形態素解析と分かち書き処理

保田　明夫
テキスト・マイニング研究会事務局長
株式会社 平和情報センター

形態素解析と分かち書き処理

はじめに

社会調査（意識調査・態度調査等）や市場調査等の自由回答型・自由記述型データ、あるいは、コールセンターやお客さま相談室などで収集された生活者や消費者の「生の声」、さらには、製造部門における工程管理や品質管理の定性情報や営業日報や会議録など、多種多様かつ膨大なテキスト型データ（textual data）を経営に活かすテキストマイニング（TM：Text Mining）への期待が高まっている。

TM あるいはテキスト型データの解析（TDA：Textual Data Analysis）に限らず、日本語の自然言語処理を行う際には、単語の分かち書き（単語分割：word segmentation） は最も基本的かかつ重要な課題である。英語やフランス語などの欧米言語の正書法（orthography）では単語と単語の間に空白を入れる「分かち書き」という習慣があるため、単語の認識は大きな問題にはならないが（一般に空白で区切られた単位が単語であるという暗黙の合意が成立している)、日本語のように単語の間に区切り記号を持たない言語（「分かち書き」という習慣を持たない）は、単語の認識が根本的な問題になる。

形態素の定義は明確に与えられているわけではないが、言語学では、意味を担う最小の言語要素（それ以上分割できない語の単位）を形態素（morpheme）と呼ぶ。この形態素を解析する処理、すなわち、自然言語の文中の単語を識別し（tokenization)、その語形変化を解析し（lemmatization, stemming)、品詞を同定する（part-of-speech tagging）処理を形態素解析（morphological analysis）と呼ぶ。形態素解析は、構文解析、意味解析、文脈理解などともに自然言語処理の基礎となる要素技術であり、仮名漢字変換、音声認識、情報検索、機械翻訳などに適用されている。

多くの形態素解析システムでは、隣り合った形態素間の結合に関する規則と形態素情報を記述した形態素辞書と形態素に関する文法の知識を用いて、文を単語単位に分かち書きし、それぞれの構文上の役割を決定する。

一方、言語の形態的分類からみると、英語やフランス語などは、主として語形変化によって、性、数・格などの文法的関係を示す言語であり屈折語（inflectional language）と呼ばれるのに対し、日本語は語の順序や語形変化よりも、助詞や助動詞などの付属語によって文法的な関係を示す言語であり、膠着語（こうちゃくご：agglutinative language）と呼ばれる。したがって、日本語の形態素解析では、語形変化の解析よりも単語の同定、すなわち、文を単語に分割することが重要な課題となる。

また、複合名詞（例えば：「自然言語処理」→「自然」＋「言語」＋「処理」）や複合動詞（例えば：「書き写す」→「書き」＋「写す」）などの複合語への対応も大きな課題である。単語を組み合わせることにより複合語は無限に生成されるため、言語学的な単語の定義の問題とは別にしても実用上の課題（とくに TM やテキスト型データ解析などでは重要）においても、これをどのように単語として分割すべきか議論が分かれる。

いずれにしても、計算機上の形態素辞書に格納できる単語の数は有限であり、辞書に登録された語に基づき文を分かち書きする処理が重要であり、従来から最長一致法（辞書内の単語と最も長く一致する単語を優先する)、文節数最小法（文全体の文節の数が最も少なくなるような単語の並びを優先する）などの形態素辞書に対する検索方法が採り入れられ、より一層改善されてきている。

TM あるいは TDA の処理の流れを大別すると、テキスト型データから情報や概念を抽出するステップと抽出された情報や概念を解析する（マイニング）２つのステップに分けることができる。この第１ステップである情報や概念の抽出ステップは、構造化・形式化されていないテキスト型データを次の解析ステップ（マイニング）で扱えるようにするための数値変換処理であり、TM や TDA の特徴的（象徴的）な機能である。

本テキストでは、TM や TDA における情報や概念の抽出に適用される分かち書き処理、及び形態素解析について、その内容やアルゴリズムの概要を紹介するとともに、TM における形態素解析の役割や位置づけ、その適用性について解説する。

[本テキストの内容]

１．言語理解と形態素解析

人間が通常用いている言葉（自然言語）を計算機で扱える内部表現に解釈できたとき、計算機が言葉を理解したといえる。自然言語の文が与えられたとき、その文が表す意味やその背後にある書き手（話し手）の意図を汲み取ることを「文を理解する」あるいは「文を解析する」といい、文の解釈を行うには、言葉に関する知識が必要であり、その言語知識のレベルに対応し、言語処理の過程は、形態素解析、構文解析、意味解析、文脈解析、そして、意図解析のフェイズに分けることができる。

ここでは、自然言語の持つ曖昧さについて触れ、計算機が言葉を理解する過程について説明する。

１．１　言葉の曖昧さ

自然言語は、人間が用いる言葉であるために、プログラミング言語とは異なり、解釈上の曖昧さがあるのが一般的である。普通の辞書を見ても、専門的な用語以外は、通常、単語にはいくつかの意味を持ち、異なった品詞が指示されていることも多い。一般に英語の単語では品詞の多様さ、日本語の単語では意味の多様さが目立つ。品詞の多様さは解釈可能な構文の曖昧さを増加させ、意味の多様さは意味解釈の複雑さを増加させる。

（１）単語の品詞の多様性から生じる解釈の多様性

構文の解釈の曖昧さに説明でよく使われる例文に、

Time flies like an arrow.

という文がある。通常は「光陰矢のごとし」という解釈に落ち着くであろうが、「fly」には名詞として「ハエ」という意味を持ち、「like」には「好き」という意味を持つ動詞とみなせば「時蝿（ハエの一種として）矢が好き」という解釈も成り立つ。

普通の辞書を見れば分かるように、英語では数個の品詞が指示されている単語も多く、品詞だけに着目し、意味を考慮に入れなければ、多くの構文の解釈が成り立つ。

このように単語の品詞の多様性から解釈の多様性が生じる。

（２）係り受けの候補の多様性から生じる解釈の多様性

たとえば、

長い髪の美しい少女。

という文では、「美しい長い髪をした少女」とも「長い髪をした美少女」とも解釈することが可能である。また、

Put the ball in the box on the table.

という文では、「ボールを箱に入れよ」（動詞「put」が前置詞「in」と係り受け関係を持った場合）とも、「ボールをテーブルに置け」（動詞「put」が前置詞「on」と係り受け関係を持った場合）とも解釈できる。

このように係り受け候補の多様性から生じる解釈の多様性は、特別なことではなく、日常の言語のなかにおいてもよく見られる。

１．２　言語を理解する過程

言語の解釈を行うには、言葉に関する知識が必要であり、その知識には幾つかのレベルがある。その利用する知識や情報の観点から、言語処理の過程を、形態素解析、構文解析、意味解析、文脈解析、そして意図解析の 5 つのフェイズに分けることができる。(図 1.)

（１）形態素解析

形態素そのものの明確な定義が与えられているわけではないが、一般に、語を構成する最小の意味のある単位を形態素と呼び、日本語では、名詞や形容詞、動詞語幹、活用語尾、助詞、助動詞などの語（構成単位）が形態素にあたる。

形態素解析では、辞書や語形変化規則などの語彙的な知識と隣り合った形態素間の結合に関する規則を用いて、文の分かち書きを行い、単語列を同定し、個々の単語の品詞の決定などを行う。

（２）構文解析

構文解析では、形態素解析の結果に基づき、語と語の関係を表した文法的知識や語の用法に関する知識を適用して、文の構文構造を特定する。

構文解析では、係り受けを明確にすることが要求される。たとえば「太郎とビールを飲んだ」という文の「太郎と」は「飲んだ」に係り、「ワインとビールを飲んだ」では「ワインと」は「ビール」に係ることを認識する必要がある。このように、実際の構文的曖昧性を解消するには意味的知識や文脈情報も必要する場合が多く、これらの意味的知識や文脈情報を利用する処理も含めて構文解析と呼ぶことも多い。

（３）意味解析

意味解析では、構文構造をもとに、語の意味や語と語の意味的な関係に関する知識を参照し、意味が解釈可能な意味構造を決定する。

したがって、上記の「太郎とビールを飲んだ」および「ワインとビールを飲んだ」という例では、「(動作主格：生物）が（随伴格：生物）と（対象格：飲料）を（行為：飲む)」という意味的知識を参照し、「太郎」は（随伴格：生物)、「ワイン」と「ビール」は（対象格：飲料）であることを意味的に決定しなければならない。

（４）文脈解析

文脈解析では、文中や文間での語の省略や照応（指示詞や代名詞が前述の文の何を示しているのか）に関する問題を解決し、文の関係を明らかにして文章全体の構造を決定する。

上記の「太郎とビールを飲んだ」という例で、たとえば、前後の文章から「一郎」をその（動作主格：生物）と同定し、「一郎は太郎とビールを飲んだ」という解釈に導くことをいう。

（５）意図解析

文が書かれた状況や慣習などを考慮して、文や文章の伝達しようとする意図を導く。

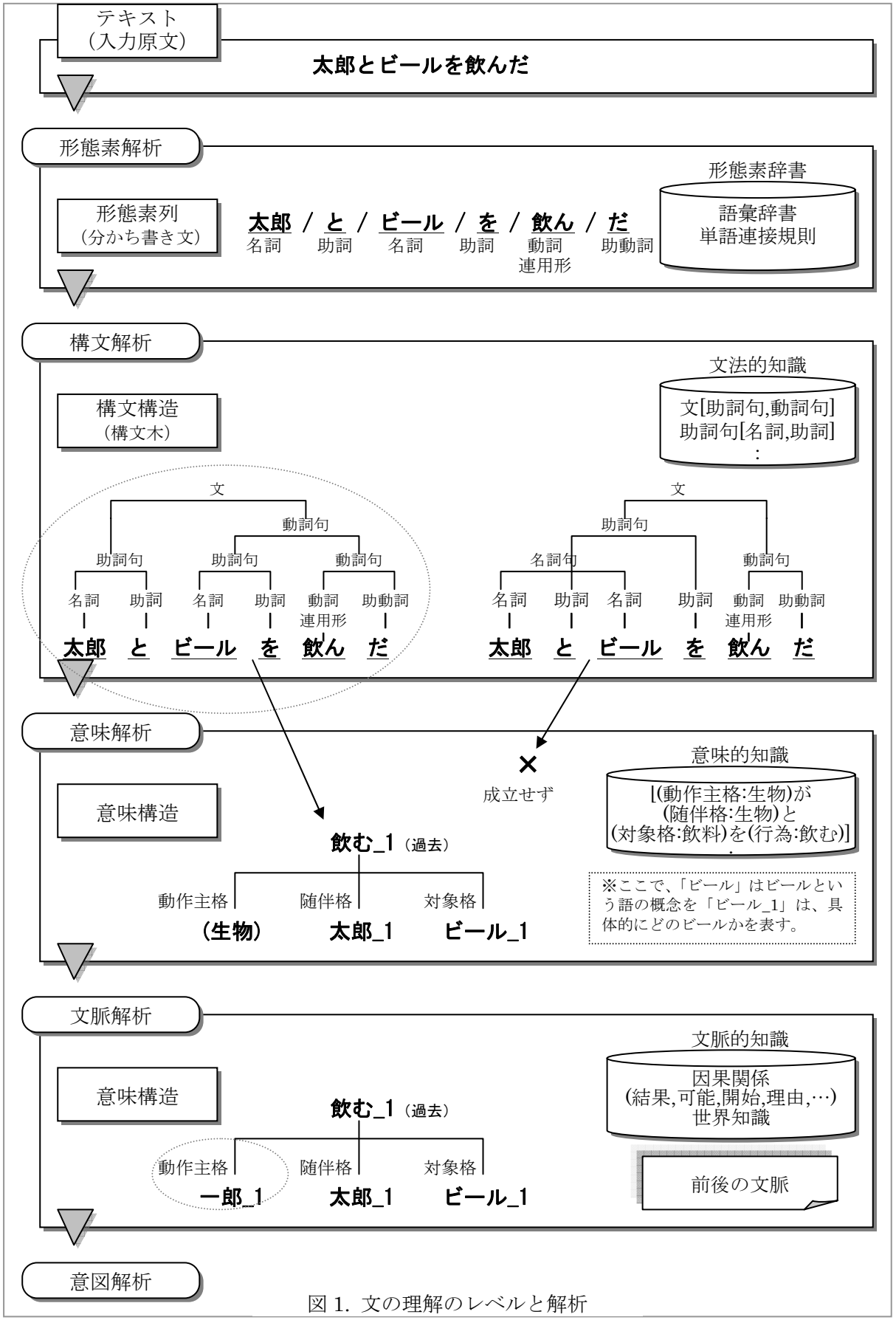


図 1. 文の理解のレベルと解析

２．形態素解析の概要

日本語は、英語やフランス語などの欧米言語のように単語と単語の間に空白を入れる「分かち書き」という習慣がなく「ベタ書き」されるのが通常である。したがって、人間が通常用いている自然言語（文）を解析するためには、まず、文を単語単位に分割する必要がある。

