6Q-2

概念知識を用いた連体助詞「の」の決定木による自動意味分類

森山 健太，古宮 嘉那子，但馬 康宏，小谷 善行
東京農工大学情報コミュニケーション工学科

１．はじめに

自然言語処理の研究で、解析が困難である現象として、連体助詞の「の」という現象がある。連体助詞の「の」とは、「A の B」いうパターンの名詞句の事である(以下「A の B」と呼ぶ)。「A の B」の意味は多様であり、例として「人間の足」という場合は「人間の一部である足」という意味であり、「木の家」という場合は「木で作られた家」という意味となる。

本研究では、機械学習により「A の B」の意味を推定することを目的とする。「A の B」は、日本語の文中に非常に高い確率で出現するため、意味解析の上で非常に重要な問題であり、現在様々な手法が提案されている。

語彙項目を名詞毎に与え、AとBの語彙項目の組み合わせにより意味を解析する[1]や、国語辞典を用いた解析を行う[2]のような研究がある。

本研究は、事前にルールを定めて解析するのではなく、EDR 概念辞書[3]から得た前後の単語の概念的知識と「A の B」の意味の関係を決定木で学習させることにより、ルールを自動生成することを目的とする。

２．「A の B」の意味定義

本実験で分類する「A の B」の意味を 19 種類定義した。定義の一覧を表１に示す。

表１　「AのB」定義一覧

意味名	「A の B」における意味	例
所有	A が持っている B	あなたのカバン
所属	A に所属する B	自民党の人
所在	A にいる（ある）B	神奈川の彼
行為の場所	場所 A における行為 B	インドの生活
時	時 A における B	春のイベント
作者	A が作った B	私の論文
行為者	A がした B	彼の釣り
関係・資格	～と A の関係にある B	左のもの
同格	A である B	政治家の太郎
性質・状態	性質 A である B	軟体の動物
材料	A でできた B	鉄の棒
数量・順序・割合	A 個（番目）の B	３個の石
対象	A に対する行為 B	害虫の駆除
所有属性	A がもつ性質 B	私の年齢
A の関係	A と B という関係にあるもの	それの左
分野・内容	内容、分野が A である B	格闘の本
部分	A が含む部分 B	刀の刃
A の状態	A が状態 B である	あなたの危険
分類	A に分類される B	類人猿の動物

*Classification of "A no B" using Decision Tree Learning of Conceptual Knowledge
Kenta Moriyama, Kanako Komiya, Yasuhiro Tajima, Yoshiyuki Kotani
Department of Information and Communication Sciences, Tokyo University of Agriculture and Technology

これらは国語辞典や事例データをもとに定義した。

３．「の」の分類をする決定木学習

3.1 決定木学習に用いる要素

Web から集めた文章ファイルを使って学習を行った。その中から、係り受け解析器 Cabocha[4]を使用し係り受け解析を行い「A の B」を抽出したところ、1465 件の「A の B」を含む文が存在した。そして、以下の要素が「A の B」の分類に役立つと仮定し、決定木の入力として使用した。学習のアルゴリズムは C4.5[5]で、作成する木は Yes-No2 分木とした。

・A と B の表層語
文中に記述されている単語の文字列をそのまま使用する。

・A と B の読み
Cabocha による解析に含まれる値を使用した。「もの」と「モノ」など、表層語が異なる語も読みが等しい場合に同一に捉えることができる。

・A と B の品詞
Cabocha による解析に含まれる値を使用した。「名詞-固有名詞-地域-一般」「名詞-副詞可能」のように、名詞を分類することができる。

・A と B の概念知識
EDR 概念辞書を使用して求める単語の概念である。概念知識に関する説明は 3.3 に記述した。

・「A の B の意味」
「A の B」に対し人手で付加した意味である。19 種類の意味定義の中から１つを選んだものを用いる。

3.2 形態素解析による単語細分化

形態素解析を行うと、人間の認識する単語より細かく分割してしまうという問題が生じる。例として「ページの閲覧数」という場合、形態素解析した場合、B が閲覧と数に分かれてしまう。先頭の「閲覧」だけを参照すると「ページの閲覧」となり、意味を「対象」と捉えてしまうが、末端の「数」をみると正しい意味である「所有属性」を推測できる。

このように名詞が連続した場合、重要な表現が末端にある場合が多いが、末端だけでは意味が特定できない場合も存在する。よって、形態素解析の結果名詞が複数連続した場合は、先頭と末端の両方を学習に用いることにする。

3.3 学習に用いる概念知識

EDR 概念辞書には、概念が図 1 のように木構造で記述されており、各概念にはその概念に最も意味が近い単語が記述されている。下位概念になるほど意味が細かく分類されていて、最も上位の概念である「全ての概念」の下位概念を１層概念としたとき、決定木には１～n 層概念を学習させる。n は可変とし、実験で変化させて最適値を調べる。単語が概念辞書に載っていなかった場合、概念を null とした。学習データの中で、概念が辞書に載っていたものは全体の 80 パーセントであった。

