英語構文解析における句動詞関連の曖昧性抑制機構

岩渕 亮　　宮崎 正弘

新潟大学大学院 自然科学研究科
{ryo,miyazaki}@nlp.ie.niigata-u.ac.jp

1　はじめに

英語構文解析では解析結果に種々の構造的曖昧さが生じ，解析結果が一意に定まらないという問題がある．本稿では，そのような構造的曖昧さの中でも特に基本動詞が前置詞や副詞と結びついて，一つの動詞に相当する意味を表す句動詞の解析誤りによって生じる曖昧さに注目し，それを抑制する手法を提案する．

英語の種々の句動詞を収集し，概念情報を組み込んだ句動詞データベースを作成する．そして英語文パーザの解析結果とデータベースを照合することにより英語文中に出現する句動詞を認定し，曖昧さを抑制する．句動詞データベースを利用しなかった場合と利用した場合の解析結果を比較し，その有効性を示す．

2　句動詞による構造的曖昧さ

英語文中に動詞と前置詞が並んで出現した場合，その動詞と前置詞が句動詞を構成するものなのか，それともそれぞれ動詞句と前置詞句が独立した構造をするのかの判断で曖昧さが発生することがある．

実際に句動詞を含む英語文を構文解析したときに，どのようにして構造的曖昧さが発生するのかを説明する．次に示すのは “the work calls for special skill” という例文を解析させた結果の解析木である．

図 1 の木構造では本来動詞 “call” と前置詞 “for” が句動詞として解析されなくてはならない部

```
|-s
  |-cl
    |-sbj
    | |-np
    |   |-det--< the >
    |   |-n--< work >
    |-vp
    | |-vpit0
    |   |-vit0--< calls >
    |-advp
      |-pp
        |-prep--< for >
        |-np
          |-adj--< special >
          |-n--< skill >
```

図 1: 誤った解析木

```
|-s
  |-cl
    |-sbj
      | |-np
      |   |-det--< the >
      |   |-n--< work >
      |-vp
        |-vpsa2
          |-vsa2--< calls >
          |-prep--< for >
          |-dobj
            |-np
              |-adj--< special >
              |-n--< skill >
```

図 2: 正しい解析木

分がそれぞれ，“vp” と “advp” といった独立したノードに分解されたかたちで解析がされている．

一方，図 2 の解析木は動詞 “call” と前置詞 “for” が 1 つの “vp” という木構造を構成しており，句動詞としての解析がされている．

以上のようにして，解析する英語文中に句動詞が含まれる場合，それが原因で解析結果に構造的曖昧さが生じる．このようにして発生する構造的曖昧さを抑制する方法について以降で詳しく論じる．

3 概念識別子を利用した抑制

上記のような句動詞が原因で発生する構造的曖昧さを抑制するために，句動詞の目的語の概念に着目した．ここでの概念とは，文や単語の意味内容が計算機内で表現されたものである [1]．概念は EDR 電子化辞書全体を通じて一意的に定められている．概念を特定するための方法が概念識別子 (概念 ID concept identifier) であり，16 進整数でによって表される．

句動詞によっては前置詞，副詞に続く目的語の概念が限定されるものが多い．例えば日本語で “～を実行する” の意味に相当する “carry out ～” を考えてみる．ここで句動詞の目的語が取り得る概念としては “もの”，“事柄”，“行為” などが考えられるが “場所”，“領域”，“方向”，“時間” などは考えにくい．このような句動詞の性質を構造的曖昧さの抑制に有効利用する．

3.1 概念体系について

各単語の概念を得るためには EDR 電子化辞書の概念体系辞書 [2] を利用する．概念間の関係のうち，特に概念の上位-下位関係を用いて概念全体を体系化したものが概念体系辞書である．

概念間の上位-下位関係として，例えば図 3 のような体系の一部を例にとる．この図は “情報が記されているもの (4445e8)” という概念の下位分類である．この概念は，“書いた物 (4445bc)”，“絵を主体とする表現物 (4445c4)”... といういくつかの下位概念に分割され，詳細化されている．さらに，“書いた物 (4445bc)” という概念は，“書簡 (4445a0)”，“文書類 (4445c2)”... という下位概念に詳細化される．