形態素解析では、辞書や語形変化規則などの語彙的な知識と隣り合った形態素間の結合に関する規則を用いて、文の分かち書きを行い、単語列を同定し、個々の単語の品詞の決定などを行う。日本語では、複合名詞や複合動詞などの複合語の分割も含めて分かち書きの多様さが重要な課題となるが、英語では単語の持つ品詞の多様さが重要な課題となる。

ここでは、日本語の形態素解析の役割と特徴について概観し、日本語固有の「分かち書き」処理における形態素の単位や基本的な形態素解析の方法について説明する。

表 1.に、日本語と英語の構造的特徴を示す。また、図 2.に、分かち書き処理と形態素解析結果のイメージを示す。

表 1. 日本語と英語の構造的特徴

言　語	言語の構造的特徴
日本語	膠着言語であり、単語の切れ目が記述表現だけでは分からない。 名詞は語尾変化しない。 多数の名詞がつながって複合名詞を形成することが多い。複合動詞も形成する。
英　語	単語間にブランクがあるので、個々の単語を容易に取り出せる。 名詞は単数・複数で語尾変化する。動詞も時制、数で語尾変化する。 Prefix（接頭辞）、suffix（接尾辞）が単語につく。

２．１　日本語の形態素解析

日本語の形態素解析は、通常、狭い意味では「分かち書き」処理、すなわち漢字仮名混じりで「ベタ書き」された自然言語（文）を単語に分割することをいい、文を構成する単語の表記や語形変化という形態論的性質の同定を行う。より広い意味では、読みやアクセントなどの音韻的性質、品詞などの統語論的性質、さらには語義などの意味論的性質を同定する処理までを含む場合がある。

形態素解析は、仮名漢字変換、音声合成、機械翻訳、情報検索などの自然言語処理の分野で、最も基本的で重要な役割を果たす技術である。

（１）仮名漢字変換処理と同音異義語

仮名漢字変換処理では、ひらがな（あるいはローマ字）で書かれた文を分かち書きし（単語分割）、各単語の漢字表記を選択する。このとき、単語分割の仕方によっては異なった解釈を生じる場合がある。たとえば、「きのうはいしゃにいった」という例では、

「きのう / はいしゃ / に / いった」　→「昨日歯医者に行った」

「きのう / は / いしゃ / に / いった」　→「昨日は医者に言った」

という意味が異なる単語分割が候補となる。

また、「いった」という単語には「行った」「言った」のほかにも、「入った」「逝った」な

ど、いくつかの同音異義語（homonym）が存在し、この選択により全く違った解釈になる。

（２）読み振り処理と同形異義語

仮名漢字変換が「読み」から「表記」を同定するのに対し、テキスト音声合成などでは逆に「表記」から「読み」を同定する読み振り処理が必要になる。漢字仮名混じり文を単語に分割し、各単語の読みを選択する。単語分割の仕方によって異なった解釈が生じるのは前述の仮名漢字変換処理の課題と同じであるが、さらに、「最中」を「サイチュウ」と読むか、「モナカ」と読むかといった同形異義語（homograph）の選択も解釈に影響を与える。

（３）情報検索と索引付け

情報検索では、文書の索引付け（indexing）において形態素解析技術が適用されている。最も基本的な処理としては、文書（原文）を分かち書きし、単語の品詞を決定し、助詞・助動詞などを不要語（stop word）として除去するという方法がある。たとえば、「日本語の形態素解析を行う。」という例では、

「日本語(名詞) / の(助詞) / 形態素(名詞) / 解析(名詞) / を(助詞) / 行う(動詞) / 。(記号)」という形態素解析結果に基づき、これから不用語（この例では、「の(助詞)」「を(助詞)」「行う(動詞)」「。(記号)」）を除き、「日本語(名詞)」「形態素(名詞)」「解析(名詞)」をキーワード（索引付け）の候補とする。

（４）自然言語インタフェイス

索引付けを行うための形態素解析の方式を検索文（問い合わせ文）の解析に応用することにより、自然言語インタフェイスを実現することができる。検索要求を「最近、発売された形態素解析システムにはどのようなものがありますか」のような自然な文章で入力し、それを形態素解析し、検索式を自動生成する機能を「自然言語検索」と呼ぶ。計算機や検索システムに不慣れな人にとって面倒な検索式の作成や入力の手間を省いてくれる。

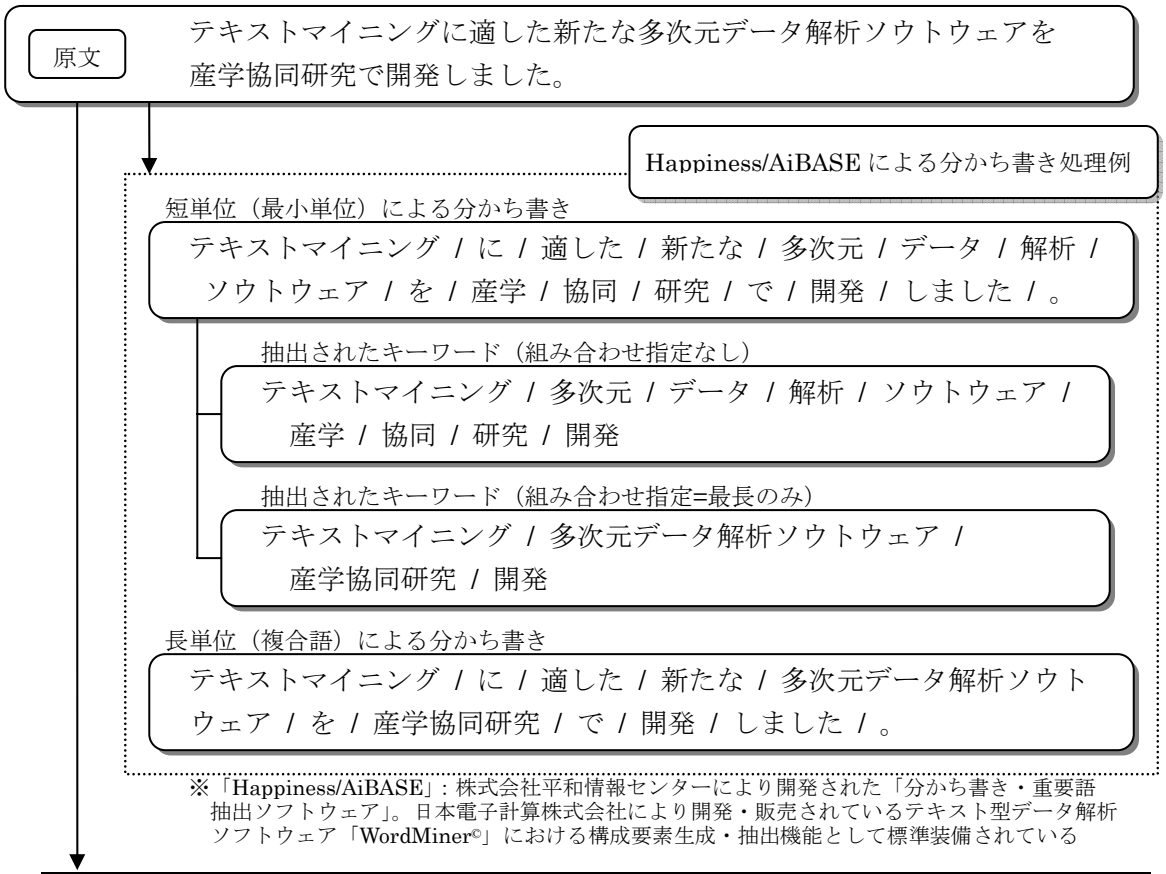
さらには、検索式の検索キーワードを単語レベルで英語などの他の言語に翻訳して、日本語で他の言語の文章を検索する「クロスリンガル検索」のような機能も実現できる。

（５）文書データベースの検索システム

文書データベースに含まれる全ての単語を次元とするベクトル空間を定義し、このベクトル空間で、登録文書はそれを特徴付ける複数の単語の重みを値とするベクトルとして表現する。ここで、検索（問い合わせ）文についても、そこに含まれるキーワード（単語）の重みを値とするベクトルとして表現すると、文書データベースのベクトル空間において、ベクトルの方向が一致するベクトルを持つものほど類似性が高いものと定義できる。これを利用して、検索結果のスコアリングや類似文書の検索、さらには、検索要求に対する評価結果をもとに、検索式をよりよくする関連性フィードバック（relevance feedback）と呼ばれる方式が実現されている。

また、検索でヒットした文書を単語に着目して類似した文書ごとにグループ分けするよう

な検索結果の自動分類や検索式を改善する作業を支援するための関連単語の提示機能（ガイダンス）なども可能である。



※「Happiness/AiBASE」：株式会社平和情報センターにより開発された「分かち書き・重要語抽出ソフトウェア」。日本電子計算株式会社により開発・販売されているテキスト型データ解析ソフトウェア「WordMiner©」における構成要素生成・抽出機能として標準装備されている

表層語	基本形	読み	品詞	活用
テキスト	テキスト	テキスト	名詞-一般	
マイニング	マイニング	マイニング	名詞-サ変接続	
に	に	ニ	助詞-格助詞-一般	
適し	適す	テキシ	動詞-自立	五段・サ行　連用形
た	た	タ	助動詞	特殊・タ　基本形
新た	新た	アラタ	名詞-形容動詞語幹	
な	だ	ナ	助動詞	特殊・ダ　体言接続
多次元	多次元	タジゲン	名詞-一般	
データ	データ	データ	名詞-一般	
解析	解析	カイセキ	名詞-サ変接続	
ソウトウェア	ソウトウェア	ソウトウェア	未知語	
を	を	ヲ	助詞-格助詞-一般	
産学	産学	サンガク	名詞-一般	
協同	協同	キョウドウ	名詞-サ変接続	
研究	研究	ケンキュウ	名詞-サ変接続	
で	で	デ	助詞-格助詞-一般	
開発	開発	カイハツ	名詞-サ変接続	
し	する	シ	動詞-自立	サ変・スル　連用形
まし	ます	マシ	助動詞	特殊・マス　連用形
た	た	タ	助動詞	特殊・タ　基本形
。	。	。	記号-句点	

茶筌(WinCha 2000 R2)による形態素解析例

※「茶筌」：奈良先端科学技術大学院大学で開発され、松本裕次氏の監督下、公式サイトで管理・配布されている日本語形態素解析システム　<http://chasen.aist-nara.ac.jp/>

図 2. 分かち書き処理と形態素解析結果のイメージ

２．２　形態素の単位

先に述べたように、形態素の定義は明確に与えられているわけではないが、一般に、語を構成する最小の意味のある単位を形態素と呼び、日本語では、名詞や形容詞、動詞語幹、活用語尾、助詞、助動詞などの語（構成単位）が形態素にあたる。

日本語では、一般的に「自立語」と「付属語」を形態素の単位とすることが多い。自立語は名詞、副詞や動詞語幹、形容詞語幹などをいう。付属語は、動詞や形容詞の活用語尾、助詞や助動詞など、それ自体では意味を持たず、自立語と組み合わさって初めて意味を持つ語である。(図 3.)

