統辞・意味情報を付加した日本語コーパスの構築 欅ツリーバンク プロトタイプ について

Alastair Butler†♠　方 采薇‡　檜山 祥太‡　周 振‡　小菅 智也*　吉本 啓♠‡
†科学技術振興機構戦略的創造研究推進事業さきがけ　‡東北大学大学院国際文化研究科
*東北大学大学院情報科学研究科　♠東北大学高等教育開発推進センター
ajb129@hotmail.com

1　はじめに

現代日本語のテクストに対して文統辞解析情報および論理意味表示 (述語論理式) をタグ付けした欅ツリーバンク (Keyaki Treebank) を構築している (Butler et al. 2012, Butler and Yoshimoto 2012, 吉本他 2013)。通常のツリーバンクにおけるように統辞解析情報を解析器と人手によってタグ付けするが、意味評価システムによって意味表示が自動的に得られるので、文の完全な依存関係が得られるという特徴を持っている。この度プロトタイプの完成を迎え、開発の現状を研究全体の俯瞰とともに報告する。

2　開発の動機

バトラーは、従来のように複雑な文の解析をすべて統辞論で行うのでなく、複雑な処理は意味解釈のレベルで行う独自の意味理論 Scope Control Theory（SCT）を提唱し、英語や日本語の意味処理に応用しようとしてきた (Butler 2010)。SCT をインプリメントしたシステムへの入力は表層的な統辞解析を行ったもので十分であり、近い将来は自動的に行える可能性が強い。将来、日本語を含む自然言語の文の意味を自動的に抽出できるようになれば、その価値は計り知れない。このことを視野に入れて、その前提として必要な、日本語統辞情報コーパス (ツリーバンク) の構築を行ってきた。またすべての文について、SCT システムにより論理意味表示を自動生成する実験を行った。

独自のツリーバンクの開発を行う動機は、現在入手可能な日本語コーパスの多く (例えば、京都大学テキストコーパス, Kurohashi and Nagao 2003) が文節にもとづくものであり、筆者らの意味自動解析という目的には適合しないためである。例えば、日本語には関係代名詞が存在せず、通常型の関係節による修飾 (例: 昨日撮った 写真) と節の埋め込み (例: 子供が泳いでいる 写真) とが区別されていない。これは意味表示としては基本的な相違にかかわり、関係節と主節との意味関係は、前者では論理積で連結される並列関係、後者では主名詞の意味への関係節の意味の埋め込みとなる。他にも、トによる条件節 (例: トンネルを抜けると、一面の菜の花畑だった) は、同じくトが導く引用節 (例: 明日は晴れると天気予報で言っていた) と形式がまったく同一であり、自動的に区別をつけることは困難である。京都大学テキストコーパスの一部に対しては主格および直接・間接目的格と照応に関する情報が付け加えられている。しかし、文法役割に関する情報全般を捉えることは出来ない。また、情報の付加に際してインデクスを使用しており、元の文節データに変更を加えることが困難である。

本ツリーバンクは、句構造にもとづいていることの他に、句に対して機能情報をタグ付けすることを最大の特色としている。このことによって、上記で触れた、論理意味表示において並列関係（論理積）の一部として扱わねばならない内容（テ節、副詞節、関係節等）と、埋め込みとして取り扱うべき内容（不定詞、引用節、疑問節、名詞節）とを明確に区別することが可能になる。この区別は、単なる項と述語の関係を超えて文の意味理解を行うには必須のものである。

3　ツリーバンク構築の手順

テクストデータはまず形態素解析器 MeCab にかけられる。形態素解析結果は、人手による修正を経た上で、2 種類の解析器、文節解析器 CaboCha および Probabilistic Context-Free Grammar にもとづくツリーバンク解析器 (Fang et al. 2014) にかけられる。その結果は多くの誤りを含み、この統辞解析結果の人

