共起関係と格の重要性を考慮した音声対話文の格解析

6Ｈ－6

堀田剛志　本田岳夫　奥村学
北陸先端科学技術大学院大学情報科学研究科

1　はじめに

音声対話では発話において，助詞は省略されることがある．さらに助詞や助動詞はといった付属語は，単語長が短く，音響的に曖昧になりやすいため，音声認識で誤認識や脱落が起きやすい．そのため，助詞は自立語と比べて認識精度が低い．これらの理由で，キーワードスポッティングなどの手法により，発話中の自立語のみで言語解析を行う研究が見られる．

本稿では，助詞・助動詞の情報を除いた自立語列を入力とし，コーパスからの共起関係を用いて，深層格を推定する方法を述べる．本手法は，コーパスからの共起関係として，< 名詞，深層格，動詞 >の３つ組の共起関係を利用する．なお，３つ組の共起関係がコーパスに出現しない場合は，< 名詞，深層格 >，< 深層格，動詞 >の２つ組から３つ組の共起関係を推定することによって，コーパスのスパースネスの問題を解消する．

2　共起スコア

コーパスから３つ組< 名詞 n，関係子 r，動詞 v >の共起頻度を元に共起スコアを計算する様々な手法が提案されている．Hindle は，相互情報量 $I(x,y)$ を基に共起スコア $SC_r(n,v)$ を計算し，このスコアを用いて名詞のクラスタリングを行った [1]．

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{1}$$

$$SC_r(n,v) = \log_2 \frac{\frac{f_r(n,v)}{N}}{\frac{f(n)}{N}\frac{f(v)}{N}} \tag{2}$$

ここで，Nはコーパス内の単語数，$f_r(n,v)$ はコーパス中の $< n,r,v >$ の頻度，$f(n), f(v)$ はコーパス中の n,vのそれぞれの頻度である．

しかし，３つ組$< n,r,v >$は必ずコーパス内に出現するわけではない．Resnik は，このようなコーパスのスパースネスの問題を解消するために，シソーラスを用いて，n の属する上位概念 Cの中の全ての名詞と動詞の共起頻度$\sum_{n_i \in C} f_r(n_i,v)$ から$< n,r,v >$の共起頻度を元に推定した [2]．これにより，式 (2) は次式のようになる．

$$SC_r(C,v) = \log_2 \frac{\frac{\sum_{n_i \in C} f_r(n_i,v)}{N}}{\frac{\sum_{n_i \in C} f(n_i)}{N}\frac{f(v)}{N}} \tag{3}$$

シソーラスを用いて，名詞の上位概念を順にたどっていき，かなり上位の概念 C_{upper}で，共起スコアが得られた場合，$\sum_{n_i \in C_{upper}} f_r(n_i,v)$ は概念 C_{upper}が様々な名詞を取るため，$< r,v >$の頻度 $f_r(v)$ に近似でき，式 (3) は次式のようになる．

$$SC_r(C_{upper},v) = \log_2 \frac{\frac{f_r(v)}{N}}{\frac{\sum_{n_i \in C_{upper}} f(n_i)}{N}\frac{f(v)}{N}} \tag{4}$$

これは，C_{upper}と vの相互情報量を計算していると解釈できるので，名詞の情報が反映されていない．

そこで本研究では，シソーラスとして EDR 概念辞書 [4] を用いて，図１のように，n の上位概念をたどる上限を設け，上限の概念で共起スコアが得られない場合には，共起スコアが式 (4) で近似されると考え，２つ組$< n,r >$ と $< r,v >$ の頻度 $f_r(n), f_r(v)$ を元に，３つ組$< n,r,v >$の共起スコアを推定する．$f_r(n), f_r(v)$ は，EDR 共起辞書 [5] より得ることができる．

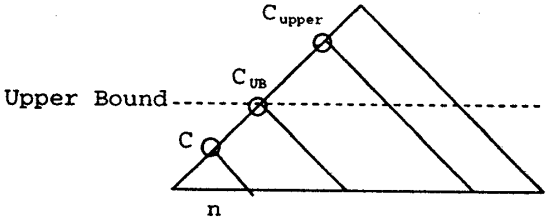

図 1: シソーラス中の上位概念をたどる上限

コーパスでの n の出現中で n と rが共起する確率 $\frac{f_r(n)}{f(n)}$ を重みとして式 (4) に導入する．

$$Score_r(n,v) = \frac{f_r(n)}{f(n)} \log_2 \frac{\frac{f_r(v)}{N}}{\frac{\sum_{n_i \in C_{UB}} f(n_i)}{N}\frac{f(v)}{N}} \tag{5}$$

ここで，C_{UB}は上限の概念を表している．また，共起スコアが上限より下位の概念で得られる場合には，式 (5)

