1L－5

# 複合語用例データベースを用いた複合名詞の構造的曖昧さの絞り込み法

太田悟　宮崎正弘

新潟大学大学院自然科学研究科

## 1　はじめに

日本語処理において、名詞、および名詞相当の接辞が結合して構成される複合名詞は、難しい処理対象の一つである。従来、複合名詞を自動分割し構造を解析する方法[1][2]に関しては様々な提案がなされてきた。また構造的曖昧さを絞り込む方法[3]についてもいくつか提案されているが、複合名詞が長くなるにつれ十分な正解率が得られないなどの点で、まだ多くの課題を残している。

本稿では、主に「名詞+名詞」の構成を持つ複合名詞の用例をデータベース化し、日本語複合名詞の構造的曖昧性の解消に利用する方法を提案し、その有効性を示す。解析対象となる複合名詞は拡張CYK法[2]により構造解析され、複合語の用例データベースに類似する木に対しては、類似度に応じた重みを付与することなどにより、曖昧さの絞り込みを行う。

## 2　複合語用例データベース

複合名詞の構造的曖昧さを絞り込むため、まず複合語の用例を集めたデータベースを用意する。これらの複合語の用例は主に新聞記事、EDR共起辞書、解析済みの複合名詞データから獲得したものである。ここでは次に示す形態素の組合せを持つ用例をデータベース化する。

1. 名詞　＋　名詞
2. 名詞　＋　接尾辞
3. 接頭辞　＋　名詞

複合名詞を解析する際に最も問題となっているのは、二文字漢字名詞が連続する場合の各名詞間の係り受け関係である。そこで本稿でも二文字漢字名詞に着目し、名詞に関しては現在のところ二文字漢字名詞のみを対象としデータベース化した。

Disambiguation of Japanese Compound Noun Structure Using Compound Noun Example Database

Satoru Ohta , Masahiro Miyazaki

Niigata University

## 3　類似度の判定

類似度の判定は、複合語用例データベースから類似用例を検索する段階と、検索された用例を基に属性推定を行なう段階に分け、これらの結果を基に類似度を決定する。

類似用例の検索は図1に示した順序で行う。各検索条件に合った用例が獲得されると、その用例を基に属性推定が行なわれる。ここでは、不確定部分の品詞とカテゴリの組合せを数え上げランク付けを行なう。最後に類似度の算出を行なうが、不確定部分の品詞・カテゴリが推定したものと合えば、次に示す点を与えその和を類似度とする。

字面　→　0.1
品詞　→　0.2
カテゴリ　→　0.7

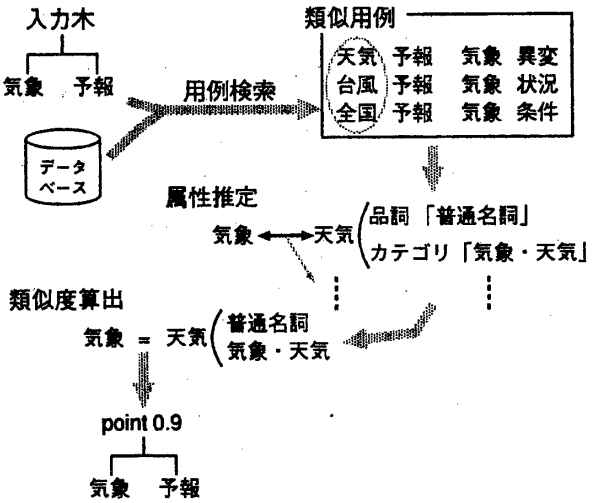


図1. 類似度の判定の流れ

この属性推定の特徴は、形態素同士の類似度を判定するだけでなく、長い複合名詞内に存在するより小さな複合語である部分複合名詞[1]を含んだ類似度

を判定できる点にある。これにより長い複合名詞に対しても、複合名詞同士の用例を用意することなく、効果的な解析が行なえる。

## 4 構造的曖昧さの絞り込み

形態素解析の結果、複合名詞内には分割や同形語の曖昧さが多数存在し、その構造的曖昧さは膨大なものになる [2][3]。そこで従来の評価法に、類似度による評価を組み込んだ絞り込み法について検討した。

