように変化してきたかを記述する。3.2節では，利用者から見て現在の主流となっている自然言語処理がどのような問題を抱えているかに言及する。3.3節では，これまでの議論を踏まえて，今後の自然言語処理に求められるものに関して提言を行う。

4章は「深層格推測手法の提案」であり，3章で考察した自然言語処理を実現するために，ユーザの援助を必要とせずに深層格の知識を学習する手法DCAPRを提案する。4.1節でDCAPRの基本的な考え方を説明した後，4.2節でアルゴリズムを記述する。4.3節では，新聞記事などを対象とした実験を行い，DCAPRの有効性を確認する。また，4.4節では，DCAPRを応用した自動要約システムを記述する。

5章は結論である。

2章　情報化社会における知識処理

2.1　情報化社会の現状と今後の動向

今日では，情報機器の低価格化や情報インフラの整備が進み，誰もが情報機器を利用できるような社会になりつつある。この傾向は，今後も拡大が続くと考えられ，将来的には，ユビキタス社会が到来することになると考えられている。ユビキタス（Ubiquitous）とは，ラテン語で「遍在」，「いつでも，どこにでも存在する」という意味であり，ユビキタス社会とは，ユビキタス・コンピューティング（ubiquitous computing）が実現されている社会のことを意味する。

ユビキタス・コンピューティングは，1988年に，ゼロックス・パロアルト研究所（Xerox Palo Alto Research Center，PARC）のマーク・ワイザー（Mark Weiser）が提唱した概念であり，いつでもどこでも，利用者が意識せずとも，情報通信技術を活用できる環境のことを意味している。情報通信機器が現実生活の場の至る所に埋め込まれ，複雑な操作なしにそれらを有機的に活用できる環境をいう。例えば，「買いたい商品を持って店を出ると自動的に代金が引き落とされる」などが挙げられる。

ユビキタス社会の実現は，国家単位で進められているプロジェクトである「e-Japan重点計画」の重要な柱の1つである。2002年6月18日に出された「e-Japan重点計画2002」では，重点政策5分野の一つとして「世界最高水準の高度情報通信ネットワークの形成」が掲げられている。その具体的な施策として，「ブロードバンド時代に向けた研究開発の推進」という項目があり，その文頭には，「すべての機器が端末化する遍在的なネットワークへの進化を目指す」と書かれている。e-Japan重点計画2002によれば，総務省が中心となって，「1つの端末にとらわれず，いつでもどこでも接続できる，十分な伝送容量を備えたネットワーク環境を目指し，メディアハンドオーバー技術(注1)，フォトニックネットワーク基礎技術(注2)，無線セキュリティプラットフォーム技術(注3)等を2005年度までに実現する」と規定された。

e-Japan重点計画2002に先駆けて2001年11月から2002年6月まで，総務省でも「ユビキタスネットワーク技術の将来展望に関す

る調査研究会」を開催し報告書をまとめている。研究会ではユビキタスネットワークを実現させるメリットとして,「新たな産業やビジネス・マーケットの創出」,「障害者・高齢者等の社会参加の促進」,「環境問題への対応」を謳っている。産業創出では日本が得意とするフォトニック(光技術),モバイル,情報家電分野で,2005年に30.3兆円,2010年には84.3兆円の市場を予測している。

このようなユビキタス社会が実現することで,享受できる恩恵として以下のような点があげられる。

例えば,思いついたとき,ちょっと時間があるときに,自分が端末を持っていなくても気軽にネットワークでオンライン・ショッピングができるようになれば,即座に消費者の間にオンライン・ショッピングが広まるであろう。

企業においては,いつでもどこでも簡単に会議ができるようになる。また,オンラインでの受発注などは現在でもマーケットプレイスで行われているが,競りや仕入れの場に,だれでも,どこからでも参加できるようになれば,中小企業にもビジネスチャンスが広がると考えられる。

さらに,ユビキタス社会には新しいコミュニティが育つ可能性がある。常時,いつでもどこでもネットワークに参加できるようになれば,人々は電話代や時間やアクセスの速度を気にせずに情報の交換をすることができる。となれば,そこから新たなビジネスのアイデアが生み出される可能性も高まるであろうと言われている。

また,e-Japanとは別に,ユビキタス・コンピューティング環境の構築を目指している団体としてT-Engineフォーラムがある。T-Engineフォーラムは,TRONプロジェクト(注4)の一つとして発足し,ビジネス基盤としてのハードウェアやミドルウェアを提供している。例えば,無線ICタグの識別コードの標準化やモノの自動認識に関する技術の研究などを行う「ユビキタスIDセンター」も,T-Engineフォーラム内に設置されている。

しかしながら,これらの活動内容は,真のユビキタス社会を実現するためには不十分であると考えられる。

ユビキタス社会実現のための研究開発プロジェクトとしてあげられているのは,「超小型チップネットワーキングプロジェクト」,「何でもマイ端末プロジェクト」,「どこでもネットワークプロジェクト」であり,これらは情報通信インフラに関するプロジェクトである。

T-Engineフォーラムの活動においても,ハードウェアやOSの規格化・標準化,あるいは,ネットワークセキュリティの強化などが中心であり,情報そのものを分析・加工する技術に関する活動は乏しい。

誰もがコンピュータを意識せずに利用できる社会を実現するためには,物や形としてコンピュータを意識させないだけでは不十分であり,利用するアプリケーションにおいて,その処理に不自然さを感じさせないことも重要である。

1章で述べたように,日本信販や野村證券における問題点は,適切な情報を抽出する技術が未発達であったことによる。情報源となるデータがどれだけ容易に入手できたとしても,適切な情報を抽出できなくては有益なものとはならない。むしろ,ユビキタス社会が実現することで情報の取得・発信が容易となり,社会に流れる情報の量が多くなるほど,情報を分析・加工する技術なしでは,適切な情報へのアクセスは困難なものとなる。

適切な情報へのアクセス手段が確立されないまま,膨大な情報源にアクセスできるようになったとしても,それは,誰もがコンピュータを意識せずに利用できるようになったとは言い難い。

そのような観点から，「ユビキタスネットワーク技術の将来展望に関する調査研究会」の報告書や，T-Engineフォーラムの活動内容を見ると，情報通信技術と比較して，必要な情報にアクセスするための技術に関する言及は極めて乏しいと言わざるを得ない。

2.2 知識処理と自然言語処理

自然言語処理は工学の一分野であり，これまで工学的な見地から論じられることが多かった。しかしながら，工学的に優れた技術が必ずしも一般に普及するわけではない。それゆえ，より広範，包括的な観点から論じるために，自然言語処理を包含する観点から論じることとする。

2.1節では，今後の社会における情報処理の重要性がますます高まっていること，その重要性は，通信・伝達手段としての技術よりも，得られる情報を効果的に分析・加工する技術の方に大きいことを主張した。情報を分析・加工するには専門的な知識が必要であり，そうした本来人間がもっている知識をコンピュータに移転して処理を行わせることが知識処理である。

自然言語処理は，知識処理の対象を言語に限定したものとみなす(注5)ことができるため，知識処理における問題点や対策などは自然言語処理にも当てはまることが多い。そこで，人間がもつ知識をコンピュータに移転させることによる問題点を考察するために，知識処理を人間のもつ知識を移転・活用したコンピュータ処理と定義し，知識処理という観点から自然言語処理の問題に言及していくこととする。

今後の知識処理は，現在よりも効果的な分析・加工を行うために，さらに知的な領域に踏み込む，すなわち，人工知能との結びつきを強める必要があると考えられる。しかしながら，コンピュータが知的な領域を扱うことは，単に技術的な問題であるだけではなく，心理的な問題なども考慮に入れる必要がある。そのような問題を解決して，はじめて知識処理が一般に普及することになると考えられる。それゆえ，今後の自然言語処理が現在よりも知的な領域に踏み込むために，次節以降，知識処理の普及における問題点を考察することで自然言語処理の問題を考察していく。

2.3 ビジネスにおける自然言語処理

自然言語処理はビジネスに必要な情報を得るためのツールとして，現在様々な場面で活用されている。

言語という豊かな表現能力を持つ形式で記述されたデータからは，単なる数字や記号で記述されたデータ以上の情報を抽出することが可能である。また，現代日本のように識字率がほぼ100%となっている社会においては，情報を発信する側にとっても，言語で発信することに要求される技能は比較的容易なものであり，多少の訓練で誰でも自由に情報を発信できるという利点がある。

したがって，今後の情報化社会において，自然言語によるオンライン・コミュニケーションの機会はますます増加すると考えられる。これは，電子商取引やオンライン会議など，ビジネスの場面においても同様である。2.1節で述べたような情報インフラの整備が進めば，オンライン・コミュニケーションのためのハードウェア的な環境は実現されることになるであろう。

しかしながら，円滑なコミュニケーションとは，単に多くの情報が流れていてアクセスが可能であるということではなく，その情報を認識した上で，それが何であるかを判断したり解釈したりする過程が達成されているという条件が必要である。

2005年2月2日現在，WWW検索エンジンの代表格であるGoogleでは80億を越す8,058,044,651のWebページが登録されている。この段階で，適切なサイトを特定し必

要な情報を入手することは既に容易ではない。今後さらに膨大な情報がネットワークを流れることを想像すると，自然言語処理の活用なくしては円滑なコミュニケーションが実現できないと考えられる。

言語は人間の知的活動における重要な要素である。これまでは，情報伝達の手段としての観点から，自然言語処理の必要性を説いてきたが，言語には，伝達手段としての言語（外言，outer speech）と，思考の道具としての言語（内言，inner speech）が存在する。

将来，コンピュータが人間と同等の言語処理を実現できるようになった社会を仮定する。そのような社会において，コンピュータは単に情報を伝達，加工する手段ではなく，人間にとってさらに有益な情報となるよう，種々の情報を分析，推論して，新たな情報価値を付加するために不可欠な存在となると考えられる。

例えば，現在ほとんどの企業で速度や労力などの理由から，コンピュータを介した報告が主流となっている。しかしながら，コンピュータを介することで容易に報告が可能となった結果，受けとる報告の数は増加する傾向にある。それに比例して，受けとった報告を自動的に分類，整理する機能の重要性も増加している。しかしながら，現在の自動分類技術では，タグや書式などの自動分類に必要な情報を報告者自身が付与しなくてはならない。しかしながら，コンピュータが報告書の内容を判断し，自動的に分類できるようになれば報告者である人間の労力は大きく軽減される。また，報告を受ける人間にとっても，コンピュータが報告書の内容を理解することができたならば，タグなどによる分類よりも効率のよい整理ができるであろう。例えば，分類する際に，以前上司の出した指示を単純に履行しただけの報告書よりも，新たに上司の指示を仰ぐ必要がある報告書の方が優先されるように分類することができたならば，報告を受ける人間の労力はさらに軽減されるであろう。あるいは，上司の指示を仰ぐような報告書の場合に，上司が判断するために必要な関連情報を提示したり，オブザーバとして参考意見を述べたりすることができるようになると考えられる。

