日本語の統語・意味コーパス「檜」

Francis Bond, 藤田 早苗, 田中 貴秋, 中岩 浩巳
日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
<{bond,sanae,takaaki,nakaiwa}@cslab.kecl.ntt.co.jp>

Abstract

本稿では詳細な統語情報と意味情報の両方が付与されたツリーバンク「檜」 を紹介する。統語情報は、HPSG に基づく日本語文法を利用して付与し、意味情報は基本語彙意味辞書 Lexeed を利用して付与した。本ツリーバンクは、語義文と例文をあわせて、12 万文、120 万語、54 万内容語からなる。

1 はじめに

我々は自然言語理解のメカニズムの解明とそれを計算機に実装する手法の研究を進めている。統語情報と意味情報を統合して扱うことのできる自然言語処理を実現するための基盤となる言語資源として構文木コーパス「檜」の構築を進めている。本研究の究極の目標は機械に自然言語を理解させることであり、そのためにテキストを統語的に解析するだけでなく、意味的に解析し、意味情報を獲得することを目指している。

我々は、この目標を達成するために、以下の言語要素を盛込む事が必要と考えている。

1. 語彙情報を集結した**辞書**
2. 単語と意味をつなぐ**文法**
3. 意味情報を体系化する**オントロジ**
4. 語の用法を示す**コーパス**

さらに、言語理解のためには、これら言語要素を有機的に統合・融合することが重要となる。すなわち、辞書に登録されている語の説明文に対して統語的情報を精緻な文法に基づいて付与するとともに、その語の意味情報をオントロジへのリンクにより表現する。また、コーパス中の各文に対し文法と辞書を用いて構文木情報と語義情報を付与することによって、語の用法を量的に分析することが可能となる。「檜」の構文木コーパスは、意味辞書である基本語意味データベース Lexeed と日本語 HPSG に基づいて、意味辞書自身及びコーパス中の文に統語情報、意味情報を付与したものであり、この「檜」の情報からオントロジの構築を行っている。このようにして、これらの四つの言語要素を一つに融合し処理する事ができるのである。

ここ十数年来、自然言語処理の分野で飛躍的に進歩しているのは、統計的自然言語処理を用いて諸問題を解決する技術である。コーパスや辞書などの電子化されたデータの増加が、この進歩に拍車をかけたといえる。統計的手法の利点は、対象とする分野に対する学習データがあれば、効率的に精度とカバー率を向上できる点である。統計的手法は形態素解析、構文解析、語義曖昧性解消の問題などで、大きな成果を上げている。

しかしその一方で、性能は学習するデータに依存するため、学習データの性質に起因する性能限界がある。多くの場合、学習データとして表層的な情報のみが使われているので、学習データ中に、低頻度でしか観測できない言語現象については、十分に学習を行うことができない場合がある。 さらに、言語現象は多様であるため、大量のデータを使用しても、対象とする表現と表層的に同じ表現が学習データ中に現れるとは限らない。

これらの問題を大幅に軽減し、自然言語の理解といった、より高度な問題を扱うためには、表層的な情報だけでなく、より詳細な統語情報や意味情報をコーパスに付与し、学習を用いることで統計的手法と詳細な意味処理を融合する手法が望まれる [1]。 しかし、現時点ではこれらの情報が付与された大規模なコーパスは存在しない。文節間の依存関係が付与されたコーパスであれば、日本語では京大コーパス [2] や EDR[3] などがある。しかし、下位範疇化構造などの詳細な統語情報や、単語の意味の区別するタグなどの意味情報を持つ日本語の大規模なコーパスは存在しない [1]。

意味情報まで付与されている大規模な日本語コーパスが現在まで存在しなかった理由は、(i) 構築が難しかったこと、(ii) 単語あるいは文節間の依存関係程度の情報で十分であると考えられてきていたこと[1]、という二点が挙げられる。

(i) に関しては、後に詳しく述べるが、近年、DELPH-IN(5.1 節参照) 等により、文法理論の実装とコーパス構築を同時に、かつ、効率的に行なえるツールが整備されてきている。そのため、文法の専門家でなくても容易に詳細な情報をもつコーパスを構築する事が可能であり、規模も統計学習に利用できるものとなってきている [4]。