Semantic analysis using co-occurrence information and caseframe preference in spoken dialogue
Koji Hotta,Takeo Honda,Manabu Okumura
Japan Advanced Institute of Science and Technology
15 Asahidai, Tatsunokuchi, Ishikawa 923-12, Japan

の $f_r(v)$ は近似する前の値 $f_r(n,v)$ であるため, 共起スコアは次式で表すことができる.

$$Score_r(n,v) = \frac{\frac{f_r(n)}{N}}{\frac{f(n)}{N}} \log_2 \frac{\frac{\sum_{n_i \in C} f_r(n_i,v)}{N}}{\frac{\sum_{n_i \in C} f(n_i)}{N} \frac{f(v)}{N}} \quad (6)$$

以上のことから, 式 (5), (6) により共起スコアを計算する. これにより, 3 つ組$< n,r,v >$の共起スコアを, 2 つ組$< n,r >$,$< r,v >$から近似できる. このようにして, コーパスのスパースネスの問題を解消する.

3 解析の手順

前節で述べた共起スコアを用いて格解析を行うために, 本稿では, 係り受け候補を抽出し, 深層格を推定するという手順で行う.

3.1 係り受け候補の抽出

本研究では,ATR 対話データベース [6] の「旅行に関する旅行会社と客の対話」の内, 意味関係コードが記述してある 5 対話から, 助詞・助動詞の品詞が付いている単語や, 助詞相当語として登録されている単語列を取り除いた品詞付きの自立語列を入力とした.

音声対話を対象とした場合, 処理単位として「文」を認定することが難しく, 音声認識から得られるポーズを処理の区切りとして発話を理解するといった研究が増えてきている.[3] 本研究でもポーズを処理の区切りとし, 係り受け候補を抽出する. ポーズを処理の区切りとした場合, ポーズ間の発話内に動詞は平均 1.5 個程度しか出現せず, その大多数の動詞に係る名詞は, 近くにある動詞に係るため, 名詞は近くにある動詞に係ることにする.

3.2 深層格の推定

関係子 r を深層格とし, それぞれの深層格 r に対して, 共起スコアを次式を用いて計算し深層格の推定を行う.

- C_{UB} より下位で,$f_r(n,v)$ が得られた場合

$$Score_r(n,v) = \frac{\frac{f_r(n)}{N}}{\frac{f(n)}{N}} \log_2 \frac{\frac{\sum_{n_i \in C} f_r(n_i,v)}{N}}{\frac{\sum_{n_i \in C} f(n_i)}{N} \frac{f(v)}{N}}$$

- C_{UB} で,$f_r(n,v)$ が得られなかった場合

$$Score_r(n,v) = \frac{\frac{f_r(n)}{N}}{\frac{f(n)}{N}} \log_2 \frac{\frac{f_r(v)}{N}}{\frac{\sum_{n_i \in C_{UB}} f(n_i)}{N} \frac{f(v)}{N}}$$

ここで,C_{UB} はたどれる上限の上位概念を表している.

一般に,1 つの動詞に係る名詞は複数であるため, 文としてそれぞれの名詞が動詞に対してどのような深層格になるかを推定する必要がある. 本研究では, 動詞の語義が同じであり, 深層格は重複しないという条件を満たしているもので, それぞれの共起スコアの和が高いものを選択する事によって,1 つの動詞に複数の名詞が係る場合に対処する.

4 おわりに

本稿では, 助詞・助動詞の情報を除いた自立語列を入力とし, コーパスからの共起関係を用いて, 深層格を推定する方法を述べた. コーパスから 3 つ組$< n,r,v >$の共起スコアが得られない場合に, 2 つ組$< n,r >$,$< r,v >$から共起スコアを計算することにより, コーパスのスパースネスの問題を解消した.

今後の課題としては次のようなことが挙げられる. 抽出した係り受け候補の中には, 名詞−名詞関係のものも含まれているため, 名詞の係り先が名詞か動詞かを決定する方法を考える必要がある. また, ポーズを越えて係り受け関係があるものの解析や, 言い直し・言い淀みなど冗長語を含んだ対話文から, これらの冗長語を取り除き, 格解析が行えるようにする必要がある.

参考文献

[1] Donald Hindle. Noun Classification from Predicate Argument Structure. ACL90-6,pp.268-275,1990.

[2] Philip Resnik. WordNet and Distributional Analysis: A Class-based Approach to Lexical Discovery. AAAI Workshop,pp.48-56,1992.

[3] 上條俊一, 秋葉友良, 伊藤克亘, 田中穂積. 休止を処理の区切りとした自由発話理解. 情報処理学会 第 50 回全国大会 3-91,1995.

[4] 日本電子化辞書研究所. EDR 概念辞書. 1994.

[5] 日本電子化辞書研究所. EDR 日本語共起辞書. 1994.

[6] ATR 自動翻訳電話研究所. ATR 対話データベース. 1990.