一方、非常に小さな単位の形態素に分割してしまうことは、自然言語解析システムにおいて逆に負担を生じることもあり、何を形態素の単位とするかは、形態素解析の結果をどのように活用するのかといった目的によって様々な工夫が見られる。一般に、自然言語解析システムの目的によって、形態素解析処理の効率、形態素辞書の容量や管理、形態素解析に引き続く構文解析や意味解析の戦略などの設計指針に基づき、形態素の単位が決定される。

たとえば、「～について」の形態素は「に（格助詞)」、「つい（動詞語幹)」、「て（動詞語尾連用形)」と分割することができるが、実際には「について」をひとつの形態素（形態素列）として扱うほうが好ましい場合もある。

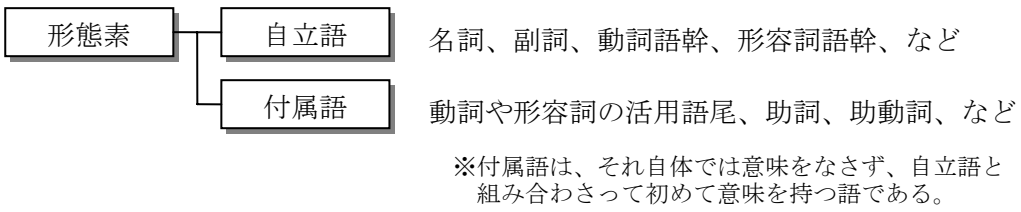


図 3. 形態素の単位

２．３　形態素解析の方式

形態素解析の主な処理は、形態素間の接続規則に基づく処理であり、原文を形態素接続規則と形態素解析辞書を用いて形態素に分割し、単語を発見するとともに、その構文上の素性を決定する。形態素解析を行うための重要な処理は、「分かち書き」「単語の品詞の同定」「辞書にない語（未知語）の処理」の３つである。

形態素に分割する際、その切り出し（単語分割）には多様な組み合わせがあり、さらに切り出された単語には品詞上多様な候補が存在する。したがって、形態素解析では、こうした多様な候補の中から、形態素に関する情報を用いて、解釈可能な候補に絞り込むことが目標となる。形態素解析システムでは、こうした形態素解析での解釈の曖昧さを解消するために、経験的優先規則や文法的接続可能性、接続コスト、統計的言語モデルなどの方法が採り入れられている。

２．４　形態素解析接続規則と形態素辞書

日本語では、自立語と付属語の間には組み合わせの規則がある。たとえば、

・五段活用の動詞には、その活用での活用語尾が結合する。

（カ行五段動詞語幹「聞」に対するカ行五段活用語尾「か,き,い,く,け,こ」）

・格助詞は名詞の後に置かれ、動詞の直後には置かれない。

というような規則を形態素間に与えることができる。

形態素解析は、こういった形態素接続規則と単語辞書（形態素辞書）を用いて、与えられた原文を形態素に分割し、単語列を同定し、個々の単語の品詞の決定などを行う。

単語辞書（形態素辞書）は、単語（形態素）の表記、読み、品詞などを記述した辞書であり、実用的なシステムでは 10 万語から 10 数万語を収録している。

形態素接続規則は品詞接続表とも言われ、隣接する単語の接続可能性を示すもので、たとえば、右接続情報を行とし、左接続情報を列とする行列（マトリックス）の形式で表現することができる。実用的なシステムでは、普通名詞、固有名詞、サ変名詞、代名詞、あるいは、五段動詞、一般動詞、さらには格助詞、接続助詞など、きめ細かい品詞区分を設けるとともに、動詞や助動詞などの活用語は未然形、連体形、終止形などの各活用形を一つの品詞区分とみなすので、品詞区分の数は 100 個程度になる。したがって、形態素接続規則（品詞接続表）はおよそ 100 行×100 列の行列となる。（図 4.）

形態素接続規則(品詞接続表)の例　※実用システムでは、およそ 100 行×100 列

品詞区分	名詞	助詞	形容詞	副詞	助動詞	動詞	語尾
名詞	1	1	0	0	1	0	0
助詞	1	0	1	1	0	1	0
形容詞	1	0	0	0	1	0	0
副詞	1	0	0	0	0	1	0
助動詞	1	0	0	0	1	1	0
動詞	1	0	0	0	1	1	1
語尾	1	1	1	1	1	1	1

表記	読み	品詞	左接続	右接続
形態素	ケイタイソ	名詞	名詞	名詞
解析	カイセキ	名詞	名詞	名詞
について	ニツイテ	(注 1)	助詞	語尾
話す	ハナス	動詞	動詞	語尾

ここでは、より柔軟な接続関係を表現するために、左接続と右接続と左右両方からの接続情報を辞書に記述している。

(注 1)
ここでは「～について」を「に(格助詞)」、「つい(動詞語幹)」「て(動詞語尾連体形)」と分解せず、「について」を一つの形態素として扱っている。

形態素接辞書(単語辞書)の例　※実用システムでは、10 数万語を収録

図 4. 形態素接続規則(品詞接続表)と形態素辞書(単語辞書)の例

２．５　単語分割の多義性とその解消方法

たとえば、「日本海運輸出産業界」という複合名詞を分割することを考える。この場合、

「日本 / 海運 / 輸出 / 産業界」

と分割するのが妥当と思われるが、「日本海」「運輸」「出産」「業界」なども語を形成する（形態素辞書に登録されている）ので、

「日本海 / 運輸 / 出産 / 業界」

をはじめ、多くの単語分割の候補が存在する。（図 5.）

このようにある原文について複数の単語分割の可能性が存在する場合、文は単語分割に関する多義性または曖昧性を持つという。

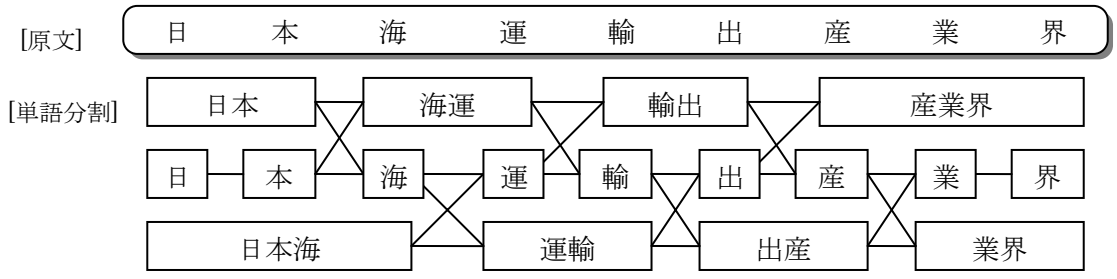


図 5.　「日本海運輸出産業界」単語分割の多義性

日本語の形態素解析では、多くの単語分割の候補の中から、最もふさわしいと思われる単語分割パターンを選択する方法として、経験的優先規則に基づく方法、文法的接続可能性に基づく方法、接続コストに基づく方法、統計的言語モデルに基づく方法などが採用されている。

（１）経験的優先規則に基づく方法

日本語の単語分割では「最長一致法」及び「分割数最小法」が有効な優先規則として知られており、多くの言語解析システムで採用されている。しかし、もともと言語の特徴にその根拠を置いていないために一長一短があり、経験的にその有効性が評価されている。

最長一致法は、文字列の先頭から解析をはじめ、後続する可能性がある単語が複数あるときは、最も長い単語を選択する方法である。最長一致法は簡便なことから多くの言語解析システムで用いられているが、分割最小法に比べて精度は劣ると言われている。

一方、分割数最小法は、原文から単語を切り出す際に、原文を構成する単語の総数が最も少ない候補を選択する方法である。一般に、分割数最小法は最長一致法に比べると制度が高く、日本語の仮名漢字変換では分割最小法が語の適切な分割を行えると言われているが、最長一致法に比べると、総当り探査（組み合わせ的爆発が生じる）となり、多くの記憶領域を必要とする。

たとえば、先の「日本海運輸出産業界」（図 5.）の例において、最長一致法では、「日本海－運輸－出産－業界」が優先され、分割数最小法では、「日本－海運－輸出－産業界」と「日本海－運輸－出産－業界」がどちらも分割数 4 単語として優先される。この例のように最長一致法による単純に長い単語を優先することには問題があり、その根拠も明確ではない。また、分割数最小法では、分割数が同じ候補に対してどの候補を優先させるかという問題が生じる。

（2）文法的接続可能性に基づく方法

最長一致法や分割数最小法では、日本語の解釈としては不適切な候補が優先される事象が多数発生する。そこで、文の単語分割の妥当性を評価するために単語間の分法的接続可能性を検査することが有効となる。

既に、2.4 項「形態素解析接続規則と形態素辞書」で説明したように、形態素解析では、形態素接続規則（品詞接続表）と単語辞書（形態素辞書）を用いて、与えられた原文を形態素に分割し、単語列を同定し、個々の単語の品詞の決定などを行う。

実用的な形態素解析システムでは、文法的接続可能性と前述の経験的優先規則を組み合わせて単語分割の精度を向上させている。

（3）接続コストに基づく方法

基本的に、文法的（品詞）接続可能性は、2.4 項、図 4.「形態素接続規則(品詞接続表)と形態素辞書(単語辞書)の例」で示したように、品詞の接続可能性を可能「1」または不可能「0」という二つの値（2 値表現）で表現したものであるが、実際に使われている言語表現に対し、その可能性をすべて表現することは難しい。

そこで、接続コストに基づく方法では、接続する可能性を可能か不可能かの 2 値ではなく、接続する可能性が高いものほど小さな値をとり、接続する可能性が低いものほど大きな値をとる「コスト」（通常は、0 から 1 の範囲）で表現する。このとき、品詞接続コストだけではなく、単語そのものの出現頻度の差に着目して、出現頻度が大きいものほど小さな「コスト」を与えるといった単語コストをあわせて採り入れる場合も多い。

（4）統計的言語モデルに基づく方法

統計的言語モデルに基づく方法とは、先の接続コストに基づく方法の「コスト」について、情報理論と確率論に基づく理論的な根拠を備え、対象領域のテキストからモデルのパラメタを学習する方法を与えるものである。代表的なモデルには「品詞二つ組モデル」や「隠れマルコフモデル」などがある。

２．６　形態素解析システムの構成

最長一致法により分かち書き候補抽出（単語分割）を行う形態素解析システムの基本構成を図 6.に示す。

① 形態素辞書の検索は、ポインタで指示されている位置からはじまる文字列をキーとして、最長一致法に従って検索する。形態素辞書に格納されている形態素（単語エントリ）の中でキーの文字列に最も長く一致する単語エントリを発見する。この単語エントリが分かち書き（単語分割）の候補となる。

② 次に、単語分割の候補とした単語エントリの文字列長文だけポインタを進め、次に検索する文字列の先頭位置を指示するようにし、①と同様に、次の分かち書きの候補を得る。

③ そして、形態素接続表を用いて、直前に得た形態素（単語エントリ）の素性と最新の形態素（単語エントリ）の素性による接続テストを行う。

④ 接続テストの結果が接続可能であれば、正しく切り出せたものとし、ポインタをすすめ、同様の処理を対象原文の末尾に至る（テキストバッファが空になる）まで繰り返す。

⑤ 接続テストの結果が接続不可であれば、最新の形態素（単語エントリ）の末尾の文字を捨てて、同様の処理（分かち書き候補の獲得、接続テスト）を行う。このとき、接続不可の状態が続くと形態素辞書を検索する文字列の長さが 0（空）になってしまうことがある。この場合は、直前に得た形態素の候補が誤って切り出された可能性があるので、この形態素の末尾の文字を捨てて、あらためて形態素辞書を検索し、分かち書き（単語分割）候補を得る。

このような処理で、形態素の切り出し（単語分割）と形態素の素性（品詞）の決定が行われる。なお、ここでは触れなかったが、接続テストの失敗は、未知語の存在（辞書に存在しない単語）が考えられるので、未知語の発見と対応が必要になる。

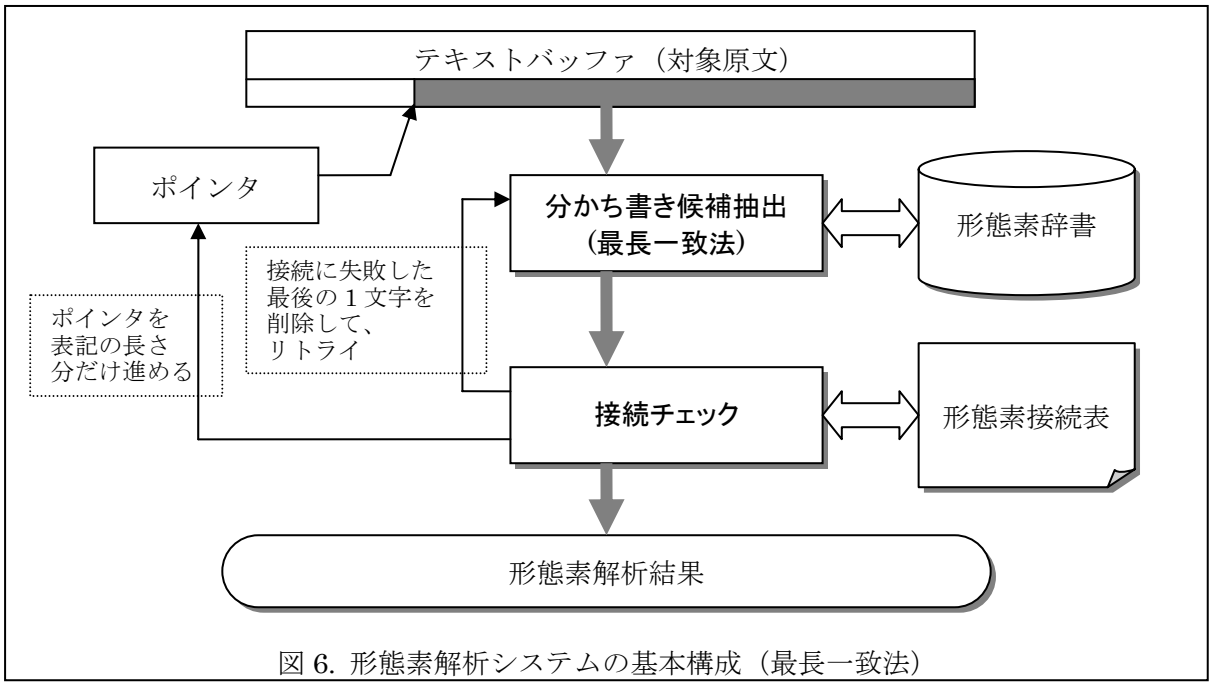


図 6. 形態素解析システムの基本構成（最長一致法）

３．テキストマイニングにおける形態素解析

テキストマイニングとは、膨大なテキスト型データを様々な観点から分析し、役に立つ知識や情報を見つけ出す技術である。一方、実務的な面からみれば、CRM（Customer Relationship Management）や KM（Knowledge Management）などと同様に、単なる技術ではなく、顧客の信頼を獲得し、維持し続けるためのマネージメントの一つとして位置づけることもできる。

