多言語専門用語抽出モデルの構築

竹内 孔一† 影浦峡‡ ダイユ・ベアトリス¶ 小山照夫‡
岡山大学工学部情報工学科†
koichi@it.okayama-u.ac.jp
国立情報学研究所‡
ナント大学¶

1 はじめに

本研究では多言語における専門用語抽出を行なうために用語の持つ文法パターンに注目してパターンベースによる用語抽出モデルの構築を行なっている．フランス語の用語抽出モデルにおいて Daille[2] はある基本単位からの派生関係を詳細な文法的パターンで記述してフランス語，英語に関する用語抽出を成功させた．そこで本稿では用語抽出における言語間での文法パターンを比較するために日本語で詳細な文法パターンを記述してどのようなパターンが得られるのかを明らかにする．

用語抽出の方法は大きく分けて統計的手法によるコーパスからの重要度測定 [6][1] による方法とパターンベースを中心に用語の派生関係を捉えて抽出する方法 [2] [7] がある．

本稿では形態素の単位を利用した用語のパターンを明らかにして，1999 年に行なわれた評価型用語抽出モデルワークショップ TMREC のの結果と比較する．その結果，本稿が提案する詳細なパターンベースの用語抽出モデルが統計的手法に対して低頻度の用語抽出で効果があることを明らかにする．

2 用語抽出モデル

本研究では Dalle が作成したパターンベースの専門用語抽出モデル ACABIT を利用する．入力文に対して形態素解析を行い形態素のパターンに適合したものを専門用語の候補として取り出し，複合語内の共起情報でフィルタリングを行う．そこで我々は日本語に対して ACABIT を利用するために日本語の語構成の文法的特性を分析し，機能範疇を作成してパターンを記述する．形態素解析システムは ChaSen[1] を利用しているが，具現化の段階で機能範疇を ChaSen の品詞体系に変換する．以下ではまず機能範疇について説明する．

2.1 文法カテゴリー

専門用語として通常用いられる表現は単語か複合名詞，あるいは短い名詞句である[2]．語構成を考えると形態素レベル，単語レベル，名詞句のレベルと要素にある結合の単位が存在することになる．どの複合レベルが単語もしくは基本単位かということは決定はできないが，用語のパターンを作成する上で形態素単位の結合パターンを機能を中心にまとめあげればよい．さらに日本語には和語，漢語，カタカナ語という語種の違いによる活用形や接続の違いがみられる．よって語種までふくめた文法カテゴリーを設定する．具体的には以下の表 1 のようになる．

これらは形態素の全品詞において，複合名詞，名詞句を構成するもののみを取り出した．機能カテゴリーは語構成において接続タイプが異なる振舞をするもについて設定した．以下ではこのカテゴリーについて簡単に説明する．

まず語種について．漢語とカタカナ語はどちらも接続に対して自由度が大きくどちらも外来語であるという観点から両方含めて「漢語」と表示してい

[1] http://chasen.naist.jp/hiki/ChaSen/
[2] 情報処理用語辞典 [4] では「の」に関しては A の B の C という 3 連続の「の」の表現は存在しなかった．

表 1: 文法機能カテゴリー

カテゴリー名	字種	例
Noun	漢と和	ブロック-長
AdjStem	漢	アクティブ-ジョブ
Prefix	漢	非-等価-演算
Suffix	漢	漸近-法
AdjStemSuffix	漢	連続-的-プロセス
NomSuffix	漢	3-型-文法
InfV	和	連続-紙-送り
InfA	和	深い-推論
AdjStemOJ	和	深-さ-優先-探索
ConOF	和	ブロック-の-頭

る．大きく分けて，上記のカテゴリーは 3 つに分類できる．1) 名詞の Noun と 2) 名詞句を接続する接続助詞「の」と，3) 活用と接辞に関するカテゴリーである．

Noun は通常の名詞，(例えば「ブロック」) だけでなく，サ変名詞 (例えば「測定」)，形容動詞 (「平行」など) もこのカテゴリーに入れる．また名詞句を作り出す接続助詞「の」の分類として ConOF をカテゴリーとして作成した．残りのカテゴリーは活用語と接辞に関する分類である．活用語は動詞 InfV，形容詞 InfA は和語であり，接辞は漢語に接続する接辞と和語に接続する接辞がある．先行研究 [5] において和語の活用と漢語と接辞による品詞の変更は対応があることがわかってきたので，こうした品詞の機能に関係する働きをもつ接辞について重点てきに詳細に分類している．

