語の共起頻度を用いた複合語の解析

Analysis of Japanese compound noun using collocational information

23－6

小林義行* 山本修司 徳永 健伸 田中 穂積
KOBAYASHI, Yoshiyuki YAMAMOTO, Shuji TOKUNAGA, Takenobu TANAKA, Hozumi

東京工業大学 工学部
Department of Computer Science, Tokyo Institute of Technology

Analyzing compound nouns is one of the crucial issues for natural language processing systems, in particular for the systems that aim wide coverage of domains. In this paper, we propose a method to analyze structures of Japanese compound nouns by using both statistics of word collocations and thesauruses. An experiment is conducted in which 160,000 word collocations are used to analyze compound nouns of which average length is 4.9. Finally, the accuracy of the method is about 80%.

1 背景と目的

複合名詞の解析は、実用的な自然言語解析システムの実現において、解決しなければならない困難な問題の1つである。なぜなら、(1) 複合名詞は名詞を組み合わせることで無限に生成される、(2) 日本語には区切り記号がないので構成要素に分割することが難しい、からである。複合名詞を構成要素である名詞に分割し、複合名詞の構造を求める方法が必要である。

複合名詞「歩行者通路」を解析する場合を例にして考察してみる。辞書を検索すると「歩行者通路」は、「歩行/者/通路」と「歩/行者/通路」の2通りに分けられることが分かる。構造を考慮すると前者には [[歩行, 者], 通路] と [歩行, [者, 通路]] の2つの構造、後者には [[歩, 行者], 通路] と [歩,[行者, 通路]] の2つの構造、合計4つの構造が考えられる。この4つの候補のなかから、正しい構造 [[歩行, 者], 通路] を選択しなくてはならない。

複数の複合語分割結果から正しい解析結果を得るために、構成要素間の係り受け関係を解析する方法が提案されている。宮崎らは、語がとりうる概念に関する知識、語の係り受け関係について規則を記述して、これらの知識を用いて複合語の正しい分割と構造を求める方法を提案している [6]。辞書を整備することでこの方法は、高い精度 (94.6%) を実現しているが、この方法では

- 新しい言語現象に対応するための規則や知識の拡張や保守が容易でない
- 領域ごとに知識を用意するのはコストが高い

などが問題になる。

藤崎らは漢字複合語の正しい分割を得るために HMM モデルを用い、正しい構造解析を得るために確率付き文脈自由文法を用いる方法を提案した [2]。平均語 4.2 の漢字複合語を精度 73%で解析することができる。この方法には以下の問題がある。

- 複合名詞の分割を完全に統計的な方法で行なっているため、実際にはありえない語を用いた分割結果が得られることがある
- 文脈自由文法で用いる規則の数が多い
- 複合語は 1 文字語と 2 文字語から構成されると仮定している

本研究では、語と語の共起に関する知識とシソーラスとを用いた複合名詞の解析方法を提案する。語と語の共起関係は 4 文字漢字語コーパスを用いて獲得する。共起知識の獲得は以下のような処理となる。

1. 4 文字漢字語コーパスから語と語の共起関係を抽出する
2. 各語をシソーラスのカテゴリで置換え、カテゴリの共起関係を獲得する
3. カテゴリの共起の頻度を求める

ここで得たカテゴリの共起頻度を用いて構造の優先度を求める。

共起知識を用いて構造の優先度を求める研究には、Hindle らの前置詞句接続の曖昧性解消がある [3]。前置詞と動詞、前置詞と名詞それぞれの意味的関係の強さを共起頻度を基に計算し、前置詞句接続の曖昧性を解消している。

2 解析に用いる共起知識

この章では、複合名詞の解析に用いるカテゴリの共起知識をどのようにして獲得するのかについて述べる。その方法の概要は

*連絡先：小林義行 東京工業大学 工学部 情報工学科 田中・徳永研究室
〒152 東京都目黒区大岡山 2-12-1 Tel: (03)5734-2831 E-mail: yashi@cs.titech.ac.jp

1. 4文字漢字語の収集
2. 4文字漢字語を2つの2文字語に2分割して語と語の共起関係を求める
3. 各2文字語をシソーラスのカテゴリで置き換え、カテゴリの共起関係を獲得する。
4. カテゴリ共起の頻度を求める

