TR-IT-0324

統計的手法による部分木併合

A Statistical Method for Merging Partial Parses

竹澤 寿幸　荒川 直哉†　森元 逞‡
Toshiyuki TAKEZAWA　Naoya ARAKAWA†　Tsuyoshi MORIMOTO‡

2000. 1.11

内容梗概

自然な話し言葉を対象に通常の句構造解析を行おうとすると、非文法的な表現に対しては全体の構造が得られず、しばしば断片的な部分木構造の集まりを得る。本稿では、これらの部分木構造を併合して、文全体の構造解析を与える試みについて述べる。頑健な併合処理を実現するために統計的手法を用いた。そのための統計データは旅行会話コーパスから得た。本稿では、併合処理について詳しく述べるとともに、文節単位に分解された木構造データベースを本手法に基づいて再構成する実験について報告する。

ATR 音声翻訳通信研究所
ATR Interpreting Telecommunications Research Laboratories
† 現在、シーエーアイ (株)
‡ 現在、福岡大学工学部電子情報工学科

1　はじめに

この論文では、通常の句構造解析による処理が困難であるような話し言葉の言語現象を統計的に解析する手法について述べる。自然な話し言葉を機械で処理しようとする場合、非文法的な表現、言いよどみなどのために通常の句構造解析は文全体の解析結果をもたらさないことがある。このような場合でも、句構造解析の結果として、発話入力中の形態素、文節、句、節などを要素とする木構造を得ることができる。複数のこのような木構造が意味上１つの文を構成する（あるいは１つの述語に支配される）と考えられる場合、それらを併合して処理できれば、意味の解析あるいは翻訳処理に役立つ。この論文で部分木とは、このような文より小さい単位の木構造を指し、部分木併合とは部分木を併合して文に相当する構造を作るようなプロセスを指す。

ここで提案する手法においては、句構造解析が出力する部分木を依存（係り受け）構造に変換したものを併合の対象とし、統計的に依存関係を持つ可能性の高い部分木同士を併合する。各部分木と、それが依存可能な他の部分木中の表現との間の依存確率を依存構造解析を施したコーパスから抽出して、併合のための選好情報として用いる。また、それらの部分木の相互位置や、文末が想定される位置なども依存尤度を計算するために用いる。

2　背景

話し言葉の機械処理においては、非文法的な表現、言いよどみなどの「非流暢性」(disfluency) のために通常の句構造解析が失敗することがしばしばある。こうした「流暢でない」発話でも、もしそれらが意味的に解釈可能であれば、解析できるような頑健な構文・意味解析過程の実現が望まれる。ここで「非流暢性」には、言い直し、「あのー」「えーと」などの間投詞の挿入、助詞などの省略といった現象がある。これらの現象の分析については文献 [10] を参照されたい。

構文解析に失敗して取り残された要素を併合する研究としては文献 [1, 2, 3] が挙げられる。Stallard et al. [2] では意味的な整合性、Nasukawa [3] では談話（テキスト）中に現れた依存構造を利用して併合を行っている。また、田代ら [4] は部分木併合手法について議論を行っている。

近年、統計情報が自然言語処理に取り入れられてきている。統計的な手法の利点は、選好情報を統計的に最適化した形（すなわちアドホックでない形）で自然言語処理過程に取り入れることができることである。厳格な制約の代わりに選好情報を利用することにより、自然言語処理はより柔軟かつ頑健になることが期待される。また、統計的な選好情報の利用により、解析結果には自然に選好順位が与えられるため、解析の際の多義性の解決に役立てることができる。文献 [6] は統計的な手法を用いた頑健な依存文法的構文解析を提案している。また、文献 [9] は文節間係り受け距離の統計的性質とその係り受け解析への応用についての研究である。

