自然言語処理の再挑戦 ～統計的言語処理を超えて～

An Introduction to Natural Language Processing: Beyond Statistical Methods

乾健太郎　　浅原正幸
奈良先端科学技術大学院大学情報科学研究科

1　はじめに

Web の爆発的な普及とともに情報の流通が加速度的に増大し，個人でも大量の情報を入手し，発信できる時代になった．膨大な情報の中から必要なものを検索したり，日々生産される情報を自動分類したり，要約や可視化によって分析したりする「情報加工技術」が今後ますます重要になると予想される．加工の対象となる情報はほとんどが日本語や英語のような自然言語 (natural language) で表現されている．したがって，上の意味で「情報を加工する」ためには，自然言語で書かれた文書の集まりから情報を抽出し，構造化し，再構成する自然言語処理 (natural language processing) 技術が極めて重要な役割をになう．

自然言語処理研究の究極の目標は，ありていに言うと，言語を理解する機械を作ることである．言うまでもなく，この目標は人間の知能を作ろうとする人工知能全体の目標と同じくらい遠大であり，我々はまだその途についたばかりだ．しかし，確かな進展もあった．とくに，1990 年代以降，確率理論や機械学習研究の発展とハードウェアの飛躍的な進歩によって種々の経験的手法が一気に実用レベルに到達し，さらに Web の登場によって大量の電子化文書が流通し，入手可能になったという事情も手伝って，経験的手法に基づく言語処理技術が革新的な発展をとげた．また，情報化社会の深化に伴って，言語処理技術を核とする応用技術も広がりを見せている．

形態素解析 (morphological analysis)，構文（係り受け）解析 (syntactic analysis) といった基礎技術の成熟はもう一つの大きな意味を持っている．1980 年代，意味解析 (semantic analysis) や意図理解 (intention understanding) の名で当時盛んに研究された高度な言語理解 (language understanding) は，それに必要な大量の言語知識や世界知識を用意する方法論を持たなかったために，実用規模には発展しなかった．しかし，形態素解析や構文解析技術が実用レベルに達しつつある現在，これをうまく利用して言語知識や世界知識を大量の電子化文書から自動的に獲得するというシナリオが現実味をおびてきている．そうした試みが成果を上げれば，今度は実用規模の知識という新しい武器を持って高度な言語理解の実現に再挑戦することができるだろう．この意味で，自然言語処理研究は今新しい段階を迎えつつある．

本稿の目的は，自然言語処理の研究を始めるのに有用な情報を自然言語処理以外の研究領域で活躍している研究者に提供することである．以下，まず 2 節で自然言語処理の基本問題について論じ，3 節でそれらの問題を解く代表的な方法論を紹介する．4 節では自然言語処理でよく用いられる言語資源やツールなどを紹介し，最後に 5 節で現在今後注力すべき研究課題の例を挙げる．

自然言語処理は，研究対象そのものに言語依存性があるという点で科学技術分野としてはいささか特異な研究分野である．日本語の解析と英語の解析では少なくとも必要な辞書が異なるので，両者は完全に同じではない．しかし，本質的な問題の多くはどの言語にも共通で，ある言語で成功した方法論は別の言語にも適用できる場合が多い．本稿では言語依存性・横断性の問題には触れず，もっぱら日本語の処理を例にとって解説する．入手可能な資源やツールも日本語のものを中心に紹介する．日本語以外の言語の資源，ツールについては，ACL，LDC，ELRA の Web サイト（表 3）を参照すると良い．

2　自然言語処理の基本問題

言語は，我々人間が情報を創造し伝達し蓄積するためのメディアである．したがって,「言語を処理する」とは，言語というメディアによって伝達・蓄積された情報を処理することと言える．本節では，近年注目を集めている言語処理アプリケーションの一つ，意見マイニングを例にとって「言語を処理する」ことの意味を考え，言語処理の中心的な問題の多くが「情報の抽出・構造化」と「同義・含意関係の認識」の２つに集約できることを示す．

2.1　意見マイニング

意見マイニング (opinion mining) とは，Web 上のブログ記事や社会調査の自由回答アンケートのような大規模な文書集合から個人が発信する意見を抽出し，構造化情報として蓄積することにより，ユーザの関心に合わせて検索したり，要約・分析したりすることを可能にする情報加工サービスである [31]．

今，文書集合中に次のようなパッセージが含まれていたとしよう．

(1) 週末に母とやまぶきに行ってきました．私はせいろを頼んだのですが，そばの旨みが活きていて絶品でした^^)

意見マイニングでは，まず入力文字列を解析し，こうした個人の意見を例えば (2) のような構造化された情報として抽出する．

(2)
$$\begin{pmatrix} \text{意見のタイプ} & = & \text{評価} \\ \text{評価者} & = & \text{著者} \\ \text{評価の対象} & = & \text{「やまぶき」} \\ \text{評価の対象のクラス} & = & \text{飲食店} \\ \text{評価の着目点} & = & \text{「せいろ」} \\ \text{評価の着目点のクラス} & = & \text{料理} \\ \text{評価の値} & = & \text{「絶品だ」} \end{pmatrix}$$

このように構造化された情報をここでは便宜的に意見ユニットと呼ぶことにしよう．単なる文字の羅列でしかなかった元のパッセージに比べると，意見ユニットは内部に明示的な構造を持っており，情報の検索や分析に都合がよい．これと同様の解析を大量の文書に施せば，大量の意見ユニットが関係データベースのように構造化された情報として得られる．

こうして大量の意見ユニットが集まると，次はそれらを様々な観点から分類し，分析することを考える．集合の要素を分類（あるいはクラスタリング）するためには，要素同士の類似性を判断するモデルが必要である．例えば,「やまぶきの蕎麦は絶対オススメです」という記述からは (3) のような意見ユニットが得られるが，これは先の意見ユニット (2) にかなり内容が近い．

(3)
$$\begin{pmatrix} \text{意見のタイプ} & = & \text{評価} \\ \text{評価者} & = & \text{著者} \\ \text{評価の対象} & = & \text{「やまぶき」} \\ \text{評価の対象のクラス} & = & \text{飲食店} \\ \text{評価の着目点} & = & \text{「蕎麦」} \\ \text{評価の着目点のクラス} & = & \text{料理} \\ \text{評価の値} & = & \text{「オススメだ」} \end{pmatrix}$$

こうした意味的な類似性に基づいて意見ユニットの集合をクラスタリングすることができれば，どのような意見がどのくらいの多いかといった俯瞰的な情報を得ることができる．また，予め決めておいた分類項目に意見ユニットを分類し要約することができれば，レーダチャートのような形に可視化することも可能になる．

上の例が示すように，意見マイニングの中心的なタスクは次の２つである．

情報の抽出・構造化　自然言語の文字列から意見情報を抽出し，構造化された形式（意見ユニット）に変換する処理 (information extraction and structurization)

同義・含意関係の認識　抽出された情報（意見ユニット）どうしの意味的な類似性あるいは包含関係を認識する処理 (paraphrase and entailment recognition)

これら２種類の技術で実現できる言語処理アプリケーションは意見マイニングに限らない．むしろ，ほとんどのアプリケーションがこれら２種類の問題に帰着する．つまり，この２つの問題が自然言語処理の基本問題であると言える．以下，この２つの基本問題を，情報伝達メディアとしての自然言語が持つ２つの重要な特性に関連づけながら概観する．

