情報抽出システムへのHPSGパーザの適用

薬師寺 あかね† 宮尾 祐介‡ 建石 由佳‡ 辻井 潤一‡
†東京大学理学部情報科学科 ‡東京大学大学院理学系研究科情報科学専攻
{akane, yusuke, yucca, tsujii}@is.s.u-tokyo.ac.jp

1 はじめに

本稿では、情報抽出システムにHPSG[1]パーザを適用するシステムについてその予備実験の結果を報告する。従来、情報抽出にHPSGパーザのようなフルパーザを適用することは解析速度と記憶容量の面で非効率であるとして避けられてきた。しかし我々は、高精度な前処理系を導入することで効率面での問題を解決し、HPSGパーザを用いた深い解析を利用することを目指す。

従来の情報抽出システム[2, 3, 4]では、タガーやシャローパーザなどの浅い解析の結果のみを利用し、それから正規表現パターンなどで直接情報を取り出していた。情報抽出においては1つの情報に対して多様な自然言語表現が存在することが問題となっているが、浅い解析のみではその多様な変形を吸収することが困難である。従って、パターンを多数列挙する必要があるが、人手でad hocに大量のパターンを書くのは現実的でない。それに対し我々は、HPSGパーザを用いることにより、多様な変形を系統的に吸収することができると考えている。

そこで本研究では、シャローパーザなどの浅い解析器をHPSGパーザの前処理系として導入し、探索空間を制限することで非効率性の問題を解決する。タガー・シャローパーザなどの、浅いが高精度な解析能力に着目し、これらをより深い解析のための前処理として用いることにより解析精度を損なうことなく探索空間を制限することが可能となる。

実際に、前処理として固有表現認識システムを利用した名詞句等のチャンキングと、シャローパーザを利用した品詞・語形・構文構造の曖昧性の除去を行うシステムを実装した。その結果予備実験では、HPSGパーザのカバレッジを下げずに7.8倍の高速化を達成した。これにさらに、現在までに提案されているHPSGパーザの高速化手法[5]を導入することにより、実用的な効率でHPSGパーザを情報抽出に適用できると考えられる。本稿では、この実験結果とその分析を報告する。

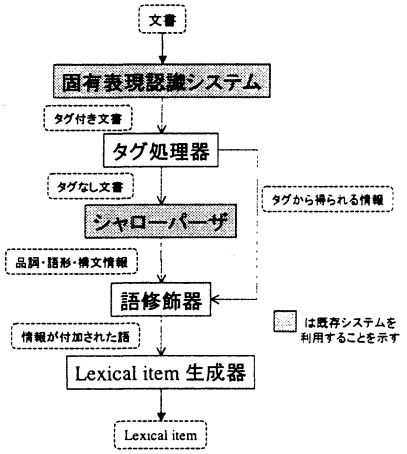

図 1: 前処理系の構成

2 システムの構成

今回実装したシステムではフルパーザとして大規模な英文解析システムであるXHPSGシステム[6]を用いた。また、前処理系は図1のように設計した。本システムの特徴は、固有表現認識システムとシャローパーザという従来の情報抽出で用いられてきた浅い解析を前処理として利用し、それらから得られる情報を積極的に活用してフルパーザの解析効率を向上させることにある。

前処理系は次の5つのモジュールからなる（図1参照）。

固有表現認識システム 文書を入力とし、その文書の分野に固有の表現（語句）に対しタグを付けた文書を出力する。

タグ処理器 固有表現認識システムからタグ付けされた文書を受け取る。文書からタグを除去し、シャローパーザに渡す。またタグによって示されていた情報（1. どの語句がチャンクであるのか、2. どのような種類の固有表現であるのか）を語修飾器に渡す。

シャローパーザ タグ処理器からタグを除去した文書を受け取り、各語の品詞・語形・構文情報を求め語修飾器に渡す。また文書を文単位に切る。

語修飾器 タグ処理器が出力したタグを除去した文書とタグの情報、シャローパーザが出力した各語の情報を受け取る。シャローパーザが出力した品詞・語形情報をフルパーザで利用できる形式（語彙項目の種類および与えられる素性）に変換し各語に付加する（図2参照）。またタグ付けされていた固有表現（語句）をチャンクして1単語とする。

Lexical item 生成器 語に付加された情報とフルパーザ自身の辞書からの情報を統合し、フルパーザで用いる語彙項目のインデックスである lexical item を生成する。