現在の自然言語処理では，文章を人間のように理解する段階に到達するのは，まだ当分先のこととと思われるが，上の例で述べたような機能が実現すれば，今後のビジネスにおける新たな賦活剤となるであろう。

現時点で，比較的，実現可能性が高い応用例のひとつとして，自動要約があげられる。自動要約とは，ひとつもしくは複数の文章の趣旨を要約して表示する方法である。自動要約を活用することで，経営者が膨大な資料に目を通すことなく，大要を把握することが可能となり，複数の資料間の関連性が理解しやすくなると考えられる。

現在の自動要約は，重要語や接続詞などを手がかりとして，重要な段落や文の要点を抜き出す方法，複合名詞や体言止めといった短縮表現に言い換える方法などが主流である。身近な例としては，Microsoft Word の要約にマーカをつけるものがある。しかしながら，精度は30%程度と言われており，結果の信頼性は得られていない。

2.4 知識処理システムの普及に向けて

2.1節では，現在の社会が，経営情報をはじめとする，あらゆる情報をネットワーク上で流通させるような方向に向かっていること，2.3節では，そのような社会において，自然言語処理システムが経営における意思決定に有益であることを示した。本節では，これらに対する反定立として，自然言語処理をはじめとする知識処理システムがユーザに受け入れられない原因を分析し，知識処理システムの普及に向けて考察していくこととする。

ユーザが知識処理システムを利用すること

の障害として，大きく以下の2点が考えられる。1点目は，変換ミスや翻訳ミスなど，コンピュータの誤りに対してユーザが過剰な拒否反応を起こす可能性があること，2点目は，知識処理システムを導入した場合に要求されるコストの大きさである。

最初の障害としてあげた過剰反応の例として，山本ゆうじ（2005）の提唱したルーウェリン反応（Llewelyn reaction）があげられる。ルーウェリン反応とは，コンピュータの認識処理，特に自然言語処理の認識の失敗に対する，人間側の，感情的な拒否反応心理のことである。理性的な有用性の判断よりも，不信感，嫌悪感が先立ち，「単に処理に失敗した」以上の拒否反応を示す可能性が知られている。ルーウェリン反応は，産業革命期のイギリスに発生した機械破壊運動，いわゆるラッダイト運動（Luddite movement）の心理に通じるものがあるといわれており(注6)，デジタル・ディバイドの一因として自然言語処理を含む高度な知識処理技術の推進・普及の妨げとなっている。

このような過剰反応の原因を更に分類すると，コンピュータが「知」を扱うことへの抵抗感によるものと，コンピュータに対する過剰な期待が裏切られたことによる失望感によるものとに分離することができるであろう。前者の抵抗感は，ルーウェリン反応やラッダイト運動の心理に見られるように，機械が人間の領域を侵犯することが原因であると考えられているが，この問題は非常に深い問題であり，ユーザへの啓蒙活動をはじめとして，様々な面から解決の手段を講じていく必要があると考えられる。後者の失望感は，コンピュータの誤りが人間では考えられないような誤りであることへの苛立ちなどが原因であると考えられる。この問題も決して浅い問題ではないが，処理技術の見直しによって解決できる部分が大きいと考えられる。

第二の障害であるコストの問題に関しては，知識処理システムを導入する時点で支払うコストと，利用している間，支払い続ける導入後のコストに分類することができる。

特に経営の第一線においては，鮮度が高い情報が求められるため，知識処理に用いられる知識も新鮮な情報に対応できるようにする必要がある。また，処理に必要とされる，システムがもつべき知識は利用者ごとに異なる場合が多く，有効に活用するためには利用者ごとにシステムの知識をカスタマイズする必要がある。このようなことから，知識処理システムを利用するために，導入後に支払うコストは，ルーチンワークを代行する類のシステムと比較して大きいと考えられる。

さらに，導入時のコストが，事前に明示的に計算することが可能であるのに対して，導入後のコストは，利用期間中常に支払わなくてはならないこと，コストの大きさもその時点にならないと不明であることなどから，導入時のコストと比較して予測することが困難である。このような問題を解決しないと，経営者が導入に踏み切るのは困難であろう。

また，支払うコストの種類の面からも，金銭的なコストだけではなく，導入したシステムを使いこなすための技能的なコストを考慮する必要がある。上で述べたように，有効に活用するためには知識を利用者ごとにカスタマイズする必要がある。利用者ごとのカスタマイズを行うためには，何らかの形で利用者側の作業が必要であるが，この作業に要求される技能があまりに煩雑なものであった場合にも，システムの導入は見送られることになるであろう。

したがって，利用者が求める理想的なシステムとは，利用者がそのシステムを使用しているうちに，システムの方で利用者に必要な知識を推測し，自動的に知識のカスタマイズを行うような形式でありと考えられる。

以上の問題を解決し，理想を実現するための手段として，コンピュータに学習を行わせ

ること，すなわち，機械学習の利用が考えられる。機械学習は，人工知能における研究課題の一つで，人間が自然に行っている学習能力と同様の機能をコンピュータで実現させるための技術・手法のことである。ある程度の数のサンプルデータ集合を対象に解析を行い，そのデータから有用な規則，ルール，知識表現，判断基準などを抽出する。また，データ集合を解析するため，統計学との関連も非常に深い。

機械学習を組み込んだシステムであれば，知識処理システムを保守するために必要なコストの内，最も不明瞭で大きいと考えられる知識整備に必要なコストを大幅に抑えることができる。したがって，残るコストは，明確に計算できる導入時点のコスト，および，他のコンピュータシステムの保守と同程度と見積もることが可能である，システム自体の保守に必要なコストとなり，利用者にとって導入しやすくなると考えられる。

以上から，知識処理システムの普及のためには，ユーザに拒否反応を抱かせないような知識処理技術を開発することと，知識の学習において人手による労力を軽減することが必要であると考えられる。このことは自然言語処理システムの普及においても同様であると考えられる。

自然言語処理システムの普及におけるコスト軽減の影響を調査するために，円滑なコミュニケーションの実現を，繰り返し囚人のジレンマ・ゲームにおける協調関係の成立とみなしたコンピュータ・シミュレーションを行った（渋木，2005)。協調と対立の問題を繰り返し囚人のジレンマ・ゲームとしてモデル化し，進化的アプローチと呼ばれるコンピュータ・シミュレーションにより解決の糸口を探る方法はAxelrod（1984）が提案したものであり，その後も多くの研究者によって取り組まれている（Axelrod 1997；Gilbert and Troitzsch 1999；高増・服部1999a；高増・服部1999b；田村2001；牛丸2004)。

渋木（2005）のシミュレーションにおいて，エージェントの意思が行動に反映されない外的要因をノイズと定義し，協調関係を目指すエージェント同士において必ずしも協調関係が成立するとは限らない状況をつくりだした。このような状況において，自然言語処理システムをノイズの発生率を抑える手段とみなし，自然言語処理システムを導入することが協調関係の成立ひいては導入したエージェントの利益にどのような影響を与えるかを調査している。結果として，多くの場合においてノイズ抑制の戦略をとったエージェントの利益が増加する傾向にあることを示した。このことからも，自然言語処理システムを普及させるための問題を解決していかなくてはならないと考えられる。

3章　自然言語処理

3.1　自然言語処理の歴史

今日において，コンピュータ上で文字情報はもちろん，画像，音楽，動画などの情報を扱えることは当然のことであるが，コンピュータが作り始められた1940年代には事情はまったく異なっていた(注1)。コンピュータは，電子計算機の名が示す通り，数の計算を高速に行う機械であった。しかしながら，1947年にウォレン・ウィーバー（Warren Weaver）とアンドリュー・ドナルド・ブース（Andrew Donald Booth）による英語とフランス語における翻訳システムを提案(注2)したことにより，文をコンピュータで扱えることが明らかになり，自然言語処理の研究が始まった。

この時代の機械翻訳の方式は一般に直接翻訳方式と呼ばれている。翻訳すべき言語対に対応して，単語レベルで一つの言語から他の言語へ置き換えを行うことが中心であった。

その後，文中の各単語の文法的役割を明らかにして翻訳する方式，すなわち，もとの文において，どの単語が主語でどの単語が目的語なのかといった各語の文法的役割を明らかにした後，これを翻訳先の言語の構造に移して，その構文構造から文を組み立てるという方法で翻訳を行うようになった。この方式の翻訳を構文翻訳といい，文の構造を明らかにすることを構文解析という。

文の構造という考え方は，1957年にノーム・チョムスキー（Noam Chomsky）によって書かれた，Syntactic Structureによって，より一般的な概念として広まった（Chomsky，1957）。

しかしながら，構文解析の研究が本格化した結果，明らかになったことは，一つの文が構文的立場からは多くの解釈を持ちうるという発見であった。この構文的な解釈のあいまいさの例としては，

He saw a woman with a telescope.

などがあげられる。この文は，前置詞句「with a telescope」が修飾する対象が動詞「saw」なのか名詞「a woman」なのか明確ではない。もしも，「saw」を修飾しているのであれば，この文は「彼は望遠鏡で女性を見た」という「手段」としての意味になる。一方，「a woman」を修飾しているとすれば，「彼は望遠鏡をもっている女性を見た」という「所有」の意味と解釈されることになる。このような曖昧性を解消するためには，「手段」なのか「所有」なのかという意味を考慮する必要がある。

この問題に対して，1968年，チャールズ・フィルモア（Charles Fillmore）によって提唱されたのが格文法（Case Grammar）の考え方である（Fillmore，1968）。

格文法では，文は動詞を中心として組み立てられているとし，文中の各主要名詞（句）が述語動詞に対してどのような格に立つものであるかに注目する。そして，その格関係を意味的に（深層的に）とらえることに重点をおいている。このような意味的な格を深層格という。

この考え方は魅力的なものであったが，コンピュータに深層格を判断させることは非常に難しかった。格文法の考え方をコンピュータ上で実現するためには，一つ一つの個別動詞について，その動詞がどのような意味の名詞をどのような深層格にとりうるかということを詳細に辞書記述する他はないということが分かり，語の意味を類型化した数十～数百の「意味素」を設定し，それぞれの名詞がどのような意味素を持ちうるかを記述する作業を行わなくてはならなかった。

このような莫大な時間と労力をかけて意味を人手で記述することにより，一定の成果を得ることが可能となったが，記述すべき事項が多くなると，記述漏れによる網羅性の問題や，記述間の整合性がとれなくなるなどの問題が発生するようになった。また，一般に使用される文を実用的な精度で網羅的に解析するために必要な記述事項の数はあまりに膨大すぎて，人手で記述することは事実上不可能であった。すなわち，人手による記述というアプローチによる文法・意味理論の限界であった。