2.2 言語の特性１：効率性が多義性の問題を生む

言語は，複雑で多様な情報を有限のアルファベットからなる文字列，それも我々人間が読める程度の長さの文字列に符号化できるという点で極めて効率的な符号である．例えば，先の意見マイニングの例では,「やまぶき」という文字列が蕎麦屋を指しているが，同じ文字列が別の文脈では特定の植物を指すかもしれないし，色を指すかもしれない．同じ文字列が２種類以上の情報を表現できるというのはかなり効率的である．また，自然言語では文脈から復元できる情報はしばしば省略される．先の例では，誰が「やまぶきに行った」のか，何が「絶品」なのかは明示的には書かれていないが，こうした情報は文脈から容易に復元できる．

この人間にとって便利な性質は，しかしながら機械にとっては極めて都合が悪い．テキストから欲しい情報を抽出するためには，上の「やまぶき」が植物でも色でもなく，蕎麦屋を指すことを識別できないといけないし，著者が「絶品」と褒めているものが「私」でも「やまぶき」でもなく,「せいろ」であることを認識できないといけない．このように言語処理では，たくさんの解釈の可能性の中から話し手あるいは書き手が意図した「真の」解釈をどのように選択するかが極めて重要な問題になる．与えられた言語表現に対して表面的にはいくつもの解釈があるように見えることを**曖昧性**(ambiguity) があると言い，いくつもの解釈の中から「真の」解釈を推定する作業を**曖昧性解消**(disambiguation) と言う．一口に曖昧性といっても，さまざまなレベルの曖昧性が存在し，それぞれの解消が図 1 に示すような言語処理の部分問題を形成している．それらを総称して**自然言語解析**(natural language analysys) と呼ぶ．テキストから情報を抽出し，構造化する処理はまさにこの自然言語解析を行うことに他ならない．

2.3 言語の特性２：冗長性が同義性の問題を生む

言語は，上の意味で効率的である一方，同じ情報を伝える言語表現がいくつも存在するという点で冗長でもある．

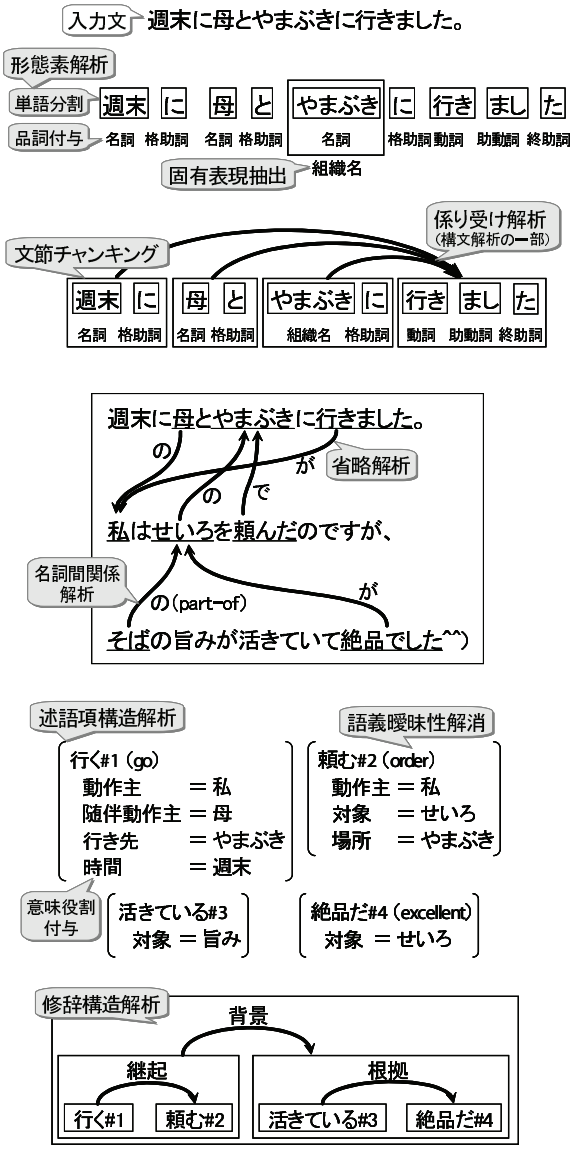

図 1: 自然言語解析の部分問題

意見マイニングの例に戻ろう．ある料理を「絶品」と評することと,「オススメ」と評することは意味的に近い.「(料理が) オススメ (だ)」という表現は，その料理がかなり「美味しい」ことを暗に伝えており,「美味しい」は「(料理が) 絶品 (だ)」と意味的に似ているからだ．このように，言語は，その効率性ゆえに，同じ情報を明示的にも暗示的にも伝えることができる．すなわち，言葉には同じ情報を表現する，あるいは含意する言い回しがいくつも用意されていることになる．

言語が持つこの冗長性は我々の豊かな言語文化の源になっているが，言語処理にとっては曖昧性

に並ぶもう一つの大問題である．同義・含意関係の認識が必要となる理由もそこにある．

例えば「V. ユーゴー」と「レ・ミゼラブル」のような《著者名，書名》の関係がいつも

(4) 「レ・ミゼラブル」の著者はV. ユーゴーだ
《書名》　　《著者名》

のような決まった言い回しで述べられるのであれば，Web 上の文書に書かれている《著者名，書名》関係を網羅的に抽出してデータベース化するといった情報抽出課題も比較的単純なパターンマッチでできるはずである．しかし，実際には，

(5) 《著者名》が「《書名》」を著す
《著者名》が「《書名》」を発表する
《著者名》の代表作「《書名》」
《書名》(《著者名》)

のように同じ関係を含意する言い回しがいくつもあり，それらを漏れなく探すことは決して容易ではない．見方を変えると，文書集合から《著者名，書名》の関係を網羅的に収集する情報抽出課題 (information extraction) は，存在限量された型付きの変数 $X_{\text{書名}}$，$Y_{\text{人名}}$ を含む文 (6) を含意する記述を文書集合から探し出す含意認識問題に帰着すると考えることができる．

(6) $X_{\text{書名}}$ の著者は $Y_{\text{人名}}$ だ

問題の本質は，情報検索 (information retrieval) や質問応答 (question answering) と呼ばれる課題でも，あるいは言語処理研究の胎動期からの古い歴史を持つ機械翻訳 (machine translation) でも同じである．質問応答は，例えば (7a) のような質問文の答えを (7b) のような記述から探し出す課題である．これは，質問文を平叙文化した (7c) が情報源のテキスト (7b) から含意されるか否かを識別する含意認識問題に帰着する．この例では，$X_{\text{人名}}$ に「坊ちゃん」を代入した文が (7b) から含意されるので，質問の答えが「坊ちゃん」であることがわかる．

(7) a. 『坊ちゃん』の著者は誰ですか？
b. 夏目漱石は明治 39 年の春に『坊ちゃん』を雑誌「ホトトギス」に発表,...
c. 『坊ちゃん』の著者は $X_{\text{人名}}$ です

機械翻訳はある言語の文を別の言語の同義な文に変換するタスクであるが，これも原理的には与えられた 2 つの文が意味的に同義か否かを識別できれば実現できる．入力文を形態素解析や係り受け解析して，構造化するのは，その方が同義性の識別が容易になるためと考えてもよい．

このように，現在の言語処理アプリケーションの多くは，異なるテキストの間の同義性あるいは含意関係を認識する問題に帰着させることができる [3, 30]．

3　経験的手法と知識

前節では，言語処理の主たる問題が曖昧性の解消と同義・含意関係の認識の 2 種類に集約できることを述べた．本節では，いずれの問題もラベル付け問題に分解することができ，それによって確率モデルや機械学習に基づく経験的手法を適用する道が開けること，またそうしたラベル付け問題の補助問題として知識の設計と集積，そして自動獲得が重要であることを述べる．

3.1　ラベル付け問題としての言語処理

言語解析（曖昧性解消）とは解釈の選択肢の中から正しい解釈を選択することであるので，どんな部分問題も原理的には分類問題あるいはラベル付け問題に帰着させることができる．