役に立つ知識や情報を見つけ出すという点では、データマイニングも同じ目的である。ただし、データマイングで扱うデータはデータベースのスキーマ（項目）等で構造化・形式化されたデータ（例えば売上情報や生産情報）を対象としている。それに対しテキストマイニングは構造化・形式化されていないテキスト型データからのマイニング（知識や情報を見つけ出す）を目的としている。そのような意味でテキストマイニングはテキスト・データマイニングと呼ばれることも多い。

身近なテキスト型データとしては、コールセンターやお客さま相談室などに寄せられたクレームや要望・意見等、あるいはアンケート調査の自由回答・自由記述文などがある。また、新聞記事や学術論文・特許公報等の文献情報もテキスト型データの代表である。

３．１　テキストマイニングの機能

マイニングの本質はデータからの知識や情報の「発見」である。既にデータ内の関係について推測（仮説）を持ち、データや情報を検索・評価して、仮説を検証するためのデータベース（情報）検索や統計手法を用いた検定などは二次的なものである。

また、テキスト型データベースやインターネットのサーチエンジンなどの概念検索やあいまい検索と呼ばれる文書の検索手法、あるいは、検索結果のスコアリング（一致度、適合度等）や要約文の表示方法などもテキストマイニングの関連技術ではあるが、知識の発見という点ではマイニングの本質とは異なる。

マイニングの主たる目的は、大規模なデータの集まりから、自明ではなく価値のある新事実や関係を自動的に発見することである。また、マイニングでは、取り扱うデータが集計値や編集加工されたものではなく、個々の生データに着目して分析を行うことに意義がある。

テキストマイニングは、人間が日常的に自然に使っている言語（文字言語・音声言語）を理解する自然言語処理技術、情報検索やデータベース技術、主成分分析、判別分析、因子分析、数量化法など多変量解析の手法など、様々な技術を組み合わせた複合的な技術であり、その手法も様々なものがある。

一方、テキストマイニングの処理の流れは、一般的にテキスト型データから情報や概念を抽出するステップと抽出された情報や概念を解析する（マイニング）２つのステップに分けることができる。この２つに、データの基本情報(背景)や解析過程・結果を分かり易く提示するための情報の分析・可視化を加えた３つの機能がテキストマイニングの主要な要素となる。（図 7.）

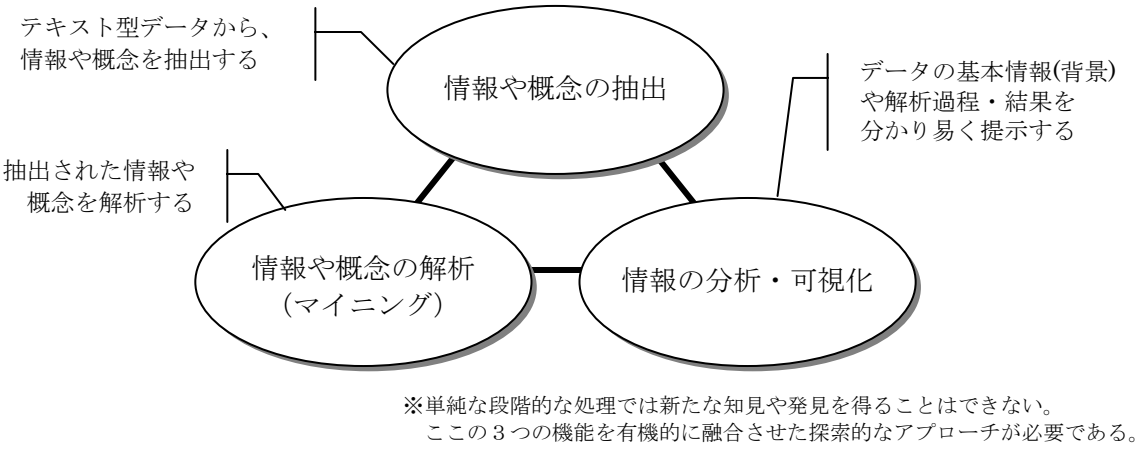


図 7. テキストマイニングの主要な構成要素

この数年，テキストマイニング関連技術の実用化が急速に進み、適用事例も数多く報告されるようになった。解説図書や学会・セミナーなどでも取り上げる機会が増えてきており、数多くのテキストマイニングツールが市販されている。

テキストマイニングツールは、自然言語処理による情報や概念の抽出技術、データマイニングや多変量解析のマイニングアルゴリズム、解析結果の可視化技術など、各々特徴的な得意とする解析手法を持っている。また、社会調査や市場調査における自由記述型回答に代表される定性的情報の解析、コールセンターやお客さま相談室に寄せられた顧客の声の解析、新聞記事や特許広報・技術論文などの文献情報の解析など、解析目的や対象とするデータ、あるいは業務プロセスなどにより、テキストマイニングの適用方法は異なる。

３．２　情報・概念の抽出

テキストマイニングにおける情報や概念の抽出ステップでは、分析対象のテキスト型データを形態素解析や構文解析などを用いて、その内容をあらわす情報や概念を抽出する。構造化・形式化されていないテキスト型データを次の解析ステップ（マイニング）で扱えるようにするための数値変換処理であり、テキストマイニングの特徴的（象徴的）な機能である。

単に類似文書（テキスト）の検索や分類を行うのであれば、形態素解析による単語分割を行わずとも、N-gram 法（N 文字が隣接して生じる文字の共起関係、2 文字隣接のときは 2-gram という）や字種切り法（句読点などの区切り符号、漢字、カタカナ、英字、数字、平仮名など、字種の切れ目を利用する）などにより切り出された文字列を活用することでも実現できる。たとえば、「TM における情報や概念の抽出ステップ」という例では、2-gram 法では、

「TM」「 M に」「にお」「おけ」「ける」「る情」「情報」「報や」「や概」「概念」
「念の」「の抽」「抽出」「出ス」「ステ」「テッ」「ップ」

が切り出される。字種切り法では、

「TM」「における」「情報」「や」「概念」「の」「抽出」「ステップ」

と切り出される。

マイニングの本質はデータからの知識や情報の「発見」であり、膨大な生データに着目して、自明ではなく価値のある新事実や関係を自動的に発見することに意義がある。

役に立つ知識や情報を見つけ出すためには、単なる単語分割情報（形態素やキーワードの抽出、品詞の同定など）だけでは不十分であり、単語の同義性や多義性を考慮に入れた概念や情報の抽出が必要になる。

たとえば、「米」「米国」「合衆国」などを「アメリカ」という一定の表現に置換することで同義性を吸収したり、「米－食物」または「米－国名」というように意味属性を加えて「米」という語（文字）の持つ多義性を解消したりする。

また、「性能と価格は良いが、デザインが悪い」といった文章の場合、何が「良い」のか、何が「悪い」のか、語句間の関係を表す係り受け情報や、文節や文章間の複合概念まで抽出することも必要になる。さらに、概念や情報を端的に表現するという意味では、文書要約や自動タイトル付けなどの技術を活用する場合もある。

３．３　情報や概念の解析（マイニング）

抽出された情報や概念に基づき、今までに知られていない新しい事実や知識を得るためのマイニングを行う。

相関分析、クラスタリング、クラス分類、時系列分析など、解析の目的とデータの質や量、抽出した情報や概念の特徴に合わせて、データマイニング・アルゴリズムや多変量解析手法、あるいは統計的手法を適用する。

ここでクラスタリングとは類似のパターンを教師なし学習アルゴリズムによりグループ化すること(clustering)をいい、クラス分類とは教師付きアルゴリズムにより入力パターンを識別する処理(判別；discrimination)をいう。例えば、前者は文書群を情報や概念の類似性により仕分けるような場合のことであり、後者は既存の分類体系に振り分けるような場合に相当する。いずれもテキストマイニングでは欠かせない機能である。

以降に、文書（テキスト）に形態素解析等を行い、その結果得られた単語（キーワード）及び単語（形態素）情報などから、単語の関連性やテキストの類型化を評価する簡単な例を示す。

（１）親近性尺度による言葉の類似性

文書（テキスト）を一つの概念として捉えたとき、一つの文書に含まれる単語（キーワード）対は、その概念上において何らかのつながりを持つ。このとき、より多くの文書に出現するキーワード対ほど、概念説明上の親近性が大きいものとみなすことができる。したがって、キーワードの出現頻度とキーワード対の同時出現（共起出現）頻度により、キーワードの親近性をキーワード間の類似度として定義できる。(図 8.)

（２）近接的共起発生関係

実際の文章中には文脈や話題の展開に従い様々な概念が多岐に表現されており、キーワード間の類似度を計算する際には注意が必要となる。とくに長文の文書の場合、キーワードの

組み合わせの数が莫大になり、概念説明上において親近性のないキーワード対が相対的に多く抽出され、キーワード間の近さの定量化精度が低下する。

そこで、文書全体のキーワード対ではなく、段落や個々の文章、あるいは文節などを一つの概念とみなして、キーワード間の類似性を定義するといった工夫が必要となる。

キーワード対としての共起出現の判定基準として、キーワード間の単語数を物理的な近さとする近接的共起発生に基づくものがある。(図 9.)

（3）出現パターンの類型化によるまとまり

たとえば、文書（テキスト）と単語の出現頻度の行列、あるいは単語と単語の共起度（同時出現）の行列を用意する。これを、出現パターンが似たものが近くにくるように行と列の入れ替えを繰り返すと、文書ならびに単語の概念が近いものほど近くにならぶとみなすことができる。このようにした得た行列は、相関関係や行列の固有値解法・スペクトル分解などにより、その並び替えによる変化を評価することができる。図 10. に、出現パターンの類型化によるまとまりについての概念を示す。

（4）類似度によるまとまり

たとえば、文書（テキスト）の類似度や単語の共起出現頻度により、文書間あるいは単語間の近さを定義する。ここで類似度の近いものでペアをつくる操作を繰り返すと、文書あるいは単語の概念の近いものが近くにならぶと考えることができる。この場合、個々の類似度の関係はツリーとして表現できるので、ツリーの切断位置によりグループ化する数を操作できる。図 11. に、類似度の近さの尺度によるまとまりについての概念を示す。

$$L(x,y) = \{ N(x) + N(y) - N(x,y) \} / N(x,y)$$

L(x,y) : キーワード x と y の類似度（ただし、N(x,y)=0 のとき L(x,y)=0
N(x) : キーワード x のテキスト出現頻度
N(x,y) : キーワード x と y のテキスト同時出現頻度

図 8. 親近性尺度による言葉の近さ

$$I(x,y) = F[P(x|y) / P(x)]$$

I(x,y) : 近接的共起発生，1 : 関連あり、0 : 関連なし
F[x] : x がある閾値(たとえば 0.6)を超えるとき 1，超えないとき 0 とする
P(x|y) : y の近傍(たとえば、後続する 5 単語以内)における x の出現頻度
P(x) : x の全出力頻度