また，文法や意味理論を抽象的な規則できれいに記述しようという試みにも限界が見え始めた。なぜならば，言語は，自然科学が対象とする外界世界とは異なり，人間の頭脳のもつ能力であり，きれいな法則性をもっているとは言い難く，必然的に言語の理論は言語の第一次近似を与えるものという位置づけにならざるを得ないからである。したがって，近似的なレベルで文はどういう構造をしていると言うことはできても厳密にあらゆる文に対して適用できる規則の集合は作成できないという考え方が主流となってきた。そこで，種々の特有の表現を例文（例句）として集め，それをあたかも規則であるかのように考えて

処理する方法が考えられた。この方法を，コーパス・統計モデルといい，現在の主流の考え方である。コーパスとは，例文を集めたデータベースのことである。

コーパス・統計モデルのさきがけは，機械翻訳における用例ベースである。用例ベースでは，中学生が英文を訳す時のように例文中の語句を置き換えて翻訳を行う。例えば，「私は学生です」を翻訳する場合を考える。「私は少年です」の訳文が「I am a boy.」であると知っており，「少年」と「学生」の訳語がそれぞれ「boy」と「student」であると分かっているならば，「I am a boy.」の「boy」を「student」に置き換えた「I am a student.」を訳文として求めることができる。これが文法や意味を考慮せずに翻訳する用例ベースの基本的な考え方である。

現在のコーパス・統計モデルでは，そのまま例文を加工して目的となる文を作り出す方法以外にも，コーパス中に出現する単語の頻度などを計算して確率的に尤もらしい選択を行う方法などが存在している。

このようなコーパス・統計モデルでは，かつての文法・意味理論で問題となった整合性の問題をあまり考慮することなく，網羅性を高めるために，ひたすら多くの用例を収集すればよいという利点がある。

しかしながら，文法・意味理論が行き詰った反動からか，構文や意味の領域に踏み込んだ知識を用いず，単語レベルでの解決を目指す方法が現在の大勢を占めている。

これは，「人手による記述」から「コーパスを用いた統計処理」へと知識付与のアプローチが変化しただけではなく，対象とする知識も「文法・意味理論」から「単語などの表層情報」へと変化したということができる。このことが，3.2節で述べる種々の問題を生み出している。

自然言語処理の大まかな歴史を示したのが図3.1である。直接翻訳方式に代表される逐語処理から始まった自然言語処理は，構文翻訳などに代表される文法，意味理論へと移り，処理対象となる知識をそれまでの単語レベルから意味のレベルへと発展させてきた。しかしながら，人手により知識を記述するというアプローチの限界により，コーパスから自動的に知識を学習するというコーパス・統計モデルへと移ることとなった。しかしながら，コーパスからの学習というアプローチに変化した際に，学習対象となる知識をかつての単語レベルに戻すこととなったまま，現在に至っている。

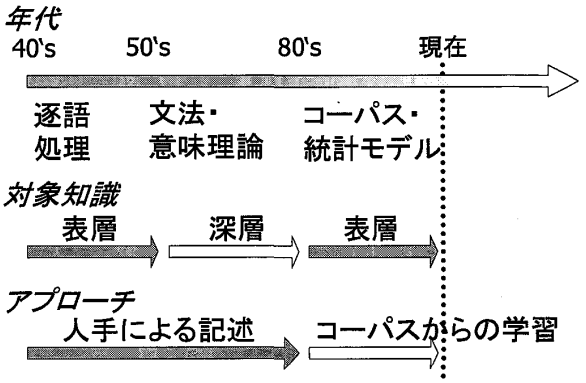

図3.1　自然言語処理の歴史

3.2　利用者から見た自然言語処理の問題

2.2節で述べたように，経営において求められる自然言語処理は，最終的には意思決定における補佐としての役割であると考えられる。そのためには，出力結果に対して，コンピュータは根拠を示すことが可能でなくてはならない。

コーパス・統計モデルによって提示される根拠は，「過去にあった○○という事例と△△%類似しているため，その時と同じ解決方法を選択した」といった類の説明となる。このような根拠の問題として，以下の3点が考えられる。

1点目は，過去にあった事例を根拠としてしか判断できないことである。もちろん，過去の事例を踏まえて今後の方針を決定するこ

とは非常に重要かつ有益なことであるが，判断の根拠が過去の事例にしかないことは問題である。過去に先例のない事案に関して参考意見を求めたとしても，コンピュータは何も答えることができないであろう。このような問題に答えるためには，コンピュータの中に何らかの理論が構築されていなくてはならない。

2点目は，判断の基準が総合的な尺度でしか測れないことである。人間の場合，総合的に見て全体の何割が類似しているかということよりも，何か一つの要因を判断基準として決定を下すことがある。理論を持たずに，コーパス中の全要素を等価値に判断するコーパス・統計モデルでは，特定の要因に着目するということが困難である。

3点目は，対策が過去の事例と同一になることである。その当時において画期的な方法であっても，現在において通用するとは限らない。むしろ，画期的であればあるほど周知の事実として広まっている可能性が高く，その事実を踏まえた新しい方法を考え出さなくてはならないであろう。1点目の問題と関連するが，意思決定においては，新規な事柄について判断したり回答したりする必要がある。このような問題を解決するためには，具体的な事例のみでは不十分であり，抽象化された理論に基づいた推論が必要であると考えられる。

3.3 今後の自然言語処理に求められるもの

3.1節で述べたように，現在の自然言語処理の主流はコーパス・統計モデルである。コーパス・統計モデルでは，単語レベルでの処理を中心に行うことで，従来の文法・意味理論に踏み込むことなく，一見，尤もらしい結果を求めることが可能である。

しかしながら，このことは自然言語処理に「文法・意味理論」が不要であることを意味しているわけではない。例えば，3.1節の用例ベースの例において，「私は宇宙飛行士です」を翻訳する場合を考える。「I am a boy.」の少年「boy」を単純に宇宙飛行士「astronaut」に置き換えただけでは，「I am a astronaut.」という訳文となってしまい冠詞の誤りが起きてしまう。この問題を解決するには，母音の前の不定冠詞は「an」であるという規則が必要である。

上の例は簡単な規則で修正することが可能であるが，基本的に，文法や意味を考慮しない単語レベルでの確率や統計による処理では，人間では考えられないような誤りとなることがある。このような誤りは，人間の言語処理と異なる処理過程を経て導き出されたものであるため，何故そのような誤りとなったのか，人間にとって理解が困難である。また，誤りを訂正しようにも，特定の規則から決定的に求められたものでないため，抜本的な解決が困難である。例えば，「I am an astronaut.」という正しい訳文を教えたとしても，「私は正直者です」を「I am an honest.」と正しく翻訳できるか保証はない。

このようなことから，コーパス・統計モデルによる誤りは，文法・意味理論による誤りよりも，利用者に不信感を与えやすく，2.3節で述べた過剰な拒否反応を引き起こす可能性が高いと考えられる。この問題を解決するためには，自然言語処理過程に，人間の言語処理過程との親和性をもたせる必要がある。

したがって，以上の問題を解決するためには，文法や意味などの深層的な情報に再び踏み込む必要があると考えられる。3.1節で述べたように，「文法・意味理論」から「コーパス・統計モデル」への変化は，対象知識とアプローチの両方の変化を引き起こした。これは，それまでの人手による知識の付与というアプローチの限界から止むを得なかったことであるが，図3.2に示すように，今後の自然言語処理には，知識の付与を「コーパス・統計モデル」と同じ機械学習のアプローチを

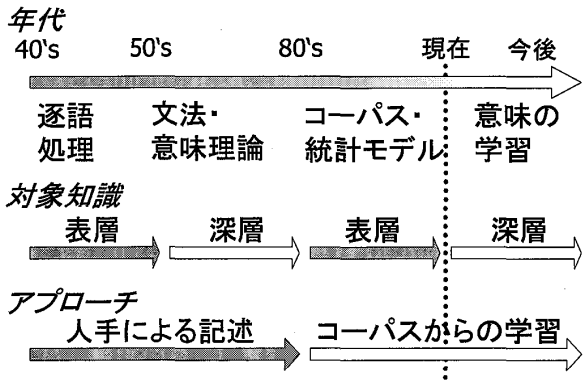

図3.2 今後の自然言語処理

維持しつつ，対象とする知識を「文法・意味理論」へと拡張していくことが求められると考えられる。

4章 深層格推測手法の提案

4.1 基本的な考え方

4.1.1 従来の深層格推測手法の問題

機械翻訳における曖昧性の解消などのために，文の骨子を表現する手段の一つとして，深層格を利用することが考えられる。深層格を推測するための知識を人手で与えることは，網羅性やコストの点で問題が残るため，コーパスから学習する手法が望まれている。このような背景から，大石らの手法（1995）をはじめとして，元木（2000），Blaheta (2000)，Gildea（2000），Gildea（2002），原田（2002），小山（2003）などに挙げられる手法が提案されている。これらの手法では，全ての学習データに深層格のタグが付与された文を用いており，タグが付与されていないデータから学習する点に関して考慮されていない。人手によるタグの付与にはコストがかかるため，コスト軽減という観点からは，タグなしの文から学習できる機構を備えた手法が望ましい。そこで著者は，一定量のタグ付きコーパスから得られた知識を中核として，タグなしコーパスから深層格推測の知識を学習する手法を提案する。これにより，多大なコストをかけずに網羅的に知識の拡充を図ることを試みる。

著者も従来手法の多くと同じく，深層格を推測するためには，最終的に助詞や語順などの表層表現と対応付けられた知識が必要であると考えている。しかしながら，係助詞のように深層格を決定できない表層表現や，「水が飲みたい」のように格助詞であっても共起する動詞により一般に解釈される深層格と異なる場合など，深層格と表層表現との対応関係は複雑であり，人手で網羅的に与えることは困難であると考えられる。この問題は，タグ付きデータを作成，利用する場合にも生じる問題である。著者は「水が飲みたい」と「水を飲みたい」が同じ深層格と判断されるのは，「水」と「飲む」という名詞と動詞の概念によるものであると考え，この傾向は言語の種類を問わずに普遍であると仮定した。このような考えから，提案手法では，名詞や動詞の概念ごとに，それぞれ特定の深層格に解釈される傾向が存在するという仮定に基づき，タグ付きコーパスから深層格の傾向を計算した後，その傾向を用いてタグなしコーパスから多様な言語表現に対処するための知識を学習する。この深層格の傾向を深層格選好と呼び，深層格選好に基づいて深層格を推測する本手法をDCAPR（Deep Case Analysis based on deep case Preference and Regularization）と呼ぶ。単語概念の深層格選好を手がかりとすることで，表層表現の違いを事前に考慮することなく学習することを可能としている。本章では，言語における表層表現の差異を取り上げ，日本語と英語を対象として実験を行う。