例えば，日本語の形態素解析（図 1 参照）は，長さ m の文字列からなる文 $\mathbf{s} = c_1 \dots c_m$ を入力として，これを単語の列 $\mathbf{w} = w_1 \dots w_n$ に分割し（単語分割），各単語 w_i に品詞 t_i（$\mathbf{t} = t_1 \dots t_n$）を割り当てる（品詞付与）タスクである．単語分割は，$\mathbf{s}$ 中の各文字と文字の間が単語の境界になっているか否かを判定する問題と見なせる．したがって，各文字 c_i の右が単語境界か否かをラベル b_i で表すとすると，シンボル系列 $\mathbf{s}$ にラベル $\mathbf{b} = b_1 \dots b_m$ を付与する**系列ラベリング問題**(sequential labeling) として定式化できる．品詞付与も同様に，系列 $\mathbf{w}$ にラベル $\mathbf{t}$ を付与する系列ラベリング問題である．

もう一つ，固有表現抽出の例を示そう．固有表現抽出は，入力テキストに対して，人名，組織名，地名といった固有表現の出現箇所を特定する問題である．固有表現は日々生産され，網羅的な辞書

北	京	大	に	入	学	す	る
ORG-B	ORG-I	ORG-I	O	O	O	O	O

図 2: 系列ラベリングによる固有表現抽出

を用意することができないので，未知の固有表現を識別する必要がある．この問題も図 2 のような系列ラベリング問題と見なすことができる．図中のラベル ORG-B は組織名の開始位置，ORG-I は組織名の内部，O はそれ以外を表わし，この例では部分文字列「北京大」が組織名であることを示している．

問題の構造は，係り受け解析 (dependency analysis) や省略解析 (elipsis or zero-anaphora resolution) など，より構造的な情報を扱う課題でも同じである．係り受け解析は入力文中の各文節について係り先の文節を一つ選ぶ問題，省略解析は入力文章中の各省略箇所について先行詞を一つ文章内あるいは外界から選ぶ問題と見なせるので，ともにラベル付け問題の組み合わせで表現できる．また，同義・含意関係認識についても明らかに同様の一般化が成り立ち，もっとも単純化すれば，与えられた 2 つの解析済みテキストが同義あるいは含意関係にあるか否かを推定する二値分類問題と見なすことができる．

3.2 統計的言語処理の発展と課題

このように言語処理の問題を一旦ラベル付け問題として定式化できれば，ラベル付け問題の解法として研究されてきた様々な統計的手法や機械学習アルゴリズムが適用できるようになる．例えば，上の形態素解析では，入力 $\mathbf{s}$ に対する単語・品詞系列 $\mathbf{w}, \mathbf{t}$ の事後確率を最大化する問題と見なし，これを次式で与えられる隠れマルコフモデル（Hidden-Markov Model; HMM）と呼ばれる確率モデルで解くのが代表的である．

$$\begin{aligned}&\arg\max_{\mathbf{w},\mathbf{t}} P(\mathbf{w},\mathbf{t}|\mathbf{s}) = \arg\max_{\mathbf{w},\mathbf{t}} P(\mathbf{w},\mathbf{t}) \\ &= \arg\max_{\mathbf{w},\mathbf{t}} \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1})\end{aligned}$$

$P(w_i|t_i)$ や $P(t_i|t_{i-1})$ のようなパラメタは，人手によって単語分割と品詞付与が行われた正解ラベル付きデータ（タグ付きコーパス，4 節参照）が大量にあれば，対応する事象の相対頻度から近似的に推定することができる．

また，最近では，条件付き確率場（Conditional Random Fields; CRF）[12] に代表される**識別モデル**(discriminative model) でラベル系列の事後確率 $P(\mathbf{w},\mathbf{t}|\mathbf{s})$ を直接最大化する手法も注目を集めている．CRF では，入力とラベルの組を特徴づける様々な特徴量[1]を使ってラベルの事後確率を求める．具体的には，入力 $\mathbf{x}$ とラベル $\mathbf{y}$ の組に対して，特徴量関数の集合 $f_i(\mathbf{x},\mathbf{y})$（$i = 1,\ldots,n$）を用意し，正解ラベル付きデータを使って次式の多クラス・ロジスティック回帰モデルを訓練する（各特徴量の重み w_i を最尤推定する）．

$$P(\mathbf{y}|\mathbf{x}) = \frac{\exp\left(\sum_i w_i f_i(\mathbf{x},\mathbf{y})\right)}{\sum_{\hat{\mathbf{y}}} \exp\left(\sum_i w_i f_i(\mathbf{x},\hat{\mathbf{y}})\right)}$$

HMM のように対象の同時分布を確率パラメタの積で近似するモデルとは異なり，CRF のような識別モデルではラベルの推定に有効と考えられる様々な手がかりを特徴量として自由に組み込むことができる．例えば，形態素解析では，部分文字列や字種といった，HMM では扱えない手がかりを採り入れることによって成果を上げてきた．言語解析ではどのような手がかりがどの場合にどの程度有効かを人間の直感だけで判断するのが困難な場合が多く，CRF のように特徴量を自由に設計できる学習モデルは極めて有用性が高い．この利点は，サポートベクタマシンやパーセプトロン，ブースティングなど，確率に基づかない分類器学習アルゴリズムを使う場合も同様である．

こうした機械学習技術の進歩がハードウェアの進歩と相まって，実用規模の高次元空間が比較的容易に扱えるようになり，とくに形態素解析，固有表現抽出，係り受け解析といった基礎技術は大きな発展を遂げた．これらのタスクは，さながら機械学習の実験場のような様相を見せている．現在，日本語新聞記事に対する精度は，形態素解析が形態素単位で約 97〜98%，係り受け解析が係り受け単位で約 90%となっており，改善の余地はあるものの，実用レベルに達しつつあると言える．

しかしこれは，4 節で紹介するように新聞記事についてはよく整備された比較的大規模な正解ラベル付きデータがあったことに多くを負っている．技術論文や小説，あるいはメールやブログ記事といった他のジャンルにはそのようなデータが存在

[1] 言語処理の分野では，パターン認識で言うところの特徴量を慣習的に「素性（feature）」と呼ぶ．本稿でもこの用語を用いる．

しないため，同様の性能を得ることは難しい．あらゆる言語のあらゆるジャンル，あらゆるスタイルについて個別に十分な規模の正解データを人手で用意するのは現実的ではない．とすれば，教師あり学習に基づく手法には，教師データが十分に得られないという**データの過疎性**(data sparseness)の問題がついてまわる．明らかに，教師あり学習だけに頼る方法には限界があると考える．意味役割付与や述語項構造解析，照応・省略解析といったより深い解析では同じ問題がさらに顕著になることも容易に想像できよう．

この問題に対するアプローチの一つは，半教師あり学習 (semi-supervised learning) あるいは能動学習 (active learning) によって，より少数のラベル付きデータからモデルを学習する方策を考えることである．半教師あり学習を指向する手法は，典型的には次のステップからなる．

1. 最初に少量のラベル付きデータ（シード）を用意し，教師データとする．
2. 教師データを使ってモデルを訓練する．
3. 得られたモデルでラベルなしデータを解析し，モデルが付けたラベルのうち正しそうなものを教師データに擬似的に加える．
4. 2 と 3 のサイクルを繰り返す．

例えば，病気の名前を抽出する固有名抽出課題 (named-entity recognition) の場合，初めに多義性の少ない病名をシードとしていくつか用意し，それらの出現文脈をコーパスから収集する．集まった事例を教師データと見なしてモデルを訓練すると，モデルは例えば「『を煩う』という文字列の直前の単語は病名である可能性が高い」といった手がかりを発見するかもしれない．このモデルでラベルなしデータを解析すると，今度はシードになかった新しい病名が見つかる可能性がある．モデルの訓練と教師データの拡張を交互に行うこの種の方式は**ブートストラッピング**(bootstrapping) と呼ばれることもある．

