3B1-01

セミオティックベースを使ったテキスト処理アルゴリズム

Algorithms for Text Processing with the Semiotic Base

伊藤紀子[*1] 杉本徹[*1] 高橋祐介[*1] 小林一郎[*1 *2]
Noriko ITO Toru SUGIMOTO Yusuke TAKAHASHI Ichiro KOBAYASHI

[*1] 理化学研究所 脳科学総合研究センター
Brain Science Institute, RIKEN

[*2] お茶の水女子大学 理学部
Faculty of Science, Ochanomizu University

We propose a Japanese text understanding and generation system, combining the existing parser and dictionary with the linguistic resource that we developed based on the systemic functional linguistics. This paper focuses on the text understanding process, which starts with morphological and dependency analysis by the non-SFL-based existing parser, followed by looking up the dictionary to enrich the input for SFL-based analysis. After mapping the pre-processing results onto systemic features, the path identification of selected features and unification are conducted with reference to the linguistic resource represented in the system networks. Consequently, graphological, lexicogrammatical, semantic and conceptual annotations of a given text are produced. We also give a sketch of the text generation process.

1. はじめに

日常言語コンピューティングの実現に向けて我々の研究の主要な目的の１つは、選択体系機能言語理論（systemic functional linguistics、以下、SFL）で提唱されている言語モデルに基づいた自然言語処理システムを実装することにある。我々がこの言語理論を採用した理由は、SFL が言語体系を包括的に記述することを目標とし、言語使用を言語が使われる状況とともにモデル化する統一的な方法を提供してくれるからである [Halliday 94]。SFL はテキスト生成システムの基礎としては多く利用されているが [Matthiessen 91, Fawcett 93]、SFL をベースとしたテキスト理解システムの開発は非常に少ない [O'Donnell 94]。

[伊藤 01] は、SFL の言語モデルをベースにした言語知識をデータベース化したセミオティックベース（Semiotic Base、以下、SB）の構造と、それを利用したテキスト理解・生成の手法の概略を提案した。[Sugimoto 02] は、SB を対話管理システムに組み込むことによって、どのように知的エージェントが現在の対話の状況を同定し、それに合わせて行動するかを提案した。我々は、彼らのアイデアを拡張・詳細化することによって、日本語のテキスト理解・生成システムを実装した。

本稿では、我々が実装した SB 内の様々な言語資源を活用するテキスト処理システムについて説明したい。始めに、SB の構造とそこに含まれる言語資源について述べる。次に、SB を参照することによってどのようにテキスト理解が進んでいくのかを具体例を挙げて説明する。テキスト生成については、紙面の都合上、処理の概略を述べるだけにする。

2. 言語資源：セミオティックベース

[Sugimoto 02] によると、SB は表 1に示したように4つの主要コンポーネントと2つの補助コンポーネントから成る。

連絡先：伊藤紀子，理研脳センター言語知能システム研究チーム，〒351-0198 埼玉県和光市広沢 2-1，Tel: 048-462-1111(ext.7408)，Fax: 048-467-6450，itoh@brain.riken.go.jp

主要コンポーネント	コンテクストベース	状況ベース
		ステージベース
		概念辞書
	意味ベース	
	語彙文法ベース	
	表現ベース	
補助コンポーネント	電子化辞書	汎用辞書
		状況特化辞書
	コーパスベース	

表 1: セミオティックベースの構造

SB は、SFL で提案されている言語体系をモデル化する際の指針に則って設計されている。その指針の1つが状況と言語の体系を層的に捉える点である。これに合わせて、SB の主要コンポーネントは、コンテクストベース（Context Base、以下、CB）、意味ベース（Meaning Base、以下、MB）、語彙文法ベース（Wording Base、以下、WB）、表現ベース（Expression Base、以下、EB）となっている。CB には、言葉がやりとりされる状況を分類するための特徴の体系と、状況タイプ毎にそこでやりとりされる言葉の意味的な特徴に関する記述が納められている。MB には、言葉の意味特徴の体系と、特定の状況下で当該の意味特徴を具現するための語彙文法的な手段に関する制約が納められている。WB には、言葉の語彙的、文法的な特徴の体系と、特定の状況下で当該の語彙文法特徴を具現するための表現的な手段に関する制約が納められている。EB は、現在は音声には対応しておらず、文字に関する特徴が納められており、句読法や、当該の語彙項目をどの表記体系（漢字、仮名、英数字など）で具現するか、文字列をどのように画面表示するかを HTML などで指定する際に利用することに主眼が置かれている。