音声認識と句構造解析を統合して音声認識率の向上を図ることが試みられている [7] が、句構造解析を音声入力に施す場合、発話の「非流暢性」や文末の不明確さなどから発話を文よりも短い単位で句構造解析することがよい場合がある。文献 [7] は自然な発話中にあらわれる短いポーズで区切られた区間が文脈自由文法で記述可能である場合が多いことを報告している。このような文より小さい発話単位の句構造は文より小さい単位の木構造、すなわち部分木となる。これらの部分木を併合して文の解析結果を得ることができれば、音声認識から文の解析まで統合した形の音声言語処理を実現でき、また計算資源の経済という点からも有利である。次頁の図１は統合化された音声言語解析の流れを部分木併合を含めて図式化したものである。

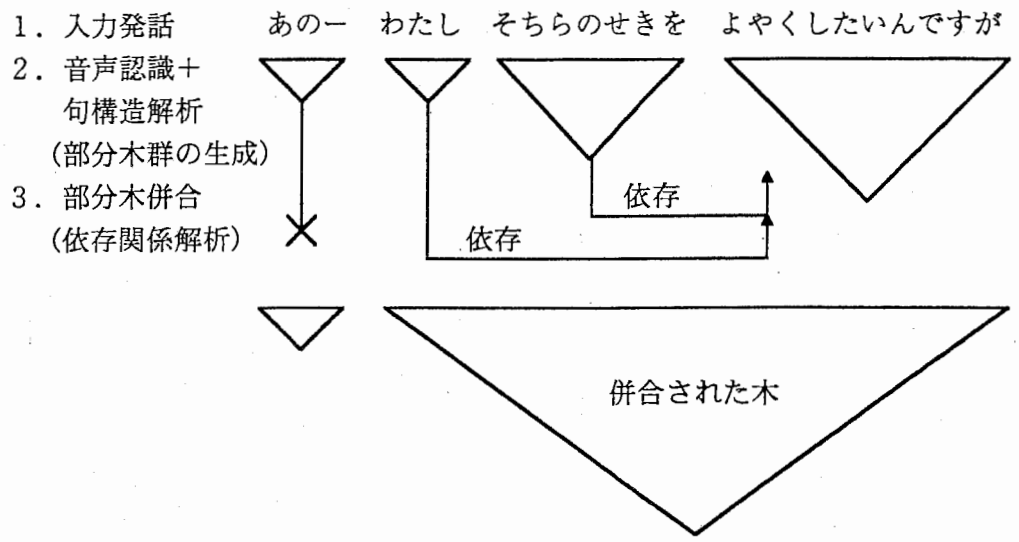

図１：音声認識と部分木併合

発話の入力から、ポーズで区切られた部分ごとの句構造解析、部分木併合までの流れ

「あのー」や「わたし」のように一単語で部分木を構成するものもあれば、「そちらのせきを」のように２文節で１部分木を構成するものもある。入力中の間投詞「あのー」を要素とする部分木は併合されずに残っていることに注意されたい。

3　部分木併合処理の実際

3.1　部分木併合処理の動作

入力として依存構造木（部分木）の列を用いる。依存構造を用いる理由は、依存構造のレベルでは併合は係り受けの決定として解釈できることが多く、人間にとって意味的に理解しやすいためである。ただし本研究で扱う依存関係は文節間の係り受けに限らず、任意の文法的支配関係を含む。

以下、併合の対象となる２つの部分木を依存候補対と呼ぶ。その一方を依存元候補と呼び、依存元候補が依存する可能性のある他方の部分木のノード（通常は単語を表現する）を依存先候補と呼ぶ。依存候補対の依存尤度は、依存元候補のトップノードを含むパターンと依存先候補の対の文法的、意味的な整合性に依存する。日本語ではトップノードを含むパターンは通常その部分木を構成する形態素列中の最後の部分（例えば最後の文節や助詞）に対応する。

例：文「わたしは鈴木と申します」の依存構造を以下に示す。依存関係は ARG というラベルで表示されている。

```
[[見出し ます]
 [品詞 <助動詞>]]]
 [ARG [[見出し 申し]
       [品詞 <本動詞>]]]
       [ARG [[見出し と]
             [品詞 <格助詞>]]]
             [ARG [[見出し 鈴木]
                   [品詞 <人名>]]]
       [ARG [[見出し は]
             [品詞 <係助詞>]]]
             [ARG [[見出し わたし]
                   [品詞 <代名詞>]]]]]]]
```

