自然言語コーパスにおける概念のマーキングルールの設定*

4B－3

土井 晃一, 大森 晃
株式会社 富士通研究所 情報社会科学研究所, 東京理科大学 工学部 経営工学科
doy@iias.flab.fujitsu.co.jp, ohmori@ms.kagu.sut.ac.jp

1 はじめに

自然な発話における発想量を定量化したい。本研究では発想量を概念数として計量する。我々はここでいう概念数を、発話の中に現れる(発話行為論[1]でいうところの)命題の数と考える。しかし、発話行為論の命題は考え方を示しているだけで、そのままでは実際に命題を数えることはできない。しかも発話行為論は、形態素・構文に関して、英語に対して考えられているので、そのままでは日本語に適用できない。そこで国語文法の品詞の概念を細かく用いることを考える。

本稿では自然な発話を書き起こしたコーパスから概念を抽出するためのマーキングルール(品詞のマーキングルールと概念抽出ルール)の設定方法について考察し、実際のその適用について述べる。

本ルールは要求の構造化の前段階となる概念抽出にも使えるので、最後にそれについても述べる。

命題は指示、あるいは述定、あるいは指示と述定の組からなる。またエコ[2]によると、指示は表象であり、指示行為の結果によって生じる言語表現である。指示行為とは指示対象に言及することである。指示対象とは言語によって表現しうる実体、抽象的概念、関係、性質などである。また述定とは指示対象の属性(動作、存在、性質、状態)、あるいは指示対象間の関係に言及した言語表現である。

我々はこれらの概念を用いてコーパスから概念を抽出する。

2 コーパスにおける品詞のマーキング

2.1 基本方針

マーキングの基本的方針は以下の通りである。

1. 品詞の切り分けを行う。名詞(代名詞・数詞を含む)・動詞・形容詞・形容動詞・副詞・連体詞をマーキングする。
2. 字面で品詞が曖昧な場合には、構文レベルから検討する。

1に関して補足する。「その本」の「その」のような連体詞は、指示行為を特定しているだけだが、品詞のマーキングのときには別々にマーキングして、次の概念抽出の時、まとめて指示とする(ルール1)。

次に2に関して補足する。はっきりしないものは分析のしようがないので、マーキングしない。品詞として同定可能なものだけを対象にする(ルール2)。どっちにもとれる場合、例えば、感動詞とも、連体詞ともとれる場合は、確定しているものだけを扱う(ルール3)。この場合、発話者の意図を優先せず、現象を優先することにする。

品詞のマーキングでは品詞ごとと全品詞に対して各々連番をふり、後々の参照に利用する(ルール4)。

例えば「この～、あの～、その～、米が...」という発話があったとする。この場合「この～、あの～」は感動詞として捨てる(ルール3と方針1より)。「その～」の部分は連体詞であるから次の概念抽出の時まとめて指示とすることにする。(ルール1より)。

2.2 句の扱い

本節では、名詞句のマーキングについての単位の問題について述べる。名詞句をどこまで分解してマーキングするかである。この問題は、格助詞を含んだ句と複合語との違いに帰着される。例えば、「米の政策」と「米政策」の場合である。この場合、両者は概念が異なると考える。格助詞は述語に対してどのような関係にあるかを示す語であるから、格助詞を伴う句は、指示と述定を構成するものと考え、「米」と「政策」をマーキングする(ルール5)。一方、複合語は熟語化していて分離する必要がないため「米政策」としてマーキングする(ルール6)。

同様にして用言句の扱いとして、用言の複合語は一つの語とみることにする(ルール8)。例えば、「飼い殺しみたいになって」という句は、動詞句であり動詞としてマーキングすることにする。

2.3 補助詞の扱い

本節では補助詞(補助形容詞・形式名詞など)の扱いを述べる。補助形容詞は2.2節の複合語と同じ扱いにし、まとめてマーキングする(ルール8)。例えば、「よくない」はまとめてマーキングする。その他の補助詞についても同様に考える。ただし、「悪くはない」のように取立詞を伴う句は別々にマーキングする(ルール9)。

