

語彙機能文法に対する Glue 意味論に基づいた 日本語解析器の実装

Implementing a Japanese Semantic Parser Based on Glue Approach for Lexical-Functional Grammar

要 旨

本論文では Glue 意味論に基づく日本語解析器の実装について述べる。この解析器はいわゆる言語学的に深い解析を行ない、テキストの意味的妥当性の検証や冗長性・整合性などの論理的・推論的な関係の検出などに利用できる。特長としては、分野に依存しない設計になっており、ストレージなどの付加的な機構なしに意味的に曖昧な文を処理でき、スコープ付きの高階内包論理式を生成する。本解析器は、英語版の語彙機能文法 (LFG) に対する Glue 意味解析器を元に構築され、LFG の言語類似性により、比較的少ない構築労力で高い解析カバー率を達成できた。本稿ではこの解析器を概説し、日本語特有の表現である、浮遊量化子の分配読みと累積読み、二重主語構文、副助詞の解析について説明する。また、構文的には対応する英語表現と同一だが意味的に異なる表現である、従属節中の相対時制に関する解析結果を示す。今後は、営業や医療といった分野のドメイン知識と組み合わせたテキストからの知識抽出などに応用することを検討していきたい。

Abstract

This paper describes the implementation of a Japanese semantic parser based on the glue approach. The parser puts linguistically deep analyses into practice and is utilized for verifying the semantic validity of text and detecting logical or deductive relations such as redundancy and consistency within a text. It is designed as domain-independent, and produces fully scoped higher-order intensional logical expressions, coping with semantically ambiguous sentences without a storage mechanism. It is constructed from an English semantic parser on top of Lexical-Functional Grammar (LFG), and it attains broad coverage through relatively little construction effort, thanks to the parallelism of the LFG grammars. I outline the parser, and I present the analyses of Japanese idiosyncratic expressions including floating numerical quantifiers, showing the distinct readings of distributive and cumulative, as well as a double-subject construction and focus particles. I also explain the analyses of expressions that are syntactically parallel but semantically distinct, such as relative tense in subordinate clauses. In the future, I would like to explore the application of the parser for knowledge extraction from text combined with domain knowledge in fields such as sales and medical care.

執筆者

梅基 宏 (Hiroshi Umemoto)

研究本部 未来ワーク研究所
(Future Work Research Laboratory, Corporate
Research Group)

1. 緒言

様々な局面で大量の電子テキスト情報にアクセスできるようになり、必要な情報を取り出すために情報検索、情報抽出、機械翻訳、テキスト要約などのテキスト処理技術が広く活用されている。キーワードやパターンなどに基づいた、いわゆる言語学的に浅い解析方法は、処理が効率的に行なえて広く使われている反面、意味的な妥当性を結果に保証していないためにおかしな結果を出力してしまうことがある[1]。また、そのような浅い解析は文書の類似性を見つけることには向いているが、冗長性や整合性などといった論理的・推論的な関係を見つけ出すことが不得意で、少し込み入った文になると正しい結果を出すことが難しくなってしまう[2]。

語彙機能文法 (LFG) および Glue 意味論は PARC を中心に研究された、いわゆる言語学的に深い解析技術である[3]。これらの技術の応用例として、PARC のユーレカデータベースの研究が挙げられる[2]。ユーレカシステムは、コピー機を修理する折にゼロックスの技術者が書いた自由テキスト形式のノウハウに対して、冗長性や整合性を自動的に検出することができるもので、ノウハウが蓄積されるデータベースの品質を管理し、関連するノウハウを提示するとともに、関連する個々のノウハウを一つの新しいノウハウにまとめあげることを目指している。ユーレカシステムの技術は、ノウハウデータにとどまらず、特許情報、法的な文書、医療文書などといった高い価値をもった複雑な情報ソースに応用することが期待できる[4]。