以上の事を踏まえ、我々は、統語情報と意味情報を融合する第一段階としてツリーバンク「檜」の構築を進めている。「檜」は、表層的な処理に頼らず利用できるた

[1]なお、ツリーバンク「檜」から依存関係のみを抽出することもできる。そのため、「檜」は、依存関係の情報のみが付与されたコーパスに対して上位互換であると言える。

め、質問応答や要約、翻訳など、あらゆる自然言語処理分野において、統計的手法と意味情報を統合した高度な処理を実現するための基盤的言語資源となる。例えば、機械翻訳に適用する場合には、意味情報を利用することで、分野や言語に依存しない、汎用的な自然言語処理技術が実現できるという利点がある。つまり、言語非依存な意味表現を介して他の言語に翻訳することができる。

本稿ではツリーバンク「檜」の特徴について、具体的に構築に用いられた日本語文法・意味表現や対象データについて述べ、構築されたツリーバンクの概略について報告する。

2 辞書：Lexeed

実際のツリーバンク「檜」の構築対象の辞書として、基本語意味データベース Lexeed[5] を選択した。Lexeed は、日本語で一般的に使用されている語を網羅的に収録し、各語義に語義文が付けられたものである。 Lexeed に収録されている語は、日本人の各語に対するなじみ深さの度合を表す「単語親密度」に基づいて選定されている。単語親密度とは語に対するなじみ深さの度合を 1 から 7 の実数で表したものであり、7 が最もなじみ深いことを示す [6]。

このうち、単語親密度が 5 以上である 28,270 語を「基本語」と定義し、Lexeed に収録されている。この基本語は、典型的な日本語の新聞に出現する一般語のうち延べ数で 75%以上をカバーしている [7]。

多くの語は複数の語義を持つので、総数で 46,347 の語義が収録されている。全ての語義には、1 文以上の語義文による説明が付与されており、Lexeed 全体では、81,100 文の語義文が存在する。全ての語義文は、辞書の中で自己完結するように、基本語のみで書き換えられている。書き換えられた語義文中で最終的に使用された基本語は、全体の 60%、16,914 語であった。また、各語義には基本語のみからなる例文も付与されている。

このような言語資源構築のアプローチは JUMAN での構築法と同じ考え方に基づくものである。具体的には、3 万単語のコアとなる語を人手で構築した後、コーパスの処理により未収録語の拡張を行った [2]。

図 1 に、「ドライバー」という基本語を例に Lexeed に含まれる情報と、ツリーバンク「檜」によって新しく付与される情報を示す。「檜」によって付与される情報は下線を引いた部分である。

3 文法：Jacy

HPSG とは、統語解析と意味解析が密接に関連した解析が可能な、主辞駆動句構造文法 (Head-driven Phrase Structure Grammar: HPSG) [8] であり、言語学および隣接科学における数多くの異なった研究に基づいている。HPSG は、重要なアイデアの多くを意味理論や情報科学の成果より得ており、その枠組は統語解析などの言語処理の基礎技術だけでなく、形式意味論との親和性も高い。HPSG は、統語解析と意味解析が密接に関係しているため、データの変更や拡大に際し、比較的両者の整合性を保持しやすい。

HPSG に立脚した日本語分析と文法実装の試みは，ICOT による JPSG[9] をはじめとして既にいくつか存在している。しかし、それらはいずれも日本語の一般的性質に関する理論の精緻化を指向したものか、一現象の説明における形式的妥当性の証明を目的とするものであった。そうした研究は理論的には重要である。しかし、システムの実用性の面から見れば、これまでに構築されたモデルは小規模で、処理できるデータが少なく、現実的な解析を行なうには不十分な規模であった [10, 11]。

いわば実験的な取組みであった先行研究に対し、本研究で利用している Jacy は、実用指向の大規模日本語 HPSG 文法である [12]。この文法は、Verbmobil プロジェクト [13] に端を発し、当初は旅行企画に関する対話文の処理を意図して記述されていた。その後、対象分野を広げつつ語彙や規則の拡張をすすめてきた。文法開発環境はオープンソースの DELPH-IN(5.1 章参照) ツールを利用しており、拡張が容易に行える。 また、Jacy を使った解析器は、形態素解析システム茶筌 [14][2]の出力を利用した未知語処理を実装しており、記述文や電子メールの自動応答など未知語を含んだテキストにも対応できる実用システムとなっている [3]。

Jacy は、もともと対話コーパスを対象としていたため、辞書語義文をドメインとした解析を行うためには文法を拡張する必要があったが、数人月の作業により辞書の語義文・例文の解析カバー率を 82%に向上させた [4].