2.2 接続パターンの構築

前節で取り上げた形態素カテゴリーを利用して接続規則を作成する．Daille は基本 2 単語接続を仮定してそれらの派生ルールを記述するという形で接続規則を作成したが，フランス語と異なり，日本語の場合は基本単位が形態素であるため「接辞」など単語より小さい単位が基本単位である．そこで，3 段階の複合化，つまり接辞などが結びつくなどの形態素から単語のレベル，単語から複合語のレベル，複合語から名詞句の結合レベルにわけて記述した．

まず，形態素が複合して単語相当になるパターンを記述する．以下では BNF を利用して漢語における接続 (WordIM) と和語における接続 (WordOJ) 関係を記述する．

$$
\begin{aligned}
&< WordIM > \\
&\quad ::= \quad < Noun > | < AdjStem >< Noun > | \\
&\qquad\quad < Prefix >< Noun > | < Noun >< Suffix > | \\
&\qquad\quad < Noun >< AdjStemSuffix >< Noun > | \\
&\qquad\quad < AdjStem >< AdjStemSuffix >< Noun > | \\
&\qquad\quad < Prefix >< Noun >< Suffix > \\
&< WordOJ > \\
&\quad ::= \quad < InfV >< Suffix > | \\
&\qquad\quad < AdjStemOJ >< NomSuffix > | \\
&\qquad\quad < InfV >< Noun > | < InfV >< SuffixV > | \\
&\qquad\quad < InfA >< Noun >
\end{aligned}
$$

この複合化規則は接辞や語のレベルの必要に応じて構成されてる．例えば漢語の接辞「的」は AdjStemSuffix というカテゴリーに分類してあり，その接続特性は接辞であり複合化して語幹 (単語未満の単位) を形成する．よって上式のように前後に名詞を必要とする．また接頭辞 Prefix ならば次に名詞を必要とするなどである．また，語種による接尾辞の接続の異なりも規則に反映している．例えば和語における名詞化接尾辞「さ」NomSuffix は和語の形容詞の語幹と接続して 1 語を成す．和語動詞の活用形 InfV や InfA については修飾詞として次の語を修飾することで (例:深い-推論) ひとつの語となるのでそうした規則を盛り込んでいる．[3]

次に単語レベルをさらに複合して複合名詞のパターンを記述する．

$$
\begin{aligned}
&< CompIM > \\
&\quad ::= \quad < CompIM >< WordIM > | \\
&\qquad\quad < WordIM >< CompIM > | < WordIM >
\end{aligned}
$$

[3] こうした語形成の単位が形態素か単語か複合語かということはこの規則自身はなにも述べていない．ただ，機能として修飾機能のある形態素は修飾先の形態素を必要とするという必要性に基づく規則を記述しているのみである．よって語の単位についてこれで定義しているわけではない．

$< CompOJ >$
$::= \quad < WordOJ > |$
$< InfA >< NomSuffix >< Noun+ > |$
$< Noun >< InfV >< Suffix > |$
$< Noun >< InfV >< Noun >$
$< Noun+ >$
$::= \quad < Noun > | < Noun+ >< Noun >$

この複合化規則も語種によってパターンを分けている．漢語による複合名詞を CompIM で示しており，和語が入った複合名詞を CompOJ と示している．漢語の場合は表現したい内容に応じていくらでも長い複合語が可能でその機能面からの接続制約はほとんどない．一方和語の接続の場合まだはっきりとはしないが漢語に比べて比較的制約のある複合語のみが観測されている．和語の複合化規則はまだ網羅できていないが，用語辞書などによく記載されているパターンを観測しながら構築した．

次に名詞句のレベルの複合を示し，用語のパターンを示す．

$< Phrase >$
$::= \quad < CompIM >< ConOF >< CompIM >$
$< TERM >$
$::= \quad < CompIM > | < CompOJ > | < Phrase >$

名詞接続助詞 (ConOF) で接続されて複合名詞を形成し用語となる．ここで提案した複合パターンを利用して用語抽出モデルを構築する．

3 用語抽出実験

人工知能分野の抄録に対して用語抽出の実験を行なう．対象は 1999 年に開催された用語抽出ワークショップ (TMREC) の評価データで，著者キーワード (jsaiK) と人手により付与されたキーワード集 (jsaiM) が存在する．提案する用語抽出モデルが取り出した用語とこれらのキーワードとを比較することで用語抽出モデルの評価を行なう．

表 2: キーワードの特性

	jsaiM(%)	jsaiK(%)
抄録内	639/639 (100)	2890/4206 (68.7)
1 語用語	19/639 (3.0)	582/2890 (20.1)
上限	620/639 (97.0)	2308/4206 (54.9)