また、Glue 意味論では形式的な論理表現を用いて意味を表現しているが、そうした論理表現はドメイン知識などと組み合わせることで、テキストに基づいた推論を自動的に利用することが可能になる[5]。すなわち、ことばの意味を論理で表現することで、ことばだけの世界にとどまらず、例えば営業や医療といった実世界のドメイン知識と論理表現を介してつなげることができる。そこでの応用としては、知識ベースの記述と注目している文書との意味的な整合性を検証したり、新しい知識を文書から抽出することなどが考えられる。この場合、実世界の知識をどう表現して構築するかということが課題になるが、近年 Semantic Web の発展などもあってオントロジー工学に注目が集まっており[6]、

将来的に利用できる知識表現は格段に増大するものと予想される。

さて、ことばの意味について、人は初めて聞いた文であってもその意味を問題なく理解できることから分かるように、人は文の部分ごとの意味とそれら相互の組合わり方から文全体の意味を決めることができると考えられている[3]。このように意味解析には、小さな部分の意味からより大きな部分の意味を構成する理論が必要になる。また、意味解釈には曖昧性があるのが普通であるが、論理表現として意味を考えた場合、スコープの曖昧性をどのように表現するのが課題の一つになる[1]。

Glue 意味論は構文と意味論のインターフェイスを提供し、意味的な合成を線形論理の導出として表現する[3]。論理的な導出の前提は、文の構文的な項目に対応しており、文から得られた前提の集合から複数の証明が導かれた場合、それは文の意味的な多義性、すなわち解釈（読み）の違いを表し、その際他の方法論で使われるストレージなどといった特別な機構を必要としない。導出に線形論理を利用することで、意味的な完全性と一貫性が結果に保証されている。すなわち、すべての前提条件は充足され、使われなかった前提が線形論理の導出結果に残ることはない。Glue 意味論は元来、LFG に対して開発されたが、特定の文法フォーマリズムに適用が制限されているわけではなく、様々な構文的フォーマリズムや意味表現に適用することができる[7]。

本稿では、Glue 意味論に基づく日本語解析器の実装について述べる。英語の LFG に対する Glue 意味論の実装はすでに構築されており、実テキストを対象とした応用システムに組み込まれている[2][8]。しかしこれまでは、英語以外に高い解析カバー率をもって適用された言語はなかった。今回説明する日本語の解析器は英語版の解析器と同じフレームワーク上で構築されている。両者は共に分野に依存しない設計になっており、伝統的な形式意味論に則り、完全にスコープが与えられた高階の内包論理表現を文の意味として生成する。この日本語解析器は、英語版の解析器と同様に、現実のテキストを幅広く扱うことを目指しており、高い解析カバー率をもつ LFG 日本語文法の上に構築されている[9]。この日本語文法は、Parallel Grammar (ParGram) プロジェクトを通じて開発されて

いる。ParGram プロジェクトでは、LFG に基づく効率的な解析器および文法開発プラットフォームである Xerox Linguistic Environment (XLE) が用いられており、LFG の言語普遍性と適用範囲について検証が行なわれている[10]。

2. 解析器の概要

Glue 意味論の解釈は 2 つのステップから構成される。ステップ 1 は、与えられた文の構文的な項目から対応する意味構成要素 (meaning constructor) を生成する。ステップ 2 では、意味構成要素を組み立てて、文全体の意味表現を求める。

ステップ 1 は、いわゆる「解析による記述 (description by analysis[3])」として実装されている。すなわち、規則に基づいて f(unctional)-structure を解析することで意味構成要素が得られる。対応づけの写像規則は、意味辞書で定義されており、LFG の言語類似性のために、英語版の写像規則の大半はそのままの形か、もしくははわずかに変更することで、日本語の解析にも適用することができる。

ステップ 2 の導出機構は言語に関係なく共通に使うことができる。意味的な合成は意味構成要素の集合に対する線形論理の証明として行なわれ、Curry-Howard 同型写像により意味項を合成し、意味表現を可能な限りすべて生成する。スコープの曖昧性といった意味的な曖昧性は、意味構成要素の同じ集合から成り立つ複数の証明で表されている[11]。

