

LFG 解析を利用した 日本語 RTE における複合語の対応

服部 圭悟^{†1} 梅基 宏^{†1} 増市 博^{†1}

本稿では、LFG 解析を利用した複合語の構造解析手法を提案する。Wikipedia の記事の LFG 解析結果を用いて複合語の語基間の述語項構造を明らかにし、複合語の言い換え規則を生成する。また、生成した規則を用いて複合語と複合語の言い換え表現を含む自然文で RTE を行い、複合語の言い換え規則を自然文に適用する上での課題を明らかにする。実験の結果、最大で F 値 77.97%、適合率 87.34%、再現率 70.41% を達成した。

Recognizing Entailment of Japanese Text including Compound Words Using LFG Analysis and Lexical Resources

KEIGO HATTORI,^{†1} HIROSHI UMEMOTO^{†1}
and HIROSHI MASUICHI^{†1}

We propose a method of semantic analysis for compound words based on LFG (Lexical Functional Grammar). We analyze all Wikipedia articles using LFG and determine the grammatical and semantic structures. Then we extract predicate argument structures of compound words. Finally semantic relations of compound words are determined. We conduct experiments to recognize the entailment of Japanese text including compound words using semantic relations and clarify the problems when applying the proposed method. Experimental results showed that we achieved up to 77.97% of f-score, 87.34% of precision, and 70.41% of recall.

1. はじめに

近年、テキスト含意認識 (Recognizing Textual Entailment; RTE) に関する研究が盛んに行われている。RTE は、質問応答、情報抽出、要約、機械翻訳などの様々な応用に役立つと同時に、意味解析を評価する指標になると考えられている^{*1}。これまで、我々は言語理論 LFG (Lexical Functional Grammar)^{1),2)} および言語解析エンジン XLE (Xerox Linguistic Environment)^{*2} を利用した日本語 RTE の研究を行ってきた^{3),4)}。文献 5) では、質問応答における重要な課題として単語や文章の言い換えを挙げている。RTE においても、人や形態 (論文、特許文書、ブログなど) により変化する表現の差異をどのようにして吸収するかは重要な課題であり、ある単語または文章を正しく言い換えることができれば、精度 (適合率や再現率) の改善を期待できる。

今回、我々は複合語の言い換えに注目した。複合語は、複数の基本的な単語 (語基) の組み合わせから構成される単語である。単語の組み合わせで容易に新しい語が加わるため、あらかじめ全ての複合語に対して言い換え表現を網羅することは困難である。よって、複合語を構成している単語の意味とその構造から、複合語全体の意味を推定して言い換えることが必要になる。

従来研究は、人手で作成した規則を用いて言い換え表現を生成する手法が多く提案されている。文献 6), 7) では、文法理論を用いることで複合語の構造を解析している。宮崎ら⁶⁾ は、名詞の表層や読み、品詞の細分類 (一般名詞、固有名詞、数詞など) を用いて構造化規則を作成し、複合語の構造解析を行っている。竹内ら⁷⁾ は、語彙概念構造 (Lexical Concept Structure; LCS) を拡張し、複合語の構造解析を行っている。規則を用いる手法は、規則の作成に熟練した知識が必要なことと、規則作成に時間がかかることが問題である。一方、野口ら⁸⁾ は、京都大学で構築された格フレーム辞書⁹⁾ ^{*3} を用いた複合語の語基間の格関係の解析方法を提案している。精度は、適合率 76.7%、再現率 39.7% と再現率が低く、格フレーム辞書の網羅性が不十分であることを指摘している。

本稿では、LFG 解析を利用した複合語の構造解析手法を提案する。LFG 解析を利用し複合語の語基間の述語項構造を明らかにし、複合語の言い換え規則を生成する。また、生成規則を用いて複合語と複合語の言い換え表現を含む自然文で RTE を行い、生成規則を評価す