2.1 語と語の共起関係の獲得

本研究では、2項共起の知識源として4文字漢字語を用いる。その理由は

1. 4以上の長さの漢字列はたいてい複合語と考えられる。分類語彙表では、記述されている漢字語のうち4%が4以上の長さの漢字語であったが、新聞など22万文から自動的に抽出した漢字列では、4文字以上の長さを持つものが71%あった。
2. 4文字漢字列を2つの2文字語に分割することで共起知識を得ることができる。分割して得た2つの2文字語が両方とも辞書に記述されている場合、精度は96%であった。
3. 田中康仁氏によって造られた4文字漢字列約16万語を含むコーパスが利用できる [7]

2.2 カテゴリの共起頻度の獲得

獲得した語と語の共起関係から、各語をシソーラスのカテゴリで置き換えることでカテゴリの共起関係を求める。単語が複数のカテゴリに属している場合、どのカテゴリの意味で用いられているか決めなくてはならない。

本研究では、カテゴリが一意に決まる語のみを含む共起データのみを用いた。本研究で用いるシソーラス「分類語彙表」では、複数のカテゴリに属する語はそれほど多くないからである。得られた語と語の共起データ中2/3では、2つの語ともカテゴリが一意に決定している。

複数のカテゴリに属す語を多く含むシソーラスを用いる場合、頻度を分配したり、統計的に正しいカテゴリを推定するなどの方法が必要となる [1, 5]。

本研究では、語の共起頻度は用いない。その理由は、(1) 十分な数の語について共起頻度を獲得するためには膨大な共起データが必要であるがそのような共起データを得ることができない、(2) 語のレベルで獲得した知識では、共起データを獲得したコーパスにない共起関係を処理することができない、からである。各カテゴリの共起頻度は、そのカテゴリ共起に置き換えられた語の共起の種類によって決まる。以下の手順によって頻度を求める。

1. 語と語の共起関係を抽出する。頻度は求めない
2. 得られた語をシソーラスで検索しカテゴリを求める
3. 得られた各カテゴリ共起の数を数える

3 共起知識を用いた解析

3.1 アルゴリズム

この章では、共起知識を用いてどのようにして、複合名詞の分割と構造を解析して結果の優先度を計算するか述べる。

その概要は、以下の通りである

1. 辞書の見出し語を用いて、可能なかぎり複合名詞の分割を求める
2. 各語のシソーラスのカテゴリを求める
3. 全ての構造について、共起頻度を基に優先度を計算する

複合名詞の構造は、二分木で表現できると仮定する。複数のカテゴリに属する語は、それぞれのカテゴリについて別々に優先度を計算する。また、日本語では左側の語が右側の語を修飾することから、各部分構造のカテゴリはその構造中で最も右にある語によって決まると仮定する。木 t の優先度 p は以下の式によって求める

$$p(t) = \begin{cases} 1 & \text{if } t \text{ is leaf} \\ p(l(t)) \cdot p(r(t)) \cdot cv(cat(l(t)), cat(r(t))) & \text{otherwise} \end{cases}$$

関数 $l(t)$, $r(t)$ はそれぞれ、木 t の左側の部分木、右側の部分木を返す。関数 $cat(t)$ は木 t のカテゴリを返す。$cv(cat_1, cat_2)$ はカテゴリ cat_1 と cat_2 が共起する頻度によって決まるある値を返す。本研究ではこの値として、以下の2つを用い比較した。ここで $P(Cat_1, Cat_2)$ は、左から cat_1, cat_2 と並ぶ共起が起こる相対頻度である。2つめの式は相互情報量の式で語順を考慮したものと考えればよい。

相対頻度 $cv_1 = P(cat_1, cat_2)$

修正相互情報量
(Modified mutual information statistics 以下 MMIS)

$$cv_2 = \frac{P(cat_1, cat_2)}{P(cat_1, *) \cdot P(*, cat_2)}$$

*はどのようなカテゴリでも良いことを表す

3.2 解析例

「歩行者通路」を例にして、どのように解析するかを説明する。

1. 全ての可能な分割を求める。
   1. 歩行/者/通路
   2. 歩/行者/通路
2. シソーラスのカテゴリを検索する
   1. 歩行 [133]/者 [110:120]/通路 [147]
   2. 歩 [119:133:145]/行者 [124]/通路 [147]

3. 優先度を計算する。曖昧なカテゴリが別々に計算されることに注意

(a) [[133,110],147], [133,[110,147]],
[[133,120],147], [133,[120,147]]
$p([[133,110],147])$
$= p([133,110]) \cdot p(147) \cdot cv(110,147)$
$= p(133) \cdot p(110) \cdot cv(133,110) \cdot cv(110,147)$
$= cv(133,110) \cdot cv(110,147)$

(b) ...