4 オントロジ

Lexeed の語義に基づいて、日本語語彙大系の構文体系と意味大系の拡張を試みた。この設計は本研究の基礎的な目的の 1 つである。本研究では、まず大枠の設計を行った後、詳細化する方針で構築を進めた。

まず最初に、語義文を用いて各語義間の関係を抽出した。最も有用で抽出が簡単な関係は「上位-下位関係」であり、解析結果の意味表象から抽出した [16].。このようにして抽出した関係は Tokunaga らの研究 [17] と同様に、語彙大系 のような他の辞書との関係付けに活用した。半自動的に獲得された関係のうち約 80%が有効な関係であることが確かめられている。残りについては人手による修正をすることで構築を進めた。同様の拡張は英語に対しても GCIDE と ERG を用いて行った。

今後は、オントロジ構築に関連して以下の研究を進める予定である。

(a) 他の関係の抽出

[2]http://chasen.aist-nara.ac.jp/を参照。

[3]Lexical Functional Grammar を理論的基盤とする ParGram プロジェクトにおいても実装環境の整備と日本語の分析 [15] がすすめられている。本稿と同様に大規模データの解析を指向した文法設計ではあるが、オープンではないため、利用できない。

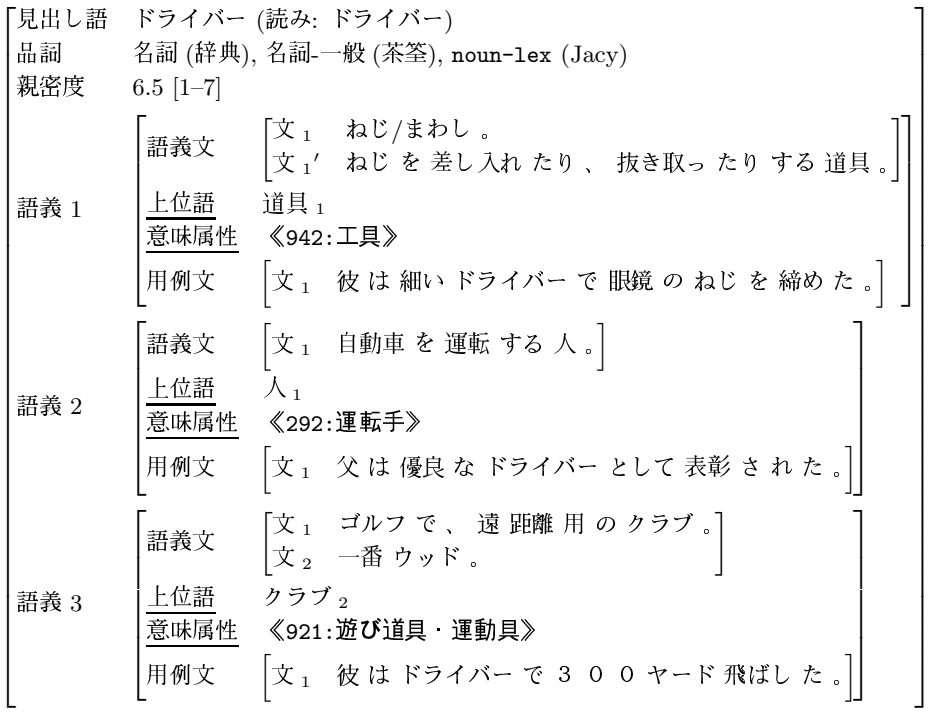

図 1: Lexeed における「ドライバー」の意味記述

(b) verb-argument 関係の抽出
(c) LSA ベースの手法との融合・比較
(d) 岩波辞書 (Senseval) との比較

5 コーパス

5.1 ツリーバンク

ツリーバンクの構築は 2 段階で行われる。第 1 段階では、解析器によりコーパスを解析し複数の解析結果候補を出力する。「檜」の場合は、解析器に PET を使用し、日本語文法 Jacy に基づいて解析した。解析結果は、木構造で表された統語情報と MRS で表された意味情報からなる。第 2 段階では、作業者が解析結果候補の中から正しいものを選択することによって、各文にタグ付けを行う。場合によっては全ての候補が誤りであると判定する。この解析結果候補の選択は `[incr tsdb()]` 上で行った。候補が適切であるかの判定は、(1) 構文木の形、(2) 統語ラベル、すなわち適用された文法規則、(3) 意味表現 (MRS) の妥当性を見て行われる。

