決定木を用いた日本語ゼロ代名詞補完

4 J － 3　山本 和英　隅田 英一郎　古瀬 蔵†　飯田 仁
ATR 音声翻訳通信研究所
E-mail: yamamoto@itl.atr.co.jp

1. はじめに

日本語対話文におけるゼロ代名詞補完について述べる。主語や目的語などの表示が義務的でない日本語の処理においては、ゼロ代名詞と呼ばれるこれら省略要素(非明示要素)を補う処理が重要である。特に、日本語から英語 / ドイツ語などへの翻訳の際には、補完処理は必須となる。

多くの言語情報を利用したゼロ代名詞補完の研究としては、村田ら [村田 97] や江原ら [江原 96] の研究などがある。[村田 97] はヒューリスティックスによって各言語現象に得点を付与し、それらの合計によって最尤のゼロ代名詞を推定しているが、得点の調整には困難を伴うと予想される。また [江原 96] は経験的に 8 項目の特徴パラメータを設定して、確率モデルによって主語補完を行なっている。

本稿では、ゼロ代名詞の正解付きコーパスから言語現象と補完すべきゼロ代名詞の関係を帰納的に機械学習し、これによって補完することを提案する。コーパスから獲得した知識は決定木 (decision tree) によって表現する。これにより、多数の言語現象と補完するゼロ代名詞の関係が自動的に決定する。本稿では特に、ゼロ代名詞のうち主語に関する補完について述べる。

2. 日本語対話文の省略現象

例として、以下の発話で「忘れる」の主語を補完することを考える。

(例 1) 部屋にカメラを忘れてきてしまったようなんですが。

この例では、動詞「忘れる」の持つ意味属性や「てくる」「てしまう」「た」「ようだ」「んです」「が」といった文末表現など、非常に多くの要素が補完すべき主語に関係する可能性があり、このうちどの要素がどの程度主語補完に影響しているかを明確に記述することは難しい。

一方、本稿で提案する決定木を用いた手法では、影響する可能性のある要素を列挙するのみでよく、学習の結果主語補完に不要な属性は決定木の属性からは自動的に排除される。さらに、従来気づかなかった複数要素の同時出現による影響を自動的に学習する可能性もある。

一般に、対話文において省略された主語を補完するためには以下の情報が必要と考えられる。

1. 文内の情報
   動詞、平叙 / 疑問、能動 / 受動、尊敬 / 謙譲など。
2. 前文以前の情報 (文脈情報)
   対話におけるこれまでの話の流れ。
3. 言語外情報
   その文が、どこで、誰が誰に対して発話されたか、など。

ここで、以上のように分類された情報は独立ではなく、相互に影響しながら主語が省略されていることに注意しなければならない。例えば、ホテルのフロントにおける受付と客の対話で、一般的に「宿泊する」の動作主は客もしくは「一般的な人」であるが、ある特殊な文脈によってはそれ以外の可能性も考えられる。

3. 決定木を用いた補完処理

前述したように、日本語対話文の省略補完に必要な情報は多岐にわたる。これら情報を統一的に、かつ自動的で一意に主語を補完する手法として、本研究では決定木を用いる手法を提案する。

決定木は、多要素が複雑に関係した概念に対する知識表現手法の一つであり、有向木で表現される。各分岐節点はある属性に対応してその属性値によって枝分かれしていき、それぞれの葉で意志決定が行なわれる。決定木は分岐節点における分岐数によって大きく二分木と多進木とに分かれるが、本研究は前者を使用した。

3.1 使用属性

本稿では、計 1533 の属性を使用した。その内訳を表 1に示す。表で、文末表現とは終助詞、助動詞などの、動詞に後接する付属語群を指す。動詞の意味属性としては角川類語新辞典を使用した。言語外知識としては、発話された文の話者が情報提供者か情報授受者か、という属性のみを使用した。

表 1: 使用属性とその要素数

属性	要素数
動詞 (正規形)	292
動詞 (意味属性)	1000
文末表現	240
言語外知識	1 (話者情報)
合計	1533

Ellipsis Resolution in Dialogues via Decision-Tree Learning.
Kazuhide YAMAMOTO, Eiichiro SUMITA, Osamu FURUSE and Hitoshi IIDA.
ATR Interpreting Telecommunications Research Labs.
†現在、NTT コミュニケーション科学研究所

