

選言的素性構造の効率的単一化手法

2Q-4

高橋 誠 谷田 泰郎

田代 敏久

(株)東洋情報システム 自然言語処理グループ

ATR 音声翻訳通信研究所

1. はじめに

本稿では、日本語文解析の文法記述の枠組みとして採用している Kasper の選言的素性構造単一化手法をベースに、1) 簡単な近似化によって単一化をさらに効率化する手法、並びに 2) 選言的素性構造から尤もらしい素性構造を優先的に高速に取り出す手法を紹介する。

2. 研究の背景

2.1 選言的素性構造の単一化に関する問題

選言的素性構造を選言標準形で表現し、選言要素の全ての組合せに対して無矛盾性の検査を行なうような素朴な方法では、選言数の指数に比例する計算量が必要となる。この問題を解決するために、様々な選言的素性構造の表現方法および単一化手法が提案されている[1][2]。本稿では、Kasperの手法をベースとして、その効率化を図る(3. 参照)。

また、選言的素性構造から、選言を含まない素性構造(以下、FS)への展開処理も、同様指数的な問題である。この手続きにはグラフの単一化という、非常に負荷のかかる処理を要する。本稿ではこの問題に取り組む、全てのFSを取り出すのではなく、優先度順に順次取り出す枠組みを検討し、効率化を図る(3. 参照)。

2.2 Kasper の手法

ここで、本稿のベースとなるKasperの手法[1]を略説しておく。Kasperは、下で説明するようなデータ構造により展開を遅延し、無矛盾性検査を段階的に行なうことで、平均的な計算量を選言数に対して線形ですむようにした。

データ表現 Kasperは、素性記述の情報を確定した部分(定部)と不確定な部分(不定部)とに分け、特殊なAND/OR木によって選言的素性構造(以下、FDESCと呼ぶ)を表現している。Kasperの手法による文法記述の例とそのデータ表現を図1、2に示す。

定部: FS。有向グラフによって表す。

不定部: 選言の集合。各選言はFDESCの集合。

単一化アルゴリズム Kasperの単一化手続きは、3つの段階により成る。第1段階では、2つの素性構造の定部を単一化して無矛盾性を検査する。第2段階では、定部の単一化結果と不定部の和との無矛盾性を再帰的に検査する(近似的無矛盾性検査)。第3段階では、完全な無矛盾性検査を行なう。但し、各段階で失敗した場合、それ以降の処理は省かれる。

```
(DEFLEX-NAMED に関して-POSTP に関して POSTP-OPTN
(<!M> ==
[[SYN [[SUBCAT [[FIRST [[SYN [[SUBCAT (:LIST )]]]
[SEM ]]]]
[REST (:LIST )]]]
[HEAD [[POS P] [FORM に関して]
[COH [[SYN [[HEAD ]]]]
[SEM ?X01[ ]]]]]]]]
[SEM ?X01]])
(:OR ;; disj1
((<!M SYN HEAD COH SYN HEAD POS> = V))
((<!M SYN HEAD COH SYN HEAD POS> = P)))
(:OR ;; disj2
((<!M SYN SUBCAT FIRST SYN HEAD POS> = N))
((<!M SYN SUBCAT FIRST SYN HEAD POS> = N-ADV))))
```

明朝体箇所が定部。ゴシック体箇所が不定部にあたる。

図1. 文法記述(辞書記述)の例

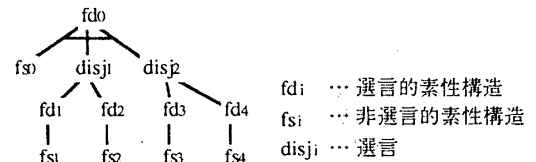


図2. 文法記述のデータ表現(図1. に対応)

査を行なう。但し、各段階で失敗した場合、それ以降の処理は省かれる。

第3段階は非常に効率が悪く、極く限定された場所で行うべきだとされる。著者らはこの段階を全く行なわない。それは、最終的にFDESCからFSを取り出す展開処理が、完全な無矛盾性検査を兼ねると見なされるからである。

3. 単一化の近似化による無矛盾性検査の遅延

3.1 アルゴリズム

本稿では、Kasperの単一化手法の第2段階に対して、さらに近似化を施し無矛盾性検査を遅延する。

Kasperの手法における第2段階の無矛盾性検査は、不定部の各選言に対して再帰的に行なわれるが、本手法では再帰を省略する。すなわち、AND/OR木の根における定部と不定部のみに注目し、定部の単一化結果と不定部の和に含まれる選言要素の定部との無矛盾性検査のみ行ない、選言要素の不定部から下の検査を行わない。

3.2 本手法の問題点

生成規則適用時に素性記述の制約が緩くなるため、構文解析で無駄な探索を行ない、場合によっては誤った構文木を作ってしまう¹⁾。このことは、Kasperの近似化手法にも言えることであるが、本手法のような大胆な近似化では、より可能性が高くなる。

