研　究　報　告　書

「自然言語テクストの高精度で頑強な意味解析とその応用」

研究タイプ：通常型
研究期間：　平成 22 年 10 月～平成 26 年 3 月
研　究　者：　アラステア・バトラー

1.　研究のねらい

ヒトが日常的に使用している自然言語の意味や思考内容を表す手段として、述語論理が確立され、またその様々な拡張が提案されている。自然言語の文をコンピュータによって自動的に述語論理式へと変換する方法が開発されれば、そのことは文の意味の中核的部分をコンピュータに理解させたことを意味し、それによって得られる恩恵ははかり知れない。特に 1970 年代にモンタギューが文の統辞（構文）構造を用いて意味解釈を行う手法を確立して以来、生成文法や句構造文法にもとづく統辞解析を利用して文の意味表示を得ようとする研究が盛んになった。しかし、少数のサンプルを実験的環境で意味処理することには成功しても、現実に存在する大量の言語データを処理するレベルには至っていない。これは、統辞解析を行うための規則がぼう大な数にのぼり、それらの間のコントロールを行うことに失敗したことが主たる原因である。

私は、従来のように複雑な文の構文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（スコープ制御理論, SCT）を編み出し、英語や日本語の意味処理に応用しようとしてきた。本研究では、それまでの研究成果を大量の無制約日本語テクストの処理に応用する研究を行ってきた。SCT をインプリメントしたシステムへの入力は簡単な（表層的な）統辞解析を行ったもので十分であり、これは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。

2.　研究成果

（1）概要

日常生活で使われている無制約の自然言語処理テクストから高精度の意味表示を自動的に得るための頑強な方法を開発することを目的として研究を行った。得られた意味表示は、知識ベース構築等に利用するために、推論に適した述語論理にもとづいている。研究の基礎となったのは、動態意味論（Dynamic Semantics）を拡張発展させて自然言語の文の意味解析を行うための形式言語理論であるスコープ制御理論（Scope Control Theory; SCT）である。SCT を実装した意味評価システムに対して英語や日本語の統辞解析器による表層統語解析結果を入力することによって、正確で深く、カバー率も高い意味表示の出力を可能にした。

本研究では、通常言語処理で行われている統辞解析結果にもとづいて英文および日本語文テクストから意味表示を得る研究を行ったが、その前提として必要な日本語統辞構造データ（ツリーバンク）の構築も行った。その結果として、研究終了までに、約 1 万 5 千文からなる日本語テクストに統辞および意味解析情報を付加したツリーバンクを完成させる予定である。

これによって、提案する方法により、生の日本語テクストから意味表示までに至るまでの「パイプライン」を構築できたことになる。

（2）詳細

1. 日本語ツリーバンクの構築

より正確でよりカバー率の高い意味表示を大量の自然言語テクストから得るためには、まず適切な文の統辞解析を得ることが不可欠である。本研究における最大の課題となったのは、いかにして大量の良質な文解析結果を得るかということであった。そのため、文に対して統辞解析結果をタグ付けしたコーパスであるツリーバンクの構築が重要になった。

日本語についてはすでに、ある種のツリーバンクと言える京都大学テキストコーパス（Kurohashi and Nagao 2003）が存在する。これは 1996 年から 2003 年までの毎日新聞の記事 4 万文にもとづいている。同コーパスは、日本語文解析研究の基礎資料として使われている。

しかし、上記のコーパスは、文の解析の基本を文節に置いているため、原理的な問題を抱えている。例えば、日本語には関係代名詞が無いため、意味表示という観点からは大きく異なる関係節と名詞節として埋め込まれた節とを区別する手掛りが存在しない。別の例として、条件表現を挙げることが出来る。条件表現は通常、後置詞ト、バ、ナラ等によって表示される。このうち、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）と形式が同じであり、自動的に区別をつけることは困難である。

この京都大学テキストコーパスの欠点を補うためにさまざまな試みが行われている。たとえば、NAIST Text Corpus (Iida et al. 2007) では京都大学テキストコーパスに対し、格および照応情報が付け加えられている。しかし、格情報は表層的なもので、しかも主格、直接目的格および間接目的格の 3 つに限定されているため、文法役割に関する情報を一般的に捉えることは出来ない。しかも、情報の付加に際してインデクスを使用しているので、元の文節データに変更を加えることが困難である。本研究では当初、意味表示構築のための入力として京都大学テキストコーパスに手を加えて利用する予定であったが、以上の理由からそれが不可能であると判断し、文節ではなく句構造（phrase structure）にもとづくコーパスを初めから構築する必要があるとの認識に至った。

