汎用的な意味解析技術への挑戦

NTTが開発した日本最大級のシソーラスである「日本語語彙大系」や単語の意味を記述した基本語意味データベース「Lexeed」を紹介し，これらの言語データベースを利用して日本語テキストの単語や文の意味を解析する汎用的なソフトウェアを開発する研究を紹介します．

永田 昌明（ながた まさあき） /藤田 早苗（ふじた さなえ）
平 博順（たいら ひろとし）

NTTコミュニケーション科学基礎研究所

「ことばの意味を理解する」コンピュータを目指して

NTTコミュニケーション科学基礎研究所では，人間のようにことばを操るコンピュータの実現を究極の目標として，ことばの意味をコンピュータ上で表現する方法や，人間が話したり書いたりしたことばをコンピュータ向きの意味表現に変換して高度な言語処理アプリケーションを実現する方法について研究しています．

この研究の一環として，これまでに日本最大級の日本語シソーラスである「日本語語彙大系」，単語の意味（語義）を定義した基本語意味データベース「Lexeed」，新聞記事など実際のテキストに対して意味解析の正解データを付与した構文意味データベース「檜（ひのき）」などを構築してきました．

本稿では，まずこれらのことばの意味に関するデータベースを紹介し，次に単語の意味および文の意味を解析する技術である語義曖昧性解消と述語項構造解析を紹介します．さらに，質問応答や要約などの高度な言語処理アプリケーションを実現するための「意味解析のミドルウェア」として近年盛んに研究されているテキスト含意認識（Textual Entailment Recognition）についても紹介します．

日本語語彙大系

日本語語彙大系は，約3 000種類の意味カテゴリを用いて約40万語の日本語の単語の意味を定義したものです(1)．一般名詞・固有名詞・用言に対して３つの異なる意味カテゴリの体系がありますが，もっとも良く使うのは一般名詞カテゴリです．

一般名詞意味カテゴリの一部を図１に示します．意味カテゴリは上位下位

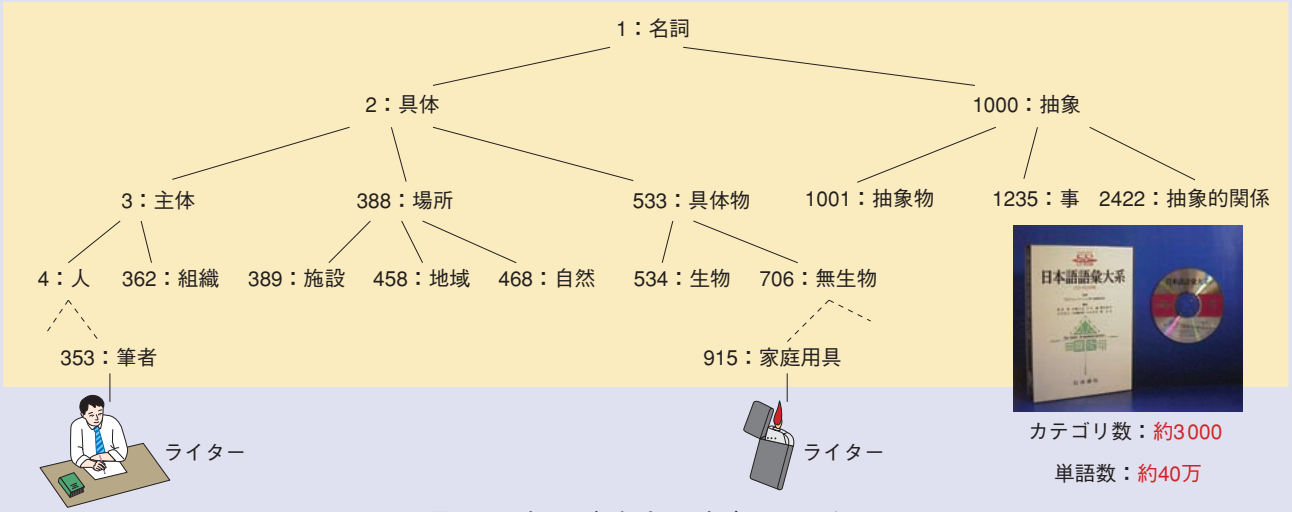


図１　日本語語彙大系の一般名詞カテゴリ

関係または全体部分関係により体系化されており，カテゴリ名とともに1から始まる固有の番号が付与されています．またそれぞれの単語にはその意味を表す意味カテゴリ（番号）が最大5つまで付与されています．

例えば「ライター」という単語には「353：筆者」と「915：家庭用具」という2つの意味カテゴリが付与されています．また意味カテゴリの親子関係をたどることによって，前者は動作の主体となり得るもの（3：主体）であるのに対し，後者は具体的な物（533：具体物）であることが分かります．