^{†1} 富士ゼロックス株式会社 研究技術開発本部
Research and Technology Group, Fuji Xerox Co., Ltd

^{*1} <http://pascallin.ecs.soton.ac.uk/Challenges/RTE/>

^{*2} http://www2.parc.com/isl/groups/nltt/xle/doc/xle_toc.html

^{*3} <http://reed.kuee.kyoto-u.ac.jp/cf-search/>

るとともに、複合語の言い換え規則を自然文に適用する上での課題を明らかにする。対象とする複合語は、従来研究を参考に、2語の名詞で構成される動詞由来複合語とする。動詞由来複合語とは主辞がサ変名詞の複合名詞のことである。

2. 含意関係の判定

2.1 日本語解析の流れ

日本語解析は、文分割、文正規化、形態素解析、LFG 前処理、LFG 解析^{1),2)}、XFR 意味解析¹⁰⁾の順に処理を進める。

形態素解析には茶筌 (ChaSen) と IPA 品詞体系辞書^{*1}を用い、曖昧性を含まない単一の解を求める。LFG 解析および XFR 意味解析は XLE 上で動作し、解析した結果生じる曖昧性は、チャートの形で異なる選択空間の中に閉じ込められて表現され、複数の曖昧性を持つ解は展開されずに閉じ込められた形のまま効率的に処理される。意味解析は XLE の XFR システム上に実装し、f-structure を入力として一連の書き換え規則を順次適応し、同時に概念辞書などを参照することで意味表現を出力する。次節で LFG および f-structure について簡単に説明する。

2.2 Lexical Functional Grammar

LFG は自然言語文の解析を行うための文法理論である。解析結果として、c(onstituent)-structure と f(unctional)-structure と呼ばれる2種の構造を出力する。c-structure は文の係り受け構造の木構造表現であり、一般に構文木と呼ばれるものに対応する。一方、f-structure は、主語 (sb) や目的語 (ob) といった文法機能の概念に基づき、文法の述語・項構造、時制、様相、話法等の意味情報を属性-属性値のリスト構造で表現するものである。LFG では、c-structure を生成するための文脈自由文法規則と、f-structure を生成するために文脈自由文法規則に付与する機能的注釈を同時に記述する。f-structure の例を図 1 に示す。

- (1) 太郎が読んだ本
- (2) 太郎の読んだ本
- (3) 太郎の本を読んだ
- (4) 太郎が本を読んだ
- (5) 太郎は読んだ本を捨てた

例えば、(1)(2)の名詞句は共に文法に則った表現であり『読む』の主語は『太郎』である。

"太郎 / NOUN-PROPER-N-NAME-GIVEN-SFX が読んだ本"

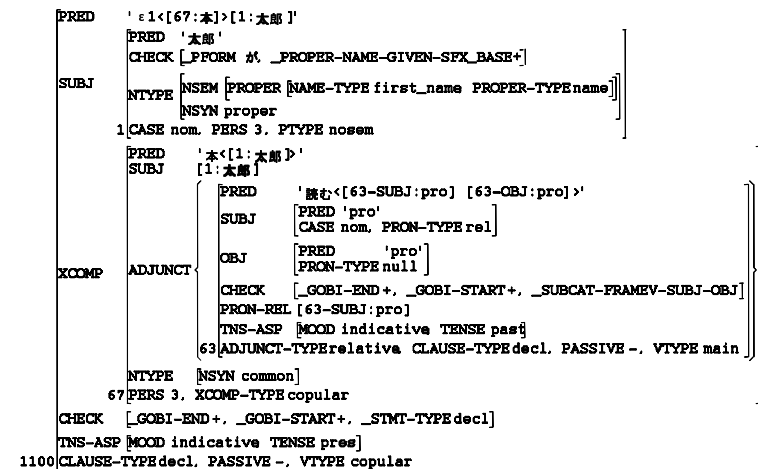


図 1 f-structure
Fig. 1 f-structure

すなわち (2) においては格助詞『の』が主格標識として機能する。しかし、(3)の文において『太郎』が『読む』の主語であるという解釈は成り立たず、この場合の『の』は主格標識となりえない。同様に、(4)の文において『読む』の主語は『太郎』である。一方で、『太郎は読んだ本』という名詞句表現が不自然であることから、(5)の『太郎』は『捨てた』に掛かると判断できる。すなわち、この場合『読む』の主語は省略されていると解釈する。これらの言語現象は、(I) 関係節内においてのみ格助詞『の』が主格標識として機能する、(II) 関係切ないにおいて係助詞『は』による主題化は生起しない、という文法規則に一般化できる。