ここで採用されている SFL に特有のもう1つの指針は、状況や言語の特徴と、特徴を選択していくことによって特定の構造を形成したり、ある特徴を選択したりする際に考慮すべき様々な条件を選択体系網（system network）と具現規則（realization statement）という一定の形式で記述することである。図1は WB 内の選択体系網と関連する具現規則の例である。

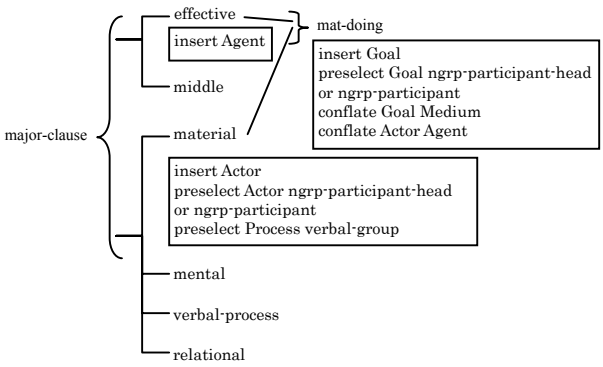

図 1:選択体系網と具現規則の例

選択体系網とは、1つ以上の選択体系(system)から構成され、選択体系の選択肢は言語的な特徴(feature)で表現されている。それぞれの選択体系からは、特徴を1つだけ選ぶことができる。選択された特徴は、他の選択体系の入力条件となることがある。例えば、図1では、major-clause は2つの選択体系への入力条件である。もしこの特徴が選ばれたら、最初の選択体系からは effective と middle のいずれかが選ばれなければならず、また、2つめの選択体系からは material, mental, verbal-process, relational の中から1つ選ばれなければならない。もし最初の選択体系から effective、二つ目の選択体系から material が選ばれたなら、さらに次の選択体系に入る。この選択体系には mat-doing という特徴しかないので、自動的に mat-doing が選ばれる。

処理中に選択された特徴はすべて記録され、特徴選択パス(feature selection path)として、インスタンス構造に含まれる。ここで我々は、インスタンス構造をユニットと呼ばれるノードを持つ木構造、ユニットを選択体系網のルートからの選択の道筋を示す特徴選択パスと親ユニットに対して当該ユニットが担っている SFL ロールから構成されるものと考える。

選択体系内の特徴には具現規則が付与されているものがある。具現規則は、当該の特徴を含んだインスタンス構造を指定するために使われる。具現規則は、具現オペレータ(realization operator)と具現オペランド(realization operand)で指定される。具現オペレータには、例えば、insert, preselect, conflate, partition, order などがある。具現オペランドは、Agent などのロール、又は ngrp-participant などの特徴である。図1で言うと、effective という特徴に付与されている insert Agent という具現規則は、この特徴が選ばれたら、インスタンス構造に Agent というロールを持つユニットを挿入しなさい、という意味である。preselect Process verbal-group(material に付与)は、Process というロールを持つユニットは verbal-group という特徴を持っていなければならないことを意味している。conflate Goal Medium (mat-doing に付与)は、Goal というロールを持つユニットと Medium というロールを持つユニットを単一化する、という意味である。

