論文の内容の要旨

論文題目 ： Preference Dependency Grammar (PDG): Sentence Analysis Method Based on Integrated Multilevel Preference and Constraint
（ 選好依存文法：多層の選好／制約知識を統合した文解析方式 ）

氏　　名 ：平　川　秀　樹

1. 自然言語解析システムの設計課題

自然言語文解析(NLA: Natural Language Analysis)システムは、入力文に対して言語知識を適用して最も適切な解釈を出力する。一般に、NLA システムは、形態素、構文、意味、文脈などの多層の言語知識を扱う必要があり、各層の知識記述のベースとして、句構造木、依存木、述語論理式など様々な文解釈記述スキーマが提案されている。これらのスキーマは、知識記述の質や能力、すなわちシステムの性能を規定するため、どの様なスキーマを採用するかは NLA システム設計上重要な課題の１つである。また、言語知識は、計算機処理の視点から見て、入力文に対して不可能な解釈を排除する「制約知識」と可能な解釈の優先順序付けを行う「選好知識」の２種類に分類される[1]。制約知識は自然言語の各層で生じる解釈の組み合わせ爆発の抑制に有効であるが、その過度の適用は正解解釈の枝刈りによる性能劣化を引き起こしてしまう。一方、自然言語の各層の選好知識は、相異なる解釈を支持することもあり、全体として最適な解を得るためには各層の知識を統合評価する必要がある。このように、多層の知識、制約/選好知識の統合をいかに行うかが NLA システム設計のもう１つの重要な課題である。

2. 選好依存文法の概要と特徴

本論文は、上記 NLA の課題に焦点をあてて設計された文解析の枠組みである選好依存文法(PDG: Preference Dependency Grammar)を提案する。PDG の文解析モデルとして、Meaning Text Theory(MTT)[Melcuk 88, Kahane 03]の考え方をベースに多層の言語知識を扱う「多レベル圧縮共有データ結合(MPDC: Multilevel Packed Shared Data Connection)モデル」を提案する。MTT は、テキストから意味までを扱う言語理論体系であり、自然言語をテキストと意味の対応を規定するものと捕らえ、形態素、構文といった複数の解釈層のデータ構造を介してテキストと意味を対応付ける。MTT は、基本的に解析/生成の双方向モデルであるが、文生成に軸足をおいており、文解析に必要な解釈多義や選好知識の扱いが不十分である。この課題に対して、MPDC では、各層の文解釈を圧縮共有データ構造で結合することにより組合せ爆発を抑制しながら解釈多義性をモデルに導入し、また、各層の選好知識を統合処理する仕組みも導入している。