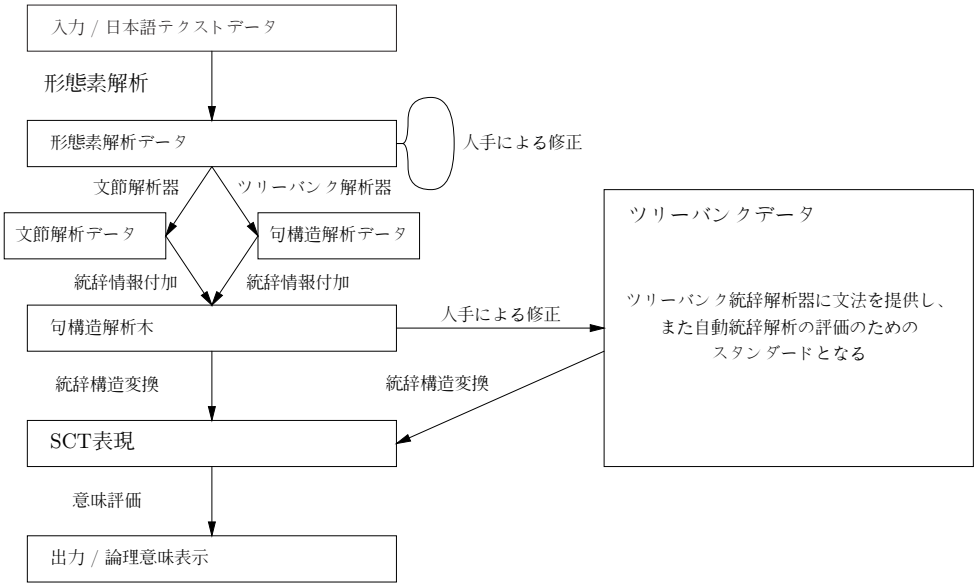

図 1: 意味表示までのパイプライン

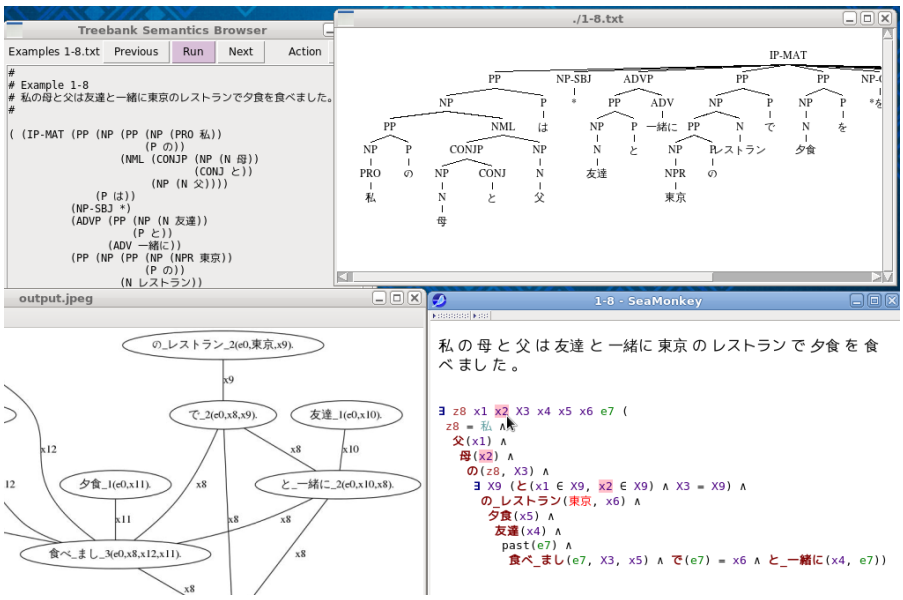

図 2: 統辞解析木と意味表示

]

手による修正が、ツリーバンクの開発における最大の課題である。こうして得られた句構造解析情報は、SCT システムへの入力として適合した SCT 表現へと自動的に変換される。SCT 表現を SCT システムが意味評価することにより、論理意味表示が出力される(図 1 を参照のこと)。