3.2 正解データと決定木学習

自動的に一意に決定木を作成するために、コーパスからの帰納的学習により決定木の作成を行なう。本研究で使用したコーパスは、チケット予約、観光案内などにおける二者の会話を収録したATR旅行会話コーパスである。このコーパスにおいて、主語が省略されている動詞に対して、一人称、二人称、三人称、特定されない人物[1]の4種類の補完すべき主語の正解情報を付与した。このとき、これらの正解は日本語のみを考慮して付与した。つまり、英語などへの翻訳時にどの主語になるかという観点では付与していない。

一般に決定木学習はNP完全であるので、本研究ではID3[Qui93]のアルゴリズムと同様、エントロピー規準による貪欲法(greedy algorithm)によって決定木学習を行なった。また、枝刈り(pruning)は行なっていない。

4. 評価実験

ATR旅行会話コーパスから226会話(総文数8961、主語省略総数3808)を使用して評価実験を行なった。

4.1 クローズドテスト

クローズドテストの補完正解率を表3右欄に示す。この実験で作成した決定木には、前述した1533属性のうち155種類の属性が各分岐節点で使用されていた。このうち比較的上位で使用された属性を表2に示す。

表2: 作成した決定木の上位で使用された属性

深さ	属性
1	動詞が「願う」である
2	動詞後に補助動詞「くださる」を含む
3	動詞後に終助詞「か」を含む
4	動詞後に助動詞「てくださる」を含む
6	動詞後に助動詞「てもらえる」を含む

実験の結果、最深の葉は深さ50であった。また、葉に到達するまでに必要な平均属性数(人称別)は、一人称16.1、二人称10.5、三人称16.7、「一般」23.1であった。この結果は、二人称が最も特徴をとらえやすく、反対に「一般」は本実験で用意した属性だけでは特徴をとらえにくいことを示している。

4.2 オープンテスト

クローズドテストと同一の会話を対象にして、オープンテストを行なった。有効な学習量を確保するため、実験は5分割の交差確認法(cross validation)により行なった。このとき、分割は会話単位で行なった。

実験の結果を表3に示す。人称の補完に関しては、全体の平均で90.7% となった。また、表に示すように出現数の多いものほど正解率が高いという結果になった。

[1]一般的な「人」を念頭において発話していると考えられる場合。以下「一般」と呼ぶ。

表3: 各人称における補完正解率

人称	出現数	オープン	クローズド
一人称	2272	95.60 %	99.34 %
二人称	1186	90.39 %	98.65 %
三人称	89	51.69 %	92.13 %
一般	261	62.45 %	87.74 %
全体	3808	90.68 %	98.16 %

実験では10分割による交差確認法も行なった。その結果、一人称、二人称の結果は表3とほぼ同じとなり[2]、三人称や「一般」はそれぞれ56.18%、66.28%となった。三人称や「一般」で結果が良くなったのは学習サンプル数増加による改善のためと考えられる。

正解と出力との関係を表4に示す。表4に示すように、主な補完誤りは「一般」とその他の人称との間で起こっている。文脈を考慮しない場合、「一般」として出現する文は他の人称主語の文と解釈することも可能であるため、本稿の属性のみでは「一般」の特徴抽出が十分に行なえなかったと考えられる。

表4: 正解と出力の相関(オープンテスト)

正解 \ 出力	一人称	二人称	三人称	一般
一人称	2172	39	10	51
二人称	67	1072	10	37
三人称	13	9	46	21
一般	64	24	10	163

5. まとめ

日本語のゼロ代名詞補完に対して決定木による機械学習を行なう手法を提案し、実験によってその有効性を確認した。本稿で提案した手法は入力として品詞付き形態素列のみを使用しており、韓国語などのゼロ代名詞補完にも有用であると推定される。

参考文献

[Qui93] QUINLAN, J. R.: *C4.5 Programs for Machine Learning*, Morgan Kaufmann (1993).

[江原 96] 江原暉将, 金淵培：確率モデルによるゼロ主語の補完, 自然言語処理, Vol. 3, No. 4, pp. 67–86 (1996).

[村田 97] 村田真樹, 長尾眞：用例や表層表現を用いた日本語文章中の指示詞・代名詞・ゼロ代名詞の指示対象の推定, 自然言語処理, Vol. 4, No. 1, pp. 87–109 (1997).

[2]一人称補完が95.51%、二人称補完が90.13%。