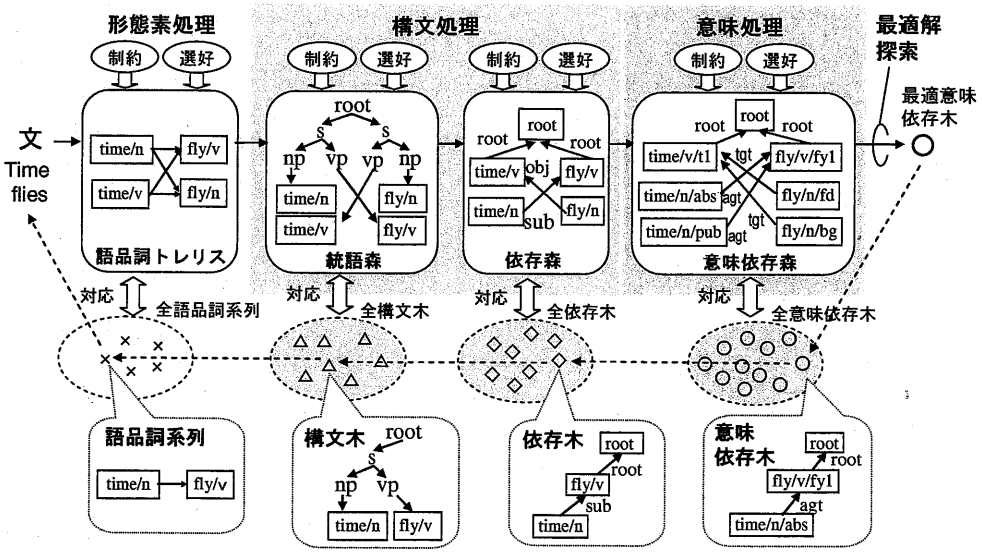

図 1 PDG の文解析実装モデル（例文:“Time flies”）

PDG の実装モデル(図 1)は形態素、構文、意味の各層の解

[1] 文解釈を生成する生成知識も存在するが、本文に述べた様に可能な文解釈を規定するという意味で制約知識に含まれる。

釈記述スキーマとして語品詞系列[2]、句構造木、依存木、意味依存木の 4 種を用い、それぞれ、語品詞トレリス、統語森、(機能)依存森、意味依存森の 4 つの圧縮共有データ構造で接続している。各データ構造は、隣接するデータ構造との対応関係を持ち、意味構造とテキスト(文)のマッピングが規定される。PDG の開発では、構文レベルまでをカバーする構文解析版 PDG と意味レベルまでをカバーする意味解析版 PDG の 2 段階を想定しており、各層の選好知識のスコアは、前者では(機能)依存森、後者では意味依存森に集約され、依存森から文に対する最適解釈が探索される。

本稿で提案する「依存森」は、次の PDG の特長を実現するキーとなるデータ構造である。
(a) 構文層において、句構造と依存構造の両者を解釈記述スキーマとして統合利用する
(b) 構文層と意味層を依存構造に基づく圧縮共有構造で橋渡しする
(c) 制約知識と選好知識を統合した最適解探索を実現する

句構造と依存構造は、文の構文構造をそれぞれ異なった側面から表現する代表的記述スキーマであり、その統合利用により両者の利点を活用できる。統合利用には、句構造と依存構造のマッピングが必要であるが、従来、これを目的とする研究はあまり例がない。構文グラフ[Seo & Simmons 89]は、両者のマッピング手法を提案しているが、対応関係が不完全であり、MPDC モデルに利用できない(後述)。依存森は、構文グラフの問題を解決することにより、従来手法では困難であった(a)を実現した[平川 06a]。これにより、(b)の構文層と意味層の橋渡しが可能となり、図 1 の実装モデルを実現している。また、依存森は、制約知識と選好知識を記述するための制約/選好マトリックスを有し(後述)、各層の制約知識と選好知識がこの上で統合される。グラフ分枝法と呼ぶ新規提案のアルゴリズムにより、任意の依存関係間の制約条件を満足する最適な解を依存森から探索可能である。

3. 圧縮共有データ構造

[問題設定] 図 1 の PDG 実装モデル実現のための圧縮共有データ構造を設定する。語品詞系列と句構造木に対しては、それぞれ既存のデータ構造、即ち、語品詞トレリスと統語森が利用できるが、依存木、意味依存木に対する適切なデータ構造が必要である。

[従来技術・課題] 構文解析を行い圧縮共有型の依存構造を生成する手法として、意味依存グラフ[平川 02]、構文グラフが提案されているが、前者は日本語を前提とし汎用性に欠け、後者は句構造木と依存木の完全で健全な対応がとれないという問題[3]があり、MPDC モデルに利用できない。

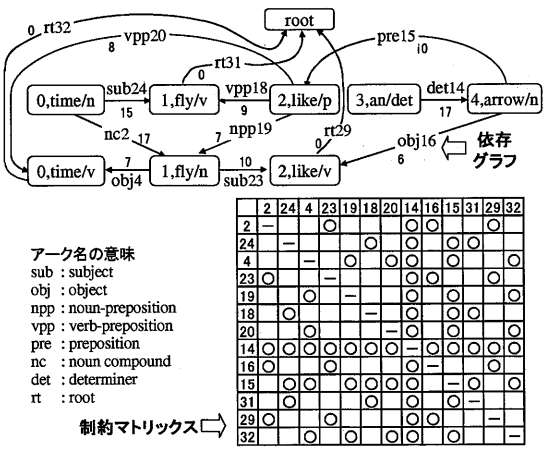

図 2 依存森の例

[新手法の提案] これらを解決するデータ構造として依存森を提案する(図 2)。依存森は、ノードとスコア付きアークからなる依存グラフとアークの共起関係を規定する制約マトリックスから構成される圧縮共有データ構造であり[4]、書き換え規則と部分依存構造よりなる拡張文脈自由文法規則を用いて、(a)チャート法に基づく構文解析によるヘッド付き統語森の生成、(b)ヘッド付き統語森からの依存森の生成、(c)依存森の縮退処理(コンパクト化)の 3 つの処理により構築される。統語森と依存森の間には、完全性と健全性が成立し[5]、MPDC モ

---

[2] 語品詞とは単語とその品詞のペアをいう(ex. 単語"time"に対し、"time/n","time/v"など)