4.1.2 DCAPRの概要

本手法は，人手による労力の軽減という目的から，機械的に処理できる情報から深層格を推測することを試みている。それゆえ，深層格は，格支配される名詞，格支配する動詞，

格支配される際の表層表現によって推測されると仮定した。本論文では，名詞と動詞を手がかりとした知識を深層格選好，表層表現を手がかりとした知識を深層格推測規則と定義する。

まず，名詞と動詞の深層格選好について説明する。例として，「太郎」と「東京」の二つの名詞が，ある文の中で使われた場合に，これらの名詞が担うであろう深層格を考える。文脈によって，どちらの単語も様々な深層格を担うことが可能であるが，一般的にagentを担う傾向が強いのは「東京」よりも「太郎」の方である。同様に，placeを担う傾向が強いのは「東京」の方と考えられる。また，これらの傾向は，動詞との関係を考慮することにより，さらに顕著なものとなる。例えば，「住む」という動詞が二つの名詞をとることを考えた場合，agentとplaceを担う二つの名詞を同時にとる傾向の方が，objectとsourceを同時にとる傾向よりも強いと考えられる。したがって，これらの傾向を用いることで，「東京，太郎，住む」のような助詞が欠落した文に対して，「東京」がplace，「太郎」がagentと判断し，「東京に太郎が住む」と復元することが可能である。本論文では，この傾向を数値化したものを深層格選好と定義し，4.1.3に述べる。

次に，深層格推測規則について説明する。深層格選好を手がかりとした推測は，一般的な傾向を反映したものであるため，多くの事例を正しく推測できる一方で，例外的な事例には対処できない場合があると考えられる。例えば，「次郎を公園で見る」という文の推測を行う場合，「次郎」，「公園」，「見る」の深層格選好に基づいて推測するため，「次郎が公園で見る」と同じように推測される可能性が高い。この問題を解決するために，機能語や語順といった，深層格を推測する手がかりとなる表層的な表現を考慮する必要がある。しかしながら，4.1.1で述べたように，表層表現と直接対応付ける知識を人手で網羅的に与えることには問題があるため，タグなしコーパスから自動的に学習する仕組みが必要である。表層表現がどのように深層格と対応しているかを学習するために，本手法では，深層格選好による推測結果の多くは正しく，推測結果に共通した表層表現は深層格推測の手がかりとなるという仮定に基づいて学習を行う。例えば，「花子が服を買う」と「部屋で本を読む」という二つの文が深層格選好に基づいて以下のように推測されたと仮定する。一番目の文が「花子」，「服」，「買う」に基づいて，二番目の文が「部屋」，「本」，「読む」に基づいてそれぞれ推測され，「服」と「本」の推測結果がobjectと正しく推測されたとする。このとき，両方の文においてobjectと推測された単語に共通した表層表現として「を」が得られることとなる。したがって，objectを推測する表層表現として「を」を学習することにより，「次郎を公園で見る」における「次郎」の深層格推測に役立たせることができる。本論文では，表層表現と深層格を対応付ける知識を深層格推測規則と定義し，4.1.5に述べる。

本手法の全体の流れを図4.1に示す。学習データとテストデータは全て構文解析済みの単文である。また，図中の番号は処理の順序を示しており，以下，番号に従って記述する。最初に，(1)全てのデータで使用される名詞と動詞のクラスタリングを，4.1.4に述べる表層情報に基づいて行う。これは，タグ付き学習データに含まれない未知語の深層格選好を，作成したクラスタから類推するための処理である。表層表現に基づくのは，人手による労力を軽減するため，機械的に処理できる情報に限定したためである。クラスタリングの詳細は，4.2.1に述べる。次に，(2)作成した単語クラスタを参照し，単語及び単語クラスタの深層格選好を，タグ付き学習データに付与された深層格タグに基づいて学習する。深層

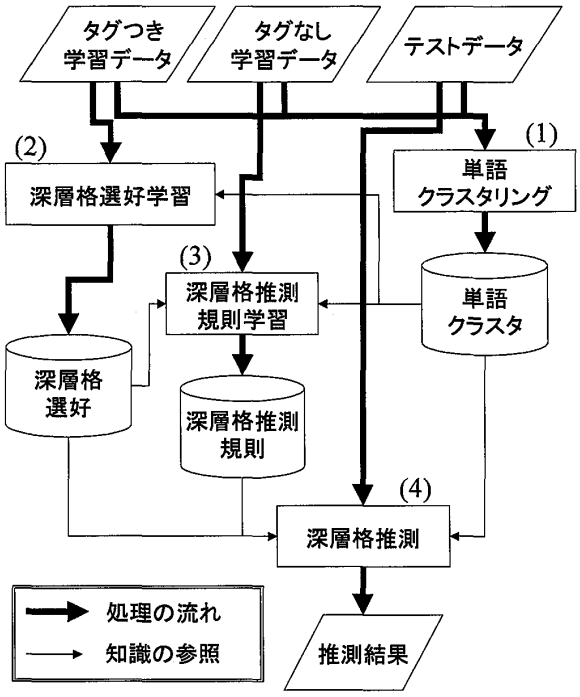

図 4.1　全体の流れ

格選好は，ある深層格と解釈される傾向を確率として表現したものである。その後，(3)学習された深層格選好を基に，タグなし学習データの深層格を推測し，推測された深層格とタグなし学習データ中の表層表現とを対応付けることで深層格推測規則を学習する。本来であれば，深層格選好の学習に用いられるタグ付きデータとは別に，規則を学習するためのタグなしデータを用意することが望ましいが，本論文では利用可能であったデータ量の関係から，深層格選好学習に用いたデータからタグを除いたものを規則学習のデータとした。深層格選好学習と深層格推測規則学習の詳細は，4.2.3 と 4.2.4 にそれぞれ述べる。最後に，(4)作成されたクラスタ及び学習された深層格選好と規則に基づいて，テストデータの深層格を推測する。深層格推測の処理は 4.2.5 に述べる。

4.1.3　深層格選好

名詞と動詞の深層格選好の例を表 4.1 と表 4.2 にそれぞれ示す。名詞の深層格選好は，

表 4.1　名詞の深層格選好の例

NOUN	agent	object	goal	condition	…	quantity
太郎	0.8	0.0	0.2	0.0	…	0.0
Tokyo	0.0	0.0	0.4	0.0	…	0.0
(ClusterN1)	0.0	0.5	0.4	0.0	…	0.1

表 4.2　動詞の深層格選好の例

VERB	agent, agent	agent, object	agent, goal	agent, condition	…	quantity, quantity
行く	0.0	0.0	0.6	0.1	…	0.0
live	0.0	0.0	0.0	0.0	…	0.0
(ClusterV1)	0.0	0.5	0.4	0.0	…	0.0

ある動詞に格支配される際にその名詞が担う深層格の傾向を表し，動詞の深層格選好は，ある名詞を格支配する際にその名詞が担う深層格の傾向を表す。深層格選好は，それぞれの深層格ごとに，その深層格と解釈される確率で表現される。また，学習データにおける単語のスパースネス問題に対処するため，「太郎」や〝Tokyo〟のような個々の単語ごとに深層格選好を学習するのに加えて，4.2.1 に述べるクラスタリングを行い，クラスタごとにも深層格選好の学習を行う。本論文では，日英における比較のために EDR コーパスを用いるため，対象とする深層格を EDR の概念関係子に基づくこととした。(大石，1995) と (小山，2003) においても，EDR の概念関係子に基づいていたため，それぞれが対象とした深層格を併せ，表 4.3 に示す 14 種類とした。

動詞は複数の名詞を格支配するため，動詞の深層格選好は表 4.2 に示すように，一つの深層格ではなく，同時に格支配する名詞群が担う深層格の組として表現する。深層格の組で表現する理由は，深層格の推測において，以下のような効果を期待するからである。便宜的に「行く」の深層格選好が，「太郎が [agent] 東京に [goal] 行く」と「東京から [source] 大阪に [goal] 行く」のように，

表4.3　深層格の一覧

深層格	意　味	例
agent	有意志動作を引き起こす主体	父が食べる
object	動作・変化の影響を受ける対象	りんごを食べる
goal	事象の主体または対象の最後の位置	東京に行く
condition	事象・事実の条件関係	雨が降ったので家に帰った
implement	有意志動作における道具・手段	ナイフで切る
material	材料または構成要素	牛乳からバターを作る
place	事象の成立する場所	部屋で遊ぶ
scene	事象の成立する場面	ドラマで演じる
source	事象の主体または対象の最初の位置	京都から来る
cause	原因	病気で倒れる
purpose	目的	映画を見に行く
basis	比較の基準	バラはチューリップより美しい
beneficiary	利益・不利益の移動先 [受益［者］と被害［者］の両方を含む]	父に買ってあげる
quantity	物・動作・変化の量	3kg痩せる

[agent, goal] と [source, goal] の組しかとらないとし，「札幌から小樽に行く」の推測を行うとする。このとき，「札幌」の深層格がsourceだと推測できた場合，深層格単位で表現されていたとするならば，「小樽」の深層格としてagentとgoalが考えられるのに対して，深層格の組による表現ならば，sourceと組であるgoalと一意に決定することができる。この深層格の組を深層格パターンと定義する。本論文では，二つの名詞と一つの動詞からなる文を対象としたため，動詞の深層格は表4.2に示すように二つの深層格の組に対して表現される。

4.1.4　表層情報

本手法では，ある名詞の深層格を推測するための表層的な手がかりとして，その名詞に接続する機能語，その名詞の文中での位置，文全体で使用されている全ての機能語の並び，という三つの情報に着目する。第一の名詞に接続する機能語は，日本語においては助詞，英語においては前置詞とした。機能語が存在しない場合には空の文字列を代用とする。第二の文中での位置は，動詞を基準とした文節または名詞句単位での相対的な位置で表現する。例えば，「太郎が本を読む」における「太郎」の位置は，「読む」の二つ前に位置するので〝Ｆ２〟という略号で表記する。また，〝Taro reads the book〟における〝book〟は，〝read〟の一つ後ろに位置するので〝Ｂ１〟となる。位置を考慮することにより，「8時に札幌に行く」のような同じ機能語が連続する文の機能語を識別することが可能となる。第三の全ての機能語の並びとは，「太郎が本を読む」を例にとると［［Ｆ２：が］［Ｆ１：を］］のように，それぞれの名詞における機能語と位置の組を並べたものである。本論文では，これを機能語パターンと定義する。機能語パターンでは，位置を考慮するため，「太郎が本を読む」と「本を太郎が読む」では別の機能語パターン［［Ｆ２：が］［Ｆ１：を］］，［［Ｆ２：を］［Ｆ１：が］］となる。