上の依存構造を文節ごとに分解した３つの部分木の依存構造は次のようになる。

```
[[見出し は]                    ←トップノード
 [品詞 <係助詞>]]]
 [ARG [[見出し わたし]
       [品詞 <代名詞>]]]]
```

```
[[見出し と]                    ←トップノード
 [品詞 <格助詞>]]]
 [ARG [[見出し 鈴木]
       [品詞 <人名>]]]]
[[見出し ます]                  ←トップノード
 [品詞 <助動詞>]]]
 [ARG [[見出し 申し]
       [品詞 <本動詞>]]]]
```

ここで「わたしは」に対応する最初の部分木を依存元候補とすると、依存先候補としてそれ以外の部分木の各ノード（「鈴木」「と」「申し」「ます」に対応）が考えられるが、「申し」に依存するのが正解である。この場合「わたしは」の最後の部分が係助詞「は」であることから、それが用言に係ることが統計的に推測できる。

部分木併合のアルゴリズムは以下のとおりである。

1．各部分木のトップノードを含む部分と後続する部分木（複数）のすべてのノードの組み合わせに対して依存尤度（下記3.2 参照）を計算する。
2．一定値以上の依存尤度を大きい順に並べ、それらの依存尤度を持つノードの組み合わせを併合する。用言の係り受け併合にあたっては、格フレームのチェックを行う（下記 3.3 参照）。

このプロセスがどう動くかを例を用いて見てみよう。

例：「私、ハンバーグを、ハンバーガーを、注文したんです。」

この文を読点のところで分割したものから作った部分木を併合することを考える。まず、代名詞の「私」および「を」で終わる２つの文節は用言「注文」へ依存するという可能性に対し高い尤度が統計的に与えられる。この例では言い直しが現われているが、「ハンバーガーを」と「注文した」が距離的に近いために「ハンバーグを」と「注文した」より強い依存尤度が与えられ、まず「ハンバーガーを」と「注文した」の併合が行われる。この併合により「注文した」の「を」格が消費されてしまうため「ハンバーグを」が「注文した」に係るという解釈は格フレームチェックにより却下され「ハンバーグを」は係り先なしのゴミと判定される。

3.2　依存尤度の計算

本実験では、依存尤度は３つの尤度関数(それぞれ０以上１以下の値を取る）の積とした。以下、各関数について説明する。

A. 統語的構成による依存関係の統計的整合性に関する関数

これは一種のバイグラム尤度関数であり、依存元候補のトップにある文節（あるいは文節より小さな部分構造)と依存先候補(通常は単語）のノードそれぞれの見出し・品詞パターンから、それらが依存関係にある尤度を返す。尤度は、コーパス中で依存候補対のパターンが依存元候補が先行する形で同一文中に生起した場合に依存が成立する条件付き確率である。この確率は依存候補対の依存元・依存先パターンをＰＰとすると次式で表わされる。

・式１：

$$\frac{\text{ＰＰが実際に依存関係にある回数}}{\text{ＰＰが同じ文中で依存元パターンが先行する形で現れる回数}}$$

この確率値は併合処理とは別に統計を取り、あらかじめテーブルに格納しておいたものを用いる（下記 3.4 参照）。見出し情報は意味情報を含んでいるから、意味と依存の関係も、この手法によりコーパス中に現われる組み合わせの範囲内であれば捉えることができる。

尤度テーブル検索の際のパターンマッチングは一種の最長一致法による。以下に例を用いて説明を試みる。

例：「予約を」が「キャンセル」に依存する尤度を尤度データベース（DB）か

ら求める。以下の例では依存候補対のパターンを上段が見出し、下段が品詞という形で示す。各段階で、パターンがＤＢ中にあれば、ＤＢ中の尤度を用い、見つからなければ次の段階へ進む。