¹⁾ バックトラックを要するため、その分非効率になる。

また、無矛盾性検査を遅延しているため、出力される FDESC に矛盾を含む冗長な選言が残り、FS の展開に負荷がかかる。この負荷は、4. で述べる方法により削減される。

4. 非選言的素性構造の優先度順取り出し手法

4. 1 概要

著者らはこれまで、FDESC を展開する際、全ての FS を取り出し、その中から解を決定していた。この全解探索のやり方では、処理の負荷が大きいうえ、単一の解を決定する基準がなかった。本稿では、この問題に取組み、FS を優先度順に順次取り出す枠組みを検討する。

4. 2 優先度情報

本稿では単純に文法記述における選言要素の使用頻度を優先度とする。この場合、意味的、語用論的に正しいものを選ぶ保証はない。著者らの使用文法では、拡張性を考慮し汎用的に記述されており、使われない選言が実は多い。これは特に、自由な語順に対応するための subcat 素性のかき混ぜ (subcat scrambling) において顕著である。例えば、主格/与格/対格の語順は、省略を考慮に入ると 24 通りあるが、実際の発話で必ずしも全ての語順が出現するとは限らない (表 4)。使用頻度を基準に、実際にはあまり使われない選言の展開を後回しにするこの方法は、対象問題に依存した戦略として位置付けられるだろう。

表4.「送る」を中心とした 必須格の語順の分布

必須格語順	比率
対格のみ	44.70%
なし	39.73%
与格のみ	8.58%
与格/対格	3.84%
対格/与格	2.03%
主格のみ	0.45%
主格/対格	0.45%
主格/対格/与格	0.23%
その他	0%

(¹) ADD 国際会議に関する問い合わせ 238 会話 / 7,968 文のうち、「送る」を含む 443 文より抽出)

4. 3 選言要素の使用頻度学習とその利用

文法記述における選言要素の使用頻度は、コーパスの発話を構文解析することにより学習したあと、図 4. のように文法記述の構造に付与され、構文解析に供される。

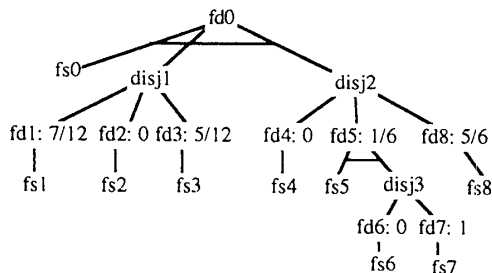


図 4. 頻度情報を付与した選言的素性構造

こうして構文解析を行なった後に得られる FDESC には、文法記述に付与されていた選言要素の使用頻度が残る。これをもとに、Best First Search により FS を取り出す。

5. 実験および考察

国際会議の予約に関するモデル会話 50 文から、選言要素の使用頻度情報を学習し、クローズ、オープン各 5 文に対して、Active Chart Parser により解析の実験を行なった。FDESC の優先度順展開では、第 1 解のみを出力する。

実験の結果を表 5. に示す。

表 5. 実験結果

テストデータ	クローズ	オープン
展開時間比 (優先度順探索/全解探索)	1.48%	1.69%
解析全体の時間比 (新 Unity & 優先度順展開/従来版)	68.90%	70.31%

FDESC の展開の実行時間については、従来の全解探索から優先度順探索に変えることにより、クローズデータで 1.48%、オープンデータで 1.69% にそれぞれ短縮できた。オープンとクローズの差は、僅かなものであるが、学習により無駄な探索を減らしうることを示している。

また、解析全体の実行時間は、本稿の単一化手法と FDESC の優先度順探索を組合せると、従来の方法に比べ、クローズデータで 68.90%、オープンデータで 70.31% にそれぞれ短縮できた。単一化の近似化による無駄な探索やバックトラックによる悪影響は、近似化自身の効果と全解探索をやめたことによる効果により、平均的に全く問題のないものとなっている。

なお、本手法による解析の精度を調べるため、優先度順展開により全解を求めて正解の順位を見たところ、構文解析の精度上の問題もあり必ずしも正解が候補 FS の中に存在するわけではないが、従来の方法に比べて致命的な精度の劣化は見られなかった。

6. おわりに

本稿では、Kasper の FDESC 単一化手法の近似化方法、並びに、FDESC を順序付きで展開する手法を紹介し、実験によりその効果を示した。

謝辞 本研究の機会を与えて下さいました ATR 音声翻訳通信研究所の山崎社長、森元室長はじめ第 4 研究室の皆様へ感謝いたします。

参考文献

[1] Kasper, R.: "A Unification Method for Disjunctive Feature Discriptions", 25th ACL, pp235-242, 1987.
 [2] Eisele, A.; Dorre, J.: "Unification of disjunctive feature descriptions.", 26th ACL, pp.286-294, 1988.