2.4 言い直しの扱い

本節では言い直しの扱いについて述べる。基本的に、発話者の意図を優先せず、現象を優先することにする。つまり、言い直し(前言の否定)の現象を尊重して、言い直しの前を捨てて、言い直し後だけをマーキングすることにする(ルール10)。ただし、部分的な言い直し、例えば、「適正に、な、価格で...」という発話があった場合、「適正に、な」までを形容動詞としてマーキングして、次の概念抽出の際に「適正な」と直すことにする(ルール11)。また、前言がはっきりしない(客観的にみて何を言っているか分からない：指示対象が不明な)場合の言い直しは、分析のしようがないので、捨てることにする(ルール12)。

3 マーキングの実例

本節ではマーキングの実例を挙げる。図1は実際の発話コーパスにマーキングを施した例である。図中"A"は発言者を表す識別子、"= = ="は発話が不明瞭で聞きとれなかったところを表す。括弧内がマーキングした箇所を表す。括弧内の最初の数字が全品詞の通し番号を表す(ルール4より)。次のアルファベットが品詞(Mが名詞、Dが動詞、Kが形容詞、KDが形容動詞)を表す。最後の数字が品詞ごとの通し番号を表す(ルール4より)。002は補助形容詞「ない」を含むのでまとめてマーキングしてある(ルール8より)。本ルールを延べ240分の発話コーパスに適用した結果、不都合はなかった。

```
A: えー、まあ、今まで、= = = とこで、えー、まあ、
   (001M001:米) が (002K001:足りなく) ても
   (003M002:生産度) が (004K002:悪く) て、まあ
   (005M003:何か)、(006D001:輸入し) たけれども、
   (007M004:何か)(008KD001:ちぐはぐな)
   (009M005:輸入) の (010M006:仕方) を
   (011D002:し) ているんじゃないかと、
```

図1: コーパス上のマーキングの例

*Establishing Marking Rules for Concept in Natural Language Corpus
Kouichi DOI, Akira OHMORI (Fujitsu Laboratories, Institute for Social Information Science, Science University of Tokyo)

項番	指示	述定	概念
1	(001M001:米)	(002K001:足りない)	(001M001:米) が (002K001:足りない)
2	(003M002:生産度)	(004K002:悪い)	(003M002:生産度) が (004K002 悪い)
3		(005M003:何か) を (006D001:輸入する)	() が (005M003:何か) を (006D001:輸入する)
4	(007M004:何か)	(008KD001:ちぐはぐだ)	(007M004:何か) が (008KD001:ちぐはぐだ)
5	(010M006:仕方)	(009M005:輸入)	(010M007:仕方) が (009M005:輸入) に関するものである
6		(010M006:仕方) を (011D002:する)	() が (010M006:仕方) を (011D002:する)

表 1: 概念の抽出の例

4 概念抽出ルール

本節では、概念抽出ルールについて述べる。概念抽出ルールの基本方針は以下のとおりである。本来、概念は発語内行為と命題の対として考えるべきである。しかし、現在のところ、日本語における発語内行為の定義には曖昧なところが多い。そこで、我々は、発語内行為を切り離して、命題を概念と考える。命題は指示、あるいは述定、あるいは指示と述定の組からなる。そこで、指示と述定のみの場合も抽出することにする。概念抽出に際しては、前節までのルールでマーキングした範囲内のみを分析対象とする。つまり、マーキングしてない部分は分析対象からはずすことにする。例えば、言い直した場合は言い直された後の方の発話のみを対象とする。

概念抽出ルールの詳細は以下のようになる。

1. 重文は区切る。
2. 副詞節などを伴う複文は、複数概念とする。
3. 取り立て助詞を伴って指示と述定が同時に現れる場合、例えば、「象は鼻が長い」の場合は「象の鼻が長い」と読み変えて、一つの概念とする [3]。この場合、原文は変えない。
4. 格助詞を二つ以上含む句は、展開して概念とする。例えば、「米の政策」は『「政策」が「米」に関するものである』とし、「政策」を指示、「米」を述定とする。ただし、用言に直接接続する格助詞は展開せず、述定の一部とする。
5. 単純に指示と述定が同時に現れる場合は、「指示＋述定」を一つの概念とする。
6. 指示のみが現れる場合は、「指示」を一つの概念とする。
7. 同様にして、述定のみが現れる場合は、「述定」を一つの概念とする。
8. 「指示＋指示＋指示」＋「述定」や「指示」＋「述定＋述定＋述定」の場合は、分離した概念として、別々に考える。