	依存元候補	依存先候補
1.	(を (予約))	キャンセル
	(格助詞 (サ変名詞))	サ変名詞

↓ 見つからなければ

自立語見出しをワイルドカード化

↓

2.	(を (*))	*
	(格助詞 (サ変名詞))	サ変名詞

↓ 見つからなければ

自立語見出しを元に戻し、依存元候補を短くする

↓

3.	(を)	キャンセル
	(格助詞)	サ変名詞

↓ 見つからなければ

自立語見出しをワイルドカード化

↓

4.	(を)	*
	(格助詞)	サ変名詞

↓ 見つからなければ

5. 尤度０

B. 依存先候補の位置による尤度の関数

依存元候補の部分木および依存先候補の相互の位置関係または依存先候補の文末との位置関係から依存尤度を計算する。このためにコーパスから依存関係の物理的な距離のヒストグラムを取り利用する（下記 1 および 2)。

1. 依存先トップノードと依存元の品詞から、依存関係が常に隣接関係である（例えば助動詞と動詞との直接依存関係）とヒストグラムにより判断できれば、隣接している依存関係候補には尤度１、隣接していない依存関係候補には尤度０を与える。
2. 上記ヒストグラムから依存関係が隣接しない依存関係を持つと判断される場合、尤度として指数関数 $e^{-k\lambda d}$ を用いる。ここでλは上記ヒストグラムから得られる依存関係の種類に関する平均依存距離の逆数、d は依存関係候補間の距離、k は定数である。この尤度設定より近い部分木同士が優先的に併合される。
3. 感動詞や接続詞など文全体に係る語は、依存構造解析では文の最後の表現に依存することが多いため、今回の実験では、接続詞、感動詞などを文末表現に係ると想定した。よって、それらの語には依存先候補が文末表現である確率（次項Ｃ参照）を依存尤度として与える。

C. 文末確率の関数

音声言語処理においては、１人が続けて複数の文を発話することもあり、文末は常に明確ではない。依存関係は文末を超えて成り立たないから、文末が依存関係候補の間にある確率が大きくなると、それらの候補の間の依存関係尤度はその分小さくなる。従って、文末確率に応じた依存尤度は２つの部分木（依存候補対）の間に文末が来ない確率として定義される。依存元候補と依存先候補の間に部分木の切れ目がいくつかあるとすると、尤度はそれらの各々の切れ目に文末が来ない確率（1－文末が来る確率）の積になる。各部分木間に文末が来る確率は、先行する部分木の最後の語の品詞と次の部分木の最初の語の品詞の関数とし、コーパスから抽出された統計データを用いて計算する。

3.3 格フレームチェック

用言（動詞、形容詞）には、取りうる格が決まっていて、さらに同じ（深層）格を複数取ることはできないという制約がある。部分

木併合の際にもこれらの制約を考慮することが望ましい。今回の実験では、コーパス中に出現する各用言について格助詞「が、を、に」に関する表層格パターンを調査し、併合過程で各用言にそれらの格助詞に支配される表現が係りうるかどうかをチェックするようにした。また、同じ表層格を持つ表現が同じ用言に複数個係ることを禁止した（この際、係助詞の持つ格の多義性も考慮した）。その他、受け身、使役、さらに助動詞「たい」に支配される動詞の格パターンの変化にも考慮して実験を行った。

3.4 コーパスから抽出するデータ

上記の併合過程で利用するデータのうち、コーパスから抽出するものをまとめると次のようになる。ここで使用するコーパスはＡＴＲ音声言語データベース [11]（旅行会話：３７５会話、１３６４７文）に依存構造解析を行ったものである（この依存構造解析は、人手で検査修正を行った正解である）。

・依存パターンごとの依存確率

依存構造データベース中のすべての依存関係を調べ、依存元パターンと依存先パターンが同じ文中で依存元パターンが先行する形で現れた場合に依存関係が成立する確率（3.2 A式1）を求める。依存先パターンは原則的に部分木のノード（＝語）の見出し・品詞対である。依存元パターンは文節あるいは文節内構造である。例えば「私は行く」という表現に対し、依存先パターンとしては（「行く」・動詞）（「は」・係助詞）の２つが得られ、依存元パターンとしては（「は」・係助詞（「私」・代名詞））および（「私」・代名詞）の２つが得られる。なお、スパースデータ問題を回避するために自立語の見出しをワイルドカード「＊」で置き換えたものについても統計を取った。