最も困難な課題である人手による統辞解析結果の修正は主としてテクストエディタを使って行うが、一部の頻出するものについては、木構造の検索・変換用ツール Tsurgeon のスクリプトを使った一括処理が可能であり、また、木の検索と表示には Tregex も利用する (Levy and Andrew 2006)。さらに、項目ごとのアノテーションのバラつきを表示するシステムを開発してアノテーションの誤りの検出・表示のために利用している。図 2 に、ツリーバンク開発のための Graphical User Interface (GUI) 環境を示す。図で、左上は文とその統辞解析のカッコ表示、右上は統辞解析の木表示であり、右下はそれから自動生成された意味表示である。左下に、意味解析から生成されるデータベース関係の視覚化結果を示す。

4 アノテーション

意味表示の生成という筆者らの最終目標にもっとも合致する統辞情報タグ付けの規約として、Annotation Manual for the Penn Historical Corpora and the PCEEC (Santorini 2010) のそれに従った。これは Penn Treebank の解析規約を修正して、ノード数の少ない平坦な統辞構造を採用したことを特色としている。動詞句、名詞句、形容詞句など句の種類を問わず、X-バー理論に従って同一の平坦な構造として解析される。句の主要部 (N, P, ADJ 等) が原則としてそれと同一のカテゴリーの句 (NP, PP, ADJP 等) を投射する。木の構造をこのように簡単にすることにより、木の検索や変換が容易になる。またこれにより、句のレベルと主要部の間の中間的なノードがスコープに干渉することを防ぎ、柔軟なスコープ包含関係の指定を可能にする。

統辞構造の提供する情報だけでは曖昧なので、名詞句、動詞句や節のノードの一部に機能情報をタグ付けして、複雑な構文が伝える意味を処理することが本ツリーバンクのもうひとつの特徴である。これにより、第 2 節で指摘した、2 種類の関係節修飾や助詞トの異なる働きを区別することが可能になる。

欅ツリーバンクのアノテーション方式のうち、独特かつ重要なものを以下に挙げる。

(1) 機能語としての連語

複数の単語が連結して、実質的に 1 つの機能語として働くものは、1 つの助詞として扱う。これは意味解析の便宜のために行う。これには

> という、として、とともに、において、について、によれば、の代わりに、やいなや、よりも、をめぐって

等が含まれる。これらの中には、'助詞+動詞+助詞' として分析される「について」のように、形態素解析に関して曖昧なものを含む。

(2) モーダル助動詞としての連語

複数の単語が連結して 1 つのモーダル的機能を果たすものについては、1 つのモーダル助動詞 (MD) としてラベル付けする。これには以下の連語が含まれる。

> うとする、かもしれない、ことだ、ざるをえない、てはいけない、なくてはならない、てもいい、にちがいない、訳にはいかない

「うとする」のような例は 1 語として扱われるものの、MD+AX+P+VB のように表記して、必要が生じた場合には形態的な内部構成に関する情報が得られるようにしてある。

(3) 文法機能の明示

主語や目的語として働く後置詞句 PP の直後に NP-SBJ, NP-OB1, または NP-OB2 のノードを付加し、文法機能を明示する。これは一つには格助詞「が/を/に」が示す文法役割が曖昧なためである。また、係助詞や副助詞が付加されて格表示がなされない場合も、これを利用して文法役割の明示を行う。

(4) ゼロ代名詞の明示

動詞の必須格としての主語や目的語が文中で表現されていない場合の多くについて、それらをゼロ代名詞 *pro* として明示する。言語理論によってはあまりに多くのゼロ代名詞を仮定するために実際の文解析やツリーバンク構築の用をなさなくなる場合も見受けられるが、そのようなことは避けられる。本ツリーバンクの方式では、無主語文も認める。また、明示されない主語等と同一指示の名詞句が文中に存在してコントロール関係にある場合、ゼロ代名詞としてのタギングは行わない。これは、SCT のスコープ操作により、意味論的処理が同一指示関係を補完するためである。

(5) スコープの扱い

複数のスコープの中で、出現順位の早いものほど広いスコープを持つというディフォールト規則を設定

し、これに反するものは明示的情報を与えることによって、柔軟にスコープ包含関係を指定する。すなわち、品詞等の文法的タイプごとにディフォールトのスコープを定めた上で、必要に応じて HIGHEST, HIGH, LOW, LOWEST のいずれかの機能タグを付加する。

