

決定木学習による述語の省略補完

小林明子 古宮嘉那子 乾伸雄 小谷善行

東京農工大学 工学部 情報コミュニケーション工学科

1. はじめに

日本語では，述語が省略されることがある．これは，別の箇所と言及されていたり，決まりきった表現があったりするために省略されることが多い．述語が省略されていると，自動翻訳などでは，正しい結果を得ることができない．人間は無意識のうちに省略を補完しながら会話したり読んだりする．しかし，計算機は与えられた文字列だけを手がかりにするため，省略があるとうまく処理ができない．そこで，本稿では，省略されている述語がどれなのかを，決定木学習を用いて推定する．

2. 日本語における述語の省略現象

述語の省略現象を次の種類に分類する．

文脈省略

補完されるべき語が，会話や文章の他の場所で言及されている[1]．本研究では，前文に補完語が明示されている場合だけを扱う．

例：「ハンカチは持った．ティッシュも（持った）．」

定型的な省略

決まりきった言い方があることで省略される．
例：「何かご迷惑でも（かけましたか）？」

その他

上記二つの種類に分類されないもの．補完語が不明の場合や，前文以外の場所にある場合．

例：「いや，待てよ．おれは催眠術のかけ方なんか……」（[4]より抜粋）

（この例の補完語の候補は「知らない」「分からない」「教わってない」など複数ある）

本研究では，上記の三つのパターンについて，述語の省略語の推定を行った．

3. 決定木

決定木生成のアルゴリズムは，ID3 アルゴリズムを使用した．また，生成する決定木は二分決定木である．適切な木を生成するためには，属性テストの識別力を判定し，その高い順に採用して木を構成する．ID3 アルゴリズムとは，その識別力の判定をエントロピーという概念を用いる方法である[3]．

作成する決定木は3種類である．補完語について，肯定／否定と疑問，テンスについて別々に決定木を生成し，出力する．それぞれの決定木の出力は，次のようになる．

(1) 補完語（9種類）

述語なし，前文の主節の述語，前文の従属節の述語，定型的な省略（ある，する，いう，なる，いる，かける，の6種類）

(2) 肯定／否定と疑問（4種類）

肯定，否定，肯定疑問，否定疑問

(3) テンス（2種類）

非過去，過去

補完語が前文以外の場所にある場合や不明の場合は，「述語なし」と判定されれば正解とする．

(1)の決定木で「述語なし」と判定されたものの以外について，(2)と(3)について決定木を生成し，出力する．(2)と(3)の決定木生成に使用する学習データには「述語なし」のデータは含まない．

4. 学習データ

4.1 データの準備

データは小説などから収集した．扱うデータは，文の最後に述語が省略されているものである．つまり，連用修飾で終わっている文である．本研究では，次に示すものを連用修飾として扱う．

副詞，動詞のテ形，動詞の連用形，形容詞のテ形，形容詞の連用形，形容動詞の連用形，助動詞のテ形，助動詞の連用形，格助詞，係助詞，副助詞，並立助詞

Predicate-Ellipsis Resolution Using Decision Tree Learning

Akiko Kobayashi, Kanako Komiya, Nobuo Inui, Yoshiyuki Kotani

Tokyo University of Agriculture and Technology

これらは形態素解析ツール ChaSen[2]の結果に基づいているが、例外として、明らかに不適切な形態素解析結果が出た場合は、正しいものに直した。

4.2 データの内容と傾向

学習データを2章の種類で分類すると、それぞれ次のような割合になった。

文脈省略：約 50%
定型的な省略：約 20%
その他の省略：約 30%

文脈省略については、話し手が1人のデータの方が2人のデータよりも多い。それは会話をする際に自分の発話した内容の補足をすぐ後に行うためだと考えられる。また、本稿では倒置の場合も文脈省略とみなしているため、全体的に話し手が1人の場合のデータが多くなる。話し手2人の場合のデータは、質問・答えの文や、相手の発話に対して質問をする文が多い傾向にある。例を次に示す。

質問・答えの文の例：

「いったい、いつから眠っていたのです」
「地球を出発してからだ（だ→眠っていた）」
（[5]より抜粋）

相手に対する質問の文：

「私、突き止めちゃった」
「何を（突き止めちゃったの）？」
（[6]より抜粋）

定型的な省略は、会話文での省略がほとんどである。また、収集したデータでは最後の形は、ほとんどが助詞であり、その助詞が定型的な省略の補完をする大きな手がかりとなる。

4.3 決定木に使用する属性テスト

決定木のノードとなる属性テストには、次のものを使用する。

述語省略文について

- ・文の最後の形（4.1節参照）
- ・会話か会話でないかの情報
（話し手1人、話し手2人、会話でない）

前文について

- ・前文があるかないか
- ・前文に存在する連用修飾について
（連用修飾の形があるか、それがかかるのは主節か従属節か）
- ・前文に従属節があるかないか

- ・主節、従属節の肯定／否定と疑問
- ・主節、従属節の過去／非過去

話し手1人か2人かについては、述語が省略されている文とその前文の話し手が同じ人か別の人がということを表している。

5. 実験

約1000件のデータで実験を行う。学習データは手作業による収集で偏りがあるため、実験方法としては交差検定を用いることにした。

6. 考察

本稿では、補完語が前文にある場合と、定型的な省略の場合以外は「述語なし」と判定している。「述語なし」と分類したものの中には、二つ前の文に補完語がある場合や、次の文に補完語が出現している場合があるので、前文だけではなく、文の前後に範囲を広げれば応用できると考えられる。その際には、決定木の出力も多くなるので、学習データはさらに多くなってはならない。また、学習データは手作業で集めたため、数が十分でない上に、出力に対してデータ数に偏りがある。データ数を多くすることにより正解率は上がっていくと考えられる。

定型的な省略については、文末の形が決まったものが多いため、文の最後の形に関する属性テストは結果に大きく影響を与えられる。本稿では収集したデータに基づいて「ある、する、いう、なる、いる、かける」の6種類を出力としたが、定型的に省略されるものは他にも数多くあるので、データが揃えばその他の述語に関しても実験できる。

7. まとめ

本稿では、述語省略文について、補完語とその肯定／否定・疑問、テンスの推定を、決定木学習を用いて行った。実験結果に関しては、発表時に報告する。

【参考文献】

- [1] 渡邊，榊井，河合，椎野：会話における省略とその補完，音声言語情報処理(2000)
- [2] ChaSen's Wiki（形態素解析ツール）
<http://chasen.naist.jp/hiki/ChaSen/>
- [3] 新田克己 著：知識と推論，サイエンス社(2002)
- [4] 星新一 著：ご依頼の件，新潮文庫(1988)
- [5] 星新一 著：なりそこない王子，新潮文庫(1986)
- [6] 赤川次郎 著：虹に向かって走れ，角川書店(1987)