# 対話データベースへの意味情報の付与

7L－10

東 郁雄　荒木 雅弘　堂下 修司
京都大学大学院工学研究科情報工学専攻

## 1　はじめに

近年、自然言語処理においては、人手による知識を用いる手法の限界から、大量のコーパスや用例を用いる手法の研究が多く見られるようになった[1][2][3]。これまで、書き言葉のコーパスに関しては大規模なものが作られているが、話し言葉のコーパスについてはこれから構築していく必要がある。

コーパスを有効に利用するためには、様々な種類のタグが付与されていることが望ましい。コーパスに付与する情報としては、形態素情報などが一般的であるが、意味的な情報の付与によって、よりコーパスを有効に利用することができると考えられる。

本研究では、Semantic tag の付与の前提として、形態素情報が付与されていることを想定している。しかし、話し言葉の形態素解析には様々な問題があるので、最初にそれらの問題について考察する。次に、形態素情報が付与されているコーパスへの Semantic tag の付与について検討する。

## 2　話し言葉のコーパスへの形態素情報の付与

コーパスに形態素情報を付与するためには、形態素解析ツールを用いるが、解析の精度が低い場合は、後に修正を行うコストが大きくなる。そのため、形態素解析の精度をあげる必要がある。

話し言葉では、間投詞、言い淀み、言い直し、相槌などが頻繁に起こり、形態素解析の際に大きな障害となる。今回は、テキストの書き起こしを行う際にそれらを括弧でくくっておき、解析時には取り除いて入力する方法をとる。

形態素解析ツールとしては、フリーのツールである JUMAN が広く使われている。これを用いて、当研究室の対話コーパスの解析を行った。用いた JUMAN は version 3.0 Beta である。コーパスはスケジューリングタスクで、人間対人間の対話が 25 対話、人間対機械の対話が 10 対話あり、合わせて 1175 発話である。形態素が正しく区切られており、品詞が正しければ、読みが間違っていても正解とした。解析を誤った形態素は 1151 個あった。解析を誤った部分について、その原因を分類すると以下のようになる。

i) 格助詞「の」「で」が判定子「だ」の活用したものと解析される (55 %)
ii) 固有名詞が辞書に登録されていない (30 %)
iii) 普段使わないような単語が辞書に登録されているので、解析結果に現れる (4.4 %)
iv) カタカナを全てサ変名詞とするために誤る (2.8 %)
v) 話し言葉に特有の言い回し (2.0 %)
vi) 未知語と判定される (0.3 %)
vii) それ以外 (4.7 %)

i) については、以前のバージョンではコメントアウトされており、そのような対応が正しいと思われる。

ii)、vi) については、JUMAN の辞書に単語を登録することによって解析精度の向上が見込まれる。v) についても同様で、よくあらわれる言い回し等を辞書に登録することで解析精度の向上が見込まれる。

vii) については、コーパスのタスクに現れる単語の偏りの知識を用いることができれば解析が成功すると考えられるものが含まれている。それらについては、タスクに応じた単語の優先度を与えることで解析精度の向上が見込まれる。

## 3　EDR 電子化辞書を用いた単語の意味的分類の検証

コーパスから知識を抽出する場合、n-gram 等の統計的な手法を用いることが多いが、単語の種類が多い場合は空間が広くなり過ぎ、有意な統計情報を得るためには非常に大量のコーパスが必要となる。単語の意味的分類を考え、その統計をとる方法により、空間を小さくすることができる。

単語の意味的分類は、意味的に近い単語の集団を見つけることによって行うので、シソーラス等を用いる方法が有望である。今回は、EDR 電子化辞書

Semantic Tagging for Dialogue Data
Ikuo AZUMA, Masahiro ARAKI,
Shuji DOSHITA
Faculty of Engineering, Kyoto Univ.

の概念体系辞書を用い、特に名詞について単語の意味的分類が可能かどうかの検証を行う。方法は、以下の通りである。

- 形態素解析されたコーパスから名詞に属するものを抜き出す。
- EDR 電子化辞書の日本語単語辞書を用い、各単語の取り得る概念を全て求め、その中から正しい概念を選ぶ。
- 概念体系辞書を用いて各概念の概念体系上の位置を調べる。特に今回は、最上位概念から 3 レベル下の層に注目し、各概念がどの概念の下位に分布しているかを調べる。

実際に分布を調べる際に、次のような問題が生じた。

- 概念体系辞書が多重継承を許しているため、3 レベルより下の概念が複数の 3 レベルの概念の下位に存在している場合が見られた。
- 上位 3 レベル以内の概念についても多重継承が見られ、ある 3 レベルの概念の上位に別の 3 レベルの概念が存在する場合も見られた。この場合は、注目している概念が、より下位に存在する方の 3 レベルの概念の下に存在しているととらえた。
- 注目している概念が、上位 3 レベル以内に入っている場合があった。

人間対人間、人間対機械の対話各 3 対話ずつから名詞を全て抜き出し、それぞれの概念の分布を概念体系の 3 レベル目の概念を基準に調べた。その結果の一部を、表 1 に示す。この結果から、出現名詞の概念の分布に偏りが見られ、このまま直接意味分類として用いることは難しいにしろ、意味分類を求める際に役に立つものと思われる。

## 4 むすび

本稿では、対話コーパスに対する形態素解析の問題点について指摘し、形態素情報が付与されたコーパスに意味的分類を行うための EDR 電子化辞書の概念体系辞書の有効性について調べた。今回のような分布の調べ方で意味的分類が直接得られるわけではないが、役に立ちそうであることが分った。

今後の課題としては、まず単語の概念をできるだけ効率よく求めることと、意味的分類を求め、コーパスに付与する手法を確立することがある。また、得られた意味タグ付きのコーパスをもちいた、

表 1: コーパスに見られた名詞の分布 (一部)

| 概念体系の 3 レベル目 | 分布数 |
|---|---|
| 年や月や日でとらえた時 | 149 |
| 抽象物 | 127 |
| 一日の流れの中の時 | 84 |
| 会議 | 52 |
| 組織 | 37 |
| 職業、肩書、役割で限定した人間 | 28 |
| 具体的あるいは抽象的存在物 | 27 |
| 性状・性向 | 26 |
| 関係 | 21 |
| 催し | 15 |
| 基準点の前後の時 | 14 |
| 機能で捉えた場所 | 12 |
| 具体物 | 11 |
| 地域 | 11 |
| ある個体を基準とした関係で捉えた人間 | 11 |
| 自動物 | 11 |

| 概念体系の 3 レベル目 1 | 概念体系の 3 レベル目 2 | 分布数 |
|---|---|---|
| 具体物 | 部屋 | 30 |
| 具体物 | 抽象物 | 25 |
| 性状・性向 | 職業、肩書、役割で限定した人間 | 25 |
| 性状・性向 | 抽象物 | 15 |

## 参考文献

[1] 永田 昌明 , 鈴木 雅実: 日英対話コーパスへの発話行為タイプ付与の試みとその統計的対話モデルへの利用, 人工知能学会研究会資料. SIG - SLUD - 9302 - 7 ( 1993-9 ).

[2] 鈴木 雅実 , 永田 昌明: 日英対話コーパスへの談話レベルの情報付与と翻訳への利用, 信学技報. TECHNICAL REPORT OF IEICE. NLC93-38 ( 1993-7 ).

[3] 樽松 明: 対話コーパスの構築と応用 スケジューリングタスクの自由発話音声の言語的性質, 文部省重点領域研究「音声対話」(第 3 年次) 研究成果報告書 (1996-3).