[3] 完全性とは「統語森中の全ての句構造木に対して対応する依存木が構文グラフに存在する」ことであり、健全性とは「構文グラフ中の全ての依存木に対して対応する句構造木が統語森中に存在する」ことである。構文グラフは、健全性が成立しない[平川 06]ため、存在しない解釈を生成する場合がある。

[4] 後述するようにスコアの部分をマトリックスで表現する、より汎用のモデルも存在する。

[5] 依存木から句構造木を生成する再帰的アルゴリズムの存在を示す事により証明する。(本論文付録 B)

デルに利用可能である。また、依存森は、任意のアーク間の共起関係を記述可能であり、非交差制約を満たさない依存構造(Non-projective dependency)なども扱えるという柔軟性を持つ。

4. 最適解探索

[問題設定] 依存森から最適な解釈（依存木）を探索するタスクは、「スコア付き依存グラフから依存木の整合性を規定する制約条件を満足する最大のスコアを持つ木の探索」として形式化される(図 3)。この枠組みのもとで従来様々な依存グラフに対する探索法が提案されている。

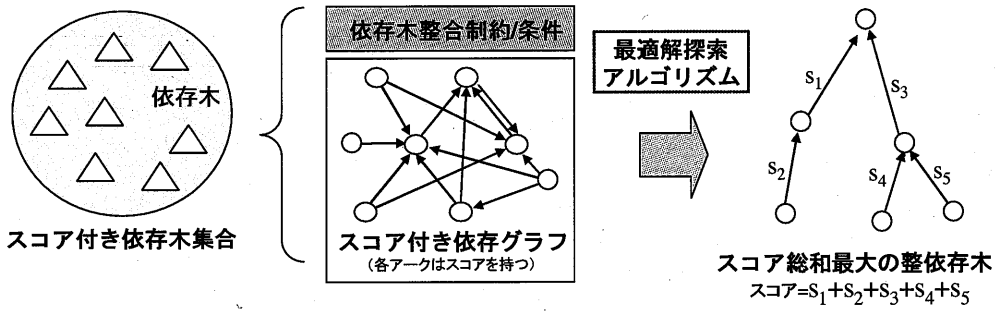

図 3 スコア付き依存グラフからの最適解探索

[従来技術・課題] 言語非依存の基本的な依存木整合制約条件には、

(a) 被覆制約 ：入力文中の全ての語に対応するノードが依存木に存在する
(b) 単一役割制約：依存木中のノードは１つの単語に対応する
(c) 非交差制約：依存関係の交差が存在しない
(d) 多重格禁止制約：述語の格スロットは１つの要素のみが占有する

などが存在する。(a),(b)を満足する Chu-Liu-Edmonds 法は高速である[McDonald 05]が、語品詞からなる依存グラフ(依存森など)に適用できないという問題がある。Dynamic Programming をベースとするアルゴリズムで(a)～(c)を満足する手法が広く提案・利用されているが、(d)や依存森の制約マトリックスの任意の共起制約を扱えないという問題がある。

[新手法の提案] 依存森は、高い制約記述能力を持つため従来手法では最適解の探索が困難である。このため、PDG では、分枝限定法[6]に基づき依存森から制約マトリックスを満足する最適解を計算する「グラフ分枝アルゴリズム」を提案する[平川 06b]。

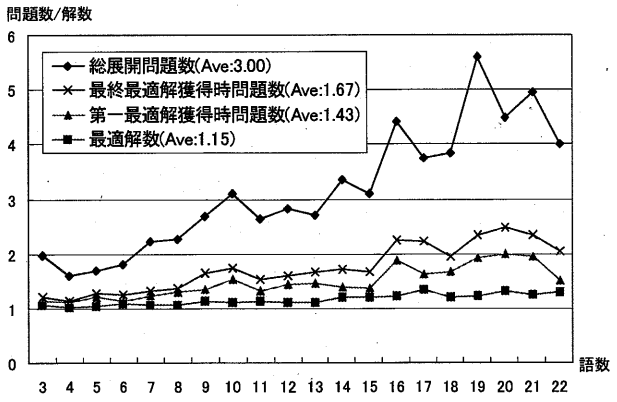

図 4 グラフ分枝法の問題展開数

更に、PDG では、選好スコアの記述の枠組についても従来型(図 3)のアークスコアのみのモデル（単項選好モデル)からアーク共起の選好スコアも加えたモデル(二項選好モデル)に拡張し、記述力を強化した。二項選好モデルの依存森は、依存グラフ、制約マトリックス、選好マトリックスより構成される。グラフ分枝アルゴリズムは、二項選好モデルに適応可能である。