候補選択は、解析木に適用されている文法規則の差異ごとに妥当性を判断することによって行う。実際の作業では、ラベル付きの構文木と、解析結果候補間で差異がある文法規則（discriminants）が提示される。この方法は Carter[18] が提案し、Redwoods で使用されている。提示された文法規則の差異に対する正誤の判定を、正しい解析結果が一つ選択されるまで繰り返す。各文で必要となる判定回数は、解析木候補数 N に対して大抵 $\log N$ のオーダーである。作業に慣れた段階では、平均 10 語の長さの文を対象にしたタグ付けを、1 時間当たり 50 文程度行うことが可能であった。

構文木候補に正しいものが存在しない場合や、冗長な曖昧性を持つ構文木が存在する場合がある。それを改善するために文法開発者が文法拡張を行う。文法を変更した場合でも、すでにタグ付け済みの文に対して、作業者が解析木の選択を最初からやり直さずに済むように、`[incr tsdb()]` が以前の作業者の判定をもとにツリーバンクを自動的に更新することができる [19]。解析木は文法に依存しているが、文法の変更によって作業者が再選択しなければならないのは、解析木の曖昧性が増大して新たな判断が必要なときか、既存のルール／語彙項目が大幅に変更されてシステムが自動的に解析木を再構築できなくなったときである [20]。

この方法の問題点は、タグ付けできる文が解析器によって解析できたものに限られるということである。文法で実装されていない現象を含む文や、システムの制約で解析に失敗した文はタグ付けできない。つまり、ツリーバンクの構築にとって、文法が高い網羅性を持つことが重要な条件となる。

5.2 語義の付与

Lexeed の語義文と例文中の各単語には基本語彙の語義タグが付与されている。語義文と例文をあわせると延べ 119 万語からなるが、機能語と語義が 1 つしかない語を除くと多義ある基本語数は 34 万語しかない（詳しい数値は表 1 を参照）。Web ブラウザベースのアノテーションツールを活用することで一人当たり 1 日 1,500 語のペースで付与できており、既に例文中の多義がある基本語は全て付与済みである。今年度中に全語に対して付与が完了する予定である。。

コーパス	文数	語数	基本語	多義あり
語義文	81,000	690,000	317,000	218,000
例文	46,000	500,000	221,000	126,000
合計	127,000	1,190,000	538,000	344,000

表 1: 檜での語義付与対象語数

同一の単語について、作業者 5 人でタグ付与を行っている。それぞれの一致率（作業者間 2 者間の平均一致率）は 82.3%であった。これは、Senseval-2 日本語の特定語タスク（lexical sample task[21]）の一致率（86.3%）より低くなっているが、Senseval-3 の英語全語タスク（all words task [22]）の 72.5%より高い。これは、特定語タスクより全語タスクの方が曖昧性が高い語も含んでいるので難易度が高いためである。

6 おわりに

本稿では、計算機による言語理解技術の確立を目指して構築を進めてきたツリーバンク「檜」について概説した。今後は、「檜」を活用して、複合語解析や構文情報と意味情報の関係の分析などを進めていきたい。

謝辞

本研究の一部は日本電信電話（株）NTT コミュニケーション科学基礎研究所とスタンフォード大学 CSLI との共同研究により実施されたものである。

参考文献

[1] Kristina Toutanova, Christoper D. Manning, and Stephan Oepen. Parse ranking for a rich HPSG grammar. In *Proceedings of The First Workshop on Treebanks and Linguistic Theories (TLT2002)*, Sozopol, Bulgaria, 2002.

[2] Sadao Kurohashi and Makoto Nagao. Building a Japanese parsed corpus — while improving the parsing system. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, chapter 14, pp. 249–260. Kluwer Academic Publishers, 2003.