3.3 言語/世界知識の設計開発と自動獲得

教師データの過疎性に対するもう一つの重要な方策は知識の整備である．様々なアプリケーションにおける様々な部分問題で共通に有用となる言語知識 (linguistic knowledge) あるいは世界知識 (world knowledge) を予め整備し，これを効果的に使うことによって，教師データの不足を補うことができる．知識の整備と共有化，そしてその方法に関する研究は，言語処理全般にわたる重要な課題である．

例えば，係り受け解析（図 1）では，動詞「食べる」は他動詞で「〜を」(「ヲ格」) をとるが「寝る」は自動詞でヲ格をとらない,「食べる」のヲ格には食べ物を指す名詞が入る，といった知識が有用である．係り受け情報を付与した大量の教師データを入手できれば，こうした規則性も自然に学習できるかもしれないが，現実はそうでない．そこで，こうした知識を教師あり学習とは違う方法で開発し，蓄積していくことが必要になる．

必ず必要なのは，**格フレーム辞書**や**シソーラス**などの語彙の知識である．格フレーム (case frame) とは，上の「食べる」の例のように，個々の用言がどのような項（ガ格，ヲ格など）をどのような種類の名詞とともに取るかに関する知識である．シソーラスは，単語間の意味的な類似性に関する知識に基づいて語彙を階層的に分類したもので,「昼ご飯/昼食/ランチ」のような表現の多様性を吸収し，学習の効率を上げるのに役立つ．格フレーム辞書は，係り受け解析や省略解析，述語項構造解析 (predicate-argument structure analysis) など，種々の言語解析に用いられる．一方，シソーラス (thesaurus) は，こうした言語解析の他，同義・含意関係の認識にも欠かせない．例えば，2.1 の意見マイニングの例で 2 つの意見ユニット (2) と (3) の類似性を認識するためには,「せいろ」と「蕎麦」,「(料理が) オススメだ」と「(料理が) 絶品だ」の意味的な類似性の知識が必要である．

同義・含意関係の認識には，さらに事態間の含意関係に関する膨大な知識も必要になる．例えば,「他人のものを盗む」行為が「罪を犯す」行為の下位事象である，あるいは「映画館に行く」行為は「映画を観る」行為の手段の一つであり，前者は後者をたいていの場合に含意する，といった事態や行為に関する豊富な常識的知識が言語処理技術の深化とともに必須の資源となってくるだろう．

語彙の知識にせよ，事態の知識にせよ，実用に耐えるには非常に大規模な資源の開発が必要である．そこで，すでにある程度成熟した形態素・構文解析技術あるいは固有表現抽出技術を利用し，大量の

テキストデータからより高度な言語処理に要する知識を獲得する研究が精力的に行われている．統計的言語処理技術の発展，爆発的なテキストデータの流通，計算資源の大規模化といった要因が重なり，こうした知識の自動獲得を実用規模で考えることが可能になってきた．

共起パターン（１次の共起）のマイニング　どの表現がどの表現と同一文中に一緒に出現（共起）しやすいか，すなわち表現間の共起 (cooccurrence) に関する統計情報を手がかりにして，個々の表現の使われ方に関する知識を獲得する [14]．例えば，上の「食べる」「寝る」の例のような用言格フレームの知識は，各用言と項の共起頻度を統計的に分析することによって得られる可能性がある．さらに，ここで「表現」を「事態」に置き換えて，事態間の共起情報を集めることにすると，事態間の関連性や因果関係を知る手がかりも得られる．「商品を《製造》し**て**《販売》する」「《交通事故》**で**《怪我をする》」といった共起を考えれば，その可能性が想像できよう [32, 20]．

出現文脈の類似性（２次の共起）の利用　意味の近い単語は同じような使われ方をする傾向がある．例えば,「日本酒」と「焼酎」はともに「を飲む」「を生産する」「を買う」ような表現とよく共起する．逆に，よく共起する表現の分布が似ている２つの単語は意味も似ていることが多い．この性質（**分布仮説**(distributional hypothesis) と呼ばれることもある）を使えば，単語間の類似度をそれぞれの単語の出現文脈の分布の類似度から推定することができる [14]．分布仮説は広い概念で，統計的言語処理の様々な場面で利用されている．先に紹介したブートストラッピングによる固有表現抽出も基本となるアイデアを分布仮説に負っている．

Web 文書の半構造情報の利用　Web 文書中の表や箇条書きといった半構造情報も知識獲得に役立つ場合がある．病名のリストを箇条書きで記載した Web 文書があれば病名抽出に直接使える知識が得られるし [18]，デジカメのスペック表からはデジカメの属性表現に関する知識が得られる．

4　入手可能な資源とツール

前節では言語処理がラベル付け問題として帰着できることを示した．本節では実際に言語処理をラベル付け問題として解くにあたってよく用いられる言語資源やツール類を紹介する．ラベルを付与する言語資源，ラベル付け問題を解く際に用いられる言語情報を付与するための言語解析ツール，自分で設計したラベルを付与したり言語解析ツールの出力を修正したりするコーパス管理ツール，実際にラベル付け問題を解く機械学習器，係り受け解析により木構造化された解析済みデータを分析する頻出パターンマイニングツールを順に紹介する．

4.1　辞書とコーパス

重要な言語資源として，辞書 (lexicon) とコーパス (corpus) の２つがある．辞書はある目的のために集められた単語の集合のことをいい，その多くは品詞や意味などの情報が付与されている．コーパスとはある目的のために集められたテキストデータのことをいい，新聞記事や小説，話し言葉の書き起こしなどの様々なデータが公開，販売されている．表 1 に広く用いられている言語資源の例をあげる．

比較的自由に利用できる大規模な辞書として，形態素解析用辞書である JUMAN 辞書と IPADIC がある．JUMAN 辞書は，益岡・田窪文法に基づく体系 [37]，IPADIC は IPADIC 品詞体系 [27] に基づく品詞が付与されている．その他，読み（かな表記，発音），活用形の情報が付与されている．

意味上の上位下位関係に基づき，木構造もしくは有向無閉路グラフの形に構造化された辞書（概念辞書，シソーラス; thesaurus）もいくつか公開されている．分類語彙表 [35]，日本語語彙大系 [28, 29] は単語を意味情報に基づき木構造に分類したシソーラスである．小規模ではあるが，用例とともに動詞，形容詞がどのような格助詞を取るか（表層格情報）の情報が付与された辞書として IPAL 辞書がある．また，EDR 辞書は，構造化されており，表層格情報を含んでいるだけでなく，実例文から抽出された共起の情報も付与されている．Web 上のフリー百科辞典 Wikipedia は約 24 万語の用語に対する解説文からなる．Wikipedia の中では用