### 4.1 類似度による評価 ($C_{similar}$)

類似度による評価は、構造解析により得られた各木に対し与えられた類似度の和で表わされる。

$C_{similar} = \sum_i (S_i + \alpha_i)$

$S_i$　各木の類似度

$\alpha_i$　頻度による加点

この評価に対しても重み $W_{si}$を与え、従来の評価式 [3] に加えたものが下式となる。

$$C_{total} = W_{co} \cdot C_{connect} + W_{mo} \cdot C_{morph} + W_{st} \cdot C_{struct} + W_{si} \cdot C_{similar}$$

$C_{connect}$　… 形態素の接続による評価

$C_{morph}$　… 形態素の評価

$C_{struct}$　… 構造による評価

$W_{co}$、$W_{mo}$、$W_{st}$　… 各評価への重み

### 4.2 解析例

上の評価式を基に、「国際鳥類保護会議」の例で実際に解析を行う。図 2 に示す分割の例では 5 つの構造的曖昧さが存在する。まず類似用例の検索により「国際|会議」「鳥獣|保護」が検出され、類似度 0.9 がそれぞれの木に与えられる。他の木に対しても同様の処理が行われるが、類似用例は見つからず加点はされない。他の評価から動作性名詞にそれぞれ 0.25、左枝分かれ構造に 0.1 が与えられ、評価値 $C_{total}$ =2.4 となり、図 2 の構造に絞り込める。なお各重みは $W_{co}$ =0.8、$W_{mo}$ =0.5、$W_{st}$ =0.1、$W_{si}$ =1.0 とした。

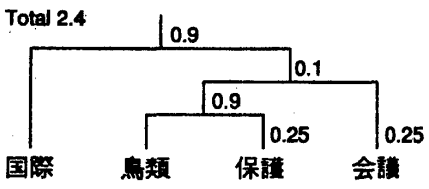


図 2. 解析例

従来の構造解析は長い複合名詞に非常に弱く、長さ 8 文字を越える複合名詞となると、現状では 50% の正解率が得られれば良いほうであった。本稿の解析を行なうことにより、約 80%の正解率が得られると思われる。

## 5 学習

構造解析、曖昧さの絞り込みを終えた後、解析結果の正誤に関わらず教師あり学習を行なう。ここでは人手により正解構造を与え、これを基に複合語用例データベースへの新規登録や頻度情報の更新をする。この処理を行なうことにより、誤った解析結果が出力されても、次の解析から正しい結果を得ることが出来るようになる (図 3)。

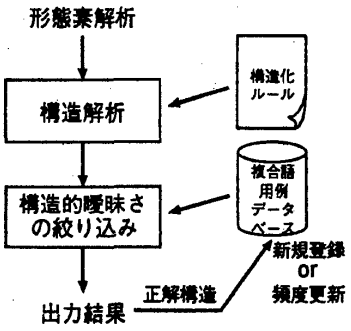


図 3. 正解構造の学習

## 6 おわりに

日本語複合名詞を対象とした構造解析において、複合語用例データベースを用意し、その用例の類似度に応じた加点をすることにより、構造的曖昧さを絞り込む方法について提案し、その有効性について示した。今後は、より多くのデータを解析し、適切な重みの獲得やデータベースを充実することが必要となる。

## 謝辞

「NTT 名詞意味属性体系データ」を提供して頂いた NTT コミュニケーション科学研究所、「EDR 日本語共起辞書」を提供して頂いた日本電子化辞書研究所の関係各位に深謝いたします。

## 参考文献

[1] 宮崎、池原、横尾：複合語の構造化に基づく対訳辞書の単語結合型辞書引き、情報処理学会論文誌、Vol.34、No.4、pp.743-754(1993)

[2] 佐野、宮崎：拡張 CYK 法による日本語複合名詞の構造解析法、信学会秋季大会、No.D-51(1992)

[3] 前川、宮崎：日本語複合名詞の構造的曖昧さの絞り込み法とその評価、情報処理学会第 49 回全国大会、No.1G-5(1994)