4 実験

4.1 実験データと解析方法

実験用のデータは、新聞のコラムと社説、用語辞典から抽出した複合名詞で、4文字語954、5文字語729、6文字語786でについて実験した。これらの実験用複合名詞は、自動的に抽出したものを人間が検査している。また、シソーラスに含まれない語を含む語は除いている。

シソーラスには、分類語彙表を用いた。分類語彙表は、木構造をしており階層が6段ある。階層とカテゴリ数の関係を、表1に示す。

表1 カテゴリ数と階層の関係

階層	0	1	2	3	4	5	6
カテゴリ数	1	4	13	94	510	833	6023

本実験では、階層3を選択した。

第2章で述べた方法によって、4文字漢字語とシソーラスから共起知識を獲得する。

第3.1章で述べたアルリゴリズムに従って解析を行なう。ここで、ヒューリスティクスとして自立語数最小法を複合名詞を構成要素に分割するさいに利用した。

4.2 結果と考察1

表2に解析結果を示す。∞は正解が得られなかったことを表す。～iはi位までに正解が含まれている場合である。正解が1位でありかつ一意に決定した場合は1の欄である。正解の90%以上がが2位以内である。確率とMMISの精度を比較すると、若干MMISのほうがよい。

複合名詞を分割するとき、自立語最小法をヒューリスティクスとして用いているが、このヒューリスティクスによって、正しい分割結果が得られなかったのは、4%であった。原因は、(1)「約二千万人」のような数詞を含む場合、(2) 適切な語が辞書に記述されていない場合、の2つであった。

表2 解析結果 [%]

	4文字		5文字		6文字	
rank	cv_1	cv_2	cv_1	cv_2	cv_1	cv_2
1	96	96	64	61	48	55
～1	97	96	70	66	54	62
～2	99	99	91	90	89	92
～3	99	99	92	92	92	94
4～	0.1	0.1	2	2	4	4
∞	1	1	6	6	5	2

複合名詞の分割において複数の候補があったのは64語であり、この方法で正しい分割と構造を得られたのはそれぞれ確率によって優先度を計算した場合18、MMISによって計算した場合19であった。

ここまでで述べた優先度の計算方法では、2つの語がどのような位置関係に現われても同じように計算しているが、この計算方法では不十分であると思われる。2語の位置関係によって変える必要があるか検討するために、複合名詞の構造ごとに出現頻度を調査した。調査結果を表3に示す。距離総和が大きい構造ほど、その出現頻度が低いことが分かる。距離は、2つの語の間にある語の数+1である。

表3 構造の分散

構造	5文字語	6文字語	距離総和
$[w_1]$	0	1	0
$[w_1, w_2]$	268	78	1
$[[w_1, w_2], w_3]$	283	406	2
$[w_1, [w_2, w_3]]$	84	160	3
$[[[w_1, w_2], w_3], w_4]$	13	43	3
$[[w_1, w_2], [w_3, w_4]]$	16	48	4
$[[w_1, [w_2, w_3]], w_4]$	4	11	4
$[w_1, [[w_2, w_3], w_4]]$	3	8	5
$[w_1, [w_2, [w_3, w_4]]]$	2	3	6

4.3 実験と考察2

距離の総和が大きい複合名詞が現われにくいという現象は、丸山が文節間の係り受け関係において、位置的に近い文節間の係り受け関係のほうが高い頻度で生じているという分析結果と関係があると考えられる[4]。丸山は、文節間の距離kと文節間の係り受け頻度の確率$q(k)$の関係を表す式を求めている。複合名詞の構造においても文節間の係り受け関係と同じ関係が成り立つと仮定して、優先度の計算に丸山の求めた以下の式を利用する

$$q(k) = 0.54 \cdot k^{-1.896}$$

上記の式を用いて2つのカテゴリの関係を以下のように再定義する。

$$cv'(Cat1, Cat2, k) = cv(Cat_1, Cat_2) \cdot q(k)$$

実験1と同じ共起知識、テストデータを用いて、距離を考慮した場合の精度を評価する実験を行なった。その結果をTable4に示す。

表4　解析結果 [%]

