特許文に含まれる複合名詞の解析

内山清子 ‡
‡ 慶應義塾大学大学院
政策・メディア研究科
〒 252-8520 神奈川県藤沢市遠藤 5322
<kiyoko@sfc.keio.ac.jp>

石崎俊 ††
† 慶應義塾大学
環境情報学部
〒 252-8520 神奈川県藤沢市遠藤 5322
<ishizaki@sfc.keio.ac.jp>

概 要

本研究は，特許文の適切な理解と検索への応用を目的とする．特許文に多く含まれる 2 つの語基から成る複合名詞を対象とし，語基の文法および意味情報を基にし，意味関係を Concept Discription Language(CDL) を用いて表現することにより意味関係解析規則を作成した．その規則を用いて 3 つの語基から成る複合名詞に適用し，高い精度を得ることができた．

1 はじめに

情報検索において，専門用語の抽出や，関連情報の効率的収集が重要な問題となる．特に特許の分野では，関連する特許の情報を漏れが無く，大量のデータから重要な情報だけを収集しなければならない．特許文書には多くの専門用語が含まれるが，専門用語に関連した複合名詞も多く存在し，その形式は一般的な語の羅列や結合，名詞句の省略など様々である．特に専門性の強い分野では簡潔な表現を好んで，多くの複合名詞を生成し，使用する傾向がある．そのため，同じ内容を異なる表現で記述する場合に，検索漏れが生じてしまう．たとえば，「構文木生成」は「構文木の生成」と「構文木を生成する」に言い換えることができ，特許文毎に表現が異なる．一方で複合語を構成している語基の羅列では不要な情報を大量に検索してしまう．検索漏れを減らし，且つ必要な情報を効率的に検索するためには，複合名詞の語基間の関係と，複合名詞間の類似度と関連度を明確に記述することが重要な問題となる．

そこで，本研究は複合名詞を構成している語 (以下語基と呼ぶ) の文法情報と意味情報に基づいた，語基間の意味解析解析手法を提案する．以下第 2 章では，複合名詞に関する先行研究などの研究背景を説明し，第 3 章は対象とする複合名詞の抽出方法と，複合名詞の構成語基の文法情報と意味情報の種類を述べ，第 4 章で意味関係解析規則の作成手順とその評価結果を考察し，最後に今後の課題を述べる．

2 研究の背景

複合名詞の研究は，日本語学と言語学の立場から語構成論として，語と語の結合，語基の品詞や語基間の関係などに関する先行研究が行われている [2][10]．自然言語処理における複合語の解析に関する研究は統計，文法理論，ルールベースの 3 つに基づいた手法が提案されている．統計に基づいた手法の特徴として，扱う問題が主に複合語の分割および抽出問題に焦点を当てている．語の共起の統計的な情報とシソーラスを用いて日本語複合名詞を解析した研究 [3] は，複合名詞の内部構造解析にも応用可能と考える．

文法理論を用いた研究として，竹内ら [11] は，主辞がサ変名詞である複合名詞の語構成において，単語間の係り関係を支配する語彙的性質に着目し，語彙概念構造 (LCS)[8][9] を利用した動詞の分類と，その構造を利用した名詞の分類に基づく複合名詞解析モデルを構築した．文法理論に基づいた手法は第 1 段階として人手によるタグ付け作業が必須であるが，記述能力が高く，詳細な言語現象に対しても対応可能であり，他言語でも共通した文法理論の枠組みで計算機上に実装できることから，今後進歩が期待される手法の 1 つである．

ルールベースに基づいた手法として内山ら [13][14] と宮崎ら [6][7] の研究がある．内山ら [13][14] の研究は，単独で出現する時の語の接続関係に基づき品詞相当カテゴリーを設定し，複合名詞の語基間の結合規則を作成し，解析を行った結果，この手法ではサ変接続名詞 (サ変名詞) を含む複合名詞の解析が難しいことがわかった．宮

崎ら [6][7] の研究は複合名詞の内部構造を，用言性名詞に対して詳細な係り受けの型と条件を設定した係り受け関係の規則を用いて適切に解析している．しかし，サ変名詞を含む場合に成立する格関係や修飾関係の詳細な規則が明らかではない．そこで本研究は，複合名詞の語基間の意味関係を解析するために，文法情報と意味情報の組み合わせを用いた手法を提案する．