Glue 意味論の解析器は XLE 内に組み込まれており、言語に依存しないフレームワークを提供している (Glue 解析フレームワークと英語版の意味辞書は、Dick Crouch、Ash Asudeh、John Fry によって作成された)。意味辞書には、f-structure の項目を意味構成要素に写像する規則が定義されていて、言語に依存することになる。各写像規則は条件部分と本体部分に分かれており、前者にはチェックされることになる形態素ラベル、述語形式、他の f-structure との関係が、後者には、前者の条件が満たされた場合に提供される、意味構成要素およびスコープ制約条件が規定されている。

今回実装した日本語の意味辞書は、258 個の写像規則で構成されている。そのうち、109 個の規則は日本語独自の意味論のために新たに作

成され、英語版の意味辞書には存在しない。これらは動詞の格フレーム (例えば二重斜格をもつ動詞)、述語形容詞、絶対時制・相対時制、副助詞、連体詞、分配読み、様態表現などに関係する。72 個の規則は、英語版と同様に日本語の辞書でもそのままの形で使われている。77 個の規則は本体部分はそのままの形で、f-structure の条件部分が日本語 LFG の出力に対応するために若干変更された。

3. 日本語文の解析

3.1 構文的に英語文と類似して意味的に曖昧な文

基本的に、英語用の写像規則であっても、構文的に類似している表現であれば、ほとんどそのままの形で日本語にも適用できる。次の例文を考えることにする。

(1) どの学生も机を運んだ。

例文 (1) には構文的な解析結果がただ一つしかないが、意味的には少なくとも二つ解釈の可能性がある。一つの解釈は対格名詞の「机」が狭いスコープをとり、文の意味が各生徒ごとにそれぞれ一つの異なる机が存在するものである。もう一つの解釈は、対格名詞が広いスコープをとり、文の意味が全ての生徒に対してただ一つの机しか存在しないものである。

例文 (1) に対応する f-structure を図 1 に示す。

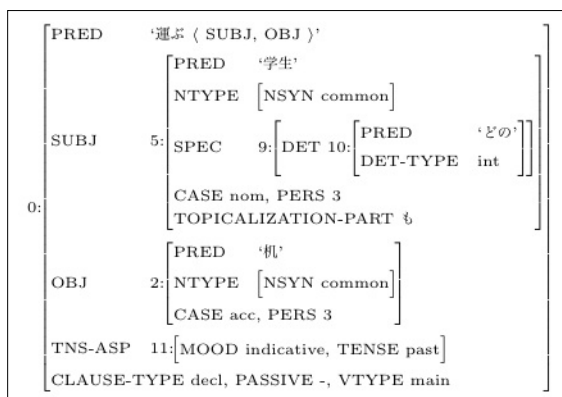


図 1. 例文(1)に対応する f-structure
F-structure corresponding to sentence (1)

f-structure は、下位の f-structure を階層的に含むことができ、図 1 で各 f-structure は開き角括弧の左側にコロンと共に番号がつけられている。