表 1: さまざまな言語資源

言語資源	特徴	URI など
IPADIC	品詞，活用形，かな表記，発音	http://chasen.naist.jp/hiki/ChaSen/
JUMAN 辞書	品詞，活用形，読み，代表表記	http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html
分類語彙表	シソーラス	http://www.kokken.go.jp/katsudo/kanko/data/
IPAL 辞書	意味情報，表層格情報	
日本語語彙大系	シソーラス，構文パターンなど	
EDR 辞書	品詞，かな表記，発音，活用情報，表層格情報，シソーラス，共起情報	http://www.iijnet.or.jp/edr/
Wikipedia	Web 上のフリー百科辞典	http://download.wikimedia.org/
毎日新聞	新聞記事	http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html
朝日新聞	新聞記事	http://www.kinokuniya.co.jp/02f/d13/2_13a001.htm
日経新聞	新聞記事	http://sub.nikkeish.co.jp/gengo/zenbun.htm
読売新聞	新聞記事	http://www.yomiuri.co.jp/cdrom/etc/oshirase.htm
青空文庫	文学作品など	http://www.aozora.gr.jp/
プロジェクト杉田玄白	文学作品などの翻訳	http://genpaku.org/
京都大学テキストコーパス	品詞，係り受け情報，格関係，照応・省略関係，共参照関係	http://nlp.kuee.kyoto-u.ac.jp/nl-resource/corpus.html
RWC テキストデータベース	形態素情報，意味情報	
EDR コーパス	形態素情報，構文情報，意味情報	http://www.iijnet.or.jp/edr/
日本語話し言葉コーパス	形態素情報，節単位情報，音韻学的情報，係り受け情報，要約・重要文情報，談話構造情報など	http://www2.kokken.go.jp/∼csj/public/index_j.html
河原らによる Web から自動獲得した大規模格フレーム	格フレーム情報	http://nlp.kuee.kyoto-u.ac.jp/nl-resource/caseframe.html
NAIST Text Corpus	照応・共参照情報	http://cl.naist.jp/nldata/corpus/

語が木構造からなるカテゴリに分類され，簡易シソーラスとしても用いることができる．

ラベルが付与されていない手に入りやすい大規模テキストデータとして新聞記事がある．毎日新聞社，朝日新聞社，日本経済新聞社，読売新聞社などが，研究目的向けに記事データを販売している．安価なテキストデータとして著作権が切れた文学作品などがある．青空文庫は，著作権が切れた文学作品を電子化し公開している．プロジェクト杉田玄白は，様々な文書を翻訳して公開している．

ラベルが付与されているデータもいくつか公開されている．京都大学テキストコーパスは，毎日新聞の記事約 4 万文に対して，品詞情報や文節情報，文節間の係り受け関係の情報が付与されたデータである．最新版では，このうちの 5000 文に対して，格関係，照応・省略関係，共参照関係などが付与されている．また，奈良先端大は同テキスト全文に対する照応・共参照情報を別に公開している．EDR コーパスは，EDR 辞書と同じく，日本電子化辞書研究所により整備されたテキストデータである．20 万文に対し，形態素情報，構文情報，意味情報などが付与されている．日本語話し言葉コーパスは，話し言葉についてのデータベースである．独話の音声データを書き起こしたものに対し，形態素情報，節単位情報，音韻学的情報，係り受け情報，要約・重要文情報，談話構造情報などが付与されている．また河原らは，Web から抽出 した例文集合 5 億文を，そこから自動獲得した格フレーム (case frame) 辞書とともに公開している [6]．

4.2 言語解析ツール

4.1 では手に入りやすい言語資源について紹介してきたが，全ての分野のテキストについて言語情報が付与されたデータが収集できるわけではない．そのようなデータに対して，形態素情報を付与するツール（形態素解析器; morphological analyzer）や係り受け関係を付与するツール（係り受け解析器; dependency analyzer）が公開されている．

ChaSen[38] は，IPADIC 品詞体系の品詞を付与することができる形態素解析器である．JUMAN は，益岡・田窪品詞体系の品詞を付与することができる形態素解析器である．表記ゆれの問題に対処するために，代表表記の情報を付与することができる．MeCab は，条件付き確率場に基づく形態

素解析器[10]である．JUMAN 辞書，IPADIC の辞書を切り換えることにより，益岡・田窪品詞体系と IPADIC 品詞体系のいずれかの品詞を出力することができる．制約付き解析に加えて，条件付き確率場が出力する周辺確率を用いて，ソフトなわかち書き（単語単位の尤度つき複数解出力）[33]が可能である．また最新版では学習器を含むようになり，ユーザが辞書と品詞ラベルつきコーパスを用意することにより，自分で形態素解析モデルを構成することも可能となった．

CaboCha は，機械学習器サポートベクタマシン（support vector machines; SVM）に基づく係り受け解析器[7, 9]である．形態素解析 ChaSen もしくは MeCab の出力を前提とし，文節まとめあげ，固有表現抽出，係り受け解析を行う．KNP は，京都大学で開発された係り受け解析器である．形態素解析 JUMAN の出力を前提とし，同形異義語の処理，文節まとめあげ，並列構造解析を経て，係り受け解析を行う．

4.3 コーパスの作成・管理ツール

コーパス中に出現する辞書中の単語を検索する際に，単純なパターンマッチによる手法では実行に時間がかかるため，大規模なデータを扱うことが難しい．ここでは，テキストから高速に文字列を検索するライブラリを紹介する．接尾辞配列 (suffix array)[13, 39]は，文字列検索 (String Search) アルゴリズムの一つである．テキスト中の接尾辞（テキスト中のある位置からテキスト末尾までの文字列）を辞書順に並べ，それに対するポインタを配列に格納したものを用いる．sary は，接尾辞配列の一実装で，オプションを用いることにより，日本語の様々な文字コードを指定してインデックスを作成することができる．別の実装として SUFARY がある．ダブル配列 (double array)[26]は，トライ (trie) 法[19]に基づく木構造を２つの並列した配列に格納し，それを基にして検索するアルゴリズムである．darts は，ダブル配列の実装である．動的更新などの機能はないが，辞書などの準静的なものを格納するのに適している．このような文字列検索アルゴリズムのライブラリは前述の形態素解析器の内部で用いられている．

4.2 では，用意したテキストに対し，形態素情報や係り受け関係を付与するツールを紹介した．いずれも新聞記事を基にしてパラメータ推定を行うことにより構成されたツールであり，異なる分野のテキストでは多くの誤りを出力する．ChaKi は，形態素解析結果および係り受け解析結果に特化したコーパスを検索するツール（コーパスコンコーダンサ）[15]であると同時に形態素解析の誤りや係り受け関係の誤りを修正する機能を持つ．検索機能は，テキスト中の文字列，形態素解析結果に対する表層文字列や品詞情報などによる絞り込み検索，係り受け解析結果に対する絞り込み検索が可能である（図 3）．修正機能は，形態素解析の層と係り受け解析の層の２つに分かれている．形態素解析のわかち書きの単位を辞書引きしながら修正したり，文節を分割・連結したり，係り受け関係を修正したりすることができる．

テキストの部分文字列を切り出してラベルを付与したり，切り出した部分文字列間に関係を付与したりする場合，これらの情報は XML により保持されることが多い．ここで XML を簡便に扱うことができる oXygen XML Editor を紹介する．oXygen は，Windows, Mac OS X, Linux, Solaris などで動作する XML エディタである．デスクトップアプリケーションとして用いられる他，統合開発環境である Eclipse のプラグインとしても動作することができる．XML スキーマ や DTD などを定義することにより，簡単にラベルづけツールを構成するだけでなく，付与された情報が妥当かどうかの検証も行うことができる．XSLT や XQuery エンジンなどにより，ラベル情報の整形や，問い合わせなどができる．

4.4 機械学習・頻出パターンマイニング

係り受け解析器を用いて木構造になったテキストデータを作成することができ，コーパス作成・管理ツールを用いて，実際に自分で分析したいラベルが付与されたデータを構成することができるようになった．ここで，木構造データに対してラベルを自動的に付与したり，木構造に頻出する部分木を分析したりするためのツールを紹介する．る部分構

従来，言語解析器の出力を基に出現する単語を二値特徴量とした機械学習による分類手法が用いられていたが，近年構文木の部分木構造を特徴量として分類する機械学習器がいくつか公開されて