3 複合名詞の解析

3.1 複合名詞の抽出

複合名詞を抽出するために，特許庁が公開している特許電子図書館の公報テキスト検索により，公報種別が「公開特許公報」のうち，発明の名称が「自然言語処理」を含む 165 個の文書を対象テキストとした．この文書を茶筅 [4] を用いて形態素解析し，品詞が名詞－一般，名詞－サ変接続，名詞－形容動詞語幹，未知語，名詞－接尾－一般，接頭詞－名詞接続，記号－一般に該当する単独あるいは連続する語を抽出した．この時，特許用語に多く使用される「上記，前記，当該，該，毎」は排除した．その結果，単独名詞を 1040 語，2 語基から成る名詞 (2 語基名詞）を 710 語，3 語基名詞を 610 語，4 語基名詞を 378 語，5 語基名詞を 126 語，6 語基名詞を 38 語，7 語基名詞を 13 語抽出した．

3.2 文法情報と意味情報

本研究では 2 語基名詞を対象に，構成語基の文法情報として名詞－サ変接続 (SN 名詞)，名詞－一般 (N 名詞)，名詞－形容動詞語幹 (NA 名詞) の 3 つを用いた．SN 名詞は自動詞，他動詞の区別と取り得る格の種類の情報を，日本語語彙体系 [1] を参照して調べた．また，従来の品詞枠における N に分類される名詞には，名詞と結合して複合名詞を生成する時，修飾語として用いられる語が含まれている [13]．たとえば,「一般知識」は「一般的常識」のように「一般」が「常識」を修飾する働きをする．従来の品詞の枠組みでは名詞に分類されるが修飾用法を持つ名詞を区別するために，本研究では「的」と「な」接続の有無によって新しいカテゴリ NAM を設定し，N と区別することにした．複合名詞の前項あるいは後項に位置する N が，公報テキスト中で連体修飾時に「的」あるいは「な」と接続する語を NAM 名詞に分類した．意味情報は，SN 名詞，NA 名詞，NAM 名詞は日本語語彙体系 [1] の用言意味属性体系を，N 名詞は一般名詞意味属性体系を参照して収集した．

3.3 語基間の意味関係

3.1 節で抽出した複合名詞の構成語基間の意味関係を表現するために，概念記述言語 (CDL:Concept Discription Language) を採用した．CDL は人間･社会とコンピュータとの間で意味を共有し，人間･社会とコンピュータにおける情報処理をつなぐ新しい技術の実現する手段として共通のインタフェースになるものである．XML(eXtended Markup Language) 上で開発されるコンテンツ (ドキュメントを含む広義のコンテンツ) の意味構造 (概念構造) を記述するための言語で，RDF(Resource Description Framework)+OWL(Web Ontology Language) と相補的関係にある．

CDL の暫定的な仕様 [5] は Universal Networking Language (UNL)[12] に基づいている．Concept Description Language for Natural Languages (CDL.nl) に記載された「関係」概念を，本研究では複合名詞を構成する語基間の意味関係の解析に用いる．CDL の「関係」概念は，基本的に文中の単語間の関係を表現するものであるが，複合名詞の語基間の意味関係を表現することにも有効だと考えた.「関係」概念の記述力や種類を検討し，CDL 仕様にフィードバックし，将来的には共通の記述仕様に従って日本語だけでなく，多言語に対応することが可能である．

CDL における「関係」概念は，事象内関係（格関係），実体間関係，限定・修飾関係の 3 つで，その中に更に 45 の関係が設定されている．本研究では，45 の関係の中から (1)Agt (agent:動作), (2)Obj (affected thing:対象), (3)Gol (goal, final state:終状態), (4)Aoj (thing with attribute:属性主), (5)Ins (instrument:道具), (6)Met(method or means:方法), (7)Mod (modification:限定), (8)Cnt (content, namely:内容), (9)Pur (purpose or objective:目的), (10)Seq (sequence:先行) を，複合名詞の語基間の意味関係とした．

4 意味関係解析規則

4.1 規則の作成手順

意味関係解析規則を作成するために，3.1 節で抽出した複合名詞を語基に分け，文法情報と意味情報を付与した．文法情報に基づく語基の組み合わせとして，SN+N，N+N，SN+SN，N+SN，NA+N，NA+SN，