我々は CB を状況ベース(Situation Base)、ステージベース(Stage Base)、概念辞書(Concept Repository、以下、CR)という3つのコンポーネントに分割した。[1]状況ベースには、言葉がやりとりされる状況のタイプとそれを分類するための特徴(フィールド、テナー、モード)の選択体系網が入っている。ステージベースには、状況タイプ毎にそこでのやりとりの進行のパターンと言葉の意味的な特徴に関する記述が納められている。概念辞書には、概念がフレーム形式で定義されており、その概念に関連する SFL 的な意味・語彙文法特徴とロールと EDR 概念識別子も書かれている。表 2は CR のレコードの例である。

見出し概念名	writing	
概念区分	class	
EDR 概念識別子	0fe07c	
MB 意味特徴	fg-creative	
WB 語彙文法特徴	creative	
上位概念名	domain-action	
	スロット名	*スロット値タイプ*	*SFL ロール*
1	agent	agent	Actor
2	object	document	Goal
3	instrument	word-processor	Means

表 2: 概念辞書レコードの例

CR のレコードは状況タイプ毎に用意されており、この点で、EDR 概念辞書の概念体系が状況に依存しない、より汎用的なタクソノミー及びシソーラスとして設計されているのとは異なる[EDR 01]。

上述の4つの主要コンポーネントに加えて、SB はタグ付きコーパスと機械可読辞書(Machine Readable Dictionaries)を備えている。コーパスには、テキストに対して CB, MB, WB, EB, MRD を参照してタグ付けした結果が入っている。辞書については、汎用辞書(General Dictionary、以下、GD)と状況特化辞書(Situation-specific Dictionary、以下、SSD)の2種類用意した。両者共に、見出しとなっている語彙項目に関する通常の辞書情報に加えて、SFL 的な意味・語彙文法特徴とロールと EDR 概念識別子と概念関係子が書かれている。両者の違いは、状況特化辞書には見出し項目が使われる状況に関する情報、見出し項目がその状況下で使われた場合にのみ観察される言語的な特徴および概念名が含まれているが、GD にはそのような情報が入っていないことである。そのため実際の処理では、GD は主に処理すべきテキストの状況が分かっていない、又は該当する状況タイプが見つからなかった場合に利用され、状況特化辞書は状況が判明しているときに利用される、という具合に使い分けがなされている。表 3は GD のレコードの例である。

見出し語	書く	
カナ	カク	
EDR 品詞	JVE	
EDR 概念識別子	0fe07c	
MB 意味特徴	fg-creative	
WB 語彙文法特徴	creative&lg-concrete or creative&lg-abstract	
見出し語の SFL ロール	Process&Predicator	
	EDR 概念関係子	*子ユニットの SFL ロール*
1	agent	Actor&Agent
2	object	Goal&Medium
3	instrument	Means

表 3: 汎用辞書レコードの例

[1] [Sugimoto 02]で提案されているコンテクストベースのコンテンツは、本稿で言うところの状況ベースに収められている。ステージベースと概念辞書については、概ね、彼らの言うプランライブラリ内のインタラクションプランと知識ベース内の概念フレームにそれぞれ相当する。状況ベースの詳細と検討課題については、[高橋 02]を参照のこと。