図 9. 近接的共起発生関係の例

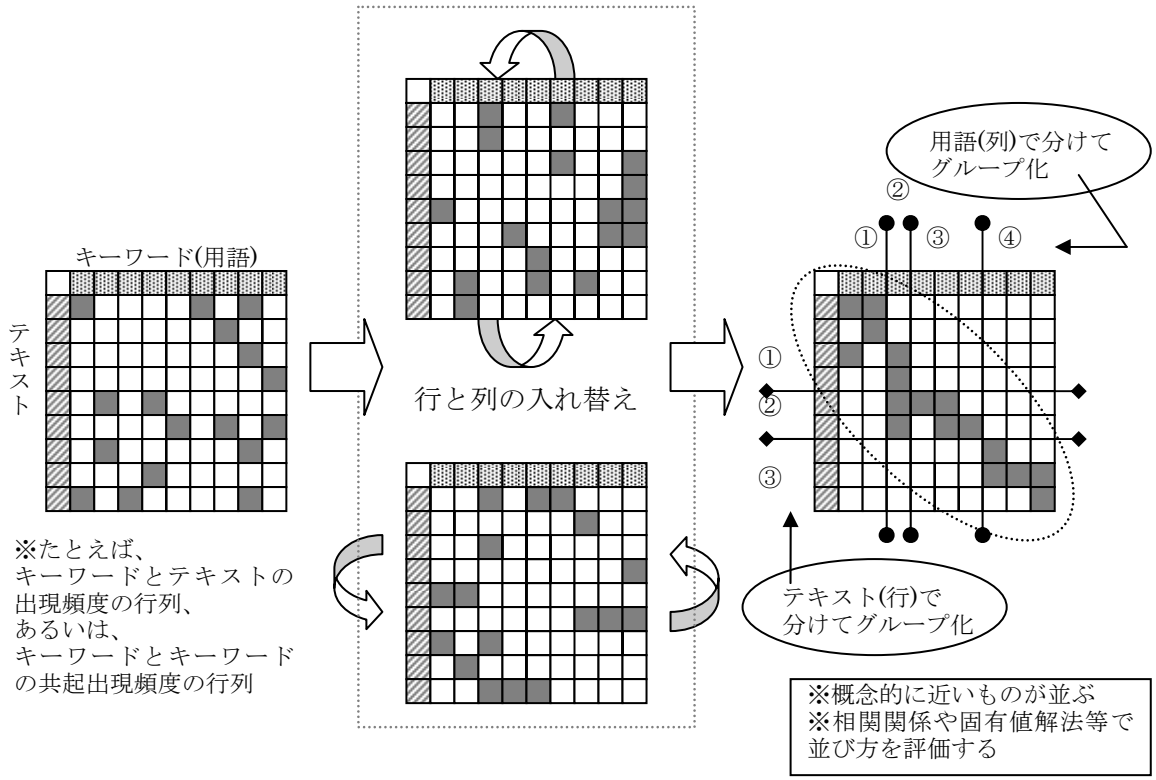


図 10. 出現パターンの類型化によるまとまり

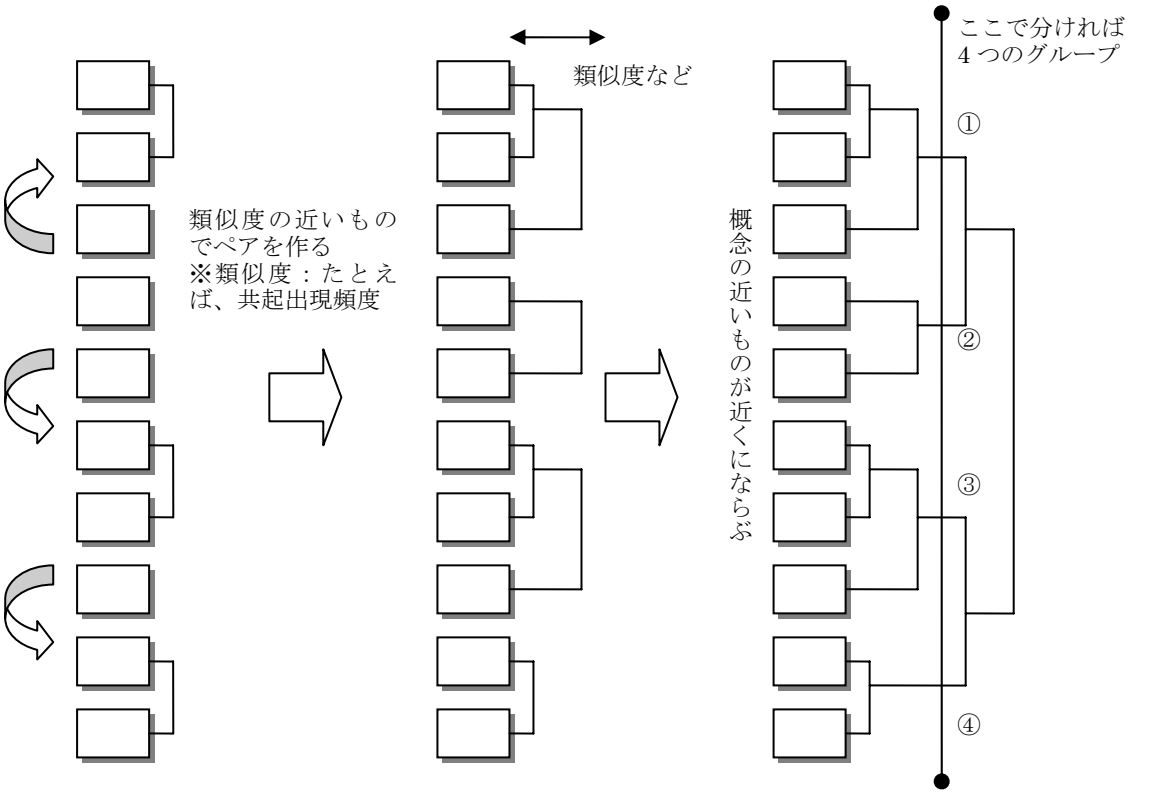


図 11. 類似度によるグルーピングの過程

３．４　テキスト（文書）の曖昧さ

テキスト型データから概念や情報を抽出する際においては、単なる単語分割情報（形態素やキーワードの抽出、品詞の同定など）だけでは不十分であり、日本語の持つ曖昧さや単語の同義性や多義性を考慮に入れなければならないことは前述したとおりである。

元々、テキストマイニングが取り扱う自然言語で記述されたテキスト（文書）は、データや情報が曖昧かつ多様な表現、多種の文字（漢字、英字、ひらがな、カタカナ、記号など）で記述されている。

テキストマイニングの狙いが「大量の」かつ「生のデータ」を対象にすることを考えれば、ワードプロセッサや手書き文字入力装置（OCR）、あるいは音声入力装置などにより作成・収集されたテキスト型データを直接的に取り扱う場面がますます増え、形態素解析や構文解析以前に、電子化される過程で発生する文書の誤りを処置する手立てが重要になる。（表 2.）

表 2. 電子化された文書の誤りパターン

誤りのパターン		説　明
元々の原文や原稿の誤り (人により作りこまれた誤り、入力装置・操作は基本的に正常)	入力されたもともとの原稿に誤りがある	単なるミスから、「はじめから手ほどきをする」といった重ね言葉や敬語の間違い、名詞と動詞の対応がとれていない場合など
	文書内の統制がとれていない(同義語の扱いが不統一)	英語・日本語の表記、長音・中点の有無、送り仮名の差異など。「AI・人工知能」、「ガラス・硝子」、「打ち合わせ・打合わせ・打合せ」など。
	本来､漢字にすべきところの字体が異なっている	本来、原文や原稿の問題であるが、仮名漢字変換時にもよく発生する。「せいりつする」、「セイリツする」、「成立スル」など（本来は「成立する」）。
	文書の用途からすると誤り	忌み言葉や放送禁止用語など。「するめ」「なし」、あるいは「片手おち」「めくら」など。
	揺れる表記	「才」と「歳」など
誤変換・誤認識による誤り (入力装置の特性・機構による誤り、元々の原文・原稿は正しい)	漢字において好ましくない選択	「最大限」に対しては「最少限」ではなく「最小限」が好ましい。「拡大」には「縮少」ではなく「縮小」が好ましい。 「的確」と「適確」(法令用語) 「年令」と「年齢」(公文書)
	[誤字]同音異義語・複合語の合成誤り	基本的に、文や熟語単位でないと正誤が決められない。ワープロの場合は、意味的に極端に誤る場合が多々発生する。「公開・後悔・更改・航海・紅海」など。四字熟語の場合、文脈によらず判断できるものもある。「人口知能」「絶体絶命」など
	]誤字]音・形・意味的に類似した字に誤る	[音]「写象(写像)」「卒直(率直)」など [形]「犬猫・大猫」、「働く・動く」など [意味的]「川・河」、「町・街」、「木・樹」、「目・眼」、「十分・充分」など
	[脱字]送り仮名、助詞や接頭・接尾辞の抜け、英語のスペリングミス	ワープロの入力ミス。OCR の読み飛ばしなど。
	[蛇足字]文字のダブり、英語のスペリングミス	ワープロの入力ミス。OCR の読み飛ばしなど。OCR の場合にはゴミや汚れの認識もある。
	認識不能	字が汚い、原稿の汚れ、複雑すぎる字（「軋轢」「躊躇」）など。

また、同義語をはじめ、類義語（言葉の形は異なるが意味的にはほぼ重なりあう語、たとえば「機械翻訳」と「自動翻訳」）や関連語・関係語（意味は異なるが関連性が強い語、たとえば「自然言語処理」と「形態素解析」）など、ひとつの用語（文字列）の表現をとっても多種多様である。表 3.に同義語のパターンを示す。

とくに E-mail や携帯電話などの世界では独特な記号や造語を容認している。例えば、文中の「(笑)」とか、記号を利用して表情や感情を表現するスマイリー（たとえば、大喜びをあらわす縦型「(^-^)」横型「8D」）、インターネットのニュースグループやメーリングリストではよく使われる発音のもじり（たとえば「F2F」face to face、「UC」you see?、「IC」I see）、頭文字をつなげたアクロニム（たとえば「FYI」For Your Information、「NRN」no reply necessary）、あるいは携帯電話のメールで多用される絵文字など、きりがない。

今後、インターネットによる Web マーケティングなどの活用の広まりとともに、避けては通れない問題となる。

表 3. 同義語のパターン

同義のパターン	説　明
カタカナ表記	[長音の有無] キャラクター=キャラクタ、ショー=ショウ [中点(ナカグロ)有無] テキストマイニング=テキスト・マイニング [音の表記の異なり] ベトナム=ヴェトナム、ロマンティック=ロマンチック アカゲザル=アカゲサル
英語表記	[大文字・小文字] Internet=internet [ルビ・読み] BEST=ベスト、DB=データベース、PC=パソコン [訳語] Dictionally=辞書
異字体表記	[外来語表記] 煙草=タバコ、麦酒=ビール [旧仮名] ゐ=い、ゑ=え [旧漢字] 醫=医、國=国、躰=体 [カタカナ・ひらがな] ネコ=ねこ=猫 [ニュアンス] 身体=体、価格=値段、賞与=ボーナス
略語（略号）	[一字漢字・英語・特殊記号・フェイスマーク] 米=アメリカ、日=日本 TV=テレビ、(株)=株式会社、S=イオウ、〒=郵便番号 (^-^)=8D =大喜び、(^0^)=:-D=大笑い
略称	アマ=アマチュア、バイト=アルバイト、JIS=日本工業規格
表記のゆれ	[送り仮名] 打ち合わせ=打合わせ=打合せ、引越=引越し=引っ越し [接頭辞] お金=金、御見舞い=お見舞い=見舞い
数字表記	二種=2 種、2=Ⅱ=②
その他	[意味的]アルコール=酒、癌=悪性腫瘍 [種類]酒=(日本酒=ワイン=ウィスキー=焼酎=ビール) [品詞]取り付け=取り付ける

略語：口語において、片方があまり使用されないもの。表記上、簡易にするために使用される。
略称：口語において使用される。語句が長い場合、語句を整えるために使用される。
スマイリー（smiley）：海外では一般に横倒しにしたものがよく使われている。

３．５　解析に用いる単語

テキストマイニングの狙いや目的、あるいは、その収集方法などにより、対象とするテキスト型データの質および量は様々であるが(表 4.)、一般に、テキストマイニングの処理では、テキスト型データから情報や概念を抽出し、抽出された情報や概念を解析する（マイニング）過程において、解析の狙いや目的に応じて、あるいは解析の操作性や解釈の容易性を確保するために、解析に用いる単語を絞り込む（単語の抽出・削除、あるいは出現頻度による抽出）ことや、同義語や類義語などを編集する（同値置換）ことが必要になる。