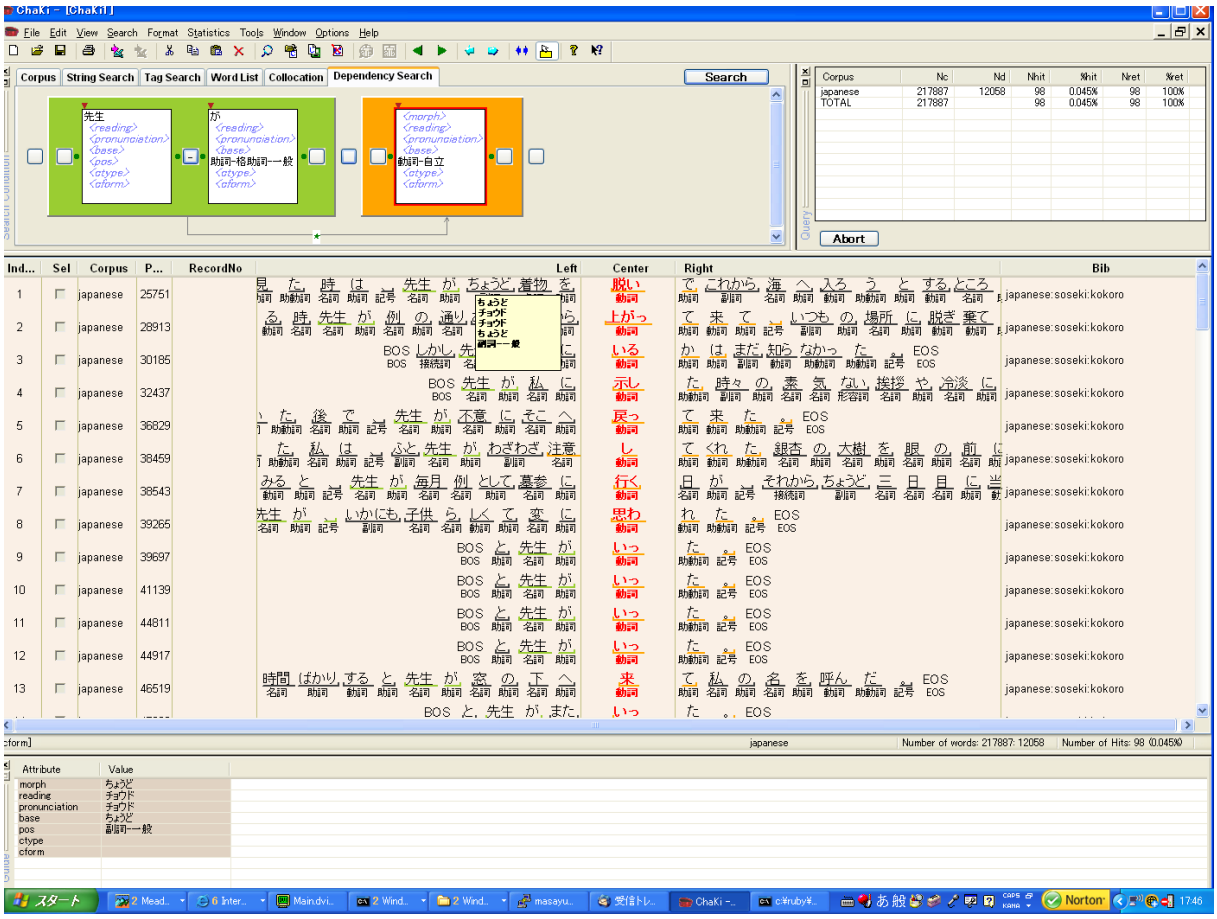

図 3: コーパスコンコーダンサ：ChaKi 検索画面

いる．amis[16] は，ロジスティック回帰による学習器であり，木構造を教師データとして与えることにより部分木を特徴量として学習することができる．libsvm は，二値分類器である SVM の一実装であり，多クラスの分類にも対応している．他の実装として，svm_light がある [4]．svm_light は，ラベルつきデータとともにラベルなしデータを用いて学習する transductive 法のための効率的なアルゴリズムが実装されている．さらに，拡張として木構造を分類するためのカーネル関数 Tree Kernel[2, 21] の実装も公開されている．木構造を分類する他の学習器として BACT がある [11]．BACT は，与えられた木構造の部分木を特徴量とした Decision Stumps を弱学習器とした Boosting アルゴリズムによる機械学習器である．学習結果として，分類に有用な特徴を持つ部分木構造を出力するため，特徴量の分析をするのに向いている．

系列ラベリングツールも汎用のものがいくつかある．YamCha は，SVM を用いた系列ラベリングツールである [8]．系列の先頭もしくは末尾から，SVM によりラベルを順次決定的に推定していくことにより，ラベルを付与する．他の系列ラベリングモデルとして，Lafferty らにより提案された条件付き確率場 (conditional random fields)[12] がある．条件付き確率場の実装としては，MALLET や CRF++ がある．CRF++ は，トークン毎の周辺確率 (marginal probability) を出力する．CRF++ の出力を基に，切り出したい部分トークン列の周辺確率の積を計算することにより，切り出された単位に全体に対する周辺確率が計算でき，その値は切り出された単位の尤度として用いることができる．

次に，与えられたテキストにおいて，頻出する部分系列や部分木を枚挙するための，頻出パターンマイニングツールを紹介する．頻出パターンマイニングツールは，例えば，ある動詞に共起する名詞

などを枚挙するためなどに用いられる．テキスト中の任意のトークンの並び（間に別のトークンが入ってもよい）を系列パターンと呼ぶ．prefixspan は [17] 系列パターンを効率よく枚挙するアルゴリズムである．この prefixspan の拡張として連続する系列パターン（閉系列パターン）のみを出力する CloSpan[24] アルゴリズムや，重複して枚挙される被覆される部分パターンを取り除き，極大パターンのみを出力する BIDE[22] アルゴリズムがある．木構造やグラフ構造などから部分構造を枚挙するアルゴリズムも提案されている FREQT[1, 25] は，木構造中に出現する高頻度の部分木を枚挙するアルゴリズムである．gSpan[23] はグラフ構造から部分構造を枚挙するアルゴリズムである．近年構文解析器の進歩により，ある程度の精度で文の構文木を得られるようになった．これらのマイニングツールを用いることにより，構文木，または，構文木に関係などを付与したものから頻出するパターンを枚挙することによる分析が可能である．

5　おわりに — 旬な重要課題

本稿では，言語処理の主たる問題が「情報の抽出・構造化」と「同義・含意関係の認識」の２つの基本問題に集約できること，いずれの基本問題もラベル付け問題に分解でき，そこに統計に基づく言語処理技術の発展があったこと，そしてそれを補う補助問題として知識の設計開発と自動獲得が重要性を増していることを述べた．また，研究の道具として入手可能な資源やツールを日本語のものを中心に紹介した．言語処理に関する入門的解説を目的としながらも，教科書的な記述にはあえて従わず，著者らの独断的視点から書いた．そのため，内容には偏りがあり，大きなトピックでも触れなかったものが少なくない．より広く知りたい読者は表 3 の読書案内を参考にされたい．

言語処理技術が一定のレベルに達した現在，言語処理技術を用いて言語処理のための知識を獲得するという一見逆説的なパラダイムが少しずつ現実味を帯びてきたように思われる．「映画館に行く」と「映画を観る」の間の＜手段・目的＞関係のような，事態や行為に関する常識的知識を広く獲得できるようになれば，その知識を使って次は話し手や登場人物の意図を理解するといった深い言語理解の問題に再度挑戦することができる．言語処理技術に質的な変革がもたらされるだろう．多くの研究者が大きな夢を持ってこの分野に参入されることを期待したい．