本研究で開発を行ってきたコーパスは、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、上記でトが導く節の解釈に関連して触れた、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。また、句の種類にかかわらず類似した構造を持っているため、木の検索と加工が容易である。

平成 26 年 3 月末の研究終了時までに日本語ツリーバンク（欅ツリーバンク）プロトタイプ 1 万 5 千文を完成させる予定である。本ツリーバンクは、様々な言語研究のためのリソ

ースとして永く利用することが出来る。試行錯誤を通じてアノテーション方式を整備し、様々な種類の言語学的情報をツリーバンクから簡単かつ正確に得ることが可能になっている。上記の例について言えば、関係節と名詞埋め込み節、また助詞トによって導かれる条件節と引用節との区別が当該ツリーバンクのアノテーションによって正確に行われ、また適切な意味表示がそれぞれ作られる。さらに、人手によって構築した本ツリーバンクにもとづいて、機械学習の方法を適用することにより、自動文解析プログラム(パーザー)を作成することも視野に入れることが出来、これが可能になればツリーバンクの構築は飛躍的に容易になる。

2. 意味表示の生成

インプリメントした意味解析システムは、統辞解析結果をSCT形式に変換し、さらにその意味評価を行うことを通じて、述語論理にもとづく意味表示を出力する。SCT の利点は、入力とする統辞解析情報(句構造)の内部において、他の理論とは異なって同一指示の表現間でインデクス付けが不要なことである。SCT システムは入力の意味評価を通じて、様々な複雑な構文の解析を行うことが出来る。このことによって、ツリーバンク構築の際の手間が著しく軽減される。日本語のテクストの入力から意味表示の出力に至るまでのパイプラインを図1に示す。

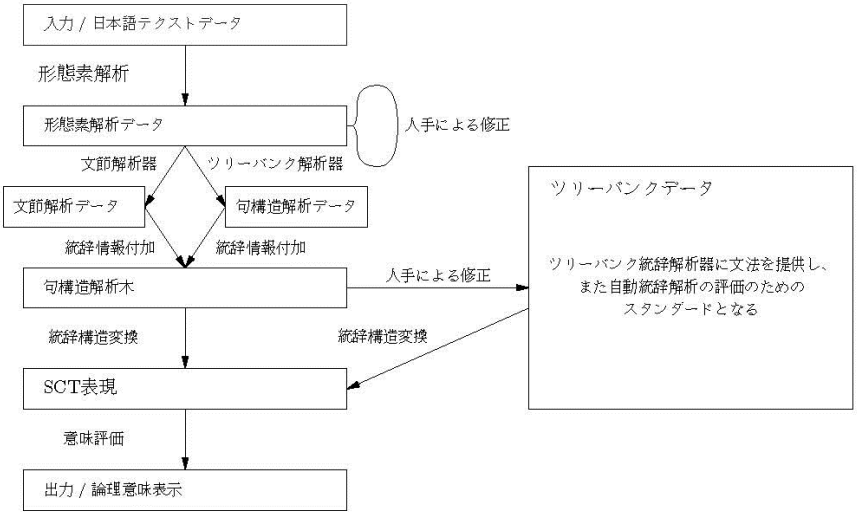

図 1: 意味表示出力までのパイプライン

SCT の理論面の研究は長期間にわたって行われたが、本研究期間では特に、高速で頑強なインプリメンテーションを可能にしたことが成果である。そのために、日本語に特有の事象について個別に対応することが必要であった。意味解析システムのドキュメンテーションとして、*Linguistic Expressions and Semantic Processing: A Practical Approach* を執筆した。

3 意味表示の使用

開発したシステムを用いて、統辞解析タグ付けを行ったすべての文から述語論理にもとづく

意味表示生成を行うことが出来た。今後の課題は、得られた意味表示をオントロジー的情報に変換して、データからの推論に利用することである。

研究成果の応用研究の 1 つとして、意味解析結果を視覚化して外国語教育に生かす方法の検討を行った。これは特に、日本語の助詞の機能についての日本語学習者の理解を助けるのに利用することが出来る。このために、Tcl/Tk を用いて、Graphical User Interface (GUI) を開発した。この GUI は同時に、アノテーションおよび意味表示の質や正確性のチェックに利用することが出来る。

GUI 環境のスクリーンショットを図 2 に示す。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。異なる種類の変項が色分けされ、さらにマウスにより矢印を当てることでハイライトされる。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。