日本語語彙大系は，もともとALT-J/Eという日英翻訳システムにおいて日本語から英語への構文変換パターンを記述するために開発されたので，パターンに基づいてテキストから情報を抽出してマイニングするタスクなどに有用です．例えば，ユーザがどこで買い物をするかを調べるために「～で買う」というパターンをブログから抽出する際に，「～」に当てはまる名詞の意味カテゴリを「388：場所」に制限すれば「ローン（1174：支出）で買う」のような用例を簡単に排除できます．

基本語意味データベースLexeedと日本語ツリーバンク檜

Lexeedは，人間にとっての馴染みの深さを表す親密度[(2)]という心理学的な尺度に基づいて日本語の基本語（2万8 000語）と語義（4万6 000語義）を選定し，各語義に対して語義文と用例文を付与した電子化辞書です[(3)]．上位・類義・連想などの語義間の関係が体系的に記述され，各語義は日本語語彙体系とも関係付けられています．

Lexeedにおける「ドライバー」の記述の例を**図2**に示します．日本語語彙大系では「ドライバー」という単語に「292：運転手，942：工具」という意味カテゴリを付与しているだけですが，Lexeedでは前者の語義に「車を運転する人」という定義文と「父はドライバーだ」という例文を付与しています．

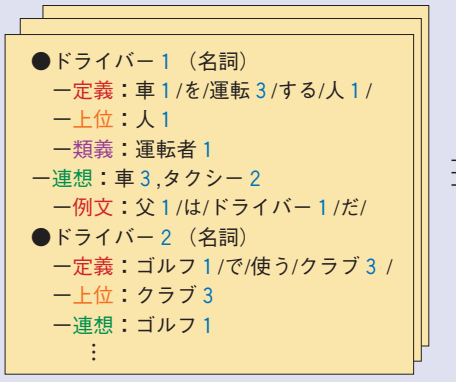


図2　基本語意味データベースLexeed

Lexeedは，語義文を基本語だけで記述し，語義文の各単語にLexeedの語義番号を付与しています．英語ではLDOCE（Longman Dictionary Of Contemporary English）の見出し語の説明が基礎語彙だけで記述されています．しかし，日本語ではこのように自己完結性があり組織化された意味の定義を持つ電子化辞書はほかにありません．

さらに単語の意味だけでなく文の意味を研究するための基礎データとして，Lexeedの語義文と用例文と新聞記事を合わせた約20万文に対して構文構造と意味構造を付与した日本語ツリーバンク檜を作成しました[(4)]．ツリーバンク（treebank）とは，名詞句と動詞句から文が構成されるといった文の構文構造を表現する構文木（syntactic tree）と呼ばれるデータを集めたデータベースのことです．

檜は，主辞駆動句構造文法（HPSG：Head-driven Phrase Structure Grammar）という言語理論に基づいて構文意味構造を定義しています．**図3**は，「車を運転する人」という文において，「車を」と「運転する」という2つの句から構成される「車を運転する」という句が「人」を修飾するという構文構造と，「運転」という動作において「人」が主体で「車」が対象であるという意味構造とが対応することを表現しています．

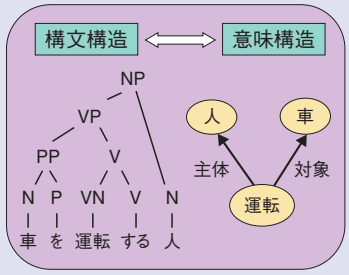


図3　日本語ツリーバンク檜

Lexeedは心理学的な知見に基づいて設計されているので，例えばマニュアルの分かりやすさを判定するなど，言語の読みやすさや理解しやすさを定量的に評価する必要がある教育やユーザインタフェース設計などの分野での応用に適しています．さらに檜はLexeedと組み合わせることにより，単語レベルから文レベルまで統合した詳細な意

味記述を持つ大規模な言語データベースを形成するので，語義と構文構造と意味構造の関係を利用した新しい意味解析技術の研究を可能にしています．

語義曖昧性解消

私たちはこれまで構築した言語データベースを利用して，現在，日本語テキストの意味を解析する汎用的なソフトウェアの研究開発を行っています．まず単語の意味を解析するための第一歩として，語義曖昧性解消（word sense disambiguation）を研究しています．

多くの語には複数の意味（語義）があり，語がどの意味で使われているかは前後の文脈で決まります．例えば「ドライバー」という語には，運転手，ゴルフ道具，ねじ回し，駆動ソフトウェアなどの語義があります．語義曖昧性解消とは，ある文脈における単語の語義を辞書に記述された語義の候補から選ぶタスクです．

