E-007

名詞と動詞の共起に着目した状況認識

Situational awareness focused on the co-occurrence of Japanese nouns and verbs

野呂 翔一† 天沼 博† 松澤 和光†
Shoichi Noro Hiroshi Amanuma Kazumitsu Matsuzawa

１．はじめに

コンピュータが会話における状況や文脈を認識するには、会話に出てきた個々の単語の意味はもちろん、それらから常識的・日常的に想起される様々な事項を含めた推定が必要となる。こうした仕組みの実現に向けて、ありふれた名詞に対して典型的に共起し易い動詞のパターンを収集することを考えた。

そのため、まず電子化辞書[1]の格パターンと概念体系について、親密度の高い名詞を中心としたデータに整理し直した。次にインターネット上の様々なコーパスを利用し、頻度の高い典型的な単語の抽出を試みた。また、抽出した結果を評価するため、寄席で行われる「三題噺」を参考にした文章の自動生成を検討した。

２．典型的な格パターンの抽出

2.1 データ抽出の狙い

自然言語処理技術が意味処理へと発展するにつれて、言語の意味情報を支える様々なデータが整備されて来ている[2]。中でも格パターンと呼ばれるデータは、動詞と共起する名詞をその格関係と意味分類で整理したもので、文の格構造解析や機械翻訳などに利用されている[3]。例えば「 (c1)が (c2)で 泳ぐ 」という格パターンでは、意味カテゴリ c1：動物、c2：位置など、共起される名詞が規定される。しかし、通常は動詞「泳ぐ」と言えば「魚が池で」等が典型例であり、各カテゴリに含まれる名詞全てが当てはまるとは限らない。

本研究では、この格パターンを利用して会話における状況を認識することを将来の目標とする。つまり、人間は会話の一部を断片的に聞いただけでも、それがどんな状況でなされた発言かを大体予想することが出来る。例えば「海、夏、水着…」とあれば、誰しも「海水浴」の文脈を想起するだろう。ただ、こうした用途では格パターンのデータには上述の典型性が必要になると考える。このような典型的な格パターンを構築しようとする研究[4]も行われてはいるが、状況認識に関わる典型性は実は固定的に決まっているものではなく、分野や場面、時代、文化などの影響で変動すると考えられる。そこで本研究では、ある程度の普遍性を期待できるカテゴリ中心の格パターンをベースとして、これに様々なコーパスを利用して典型性を与える方策をとる。

2.2 ベースとする格パターンの整理

EDR 電子化辞書[1]は、日本語の概念(意味)について様々な情報が記述された複数の辞書によって構成された辞書データである。主に、概念の上下関係を記述している概念辞書と、ある動詞と共に使われることが多い単語の情報を記述した共起辞書によって構成される。本研究ではこのうち以下の 3 つを使用する。

- **概念体系辞書：**
  データ数は約 42 万行。概念の上下関係を概念番号というコードにより、上位と下位の概念を 1 対 1 で記述。

- **概念見出し辞書：**
  データ数は約 41 万語。概念番号の名称や意味などを日本語／英語で記述。

- **日本語動詞共起パターン副辞書：**
  データ数は約 1 万 4000 語。日本語のある動詞と共に使われることが多い概念群を記述。

ただし、このままではデータ量が非常に膨大であり、動詞を基準にした多種多様なデータが付与された特殊な記述形式である等、本研究の目的には不都合なため、以下の整理を行った。

まず、親密度[5]を用いてあまり一般的でない単語を削除した。親密度とは、ある単語に対して一般の人がどの程度なじみがあると感じられるかを表した指標データである。この結果、概念見出し辞書は約 41 万語のデータを約 2 万語にまで、共起パターン副辞書は約 1 万 4000 語のデ

†神奈川大学大学院工学研究科電気電子情報工学専攻

ータを約 6300 語にまで削減した。

次に、記述形式については日本語文章の状況認識に使う事を考慮して、以下のように簡略化した。

・概念見出し辞書：

……

貯金　6.250　0fbec6　貯金
貯水　5.125　0fe07d　貯えた水
兆候　5.062　3cee2d　物事の兆候
……

概念名「貯金」という単語に対しては、親密度の数値が「6.250」であり、概念番号が「0fbec6」であり、意味が「貯金」という事を表している。

・共起パターン副辞書：

……

JCP0000582 泳ぐ【およぐ】"手足やひれを動かして,水面を進む" agent が;30f6b0;30f6bf;
JCP0000582 泳ぐ【およぐ】"手足やひれを動かして,水面を進む" place で;30f751;
……

1 行目は「泳ぐ」に対して、agent 格(その事象・動作を行う主体)としては概念番号「30f6b0」と「30f6bf」の下位概念の名詞が出現し、助詞は「が」を取るといった意味となる。2 行目も同様に、place 格(その事象・動作の成立する場所)としては概念番号「30f751」の下位概念の名詞が出現し、助詞は「で」を取るという事を表している。

2.3 コーパスを利用した典型単語の抽出

整理した概念見出し辞書に記載されている名詞について、インターネット上のブログなどの一般の人が作成したコーパスから共起頻度の高い動詞のパターンの抽出を行った。

さらに、会話の文脈や状況を認識するための情報として、名詞・動詞の共起パターンの前後に出現する動詞の出現パターンも、上記と合わせて収集を行った。この結果、出現頻度の高い動詞について、前後に出現する動詞のデータが得られた。

３．抽出した格パターンの評価

本研究で抽出した格パターンの妥当性を評価するため、「三題噺」風の文書の自動生成を検討した。「三題噺」とは、寄席で行われる余興の一つであり、観客から提供された 3 つのお題（単語）を使用して即興で落語噺を作成するものである。これを真似て、入力された 3 つの単語に対し主語や動詞などを自動で補って、短いが意味の通った文章を生成する。例えば「デパート」、「にこにこ」、「ローラースケート」という単語を入力した場合には、格パターンを利用して次の様な文章を自動で生成する。

「少女は<u>にこにこ</u>と笑っている。
なぜなら、少女は<u>デパート</u>へ父と行く。
そして、父に<u>ローラースケート</u>を買ってもらう。」

この自動生成は入力された単語の 1 つから格パターンを用いて動詞を選定し、文章を生成する。その後、生成した文章の名詞・動詞の共起パターンの前後に出現する動詞を用いて、残りの 2 単語に対して文章を自動生成する。そして、自動生成した文章を人間の感覚と照らし合わせ、どのくらい人間の感覚に近いのかを評価する予定である。

４．おわりに

会話の状況や文脈をコンピュータに認識させることを狙いとして、ネットコーパスを利用した典型的な格パターンの抽出を試みた。また、抽出結果を評価するため、「三題噺」風の文章の自動生成を検討した。今後は自動生成した文章を人間の感覚と比較するとともに、文章中の個々の単語から常識的・日常的に想起される様々な事項を含めた推定を行う仕組みを検討する。

<u>参考資料</u>

[1] ＥＤＲとＥＤＲ電子化辞書
http://www2.nict.go.jp/r/r312/EDR/J_index.html
[2] 日本の言語資源・ツールのカタログ
http://nlp.kuee.kyoto-u.ac.jp/NLP_Portal/lr-cat-j.html
[3] 日本語語彙大系
http://www.kecl.ntt.co.jp/icl/lirg/resources/GoiTaikei/
[4] 言語処理のための日本語の動詞辞書
http://cl.it.okayama-u.ac.jp/rsc/data/index.html
[5] 日本語の語彙特性,三省堂,1999