[3] EDR. Concept dictionary. Technical report, Japan Electronic Dictionary Research Institute, Ltd, April 1990.

[4] Francis Bond, 藤田早苗, 橋本力, 成山重子, Eric Nichols, 大谷朗, 田中貴秋. 精細な文法に基づいたツリーバンク「檜」の構築. In *2004-NLC-159*, pp. 91–98, 2004.

[5] 笠原要, 佐藤浩史, Francis Bond, 田中貴秋, 藤田早苗, 金杉友子, 天野昭成. 「基本語意味データベース:Lexeed」の構築. In *2004-NLC-159*, pp. 75–82, 2004.

[6] 天野成昭, 近藤公久. 日本語の語彙特性. 三省堂, 東京, 1999.

[7] 金杉友子, 笠原要, 稲子希望, 天野昭成. 単語親密度に基づく基本的語彙の選定築. In *IEICE Technical Report NLC-150*, No. 27, pp. 21–26, 2002.

[8] Carl Pollard and Ivan A. Sag. *Head Driven Phrase Structure Grammar*. University of Chicago Press, Chicago, 1994.

[9] Takao Gunji. *Japanese Phrase Structure Grammar: A Unification-Based Approach*. D. Reidel (Kluwer), Dordrecht, 1987.

[10] Kei Yoshimoto. *Tense and Aspect in Japanese and English*. Peter Lang, Frankfurt am Main, 1998.

[11] 大谷朗, 宮田高志, 松本裕治. HPSG にもとづく日本語文法について — 実装に向けての 精緻化・拡張—. 自然言語処理, Vol. 7, No. 5, pp. 19–49, 2000.

[12] Melanie Siegel and Emily M. Bender. Efficient deep processing of Japanese. In *Procedings of the 3rd Workshop on Asian Language Resources and International Standardization at the 19th International Conference on Computational Linguistics*, Taipei, 2002.

[13] Wolfgang Wahlster, editor. *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, Berlin, Germany, 2000.

[14] Matsumoto Yuji, Kitauchi, Yamasita, Hirano, Matsuda, and Asahara. *Nihongo Keitaiso Kaiseki System: Chasen*, version 2.2.1 manual edition, 2000. url=http://chasen.aist-nara.ac.jp.

[15] 増市博, 大熊智子. Lexical functional grammar に基づく実用的な日本語解析 システムの構築. 自然言語処理学会論文誌, Vol. 10, No. 2, pp. 79–109, 2003.

[16] Francis Bond, Eric Nichols, Sanae Fujita, and Takaaki Tanaka. Acquiring an ontology for a fundamental vocabulary. In *20th International Conference on Computational Linguistics: COLING-2004*, pp. 1319–1325, Geneva, 2004.

[17] Takenobu Tokunaga, Yasuhiro Syotu, Hozumi Tanaka, and Kiyoaki Shirai. Integration of heterogeneous language resources: A monolingual dictionary and a thesaurus. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, NLPRS2001*, pp. 135–142, Tokyo, 2001.

[18] David Carter. The TreeBanker: a tool for supervised training of parsed corpora. In *ACL Workshop on Computational Environments for Grammar Development and Linguistic Engineering*, Madrid, 1997. (`http://xxx.lanl.gov/abs/cmp-lg/9705008`).

[19] Stephan Oepen, Dan Flickinger, Kristina Toutanova, and Christoper D. Manning. LinGO redwoods: A rich and dynamic treebank for HPSG. In *Proceedings of The First Workshop on Treebanks and Linguistic Theories (TLT2002)*, Sozopol, Bulgaria, 2002.

[20] Stephan Oepen, Dan Flickinger, and Francis Bond. Towards holistic grammar engineering and testing — grafting treebank maintenance into the grammar revision cycle. In *Beyond Shallow Analyses — Formalisms and Satitistical Modelling for Deep Analysis (Workshop at IJCNLP-2004)*, Hainan Island, 2004. (`http://www-tsujii.is.s.u-tokyo.ac.jp/bsa/`).

[21] 白井清昭. Senseval-2 日本語辞書タスク. 自然言語処理, Vol. 10, No. 3, pp. 3–22, 2003.

[22] Benjamin Snyder and Martha Palmer. The English all-words task. In *Proceddings of Senseval-3*, pp. 41–44, Barcelona, 2004. ACL.