日本語テキストからの複合語用語抽出

Composite Term Extraction from Japanese Texts

小山照夫[*1]

Teruo KOYAMA[*1]

1 国立情報学研究所

National Institute of Informatics

〒101-8430 東京都千代田区一ツ橋 2-1-2

E-mail: t_koyama@nii.ac.jp

*連絡先著者 Corresponding Author

用語は研究成果を記述するために用いられる言語記号であり，研究成果の記述された文献を高度利用し，研究のさらなる発展を期する上で重要な役割を担うものである．本論文では，日本語研究文献テキストから，自然言語処理技術を応用した手法により用語候補を機械的に抽出するための方法について論じる．日本語では多くの用語は語幹レベルでの複合語として出現するが，現在の形態素分類に従う複合語抽出をそのまま適用するのでは，再現率を保ちつつ用語抽出の適合率を高めることは容易ではない．本研究では，抽出対象となる複合語の内部構造と，テキスト内での複合語の前後に対する接続関係に制約を設けることにより，適合率を下げることなく多くの複合語用語候補を抽出することが可能となることを示す．また，抽出された複合語用語候補について，候補間に成立する入れ子関係や候補が関連すると考えられる部分研究テーマの側面から整理し，体系化する試みについて述べる．

Terms are used to describe important research concepts in academic documents, and are important to utilize the information in various research fields. In this paper, the author discuss about a method for extracting terms from academic texts based on natural language processing technique. Most of Japanese terms take composite word form, yet simple methods to extract composite terms based on current Japanese morpheme classification cannot attain enough precision. Considering internal structure of composite term candidates and the backward/ forward connective relations of the candidates in the texts, most of composite terms can be extracted with high precision. The author also discuss about the systematization of term candidates based on the nesting relations and the relationships of the candidates to various research sub-domains.

キーワード: 用語抽出，用語認識，自然言語処理， 用語体系化，形態素解析

Term extraction, Term recognition, Natural language processing, Term systematization, Morphological analysis

1 用語と用語抽出

用語は，学術研究を記述する上での重要な言語記号である．学術研究はその本来の目的からして，新しい概念を提案することが要請される．学術研究の成果を記述するためには，そこで提案される概念を表す言語記号が用いられなければならない．この役割を担うのが用語である．

用語は基本的には文章内で名詞的機能を持つ言語記号と位置付けられるが，これらは，単位的記号（形態素）と複数の形態素からなる複合的構造を持つものに分類できる．さらに後者は，複合語，句構造，節構造などに分類することが可能である．ただし，これらの全てが用語として適切なものと考えられるわけではない．

用語は概念記述の基本的な単位という側面も持っており，研究分野で取り上げられる概念を表すと同時に，一つのまとまりとして位置付けられることも要請される．影浦はこのあたりの事情をTermhoodおよびUnithoodとして提示しているが[1]，しかし，ある言語記号が用語として認められるかどうかは，多分に主観的な判定基準が入ってくる側面も存在する．

用語は様々な学術文献を体系的に関係づけるためにも重要な手がかりとなるものであり，学術文献を高度に活用する上でも重要な役割を担っているといえる．用語を研究分野ごとに整理して活用を促進することは，それぞれの研究分野においてさらなる研究発展を推進する上で重要な課題となる．

1.1 自然言語処理を活用した用語抽出

用語の整理はこれまで主に人手で収集と体系化が行われてきた．しかしながら，多くの文献を参照しながら多様な言語記号を抽出し．整理することは容易ではなく，すべての研究分野において十分な用語整理が行われているとは言えない状況が存在している．

この問題を解決する試みの一つとして，自然言語処理技術の活用により，研究文献テキストから機械的に用語を抽出する試みが行われてきており，用語抽出ないしは用語認識の問題と呼ばれている[2]．

用語抽出の問題では，テキストに含まれる用語をどの程度網羅的に抽出できるか，また，用語でないものを誤って用語と判定する割合をどの程度低く抑えられるかが問題とされる．これは基本的には情報検索等で用いられる評価指標である再現率や適合率と等価な指標と考えてよい．

