日英機械翻訳における複合名詞翻訳処理

内野一　横尾昭男　池原悟
ＮＴＴコミュニケーション科学研究所

1　はじめに

日本文中には，助詞などを介さないで名詞が連続した名詞連続複合語（本稿では単に「複合名詞」と呼ぶ）が頻出する。次々と形成される新語に，辞書登録によってのみ対応することは困難であるため，複合名詞の要素間の関係を基に解析し，翻訳する[(1)]ことになるが、新聞記事などによく現れる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，要素合成による方式では，高品質な翻訳結果を得ることは難しい。構成要素数の多い複合語にも適用できる方法として，複合語間で共通した字面をパターン化して木構造として登録しておくことによって文法的に解析できない複合語に対処する方法[(2)]などが提案されているが，パターン化が表記のレベルであり，パターン数が収斂しないという問題点がある。パターン化によって，複合名詞を処理するためには，抽象化可能な範囲を定め，複合名詞の意味的な特徴に基づき，それら全体の構造をとらえて処理を行う必要がある。本稿では，複合名詞の意味的特徴に基づいた分類と，それぞれの特徴に対する処理の概要について述べ，機能語支配型複合語のテンプレート型翻訳を，新聞記事約11,000文に対して適用した結果を報告する。

2　複合名詞の構造特徴による分類と処理概要

2．1　新聞記事における複合名詞の特徴

新聞記事に出現する複合名詞をその構成文字数と構成要素数によって４つの分割すると，各部分に現れる複合名詞には次のような特徴があることが分かる。

A：構成要素数，構成文字数とも小の部分
　構造的な特徴がなく，パターン化しにくい複合名詞が多い。
B：構成要素数が小，構成文字数が大の部分
　固有名詞を中心として構成される複合語が多い。
C：構成要素数が大，構成文字数が小の部分
　数詞，時詞を中心として構成される複合名詞が多い。
D：構成要素数，構成文字数とも大である部分
　特定の接辞（「型」，「用」など），サ変名詞（「対応」など）を中心として構成される複合名詞が多い。

上記の特徴を持つ複合名詞を，それぞれ“一般複合名詞”，“固有名詞表現型複合名詞”，“数量表現型複合名詞”，“機能語支配型複合名詞”と分類する。この分類を実際の新聞記事約11,000文に対して適用した結果を表１に示す。

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
日経	46	295 (10.0)	320 (10.8)	592 (20.0)	1,746 (59.1)	2,953 (100)
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	4,029 (12.1)	2,758 (8.3)	9,180 (27.5)	17,432 (52.2)	33,399 (100)

（）内は各文群における割合

今回、調査を行ったのは新聞記事の、以下の５つの分野、各1,000文と日経産業新聞リード文965文（以下，「日経」と省略）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）

表１によれば，「日経」，「情報」、「経済」、「政治」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が，「社説」、「投書」では、出現頻度は1文当たり1.9件と半分程度となっている。これは，前者が基本的に読者に情報を伝達するための文であり、限られた紙面の中で多くの情報を伝えるため複合名詞を多用するのに対し，後者は会社・個人としての意見を陳述した文であり、比較的一般的な語を多用するといった文章の性質の違いによるものである。

また，各タイプの複合名詞の出現頻度も分野にばらつきがあり，翻訳の対象となる文章の分野によって強化していくべき処理が異なってくることが分かる。

２．２　複合名詞種別に応じた処理の概要

上述した複合名詞の構造的特徴に応じた以下の４つの処理について，その概要を述べる。

（１）一般名詞翻訳処理

構造的な特徴を持たない一般名詞に対しては，各構成要素の品詞種別及び意味を用いた要素合成による処理を行う。基本的には以下のような処理が必要となる。

・並列関係処理
　「研究開発」のように複合名詞内のそれぞれの名詞が並列に並んだ語の処理
・格修飾関係処理
　「心臓手術」のように用言と同じように他の名詞と格関係をとる名詞（用言性名詞）によって作られた関係を持つ語の処理
・修飾関係処理
　「映画音楽」のように一般名詞が連続するなど上記の二つの関係以外で結ばれた語の処理
・接辞処理
　接頭、接尾辞に関する処理