図 1 の f-structure から図 2 に示す意味構成要素が得られる。

[過去]	1	$\lambda A.\lambda B.\lambda C.[\text{and}, [C, B], [A, B]] :$ $(\text{VAR}_e(11) \rightarrow 11_t) \rightarrow \text{EVAR}_e(0) \rightarrow (\text{EVAR}_e(0) \rightarrow 0_t)$ $\rightarrow 0_t$
	2	$\lambda A.[\text{past}, A] : \text{VAR}_e(11) \rightarrow 11_t$
[出来事]	3	$\lambda A.[\text{quant}, \text{exists}, \text{sg}, B \setminus [\text{and}, [\text{event}, B],$ $\text{cxr}(C, B, \text{event})], A] :$ $(\text{EVAR}_e(0) \rightarrow 0_t) \rightarrow 0_t$
[机]	4	$\lambda A.\lambda B.[\text{quant}, \text{exists}, \text{sg}, \text{pl}, C \setminus [\text{and}, [A, C],$ $\text{cxr}(D, C, \text{barenoun})], B] :$ $(\text{VAR}_e(2) \rightarrow \text{RESTR}_t(2)) \rightarrow (\forall H_t.(2_t \rightarrow H_t) \rightarrow H_t)$
	5	$\lambda A.[[\text{noun}, \text{机}], A] : \text{VAR}_e(2) \rightarrow \text{RESTR}_t(2)$
[どの]	6	$\lambda A.\lambda B.[\text{quant}, \text{forall}, \text{sg}, C \setminus [\text{and}, [A, C],$ $\text{cxr}(D, C, \text{int}(\text{どの})), B] :$ $(\text{VAR}_e(5) \rightarrow \text{RESTR}_t(5)) \rightarrow (\forall H_t.(5_t \rightarrow H_t) \rightarrow H_t)$
[学生]	7	$\lambda A.[[\text{noun}, \text{学生}], A] : \text{VAR}_e(5) \rightarrow \text{RESTR}_t(5)$
[運ぶ]	8	$\lambda A.\lambda B.\lambda C.[\text{and}, [[\text{verb}, \text{運ぶ}], A],$ $[\text{and}, [\text{subcat}, A, 'V-SUBJ-OBJ'], [\text{and}, [\text{xle_gf}, A, B, \text{subj}],$ $[\text{xle_gf}, A, C, \text{obj}]]]]] : \text{EVAR}_e(0) \rightarrow 5_t \rightarrow 2_t \rightarrow 0_t$

図 2. 図 1 の f-structure に対応する意味構成要素
Meaning constructors corresponding to
the f-structure shown in figure 1

図 2 の欄は順にラベル、通し番号、意味構成要素を示す。意味構成要素のうち、コロン記号の左側は意味項を、右側は論理式をそれぞれ表しており、意味項の本体は角括弧内に表され、大文字は変数を、逆スラッシュ記号はラムダ抽象化記号をそれぞれ表している。太文字の数字は、その数字でタグ付けされた f-structure の意味を表している。EVAR、VAR、RESTR はそれぞれ事象の変数、個体の変数、個体の制約を表している。項の右下にある小文字の e と t は、それぞれ個体のタイプと真理値のタイプを表している。棒と丸からなる記号は線形論理の含意を表している。

一般化量子が Prolog のリスト記法で表現されており、要素は識別子 quant、識別子タイプ exists または forall、単数を表す sg といった数の情報、制約部分、本体部分である。関数 cxr は文脈の依存性をマークするためのフックとして置かれている。下位範疇化の情報は subcat と xle_gf の項に明示的に表現されており、後者は XLE の文法的機能を表している。

基本的に日本語の一般名詞は数に関してマークがつけられていない。例文 (1) の「机」を表す名詞の数は sg_pl に設定され、単複の両方を示している。一般名詞はしばしば裸名詞として表現され、限定性は通常特定されていない。

例文 (1) に対しては、図 3 に示す二つのスコー

プ制約条件が図 2 の意味構成要素の他に課せられる。

$\text{notscopes}(\text{EVAR}_e(0), \text{RESTR}_t(2))$
$\text{notscopes}(\text{EVAR}_e(0), \text{RESTR}_t(5))$

図 3. 例文 (1) に課せられるスコープ制約条件
Scope constraints for sentence (1)

図 3 の制約条件は、ノード 0 の事象変数がノード 2 とノード 5 の制約をスコープ内に含めてはいけないことを意味している。

f-structure に対応するすべての意味表現が、論理的な導出を通じて得られる。この例の場合、いくつかの証明が可能である。一つの証明は、意味構成要素を (1, 2), 3, 8, (4, 5), (6, 7) の順序で消費するものである。ここで、丸括弧で囲まれた意味構成要素の列が先に消費され、次に括弧内の結果の式がその右隣の意味構成要素と消費されるものとする。この証明は狭い読みに対応している図 4 に示す意味表現を導く。