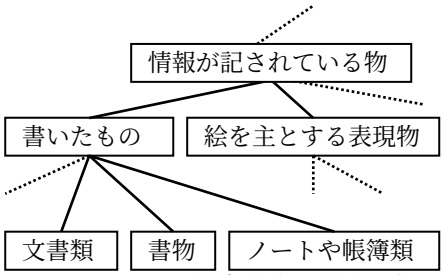

図１　EDR 概念辞書の構造の例

しかし、単語１つに対し複数の概念が存在する場合がある。例として「下」の概念を調べると、

・(地位が)下であること
・順序が後であること
・心の内側に隠れた
・ある人の支配が及ぶ範囲

の 4 通りが存在する。そこで、１つの単語に対し複数の概念を持たせ、決定木の分岐で「A は～という概念を持っているか？」と分岐することにする。

3.4 学習結果

5 分割クロスバリデーション方式で評価を行った。ベースラインは最も多い意味である所有属性が占める割合 20.2%とする。

学習させる概念の深さを 1 から 10 まで変化させて実験を行った。その結果を表 2、図 2 に示した。

表２　概念の深さによる精度の変化

概念深さ	1	2	3	4	5	6
精度	53.8	56.3	58.9	59.5	60.1	61.0
概念深さ	7	8	9	10	ベースライン	
概念深さ	61.2	61.5	61.5	61.5	20.2	

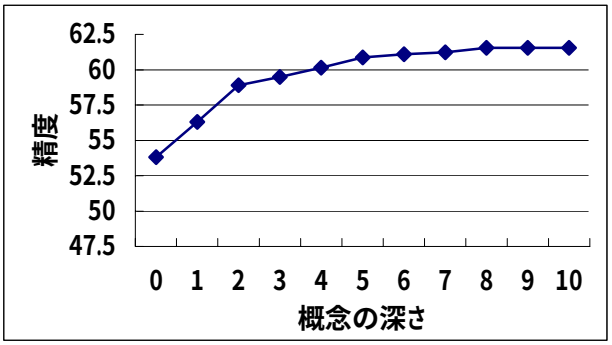

図２　概念の深さによる精度の変化

また、今回出力された決定木の 1 部を図 3 に示した。

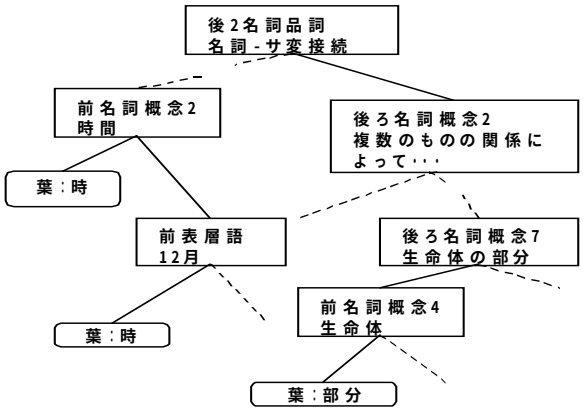

図３　出力された決定木の一部（点線部は省略）

4. 考察

概念の深さが深いほど精度が高いという結果になった。しかし、8 層以上は精度が変わっていないことから、9 層以上は分類に必要ないということが分かる。精度は最大で 61.5%だが、ベースラインと比べると 40%程上回った。

出力された決定木を見てみると、最初に品詞や浅い概念で大まかに分割し、その後に深い概念で具体的に分割する傾向があることが分かる。図 3 の左側を見ると、12 月という単語は概念辞書に載っていないため、概念「時間」とは別に表層語で分割している。このように、表層語は概念辞書に載っていない単語を学習するために用いられていることが分かる。また、図 3 の右側のように「後ろの概念=生命体の部分」かつ「前の概念＝生命体の部分」の場合、意味が「部分」となるなど、適切な概念の組み合わせが見られる箇所も多く存在した。

5．おわりに

本研究では、連体助詞「の」の決定木による解析により、精度 61.5%で正解の意味を導くことができ、ベースラインを 40%ほど上回った。

また、概念が辞書に載っていた確率が 80%であったので、概念辞書に変わる名詞の情報を獲得する手法も考える必要がある。

参考文献

[1]植村将人：生成語彙論に基づく名詞句「の」の意味解釈 北陸最先端科学技術大学院大学　修士論文, 2005
[2]黒橋禎夫,酒井康行：国語辞典を用いた名詞句「A の B」の意味解析 情報処理学会研究会 自然言語処理 129-16, pp.109-116, 1999
[3]日本電子化辞書研究所: EDR 電子化辞書使用説明書, 1995
, http://WWW.iijnet.or.jp/edr/TG.html
[4]奈良先端科学技術大学院大学自然言語処理学講座：日本語係り受け解析器Cabocha
http://chasen.org/~taku/software/cabocha/
[5]J.R.Quinlan, C4.5:Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993