（２）数量表現型複合名詞翻訳処理

数詞及び時詞に承接する接尾辞を中心として，パターン化を行い，対応する英語表現への変換を行う。基本的に次の２つの処理から構成される。

・時間表現変換処理
　「９４年度第３四半期」のように時詞を中心として構成される複合名詞をその要素の並びによってパターン化し翻訳する。
・数量表現変換処理
　「月産３万台程度」のように定型的なパターンで表される複合名詞を要素の並びによってパターン化し翻訳する。

（３）固有名詞表現型複合名詞翻訳処理

“住所”，“人名”，“役職”などを要素の意味属性の並びによってパターン化し，訳出順を決定して翻訳する。

（４）機能語支配型複合名詞翻訳処理[3]
（以下，機能語処理と呼ぶ）

機能語支配型複合語は「型」，「用」，「向け」，「製」などの固有名詞承接型接尾辞，連体詞型接尾辞，もしくは，「対応」など一部のサ変名詞を中心としてその前後に複合名詞が接続した構造を持っている。機能語支配型複合名詞の構造は，これらのキーとなる単語を中心として前後に複合名詞が付くというパターンと捉えることができる。

また，これらの複合語の対訳構造も同様にパターン化することが可能であり，例えば「アメリカ製腕時計」という複合名詞の対訳は

”A wrist watch made in America.”

となり，

（後部複合名詞の訳）made in （前部複合名詞の訳）

のようにパターン化できる。これらをキーとなる単語ごとに，単語の前後の構造を条件部としてルールを構築する。ルールに必要となるのは，キー単語，前方，後方構造の条件情報，対訳構造である。このルールは形態素解析の結果として得られる意味カテゴリを単純に採用していくことによって，ほぼ機械的に作成していくことができる。以降ではこの処理を前掲の新聞記事に対して適用した結果について述べる。

3　新聞記事への機能語処理の適用

3．1　実験方法

ルール数の収斂，効果，分野による影響を確かめるため，以下のように段階的に実験を行った。

①日経産業新聞リード文，情報１，２，経済１，２の文中に出現する機能語支配型複合名詞を抽出し，その中ですでに辞書に一語として登録されている単語を取り除く。
②リード文から抽出した複合名詞の対訳，および形態素解析結果を基にルールを作成し，リード文自身に適用し，正しく適用できるよう修正する。
③前のステップで作成したルールをその他の文の複合名詞に適用し，評価する。
④以下，情報１，２，経済１，２の順で②③を繰り返し，ルールを追加していく。

3．2　ルール数の推移

対象となった複合名詞数と作成したルール数を表２に示す。処理を行う都合上，辞書に登録されている語は対象外としたため，複合名詞の数が各対象文で大きく異なっている。これを補正するため複合名詞数とルール数の比を示してある。これを見ると，各対象文において作成したルール数は明らかに減少している。これにより，対象とする文章の分野が異なっても，最終的にルール数は収斂していくことが予測できる。

3．3 ルールの適用

②③④においてルールを適用した結果を表３に示す。上段が適用された複合名詞の数，下段が全体からの比である。「日経」から作ったルールの結果以外は，ルールが増えたことによる増分が示してある。これを見ると，リード文から作成したデータは，分野的に近いと思われる「情報」，「経済」の複合名詞に対しては，ある程度の適応を示しているが，「政治」，「社説」，「投書」に対しては一割以下しか適用されていない。「情報」から作成したルールも同じような傾向となっている。それに対し，「経済」から作成したルールは，比較的高い適応率を見せている。

本実験では実際に出現した複合名詞のみからルールを作成したため，キーとなる語の出現頻度に差があると，適応率が大幅に変わると考えられる。対象とした文章に出現したキーとなる語の分布を表４に示す。リード文での出現頻度が高い“型”，“用”，“向け”などの語が「政治」，「社説」，「投書」ではほとんど出現しなくなっている。この出現頻度の差が，ある分野で作成したルールを他の分野に適用した場合の適応率の差に大きく影響している。出現頻度の差が比較的小さな分野に対しては，他の分野で作成したルールが有効に働くと予想することができる。

表２　作成ルール数の推移

対象文	機能語支配型複合名詞数	作成したルール数	複合名詞数／ルール数
リード文	２１１	７０	3.0
情報１	２６５	４７	5.6
情報２	２７７	２９	9.6
経済１	７６	６	12.7
経済２	１１７	４	29.3
全体	９４６	１５６	6.1