2.3 意味表現を用いた含意関係の判定方法

含意関係は、テキスト P (Passage) とテキスト Q (Query) とからそれぞれ意味表現を求め、P の意味表現から Q の意味表現が論理的に含意されるかどうかを判定する。

図 2 に、含意関係の判定の様子を示す。ここでは、意味役割 (Role) と主辞 (Head)、引数 (Argument; Arg) の三組を1つの要素 (role(Role, Head, Arg); Fact) と定義する。「role(Role, Head, Arg)」は、主辞 Head と引数 Arg が意味役割 Role の関係にあることを表す。意味役割の例を表 1 に示す。また、word(Head, Concept) は、主辞 Head が概

*1 <http://chasen-legacy.sourceforge.jp>

表 1 意味役割
Table 1 Semantic roles

Role	説明
sb	Arg が意味的な主語, Head が述語の関係
ob	Arg が意味的な目的語, Head が述語の関係
の	ノ格「~の」という形式で文に現れる
に	二格「~に」という形式で文に現れる
eid	Head と Arg が同格の関係 explicit identical

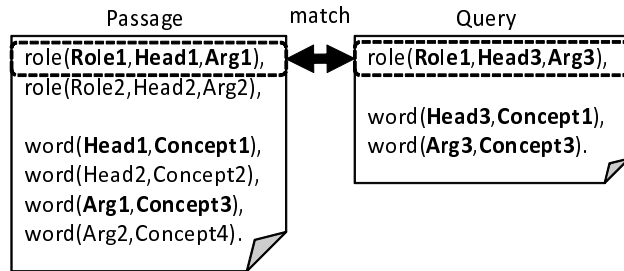


図 2 含意関係の判定
Fig. 2 Judgement of entailment relations

表 2 Wikipedia から得られた意味表現
Table 2 Semantic terms made from Wikipedia

Role	Head	Arg	%
ob	開発する	システム	42.86
の	開発	システム	20.41
sb	開発する	システム	9.52
amod	開発	システム	3.40
として	開発する	システム	2.72
の	システム	開発	2.72

パス中で出現頻度の高い *Fact* を複合語の語基間の述語項構造とみなし、複合語の言い換え規則として採用する。今回、大規模コーパスとして Wikipedia の内部テキストを利用した。

表 2 に、Wikipedia から得られた「システム開発」の言い換え規則の一部を示す。表 2 の上位の *Fact* を用いて言い換え表現を作ると、「システムを開発する」や「システムの開発」となり、「システム開発」の言い換えとして正しい。しかし、例えば表の「role(sb, 開発する, システム)」から言い換え表現を作ると「システムが開発した」や「開発したシステム」となり、言い換えとして適切ではない。また LFG 解析や XFR 意味解析に失敗した場合も、複合語の言い換え表現にならない *Fact* が生成されてしまう。そこで、筆者は『複合語の言い換え表現にならない *Fact* は、言い換え表現になる *Fact* に比べてコーパスでの出現頻度が低い』という仮説を立て、本稿では、複合語の言い換え規則として出現頻度 10% 以上の *Fact* を採用した。