シャローパーザには ENGCG[7] を利用し、品詞・語形・構文情報の曖昧性を削減する。ENGCG は、曖昧性が除去しきれないときにはその曖昧性を残しておくため、正解を落とさずに、明らかに不要である解のみを捨てることができる。したがって、前処理によってフルパーザのカバレッジを落とす危険が少なくなる。また、局所的な構文解析を行うため、多くの未知語に対して正しい品詞を割り当てることができ、専門語が多く辞書が整備されていない分野でも安定したパフォーマンスを持つ。例えば医学論文アブストラクトにおいて、全体の語の99%に対して少なくとも1つ正しい品詞を割り当てるという結果が報告されている [4]。

固有表現認識システム [4, 8] では主に名詞句のチャンキングを行う。これによって、文中の見かけの語数を減らすとともに、各文書分野に特有の表現を一般的な文と同様に解析してしまうのを避ける。

構文解析に成功	89
正しい木を出力	78
前処理系の導入により解析に成功	14
誤った木のみ出力	11
前処理系の導入により誤った木のみに	3
構文解析に失敗	90
前処理系の導入により	11

表 1: 構文解析の結果

3 実験と分析

2節で説明したシステムは、Perl および LiLFeS[9, 10] 言語を利用して実装した。以下の実験は Pentium III Xeon 500MHz, Solaris 7 上で行った。テストデータには、MEDLINE データベース [11] に収録された医学生物学論文アブストラクトに対し専門家が固有表現にタグを付けたもの [12] を利用した。実験にはすでにタグ付けされた文書を利用したため、固有表現認識システムは使用していない。9つのアブストラクト中の、179文（ただし "than"、"as"、"(so) that" 以外の ENGCG が判断した従属接続詞の部分で文を切断したもの）の構文解析を試みた。

前処理系の効果 構文解析の実験結果は表1のとおりである。構文解析が成功し、かつ得られた構文木集合の中に正しい構文木を含む文は78文（44%）であった。前処理を行わない場合には正しい結果が得られないが前処理を行うと正しい結果が得られた文は14文、逆に前処理を行ったことにより正しい結果が得られなくなった文も14文であった。すなわち前処理系の導入によるカバレッジの低下は起こらなかったといえる。

表2は前処理系の導入による効率化の結果を示している。構文解析に成功した文において、エッジ数の平均値は2644から763へ減少し、29%にまで削減された。また構文解析時間の平均値は20秒から2.6秒にまで減少し、速度は7.8倍の高速化を達成した。表2に ENGCG、固有表現の置き換えそれぞれの導

```
The$the~Det~default_det~3rd~definite+~GEN-|default_GEN-
concentrations$concentration~N~default_n~GEN-|default_GEN-~3pl
```

図 2: 語修飾器の出力例

	エッジ数（改善度）		解析時間（高速化率）	
前処理系なし	2644	-	20.0 秒	-
ENGCG による曖昧性削減	1151	44%	4.68 秒	4.3 倍
固有表現語句のチャンキング	2013	76%	13.0 秒	1.5 倍
前処理系全体	763	29%	2.57 秒	7.8 倍

表 2: 前処理系の導入による構文解析効率の改善

メモリ不足	7
前処理系の誤り	34
ENGCG と XHPSG の品詞の不整合	16
ENGCG の誤り	12
文の切断の誤り	5
中間処理での誤り	1
文法の非対応	68
辞書のエントリ不足	13
元テキストのバグ	1

表 3: 構文解析の失敗原因（101 文、ただし原因の重複あり）

入による結果も示す。また語数ごとにまとめた各文での解析時間を図 3 に示す。

現在までに、高速な HPSG パーザ [5] を利用すると解析速度が約 50 倍高速化されるという結果が報告されている。したがって、高速なパーザと本研究で提案した前処理系を統合することにより、実用的な効率で HPSG パーザを情報抽出に適用できると考えられる。

誤りの分析　表 3 は、構文解析に失敗した文について、その失敗原因を分析した結果である。これらの原因のうち、ENGCG と XHPSG で割り当てる品詞の不都合の問題は、特定の語に対して起こる問題であるため、対処は容易にできると考えられる。文法が非対応であったための構文解析の失敗が無視できないが、前処理系では対応できないため、文法の改良を期待したい。

また 202 種類の固有表現のうち、14 種類で語修飾器による主辞の選択に誤りがあったが、XHPSG システムで構文解析する際には、それに起因する誤りは生じなかった。固有表現である語句からその主辞を選択する際の誤りについては、“名詞句 of 名詞句”の場合および括弧が関わる場合に対しては選択すべきでない語の判断が容易につくため、これらに特化した処理は今後試みる価値がある。

4　まとめ

本稿では、HPSG パーザの非効率性を軽減するために、シャローパーザなどを前処理系として利用する手法について述べた。本稿で提案したシステムにより、MEDLINE アブストラクトを用いた実験の結果、精度・カバレッジを犠牲にせずに実用的な速度で HPSG パーザによる構文解析ができる見通しがついた。