N+NA，NAM+N，NAM+SN の 9 つに分類した．このパターン毎に 3.3 節で定義した意味関係を人手で判断して記述した．語基間の関係は言い換えを基準に判断を行った．言い換えは先行研究 [13] を基に，SN 名詞を動詞化したり，語基間に語を補って行う．その言い換えが実際に可能であるかどうかを公報テキスト検索を用いて調べ，意味関係を決定した．表 1 に文法情報別語基パターンと意味関係の対応を示す．[1]

次に，同じ意味関係を持つ語基パターンをまとめて，各語基の文法情報と意味情報を比較して共通属性を抽出した．その際，形態素解析では 2 語に分割されたが,「決定木，関係子，喩辞，指示詞」などと，第 2 語基が「接尾-一般」の品詞にあたる「語，度，値，先」などを含む語は 1 語として扱うべき語として語基間の関係を記述せず，最初から結合する原則とした．表 2 に前項 (複合名詞の第 1 番目の語基) を α，後項 (複合名詞の第 2 番目の語基) を β として，優先順に意味関係解析規則とその解析規則に対応した言い換え例を示す.

表 1: 語基パターンと意味関係

	文法情報別語基パターン						
意味関係	N+N	SN+N	N+SN	SN+SN	NA+N	NA+SN	N+NA
cnt	210	109	9	18			
pur	29	56	1	4			
met	1	1	1				
agt			4				
obj			76	36			
gol			3	3			
ins			2				
mod					23	6	
aoj							2
seq				1			
合計	240	166	96	61	23	6	2

4.2 評価

作成した意味関係解析規則に基づいて，2 語基間の関係を解析する規則は 3 語基間にも有効であると考え，3 語基から成る複合名詞 (3 語基名詞）に対して解析実験を行った．3 語基名詞のうち，出現頻度が高い上位 100 語を対象として，意味関係解析規則を適用した．規則を適用する前に，接頭詞，接尾を前接あるいは後接する語と結合させ 1 語とするが，結合後の品詞は接尾-サ変接続は SN，接尾-形容動詞語幹と「用」は NA とし，それ以外は N とした．3 語基名詞の場合，2 語基名詞と異なり係り受けの問題があるが，今回は先頭の 2 語基から順に意味関係解析規則に従って解析を行い，次に第 2 語基と第 3 語基の関係を解析した．ただし，NAM 名詞は後続する語と先に解析を行った．

4.3 考察

評価の結果は，意味関係解析規則と一致しない複合名詞が「構文意味解析」と「分割接続候補」の 2 つだけであった．両方とも第 1 語基と第 2 語基が並列の関係にあるが，解析規則では第 1 語基は第 2 語基の cnt(内容) を示す関係となる．この並列関係は意味情報を用いても判定が困難である．しかし,「構文・意味解析」と異なる表記の場合もあり，この表記から並列の意味であることを判断することが可能である．SN 名詞の連続は同じ意味情報の場合は And(連結) か Seq(先行) と判断できるが，その条件を満たさない場合は，共起関係や文脈から統計的に判断する方法を考える必要があり，今後の課題となる．今回評価に用いた 100 語は前方 2 語基が先に結合し，第 3 語基に係るパターンがほとんどであったが，後方 2 語基が先に結合する例もあることから，係り受けの優先度も考慮しなければならない.

表 2: 意味関係解析規則

優先度	α		β		意味関係	言い換え	関係の種類
	文法情報	意味情報	文法情報	意味情報			
1	形容動詞語幹 or 連体修飾が「的接続」	*	*	*	Mod	α な β，α 的な β	限定・修飾関係
2	名詞-一般	*	連体修飾が「な接続」	*	Aoj	α が β である	格関係
3	名詞ーサ変接続	β と同じ意味情報	名詞ーサ変接続	α と同じ意味情報	Seq	α して β する	実体間関係
4	*	*	*	0963 機械,0893 道具,1003 知的生産物 (思考/学習),1080 言語 (型式),1154 抽象物 (行為),2443 関連,2585 数量,2670 時間	Pur	α (する) の (ための) β	実体間関係
5	*	*	*	モデル	Met	α を用いた β	格関係
6	*	0003 主体	自動詞 or ニ格を取る他動詞	*	Agt	α が β する，α の β	格関係
7	*	≠0003 主体	二格を取る他動詞	*	Gol	α に β する，α の β	格関係
8	*	0962 機械	他動詞	*	Ins	α が β する，α で β する，α の β	格関係
9	*	*	他動詞	*	Obj	α を β する，α の β	格関係
10	*	*	*	*	Cnt	α の β	限定・修飾関係