$[\text{quant}, \text{forall}, \text{sg}, J \setminus [\text{and}, [[\text{noun}, \text{学生}], J],$ $\text{cxr}(K, J, \text{int}(\text{どの})),$ $L \setminus [\text{quant}, \text{exists}, \text{sg}, \text{pl}, M \setminus [\text{and}, [[\text{noun}, \text{机}], M],$ $\text{cxr}(N, M, \text{barenoun})],$ $O \setminus [\text{quant}, \text{exists}, \text{sg}, P \setminus [\text{and}, [\text{event}, P],$ $\text{cxr}(Q, P, \text{event})],$ $R \setminus [\text{and},$ $[\text{and}, [[\text{verb}, \text{運ぶ}], R],$ $[\text{and}, [\text{subcat}, R, 'V-SUBJ-OBJ'],$ $[\text{and}, [\text{xle_gf}, R, L, \text{subj}], [\text{xle_gf}, R, O, \text{obj}]]]]],$ $[\text{past}, R]]$ $]]]$

図 4. 例文 (1) の狭い読みに対応している意味表現
Meaning expression corresponding to the
narrow reading of sentence (1)

別の証明は (1, 2), 3, 8, (6, 7), (4, 5) の順序で意味構成要素を消費するものである。この証明は広い読みに対応している意味表現を導く。その他の証明として例えば (4, 5), 8, (6, 7), (1, 2), 3 や、(6, 7), 8, (4, 5), (1, 2), 3 も考えられるが、スコープの制約条件のためにこれらを導くことはできない。

3.2 日本語固有の表現

3.2.1 浮遊数値量子

主語に対する浮遊量子は構文的に動詞句と組み合わせられており、目的語に対する浮遊量子は動詞のみと組み合わせられている。前者の構文は集合読みをとることができず、分配読みをとることになる。事象から個体への写像が存在

し、測定関数が事象に対応する個体を測ることで間接的に事象を測っている[12]。次の文 (2) は両方の浮遊数値量子子を含む例である (例文 (2) は[13]に挙げられている例を改変したものである)。

(2) 学生が 3 人ビールを 6 杯飲んだ。

今回の解析システムの実装では、分類辞を二つのタイプに分けている。一つのタイプは数え上げられる対象の数を特定するもので、他方は数え上げられない対象を測るものである。主語に対する浮遊量子子が数え上げられる対象をもつ場合、分配的な複数の事象が存在し、各事象は対象とする名詞が表す要素のうち、ある一つの個体に対応している。また、意味構造の素性を二つ新たに導入した。分配読みされる対象名詞の個体と事象変数であり、それぞれ D、

```

[過去] 1 λA.λB.λC.λD.[and,[D,C],[A,C]] :
      (VARe(19) → 19t) → marke(19)
      → EVARe(0) → (EVARe(0) → 0t) → 0t
2 λA.[past,A] : VARe(19) → 19t
[事象] 3 λA.[quant,exists,pl,B\and,[event,B],
      cxr(C,B,event)],A] :
      (EVARe(0) → 0t) → 0t
[分配読み] 4 λA.A : (EVARe(0) → 16t → 0t) → DEVARe(0)
      → De(16) → 0t
5 λA.λB.λC.[quant,exists,sg,
      D\evmember,D,A],B] :
      EVARe(0) → (DEVARe(0) → 0t)
      → marke(0) → 0t
6 λA.λB.[quant,forall,sg,C\member,C,A],
      D\B,z,D]] :
      16e → (marke(0) → De(16) → 0t) → 0t
[3] 7 λA.λB.λC.λD.[quant,exists,D,E\and,[and,
      [A,E],[B,E]],cxr(F,E,numeral)],C] :
      (VARe(16) → RESTRt(16))
      → (VARe(2) → RESTRt(2))
      → (∨Ht.(16e → Ht) → 5t → Ht)
8 measure : 2e
9 λA.λB.[adverbial,A,B] : 2e → VARe(2) → 2t
10 λA.λB.λC.λD.[and,[B,z,D],[A,D]] :
      (VARe(2) → 2t)
      → (marke(19) → DEVARe(0) → 0t)
      → marke(2) → DEVARe(0) → 0t
11 λA.[classifier,人,A] : VARe(2) → RESTRt(2)
12 3 : 5e
    
```

