# 新聞記事における複合名詞の構造に着目した分析

7Q－4

内野　一　横尾　昭男　池原　悟

ＮＴＴ情報通信網研究所

## 1．はじめに

機械翻訳システムの開発においては、各種の文を翻訳できるようにすることが理想的であるが、現状ではすべての分野について正しい訳を生成することは困難な状況にある。したがって、カスタマイズ等により特定の分野に依存した表現を翻訳できるようにすることになるが、その分野における文の特徴を正しく把握し、それに対応した処理を強化することが必要となる。本稿では、日英機械翻訳の課題の一つである複合名詞の処理の立場から、５分野の新聞記事、１万文に出現した複合名詞に対してその意味的構造に着目した分析を行い、文章の特性と複合名詞の関係について考察する。

## 2．分析概要

### 2．1　複合名詞の意味的構造

複合名詞を処理する場合、複合名詞を構成する単語間の関係によって解析を行い翻訳する[1]ことになるが、この方法では、構成単語数の多い複合名詞においては意味上分割が困難な部分を含むことが多く、高い翻訳品質を得ることは難しい。したがって、そのような複合名詞に対しては、複合名詞全体の構造的特徴に応じた処理を行う必要がある。複合名詞はその意味的構造により以下の４つに分類することができる。[2]

A：数量表現型複合名詞
　数詞、時詞を中心として構成される複合名詞
B：機能語支配型複合名詞
　「製」，「用」，「対応」等のキーとなる単語を中心として構成される複合名詞
C：固有名詞表現型複合名詞
　固有名詞を中心として構成される複合名詞
D：一般複合名詞
　構造的特徴がなくパターン化できない複合名詞

Analysis of the structure of compound nouns in newspaper articles
Hajime UCHINO, Akio YOKOO, Satoru IKEHARA
NTT Network Information Systems Laboratories

A，B，Cの複合名詞は単語単位に分割して解析すると、その意味を正しく把握することが困難となるため、複合名詞全体の構造的特徴に対応した処理が必要となる。

### 2．2　対象文

今回、調査を行ったのは新聞記事の、以下の５つの分野、各１０００文である。情報欄、経済欄、及び政治欄は基本的に読者に情報を伝達するための文であり、社説、投書は、会社・個人としての意見を陳述した文である。

情報１（日経産業）、２（日刊工業、日本工業）
経済１（朝日、毎日、読売）、２（日経）
政治１（朝日、毎日、読売）、２（日経）
社説１（朝日、毎日、読売）、２（日経）
投書１（朝日、毎日、読売）、２（朝日、毎日、読売）

## 3．分析結果

### 3．1　各分野における出現頻度

各分野に出現した複合名詞の種類別の頻度を表１に示す。対象文全体における、複合名詞の出現頻度は1文当たり約3件であった。

表１　複合名詞の出現頻度

| 文群 | 平均文長 | 複合語数 | | | | |
|---|---|---|---|---|---|---|
| | | 数量表現 | 機能語 | 固有名詞 | 一般 | 合計 |
| 情報１ | 52 | 376 (10.9) | 368 (10.7) | 762 (22.1) | 1,935 (56.2) | 3,441 (100) |
| 情報２ | 65 | 516 (12.4) | 392 ( 9.4) | 1,031 (24.7) | 2,231 (53.5) | 4,170 (100) |
| 経済１ | 67 | 593 (14.8) | 252 ( 6.3) | 1,144 (28.6) | 2,006 (50.2) | 3,995 (100) |
| 経済２ | 54 | 442 (12.8) | 246 ( 7.1) | 860 (25.0) | 1,894 (55.0) | 3,442 (100) |
| 政治１ | 68 | 524 (13.0) | 269 ( 6.7) | 1,696 (42.2) | 1,530 (38.1) | 4,019 (100) |
| 政治２ | 62 | 484 (12.3) | 298 ( 7.6) | 1,531 (38.9) | 1,622 (41.2) | 3,935 (100) |
| 社説１ | 46 | 145 ( 6.5) | 208 ( 9.3) | 632 (28.3) | 1,252 (56.0) | 2,237 (100) |
| 社説２ | 47 | 161 ( 7.2) | 243 (10.8) | 536 (23.9) | 1,304 (58.1) | 2,244 (100) |
| 投書１ | 40 | 245 (15.5) | 81 ( 5.1) | 243 (15.4) | 1,010 (64.0) | 1,579 (100) |
| 投書２ | 41 | 248 (17.9) | 81 ( 5.9) | 153 (11.1) | 902 (65.2) | 1,384 (100) |
| 合計 | 54 | 3,734 (12.3) | 2,438 ( 8.0) | 8,588 (28.2) | 15,686 (51.5) | 30,446 (100) |