最後に，今後取り組むべき課題のうち，著者らが重要かつタイムリーと考えるものをいくつか挙げる．言語処理の内外の研究者にひろく刺激となれば幸いである．

重要課題中級編：

- 現在の形態素・係り受け解析技術は正解タグ付きコーパスによる教師あり学習に基づくもので，教師データとは異なる領域（話題，スタイル）のテキストに対して解析精度が落ちる場合が少なくない．とくに，形態素解析における未知語（辞書にない語）の問題は深刻である．タグ付きコーパスの存在しない領域でも，ラベルなしテキストであれば大量に入手できる．ラベルなしデータを使って既存の辞書や解析モデルを新しい分野に適合させる技術が重要な研究課題になっている．

- 照応省略解析，名詞間関係解析，修辞構造解析といった文の境界をまたぐ解析は，文内の解析に比べて研究の蓄積そのものも浅く，実用レベルに遠く及ばないのが現状である．しかし，照応照応関係や名詞間関係の認識は，冒頭の意見マイニングの例が示唆するように，テキストからの情報抽出を中核とする多くのアプリケーションで必須の技術であり，この分野の研究は急務の課題である．

- Web の爆発的な拡大とともに，飛躍的に大規模な文書集合を処理することが求められる時代になった．超大規模文書集合の処理に耐えるスケーラビリティをいかに確保するかも重要な課題である．

重要課題上級編：

- テキストからより有用な情報を抽出し，高度に構造化するためには，テキストが陽に暗に伝えている事態の構造を認識する技術が必要になる．どのような事態が起こったのか，何が原因であったのか，結果としてどんな事態が生じたのか．個々の事態を認識し，それら

表 2: さまざまなツール

ツール	特徴	URI など
JUMAN	形態素解析器	http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html
ChaSen	形態素解析器	http://chasen.naist.jp/hiki/ChaSen/
MeCab	形態素解析器	http://mecab.sourceforge.jp/
KNP	係り受け解析器	http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html
CaboCha	係り受け解析器	http://chasen.org/∼taku/software/cabocha/
sary	辞書引きモジュール (Suffix Array)	http://sary.sourceforge.net/
SUFARY	辞書引きモジュール (Suffix Array)	http://nais.to/∼yto/tools/sufary/
darts	辞書引きモジュール (Double Array)	http://www.chasen.org/∼taku/software/darts/
ChaKi	コーパスコンコーダンサ	http://chasen.naist.jp/hiki/ChaKi/
oXygen	XML エディタ	http://www.oxygenxml.com/
maxent	ロジスティック回帰モデル	http://maxent.sourceforge.net/
amis	ロジスティック回帰モデルによる木構造分類 (feature forests)	http://www-tsujii.is.s.u-tokyo.ac.jp/∼yusuke/amis/
libsvm	SVM (多クラス分類)	http://www.csie.ntu.edu.tw/∼cjlin/libsvm/
svm_light	SVM (trunsductive 法)	http://www.cs.cornell.edu/People/tj/svm_light/
tree kernel for svm_light	SVM による木構造分類 (tree kernel)	http://ai-nlp.info.uniroma2.it/moschitti/Tree-Kernel.htm
bact	Boosting による木構造分類	http://chasen.org/∼taku/software/bact/
YamCha	SVM による系列ラベリング	http://chasen.org/∼taku/software/yamcha/
MALLET	条件付き確率場（系列ラベリング）	http://mallet.cs.umass.edu/
CRF++	条件付き確率場（系列ラベリング）	http://www.chasen.org/∼taku/software/CRF++/
prefixspan	系列パターンマイニング	http://chasen.org/∼taku/software/prefixspan/
FREQT	木構造マイニング	http://chasen.org/∼taku/software/freqt/
ILLIMINE	グラフ構造マイニング他 (gSpan など)	http://illimine.cs.uiuc.edu/download/

の間の因果関係や時間関係を理解する技術が基礎技術とアプリケーションのギャップを埋める重要な中間技術になると考えられる．

- 3 節で述べたように，言語知識および世界知識の自動獲得は今後ますます重要な課題になるだろう．とくに，同義・含意関係の認識には，物に関する知識だけでなく，事態に関する上位下位関係，部分全体関係，因果関係などを記述したオントロジー的な資源の設計と開発，そしてそれを支える知識獲得技術が不可欠である．

- 2.3 で述べたように，言語処理アプリケーションの多くは，テキスト間の同義性あるいは含意関係を認識する問題に帰着させることができる．同義・含意関係認識は，言語処理の要素技術として最上位の目標であると言ってもよい．統計的手法によって言語解析技術が大きく発展した今，含意認識の問題は次の大目標として多くの関心を集め始めている．

- 形態素解析と固有表現抽出と未知語認識，あるいは述語項構造解析と照応省略解析など，言語処理の部分問題は多くの場合たがいに依存する．こうした複数の相互依存問題を融合的に処理し，大域的な最適解を求める技術が必要になるが，融合型処理に関する研究はまだ途についたばかりである．

重要課題トレンド編：

- 冒頭でも触れたように，情報化社会の成熟にともなって，膨大な文書集合に分散して埋もれている情報，すなわち我々人類の知を掘り起こし，構造化し，目的に合わせて編集し，新たな知を創造する言語情報加工技術が今後ますます重要性を増すと考えられる [36]．Web マイニング，ブログマイニング，意見マイニング，動向分析といった分散知を再構成する技術は，今後の言語処理研究を駆動する大きな柱になるであろう．

- 人間対機械のインタラクションも古くて新しいテーマである．ロボットとのインタラクションや情報アクセスサービスとのインタラクションを実現するためには，膨大な世界知識を獲得する技術，そしてそれを使いこなす技術がなくてはならない．自然言語処理が再挑戦するテーマとして不足はあるまい．

謝辞

執筆の機会を与えてくださり，また本稿に対し示唆に富むコメントくださった加藤恒昭氏（東京大学）と松下光範氏（NTT コミュニケーション科学基礎研究所）に深く感謝いたします.

参考文献

1) K. Abe, S. Kawasoe, T. Asai, H. Arimura and S. Arikawa. “Optimized Substructure Discovery for Semi-structured Data”, In Proc. of PKDD-2002, p.p. 1–14, 2002.

2) M. Collins and N. Duffy. “New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron”, In Proc. of ACL-2002, p.p. 263–270, 2002.

3) I. Dagan, O. Glickman and B. Magnini. “The PASCAL Recognising Textual Entailment Challenge”, In Proc. of the PASCAL Challenges Workshop on Recognising Textual Entailment, 2005.

4) T. Joachims. “Making large-Scale SVM Learning Practical”, In Advances in Kernel Methods - Support Vector Learning, B. Scholkopf and C. Burges and A. Smola (ed.), MIT-Press, p.p.41–56, 1999.

5) D. Kawahara and S. Kurohashi. “A Fully-Lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis”, In Proc. of HLT-2006, p.p. 176–183, 2006.

6) D. Kawahara and S. Kurohashi. “Case Frame Compilation from the Web using High-Performance Computing”, In Proc. of LREC-2006, p.p. 1344–1347, 2006.

7) T. Kudo and Y. Matsumoto. “Japanese Dependency Analysis Based on Support Vector Machines”, In Proc. of EMNLP/VLC-2000, p.p. 18–25, 2000.

8) T. Kudo and Y. Matsumoto. “Chunking with Support Vector Machines”, In Proc. of NAACL-2001, p.p. 192–199, 2001.

9) T. Kudo and Y. Matsumoto. “Japanese Dependency Analysis using Cascaded Chunking”, In Proc. of CONLL-2002, p.p. 63–69, 2002.

10) T. Kudo, K. Yamamoto and Y. Matsumoto. “Applying Conditional Random Fields to Japanese Morphological Analysis”, In Proc. of EMNLP-2004, p.p. 230–237, 2004.

11) T. Kudo and Y. Matsumoto. “A Boosting Algorithm for Classification of Semi-Structured Text”, In Proc. of EMNLP-2004, p.p. 301–308, 2004.