3.2 単語の係り受け傾向に基づく複合語の言い換え対応

3.1 節で提案した手法は、大規模コーパスで語基 $W1$ と $W2$ が係り受け関係を持たない場合、または少ない場合に言い換え規則を生成できない。そこで、本節では、大規模コーパスの解析結果から $W1$ と $W2$ それぞれがとりやすい *Role* と現れやすい位置 (*Head*, *Arg*) を分析し、語基の情報から言い換え規則を生成する手法を提案する。表 3 ~ 表 6 に、「システム開発」の語基の分析結果を示す。表 3 ~ 表 6 を利用して *Role* の一致する *Head* と *Arg* でペアを作り、以下の式によって *Score* を計算する (表 7)。

$$Score = \frac{C_n}{N_{head} + N_{arg}} \times \frac{C_m}{M_{head} + M_{arg}} \quad (1)$$

上式は、各語基の出現数の違いを正規化するように考慮している。本稿では、表 7 の出現頻度 10% 以上の *Fact* を、言い換え規則として採用した。

念 *Concept* をもつ語であることを表す。まず、 P および Q の一行目の *Fact* は同じ意味役割 *Role1* をもつことに注目する。次に、語彙に注目すると、 P の主辞 *Head1* と Q の主辞 *Head3* は、単語の概念を表す *word* 項から、同じ概念 *Concept1* をもつことが分かる。引数に関しても、 P と Q の引数 *Arg1* と *Arg3* で同じ概念 *Concept3* をもつことが分かる。以上から、 P と Q の一行目の *Fact* が照合し、 Q の全ての *Fact* が P の *Fact* と照合できるとき、 P が Q を含意するとみなす。

3. 複合語の言い換え対応

3.1 語基間の係り受けに基づく複合語の言い換え対応

今回の解析対象である動詞由来複合語は、一方の単語が他方の単語に対して動詞的な役割をするという特徴がある。そのため、複合語の語基間の述語項構造を明らかにすれば、複合語の言い換え表現が生成できる。提案手法では、大規模コーパスに対して LFG 解析と XFR 意味解析を行い、複合語の語基 $W1$ と $W2$ が係り受け関係にある *Fact* を収集する。コー

表 3 「システム」が Head
 Table 3 「システム」is Head

Role	基本形	Count
の	システム	3619
id	システム	3079
amod	システム	1950
合計		10921

表 4 「システム」が Arg
 Table 4 「システム」is Arg

Role	基本形	Count
sb	システム	2879
ob	システム	2852
eid	システム	2503
合計		15936

表 5 「開発」が Head
 Table 5 「開発」is Head

Role	基本形	Count
ob	開発する	16737
sb	開発する	12701
の	開発	8558
合計		63119

表 6 「開発」が Arg
 Table 6 「開発」is Arg

Role	基本形	Count
の	開発	4612
eid	開発	3350
に	開発	3035
合計		23772

表 7 表 3 ~ 表 6 の統合結果
 Table 7 Unified 表 3 ~ 表 6

Role	Head	Arg	Score	%
ob	開発する	システム	0.0205	29.67
sb	開発する	システム	0.0157	22.73
の	開発	システム	0.0092	13.28
の	システム	開発	0.0072	10.37
sb	開発	システム	0.0051	7.33

表 8 実験結果
 Table 8 Results

	F 値	適合率	再現率
文献 3) 手法	32.26	83.33	20.00
3.1 手法	77.97	87.34	70.41
3.2 手法	70.37	83.82	60.64

3.3 意味表現上での複合語の言い換え対応

複合語を言い換えるとき、単純に複合語をその言い換え表現に置換するだけでは、文として不自然になる。複合語を別の表現に言い換えるためには、言い換え後の文が自然文になるような、複雑な処理を行わなければならない。提案手法では、図 3 に示すように、生成した言い換え規則を意味表現上で当該複合語に付与することで、複合語の言い換え対応を行う。

図 3 は「システム開発」に「role(の, 開発, システム)」の言い換え規則を付与する様子を示している。矢印が係り受け関係、向きが係り受け方向、各矢印のラベルが意味役割 Role を表している。言い換え対象の複合語を持つ係り受けは、基本的には言い換え規則の Head に付与すればいいことが、事前に行った実験により明らかになっている。しかし、事前実験では Arg に係り受けを付与しても誤検出が増加しなかったため、本稿では、言い換え対象