（）内は各文セットにおける割合

・情報伝達型文章（情報、経済、政治）

これら情報伝達型文章においては、複合名詞の出現頻度が1文当たり3.8件と大変高くなっている。基本的に限られた紙面内で多くの情報を伝えるために、省略可能である助詞等をできる限り省いているためと推測される。このような情報提供型の文の翻訳においては、複合名詞の処理が重要な役割を果たすことがわかる。たとえば「情報１」においては、複合名詞が文中に占める文字の割合は約４割にも上る。

また「政治」の分野においては固有名詞表現型複合名詞の出現頻度がとくに高くなっている。これは、政治分野の翻訳を行う場合は人名、地名に関する複合名詞処理を強化すべきであることを示している。

・意見陳述型文章（社説、投書）

これら意見陳述型文章においては、複合名詞の出現頻度は1文当たり1.9件と情報伝達型に比べ半分程度になり、出現した複合名詞の中の一般複合名詞の比率が上がっている。このことから、意見陳述型の文章においては、無理に複合名詞を作らず、一般に使われている語を使用し、文を構成していることが推測される。とくに「投書」においては、一般の人の書いた文章であり、年齢、職業等も幅広いため、その傾向が強く現れている。

また、「投書」では、機能語支配型複合名詞の比率が下がっていることがわかる。このタイプの複合名詞は、説明的な文章にはよく使用されるが、通常の文章での出現頻度は低いといえる。

## 3．2　各分野の平均文字長

各分野における複合名詞の種類別の平均文字長を表２に示す。

・情報伝達型文章（情報、経済、政治）

これらの文章は、複合名詞の出現頻度も高かったが、複合名詞自体も長く、構造が複雑であることがわかる。

とくに「情報」において、その傾向が強く、中でも数量表現型複合名詞、機能語支配型複合名詞はとくに長くなっている。これらの複合名詞は構成単語数が全体的に多く、複合名詞の処理における問題点となりやすいため、この分野の翻訳を行う際には強化すべきポイントとなる。

・意見陳述型文章（社説、投書）

全般的に複合名詞の長さは短くなっている。また、機能語支配型複合名詞の出現頻度の低さとあわせて考えると、意見陳述型文章では複合名詞処理単独での重要性は低く、むしろ名詞句と複合名詞の関係を重視した処理が必要となる。

表２　複合名詞の平均文字長

| 文群 | 複合語平均長 | | | | |
|---|---|---|---|---|---|
| | 数量表現 | 機能語 | 固有名詞 | 一般 | 全体 |
| 情報１ | 6.49 | 7.10 | 7.52 | 5.45 | 6.20 |
| 情報２ | 6.57 | 7.54 | 8.31 | 5.53 | 6.54 |
| 経済１ | 5.45 | 5.15 | 6.46 | 4.69 | 5.34 |
| 経済２ | 6.09 | 5.14 | 7.28 | 5.12 | 5.79 |
| 政治１ | 5.78 | 4.26 | 7.51 | 4.51 | 5.93 |
| 政治２ | 5.28 | 4.24 | 7.63 | 4.60 | 5.83 |
| 社説１ | 5.25 | 3.97 | 6.89 | 4.34 | 5.08 |
| 社説２ | 5.62 | 3.37 | 6.47 | 4.51 | 4.94 |
| 投書１ | 4.33 | 4.20 | 5.24 | 4.21 | 4.39 |
| 投書２ | 4.12 | 3.81 | 5.39 | 4.31 | 4.36 |
| 全体 | 5.65 | 5.26 | 7.25 | 4.83 | 5.65 |

## 4．おわりに

新聞記事に出現する複合名詞に対し、各分野ごとに複合名詞の種類別出現頻度、平均文字長を調査し分析を行った。

情報伝達型の文章では出現頻度が高く、複合名詞の処理が重要な役割を占めることがわかった。また、数量表現型、機能語支配型、固有名詞型の複合名詞の平均文字長も長く、これらの意味的構造を持った複合名詞の処理を、このような文章ではとくに強化する必要がある。

意見陳述型の文章では、複合名詞の出現頻度も低く、文字長も短い。そのため複合名詞単独の処理より名詞句との関係を考える必要がある。

今後は、今回抽出した数量表現型、機能語支配型、固有名詞型のパターン的複合名詞に対する、半自動的なテンプレートルール作成方法の検討を行う。また名詞句と複合名詞の関係を把握するため、一般複合名詞を構成する単語間の意味的関係を調査する。

［参考文献］

[1]石崎:「日本語複合名詞の解析」,
情報処理学会第35回全国大会, 1T-1, pp. 1315-1316(1987)

[2]内野, 横尾:「テンプレートを用いた複合語翻訳方式」,
情報処理学会第44回全国大会, 2P-1, 3-135-136(1992)