用語抽出そのものは用語となりうる言語記号の抽出までを目的とするが，実際に用語を活用する視点からは，抽出された用語を整理する枠組みも重要になると考えられる．ある程度大規模なテキスト集合では，抽出される用語候補の数も膨大なものになるため，抽出結果を有効に活用するためには，候補間の関係を考慮し，用語の重要性や標準化なども意識しながら，体系的な整理を行うことが不可欠の問題となる．

1.2 用語抽出と統計的用語性判定

用語候補は名詞的概念を表現する文字列であり，先に述べたとおり，構造を持たない単一の形態素の形をとるものと，複数の形態素からなる複合的な構造を持つものとが考えられる．しかし用語としてみる限り，重要性のより高いのは複合的構造を持つものである．単一形態素からなる候補も確かに重要ではあるのだが，一般に単一形態素は概念粒度が大きく，かつ種類数としても限定されることとなる．

用語の多くは，各分野のより詳細化された

概念を表すものであり，多くの場合複合的な構造を取ると考えられる．複合的構造を持つ用語候補にはいくつかの種類が考えられるが，自然言語処理の視点からは，テキストの中で名詞的まとまりを持つ形態素列を見つけ出せば良いこととなる．

このような形態素列は一般には句構造を構成すると考えられるが，日本語の場合はやや特殊な事情があり，いくつかの例外を除いては，複合語（複合名詞）を考えておけば主要な用語のほとんどをカバーできると考えられる．これは日本語では，形態素を語幹レベルで取り扱うことにより，例えば英語などでは句構造として記述される内容を複合語として記述することが可能であり，また，用語として扱う場合，複合語化できるものはできる限り複合語としてしまう傾向が強いことによる．

このことは，日本語では用語候補として抽出すべきパタンが明確であり，例えば英語などと比較すると用語候補の抽出自体は容易であると考えられる．しかしながら，現在一般的に用いられている形態素情報に基づいてテキストから複合語用語候補を抽出してみると，そこにはかなりの割合で用語として不適切な文字列が含まれてくる．そこで，どの候補が複合語用語として適切であるかを判断する基準が必要となる．

この問題に対して，従来から広く提案されている方法に，候補のテキスト内出現に関わる統計的指標を用いる方法がある．候補の総出現頻度や，Tf-Idfなどで代表される特定文書への偏った出現傾向，特定形態素との共起傾向，候補の出現する文書の語彙的特異性などが，用語性を判定する有力な基準として提案されてきている[3-5]．

しかしながら，複合語用語候補に対してこれらの統計的判定基準を適用しようとすると，しばしば問題が生じる．統計的評価指標の適用に当たっては，評価対象が一定数以上出現していることが一つの要件となるが，多くの複合語の場合，その出現数はそれほど多いものではない．例としてNTCIR-Iに収録されている情報処理学会研究会抄録を見るならば，出現頻度がたとえ1であっても，用語としての価値が認められるものは決して少なくない．これらの低頻度用語候補について，その用語性を統計的指標で評価することにはもともと無理があると言わざるを得ない．

この問題を緩和するために，形態素間の結合の強さないしは弱さに注目して用語の境界を統計的に判定しようとする試みもなされているが[6,7]，この場合も形態素を品詞レベルで扱うのでは信頼性が欠けるし，形態素を個別に扱おうとすると特定の組み合わせについて出現数が十分ではないという問題もあり，低頻度候補の評価には限界があると考えられる．

1.3 外形的特徴による用語性判定

我々は，この問題に関して，統計的指標の利用を考えず，複合語用語候補の外形的特徴だけを手がかりに，候補の用語性を評価する方法を検討している[8]．

この背景には，日本語においては複数形態素が語幹レベルの結合として複合語化される多くの場合，ある程度用語的なニュアンスが入り込んでおり，逆に用語としてのニュアンスが少ない場合には句構造や節構造で記述が行われるという直観が存在する．

複数の形態素の集合として名詞的概念を記述する用語候補を，外形的特徴に基づいて抽出しようとする試みには，Daille等のフランス語用語抽出の試みなどがある[9]．フランス語では，複合語のみを対象とするのでは不十分であり，最低限句構造まで抽出する必要が

ある.