例: 私が御社のシステム開発を請負う

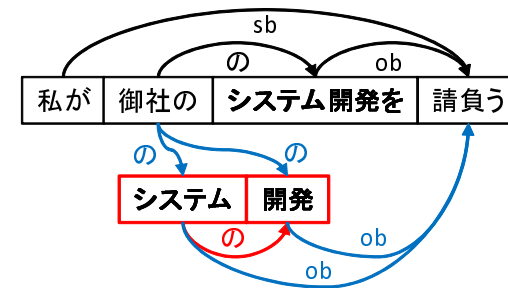


図 3 複合語の言い換え対応
 Fig. 3 Compound word paraphrasing

の複合語を持つ係り受けを、言い換え規則の Head と Arg の両方に付与することにした。

4. 実験

今回の言い換え対象とする複合語は、新聞記事（毎日新聞 7 年分、読売新聞 2 年分）に出現する複合語のうち、上位 5 件（「地価高騰」「市場開放」「システム開発」「共同開発」「税制改革」）である。評価データは、新聞記事から対象複合語を含む文を無作為に 200 文取得し、人手で正例（100 件）と負例（100 件）に言い換えたものを用いる（表 9）。

実験結果を表 8 に示す。我々の従来手法³⁾では、複合語を単語分割し、語間に意味役割「の」を付与したり、また一方がサ変名詞の場合には意味役割「sb」を付与するなど、いくつかのヒューリスティクスを用いることで複合語に対応している。しかし、ルールが不十分であるため、再現率は低い。提案手法は、文献 3) 手法に比べて再現率が大幅に改善している。以降では、実験結果の分析と考察を行う。

5. Error 分析

5.1 誤検出

誤検出 (False Positive; FP) は、3.1 手法で 10 件、3.2 手法で 11 件、重複を除いて合計 12 件を確認した。課題を 3 タイプ (Type1 ~ 3) に分類する (表 10)。

Type1 は、文献 3) 手法の含意判定機能に関する課題で、7 件確認した。例えば、表 10

の Type1 では、文献 3) 手法の複合語解析機能により Query の「共同材料」が「共同」と「材料」の複合語と判断され、「開発」とともに並列の関係となり、Passage の *Fact* と照合した。

Type2 は、LFG 解析または XFR 意味解析に関係する課題で、4 件確認した。例えば、表 10 の Type2-1 では、何らかの理由で Query の LFG 解析または XFR 意味解析が途中で終了し、「大統領府」「評議会」「議長」の 3 単語間の *Fact* のみが意味解析結果として出力された。そして、解析途中の少ない *Fact* で Passage と RTE を行い、FP となった。また、表 10 の Type2-2 では、「の」の意味解析の曖昧性が原因で FP となった。

Type3 は、提案手法で拡張したルールに余計なルールが含まれていることが原因で FP となっており、3.2 手法で 1 件確認した。例えば、表 10 の Type3 では、言い換え規則に「role(sb, 開発する, システム)」を用いたことが原因で FP となっている。

5.2 検出漏れ

検出漏れ (False Negative; FN) は、3.1 手法で 29 件、3.2 手法で 37 件、重複を除いて合計 40 件を確認した。課題を 5 タイプ (Type4 ~ 8) に分類する (表 11)。

Type4 は、作成したテスト自体に問題があるタイプで、7 件確認した。例えば、表 11 の Type4 では、「地価が高騰する波」というのは、「地価が高騰する (という現象が) 波 (のように押し寄せてくる)」といった解釈を与えることで意味を推測することはできるが、そもそも日本語として不自然である。不自然な日本語文をどう扱うかは、日本語文の意味解析における課題である。