表 4. 分かち書き・キーワード抽出結果例

項 目		1) 読売新聞記事「小泉首相」2001 年検索結果	2)文藝春秋「幸福論」書下ろし作品抜粋	3) 放送提供業コールセンター/お客様の声	4)DVD 商品/価格アンケート	5) スキンケア商品/Web マーケティング
サンプル数		4,619	40	22,704	1,204	600
総文字数		3,111,801 (673.7)	98,826 (2,470.7)	2,271,856 (100.1)	31,826 (26.4)	21,625 (36.0)
分かち書き (最長分割)	総単語数	1,469,977 (318.2)	51,932 (1,298.3)	1,126,942 (49.6)	15,849 (13.2)	10,478 (17.5)
	異なり数	96,725	8,497	27,174	1,758	1,906
	異なり率	0.066	0.164	0.024	0.111	0.088
キーワード (最長語)	総単語数	368,655 (79.8)	8,276 (206.9)	271,962 (12.0)	3,390 (2.8)	2,591 (4.2)
	異なり数	74.443	5,073	16,539	842	985
	異なり率	0.202	0.613	0.061	0.248	0.380
分かち書き (最短分割)	総単語数	1,691,804 (366.3)	53,241 (1,311.0)	1,075,007 (47.3)	15,605 (13.0)	10,759 (17.9)
	異なり数	54,533	8,439	33,398	1,806	1,881
	異なり率	0.032	0.159	0.031	0.116	0.175
キーワード (最短語)	総単語数	443,212 (96.0)	8,730 (218.3)	241,568 (10.6)	3,205 (2.7)	2,811 (4.7)
	異なり数	36,576	5,011	21,921	892	964
	異なり率	0.083	0.574	0.091	0.278	0.343

1) 読売新聞記事：2001 年(1 月~12 月)の記事 DB をキーワード「小泉首相」で検索し、その検索結果として得た 4,619 件について、記事タイトルと本文を併合し、サンプルとした。
2) 文藝春秋：平成 13 年 9 月臨時増刊号「新幸福論－ほんとうの幸せとは」(123 人の「自分らしく生きるヒント」の全篇書下ろし）に掲載された執筆文から取り出した 40 名の執筆文をテキスト化したデータで 1 著作(作品)を 1 サンプルとした。
3) コールセンター（放送提供業）：平成 13 年及び平成 14 年の 1 月~6 月に寄せられたお客様の声。(問い合わせ、苦情、要望など)
4) DVD 商品価格アンケート：DVD 商品の価格についての消費者アンケートの自由記述回答。
5) スキンケア商品 Web マーケティング：お気に入りのスキンケア商品の理由を Web サイトにて収集したもの。
※ここに上げたサンプル(データ)は、とくに何らかの区分(分野・対象、タイプなど)を代表しているものではない。
※分かち書き処理、およびキーワードの抽出は、WordMiner の Happiness/AiBASE を利用した。最長とは複合名詞を連結し 1 つの単語として扱い、最短とは複合名詞をそれを構成する複数の形態素)に分割して扱う。キーワードは分かち書き結果から不用語を削除して得られたもの。
[異なり率=異なり数 / 総単語数]

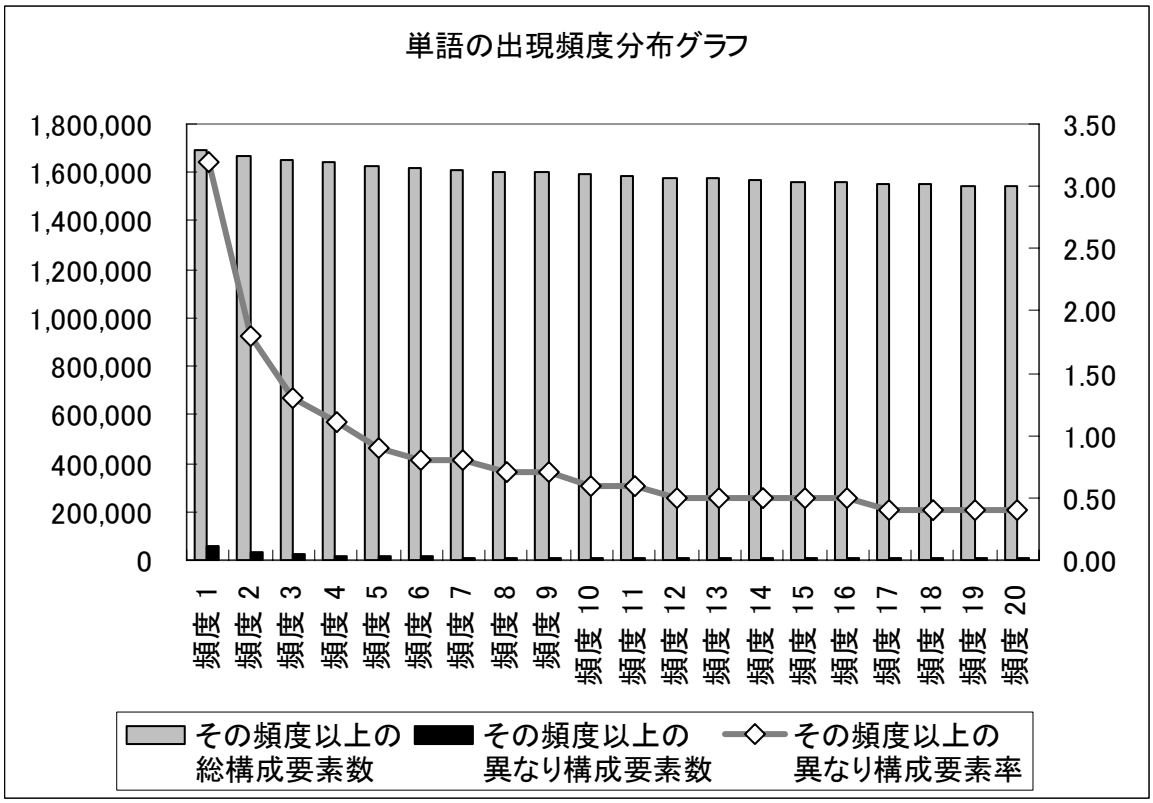
（1）閾値による単語の抽出

ここで閾値（いきち）による単語の抽出とは、閾値で指示する出現頻度以上の単語に絞り込むことをいう。たとえば、閾値 3 による単語の抽出では、各単語の出現頻度が 3 以上の単語のみ抽出される。すなわち、閾値未満の出現頻度である単語は解析から除外される。

たとえば、図 12.の新聞記事の例では、閾値 20 により単語を絞り込むと、解析対象となる単語は、総構成要素数 1,542,966、異なり構成要素数 6,673、異なり構成要素率 0.4 となる。ここでいう「構成要素」とはデータ解析上の処理単位を表し、一般に言う単語や語句、文節等より緩やかな意味として位置づけている。また、異なり構成要素数とは総構成要素の中の相異なる構成要素の数であり、異なり構成要素率はその割合である。

図 12.の例は、その対象とするデータ量がやや多い例であるが、閾値を 100 としても取り扱う総構成要素数は 1,353,118、異なり構成要素数は 1,684 となる。

ここで、ある閾値（たとえば出現頻度 100）に満たない語を解析対象から除外するとはどういうことなのか、また、同義語処理などにより解析対象の単語の持つ頻度は変化することも注意する必要がある。すなわち、解析の操作性や解釈の容易性から、対象とする構成要素（単語）を閾値で絞り込む際には、閾値指定の明確な定式や基準はないので、解析の狙いや目的、対象データの特性などを十分考慮し、探索的に行うことが重要である。



※表 4.1)の「読売新聞記事」のデータについて頻度 20 までグラフ化した。
以降、[その頻度以上の総構成要素数:異なり構成要素数:異なり構成要素率]は、
頻度 50 のとき[1,452,006：3,112：0.3]、頻度 100 のとき[1,353,118：1,684：0.1]となる。

図 12. 単語の出現頻度グラフ（新聞記事の一例）

（2）用語辞書による単語の抽出・削除

解析対象とする単語を絞り込む際に用語辞書を用いることがある。用語辞書の活用方法には、用語辞書に収録された単語あるいは単語の形態素的特性（たとえば、品詞や文字種など）に基づき、登録単語と同一のものを抽出する場合（ここでは「統制語抽出方式」と呼ぶ）と、逆に、登録単語と同一のものを除外する（ここでは「不用語除外方式）と呼ぶ）という場合がある。

たとえば、統制語抽出方式では、解析の目的や対象データの特性から、解析上、重要と思われる単語をあらかじめ辞書に登録し、これと照合したものを解析対象として絞り込むことがあげられる。この場合、解析者の意図にしたがった仮説の検証は評価しやすくなると思われるが、テキストマイニングの狙いである「新たな情報や知識の発見」にはつながりにくい。

一方、不用語除外方式では、記号や句読点、助詞・助動詞などの品詞情報に基づく単語や解析上、不要とする単語（たとえば高出現頻度）などを解析対象から除外する。ただし、記号や助詞・助動詞なども、テキスト型データの記述上の特徴をあらわす場合が少なくないので、登録内容も含めて不用語辞書の適用を適宜切り換えながら利用することが好ましい。

（3）単語の置換

人により作り込まれた誤りや機械的な操作上の特性から発生した誤りをはじめ、抽出された単語には同義や関連を持つものが多いので、テキスト型データの解析にあたっては同義語（関連語）の処置が必要である。

表記上の文字種や表記のゆれ（表 2.、表 3.参照）などの一部は機械的な処理によりに自動的に同義語を置き換えることができるが、一般的には、置換前の単語と置換後の単語を定義した同義語辞書（上位・下位の関係や関連語、さらには複合名詞の置換も含める）を用いる。

ただし、同義語でも使う人の環境、使われる場面や文中の流れ（文脈）により、その意味合いが異なる場合があるので注意を要する。たとえば、「鮨」には、「寿司」「お寿し」「スシ」「すし」「SUSHI」など、いろいろな表現があるが、アンケートなどの自由記述回答においては、回答者の「鮨」に対する思いや年代や生活環境によって差異が生じる場合も少なくない。また、「ナポリタン」や「カルボナーラ」を「スパゲッティ」として置換（同値化）してよいか、さらには「パスタ」や「イタリア料理」とまで拡げて置換してよいかどうかなど、解析の目的や特性を十分に考慮する必要がある。

また、単語の置換（同値化）を行うと単語の出現頻度が変化する。たとえば、「鮨」「寿司」「お寿し」「スシ」「すし」「SUSHI」という各単語の頻度がすべて 2 回のとき、閾値を 10 とすれば、すべての単語が解析対象の候補とはならないが、閾値指定の前に、これらのすべての単語を「鮨」に置換すると、「鮨」の出現頻度は便宜上 12 となり、閾値 10 の指定でも解析対象となる。

このように、閾値指定による単語の抽出や用語辞書による単語の抽出・削除、単語の置換（同値化）は、互いに密接に関連しているので、解析の目的や狙いに応じて、その内容や適用順序を考慮する必要がある。

４．　まとめ

本テキストでは、テキストマイニング（テキスト型データ解析）への活用を前提として、分かち書き処理と形態素解析の概要について解説した。

（１）形態素そのものの明確な定義が与えられているわけではないが、一般に、語を構成する最小の意味のある単位を形態素と呼び、日本語では、名詞や形容詞、動詞語幹、活用語尾、助詞、助動詞などの語（構成単位）が形態素にあたる。

（２）日本語の形態素解析は、通常、狭い意味では「分かち書き」処理、すなわち漢字仮名混じりで「ベタ書き」された自然言語（文）を単語に分割することをいい、文を構成する単語の表記や語形変化という形態論的性質の同定を行う。形態素解析は、仮名漢字変換、音声合成、機械翻訳、情報検索などの自然言語処理の分野で、最も基本的で重要な役割を果たす技術である。

（３）形態素解析の主な処理は、形態素間の接続規則に基づく処理であり、原文を形態素接続規則と形態素解析辞書を用いて形態素に分割し、単語を発見するとともに、その構文上の素性を決定する。形態素解析を行うための重要な処理は、「分かち書き」「単語の品詞の同定」「辞書にない語（未知語）の処理」の３つである。

（４）形態素に分割する際、その切り出し（単語分割）には多様な組み合わせがあり、さらに切り出された単語には品詞上多様な候補が存在する。したがって、形態素解析では、こうした多様な候補の中から、形態素に関する情報を用いて、解釈可能な候補に絞り込むことが目標となる。形態素解析システムでは、こうした形態素解析での解釈の曖昧さを解消するために、経験的優先規則や文法的接続可能性、接続コスト、統計的言語モデルなどの方法が採り入れられている。