表4.3に示した深層格が助詞などの表層情報とどの程度一意に対応しているかを調査す

るために，EDR 日本語コーパスを用いて調査を行った。助詞が省略されておらず，コーパス中で3回以上出現する助詞をもつ5,508の係り受け関係を抽出し調査対象とした。また，抽出された係り受け関係に含まれた助詞は52種類であった。その結果，ある助詞が同一の深層格に解釈される割合は平均76.6%であったが，位置と機能語パターンを考慮することで平均83.8%まで向上させることが確認できた。また，ある助詞が何種類の深層格と解釈されうるかの調査では，平均して4.3種類から1.9種類へ絞り込むことが可能となった。

4.1.5 深層格推測規則

4.1.4で述べた表層情報と深層格を対応付けた知識を深層格推測規則とする。4.1.4で述べたように三つの表層情報を考慮することで，深層格の推測候補を絞り込むことが可能となるが，一対一まで絞り込めるわけではない。したがって，深層格選好と同様に，規則を適用した場合に解釈される深層格の傾向を確率として表現することとした。規則の例を表4.4に示す。三つの［ ］で括られた規則の適用条件のうち，最初の二つは機能語パターンを示し，三番目の［ ］は規則を適用する名詞の位置を表す。表4.4の最初の規則は，例えば，「太郎が本を読む」において「本」の深層格を推測するために利用され，次の規則は〝I live in Tokyo.〟において〝I〟の深層格を推測するために利用される。

表4.4 深層格推測規則の例

RULE	agent	object	goal	condition	…	quantity
[が：F2] [を：F1] [2]	0.0	0.7	0.0	0.1	…	0.0
[φ：F1] [in：B1] [1]	0.9	0.0	0.0	0.0	…	0.0

4.2 アルゴリズム

4.2.1 単語クラスタリング

本手法の単語クラスタは，学習データに含まれない未知語の深層格選好を類推するために利用される。したがって，深層格選好が類似した単語群を同じクラスタにまとめる手法が必要である。これまで，Caraballo (1999) など，表層的な情報を手がかりに，意味的に類似した単語のクラスタリングを行う手法が提案されている。4.1.4で述べたように，本論文の表層情報は深層格と比較的対応関係がとれているため，表層情報を利用して深層格推測に特化したクラスタを作成することができると考えられる。このような背景から，著者は，4.1.4で述べた表層情報に基づいて単語をクラスタリングする手法を提案しており，日本語を対象とした深層格推測実験では，分類語彙表（1964）を用いた結果よりも自動的に作成されたクラスタを用いた結果の方が高い精度であったことを確認している（渋木，2004)。クラスタリングされる単語の定性的傾向は，名詞，動詞ともに，類似した文における類似した構成要素として用いられる単語群がクラスタリングされることになる。したがって，対象となる名詞に付属する表層情報の頻度分布が類似の傾向にあるだけではなく，共起する名詞に付属する表層情報の頻度分布も類似の傾向にあることになる。

本クラスタリングは，図4.2に示す流れで行われる。最初に，名詞と動詞それぞれの単語ベクトルを作成する。単語ベクトルの要素は，コーパス中でその単語と共に出現した表層情報の頻度である。表層情報は，機能語，位置，機能語パターンの三つであるが，動詞の場合は，接続する助詞や前置詞がないため機能語の情報が存在せず，位置情報も動詞を基準とした相対的な位置であるため有益な情報とならない。それゆえ，動詞ベクトルの要素数は機能語パターンにのみ依存するとし，名詞ベクトルの要素数は三つの表層情報の組

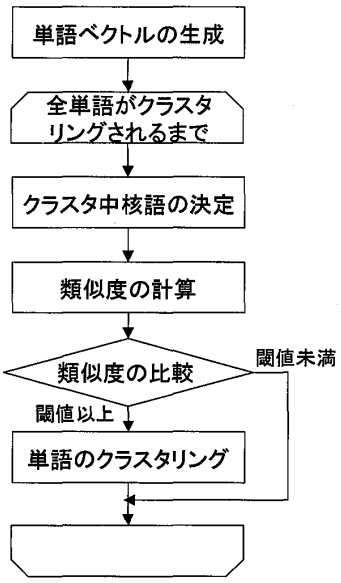

図4.2　単語クラスタリングの流れ

み合わせに依存するとした。本論文では，コーパスにおける頻出機能語上位30種類を用いてベクトルを作成した。したがって，機能語パターンは二つの機能語の組み合わせであるため，動詞ベクトルの要素数は，30×30＝900，二つの名詞を格支配する文を対象としているため，名詞ベクトルは，機能語パターンに名詞の位置情報2通りを加えて，30×30×2＝1,800となる。

次に，コーパス中での最頻出単語をクラスタの中核語として，中核語との類似度が閾値T_1以上となる単語を同一のクラスタにまとめる。単語間の類似度は，Caraballo (1999) と同様に，二つの単語ベクトルの余弦で定義する。

$$sim(\vec{v},\vec{w})=\frac{\vec{v}\cdot\vec{w}}{|\vec{v}||\vec{w}|} \quad (1)$$

閾値以上の単語を全てクラスタリングした後，残りの単語の中で最も出現頻度が高い単語を次のクラスタの中核語とし，残りの単語の中で閾値以上の類似度をもつ単語をクラスタリングする。以上の処理を，全ての単語がクラスタリングされるまで繰り返す。クラスタリングの閾値T_1は，渋木（2004）の結果から0.8とした。

4.2.2　尤度(注1)

4.1.2に述べたように，本手法の深層格dは，格支配される名詞n，格支配する動詞v，格支配する際の表層表現gによって推測される。それぞれの要素から深層格dとなる確率P(d,n,v,g)を式(2)のように近似する。

$$P(d,n,v,g)=P(d|n)\times\sum_{i\in D}P(d|v,i)\times P(d|g) \quad (2)$$

Dは，全ての深層格の集合であり，動詞における条件付き確率P(d|v,i)は，深層格パターンを考慮するため，共起する深層格iも条件に含まれている。

しかしながら，深層格推測において，ある名詞句は必ず何らかの深層格を1つだけ担うことが前提であるため，P(d,n,v,g)の値によって，深層格を一意に決定することができないという問題がある。例えば，agentとobjectに対するP(d,n,v,g)が共に0.9であるような場合には，どれほど高い確率値であったとしても一意に決定することはできず，逆にagentに対するP(d,n,v,g)が0.1しかなくとも，objectを含む他の深層格に対するP(d,n,v,g)が0.01のように相対的に低い値であった場合にはagentであると推測することが妥当であると考えられる。このような考えから，P(d,n,v,g)をベクトル要素として，単位ベクトル化することにより，P(d,n,v,g)間の相対的な大小関係を表す尤度pl(n,v,g,d)に変換することとした。尤度pl(n,v,g,d)は，以下の式(3)により計算される。

$$pl(n,v,g,d)=\frac{P(d,n,v,g)}{\sqrt{\sum_{i\in D}P(i,n,v,g)^2}} \quad (3)$$

4.2.3　深層格選好学習

4.1.2で仮定した深層格を推測する手がかりのうち，名詞と動詞の深層格選好を，タグ付きコーパスから以下の手順で学習する。ある名詞nにタグ付きコーパス中で付与され

ている深層格dの頻度を$fq_N(n,d)$とする。以下の式(4)に従って条件付き確率となるよう深層格選好P(d|n)を計算する。

$$P(d|n)=\frac{fq_N(n,d)}{\sum_{i\in D} fq_N(n,i)} \quad (4)$$

Dは表4.3に示す深層格の全集合である。

この値は単語ベクトルの各要素の頻度を正規化したものである。

動詞vの場合も同様に，二つの深層格d, d_2の頻度を$fq_V(v,d,d_2)$とし，以下の式(5)に従って深層格選好$P(d|v,d_2)$を計算する。

$$P(d|v,d_2)=\frac{fq_V(V,d,d_2)}{\sum_{i,j\in D} fq_V(v,i,j)} \quad (5)$$

また，4.2.1で作成した名詞クラスタncおよび動詞クラスタvcの深層格選好P(d|nc)及びP(d|vc,d_2)を，以下の式(6)及び式(7)に基づいて$fq_{NC}(nc,d)$及び$fq_{VC}(vc,d,d_2)$を求めた後，同様にして計算する。

$$fq_{NC}(nc,d)=\sum_{i\in nc} fq_N(i,d)$$

$$P(d|nc)=\frac{fq_{NC}(n,d)}{\sum_{i\in D} fq_{NC}(nc,i)} \quad (6)$$

$$fq_{VC}(vc,d,d_2)=\sum_{i\in VC} fq_V(i,d,d_2)$$

$$P(d|vc,d_2)=\frac{fq_{VC}(vc,d,d_2)}{\sum_{i,j\in D} fq_{VC}(vc,i,j)} \quad (7)$$

4.2.4 深層格推測規則学習

タグなしデータから，深層格と対応付けられる表層情報を学習するため，まず，深層格選好を手がかりとして，タグなしコーパスの深層格を推測する。この推測では，表層情報を用いることができないため，格支配される名詞と格支配する動詞の深層格選好のみを用いることとした。したがって，動詞vに格支配されている名詞nが深層格dとなる確率P(d,n,v)を式(8)のように近似する。

$$P(d,n,v)=P(d|n)\times\sum_{i\in D} P(d|v,i) \quad (8)$$

4.2.2に述べたように，尤度をP(d,n,v)間の相対的な大小関係として表現する必要があるため，尤度$pl_L(n,v,d)$を以下の式(9)により計算する。

$$pl_L(n,v,d)=\frac{P(d,n,v)}{\sqrt{\sum_{i\in D} P(i,n,v)^2}} \quad (9)$$

この尤度は，表層表現と対応付けられる深層格を推測するために用いられるものであり，尤度が閾値T_2以上となった深層格を学習データにおける推測結果とする。このとき，複数の深層格の尤度が閾値以上となる場合や，どの深層格の尤度も閾値未満となる可能性があるため，推測結果が一意に定まるとは限らない。本論文では予備実験の結果から，閾値T_2の値を0.7とした。尤度を単位ベクトルとして捉えている本手法において，0.7という閾値は，推測結果を最大でも二つまでしか出力できず，基本的には一つの深層格を出力することを意味している。

深層格選好は個々の単語に関する知識であり，文単位における深層格の関係を考慮していない。文単位における深層格の関係を扱った原理として，一文一格の原理（the one-instance-per-clause principle）が広く知られている（Filmore, 1968)。一文一格の原理を考慮して学習することで，学習された規則は，深層格選好のみでは考慮できない文単位の関係を補う知識となることが期待できる。したがって，ある動詞に格支配されている全ての名詞の深層格が一意に推測され，かつ，推測された深層格が一文一格の原理を満たしている文を抽出し，これらの文を用いて規則の学習を行う。抽出された文中の名詞nに付属する表層情報をgとし，深層格dが推測された頻度$fq_R(g,d)$を求める。深層格選好と同様に条件付き確率となるよう以下の式