図 5. 例文(2)の意味構成要素の一部
Part of meaning constructors corresponding to sentence (2)

DEVAR で表した。例文 (2) に対応する意味構成要素を部分的に示すと図 5 のようになる。

例文 (2) に対しては累積読みと分配読みの少なくとも二つの解釈が可能である。飲まれたビールの合計は、前者の解釈では 6 杯で、後者の解釈では 18 杯になる[13]。例文 (2) に対応する意味構成要素とスコープ制約条件とから、その二つの解釈を含む 7 つの解析が得られる。累積読みに対応する意味表現は図 6 の通りである。

```

[quant,exists,3,
A\and,[and,[noun,学生],A],[classifier,人,A]],
cxr(B,A,numeral)],
C\quant,exists,pl,D\and,[event,D],cxr(E,D,event)],
F\quant,exists,6,
G\and,[and,[noun,ビール],G],[classifier,杯,G]],
cxr(H,G,numeral)],
I\quant,forall,sg,J\member,J,C],
K\quant,exists,sg,L\evmember,L,F],
M\and,
[and,
[and,
[and,[verb,飲む],M],
[and,[subcat,M,'V-SUBJ-OBJ'],
[and,[xle_gf,M,K,subj],[xle_gf,M,I,
obj]]]],
[past,M]],
[adverbial,measure,M]],
[adverbial,measure,M]]
]
]
]
]
]
]
    
```

図 6. 例文(2)の累積読みに対応する意味表現
Meaning expression corresponding to the cumulative reading of sentence (2)

3.2.2 二重主語構文

形容詞が表面上二つの主格をもつ文は日本語では普通に見られ、二重主語構文の文とよばれる[14]。二重主語構文の基本的な文型は、「A は B が… (述語)」である。ただし A と B は名詞句を表している[15]。

(3) 象はすべて鼻が長い。

二重主語構文はいくつかのタイプに分類されるが、例文 (3) の場合は主題辞「は」が連体修飾辞「の」の代用になっている[14]。すなわち、主語の「鼻」と主題語の「象」は部分と全体の関係にある。例文 (3) の意味表現は図 7 の通りである。

```
[quant,exists,sg_pl,A\and,[noun,象],A],
  cxr(B,A,barenoun)],
C\quant,exists,sg_pl,
  D\and,[and,[noun,鼻],D],[of_topic,C,D]],cxr(E,D,
  barenoun)],
F\quant,exists,sg,G\and,[event,G],cxr(H,G,event)],
  I\and,
    [and,
      [and,[adjective,長い],I],
      [and,[subcat,I,'A-SUBJ'],[xle_gf,I,F,subj]]],
      [pres,I]],
      [postposition,topic,は,I]]
    ]
  ]
]
```

図 7. 例文(3)の意味表現
Meaning expression of sentence (3)

3.2.3 副助詞

日本語の副助詞には多義性があるが、[16]は、形態素と文法機能の二つのレベルの規則によって、「まで」「など」「だけ」について高い精度で多義性を判別できることを報告している。次の例文を取りあげる。

- (4) 太郎だけが賢い。
(5) 太郎だけに賢い。

助詞の「だけ」は、格助詞に先行している場合に PRED (icate) を制限する機能を示す傾向にある。その一方で、付加句に含まれている場合は理由を示す傾向にある[16]。例文(4)と(5)の意味表現をそれぞれ図8と図9に示す。

```
[[だけ,only],name(sigma(3),A\and,[name,太郎],A],
  [name_type,A,name]]],
B\quant,exists,sg,C\and,[event,C],cxr(D,C,event)],
  E\and,
    [and,[adjective,賢い],E],
    [and,[subcat,E,'A-SUBJ'],[xle_gf,E,B,subj]]],
    [pres,E]]
  ]
]
```