（５）テキストマイニングでは、分析対象のテキスト型データを形態素解析や構文解析などを用いて、まず、その内容をあらわす情報や概念を抽出する。構造化・形式化されていないテキスト型データを次の解析ステップ（マイニング）で扱えるようにするための数値変換処理であり、テキストマイニングの特徴的（象徴的）な機能である。単なる単語分割情報（形態素やキーワードの抽出、品詞の同定など）だけでは不十分であり、日本語の持つ曖昧さや単語の同義性や多義性を考慮に入れなければならない。

（６）テキストマイニングの狙いや目的、あるいは、その収集方法などにより、対象とするテキスト型データの質および量は様々である。一般に、テキストマイニングの処理では、解析の狙いや目的に応じて、あるいは解析の操作性や解釈の容易性を確保するために、解析に用いる単語を絞り込む（単語の抽出・削除、あるいは出現頻度による抽出）ことや、同義語や類義語などを編集する（同値置換）ことが重要な課題となる。

テキストマイニングは、膨大なテキスト型データを様々な観点から分析し、役に立つ知識や情報を見つけ出す技術である。概念的にはデータマイニングの一部として捉えることもでき、自然言語処理とデータマイニングの融合技術（広くはこれに可視化技術も要素となる）である。一方、実務的な面からみれば、CRM（Customer Relationship Management）やKM（Knowledge Management）などと同様に、単なる技術ではなく、顧客の信頼を獲得し、維持し続けるためのマネージメントの一つとして位置づけることもできる。今後、テキスト型データの活用の声が高まるとともに、テキストマイニングへの期待もますます高まるものと思われる。

分かち書き処理や形態素解析は、自然言語処理における基本要素であり、テキスト型データを取り扱うテキストマイニングにおいても重要な役割を担うものである。日本語は、複合名詞や複合動詞などの複合語の分割も含めて分かち書きの多様さへの対応が課題となる一方で、消費者や生活者、あるいは顧客の「生の声」を活用することを考慮すると、形態素解析や構文解析以前に、電子化される過程で発生する文書の誤りを処置する手立ても重要になる。

また、学術文献情報や特許情報などは、元来、新しい事実や知見を報告するものであり、記載したとき（書き手）と読むとき（読み手）の概念や意識の差は大きく、これを自動的に認識し理解するのは難しい。

そもそも自由回答・自由記述とは、単に自由に書いて（発言して）貰うだけでは不十分であり、周到に実験計画された環境下で「いかにデータを取得するか」といったデータ取得方法の研究とも密接に結びついている。適切なデータ取得法があって、はじめて解析が意味を持つ。即ち、現象解析の基本はデータにあるというデータ科学（data science）のアプローチを尊重しつつ、テキストマイニングを使いこなすための知恵と工夫がこれからますます重要となる。

【参考・引用文献】

[1] 松本祐治，他（1997），単語と辞書，岩波講座言語の科学 3，岩波書店.

[2] 野口正一（監修），牧野武則（1991），図解 自然言語処理，COM シリーズ，オーム社.

[3] SSC 編（1996），インターネット 即[書く]英語術，SSC.

[4] 冨浦洋一（編集），渡辺日出雄（編集）（2000），特集 ここまできた自然言語処理 –例文の収集とその利用–，情報処理，**41**，7，761-796.

[5] 那須川哲哉（編集），久光徹（編集），李航（編集）（2000），特集 使いやすくなった自然言語処理のフリーソフト –知っておきたいツールの中身–，情報処理，**41**，11，1201-1238.

[6] 全文検索システム協議会編（2000），全文検索システムとは何か？2000 年版

[7] 保田明夫, 大沼美佐 (1998), 言葉の関連性による文書の類似検索, 情報管理, 41, 7, 517-528

[8] 保田明夫（2000），特集 テキストマイニング，薬学図書館，**48**，4，247-252

[9] 大隅昇，Ludovic Lebart（2000），調査における自由回答データの解析–InfoMiner による探索的テキスト型データ解析–，統計数理，**48**，2，339-376.

[10] 大隅昇（2000），定性情報のマイニング–自由回答データの解析–，ESTRELA，74 号，2000 年，5 月号，14-26.

[11] Ludovic Lebart, André Salem, and Lisette Berry (1998), *Exploring Textual Data*, Kluwer Academic Publishers.

付　録

Happiness/AiBASE でみる「分かち書きとキーワード抽出」

ここでは、分かち書き、ならびにキーワード抽出を行うソフトウェアである Happiness©（Happiness/AiBASE(注 1)）の概要を紹介する。

Happiness は、日本語自然文を入力し、品詞に従った「分かち書き」行い、キーワードを抽出する。また、分かち書き文やキーワードに対する「フリガナ付け」を行う。さらには、キーワードの重み付けを行い、文書の内容にあった「重要語」を抽出することもできる。

本付録では、Happiness の機能構成や処理概念の概要を紹介するとともに、辞書登録の実際の例題をまじえながら、分かち書きとキーワード抽出処理について説明する。

なお、Happiness の詳細機能や操作説明、あるいは、「フリガナ付け」や「重要語抽出」機能などについては、必要に応じて開発元などに問い合わせていただきたい。

(注 1)「Happiness/AiBASE」:「WordMiner」(注 2)における構成要素生成・抽出機能として標準装備（Happiness の機能をテキスト型データ解析向けに特化）されている分かち書き・キーワード抽出ソフトウェア。株式会社平和情報センター製。

(注 2)「WordMiner© 」: 産学協同研究成果に基づき、日本電子計算株式会社より開発・販売されているテキスト型データ解析ソフトウェア

１．Happiness の概要

Happiness は、自然言語処理の基幹技術の実現に向けた基本機能を提供し、情報検索や自動翻訳、言語調査研究などの分野に広く活用することができる。

Happiness が提供する基本機能には以下のようなものがある。

・日本語自然文を入力し、品詞に従った「分かち書き」を行う。
・分かち書き文の中から、「キーワード」を抽出する。
・分かち書き文やキーワードに対する「フリガナ付け」を行う。
・キーワードの重み付けを行い、文書の内容にあった「重要語」を抽出する。

１．１　Happiness の機能構成

Happiness は、基本処理機能と重要語抽出機能の２つの機能から構成される。（図 1.）

BASE ルーチンは、利用者が高級言語を用いて呼び出すサブルーチン群であり、「語彙辞書」、「分かち書き辞書」、「不要語辞書」という３つの辞書を使用する。これらの各辞書について標準的に設定したものを基本辞書といい、専門用語や固有名詞、あるいは人名など、目的や用途別に用意する辞書を利用者辞書という。

重要語抽出機能は、キーワードに重み付けを行い、より文書の内容にあった重要語を抽出する機能であり、サブルーチン群と重要を抽出するルールからなる。

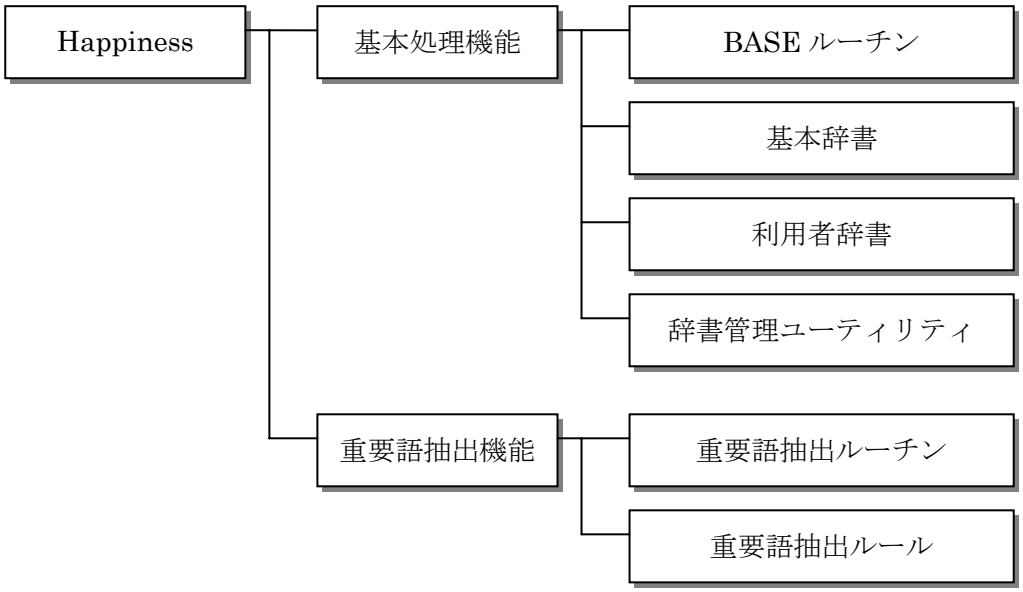


図 1. Happiness の機能構成

１．２　Happiness の処理の概念

Happiness の基本処理では、日本語自然文を入力し、まず、品詞単位に空白を挿入した「分かち書き文」を作成し、その中から不要語を取り除き、キーワードを抽出する。また、これらの分かち書きやキーワードに対してフリガナ付けを行う。分かち書き結果を得るために「語彙辞書」と「分かち書き辞書」を適用し、キーワード抽出には「不要語辞書」、フリガナ付けには「語彙辞書」が適用される。

重要語抽出処理では、分かち書き結果から重要語抽出ルールに従い単語の重み付け（重要度計算）を行い、重要語を抽出する。

図 2.に、Happiness の処理の概念を示す。

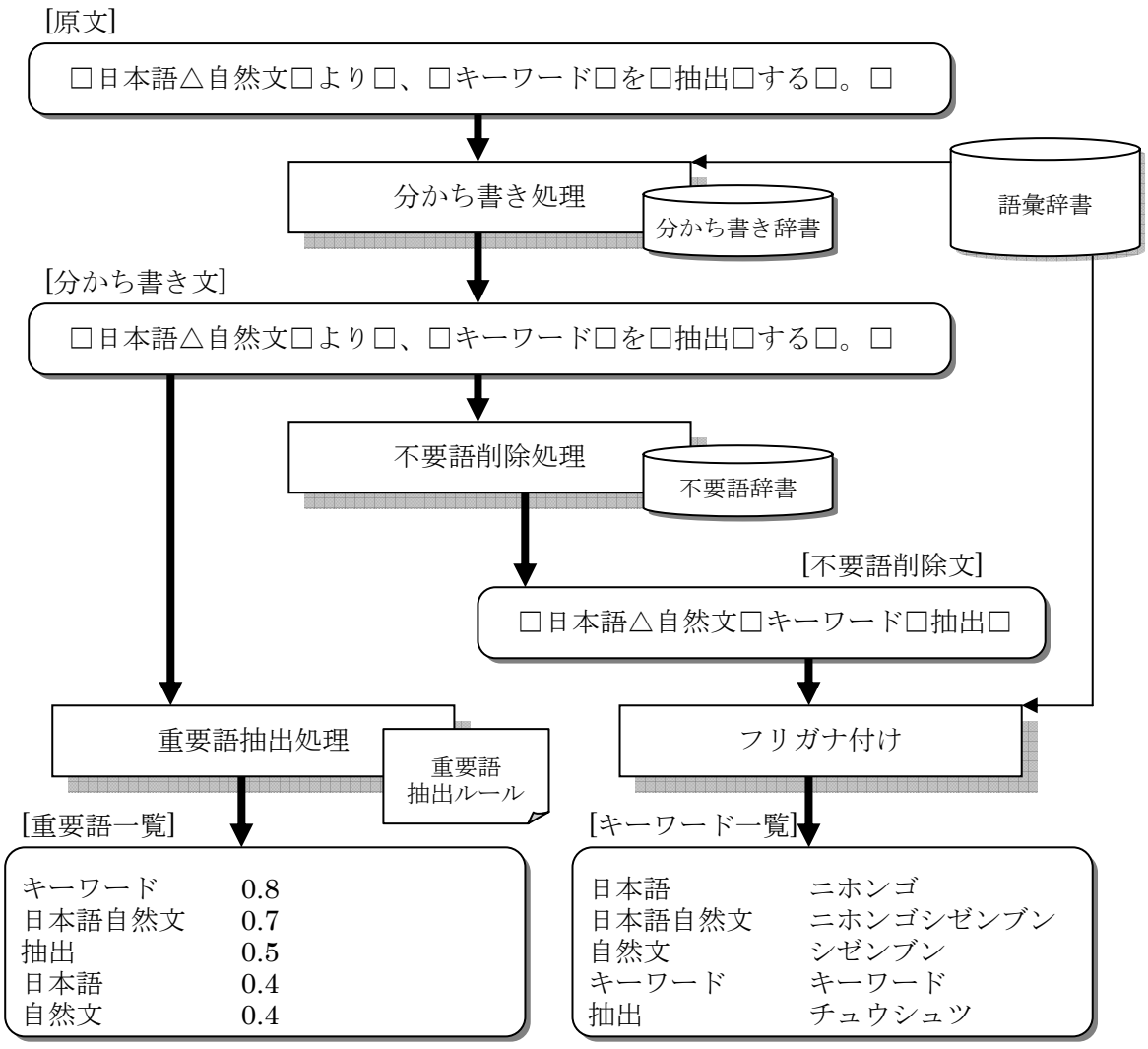


図 2. Happiness の処理の概念

２．　分かち書き処理とキーワード抽出処理

ここでは、Happiness の基本機能である分かち書き処理とキーワード抽出処理の概要について説明する。

２．１　分かち書き処理

分かち書き処理は、最も基本的な処理であり、辞書の編纂基準に従って原文を切断する。

分かち書き結果は、２種類の空白（第１空白と第２空白）を使用し、第２空白で区切られた単語同士は、後述する組み合わせキーワードの対象となる。また、原文中に連続する空白があっても、分かち書き結果は１個の空白に調整する。また、句読点や記号は、１文字単位に第１空白で切断する