[1]NAM 名詞は NA 名詞にまとめて表示

5　まとめと今後の課題

特許文に多く含まれる複合名詞を対象にして，複合名詞を構成する 2 語基間の意味関係を解析するために，各語基の文法情報と意味情報と CDL の意味関係記述を用いて意味関係解析規則を作成した．3 語基名詞に解析規則を適用して評価を行った結果，高い精度を得た．今後，3 語基以上の名詞を解析する段階では係り受けの規則も重要な問題となるため，語の共起関係や接続関係を統計的手法を用いて計算し，解析規則を詳細化し，同時に自動拡張できる仕組みを検討し，特許文の適切な理解を深める．

本研究では，CDL を用いて複合名詞の語基間の意味関係だけを記述したが，CDL は語基間の関係だけでなく，複合名詞間の関係にも利用できると考えられる．たとえば，分野オントロジーを構築するために，以下のような関係を記述することが可能である．

- Icl(included/kind of:上位)
  - 音声合成処理装置 ≫icl 音声処理装置
  - 音声認識処理装置 ≫icl 音声処理装置
- Equ(equivalent:同義)
  - 構文・意味解析 ≫equ 構文意味解析
  - 英字キャラクタ ≫equ 英字文字
- Pof(part-of:部分)
  - 検索エンジン ≫pof　検索システム
  - 翻訳モジュール ≫pof　機械翻訳システム

今後は，本研究の結果に基づいて，特許文に含まれる複合名詞の重要度から重要語を決定し，重要語に基づくオントロジー構築と，オントロジーを用いた効率的な特許文の検索に応用していきたい．

謝辞

この研究は総務省委託研究「戦略的情報通信研究開発推進制度 (SCOPE)」により実施したものです．

参考文献

[1] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦 (1999)．日本語語彙大系. 岩波書店.

[2] 影山太郎 (1993)．文法と語形成. ひつじ書房.

[3] 小林義行, 山本修司, 徳永健伸, 田中穂積 (1994)．“語の共起を用いた複合名詞の解析.” 自然言語処理研究会, 1994-NL-101, pp.1-8, 情報処理学会.

[4] 松本裕治，北内啓，山下達雄，平野善隆，松田寛，高岡一馬，湯原正幸 (2000)． 日本語形態素解析システム茶筅 (Version2.2.1) 使用説明書.

[5] Meiying ZHU and Hiroshi UCHIDA and Toshio YOKOI (2006)． Concept Description Language Specifications and Simple Syntax of CDL.nl. ISeC Technical Report.

[6] 宮崎正弘 (1984)．“係り受け解析を用いた複合語の自動分割法.” 情報処理学会論文誌, vol.25, no.5, pp.970-979.

[7] 宮崎正弘, 池原悟, 横尾昭男, (1993)．“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, vol.34, no.4, pp.743-753.

[8] Ray Jackendoff (1990)．Semantic Structure,MIT Press.

[9] Ray Jackendoff (1996). Conceptual Semantics and Cognitive Semantics Cognitive Linguistics, vol.7, pp.93-129.

[10] 斎藤倫明 (2004)．語彙論的語構成論. ひつじ書房.

[11] 竹内孔一, 内山清子, 吉岡真治, 影浦峡, 小山照夫 (2002)．“語彙概念構造を利用した複合名詞内の係り関係の解析.” 情報処理学会論文誌, vol.43, No.5, pp.1446-1456.

[12] UNL Center of UNDL Foundation (2005)． Universal Networking Language(UNL) Specifications (Version 2005)． `http://www.undl.org/unlsys/unl/unl2005/`

[13] 内山清子, 竹内孔一, 吉岡真治, 影浦峡, 小山照夫 (2001)．“専門分野における複合名詞解析のための名詞文法属性の分類について.” 計量国語学会, 第 23 巻 1 号, pp. 1-24.

[14] 内山清子 (2005)．コーパスを用いた日本語複合語の解析モデルの研究慶應義塾大学政策・メディア研究科博士論文, 慶應義塾大学