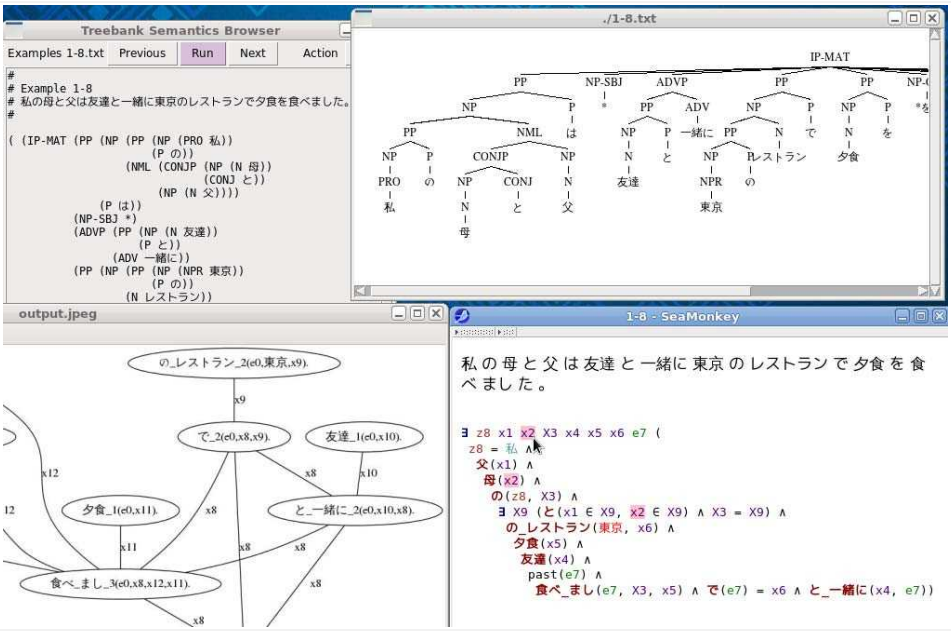

図 2: GUI のスクリーンショット

3. 今後の展開

アノテーションの方式は今や安定した段階を迎えている。過去 6 か月間で 2 倍のサイズとなり、構築にますます加速度のつく傾向が続くものと期待される。本研究で確立されたアノテーション体系を利用して、NTT コミュニケーション科学基礎研究所と 10 万文よりなる話し言葉データのツリーバンクを共同構築する協定を結んでいる。

研究を通じて確立された意味表示方式は、データベース・エントリーを生成するための情報抽出への応用が可能である。現状の意味表示には、すでにソートされたオントロジーの情報（“entity”, “group”, “attrib”, “degree”, “time”, “event”, “situation” 等）が含まれており、これを基礎として、一般的推論を可能とする情報の抽出につなげたい。

主として外国語教育の目的のために開発された GUI ソフトウェアは、文法語（助詞、助動詞等）の働きを視覚化という形で具体的に示せるという点でユニークな意義を持つ。次の研究

段階では、実際に教室での外国語教育における教材として使用し、その効果を確認したい。

構築されたツリーバンクは、コーパス言語学で必要とされるスケールで、言語を深いレベルで検索したり比較することを可能にする。この特性を利用して、アジアの諸言語の特性・類型をツリーバンクを用いて研究する共同プロジェクトを国立国語研究所とともに近年中に開始する予定である。

研究のもう１つの重要な発展の方向として、「深い」変換をともなう機械翻訳への応用が考えられる。このためには、意味表示から自然言語文への生成をいかに行うかが課題となる。しかし、本研究により自然言語文と論理意味表示との対（ペア）が大量にデータベース化されたことは今後の研究に対して基盤を与えることになるだろう。

4. 評価

（1）自己評価

自然言語のテクストから質の高い意味表示を得るための完全なシステムを構築したという点で、本研究の主要目的は達することが出来た。本研究の成果は、より高度の文解析や機械翻訳の基礎を与えるという意義がある。また、深い意味情報を十分なスケールで提供できるという点で、コーパスによる言語研究の新時代を開くものである。

他方、意味解析実験への入力としての統辞解析データベースを自前で構築する必要のあることが途中で分かり、これに多大の労力を費やした結果、意味表示からオントロジー情報を抽出するというもう１つの課題については十分な検討を行うことが出来なかった。しかし、そのための道筋は付けることが出来たので、次の研究段階で行いたい。

（2）研究総括評価（本研究課題について、研究期間中に実施された、年２回の領域会議での評価フィードバックを踏まえつつ、以下の通り、事後評価を行った）。

構文解析と直結したスコープ制御理論に基づく意味解析を行い、記述論理に基づく知識表現につなげようという研究である。日常生活で使われている無制約の自然言語処理テクストから、高精度の意味表示を自動的に得るための頑強な方法を開発することが目的である。