以下に基本辞書の主な編纂基準について説明するが、目的や用途によって編纂基準は辞書（基本辞書、利用者辞書）によって変更が可能である。

（１）基本辞書の編纂基準

①　名詞に付く助詞は、単独に切断する。

[例]　私は家にいる　　⇒　□私□は□家□に□いる□

②　用言（動詞・形容詞・助動詞）の連用形や仮定形に付く接続助詞（「て」「ても」「でも」「ば」）は、用言とつなげて、その後を切断する。

[例]　頑張れば出来なくてもいい　　⇒　□頑張れば□出来なくても□いい□

塞いでもだめだ　　⇒　□塞いでも□だめ□だ□

（注意）名詞に付く係助詞「でも」は単独で切断する。

お茶でも飲もう　　⇒　□お茶□でも□飲もう□

③　用言の終止形や連体形の後を切断する。

[例]　美しい日本に住めるなら　　⇒　□美しい□日本□に□住める□なら□

④　動詞の連用形は、助動詞がつながる場合は切断せず、名詞がつながる場合は切断する。

[例]　語学を活かしたいから金を貯めアメリカに行く

⇒　□語学□を□活かしたい□から□金□を□貯め□アメリカ□に□行く□

⑤　動詞の連用形に名詞がつながる場合でも、動名詞となる場合は、ひとつの名詞とする。

[例]　壁に張り紙をする　　⇒　□壁□に□張り紙□を□する□

（注意）辞書に登録されていない場合は切断される。ただし、つながる名詞が接尾辞と認識される場合は第２空白で切断する。

□壁□に□張り△紙□を□する□　（「張り紙」が未登録、かつ「紙」が接尾辞）

⑥　連動詞（動詞連用形＋動詞）は、ひとつの動詞と見なす。

[例]　今後話し合おう　　⇒　□今後□話し合おう□

（注意）辞書に登録されていない場合は切断される。

彼も飲み始めた　　⇒　□彼□も□飲み□始めた□

⑦　名詞に「する」が付くサ変動詞の場合、その語尾を切断する。

[例]　開発し運用する　　⇒　□開発□し□運用□する□

⑧　副詞・接続詞・感動詞は単独で切断する。

[例]　まさかいきなり飛び出すとは　　⇒　□まさか□いきなり□飛び出す□と□は□

まあしかし助かった　　⇒　□まあ□しかし□助かった□

⑨　連体詞＋名詞が慣用句となっているものは、ひとつの名詞とする。

それ以外は第２空白で切断する。

[例]　わが国の実情　　⇒　□わが国□の□実情□

わが祖国に栄光を　　⇒　□わが△祖国□に□栄光□を□

⑩　形容動詞の終止形・連体形、及び連用形の「に」は語幹に付け、その後を切断する。

ただし、名詞に「だ」「に」「な」が付いて成立する形容動詞は、その語尾を単独に切断する。

[例]　静かな雰囲気だ　　⇒　□静かな□雰囲気□だ□

⑪　名詞と名詞の間は第２空白で切断する。

[例]　要求仕様をまとめる　　⇒□要求△仕様□を□まとめる□

⑫　接頭辞・接尾辞には、切断するものとしないものがある。（辞書に従う）

切断する場合、名詞との間は第２空白となる。

[例]　元日本兵の副長官達　　⇒□元□日本兵□の□副長官□達□

⑬　接尾辞のうち、助数詞は数詞（数字）に付ける。

[例]　二十五日に１００万円を渡す　　⇒□二十五日□に□１００万円□を□渡す□

（2）パスコード指定

分かち書きに対して、一部の文字列を切断禁止にすることができる。これを「パス指定」と呼び、原文中の切断禁止とする部分を「パス開始文字」と「パス終了文字」で挟む。

このとき、分かち書きの結果は、切断禁止部分は両端のパスコード（パス開始・終了文字）を含めてひとつにつながった状態になり、キーワードは両端のパスコードを取り除いたものになる。

[例]　『赤い靴』の作者とその時代背景－『NHK 特集』

⇒　□『赤い靴』□の□作者□と□その□時代△背景□－□『NHK 特集』□

（注意）パスコードはパス開始文字“『”、パス終了文字“』”

（3）接尾辞解釈

日本語は、名詞に接尾辞を付加して意味を拡張する場合が数多くある。

たとえば、「漫画家」は名詞「漫画」に接尾辞「家」がついて作られた言葉である。

辞書に「画家」が登録されていて、「漫画家」がないとき、「漫△画家」という切断ミスが発生しないように接尾辞解釈を適用する。

[例]　彼は漫画家　⇒　□彼□は□漫画家□

浪漫画家　⇒　□浪漫△画家□

（注意）索引した用語（例では「画家」）の直前を参照し、熟語となっていれば、索引した語を確定させる。

ただし、遡るのは直前の単語までであり、次のような場合は切断ミスが発生する。

この場合は「漫画家」を辞書に登録する必要がある。

[例]　放浪漫画家　⇒　□放△浪漫△画家□

２．２　キーワード抽出

Happiness のキーワード抽出は「不要語除去方式」である。

「不要語除去方式」とは、キーワードとして必要な単語を取り出すのではなく、キーワードとして不要な単語を取り除き、残った単語をキーワードとする方式である。

（１）不要語の基準

キーワードとして不要であるとする単語は、不要語辞書に登録された用語によって選別される。基本辞書では、概ね名詞以外を不要語辞書に収録している。

なお、辞書によらず、１文字の非漢字は不要語として扱う。

（２）組み合わせキーワード

組み合わせキーワードとは、名詞が連続している場合、それらをつなぎ合わせてキーワードとすることをいう。

名詞の連続は分かち書き結果の空白の種類で判定する。つまり、第２空白で切断された単語同士が組み合わせの対象となる。

組み合わせ条件には、「短単位」、「長単位」、「組み合わせる単語の数の範囲」の３つの指定方法がある。

「短単位」とは、前後に名詞がなく、独立している単語である。すなわち、前後が第１空白に挟まれているものである。また、「長単位」とは、第２空白でつながっている範囲の単語をすべてつなぎ合わせたものである。

たとえば、以下の分かち書き結果が得られたとき、

□AIDS□は□後天性△免疫△不全△症候群□の□こと□で□ある□

「AIDS」は「短単位」、「後天性免疫不全症候群」は「長単位」はとなる。

また、組み合わせ単語の範囲を「１～２」と指定すると、

「AIDS、後天性、後天性免疫、免疫、免疫不全、不全、不全症候群、症候群」

がキーワードとして取り出される。

組み合わせ単語の範囲を「２～２」と指定すると、

「後天性免疫、免疫不全、不全症候群」

だけが取り出される。

３．　分かち書きの問題と辞書登録

分かち書き処理およびキーワード抽出処理は、「語彙辞書」「分かち書き辞書」「不要語」の３つの辞書に登録された用語情報に基づいて行われる。

利用者は、専門用語や固有名詞、あるいは人名などを利用者辞書として組み込むことにより、対象とする領域において、より精度の高い結果を得ることができるようになる。一方、辞書に登録した用語によっては、逆に、弊害となる場合もあるので、用語の登録については、充分に検討・確認する必要がある。

以下に、よくある用語登録の事例について、その注意点も含めて説明する。

①　長単位語が切断されない。

長単位語が切断されない場合は、語彙辞書に短単位の用語を登録する。

カタカナや英字で構成される単語（たとえば、化学物質名や新しい政治用語など）でしばしば発生する事象である。

[例]　□バスケットボール□　　⇒　□バスケット□ボール□

（説明）「バスケット」と「ボール」を語彙辞書に登録する。

②　切断位置が不適切である。

[例]　□女子大△回転△競技□　　⇒　□女子△大回転△競技□

（説明）「大回転」を語彙辞書に登録する。

ただし、いずれの場合も長単位指定のキーワードは「女子大回転競技」となる。

③　ひらがな交じりの名詞や固有名詞に切断ミスが発生する。

[例]　□こ□が□ね□むし□　　⇒　□こがねむし□

（説明）この例は、「こ（名詞）＋が（格助詞）＋ね（終助詞）＋むし（名詞）」と

と解釈した結果である。「こがねむし」を語彙辞書に登録する。

④　「日付け」の使い分けができない。

[例]　□日付け□は□三△日付け□で□　⇒□日付け□は□三日□付け□で□

（説明）このような単位を表す接尾辞には、直前の文字が数字（漢数字を含む）であるか

どうかをチェックし、それによって切断パターンを変える助数詞指定を適用する。

⑤　助詞を含んだ固有名詞が切断されてしまう。

[例]　□杜□の□都□　　⇒　□杜の都□

（説明）このような場合、長単位の指定をしても「杜の都」というキーワードは抽出

できない。したがって「杜の都」そのものを語彙辞書に登録する

⑥　人名がうまく切断できない。

[例]　□西□川□き□よし□　　⇒□西川□きよし□

⇒□西川きよし□

（説明）人名においては、登録する用語（語彙）により、姓と名を切断する場合と姓と名は切断しない場合を操作できる。ただし、単語長（文字列長）の短いもの（とくに、ひらがなの名前など）を辞書に登録すると、弊害を生じる可能性が高まるので注意を要する。

⑦　専門用語や略語が切断ミスを起こす。

[例]　□ヘリ□の□空△撮△映像□　　⇒□ヘリ□の□空撮△映像□

（説明）「空撮」を語彙辞書に登録する。ただし、辞書に登録しなくても、組み合わせ指定により「空撮」ならびに「空撮映像」もキーワードとして得ることができる。

⑧　接尾辞の関連で切断ミスを起こす。

[例]　□給△食用□　　⇒□給食用□

⇒□給食△用□

（説明）この場合は接尾辞解釈の指定をする。接尾辞解釈には、直前の用語と分かれるものと、分かれないものの指定が行える。

⑨　動名詞が切断される。

[例]　□張り△紙□　　⇒□張り紙□

（説明）「張り紙」を語彙辞書に登録する。

⑩　不要語にしたい。（キーワードとして抽出したくない）

（説明）不要語に登録する。たとえば、「研究開発報告」の論文集における「研究」や「開発」「報告」「目的」「効果」などは、どの論文（原文）にも出現するので、不要語として扱いたい場合がある。

なお、不要語辞書に登録した場合には、キーワードとして抽出されないだけで、分かち書きには影響しない。

ただし、用語は、文章の内容や検索（解析）する目的によって不要となったり、重要となったりするので、一意に不要語とするのは慎重を要する。

⑪　前方一致で不要語にしたいが例外がある。

（説明）たとえば、「～的」は不要語とするが、「標的」はキーワードとするような場合が考えられる。この場合、不要語「的」による前方一致指定を登録するとともに、例外救済として「標的」を不要語辞書に登録する。

⑫　助数詞を不要語としたいが、例外がある。

（説明）たとえば、「～番」は不要語としたいが、「１１０番」はキーワードとしたい場合などがある。助数詞として不要語「番」による前方一致指定を登録するとともに、例外救済として「１１０番」を登録する。

なお、「当番」「順番」などは助数詞とはならないので、ここの登録では不要語にはならない。

⑬　組み合わせキーワードの対象から除外したい

（説明）たとえば、「最近」「今年」「当該」などが考えられる。語彙辞書に他の用語とは独立する用語として登録する。