[評価実験] グラフ分枝アルゴリズムの計算量は、指数関数オーダーであるが、分

6 問題を解決容易な子問題に展開する操作(分枝操作)と不要な子問題を枝刈りする操作(限定操作)で最適解を探索する。NP 完全問題など Greedy な手法では解けない問題に適用可能。

枝限定法の上界値関数や許容解関数を最適化することにより、早期の段階で最適解に到達する(平均問題展開数が 1.43/文)ことが実験より示され(図 4)、性能的見通しが得られた。また、「部分問題展開数を 10 に限定する」という戦略を採用した場合に、実験では、99.8%の文において 1 つ以上の最適解が得られた。

5. スコアリング

スコアリングは、各レベルの選好知識を依存森の選好マトリックスのスコア(単項モデルの場合はアークスコア)に統合する処理である。現状のスコアリング方式では、各レベルの選好知識を単項ノード、単項アーク、二項ノード、二項アークの 4 種類の選好スコアにヒューリスティック関数を用いて変換し、統合する。

6. 実験評価

英語技術文 12.5 万文をクローズドデータとオープンデータに分割し、既存の文解析システムを用いて前者より選好知識データ(語品詞頻度、依存アーク頻度、語品詞 BIGRAM 頻度、依存アーク共起頻度)を、後者より正解依存木データを抽出した。また、構文解析版 PDG を実装し、英語基本文法(907 CFG 規則)を用いて、選好知識統合による解析精度の向上実験、選好知識による解の絞込み能力測定実験などを実施した。選好知識統合実験(図 5)では、単項ノードと単項アークならびに単項アークと二項ノードの 2 つの組合せが最高のアーク正解率[7]88.3%を示した。これは、ベースライン(選考知識無し)に対して 10.9%の改善となり、選好知識統合利用の効果が確認された。

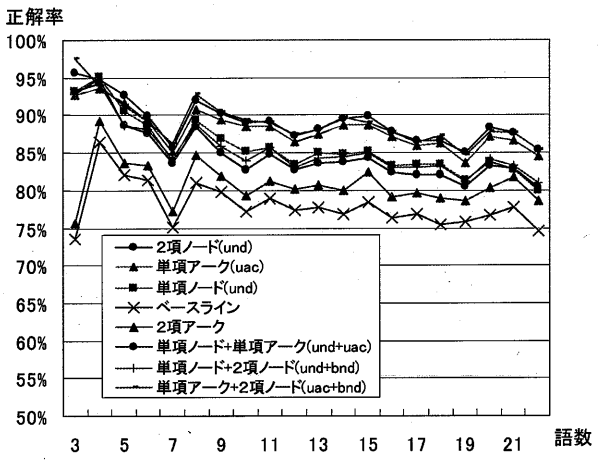

図 5 選好知識別アーク正解率

7．今後の課題

各処理の最適化や C 言語実装などによる PDG 実システムの開発、意味構造への展開、双方向モデルの検討などが今後の主な研究課題である。

参考文献

[Melcuk 88] Mel'cuk, I.A., "Dependency Syntax: Theory and Practice", State University of New York Press, 1988

[Kahane 03] Kahane, S., "The Meaning-Text Theory, Dependency and Valency", Handbooks of Linguistics and Communication Sciences 25 : 1-2, Berlin/NY: De Gruyter, 32 p.546-569, 2003

[Seo & Simmons 89] J. Seo and R. F. Simmons, ``A Syntactic Graphs: A Representation for the Union of All Ambiguous Parse Trees", Computaional Linguistics, Vol.15, 1989.

[平川 02] 平川, "最適解探索に基づく日本語意味係り受け解析", 情報処理学会論文誌,Vol.43, No.03, pp.696-707, 2002

[平川 06a] 平川, "統語森に対応する圧縮共有型依存構造「依存森」について",自然言語処理, Vol.13, No.3, pp.37-90, 2006

[McDonald 05] McDonald, R., Pereira, F., Ribarov, K. and Hajic, J.,"Non-projective Dependency Parsing using Spanning Tree Algorithms", Proceedings of HLT-EMNLP, pp.523-530, 2005

[平川 06b] 平川, " Graph Branch Algorithm: An Optimum Tree Search Method for Scored Dependency Graph with Arc Co-occurrence Constraints",自然言語処理, Vol.13, No.4, pp.3-32, 2006

[7] 出力依存木中の正解アークの割合