(10)に基づいて計算し，表層情報gを満たす名詞がdとなる傾向を示す一確率P(d|g)を深層格推測規則として学習する。

$$P(d|g)=\frac{fq_R(g,d)}{\sum_{i\in D} fq_R(g,i)} \quad (10)$$

4.2.5　深層格推測

最終的な深層格の推測は，4.2.3と4.2.4で学習された深層格選好と深層格推測規則の両方を考慮して行われる。尤度を式(2)と式(3)に従って計算する。ただし，nやvが学習データに存在しない場合には，P(d|n)やP(d|v,d_2)の代わりに，クラスタ単位での深層格選好P(d|nc)やP(d|vc,d_2)を用いてP(d,n,v,g)を求める。P(d,n,v,g)は，P(d,n,v)とP(d|g)の積で近似されているが，P(d,n,v)は，名詞と動詞の深層格選好の積であるため，結果として全ての深層格に対する尤度が0となる場合が考えられる。その場合には，どの深層格にも解釈され得るとして，P(d,n,v)の値を1として計算する。P(d|g)に関しても学習データから有効な値が学習できなかった場合が考えられる。その場合には，P(d,n,v)と同様に，P(d|g)の値を1として計算する。また，P(d,n,v)とP(d|g)の積を計算した結果，全ての深層格に対する尤度が0となった場合には，表層情報による規則を優先し，P(d|g)の値を最終的な尤度とする。4.2.4で述べたように，閾値T_2以上の尤度となった深層格を推測結果とした。

4.3　実験と考察

4.3.1　日英比較実験

本手法が，対象となる言語に依存せずに深層格を推測できることを確認するために，EDRの日本語コーパスと英語コーパスを用いた実験を行った。実験で使用した文は，日本語と英語それぞれに対して以下の手順で抽出した。コーパスから表4.3に示す深層格を含む意味フレームを抽出する。二つの名詞を格支配している意味フレームの単語を基に，構文木から対応する文節または句を特定し，二つの名詞と一つの動詞で構成される文を再構築する。この際，「れる」や「せる」などの深層格に変化を及ぼす助動詞を含む文，及び，be動詞＋〝－ed〟などの受動態の文を除外した。再構築された文の中で，それぞれの文を構成する，名詞，動詞，機能語の出現頻度が全て3以上の文のみを実験で使用する文とした。最終的に，日本語2,492文，英語7,485文が抽出された。日本語のデータに関しては量的に十分ではないため，11分割交差検定法による実験を行った。英語のデータに関しては，抽出された文の中から，11分の1に相当する681文をランダムに選択しテストデータとし，残りの6,804文を学習データとした。日英ともに学習データは同一の文を二組用意し，一方は深層格推測規則学習のために深層格のタグを取り除き，もう一方は深層格選好学習のためにタグをそのまま残した。

推測結果の評価は，名詞単位で行い，以下の式にしたがって再現率，精度，F値を求めた。

$$recall=\frac{NCO}{NA} \quad (11)$$

$$precision=\frac{NCO}{NO} \quad (12)$$

$$Fvalue=\frac{2\times recall\times precision}{recall+precision} \quad (13)$$

NAはコーパス中で付与されている深層格の総数であり，NOは推測された深層格の総数である。NCOは，推測された深層格が正解であった数である。正誤の判断として，EDRコーパスで付与されている概念関係子と一致している推測結果を正解とした。

本手法は，人手による労力の軽減を目的としたMinimally supervised learningであり，深層格推測を対象として，Minimally super-

vised learningを用いた従来手法は存在していない。したがって，Supervised learningである，正解タグを用いて深層格推測規則を学習する手法をベースライン手法として設定し，提案手法がどの程度Supervised learningに迫れるかという観点から評価することとした。

4.3.2 結果と考察

日本語と英語における実験結果を表4.5と表4.6にそれぞれ示す。クローズドデータである学習データに対する結果を上段に，オープンデータであるテストデータに対する結果を下段に記している。また，本手法の推測は，式(3)に示すように，深層格選好P(d,n,v)による推測と，深層格推測規則P(d|g)による推測に大別できるため，それぞれ単独で推測した場合の結果についても，ベースライン手法による結果とともに併記している。

4.3.3 日本語の結果

まず，日本語の結果について考察する。表4.5の対象数と出力数の関係をみると，クローズドデータとオープンデータともに同程度の値であり，精度と再現率の間に大きな差は存在しない。ベースライン手法とのF値の比較において，クローズデータでは1.5ポイント，オープンデータでは3.5ポイントの差であった。Minimally supervised learningとSupervised learningの一般的な差に関する知見は存在しないため，主観的な判断をせざるを得ないが，1.5から3.5ポイント差というのは，タグなしデータから規則を学習する提案手法が，タグ付きデータから規則を学習したベースライン手法に迫る結果であると考えられる。

本手法が，人手による労力の軽減を目的として，2つの名詞を格支配する動詞句に対象を限定しているのに対し，大石らの手法(1995)は，より複雑な統語構造をもつ文も

表4.5 日本語における実験結果

		対象数	出力数	正解数	精　度	再現率	F　値
クローズドデータ	提案手法	49,840	47,840	38,858	81.2%	78.0%	79.6%
	深層格選好のみ	49,840	48,034	35,374	73.6%	71.0%	72.3%
	深層格推測規則のみ	49,840	46,046	35,478	77.1%	71.2%	74.0%
	ベースライン手法	49,840	48,234	39,783	82.5%	79.9%	81.1%
オープンデータ	提案手法	4,984	4,458	3,097	69.5%	62.1%	65.6%
	深層格選好のみ	4,984	4,703	2,970	63.2%	59.6%	61.3%
	深層格推測規則のみ	4,984	4,249	3,117	73.4%	62.5%	67.5%
	ベースライン手法	4,984	4,427	3,251	73.4%	65.2%	69.1%

表4.6 英語における実験結果

		対象数	出力数	正解数	精　度	再現率	F　値
クローズドデータ	提案手法	13,608	13,487	10,589	78.5%	77.8%	78.2%
	深層格選好のみ	13,608	13,389	10,047	75.0%	73.8%	74.4%
	深層格推測規則のみ	13,608	13,303	10,006	75.2%	73.5%	74.4%
	ベースライン手法	13,608	13,282	10,487	79.0%	77.1%	78.0%
オープンデータ	提案手法	1,362	1,360	999	73.5%	73.3%	73.4%
	深層格選好のみ	1,362	1,350	953	70.6%	70.0%	70.3%
	深層格推測規則のみ	1,362	1,336	982	73.5%	72.1%	72.8%
	ベースライン手法	1,362	1,362	1,054	77.4%	77.4%	77.4%

対象としているため，直接比較することは適当ではない。しかしながら，同じ深層格推測に関する研究であるため，参考として大石らの結果について以下に言及する。大石らの結果では，学習による知識のみを用いた推測結果がEDRコーパスと一致した正解率は80.32%であった。本手法のクローズドデータに対する精度は81.2%，再現率は78.0%であり，対象の違いを考慮する必要があるが，大石らの手法と大きな差はない。大石らの手法は表層的な表現を学習の手がかりとしているため，日本語以外の言語に適用するためには，コーパスとは別に表層表現を事前に与えなおす必要がある。本手法は，言語ごとにコーパスを用意するだけでよく，汎用性に優れている。対象の違いによる影響はあるが，このように汎用性の高い枠組みの中で，従来手法と同程度の精度であったことは本手法の有効性を示すと考えられる。

本手法のオープンデータに対する精度は69.5%，再現率は62.1%であり，ベースライン手法と比較すると精度で3.9ポイント，再現率で3.1ポイントの低下であり，クローズドデータにおける差と同程度であった。

深層格選好のみを用いた推測のF値を調べると，クローズドデータで72.3%，オープンデータで61.3%という値であり，単語の概念がもつ深層格の傾向を手がかりとすることで，6割以上の精度で推測可能であることが確認できる。この深層格選好の推測結果から学習された深層格推測規則のみを用いた推測のF値を調べると，クローズドデータで74.0%，オープンデータで67.5%という値であった。このことは，一文一格の原理を考慮することで，深層格選好よりも精度の高い規則をタグなしコーパスから学習できたことを示している。

深層格選好と規則を併用した提案手法のF値は，クローズドデータにおいては最も高い値を示したが，オープンデータにおいては規則を単独で用いた場合よりも僅かに低い値となった。しかしながら，クローズドデータとオープンデータにおけるベースライン手法の結果の差から，今回の実験で用いた学習データの量が，オープンデータをクローズドデータと同程度の精度で推測するには不足していたと考えられる。したがって，学習データの量を増加させることにより，提案手法が規則を単独で用いた場合よりも向上することが期待できる。今回の実験では，十分な量の学習データを確保できなかったが，今後，学習データの量を増加させ精度が向上するか調査したいと考えている。

4.3.4 英語の結果

次に英語の結果について考察する。表4.6に示すように，クローズドデータに対する本手法の精度は78.5%，再現率は77.8%であり，オープンデータに対する精度は73.5%，再現率は73.3%であった。日本語の場合と同様に，精度と再現率の間に大きな差はない。また，ベースライン手法とのF値の差は，提案手法がクローズドデータで0.2ポイント上回り，オープンデータで4.0ポイント下回った結果となり，日本語と比較して大きな差はないと考えられる。

深層格選好のみを用いた推測のF値は，クローズドデータで74.4%，オープンデータで70.3%であり，7割以上の精度で推測可能であることが確認できる。日本語の場合と比較して，オープンデータとクローズドデータの間に顕著な差が表れなかったのは，英語の方が学習データの異なり数が多く十分に学習できた結果と考えられる。学習が比較的十分に行われたと考えられるクローズドデータでのF値を比較すると，日本語の場合と同程度であり，単語概念の深層格選好を手がかりとした推測は，日本語と英語の違いに影響を受けないことが確認できた。

深層格推測規則のみを用いた推測のF値は，

クローズドデータで74.4%，オープンデータで72.8%であった。この結果は，日本語の場合と同じく，深層格選好よりも精度の高い規則が学習できたことを示している。深層格選好と規則を併用した場合のF値と比較すると，日本語の場合と異なり，クローズドデータ，オープンデータともに，併用した手法の方が高い値となった。これは，日本語のときと異なり，十分な量の学習データが存在したために，クローズドデータと同程度の精度でオープンデータを推測することが可能になったためだと考えられる。