(6) インデクスの不使用

非境界依存 (unbounded dependency) のような複雑な構文を意味理解を目的として統辞解析する場合、通常はインデクス付けを行うが、本ツリーバンクでは不要である。これは SCT の意味評価によって、同一指示関係が自動的に得られるからである。結果として、意味表示の形で文の構成要素間の依存関係が全て得られるにもかかわらず、ツリーバンク開発のコストは現実的な範囲内に抑えることが出来る。また、SCT システムへの入力は表層的な解析結果で済むので、意味表示出力までの処理の全過程の自動化も視野に入れることが出来る。ただし、外置、数量詞遊離および主要部内在型関係節の構文については、インデクスにより語句をそれ以外の場所と関係づける必要がある。

5 本ツリーバンクの意義

現状の一般のツリーバンクのアノテーションは、表層的な統辞解析情報の提供にとどまる。ツリーバンクも含め、現在世界に存在するコーパスは、対象とする言語にかかわらず語句間の**共起 (co-occurrence)** に関する手掛かりを与えるにすぎず、それらの間の**依存関係 (dependency)** について正確な情報を与えることは出来ない。欅ツリーバンクは、意味表示の形で文の依存関係についての情報を全て提供する最初のコーパスである。

6 開発の現状と今後の計画

平成 26 年 1 月末の時点で、現代日本語の書かれたテクストの 13,675 文に統辞・意味解析アノテーションを施した。同年 3 月末の時点で、約 15,000 文のアノテーションを含む欅ツリーバンク プロトタイプ が完成する予定である。現状の内訳は以下の通りである。

1	ブログ記事	217
2	法律条文	484
3	新聞記事	1,600
4	電話会話	1,177
5	日本語文法教科書	7,733
6	ウィキペディア記事	2,464
	合計	13,675

これらのうち、6 は英語の対訳が存在するものが多い。また、5 は、日本語の基本的な構文や文法語を網羅した益岡・田窪 (1992) の約 1,300 個の文にすべてタグ付けしたもので、欅ツリーバンクのアノテーション方式を知るのに便利である。さらに、アノテーション作業用マニュアルを編纂中である。

今後は、平成 28 年度末を目標として、日本語の書き言葉の文 4 万文に統辞・意味解析情報をタグ付けした欅ツリーバンクを完成させる予定である。また、これとは別に、対話データのタグ付けも計画している。開発したツリーバンクのうち公開可能なものは全て以下のサイトで配布する予定である。

http://www.compling.jp/keyaki/

引用文献

Butler, A. 2010. *The Semantics of Grammatical Dependencies.* Emerald.

Butler, A., et al. 2012. Treebank Annotation for Formal Semantics Research, *Proceedings of the Ninth International Workshop on Logic and Engineering of Natural Language Semantics*, pp. 210-222. JSAI International Symposia on AI.

Butler, A., and K. Yoshimoto. 2012. Banking Meaning Representations from Treebanks, *Linguistic Issues in Language Technology* 7.

Fang, T., A. Butler and K. Yoshimoto. 2014. Parsing Japanese with a PCFG Treebank Grammar. 『言語処理学会第 20 回年次大会発表論文集』.

Kurohashi, S. and M. Nagao. (2003) Building a Japanese Parsed Corpus – While Improving the Parsing System, A. Abeillé, ed., *Treebanks: Building and Using Parsed Corpora*, chap. 14. Kluwer Academic Publishers.

Levy, R. and G. Andrew. 2006. Tregex and Tsurgeon: Tools for Querying and Manipulating Tree Data Structures. In 5th International Conference on Language Resources and Evaluation.

益岡隆志・田窪行則 (1992) 『基礎日本語文法・改訂版』くろしお出版.

吉本啓, 他. 2013.「日本語ツリーバンクのアノテーション方針」『言語処理学会第 19 回年次大会発表論文集』, pp. 924-927.