表 3: 用語抽出の結果

	jsaiM(%)	jsaiK(%)
recall	450/639 (70.4)	1867/4206 (44.4)
recall (UB)	450/620 (72.6)	1867/2308 (80.9)
precision	450/837 (53.8)	1867/19352(9.6)

用語抽出モデルは抄録の中から用語らしいものをパターンから抽出するので著者キーワードなど抄録には必ずしも出現していない語が存在する．そこで表 2 ではキーワードの統計的性質として抄録に含まれている語がどの程度あるのか，1 形態素からなる用語がどの程度あるのかについて明らかし，提案する用語抽出モデルの限界値を明らかにする．

表 2 において人手により同定された用語 jsaiM の上限値が高い．これは抄録の中から選択された用語であるためすべての用語が基本的に抽出可能であるからである．反対に著者が自由につけたキーワードは抄録内にある場合は 7 割以下で 1 形態素の用語を取り除くと本提案手法では約 55%程度が上限値となる．こうした背景を踏まえて用語抽出結果を表 3 に示す．

どの程度キーワードを取り出すことができたかという再現率 (recall) と，システムが出力した候補のうちいくつ正解だったかという適合率 (precision) を計算した．その結果人手によるキーワード (jsaiM) の場合，抽出元の抄録に全て記載されているため，再現率，適合率共に高い値を示した．一方，著者キーワード (jsaiK) は再現率は低いが，表 2 に示し

表 4: TMREC 用語抽出の比較

	A	B	C	D	本手法
recall	23.3	36.7	25.0	30.5	**44.4**
precision	10.7	6.4	7.0	8.3	**9.6**

た上限値から計算すると約 80%の再現率が得られた．抽出できるはずのキーワードのうち，8 割程度をパターンで抽出できることを示した．ただし，その場合適合率にあるようにほぼ 10 倍の解候補を出力する．抽出されたキーワードを以下に示す．

(正) 遺伝的アルゴリズム、知識ベース

(誤) 本研究、本論文、筆者ら

誤りにおける出力に対しては統計に基づくフィルタリングを行う予定である．

次に，本手法の結果と TMREC で観測された結果を比較しよう．

TMREC[3] では統計的手法とパターンベースの手法を融合した手法が 4 システム参加して用語抽出実験を行った．表 4 の A から D のシステムは具体的には提示されていないがこうした混合手法を利用したシステムである．先ほどの本手法の結果と比較してみると規則のみのシステムであるにもかかわらず再現率が高く，適合率も大きくは下がっていない．

この原因は対象となる抄録は 400 語程度の要約文であるため統計システムが十分発揮できる量ではなかったのではないかと推察できる．これはパターンベースが低頻度の用語抽出に強いことを示している．ただし，A から D のモデルも簡易なパターンは記述していることから，本手法のように詳細な接続関係を考慮したパターンを記述が良い精度を得る必須条件であることがわかる．

4 おわりに

規則ベースの専門用語抽出モデルにおいて語構成規則を形態素の機能を中心に分析して構築を行った．日本語の抄録に対して用語抽出実験を行った結果，統計的手法を利用したモデルを上回る結果を得た．今後このパターンを多言語で比較し多言語用語抽出モデルの構築を目指す．

参考文献

[1] Ananiadou, S.: A Methodology for Automatic Term Recognition, *In Proceedings of the 15th International Conference on Computational Linguistics (COLING)*, pp. 1034–1038 (1994).

[2] Daille, B.: Terminology Mining, *Information Extraction in the Web Ea* (Pazienza, M.(ed.)), Springer, pp. 29–44 (2003).

[3] Takeuchi, K., Yoshioka, M., Koyama, T. and Kageura, K.: Evaluation of the Keyword Extraction Task, *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp. 435–436 (1999).

[4] 相磯秀夫: 情報処理用語辞典（コンパクト版），オーム社 (1993).

[5] 竹内孔一, 影浦峡, 小山照夫, ロマリーローレント, ダイユベアトリス: 専門用語抽出における日本語の語の単位に関する考察, 第 *10* 回言語処理学会年次大会併設ワークショップ「固有表現と専門用語」発表論文集, *pp21–24* (2004).

[6] 中川裕志, 湯本紘彰, 森辰則: 出現頻度と連接頻度に基づく専門用語抽出, 言語処理学会論文誌,Vol.5,No.4, pp. 27–45 (2003).

[7] 佐藤理史, 佐々木靖弘: ウェブを利用した関連用語の自動収集, *2003-NL-153*, pp. 57–64 (2003).