	4文字		5文字		6文字	
rank	cv_1	cv_2	cv_1	cv_2	cv_1	cv_2
1	96	97	76	82	63	72
～1	97	97	82	83	63	73
～2	99	99	94	95	92	95
～3	99	99	95	96	95	96
～4	0.1	0.1	2	2	5	3
∞	1	1	5	5	1	1

係り受けの構造によって出現頻度に違いのあることを、距離を尺度として解析に導入することで解析精度を向上できたことが分かる。

複合名詞の分割において複数の候補があったのは64語であった。この方法で正しい分割と構造を得られたのはそれぞれ確率によって優先度を計算した場合26、MMISによって計算した場合28であった。

5　まとめと今後の課題

本論文では、コーパスから共起知識を獲得する方法と、獲得した共起知識とシソーラスを用いて複合名詞を解析する方法について述べた。4文字漢字語を共起知識源として利用することで容易に高い精度の共起知識を自動的に得ることが可能になった。また、複合名詞の構造について、距離の総和が小さいものほど出現しやすいという分析結果を得た。共起頻度と語間距離を用いた解析実験によって精度を評価し、平均語長4.9の語に対して、確率を尺度として78%、MMISを尺度として83%で正解の優先度が1位になった。

統計的な知識は、詳細な規則を記述するのに比べ獲得が簡単であるが、統計的な知識のみではある程度の精度しか得られない。コストと精度をバランスさせ、コーパスから得られる統計的な知識、辞書などから抽出可能な言語学的な知識など半自動的に得られる知識と、人間が自然言語処理用に記述する詳細な規則を組み合わせることが重要な問題と考えられる。

今後の課題としては

- より詳細な意味分類の利用。意味知識源としてはEDRの概念体系が考えられる。ただし、大規模な知識源用コーパスが必要である。また、サ変名詞の選択制約などの意味知識の利用も考えることができる。
- 構造の偏りを反映させるのに今回は語間の距離を用いたが、これはアドホックな方法であり、さらに検討することが必要である。
- 接尾辞に関する知識などの統語的知識の利用

  例えば、複合名詞の構造は二分木で表現でき、木のカテゴリは最も右の語のカテゴリによって決まると仮定している。しかし、接尾辞のように振舞う語が含まれている場合、この接尾辞のカテゴリが構造のカテゴリに反映してしまう。また接頭辞で終わる語や接尾辞で始まる語を許している。
- 固有名詞を含む場合の対策
- 辞書の整備

  「許可」と「認可」から「許認可」が構成される場合が問題になる。このような語は辞書に記述するべきである
- 本手法を他の係り受け関係の曖昧性解消に適用

謝辞

4文字漢字列コーパスを提供して下さいました愛知淑徳大学の田中康仁教授に感謝いたします。

参考文献

[1] J. Cowie, J. A. Guthrie, and L. Guthrie. Lexical disambiguation using simulated annealing. In *COLING p310*, 1992.

[2] T. Fujisaki, F. Jelinek, J. Cocke, and E. Black T. Nishino. A probabilistic parsing method for sentences disambiguation. In *Current Issues in Parsing Thchnology*, chapter 10. Kluwer Academic Publishers, 1991.

[3] D. Hindle and M. Rooth. Structual ambiguity and lexical relations. In *ACL p229*, 1991.

[4] H.Maruyama and S.Ogino. A statistical property of japanese phrase-to-phrase modification. 計量国語学 18-7, 1992.

[5] D. Yarowsky. Word-sense disamibiguation using stastistical models of roget's categories trained on large corpora. In *COLING p454*, 1992.

[6] 宮崎正弘, 池原悟, 横尾昭男. 複合語の構造化に基づく対訳辞書の単語結合型辞書引き. 情報処理学会 論文誌 34-4 p743-p753, 1993.

[7] 田中康仁. 自然言語の知識獲得.-四文字漢字列-. 第45回情報処理学会全国大会, 1992.