句構造による名詞的概念の記述は必ずしも専門性が高くない可能性があり，適合率の低下が懸念されるが，実際にはそれほど適合率は下がらないという報告もされている．この背景にはおそらく複合的構造で名詞概念を表記する場合，専門分野の概念を表現するものの数が圧倒的に多数で，一般語が入ってきてもその種類は相対的に少ないことがあるのではないかと推察される．

日本語では，基本的に複合語のみを扱うことで用語のかなりの部分をカバーでき，かつ，一般語が複合語の形を取ることはそれほど多くないと考えられるため，外形的特徴だけで十分な適合率が得られると予想されるが，実際に現在広く用いられている形態素辞書に基づいて，いわゆる複合語候補を抽出してみると，そこには相当数の不適切な文字列が含まれている．

我々は，誤って候補として抽出された不適切な文字列について詳細な検討を行った結果，不適切な文字列を抽出してしまう原因として，いくつかのものを洗い出した．その内でも主要なものは，現在一般に考えられている形態素分類が必ずしも充分でないという点と，形態素解析処理の誤りがしばしば不適切な文字列を抽出する原因となっているという事実である．そこでこれらの問題を回避する方法を検討した結果，文字列のパタンにいくつかの制約を設けることにより，不適切な候補の相当部分を排除できることを明かにした．さらに，候補の出現する前後の要素を調べることにより，候補として抽出すると誤りを生じやすいものを排除できることも明かになった．

以下では以上の考察に基づいて実際に専門分野の研究抄録テキストから用語を抽出する試みと，その結果について述べる

2　日本語複合語用語抽出

現在広く行われている用語抽出の試みでは，まず当該分野のテキストを収集し，これに形態素解析を施す．ここでさらに構文解析まで行うかどうかが一つの選択肢として考えられるが，その得失については後に論じる．以下では構文解析は適用しないこととする．

用語候補となりうる形態素列のパタンをあらかじめ決定しておき，解析結果中の，想定するパタンに合致する形態素列を抽出すれば，用語候補が得られることになる．しかしながら，先に述べたとおり，そのままでは適合率を上げることは難しく，形態素解析の誤りと，形態素分類が不適切であるという二つの大きな問題に対する対処が必要である．さらに候補形態素列の前後に制約を設けることも有効である．以下ではこれらの問題について述べる．

2.1 形態素解析誤り

形態素解析の誤りは，形態素辞書に登録されていない形態素が出現した場合や，形態素解析に多義性が存在する場合に，誤った解釈をとることに起因する．形態素解析誤りが生じると，不適切な形態素／形態素分類列が得られることとなり適切な複合語の抽出が困難になる可能性がある．

ただ，日本語で複合語の用語候補抽出を問題にしている限りでは漢字文字列やカタカナ文字列の形態素解析誤りはそれほど大きな問題とならないことが多い．これは日本語では文字種の相違は一般的に形態素の区切りを意味しており，また，形態素解析に失敗した場合にも，結果として得られる分類は未知語を含めて名詞系の分類となるため，複合語として合成すると結果として得られる文字列としては同一のものになることによる．

用語候補抽出において特に問題となるのは，平仮名文字列を含む形態素の解析誤りである．この種の誤りが生じた場合には，様々な形で複合語候補として不適切な文字列が切り出されてくることになる．

ただし，専門文書に出現する平仮名文字の名詞形態素の多くは複合語を構成すると考えにくいものであり，平仮名のみからなる名詞形態素を候補文字列から排除することを原則とし，分野によって重要なものがあれば，例外形態素として扱うことを考えることで，この問題はほぼ回避できる．

形態素解析誤りのもう一つの問題として，抽出対象となる候補の直近で形態素誤りが生じている場合，候補と考える文字列の前後が，本当に適切な文字列区切りとなっているかどうか，必ずしも保証できないことが考えられる．候補の前後の形態素を調べて，候補を切り出す位置で明確に文字列の区切りが存在するとは言えない場合，このような文字列を候補として切り出すことは，得られた候補集合の適合率を低下させる恐れがある．

適合率を維持するという視点からは，このような文字列は候補としない方が好ましい．しかし，一方でこれらの文字列を採用しない場合に再現率が低下するのではないかという懸念がある．

現在の我々の考え方では，まず，用語となるような重要な文字列は，少なくとも一度は独立した形で「提題的」に文書内に出現するという仮説を設けている．

ここで「提題的」というのは，候補の表す概念そのものを操作の主体，対象，手段などとして提示している状況を想定しており，典型的には文中で文節の先頭から始まり，「が」，「は」，「を」，「で」，「に」などの助詞に接続する形を想定している．