図 4 のように各単語に対して人手でLexeedの語義を付与したコーパス（約20万文）を訓練データとして，ある単語の前後の単語の表記・品詞・語義などからその単語の語義を判別する特徴を機械学習手法により学習し，任意の入力文の各単語に対してLexeedの語義を付与するソフトウェアを開発しました．私たちはこれを「語義タガー」と呼んでいます[5]．

ことばの意味は多種多様であり，判別すべき対象をLexeedの語義（約５万個）に制限しても，各語義の学習データは数例しかありません．そこで日本語語彙大系で表現されている名詞の意味の階層構造を利用して，語義の判別を大まかな意味の判別と詳細な語義の判別の２段階に分けることにより語義曖昧性解消の精度を向上させる工夫をしています[6]．

語義曖昧性解消の応用としては語義に基づく情報検索などが考えられます．例えば現在の検索エンジンでは「ドライバー」というキーワードを入力すると，車に関する文書とゴルフに関する文書が区別されずに表示されますが，これを語義別に分けて表示すれば，検索結果の絞り込みが容易になります．

述語項構造解析

次に文の意味を解析するための第一歩として，私たちは述語項構造解析（ predicate argument structure analysis）を研究しています．

動詞や形容詞などの「述語」は文の中心的な要素であり，動きや状態などの事態を表現します．「名詞＋格助詞」という形式で述語が表す事態に関係する人やものを表現する要素を「項」と呼びます．述語項構造解析とは，文中の各述語について述語が表す意味を補う働きをする項を同定するタスクです．

例えば**図 5** に示す「きのう買ったソフトのインストールに苦労した」という文は，それぞれ「買った」および「苦労した」という述語を中心とする２つの事態を表現しています．さらにこの文では「インストール」という名詞も事態を表現しており，このような名詞を事態性名詞と呼びます．

私たちは述語項構造解析を，単語間の係り受け構造から述語項構造への構造変換であるととらえ，係り受け構造と述語項構造が人手により付与されたコーパスを訓練データとして，構造学習手法により述語項構造を決定する手法を考案しました[7]．