4.3.5　言語非依存性の考察

4.3.3，および，4.3.4の結果から，本手法は，日本語と英語の違いによる影響を殆ど受けることなく，7割から8割程度の精度で推測可能であることが確認できた。また，言語非依存性を実現するための中核となる深層格選好による推測は，クローズドデータに対して7割以上，オープンデータに対して6割以上の精度で推測可能であり，その結果を基にタグなしコーパスから学習された規則を併用することで，全てのデータにおいて精度と再現率を向上できたことも確認できた。本実験では，日本語と英語のみを対象としたが，日英の比較において顕著な差が確認できないことから，他の言語に対しても適用できる可能性がある。今後の課題として，他の言語に対しても実験を行い，日英の場合と同程度の精度で推測可能であるか調査したいと考えている。

4.3.6　推測結果の例と今後の課題

本手法の問題点を明らかにし今後の改善に役立てるため，テストデータ中の具体的な推測結果を例示して検討する。

・正解例

まず，深層格選好による推測の段階で正解であった例として，日本語では「父を[object] 戦場で [place] 失い」や「私たちは [agent] 握手を [object] 交わした」など，英語では〝Japan [agent] started from scratch [source]〟や〝those [agent] refuse appointment [object]〟などが挙げられる。ただし，ここで例示する文は名詞，動詞，機能語から再構築した文であり，前後の文脈や，形容詞，副詞，冠詞などは省略されている。これらの例は，日本語で2,613例，英語で913例が存在した。

これらの例において，〝Japan〟などはobjectやplaceとしても解釈されていたがagentとして解釈される率が高かったため，深層格選好のみを用いて正しく推測することができた。また，規則による推測も，深層格選好による推測と同じ深層格を導くものであるか，式(3)に示すように深層格選好との積として尤度を求めた場合に異なる深層格を導くようなものではなかった。

・改善例

次に，深層格選好による推測の段階では誤りであったが，規則と併用することにより改善された例を示す。日本語では，「消費者が [agent] 商品を [object] 買うとき」や「自然を [object] 相手に [goal] する」など，英語では〝I [agent] 'm going to bank [goal]〟や〝sell them [object] in Japan [place]〟などが挙げられる。このような例は，日本語で484例，英語で86例が存在した。

殆どの例は，尤度計算において規則との積をとった結果，深層格選好による誤った推測結果が抑制されたものであった。例えば，「消費者が　商品を　買うとき」において，深層格選好による推測段階では，「消費者」の尤度はobjectが最も高い値であり，正解であるagentの尤度は二番目であった。しかしながら，[[が：F2] [を：F1]] の機

能語パターンにおいて［が：F2］に適用される規則は，agent，object，goalの順に高い値であり，objectとgoalに対する値は殆ど0に近い値であった。したがって，規則との積を計算することでobjectの尤度が0に近い値となった結果，agentの尤度がobjectの尤度を超えることとなり，「消費者」の深層格を正しく推測することができた。

また，改善された結果の中で，深層格選好と規則をそれぞれ単独で用いた場合には誤りとなるが，両方を併用した場合にのみ正解となった例は，日本語で6例，英語で2例が存在した。日本語では「電話で［implement］注文を［object］出し合い」など，英語では〝reduce stress［object］at work［scene］〟などが挙げられる。

これらの例では全て，規則との積で尤度を求めた結果，併用した場合にのみ正しい推測を行うことができた。規則との積による尤度の抑制以外で改善された例としては，規則との積をとった結果，全ての深層格に対する尤度が0となり，4.2.5で述べたように規則の尤度を優先したため正解となった例などが挙げられる。

・悪化例

反対に，深層格選好による推測の段階で正解であったにも関わらず，規則と併用することにより誤りとなった例は，日本語で357例，英語で40例が存在した。

日本語では「地方に［place（*goal*）］本拠を［object］おく」や「日本では［place］恩赦は［object（*agent*）］及ばなかったが」など，英語では〝Kaifu［agent］becomes prime-minister［goal（*object*）］〟や〝it［object（*agent*）］turns profit［goal（*object*）］〟などが挙げられる。例文中の（ ）内に記されたイタリック体の深層格は，誤って推測された深層格である。

悪化した原因は，改善された原因と全て同じ理由によるものであった。すなわち，「地方に」のように，規則との積をとることで正解であった深層格の尤度が抑制されたことによるものか，「恩赦は」のように，全ての深層格の尤度が0となったために規則の尤度を優先した結果として誤りとなったものである。したがって，この問題を解決するためには，規則の学習精度を向上させる必要があり，一文一格の原理以外の制約を考慮した学習を行う必要があると考えられる。また，英語の例のように，機能語が存在せず語順のみしか表層情報として利用できない場合には，規則の粒度が粗いものとなり精度の低下を招くため，規則を細分化する手段が必要になると考えられる。

ただし，規則の併用により改善された場合と異なり，それぞれ単独で用いた場合には正解であるが，併用した場合のみ誤りとなるような例は，日英ともに存在しなかった。

・誤り例

最後に，深層格選好による推測の段階で誤りであり，規則と併用することによっても改善されなかった例を示す。日本語では「交通事故で［scene（*cause*）］けがを［object］して」や「トヨタに［beneficiary（*agent*）］エンジンを［object］供給したことも」など，英語では〝imports［object］are increasing from Southeast Asia［source（*place*）］〟や〝you［agent］stay for dinner［scene（*object*）］〟などが挙げられる。誤りであった例の殆どはこのグループに属し，日本語で1,530例，英語で323例が存在した。

改善できなかった原因の一つとして，sceneやbeneficiaryなど比較的稀な深層格が正解であったことが挙げられる。深層格選好による推測は，統計的手法の一種であるため，稀な事例に対する推測は一般的な事例に対する推測と比較して低い精度となる傾向に

ある。また，規則の学習も，深層格選好による推測結果に基づくため，稀な事例に対する影響を免れない。本手法は，深層格選好や規則を単独で用いた場合に低い尤度であっても，両者の積をとった後に単位ベクトルに再変換するため，稀な深層格が推測されないわけではない。例えば，「交通事故で」におけるcauseの尤度は，深層格選好のみの場合には0.02，規則のみの場合には0.04であるが，併用することで1.00に尤度が向上している。しかしながら，「交通事故で」のsceneや「トヨタに」のbeneficiaryなどは，深層格選好と規則のどちらの尤度も0であるため，積を計算しても対処できなかった。また，〝from Southeast Asia〟の場合には，規則による推測ではsourceの尤度が存在していたが，深層格選好による推測ではplaceにしか値が存在しなかった。このような統計上の不備を解決するために，統計的に得られる情報以外の知識を人手で与えることが考えられるが，4.1.1で述べたように，網羅性やコストの面で問題が残る。それゆえ，一般的な原理や制約から統計情報以外の知識を学習する機構が必要となるが，この点に関しては今後の課題である。

・規則の学習に対する考察

深層格選好のみを用いた推測結果から，深層格推測規則を併用することで，改善または悪化した例の数を表4.7にまとめる。表4.7に示すように，日英ともに悪化した例よりも，改善された例の方が多かった。日本語，英語ともに，改善効果のある規則を学習できたという結果は，本手法の学習の有効性を示すと考えられる。

表4.7　規則による改善と悪化の数

	正解のまま	改善	悪化	誤りのまま
日本語	2,613	484	357	1,530
英語	913	86	40	323

4.4　深層格推測手法の応用

提案した深層格推測手法を応用先として自動要約システムを選択した。インターネットの普及などにより，今日，入手可能な情報量は，人間が通常処理できる量を大きく超えており，性能の良い自動要約システムへの期待が高まっている。しかしながら，現在の自動要約システムは，文章の中で重要と思われる文をそのまま抜き出すなどの処理や，文の表現を体言止めにするなどの言い換え的な処理など，表面的な加工に留まっている。高度な要約を行うためには意味処理が必要であり，提案した深層格推測の技術を応用する対象として妥当であると考えられる。自動要約には照応処理が不可欠であるため，深層格推測結果を応用した照応システムを作成し，理想的環境の下での動作確認を行った。

照応システムは，意味的に関連している2文を入力し，入力された2文を1文に合成するものである。2番目に入力された文の要素のうち，欠落している情報を1番目に入力された文の中から適切に抽出するという処理を以下の手順で行う。

まず，1番目に入力された文に対して，DCAPRを用いて深層格を推測し，推測された深層格を含む深層格パターンから，文中の名詞の意味的カテゴリーを特定する。次に，2番目に入力された文に対して，DCAPRを用いて深層格を推測する。推測された深層格を含む深層格パターンから，欠落している深層格，及び，その深層格を担う名詞の意味的カテゴリーを特定する。2文目で特定された意味的カテゴリーと同一の意味的カテゴリーとなりうる名詞が1文目に存在しているならば，その名詞を2文目の欠落している情報の候補とみなす。最後に，その候補が一文一格の原理などの制約に違反していないかを調査して，深層格に対応する名詞を決定する。

例えば，「太郎は公園で遊んでいた。そこで花子は太郎と話した。」という2文が入力

されたと仮定する。まず，1文目「太郎は公園で遊んでいた」に対して，DCAPRを用いて，「遊ぶ」のagentが「太郎」，placeが「公園」であると推測を行う。次に，2文目「そこで花子は太郎と話した」に対しても同様に，「話す」のagentが「花子」，goalが「太郎」であると推測を行う。「話す」の深層格には，agentとgoalの他に，「どこで」話したかというplaceと，「何を」話したかというobjectなどがあるので，これらの情報が欠落していると判断する。1文目において「公園」がplaceとして用いられていること，および，「公園」の方が「太郎」よりもplaceを担う傾向が強いことから，placeの候補として「公園」が選択される。また，「太郎」の方が「公園」よりもobjectを担う傾向が強いことから，objectの候補として「太郎」が選択される。最後に，これらの候補が「花子は太郎と話した」という文において妥当であるかを調査する。「公園」がplaceを担うことには問題がないが，「太郎」がobjectを担うことは一文一格の原理に違反する。したがって，「太郎」は照応の先行詞として相応しくないと判断される。以上から，2文目の「花子は太郎と話した」は，「花子は太郎と公園で話した」という意味であると判断することができる。

しかしながら，深層格は文における意味の一面だけを捉えたものであり，人間と同じような処理を行わせるためには，さらに多くの様々な要因を考慮する必要がある。その意味で，現在のシステムはトイシステムの域を出ておらず，今後の実用化に向けて深層格以外の知識を扱えるよう改善が必要である。しかしながら，2章や3章で述べたように，深層格以外の知識をコンピュータに与える際にも機械学習による労力の軽減が必要であると考えられる。

5章　結　　論

本論文では，情報化社会における円滑なコミュニケーションの実現を目指して，情報化社会の現状と今後の動向を踏まえ，自然言語処理に求められる機能に関する考察を行った。

2章では，情報化社会における自然言語処理の役割を考察し，今後の自然言語処理には，ユーザに過剰反応を起こさせないような技術の見直し，意味の領域にまで踏み込んだ処理を行う必要性，対象となる知識をユーザの補助なしでシステムが自動的に学習する必要性があることを提言した。