この仮説に従えば，文字列の前後に区切りの明確でない形態素が存在するものは候補として採用しなくても，再現率は低下しないことになる．もちろん実際には出現頻度の低い候補については，取りこぼしが生じる可能性を否定できないが，もともと頻度の低いものは抽出自体が困難であると考えるなら，前後の接続関係にある程度の制約を設けることにメリットがあると考えられる．

2.2 形態素分類の詳細化

次に問題となるのが，既存の日本語形態素辞書に登録されている形態素分類の問題である．特に顕著なものとして，日本語で空間的／時間的相対位置を表す形態素（上，下，前，後など）は，抽象的な対象分類を表す形態素（器，機，法など）とともに，名詞性の接尾辞と分類されており，複合語を構成する上では同等のものとして分類されていることが挙げられる．用語抽出の観点からは，この分類は不十分なものと考えざるを得ない．

例えば「測定－器」がある意味での用語として認められるのに対して，「航空－機－上」は用語としては認めにくいと考えられる．これら相対的空間（時間）関係を示す接尾辞は，例えば英語では前置詞に相当するものであり，英語で前置詞を含めた形の形態素列は通常は複合語とみなさないのと同様に，これらが末尾にくる列は複合語とはみなさない方が適切であることが多い．

この他にも，和語系の形態素で複合語構成要素になりにくいものや，特定の接頭辞で名詞句を構成するというよりは英語で言う指示詞に近い機能を持つもの，日本語特有の丁寧表現に関連する接頭辞など，原則として複合語構成要素から排除した方が良いものが相当数存在している．

2.3 実質内容を構成する形態素

外形的な判定基準としてもう一つ考えられるのが，実際に抽出された候補が明確な内容を持つと考えられるかどうかである．名詞的形態素の中には自立性の高い形態素と自立性の低い，いわば補助的／形式的な形態素とが存在する．

本来は名詞句を構成する単語（単一形態素ないしは複合語）は自立的な内容を持つはずであるが，形態素解析が必ずしも全面的に信頼できない状況では，自立的要素をまったく含まない文字列が候補として切り出される可能性を完全には排除できない．

以上の検討結果を踏まえて，テキストコーパスから実際に複合語用語候補を抽出する方法を実装した．以下ではその詳細と，方法を適用することによって得られた結果の概要について述べる．

3 日本語複合語用語抽出

これまでに我々の行ってきた実験では，用語抽出の対象とするテキスト集合（コーパス）として主に，NTCIR-Iテストコレクション[10]に収録された情報処理学会抄録集を用いている．このコーパスは全体で26,803抄録を含んでおり，タイトルを加えた抄録あたり文字数は平均約290文字で，標準偏差は約74.7文字である．

最初にこのコーパスに対して形態素解析を行う．現在は形態素解析器としてChasenを用いており，形態素辞書としてIPADIC-2.7を利用している．なお現在のところ構文解析は行っていない．

形態素解析の結果に基づいて複合語用語候補となる形態素列を抽出する．抽出に当たっては名詞系形態素の連接の内で他の連接に対して入れ子になっていない，独立したものだけを考える．これは，用語候補は少なくとも一度は独立した形でテキスト中に出現することを要請するものである．実際にはさらに候補形態素列の取る形式に制約を設ける．

3.1 候補形態素列から排除するもの

日本語では，名詞系の形態素の並びはすべて複合語とみなせるというのが基本的な考え方であり，いくつかに分類される名詞形態素に加えて接頭詞，名詞性接尾辞，形容動詞語幹，動詞連用形などが連接した文字列を複合名詞ととらえている．我々は名詞系に分類される一部の形態素を複合語の構成要素から排除することにより，形態素解析誤りの影響を緩和するとともに，日本語の複合語専門用語の構成から見て可能性の低いものを取り出さない方法を試みている．

具体的には1.平仮名一文字の形態素，2.平仮名のみからなる名詞形態素，3.平仮名を含む名詞接尾辞で例外指定されていないもの，4.補助動詞と考えられる動詞の連用形，5.指定される特定接頭詞，は複合語候補の要素とはしない．

これらの形態素は，一般に複合語を構成する形態素とはなりにくいと考えられる．ただし，分野によっては個別の形態素について例外的指定を行う必要が生じることも考えられる．