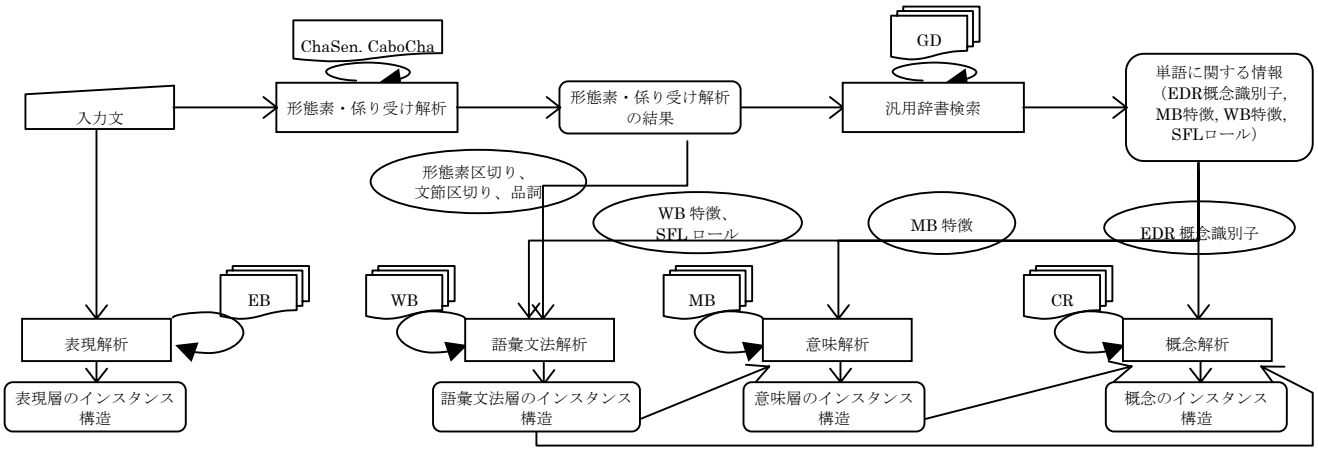

図 2: テキスト理解処理の流れ

3. テキスト処理

3.1 テキスト理解

図 2 は SB を使ったテキスト理解処理の流れを示している。このセクションでは、表4に示したような例文に対する我々のシステムからの出力を得るまでの処理を、図に示した段階ごとに簡単に説明していきたい。

形態素・係り受け解析 [工藤 02] と汎用辞書検索が行われた後、EDR 日本語動詞パターン副辞書、日本語共起辞書、概念辞書を参照することによって、文節の各ペアに対して EDR 概念関係子が割り当てられる。そして、GD レコードの対応表を参照することによって各文節に対する SFL ロールの候補が得られる。

語彙文法解析と意味解析は [O'Donnell 94] の WAG システミックパーザーで用いられているアイデアをベースにしている。彼のアイデアとは部分構造（partial-structure）というデータ構造とボトムアップ・チャート解析を用いるものである。部分構造は2種類あり、1つはリンク用部分構造（linking partial-structure）と呼ばれ、親ユニットと子ユニットの可能な組合せを表したもので、MB と WB 内にある insert, conflate, preselect という具現オペレータを含んだ具現規則からコンパイルされる。もう1つは順序付け用部分構造（ordering partial-structure）と呼ばれ、子ユニット間の順序関係を指定したもので、insert, conflate, order という具現オペレータを含んだ具現規則からコンパイルされる。このようなデータ構造とユニットの単一化アルゴリズムを用いることによって、形態素ランクからボトムアップ式にインスタンス構造が構築される。

我々は [O'Donnell 94] で採用されているパージングの方法をいくつか改良した。第一に、我々はインスタンス構造に情報を追加したり不適切な解釈を排除したりするために、前処理の結果を統合して利用している。まず新規の不活性弧がチャートに追加されると、弧のルートユニットへ単語情報が取り込まれる。続いて、より大きな活性弧を構築するために、不活性弧のルートユニットと部分構造又は活性弧に含まれるユニットの間で単一化が行われると、通常の選択パスの単一化 [Kasper 87] に加えて、特徴・ロール情報の単一化と、現在のインスタンス構造と前処理結果で得られた係り受け構造との間の一貫性のチェックが行われる。

2つ目の改良点は、語彙文法ユニットと意味ユニットの間の関係を宣言的に表現する層間リンク用部分構造（inter-stratal linking partial-structure）を導入したことである。これは語彙文法層から意味層へインスタンス構造をマッピングする際に使われる。3点目は、co-selection constraints とデフォルトの特徴選択を処理できるようにするために、前向き連鎖推論も用いていることである。

最後に、概念解析では入力文の概念内容を表現するインスタンス概念フレームを構築する。インスタンス概念のスロットは他のインスタンス概念で埋められており、それらは入力文のどこかの部分に対応している。スロットのフィラーに関するタイプ制約については、CR の概念体系だけでなく一般的な概念に関するより豊富な情報を備えた EDR の概念体系をも参照して、チェックされる。