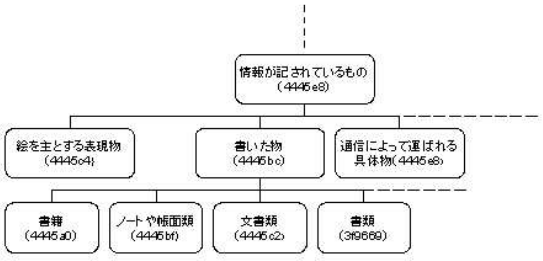

図 3: 概念体系の上位-下位関係

概念体系を上位方向へ登っていくと，最終的にはルートノードである “概念 (3aa966)” という概念に辿り着く．このルートノード “概念 (3aa966)” の 1 つ下のノード (これを “深さレベル 1” のノードとする) の概念は次の 5 つに分類される．1-人間または人間と似た振る舞いをする主体 (3aa911)，2-ものごと (3d017c)，3-事象 (30f7e4)，4-位置 (30f751)，5-時 (30f776)．句動詞データベースで利用する概念識別子はこの “深さレベル 1” のものを利用する．句動詞の目的語の概念識別子から “深さレベル 1” の概念識別子を得るために次の手順を踏む．

まず EDR 電子化辞書の英語単語辞書 [3] から目的語の概念識別子を得る．次に EDR 電子化辞書の概念体系辞書から，この概念識別子を下位概念とする上位概念識別子を検索する．得られた上位概念識別子から再び，この概念識別子を下位概念とする上位概念識別子を検索する．得られた上位概念識別子が，深さレベル 1 の 5 つの概念識別子のどれかと等しくなるまで再帰的に繰り返す．

3.2 概念識別子を利用した句動詞データベースの作成

概念識別子を組み込んだ句動詞データベースは以下のリスト構造に沿って作成する．

((vsa2:VERB),(prep:PREP),(ci:CI))

リストの第 1 要素中の “vsa2” は動詞品詞コードの 1 つで，特定の前置詞と組となることのできる動詞に割り振られている．“VERB” は句動詞を構成する基本動詞の原形が字面で入る．同様にしてリストの第 2 要素中の “prep” は前置詞を表わす品詞コードであり，“PREP” は句動詞を構成する前置詞の字面が入る．

リストの第 3 要素中の “ci” は概念識別子を表わすコードである．“CI” には “深さレベル 1” の概念が入るが，16 進数 6 桁の形式ではなく，記述者の負担を軽減する目的で下の表 1 に示す対応表に従った略式を用いる．また “CI” には複数の概念を並列に記述できる．

概念	概念識別子	略式
人間または...	3aa911	sub
ものごと	3d017c	obj
事象	30f7e4	phe
位置	30f751	loc
時	30f776	time

表 1: 概念識別子と略式の対応表

3.3 概念識別子を利用した抑制の実際

以下に示す手順に従って，概念識別子を組み込んだデータベースを活用しながら句動詞による構造的曖昧さを抑制する．

手順 1 構文解析の結果として生成される木構造のリストをファイルへ出力する．

手順 2 ファイルの中から 1 つのリストを抽出し “手順 3” へ．全てのリストのチェックが終了している場合は “手順 9” へ．

手順 3 このリストを特定のキィで検索し，句動詞としての解析がされているかどうかを調べる．リストが句動詞として解析されている場合は “手順 4” へ．そうでない場合は，句動詞が原因で発生する構造的曖昧さはないものとして，抑制処理を中断し，次のリストの処理のため “手順 2” へ．

手順 4 “手順 2” で選定されたリストの中から句動詞を構成する動詞と前置詞の組を抽出し，句動詞データベースとの照合を行う．句動詞データベースにその動詞と前置詞の組が存在する場合は “手順 4” へ．そうでない場合はその動詞と前置詞は句動詞ではないものと見なして処理を中断し，次のリストの処理のため “手順 2” へ．

手順 5 再びリストを検索し，さきほどの動詞と前置詞の組の目的語を選定する．

手順 6 “手順 5” で選定した目的語の概念識別子を EDR 電子化辞書の英語単語辞書から調べる．

手順 7 “手順 6” で得た概念識別子を基にして，上位概念識別子方向へ向って再帰に EDR 電子化辞書の概念体系辞書を検索し，深さレベル 1 の概念識別子を得る．

手順 8 “手順 7” で得た目的語の深さレベル 1 の概念識別子と，句動詞データベースの概念識別子の比較を行う．もし 2 つの概念識別子が一致した場合，その動詞と前置詞の組は句動詞であると判断し，このリストを保存し “手順 2” へ．そうでない場合はこのリストを破棄し “手順 2” へ．