図 8. 例文(4)の意味表現
Meaning expression of sentence (4)

```
[quant,exists,sg,A\and,[nullPro,A],cxr(B,A,nullPro)],
C\quant,exists,sg,D\and,[event,D],cxr(E,D,event)],
  F\and,
    [and,
      [and,[adjective,賢い],F],
      [and,[subcat,F,'A-SUBJ'],[xle_gf,F,C,subj]]],
      [pres,F]],
      [postposition,
        G\[[だけ,reason],
          name(sigma(4),H\and,[name,太郎],H],[name_type,
            H,name]]),G],
        に,F]]
    ]
  ]
]
```

図 9. 例文(5)の意味表現
Meaning expression of sentence (5)

3.3 構文的に類似しているが意味的に異なる表現

構文的に類似しているにも関わらず英語版の写像規則をそのまま適用できない場合がある。一つの例は、従属節中の相対時制である。[17]は、日本語の時制は相対的であり、時制の形態素は局所的に c-command の関係にある時制との関係で解釈されると主張している。次の例を取りあげる。

- (6) 私が日本へ行くときにプレゼントを買った。

主節が現在形であるか過去形であるかに関わらず、主節に表現されている行動が起こる前に形式名詞「とき」を含む従属節の行動がまだ完了していない場合、従属節は現在時制をとる。しかし主節の行動が起こる前に従属節の行動が完了していた場合、従属節は過去時制をとる。[15]

- (7) 私が日本へ行ったときにプレゼントを買った。

「とき」を含む従属節が状態動詞や進行形の動詞をとる場合、主節と従属節の事象は同時に起こる。そこで上の例文(7)の場合、主節の事象は必ずしも従属節の事象に後続するとは限らず、前者が後者の前に起こることはない("not-precedes")とした。

しかし、関係節の時制を表す形態素は常に相対的とは限らない。次の文は、関係節の時制が絶対的となっている一つの例である。

- (8) 自殺した人がタクシーに乗った。

例文(8)では、死人はタクシーに乗れないため、自殺した事象が乗車した事象に先行すると解釈することはできない。したがって、文脈に応じて関係節の時制を絶対的な時制として扱う必要がある。主節と関係節がともに過去形をとる場合、関係節の事象は事実であり、過去に実際に起こったことであるが、主節の事象に先行しているとは限らない。例文(9)のように、対称的に両方の節が現在時制をとる場合、関係節の事象は現在起こっているかこれから起こることであるが、主節の事象に後続しているとは限らない。

(9) 自殺する人がタクシーに乗った。

4. 解析カバー率

今回実装した解析器を使って、新聞記事や雑誌、辞書から集められた 20 万文以上を含む EDR 日本語コーパス[18]から取り出した文集合を対象に解析実験を行なった。対象としてコーパスから 1,000 文をランダムに選択したところ、文あたりの平均語数は 21 語であった。日本語 LFG により 1,000 文中 942 文が解析でき、残りの 58 文は制限時間内に解析できなかった。日本語 LFG は部分解析用の文法も備えているが[9]、解析できた 942 文のうち、完全な解が 857 文、部分解析用の文法によって適格と規定された部分解が 85 文についてそれぞれ得られた。少なくとも一つの構文解析結果に対して意味解析結果が 675 文について得られたので、対象の 1,000 文に対する解析カバー率は 65.7%、LFG の部分解を含む解に対する解析カバー率は 71.7%であった。LFG の完全解が得られた 857 文のうち、意味解析結果が 668 文に対して得られたので、LFG の完全解に対する解析カバー率は 77.9%であった。なお、意味解析器も LFG と同様に部分解析用の意味辞書項目を備えており、部分解を出力することができるが、ここでは部分解は考慮していない。

表 1. 意味解析器の解析カバー率
Coverage of the semantic analyzer

意味解析対象	(文数)	意味解析解 (文数)	解析カバー率 (%)
実験対象全体	1000	675	65.7
LFG 解	942		71.7
LFG 完全解のみ	857	668	77.9
LFG 部分解のみ	85	7	8.2
LFG 解析不能	58	0	0