3.2 テキスト生成

始めに、入力文に理解処理を施した結果として得られたインスタンス構造を基に、ステージベースを参照して、これから生成するテキストの概念情報を決定する。ステージベース内には、対話の状態遷移のフェーズ毎に典型的な概念情報と言語特徴が記載されているので、理解したテキストの内容を対話の状態遷移の中に位置づけることができる。その次のフェーズにある概念情報を基に、生成テキストの概念インスタンス構造が生成される。これは、テキスト理解の結果出力される概念のインスタンス構造と同様の形式のものである。

次に、統語構造を決定する。すなわち、ステージベース内にある対話の状態遷移のフェーズ毎に付与された言語特徴を基に、MB、WB を参照して、意味層と語彙文法層のインスタンス構造を生成する。

最後に、MRD を利用して語彙項目を決める。その際には、まず SSD を検索し、該当するレコードがない場合には GD を用いる。このうち、SSD を用いた語彙項目の決定は、次のように行われる。語彙項目の決まっていない語彙文法層のユニットに対応する概念ノードに付与された概念名をすべてピックアップし、概念名に基づいて SSD を検索する。その際には、SSD のレコードに記載されている状況情報（状況タイプ、フィールド、テナー、モードの特徴）を参照して、現在の発話の状況情報に合うレコードだけを抽出する。1つの概念名に対して複数のレコードが抽出された場合には、レコードに記載されている状況内頻度情報を参照して、最も数値の高いものを選ぶ。このようにして抽出されたレコードの見出し語を語彙文法層のユニットに割り当てると、語彙項目の生起順序が決まり、生成テキストの文字列が確定する。

前処理	形態素・係り受け解析		私	は	クリスマス	会	の	招待	状	を	書き	たい	。
			0 4D		1 2D			2 4D			4 −1O		
	GD検索	CID	0e7e95, 2dc301	−	3ceda8		−	0e3e07		−	0e910d, 0e910f	2621c8	−
		CRL	agent		scene			object			−		−
語彙文法解析	clause simplex rank		clause clause−simplex major−clause independent−clause non−conjunct cls−neutral no−circumstance material mat−doing creative lg−concrete effective thematic−agent−subject operative−voice thematic relative−theme unmarked−theme explicit−topical−theme free cls−positive cls−hon−default cls−informal sbj−explicit md−interactant md−non−addressee md−speaker non−addressee−option non−indicative cls−optative unkeyed …									−
	group complex rank		−		Goal/ Medium/ Complement/ Rheme1			−			−		−
			−		ngrp−complex ngrp−projection ngrp−hypotactic ngrp−matter…			−			−		−
	group simplex rank		Actor/ Agent/ Subject/ TopicalTheme		Qualifier/ Modifier1			Thing/ Head1			Process/ Predicator/ Rheme3		−
			ngrp−simplex specific determinative nominal−head thematic−ngrp personal pronominal ngrp−participant−head general−theme−marker speaker ngrp−part−ga non−qualified …		ngrp−simplex non−specific nominal ngrp−abstract non−thematic−ngrp ngrp−qualifier noun−qualifier non−qualified …			ngrp−simplex non−thematic−ngrp nominal−head non−specific ngrp−participant−head nominal ngrp−part−o ngrp−concrete qualified …			vgrp−simplex zgrp−positive temporal non−past modal modulation redinessinclination optative option−tai non−causative active−voice …		−
	word simplex rank		Thing/ Head1	Thematic−marker/ Nominal−marker/ Modifier1	Thing/ Head1		Binder/ Modifier1	Thing/ Head1		Nominal−marker/ Modifier1	Event/ Head1	Tense/ Modality/ Modifier1	−
			pronoun−ippan …	j−kakari−zyoshi kakari−wa …	common−noun−ippan hasei−go suffixation …		j−zyoshi−rentai−ka …	common−noun−ippan hasei−go suffixation …		j−case−marker case−o …	lexical−verb …	auxiliary−verb aux−tai …	−
	morpheme rank		1	1	Head1	Modifier1	1	Head1	Modifier1	1	1	1	−
			base …	base …	base …	suffix …	base …	base …	suffix …	base …	base …	base …	−
意味解析	figure/move		ph−figure fg−non−projected fg−doing−to−with fg−creative agentive move−simplex mb−goods−and−services command mb−demanding initiating …										
	sequence		−		Goal/ Medium			−			−		−
			−		sq−element sq−unequal sq−projection sq−locution …			−			−		−
	element		Actor/ Agent		Qualifier/ Modifier1			Thing/ Head1			Process		−
			ph−participant simple−thing mb−conscious …		ph−participant macro−thing …			ph−participant simple−thing mb−non−conscious mb−material−object…			ph−process…		−
概念解析			want−action{speaker = user, hearer = system, content = writing{agent = user, object = report{concern = business−trip{}}, instrument = word−processer{}}}										