3.2 候補列の先頭／末尾要素の制約

特定の形態素を排除した形で得られる，独立した形態素列について，その先頭要素と末尾要素について制約を設ける．制約には，接尾辞で始まるものや接頭詞で終わるものを排除するという自明の制約に加えいくつかの制約を設ける．

まず，末尾の要素について，数詞，助数詞が来ることは許さない．これは，全体として数値を意味する語は用語とはみなしにくいと考えるものである．また，末尾が「上」「下」などの，英語表現では前置詞となると考えられる形態素で終わることも許さない．

一方，先頭要素については，副詞可能名詞や動詞連用形の内，あらかじめ指定したものが来ることを許さない．副詞可能名詞は，文脈によって機能的に副詞とみなされる場合と名詞とみなされる場合があるが，副詞として出現している場合には複合語の一部と考えるべきではない．同様に動詞連用形も，動詞の連用中止として機能している場合と名詞として機能している場合があるが，連用中止となっている場合には複合語の一部ではない．

これらの形態素が実際の文脈の中でどちらの機能を持っているかは確実には判断できないが，個別の形態素をみると名詞として出現することが極めてまれなものが存在する．このような形態素をあらかじめ先頭要素とならないものとして指定しておくことが有効である．

実際にはどの形態素について先頭に来ることを禁止するかは，分野によって調整する必要があろう．

3.3 全体構造の制約

以上の手続きによって残った形態素列について，全体の構成に関するチェックを行う．ここでは得られた形態素列が，具体的記述内容を持つことを要請する．

このためには，形態素列の構成要素の内，少なくとも一つは，一般名詞，固有名詞，サ変名詞，形容動詞語幹，動詞連用形，未知語のいずれかに分類されるものとなっていることを要求する．これ以外の，形式的な名詞系要素だけからなる列は，記述内容がないと判断する．

ここではまた，2形態素の列で，先頭要素が数名詞であるものも排除する．このような列は，後ろの要素が数接尾辞に分類されていなくても，ほとんどの場合数概念を表していることによる．

3.4 形態素列の前後接続関係

これまでの操作により抽出された形態素列のそれぞれについて，テキスト内でその前後にどのような要素が出現するかを調べる．

日本語では通常，名詞系形態素の連接は複合語を構成すると考えられるが，一つには形態素解析に誤りが生じる可能性を否定できず，また一つには例外的形態素を設けていることから，得られた形態素連接が複合語となっていることを必ずしも保証できない．そこで，前後の接続関係から明確に複合語として区分されていると考えられる連接だけを抽出の対象とする．

まず，連接の直後については，1.文末，2.助詞，3.接続詞，4.「・」，「／」以外の区切り記号，5.連接の最終要素が動詞連用形でない場合に限り助動詞，のいずれかであることを要求する．一方直前の要素については，1.文頭であるか，2.名詞，動詞連用形，「・」，「／」以外の形態素，であることを要求する．

連接直前の要素については，実は連接の先頭が文節の先頭になっていれば良いという事情がある．このことからすると，構文解析を行えばかなりの確度で文節先頭は認識できるように思われるが，実際には必ずしも構文解析を行うことが適切であるとは言い切れない．

現在広く用いられている，学習結果に基づく構文解析では，たとえば動詞連用形が一般名詞に接続する場合，動詞連用形は連用中止となっていると判断して，文節区切りを挿入

することがほとんどである. しかし実際には動詞連用形が名詞として用いられ, 後ろと連接して複合語を形成する場合も多い. 構文解析を実施することにより, この形の用語は抽出できなくなる可能性が高い. また, 形態素解析誤りが生じると, いずれにしても文節先頭位置を確実に決定することは難しくなる. 今回の試みでは, 構文解析なしに, 文節先頭である可能性が高いものだけを選択する方針を採用した.

4 用語抽出結果の評価

3節で述べた方法を実際に適用して, 先に述べた情報処理学会抄録 26,803 件を集めたテキスト集合から用語候補の抽出を試みた.

実際に今回の方法を適用した結果, 130,876 の用語候補が抽出できた. 抽出結果の適合率を見積もるために, このうちからランダムに 500 サンプルを抽出して目視による評価を行った結果, 複合語として成立しているかどうか, また, 情報処理分野の用語と認められるかどうかについて, やや甘めの基準で, 423 候補(84.6%)が情報処理分野の用語として認められるという結果であった.