Type5 は、複合語の言い換えとは別の言い換えが発生している例で、4 件確認した。例えば、表 11 の Type5-1 では、「『サ変名詞』が続けば続くほど」から「『サ変名詞』するほど」への言い換えに対応していなかったために FN となった。また、表 11 の Type5-2 では、「岩田さん」の係り先が Passage の「手掛ける」から Query の「開発する」へ言い換えられたため、FN となった。複合語の言い換え以外にも、多くの言い換え表現に対応する必要がある。

Type6 は、LFG 解析または XFR 意味解析に帰属する課題であり、8 件確認した。例えば、表 11 の Type6 では、Query の「開放させる」の「せる」に相当する部分が Passage になかったために FN となった。

Type7 は、提案手法で拡張したルールが足りなかったために FN となった例で、13 件確認した。内訳は 3.1 手法が 2 件、3.2 手法が 11 件である。例えば、表 11 の Type7-1 では、「role(の, 開発, 共同)」の言い換え規則に未対応であった例である。表 11 の Type7-2 は、

「role(で, 開発する, 共同)」の言い換え規則に未対応であった例である。

Type8 は、3.3 節の言い換え対応方法に帰属する課題で、8 件確認した。表 11 の Type8 では、Query は「市場を開放する要求をする」を、XFR 意味解析により「要求が市場を開放する」と解釈し、「role(sb, 開放する, 要求)」という *Fact* を生成した。提案手法は XFR 意味解析が終了した後に 3.3 節の言い換え対応を行っている。

5.3 課題のまとめと改善検討

5 節で抽出した課題のうち、提案手法に関する課題は、Type3 および Type7 の言い換え規則の生成方法 (3.1 節, 3.1 節) に基づく課題と、Type8 の複合語の言い換え対応 (3.3 節) に基づく課題の 2 つである。

言い換え規則の生成方法に基づく課題 (Type3 および Type7) は、余計な言い換え規則の適用や、逆に言い換え規則が不足したことが原因である。よって、対象に応じて適切な閾値を設定できるような手法を検討することが、本課題の解決方法として挙げられる。しかし、例えば Type3 の課題では余計な規則として「role(sb, 開発する, システム)」を適用したことを FP の原因として挙げたが、表 2 でもこの言い換え規則は上位にある。つまり、*Fact* の出現頻度の高さと言い換え規則となる *Fact* には、相関関係がない可能性がある。以上を踏まえ、より多くの複合語に対して提案手法を適用して多くのデータで実験結果を分析し、3 節で立てた仮説の検証と適切な閾値の設定方法について検討を進めたい。

複合語の言い換え対応に基づく課題 (Type8) は、XFR 意味解析の適用規則の順番を工夫することで対応する。例えば、Type8 の課題であったように、主語を推定する規則を適用する前に本手法を適用すれば、本課題は解決する可能性がある。

その他の課題について、まず Type1 の課題は、文献 3) 手法の複合語解析機能をオフにすることと、言い換え対象の複合語を増やす、固有名詞を辞書に登録するなどにより対応する。Type2 および Type6 の課題は、LFG 解析や XFR 意味解析が抱える課題で、曖昧性の解消や Modality の扱い、バグフィックスなど、継続して改善に取り組む。Type4 および Type5 の課題は、別の研究テーマとしてそれぞれ別途検討する必要がある。

6. おわりに

本稿では、LFG 解析を利用した複合語の構造解析手法について述べた。Wikipedia の記事に対して LFG 解析と XFR 意味解析を行い、複合語の語基間の述語項構造を明らかにすることで、複合語の言い換え規則を生成した。また、複合語と複合語の言い換え表現を含む自然文で RTE を行い、複合語の言い換え規則を自然文に適用する上での課題を明らかにし