・依存距離

２つの品詞が互いに依存関係にある場合の相互の距離のヒストグラムを求めた。多くの機能語（例えば助詞）の場合、直前の語にのみ依存を受けるが、動詞などは係り受けの形で遠距離の依存関係を持つ。上で述べたように、コーパスから得られた依存距離の情報は、依存尤度を計算するために用いられる。

・文末モデル

２つの品詞の間に文末が来る確率をコーパスから求めた。

・格パターン

コーパス中に現われる用言が表層格「が」「を」「に」を取り得るかどうかを調べ、併合の際の制約として利用した。（例えば動詞「会う」は格助詞「が」「を」「に」を取る）。

4 実験

4.1 実験データ

併合対象となる部分木の列は、前述のＡＴＲ音声言語データベースの依存構造解析を、文節に分解したものを用いた。このデータ中には助詞などの省略は現われるが、言い直し、および「あのー」「えーと」などの間投詞は取り除いてある。オープンデータとしては、統計用に用いなかった１２４対話 4,758 文を用い、クローズドデータとしては、統計を取るのに用いた３７５会話 13,647 文を用いた。話し言葉のポーズ単位は通常文節以上のまとまりであるので、文節に分解したものを用いることは差し支えがないと考えられる。また、入力は文ごとにファイルに分割して与えた。

4.2 評価方法

併合された部分木列を正解（文節ごとに分解される前の依存構造）の依存構造と比較して次の率を得る。

- 文正解率：併合結果が完全に正解と一致する割合
- 適合率：併合結果中の依存関係が正解の依存関係と一致する割合
- 再現率：正解中の依存関係が併合結果の依存関係と一致する割合
- 復元率：正解中にあり、併合前のデータに

ない依存関係が復元された割合。

4.3 実験結果

・クローズドデータによる結果

	文正解率	適合率	再現率	復元率
併合前	21.0%	100%	66.5%	0%
併合後	81.4%	93.6%	93.1%	79.4%
自立語見出し＊	64.0%	87.4%	84.7%	54.2%
関数A不使用	56.2%	80.9%	80.9%	43.0%

・オープンデータによる結果

	文正解率	適合率	再現率	復元率
併合前	16.2%	100%	65.0%	0%
併合後	69.8%	89.0%	88.0%	65.8%
自立語見出し＊	56.3%	84.4%	81.4%	46.9%
関数A不使用	49.2%	76.9%	76.9%	33.9%

ここで「併合前」というのは併合の対象となる文節に分解された依存構造である。その適合率が 100% ということは、間違った併合がないことを示す。併合前の文正解率（21.0%／16.2%）は、１文節からなる文の割合である。

「関数A不使用」という項目は、3.2 で述べた関数A（統語的構成による依存関係の統計的整合性に関する関数）の代わりに恒常関数 (=1) を用いた場合の結果を示している。また「自立語見出し＊」という項目は、関数Aで自立語の見出しの代わりにワイルドカード＊を使用した場合の結果で、意味的な情報を用いない場合の結果であると考えられる。

・関数B（3.2 参照）のパラメータ

近距離の依存関係候補を優先するように距離に対する減衰関数 $e^{-k\lambda d}$ を設定したが、係数kは約 0.02 の場合に最適であるという実験結果を得た。

4.4 問題点

・スパースデータ問題

代名詞（「わたし」など）はあまり旅行会話コーパス中に現われないので、代名詞が直接、間接的に動詞に係る用例が十分に採取できず、部分木併合失敗の原因となることがあった。

・接続詞などの扱い

本実験では接続詞などの文修飾詞は文末に係ると仮定したが、埋め込み文では節末に係る場合もある。

・格フレームチェックの問題

今回、同じ種類の格助詞に支配される複数の文節が同一の動詞に係ることを禁止したが、実際には同じ種類の格助詞に支配される複数の文節でも深層格が異なれば次の文「水曜日に学校にいく。」に見るように同一の動詞に係ることができる。この問題は、格助詞が支配する語の意味素性等から深層格を推定することによりある程度解決することができる。