12) J. Lafferty, A. McCallum and F. Pereira. “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”, In Proc. of ICML-2001, p.p. 282–289, 2001.

13) U. Manber and G. Myers. “Suffix arrays: A new method for on-line string searches”, SIAM Journal on Computing, 22 (5), p.p. 935–948, 1993.

14) Y. Matsumoto. “Lexical Knowledge Acquisition”, The Oxford Handbook of Computational Linguistics, Chapter. 21, p.p. 395–413, 2005.

15) Y. Matsumoto, M. Asahara, K. Hashimoto, Y. Tono, A Ohtani and T Morita. “An Annotated Corpus Management Tool: ChaKi”, In Proc. of LREC-2006, p.p.1418–1421, 2006.

16) Y. Miyao and J. Tsujii. “Maximum Entropy Estimation for Feature Forests”, In Proc. of HLT-2002, 2002.

17) J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal and M-C. Hsu. “PrefixSpan: Mining Sequential Patterns Efficiently by PrefixProjected Pattern Growth”, In. Proc. of ICDE-2001, p.p. 215–224, 2001.

表 3: 読書案内

教科書・参考書	
Foundations of Statistical Natural Language Processing. C. Manning and H. Schuetze. MIT PRESS. (1999).	
The Oxford Handbook of Computational Linguistics. R. Mitkov. Oxford Univ. Press. (2003).	
Handbook of Natural Language Processing. R. Dale, H. Moisl and H. Somers. Marcel Dekker Ltd. (2000).	
Modern Information Retrieval. R. Baeza-Yates and B. Ribeiro-Neto. Acm Press. (1999).	
「自然言語処理 -基礎と応用-」 田中穂積監修．電子情報通信学会．(1999).	
「自然言語処理」(岩波講座ソフトウェア科学 15) 長尾真 (編)． 岩波書店．(1996).	
「確率的言語モデル」 北研二． 東京大学出版．(1999).	
「情報検索」 徳永健伸． 東京大学出版．(1999).	
「人工知能学辞典」人工知能学会．共立出版．(2005).	
「A I 事典 第 2 版」土屋 俊 他．共立出版．(2003).	
学会・研究会	
言語処理学会	
情報処理学会 自然言語処理研究会	
情報処理学会 情報学基礎研究会	
電子情報通信学会 言語理解とコミュニケーション研究会	
電子情報通信学会 思考と言語研究会	
人工知能学会	
計量国語学会	
ACL (The Association for Computational Linguistics)	
ICCL (International Committee on Computational Linguistics)	
AFNLP (Asia Federation of Natural Language Processing)	
論文誌	
「自然言語処理」 (言語処理学会論文誌)	
「情報処理学会論文誌」	
「電子情報通信学会論文誌」	
「計量国語学」	
Computational Linguistics	
ACM Transactions on Speech and Language Processing	
ACM Transactions on Asian Language Information Processing	
Natural Language Engineering	
International Journal of Computer Processing of Oriental Languages	
ACL Anthology (論文，予稿集のアーカイブ） http://acl.ldc.upenn.edu/	
リソースカタログ	
日本の言語資源・ツールのカタログ	http://nlp.kuee.kyoto-u.ac.jp/NLP_Portal/lr-cat-j.html
自然言語データに関する情報	http://cl.naist.jp/
LDC	http://www.ldc.upenn.edu/
ELRA	http://www.elra.info/
GSK	http://www.gsk.or.jp/
リンク集	
言語処理ポータル	http://nlp.kuee.kyoto-u.ac.jp/NLP_Portal/
私のブックマーク	http://www.ai-gakkai.or.jp/jsai/journal/mybookmark/
長岡技科大	http://pub.bookmark.ne.jp/nlp/
LT-world	http://www.lt-world.org/

18) K. Shinzato and K. Torisawa, “Acquiring Hyponymy Relations from Web Documents”, In Proc. of HLT-NAACL-2004, p.p. 73–80, 2004

19) T. A. Standish, “Data Structure Technique”, Addison-Wesley, Addison-Wesley, Reading, Massachusetts, 1980.

20) K. Torisawa. “Acquiring Inference Rules with Temporal Constraints by Using Japanese Coordinated Sentences and Noun-Verb Co-occurrences” In Proc. of HLT-NAACL-2006, p.p. 57–64, 2006.

21) S. V. N. Vishwanathan and A. J. Smola. “Fast Kernels on Strings and Trees”, In Proc. of NIPS-2002, p.p. 585–592, 2003.

22) J. Wang, J. Han. “BIDE: Efficient Mining of Frequent Closed Sequences”, In Proc. of ICDE-2004, p.p.79–90, 2004.

23) X. Yan and J. Han. “gSpan: Graph-Based Substructure Pattern Mining”, In Proc of ICDM-2002, p.p. 721–724, 2002.

24) X. Yan, J. Han and R. Afshar. “CloSpan: Mining Closed Sequential Patterns in Large

Datasets”, In Proc. of SDM-2003, p.p. 166–177, 2003.

25) M. J. Zaki. “Efficiently Mining Frequent Trees in a Forest”, Proc. of KDD-2002, p.p. 71–80, 2002.

26) 青江 純一. “ダブル配列による高速ディジタル検索アルゴリズム”, 『電子情報通信学会論文誌 D』, Vol. J71-D, No. 9, p.p.1592–1600, 1988.

27) 浅原 正幸, 松本 裕治. 『ipadic version 2.7.0 ユーザーズマニュアル』, 奈良先端科学技術大学院大学, 2003.

28) 池原 悟, 宮崎 正弘, 白井 諭, 横尾 昭男, 中岩 浩巳, 小倉 健太郎, 大山 芳史, 林 良彦. 『日本語語彙大系』, 岩波書店, 1997.

29) 池原 悟, 宮崎 正弘, 白井 諭, 横尾 昭男, 中岩 浩巳, 小倉 健太郎, 大山 芳史, 林 良彦. 『日本語語彙大系 CD-ROM 版』, 岩波書店, 1999.

30) 乾健太郎, 藤田篤. “言い換え技術に関する研究動向” 『自然言語処理』, Vol. 11, No. 5, pp. 151–198, 2004.

31) 乾孝司, 奥村学. “テキストを対象とした評価情報の分析に関する研究動向” 『自然言語処理』, Vol.13, No.3, pp.201–241, 2006.

32) 乾孝司, 乾健太郎, 松本裕治. “接続標識「ため」に基づく文書集合からの因果関係知識の自動獲得” 『情報処理学会論文誌』,Vol.45, No.3, pp.919–933, 2004.

33) 工藤 拓. “形態素周辺確率を用いた分かち書きの一般化とその応用”, 言語処理学会全国大会, NLP-2005, p.p. 592–595, 2005.

34) 黒橋禎夫, 長尾 眞. “並列構造の検出に基づく長い日本語文の構文解析”, 『自然言語処理』, Vol.1, No.1, pp.35–57, 1994.

35) 国立国語研究所. 国立国語研究所資料集 14 『分類語彙表 －増補改訂版－』 大日本図書, 2004.

36) 難波英嗣. “情報抽出を利用した複数文書要約”, 同巻, 2006.

37) 益岡 隆志, 田窪 行則. 『基礎日本語文法. 改訂版』, くろしお出版, 1992.

38) 松本 裕治, 北内 啓, 山下 達雄, 平野 善隆, 松田 寛, 高岡 一馬, 浅原 正幸. 『形態素解析システム『茶筌』 version 2.3.3 使用説明書』, 奈良先端科学技術大学院大学, 2003.

39) 山下 達雄. “用語解説 「Suffix Array」” 『人工知能学会誌』, Vol. 15, No. 6, p. 1142, 2000.