5. 結言

Glue 意味論に基づく日本語解析の実装を示し、構文的に英語文と類似している日本語文が英語の場合と同様に解析できることを示した。次に、浮遊量子や二重主語構文、副助詞といった日本語固有の表現と、構文的に類似しているが意味的に異なる、日本語の相対時制を解析した結果を示した。解析器を構築する際、LFG 文法の言語類似性のために、英語版の意味辞書を移植することで、比較的少ない構築労力で高い

解析カバー率を達成できた。

今後は解析カバー率をさらに高めていき、情報検索・情報抽出・機械翻訳・テキスト要約などの各種テキスト処理の結果に対する意味的妥当性の検証、文書コレクションにおける冗長性・整合性の検出、営業や医療といった分野のドメイン知識と組み合わせたテキストからの知識抽出などに応用することを検討していきたい。

6. 謝辞

本研究を進めるにあたって適切な支援をいただいた Palo Alto Research Center Inc. の Dick Crouch 氏および Tracy Holloway King 氏に感謝いたします。

7. 参考文献

- [1] J. Allen. Natural Language Understanding 2nd Ed. CA, USA. The Benjamin/Cummings Publishing Company, Inc.. 1994. 654p
- [2] D. Crouch, C. Condoravdi, R. Stolle, T. King, V. de Paiva, J. O. Everett and D. G. Bobrow. Scalability of redundancy detection in focused document collections. Proceedings First International Workshop on Scalable Natural Language Understanding (SCANALU-2002). (2002)
- [3] M. Dalrymple 編. Lexical Functional Grammar. San Diego, CA, USA. Academic Press. 2001. 484p
- [4] N. Dempsey and N. Evans. An focus report: Futuregazing. 2003
- [5] P. Blackburn and J. Bos. Representation and inference for natural language. CA, USA. CSLI Publications. 2005. 350p
- [6] S. Staab and R. Studer 編. Handbook on Ontologies. Springer. 2004. 660p
- [7] A. Asudeh and R. Crouch. Glue semantics for HPSG. Proceedings of the 8th Intl. HPSG Conference. (2001)
- [8] D. Crouch. Packed rewriting for mapping semantics to KR. Proceedings of the 6th International Workshop on Computational Semantics. (2005)

- [9] 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. Vol.10, No.2, p.79-109 (2003)
- [10] M. Butt, H. Dyvik, T. H. King, H. Masuichi and C. Rohrer. The parallel grammar project. Proceedings of COLING2002, Workshop on Grammar Engineering and Evaluation. p.1-7(2002)
- [11] A. Asudeh. Resumption as Resource Management. PhD thesis, Stanford University. 2004
- [12] K. Nakanishi. The semantics of measure phrases. the Proceedings of the 33rd Conference of the North East Linguistic Society (NELS 33). p.225-244 (2003)
- [13] T. Gunji and K. Hasida. Measurement and quantification. Topics in Constraint-Based Grammar of Japanese (Eds. by T. Gunji and K. Hasida). Dordrecht. Kluwer Academic Publishers. chapter 3 (1998)
- [14] M. Oku. Analysing Japanese double-subject construction having an adjective predicate. COLING-96. p.865 - 870 (1996)
- [15] H. Storm. A Handbook of Japanese Grammar. LINCOM GmbH. 2003
- [16] 大熊智子, 増市博, 吉岡健. 日本語 LFG を用いた副助詞の多義性解消. 自然言語処理. Vol.13, No.1, p.27-52 (2006)
- [17] T. Ogihara. Tense and aspect. The Handbook of Japanese Linguistics (Ed. by N. Tsujimura). Blackwell Publishers. chapter 11, p.326-348 (1999)
- [18] 日本電子化辞書研究所. EDR 電子化辞書説明仕様書. 1996

筆者紹介

梅基 宏

研究本部・未来ワーク研究所に所属。情報処理学会会員。
専門分野：計算言語学（意味論）、自然言語処理