手順 9 この段階で保存されているリストを木構造表示ルーチンへ渡して抑制ルーチンは終了．

4 評価実験

提案手法を評価する実験を行なった．解析には高校 1，2 年生向けの英語参考書 [4] などから 400 文の例文をを抜粋し，試験文として用いた. 試験文はランダムに選定されており，句動詞を含まない英文もこの中には存在する．試験文に含まれる平均単語数は 7.5 語である．

解析には並列型チャート法を計算機上に実装した英語構文解析システム [5] を用いた．並列型チャート法では入力文中の全ての単語を辞書引きすることから解析を始める．これにより，入力文中の先頭から 1 語ずつ順に読み込みながら解析する逐次型チャート法よりも高い並列性をひき出すことができる [6]. 並列型チャート法は入力文中の各単語に対して，CFG 文法を参考にエッジ (弧) と呼ばれるデータを生成し，これを成長させることで解析を進める．また，データベースには 350 個

の句動詞が実装されている．

400 文の試験文をこの英語文解析システムで解析し，解析成功率と発生した多義数の平均を明らかにした．図 4 は試験文中に含まれる単語数と発生した構造的曖昧さの相関関係をグラフで表わしたものである．1 文に含まれる単語数が増加するに従い，発生する構造的曖昧さも増加する傾向にあることがわかる．

次に句動詞の字面情報のみでデータベースに収録されている句動詞とマッチングを行なうことによって構造的曖昧さを抑制した．前節で説明した概念識別子を用いた抑制はここでは行なわない．単語の字面のみで抑制を行なった場合，文脈から判断して，本来句動詞ではない動詞と前置詞の組を誤って句動詞として判別してしまう可能性が大きい．結果として，正しい解析木を排除してしまうことになり，解析成功率の低下に繋がることになる．

最後に概念識別子データベースを用いた抑制を行なった．抑制を行なわない場合や字面情報のみで抑制を行なった場合とで解析成功率，構造的曖昧さの発生率の比較をした．特に，字面情報のみによる抑制の際に発生した副作用がどの程度改善されたかに注目する．その結果を表としてまとめたものが表 2 である．

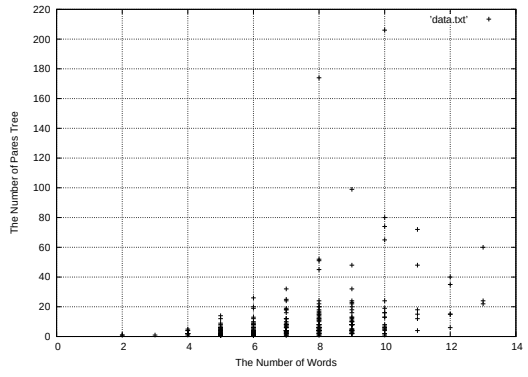

図 4: 単語数と多義数の相関関係図

抑制方法	解析成功率	平均多義数	副作用
なし	64.3%	11.6	
字面情報	59.0%	9.78	21
概念識別子	63.8%	9.61	6

表 2: 解析結果の比較

5 考察

概念識別子を用いて構造的曖昧さの抑制を行った場合，抑制を行わなかった場合と比較すると，平均多義数を約 2 減少させることができた．字面情報による抑制の場合，平均多義数の抑制だけを見ると概念識別子による抑制とほぼ同等の結果が得られたが，副作用，つまり本来句動詞ではない動詞，前置詞の組み合せを誤って句動詞として扱ってしまうという解析が 400 文中 21 文で発生してしまっている．また，このことが原因となり，抑制を行なわなかった場合と比較して解析成功率が 5%程度低下した．一方，概念識別子を用いて抑制を行った場合，副作用の発生は 6 文に抑えることができ，解析成功率も抑制を行わなかった場合とほぼ変わらないことがわかる．結果として，概念識別子を用いた場合，字面情報のみの場合と比較して，より正確な構造的曖昧さの抑制を実現することが可能となった．

参考文献

[1] 株式会社日本電子化辞書研究所; EDR 電子化辞書仕様説明書 (1993)

[2] 株式会社日本電子化辞書研究所; EDR 電子化辞書 概念体系辞書 (1993)

[3] 株式会社日本電子化辞書研究所; EDR 電子化辞書 英語単語辞書 (1993)

[4] 長谷川潔，千種基弘; 高校英文法，啓林館 (2002)

[5] 川辺，宮崎; 構造を含む生成規則を扱える拡張型チャートパーザ, 言語処理学会第 11 回年次大会発表論文 (2005)

[6] 田中穂積; 自然言語処理の基礎，産業図書 (1989)