表4: テキスト理解処理結果の例

4. まとめ

本稿では、SB を使った日本語テキスト処理システムについて説明した。ここで示した SB とテキスト処理アルゴリズムは Java で実装されており、解析結果は入力文に対する XML アノテーションとして出力される。SB は現在、約 670 個の選択体系、約 1460 個の選択肢、約 900 個の具現規則を備えている。

我々は SFL をベースにした資源と処理に SFL に特化していない自然言語処理ツールの結果を統合する方法を開発することによって、[O'Donnell 94]のアイデアを拡張した。この点において、我々のシステムは、ハイブリッドパーザーとみなすことができる。システムの基礎として SFL を採用することによって、我々はより広い視点で言語分析を行うことができるし、既存のパーザーや電子化辞書を SFL 的な資源と組み合わせることによって、開発コストの削減や標準的な分析精度の維持が期待できる。

参考文献

[Fawcett 93] R. P. Fawcett, T. H. Gordon, and Y. Q. Lin: “How systemic functional grammar works: the role of realization in realization,” in New concepts in natural language generation: planning, realization and systems, H. Horacek and M. ZockEds. London: Pinter, 1993, pp. 114-186.

[Halliday 94] M. A. K. Halliday: An introduction to functional grammar, 2nd ed. London: Edward Arnold, 1994.

[Kasper 87] R. T. Kasper: "A unification method for disjunctive feature descriptions," 26th annual meeting of ACL, 1987.

[Matthiessen 91] C. M. I. M. Matthiessen and J. A. Bateman: Text generation and systemic-functional linguistics: experiences from English and Japanese. London: Pinter, 1991.

[O'Donnell 94] M. O'Donnell: Sentence analysis and generation: a systemic perspective, Ph.D. dissertation, Sydney: University of Sydney, 1994.

[Sugimoto 02] T. Sugimoto, N. Ito, H. Fujishiro, and M. Sugeno: "Dialogue Management with the Semiotic Base: A Systemic Functional Linguistic Approach," SCIS & ISIS 2002, Tsukuba, Japan, 2002.

[伊藤 01] 伊藤紀子, 小林一郎, 菅野道夫: "セミオティックベースとそれを利用したテクスト処理について," JASFL Occasional Papers, vol. 2, pp. 63-71, 2001.

[工藤 02] 工藤拓, 松本祐治: "チャンキングの段階適用による日本語係り受け解析," 情報処理学会論文誌, vol. 43, pp. 1834-1842, 2002.

[杉本 03] 杉本徹, 岩爪道昭, 小林一郎, 伊藤紀子, 高橋祐介, 岩下志乃, 菅野道夫: "セミオティックベースを使った日常言語アプリケーションシステム," 第17回人工知能学会全国大会, 新潟, 2003.

[高橋 02] 高橋祐介, 伊藤紀子, 藤城浩子, 菅野道夫: "セミオティックベースにおけるコンテクスト層の検討," 第 16 回人工知能学会全国大会, 東京, 2002.

[EDR 01] 日本電子化辞書研究所: EDR 電子化辞書, 2 版, 2001.