また、現在の格フレームチェックは、次のような等位表現に対応していない：「AがBに、CがDにEする。」。ここで、AとCはEの主語、BとDは目的語になっているが、CとDがEに併合された後では、Eの格スロットが消費されてしまい、AとBはEに併合できなくなってしまう。

5 今後の課題

今後検討すべき課題について述べる。

・ 音声認識器との接続・文末判定

本研究の背景として頑健な音声言語処理を目的としていることを述べた。音声言語処理は音声入力に対して文法的、意味的な解析を行うのであるから、本研究の部分木併合は、音声認識、句構造解析の次段の過程としてリアルタイムで行われることが望ましい。

実際の音声入力には文の終わりを示す明確なマーカーが存在しないが、部分木併合を正しく効率的に行うためには文末を適切に判定する必要がある。本実験で用いた品詞バイグラムの他に韻律情報などを文末判定に役立てることが課題となる。

・　「非流暢性」現象への対処

本研究の目的の１つは音声言語の持つ「非流暢性」に対し頑健な解析手法を考案することであったが、今回の実験では「非流暢性」に対する分析を行っていない。言い直し、間投詞、倒置などの現象の個々について、本手法を用いた実験を行い、問題点を検討することが課題となる。

・　意味情報による解析

意味的な連関が高い依存候補対同士は依存尤度が高いはずである。現在の併合アルゴリズムは、意味的な連関の情報を主に自立語の字面の共起関係から得ている。しかし、この手法では、頻度の低い語ではスパースデータ問題が起き、また訓練データとは異なる語彙が用いられる場合には有効でなくなる。これらの問題を回避するため、意味素性やシソーラスを用い、その意味情報によって一般化を行うことが考えられる。

・　文脈からの意味情報の抽出

依存関係候補と同じ依存関係がそれまでの会話中に含まれている場合、その依存関係候補の尤度は高くなる [8]。すでに言及されている関係が再言及される可能性は高いからである。こうした情報を部分木併合的な処理に用いて解析の頑健性を増す研究としては [3] が挙げられる。同様の手法を採用することにより、部分木併合の成功率が上がることが期待される。

謝辞

本研究を進めるにあたり、実験上の支援をいただいた谷田泰郎氏に感謝いたします。

6　参考文献

[1] Jensen, K. and Heidorn, G. E.: "The Fitted Parse: 100% Parsing Capability in a Syntactic Grammar of English" ANLP83 (1983).

[2] Stallard, D. and Bobrow, R.: "The Semantic Linker - A New Fragment Combining Method" ARPA HLT-93 (1993).

[3] T. Nasukawa. "Robust Parsing Based on Discourse Information: Completing partial parses of ill-formed sentences on the basis of discourse information" ACL-95 (1995).

[4] 田代敏久，竹澤寿幸，森元逞: "音声言語処理のための部分木併合手法" 自然言語処理 109-4 (1995)

[5] 田代敏久，森元逞: "日本語会話文の言語解析実験" 信学技法 NLC95-24 (1995).

[6] Den, Y. "A Unified Approach to Parsing Spoken Natural Language" in the Proceedings of Natural Language Processing Pacific Rim Symposium '95, Vol. 2, pp. 574-579 (1995).

[7] 竹澤寿幸，田代敏久，森元逞: "自然発話の言語現象と音声認識用日本語文法" 情報処理学会音声言語情報処理研究会資料，6-5 (1995).

[8] 長尾確: "制約と選好による構造的多義性の解消" 情報処理学会研究報告 90-NL-78 (1990).

[9] 張玉潔，尾関和彦: "文節間係り受け距離の統計的性質とその係り受け解析への応用" 信学技法 NLC95-67 (1995).

[10] 伝康晴: "話し言葉における非文法的現象とその機械処理" 人工知能学会研究会資料，SIG-SLUD-9503 (1996).

[11] Morimoto, T. et al.: "A Speech and Language Database for Speech Translation Research" Proc. of ICSLP '94 pp. 1791-1794 (1994).