表３　ルール適用結果

対象複合名詞数／ルール数		情報1	情報2	経済1	経済2	政治	社説	投書	合計
		265	277	76	117	163	158	71	1127
日経	70	172 64.9	140 50.5	30 39.5	49 41.9	15 9.2	6 3.8	5 7.0	417 37.0
情報1	47		38 13.7						
情報2	29			14 18.4	15 12.8	13 8.0	7 4.4	5 7.0	92 10.7
経済1	6				19 16.2				
経済2	4					59 36.2	72 45.6	22 31.0	172 29.4
合計	156	172 64.9	178 64.3	44 57.9	83 70.9	87 53.4	85 53.8	32 45.0	681 60.4

3．4　今後の課題

現在の処理においては，１つの複合名詞の中に複数のキーとなる単語が現れ，複数のルールに合致する場合は，デフォルトで後方に現れた語を先にキー

として処理を行っている。そのためルールの条件指定が複雑となり，適応性が少し下がっている。今後は，どのキー単語を先にマッチさせるかの基準を作成し，同時に外部の名詞からの情報を使用して，ルールを適応させる必要がある。

4　おわりに

本稿では，意味的特徴に基づいた複合名詞の分類と，機能語支配型複合語処理を適用した場合の翻訳分野の影響について，新聞記事11,000文を対象に実験した結果を述べた。ルール数は分野が複数に跨っていても減少傾向にあり，収斂する。ただし，その適応性にキーとなる単語の出現頻度が影響し，分野の影響が現れることが分かった。

<謝辞>

複合名詞の抽出，分類，ルールの作成をしていただいた山本弥生さんをはじめとするＮＴＴアドバンステクノロジの皆様に感謝します。

［参考文献］

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

(2)藤田，辻井，長尾：「漢字連続複合語の解析」，第28回情報処理学会全国大会，7M-3,pp.1287-1288

(3)内野，横尾：「テンプレートを用いた複合語翻訳方式」，第44回情報処理学会全国大会，2P-5,3-135-136

表４　キーとなる語の分布

	日経	情報１	情報２	経済１	経済２	政治	社説	投書	合計
型	9.4 30	12.5 46	13.5 53	4.4 11	5.3 13	0.5 3	1.6 7	3.7 6	6.1 169
製	2.8 9	3.8 14	4.1 16	2.0 5	2.4 6				1.8 50
付き	0.9 3	0.5 2		1.2 3	1.6 4	0.4 2		0.6 1	0.5 15
式	3.8 12	0.5 2	0.5 2		0.8 2	1.2 7		1.9 3	1.0 28
専用	6.6 21	2.4 9	4.1 16					0.6 1	1.7 47
向け	15.0 48	19.8 73	11.5 45	4.4 11	8.1 20	1.4 8	0.4 2		7.5 207
性	5.9 19	6.8 25	9.7 38	15.5 39	15.0 37	19.9 113	14.9 67	9.9 16	12.8 354
用	20.9 67	20.4 75	17.9 70	2.4 6	7.3 18	0.9 5	0.2 1	3.1 5	9.0 247
対応	4.4 14	1.9 7	3.3 13	0.8 2	0.4 1	0.7 4	1.3 6		1.7 47
関連	4.4 14	4.1 15	2.8 11	6.0 15	5.7 14	5.8 33	1.6 7	3	4.0 109
関係	2.2 7	3.0 11	1.5 6	13.5 34	3.7 9	12.5 71	9.1 41	9.9 16	7.1 195
的	16.6 53	17.1 63	21.4 84	44.4 112	44.3 109	54.5 309	69.4 313	65.4 106	41.7 1149
系	3.8 12	5.4 20	6.9 27	3.2 8	3.7 9		0.4 2	1.9 3	2.9 81
版	3.4 11	1.1 4	1.5 6	1.2 3	0.8 2	1.4 8	0.7 3	0.6 1	1.4 38
調					0.4 1				0.0 1
状			1.3 5					0.6 1	0.2 6
入り		0.3 1		1.2 3	0.4 1	0.7 4	0.4 2	1.9 3	0.5 14
風		0.3 1							0.0 1
合計	320	368	392	252	246	567	451	162	2758

上段は各対象文での比率，下段は出現数