表 9 評価データ
Table 9 Test data

Label	Passage	Query
+1	規制で地価高騰の頭打ちを狙った	規制で地価の高騰の頭打ちを狙った
+1	規制緩和と市場開放で民間活力を高める	市場を開放することで民間活力を高める
+1	韓国の水質改善に向けてシステム開発などのプロジェクトが進んでいる	韓国の水質改善に向けてシステムを開発する
-1	地価高騰が下落よりも問題になる	地価の下落が高騰よりも問題になる
-1	日本の一段の市場開放と土地の保守	日本の市場の保守と土地の一段の開放
-1	コンピューターのシステム開発に1年以上かかる	コンピュータの開発に1年以上かかるシステム

表 10 誤検出
Table 10 False positive

Type	Passage	Query
1	表示部の基礎設計や材料の共同開発	表示部の基礎設計や共同材料の開発
2-1	共同開発可能な米大統領府持続評議会議長	米大統領府持続可能な開発評議会共同議長
2-2	写真フィルムの市場開放の問題	市場の写真フィルムの問題の開放
3	システム開発	開発したシステム

表 11 検出漏れ
Table 11 False negative

Type	Passage	Query
4	地価高騰の波が遅れてやって来た	地価が高騰する波が遅れてやって来た
5-1	異常な地価高騰が続けば続くほど日本企業の株価は高くなる	地価が高騰するほど株価は高くなる
5-2	岩田さんがシステム開発を手掛ける	岩田さんがシステムを開発する
6	アジアの市場開放を強硬に求めた	アジアの市場を開放させる
7-1	ロケットで初の日米の共同開発となる	初の日米の共同の開発となる
7-2	共同開発に踏み切ることにした	共同で開発する
8	他国に市場開放を要求する	他国に市場を開放する要求をする

た。提案手法の精度は、最大で F 値 77.97%、適合率 87.34%、再現率 70.41% を達成した。

今後の課題として、我々は格フレーム辞書⁹⁾の利用を検討している。文献 8) で指摘されているように再現率の改善には期待できないが、大規模な Web 文書の構文解析データであるため全体の精度改善に寄与する可能性がある。その他には、動詞由来複合語以外への対応や 3 語以上の語基を含む複合語への対応が課題として挙げられる。

参 考 文 献

- 1) H.Masuichi, T.Ohkuma, H.Yoshimura and Y.Harada: Japanese parser on the basis of the Lexical Functional Grammar formalism and its evaluation, *Proc. of the 17th Pacific Asia Conference on Language, Information and Computation*, pp.298-309 (2003).
- 2) 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, 言語処理学会, 2, Vol.10, pp.79-109 (2003).
- 3) 梅基 宏, 杉原大悟, 大熊智子, 増市 博: LFG 解析と語彙資源を利用した日本語含意関係判定, 情報処理学会研究報告, 113, Vol.2008, pp.57-64 (2008).

- 4) H.Umemoto and K.Hattori: Experiments of FX for NTCIR-9 RITE Japanese BC Subtask, *Proc. of NTCIR-9 Workshop Meeting*, pp.412-417 (2011).
- 5) 高橋哲朗, 乾健太郎, 関根 聡, 松本祐治: 質問応答に必要な言い換えの分析, 言語処理学会第 10 回年次大会発表論文集, pp.309-312 (2004).
- 6) 宮崎正弘, 五百川明, 川辺 諭: 構造化チャートパーサを用いた日本語複合名詞構造解析器, 言語処理学会年次大会発表論文集, pp.229-232 (2008).
- 7) 竹内孔一, 内山清子, 吉岡真治, 影浦 峯, 小山照夫: 語彙概念構造を利用した複合名詞内の係り関係の解析, 情報処理学会論文誌, 5, Vol.43, pp.1446-1456 (2002).
- 8) 野口慎一郎, 徳永健伸: 格フレーム辞書を用いた日本語複合名詞の解析, 情報処理学会研究報告, pp.67-72 (2007).
- 9) 河原大輔, 黒橋禎夫: 高性能計算環境を用いた Web からの大規模格フレーム構築, 自然言語処理研究会, 1, Vol.2006, pp.67-73 (2006).
- 10) D.Crouch and T.H.King: Semantics via F structure Rewriting, *Proc. of LFG06 Conference* (2006).