述語項構造解析の応用としては事態

図 4　語義曖昧性解消

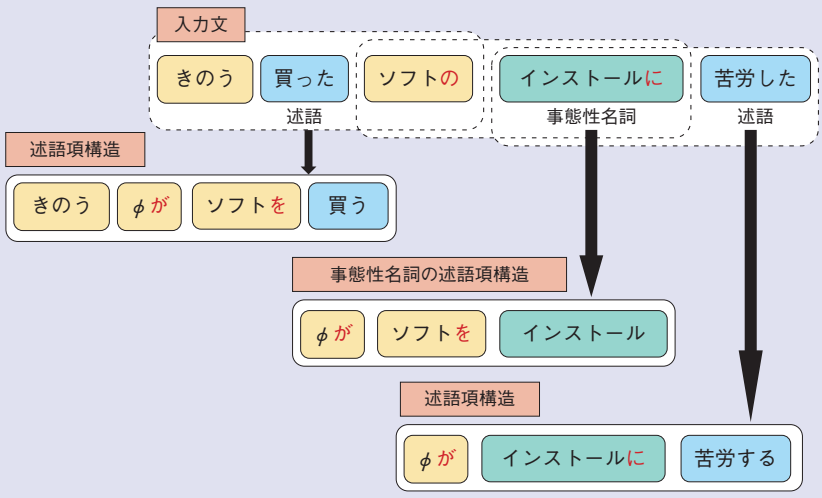


図 5　述語項構造解析

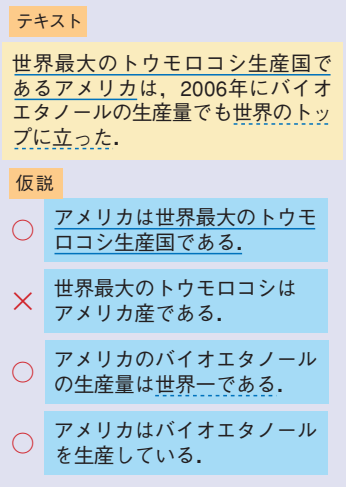


図 6　テキスト含意認識

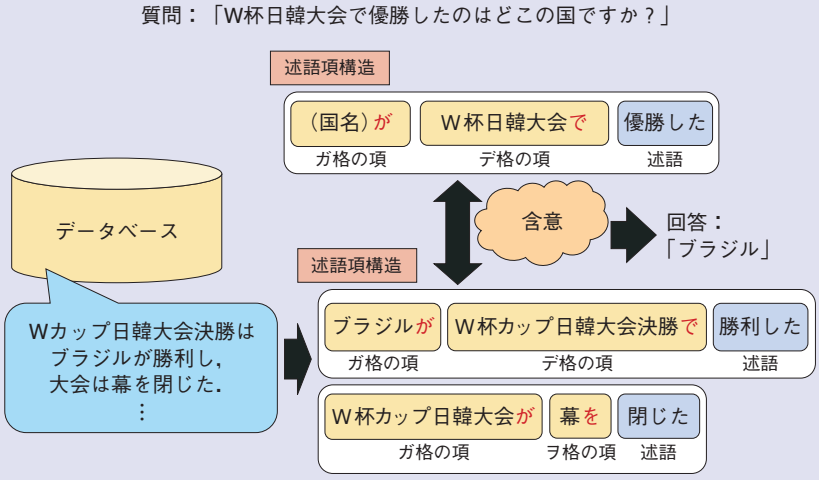


図 7　テキスト含意認識を利用した質問応答

に基づく情報検索などが考えられます．例えば現在の検索エンジンでは「ソフトの購入方法」のようなHow-Toを検索することが難しいですが，述語項構造解析を使えば，平叙文「ソフトを購入する」，関係節「購入したソフト」，名詞句「ソフトの購入」，複合名詞「ソフト購入」などの多様な表現を事態レベルで統一的に扱うことができます．

テキスト含意認識

コンピュータによる「ことばの意味」の理解度はどうすれば計測できるでしょうか？　近年，テキスト含意認識というタスクが提案され，この問題への回答として注目を集めています[8]．

テキスト含意認識は，「テキスト」と「仮説」と呼ばれる２つの文章が与えられた際に，テキストが仮説を含意（entail）するか否かを判定するタスクです．これはマークシート方式の国語のテストで「次の文章を読んで正しいものには○，間違っているものには×を記入せよ」という問題と同じです．

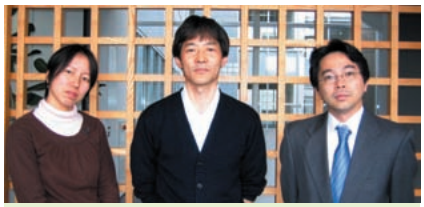
図 6 に示すテキストが仮説を含意しているかを判定するには，２つの下線部が同じ事態を表すという述語項構造解析の能力や，点線部が互いに言い換えであるという語彙知識，「生産量がトップ」なら当然「生産」しているという論理的推論の能力などが必要です．

またもし汎用的なテキスト含意認識モジュールを実現できれば，これを部品として用いて質問応答や要約などの高度な言語処理アプリケーションを構築することが可能です．質問応答の例を**図 7** に示します．質問文を平叙文に変換した仮説を含意するテキストがデータベースにあれば，疑問語に対応する項を抽出することにより質問に対する回答を得ることができます．

日本語のテキスト含意認識の研究は始まったばかりです．私たちは約200個のベンチマークデータを作成して，2007年から研究を始めましたので，近い将来，別途成果を報告したいと思います．

■参考文献

(1) 池原・宮崎・白井・横尾・中岩・小倉・大山・林：“日本語語彙大系，”岩波書店, 1997.
(2) 天野・近藤：“日本語の語彙特性　第１巻 単語親密度，”三省堂, 1999.
(3) 笠原・佐藤・ボンド・田中・藤田・金杉：“「基本語意味データベース：Lexeed」の構築，”自然言語処理研究会, pp.75-82, 2004.
(4) F.Bond, S.Fujita, and T.Tanaka: “The Hinoki Syntactic and Semantic Treebank of Japanese，” Language Resources and Evaluation, 2006.
(5) T.Tanaka, F.Bond, T.Baldwin, S.Fujita, and C.Hashimoto：“Word Sense Disambiguation Incorporating Lexical and Structural Semantic Information，”EMNLP-CoNLL, pp.477-485, 2007.
(6) 藤田・ボンド・藤野：“上位意味クラス推定を用いた語義曖昧性解消，”言語処理学会第14回年次大会, 2008.
(7) 平・永田：“構造学習を用いた述語項構造解析，”言語処理学会第14回年次大会, 2008.
(8) D.Giampiccolo，B.Magnini，I.Dagan，and B.Dolan：“The Third PASCAL Recognizing Textual Entailment Challenge，”ACL-PASCAL Workshop, 2007.



（左から）藤田　早苗／永田　昌明／
平　博順

ことばの意味の世界は混沌としていますが，私たちはこれを辛抱強く整理して，誰でも簡単に賢いテキスト処理アプリケーションをつくれるような辞書やライブラリを提供していきたいと思っています．

◆問い合わせ先
NTTコミュニケーション科学基礎研究所
TEL　0774-93-5149
FAX　0774-93-5345
E-mail　nagata.masaaki@lab.ntt.co.jp