研究期間中、自然言語の文の意味解析を行うための形式言語理論のスコープ制御理論(Scope Control Theory)を実装した意味評価システムに対して、英語や日本語の統辞解析器による表層統語解析結果を入力することによって、正確で深く、カバー率も高い意味表示の出力を可能にした。また、日本語テクストに統辞および意味解析情報を付加したツリーバンクを完成させる予定であり、生の日本語テクストから意味表示までに至るまでの「パイプライン」を構築できたことになる。これらは公開予定であり、主要目的を達成していると考える。

5. 主な研究成果リスト

（1）論文（原著論文）発表

1. Butler, Alastair, Ruriko Otomo, Zhen Zhou and Kei Yoshimoto. 2013. Treebank Annotation for Formal Semantics Research. In Motomura et al., eds., New Frontiers in Artificial Intelligence: JSAI-isAI 2012 Workshops, Lecture Notes in Computer Science, Volume 7856,

pages 25–40, Berlin Heidelberg: Springer.
2. Butler, Alastair and Kei Yoshimoto. 2012. Banking Meaning Representations from Treebanks. Linguistic Issues in Language Technology – LiLT 7(6): 1–22.
3. Butler, Alastair and Kei Yoshimoto. 2012. Towards a Self-selective and Selfhealing Evaluation. In Okumura et al., eds., New Frontiers in Artificial Intelligence: JSAI-isAI 2011 Workshops, Lecture Notes in Computer Science, Volume 7258, pages 96–109, Berlin Heidelberg: Springer.
4. Butler, Alastair and Kei Yoshimoto. 2011. Interpreting Japanese Dependency Structure. In Onada et al., eds., New Frontiers in Artificial Intelligence: JSAIisAI 2010 Workshops, Lecture Notes in Computer Science, Volume 6797, pp. 30–44, Berlin Heidelberg: Springer.
5. Yoshimoto, Kei, Zhen Zhou, Tomoya Kosuge, Ruriko Otomo and Alastair Butler. 2013.「日本語ツリーバンクのアノテーション方針」. In『言語処理学会第 19 回年次大会発表論文集』Proceedings of the Nineteenth Annual Meeting of the Association of Natural Language Processing, pages 924–927, The Association of Natural Language Processing.

（2）特許出願

研究期間累積件数：0 件

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

・ Butler, Alastair. Linguistic Expressions and Semantic Processing: A Practical Approach. 200 pp.

・ Zhou, Zhen, Alastair Butler and Kei Yoshimoto. 2013.「中国語統語解析木の形式変換及びその応用に関する研究—Penn Chinese Treebank (3.0) を対象として—」.『言語処理学会第 19 回年次大会発表論文集』Proceedings of the Nineteenth Annual Meeting of the Association of Natural Language Processing, pages 920–923, The Association of Natural Language Processing.

・ Butler, Alastair, Zhen Zhou and Kei Yoshimoto. 2012. Problems for successful bunsetsu based parsing and some solutions.『言語処理学会第 18 回年次大会発表論文集』Proceedings of the Eighteenth Annual Meeting of the Association of Natural Language Processing, pages 951–954, The Association of Natural Language Processing.

・ Zhou, Zhen, Alastair Butler and Kei Yoshimoto. 2012. Combining and splitting bunsetsu of the Kyoto Text Corpus.『言語処理学会第 18 回年次大会発表論文集』Proceedings of the Eighteenth Annual Meeting of the Association of Natural Language Processing, pages 381–384, The Association of Natural Language Processing.

・ Butler, Alastair, Zhen Zhou, Tomoko Hotta, Su Zhang and Kei Yoshimoto. 2011. Development of Corpora Tagged with High-Precision Semantic Information.『言語処理学会第 17 回年次大会発表論文集』Proceedings of the Seventeeth Annual Meeting of the Association of Natural Language Processing, pages 713–716, The Association of Natural Language Processing.

・ Miyao, Yusuke, Alastair Butler, Kei Yoshimoto and Jun'ichi Tsujii. 2010. A Modular Architecture for the Wide-Coverage Translation of Natural Language Texts into Predicate Logic Formulas. In Proceedings of PACLIC 24: the 24th Pacific Asia Conference on Language, Information and Computation, pages 481–488, Sendai, Japan.
・ Butler, Alastair, Yusuke Miyao, Kei Yoshimoto and Jun'ichi Tsujii. 2010. A Constrained Semantics for Parsed English.『言語処理学会第 15 回年次大会発表論文集』Proceedings of the Sixteenth Annual Meeting of the Association of Natural Language Processing, pages 836–839.