今後は、HPSG パーザの構文解析結果から実際に情報抽出を行う予定である。このとき、得られた構文木から、着目する語句に関わる部分木のみを効率よく抽出する必要がある。例えば動詞に着目してイベント情報を抽出する場合、名詞句内部の詳細な構造は不要である。現在行っている名詞句のチャンキングも詳細な構造の解析をしない処理の一つであるが、これを一般化して、不要な部分の曖昧性解消を行わずに済ませることができれば、効率よく情報抽出ができるようになると考えられる。

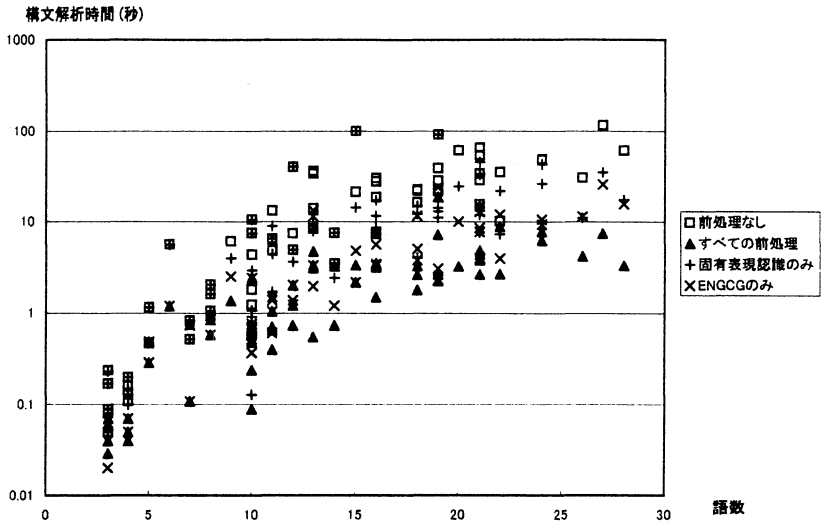

図 3: 前処理系の導入による構文解析時間の減少

参考文献

[1] Carl J. Pollard and Ivan A. Sag. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, 1994.

[2] Roman Yangarber and Ralph Grishman. NYU: Description of the Proteus/PET System as Used for MUC-7 ST. In Science Applications International Corporation, editor, *Message Understanding Conference Proceedings MUC-7*, 1998.

[3] Chinatsu Aone, Lauren Halverson, Tom Hampton, and Mila Ramos-Santacruz. SRA: Description of the IE^2 System Used for MUC-7. In Science Applications International Corporation, editor, *Message Understanding Conference Proceedings MUC-7*, 1998.

[4] Takeshi Sekimizu, Hyun-Seok Park, and Jun'ichi Tsujii. Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in Medline Abstracts, 1998. In Proc. GIW.

[5] Kentaro Torisawa, Kenji Nishida, Yusuke Miyao, and Jun'ichi Tsujii. An HPSG parser with CFG filtering. *to appear in journal of Natural Language Engineering Special Issue — Efficient Processing with HPSG: Methods, Systems, Evaluation*, 2000.

[6] Yuka Tateisi, Kentaro Torisawa, Yusuke Miyao, and Jun'ichi Tsujii. Translating the XTAG english grammar to HPSG. In *Proceedings of TAG+4 workshop*, 1998.

[7] Atro Voutilainen. Designing a (finite-state) parsing grammar. In Emmanuel Roche and Yves Schabes, editors, *Finite-State Language Processing*. A Bradford Book, The MIT Press, 1996.

[8] C. Nobata, N. Collier, and J. Tsujii. Automatic Term Identification and Classification in Biology Texts. In *Proc. NLPRS*, 1999.

[9] T. Makino, M. Yoshida, K. Torisawa, and J. Tsujii. LiLFeS — towards a practical HPSG parser. In *Proc. COLING-ACL, '98*, pages 807–811, 1998.

[10] Yusuke Miyao, Takaki Makino, Kentaro Torisawa, and Jun'ichi Tsujii. The LiLFeS abstract machine and its evaluation with the LinGO grammar. *to appear in journal of Natural Language Engineering Special Issue — Efficient Processing with HPSG: Methods, Systems, Evaluation*, 2000.

[11] National Library of Medicine. MEDLINE, 1999. available in `http://www.ncbi.nlm.nih.gov/PubMed/`.

[12] T. Ohta, Y. Tateishi, N. Collier, C. Nobata, K. Ibushi, and J. Tsujii. A Semantically Annotated Corpus from MEDLINE Abstracts. In *Genome Informatics*. Universal Academy Press, Inc., 1999.