3章では，現在の自然言語処理の主流であるコーパス・統計モデルに至るまでの歴史を説明し，利用者の側から見た場合に，コーパス・統計モデルが抱えている問題点を考察した。コーパス・統計モデルの問題点である，直観的な理解の困難さ，誤り訂正の不確実さといった部分を解決するためには，かつての文法や意味の理論に再び踏み込む必要があることに言及した。また，これらの理論は機械学習により自動的に構築されることが望ましく，今後の自然言語処理には，このような観点からの手法が必要であることを提言した。

4章では，このような観点による手法のひとつとして，深層格推測手法DCAPRの提案を行った。推測に用いる知識は，網羅性とコストの問題に対処するため，一定量のタグ付きコーパスから得られた知識を中核として，タグなしコーパスから深層格推測の規則を学習する。提案手法は，単語の概念ごとに，それぞれ特定の深層格に解釈される傾向（深層格選好）をもつという仮定に基づき，一定量のタグ付きコーパスから深層格選好を計算した後，その値を用いてタグなしコーパスから多様な言語表現に対処するための規則を学習する。深層格選好を手がかりの主体とすることにより，提案手法は言語の違いに依存せずに深層格を推測することが可能となる。言語

非依存性を検討するため，日本語と英語のEDRコーパスを用いて，二つの名詞をとる単文を対象に実験を行った。その結果，クローズドデータにおける日本語の精度81.2%（再現率78.0%），英語の精度78.5%（再現率77.8%），オープンデータにおける日本語の精度69.5%（再現率62.1%），英語の精度73.5%（再現率73.3%）となり，言語の差異による影響を回避する枠組みの中で，教師つき学習と同程度の精度による推測を行うことができた。また，オープンデータの内，規則と併用することにより，深層格選好のみを用いた推測と比較して，日本語では484例，英語では86例が全体として改善されており，タグなしコーパスから学習した規則の有効性を確認することができた。また，DCAPRの応用例のひとつとして自動要約システムをあげ，そのための基礎研究として，深層格に基づく照応システムを記述した。

今後の課題として，実用化に向けた自然言語処理技術の改善，シミュレーションを用いて知識処理技術の普及を考察することがあげられる。

注

1章注釈

(注1) 顧客の自発的な記述を促すための選択肢の効果を否定するものではない。ここで問題としているのは，最終的に得られる情報を抽出するためのソースとしての価値である。

(注2) 情報処理学会では，研究領域を，「コンピュータサイエンス領域」，「情報環境領域」，「フロンティア領域」の3領域に分割しているが，自然言語処理研究会はフロンティア領域に属している。

(注3) 人工知能などにおいて，知識とは，「ある目的のために必要な真であるとして与えられたデータや手続き，あるいはそれらのついてのデータや手続き」のことを指している。知識処理の例としては，データベースからの知識獲得，エキスパートシステム，オントロジーなどがあげられる。

2章注釈

(注1) 複数の無線アクセス網（同種及び異種無線アクセス網）をシームレスかつ高速に切り替えるための技術。

(注2) フォトニックネットワークとは，伝送，多重化，多重分離，交換，経路制御などのネットワーク機能を，すべて光技術だけで行うネットワークのこと。

(注3) 複数の無線アクセス網を安全に接続するための技術。

(注4) 理想的なコンピュータアーキテクチャの構築を目指して，1984年に東京大学の坂村健氏が始めたプロジェクト。近未来の高度にコンピュータ化された社会におけるコンピュータやネットワークのあり方を研究している。その究極的な目標は，コンピュータが内蔵されネットワークに接続された機器が協調動作する超機能分散システムの実現である。現在6つの基礎サブプロジェクトといくつかの応用サブプロジェクトが進行している。

(注5) 知識処理と自然言語処理の定義は明確なものではなく，その境界を特定することは困難である。したがって，知識処理が完全に自然言語処理を包含した概念であるとは断言できない。しかしながら，「人間のもつ知識を利用した情報処理」という観点からは，知識処理の対象を言語情報に限定したものを自然言語処理とみなしても大きな齟齬はないと思われる。

(注6) ラッダイト運動の心理に通じるものがあるという主張は，ルーウェリン反応という概念を提唱した山本ゆうじ（2005）による。

3章の注釈

(注1) コンピュータを「電子素子を用いた演算装置」と定義すると，黎明期のコンピュータには，1942年のABC，1943年のColossus，1946年のENIAC，1948年のThe Baby，1949年のEDSACがあげられる。

(注2) 機械翻訳の考え方自体は，最初のコンピュータが作られるよりも前の1930年代にアルメニア系フランス人のジョルジュ・アツルーニ(Georges Artsrouni）とロシアの技術者ペトル・ペトロビッチ・トロヤンスキー（Petr Petrovic Trojanskij）によって提案されている。

4章注釈

(注1) 自然言語処理における尤度（plausibility）は必ずしも確率論における尤度（likelihood）と同じ使われ方をするわけではない。4.2.2で用いられているように相対的な尤もらしさに対しても尤度という言葉を用いることがある。

参考文献

Axelrod, R. (1984) The Evolution of Cooperation, Basic books.（松田裕之訳「つきあい方の科学――バクテリアから国際問題まで――」ミネルヴァ書房，1998）

Axelrod, R. (1997) The Complexity of Cooperation, Princeton University Press.（寺野隆雄監訳「対立と協調の科学――エージェント・ベース・モデルによる複雑系の解明――」ダイヤモンド社，2003）

Blaheta, D. and Charniak, E. (2000) Assigning Function Tags to Parsed Text, Proc. 1st NAACL, pp.234-240.

Caraballo, S.A. (1999) Automatic construction of a hypernym-labeled noun hierarchy from text, Proc. of 37th ACL, pp.120-126.

Chomsky, N. (1957) Syntactic Structures. Mouton Publishers, Paris.

EDR（1995）㈱日本電子化辞書研究所，EDR電子化辞書使用説明書。

Fillmore, C. J. (1968) The case for case in E. Bach & R. T. Harms (eds.) Universals in Linguistic Theory, NewYork, Holt, Rinehart & Winston: 1-88.

Gilbert, N. and Troitzsch, K.G. (1999) Simulation for the Social Scientist, Open University Press.（井庭崇・岩村拓哉・高部陽平訳「社会シュミレーションの技法　政治・経済・社会をめぐる思考技術のフロンティア」日本評論社，2003）

Gildea, D. and Jurafsky, D. (2000) Automatic Labeling of Semantic Roles, Proc. 38th ACL.

Gildea, D. and Jurafsky, D. (2002) Automatic labeling of semantic roles, Computational Linguistics, Vol.28, No.3, pp.245-288.

Holland, J.H. (1992) Adaptation in Natural and Artificial Systems. MIT Press.（嘉数侑昇監訳，皆川雅章・三上貞芳・横井浩史・高取則彦・鈴木恵二・川上敬共訳「遺伝アルゴリズムの理論――自然・人工システムにおける適応――」森北出版，1999）

Howard, N. (1971) Paradoxes of Rationality: Theory of Metagames and Political Behavior, MIT Press.

JUMAN（2005）黒橋禎夫，長尾真，日本語形態素解析システムJUMAN version 5.1。

KNP（2005）黒橋禎夫，日本語構文解析システムKNP version 2.0 b6。

Russell, S.J. and Norvig, P. (1995) Artificial Intelligence: A Modern Approach, Prentice-Hall, Inc.（古川康一監訳「エージェントアプローチ人工知能」共立出版，1997）

Turban, E., Lee, J., King, D. and Chung, H.M. (2000) Electronic Commerce: A managerial Perspective, Prentice-Hall, Inc.（阿保栄司・麻田孝治・秋川卓也・木下敏・島津誠・浪平博人・牧田行雄訳「e-コマース　電子商取引のすべて」ピアソン・エデュケーション，2000）

Wallace, P. (1999) The Psychology of the Internet, Cambridge University Press.（川浦康至・貝塚泉訳「インターネットの心理学」NTT出版，2001）

芥川龍之介（1986），ちくま文庫　芥川龍之介全集2，筑摩書房，東京。

石原英樹・金井雅之（2002）「シリーズ意思決定の科学5　進化的意思決定」朝倉書店。

牛丸元（2004）「戦略的提携のガバナンス」北海学園大学『経営論集』第1巻第4号：1-9。

大石亨，松本裕治（1995）「格パターン分析に基づく動詞の語彙知識獲得」情処学論，Vol.36，No.11，pp.2597-2610。

岡田章（1996）「ゲーム理論」有斐閣。

桑島健一・高橋伸夫（2001）「シリーズ意思決定の科学3　組織と意思決定」朝倉書店。

小山正太，乾伸雄，小谷善行（2003）「「名詞と表層格」パターンに対する深層格対応の推測」情処学研報，NL-154-22。

渋木英潔・荒木健治・栃内香次（2003）「一文一格の原理と規則化に基づいた深層格の自動推測手法」FIT 2003情報科学技術フォーラム情報技術レターズ，第2巻：91-92。

渋木英潔・荒木健治・桃内佳雄・栃内香次（2004）「深層格の推測手法における自動クラスタリングの利用」FIT 2004情報科学技術フォーラム情報技術レターズ，第3巻：79-80。

渋木英潔（2005）「繰り返し囚人のジレンマ・ゲームにおけるノイズ抑制戦略」北海学園大学大学院経営学研究科研究論集第3号，pp.1-18。

高橋伸夫（1999）「生存と多様性――エコロジカル・アプローチ――」白桃書房。

高増明・服部純典（1999a）「協調の進化：繰り返し囚人のディレンマゲームのコンピューター・シミュレーション」大阪産業大学論集社会科学編，第111号：55-68。

高増明・服部純典（1999b）「繰り返し囚人のディレンマ・ゲームにおけるノイズとメモリーサイズ」大阪産業大学論集社会科学編，第113号：91-103。

田村誠（2001)「空間配置型N人版囚人のジレンマゲーム」ABSコンペティション。
長尾真（1996）長尾真（編)「岩波講座ソフトウェア科学15　自然言語処理」岩波書店，東京。
原田実，田淵和幸，大野高宏（2002)「日本語意味解析システムSAGEの高速化・高精度化とコーパスによる精度評価」情報処理学会論文誌，Vol.43，No.9，pp.2894-2902。
分類語彙表（1964（1993))国立国語研究所，分類語彙表，秀英出版。
本木実，鳴津好生，高橋直人（2000)「階層型ニューラルネットによる深層格解析」情処学論，Vol.41，No.10，pp.2852-2862。
山本ゆうじ（2005)「翻訳工学に向けて－MT＋TM翻訳ワークフロー－SATILA」AAMTジャーナル，no.3。