係受けの不明瞭な副詞 (句) の扱いについて述べておく。副詞 (句) は用言が複数ある時、どの用言を修飾するかが曖昧になることがある。また文修飾なのかどうかも曖昧になる。これを命題に埋め込むか、あるいは態度として扱うかという問題が生じる。これは係受けを考えることにより、解決する。例えば、「いっそのこと，全員帰化して，全員外国人チームっていうのがひとつできたらおもしろいんじゃないかと思います」という例では、「帰化する」にかかる。どちらか判断に困る時、つまりどちらとも判断できる時は、発語内行為 (態度) として命題と対にすることにする。

本方針で概念抽出を行ない不都合が無いかを調べることにする。

5 概念抽出の実例

本節では前節の概念抽出ルールの実適用について述べる。3節の例から概念を抽出した例を表 1に示す。項番 3 の例では、前節の概念抽出ルール 4 の規則を適用して、「何か」をまとめて述定の一部にしてある。また、項番 5 の例でも、前節のルール 4 を適用して、格助詞の「の」を展開してある。本抽出ルールを延べ 30 分の発話コーパスに適用した結果、不都合はなかった。

6 構造化に向けて

本節では、前節までのルールをソフトウエア要求の構造化に使えないかどうかを検討する。前節までのルールを使うと、構造化の前段階に当たる、発話における概念がほぼ完全に抽出できる。

ソフトウエアに限らず、要求はほぼ次のように書ける [4]。

(名詞) は (形容詞) なので (よい/好き)
(名詞) は (形容詞) なので (悪い/困る/嫌い)

ここで、(形容詞) の部分は厳密に言えば (用言) となる。すると、(名詞) は指示になり、(用言) は述定になる。つまり、要求の表現形式の前半部分「(名詞) は (用言)」という部分は前述の概念に当たる。

つまり、発話コーパスから次の手順で概念を抽出できる。

1. 名詞 (代名詞・数詞を含む)・形容詞・形容動詞・副詞・連体詞をマーキングする。
2. 指示と述定 (つまり命題) と発話内行為を抽出する。

命題と発語内行為を合わせて「原要求」とする。

例えば、前述の例を取ると表 2のようになる。表中"「」" 内に入っている部分はコーパスから直接引用した語句を、"『』" 内に入っている部分は分析者の判断した結果をそれぞれ示す。

項番	原要求
1	「米」が「足りない」のは『悪い』
2	「生産度」が「悪い」のは『悪い』
3	「何かを輸入した」のは『よい』
4	「仕方」が「輸入」に関するものであるのは『よくも悪くもない』
5	『政府』が「輸入の仕方をしている」のは『悪い』

表 2: 原要求の例

ただし、実際の適用には、

1. 発語内行為 (態度) の確定
2. 指示の同定
3. 話されていない指示、あるいは述定の補足
4. 同じものを参照している指示の見極め

が必要になる。これらの問題に解決策が見つかれば、要求の構造化の前段階となる概念の抽出にも応用できよう。

7 おわりに

自然な発話における発想量を定量化するため、発話の中に現れる命題の数を数えることにした。命題の数を数えるために、命題抽出をまず行なう。我々は命題抽出のために品詞のマーキングルールと概念抽出ルールを実際に適用しながら、確定した。

また本ルールは要求の構造化の前段階となる概念の抽出にも使えるので、その考え方を述べた。

謝辞

大森研究室所属の山口幸一君には実際にルールの適用をしてもらったことに深謝する。

参考文献

[1] John R.Searle, 坂本 百大・土屋俊訳. 発話行為 -言語哲学への試論-. 勁草書房, 1986.

[2] ウンベルト・エコ著；谷口勇訳. テクストの概念. 而立書房, 1994.

[3] 三上章. 象は鼻が長い. くろしお出版, 1994.

[4] 土井晃一. 要求獲得オフライン法での非機能/未分化要求の抽出. ソフトウエア工学研究会, 1 月 1996.