情報処理分野の用語でないと判定されたものの内訳をみると, 分野外の複合語が 40 (8.0%), 複合語として不適切なものが 37 (7.4%) であった. 複合語として不適切なものが抽出された理由としては, 形態素解析誤りの影響を完全にはカバーできなかったことや, 用語判定にあたって用いてきたいくつかの基準で, 特別扱いをする形態素の範囲が狭すぎたと考えられることが挙げられる. 特別扱いをする形態素の範囲を厳しくすれば, 適合率は向上すると期待できるが, 一方で抽出できなくなる用語もあると考えられるので, 分野によって, あるいは目的によってある程度の試行錯誤により, 形態素の範囲を決定する必要があろう.

5 抽出された候補の体系化

これまでの我々の検討結果から, テキスト集合から相当数の用語候補を, 適合率を確保しながら抽出できることが明らかとなった.

しかしながら, 実際に抽出された候補を調べると, 抽出されたそのままでは利用しにくい形となっていることがわかる. 抽出された結果には, 用語として価値の高いものもそうでないものも含まれており, そのすべてを一様に扱うことは適当ではない. 実際には抽出された候補の中から, 不必要なものを削除し, さらに残った候補がどのように分類でき, 候補間の関係にどのようなものが考えられるかなど, 体系的整理を行うことが必要となる.

この整理のすべてを人手で行うことには非常に大きな労力が必要となることが予想される. そこで, 予めある程度の候補間の関係整理が行えないかが次の課題となる. 以下ではこれまでに我々が試みてきた用語候補の間の関係整理の試みについて述べる.

5.1 複合語入れ子関係に基づく関係整理

複合語は複数の形態素からなるが, 要素数3以上のものについては, より要素数の少ない複合語を入れ子の形で含むものがある. ある複合語が他の複合語を入れ子の形で含む場合, 複合語内の係り関係が入れ子関係と整合していれば, 入れ子関係にある複合語動詞は関連を持つことになる. この場合, 末尾部分が一致するなら相互の関係は上位語－下位語の関係になるし, 先頭部分が一致するなら関連語関係になると考えられる.

実際には係り関係が整合的なものであるという保証は必ずしもないが, とりあえず入れ子

関係に従って候補の間の関連付けを行うことはできる[11]. 実際に関連付けたものを調べると, それなりの妥当性はあるが, かなり見にくいものとなっていることが多い. 入れ子として含まれている候補に対して付加されている形態素の品詞に従って分類してみると, やや見やすい関連が得られる. しかし, 実際には現在の品詞分類よりは詳しい, 分野を考慮した形態素分類による並べ替えを考える必要があると考えられる.

5.2 部分研究領域に関連付けた関係整理

一つの学会に関連する研究文書は, 学会単位というよりはもう少し詳細な研究テーマを扱うものである. これらのテーマを仮に部分研究領域と呼ぶこととする.

部分研究領域はGenreの一種であり, 必ずしも明快に定義可能なものではないが, 文書を分類する上では便利に利用できる枠組みであり, 多くの学会で分類のための研究テーマ分類や, キーワード分類が行われている.

我々は, 特定の学会における部分研究領域が, 語彙分類の視点からどの程度認識可能であるか, また, テキストから抽出された用語がどのような領域と関連があるかについて検討を進めている.

一つの学会程度の比較的広い領域に属するテキスト集合が与えられた場合に, そこにどのような部分領域が存在するかを推定する方法として, クラスタリング手法の適用を考えることができる. クラスタリングは, 例えばWeb文書の分類などでも採用される手法であり, その効果はいくつかの研究でも確認されている. ただ, 文書その物をクラスタリングする手法では, 結果として得られた文書グループの意味づけに苦労することも多い.

これに対して語彙のクラスタリングは, クラスタとして現れた語彙に共通する部分領域を考えるという点で, より解釈が容易であると期待できる. 我々は過去に領域テキストに出現する語彙についてクラスタリングを試みている. クラスタリングを行う対象としては, 幅広い語彙を考えることもできるが, 比較的広い部分領域を推定するには, ある程度対象とする語彙を制限した方が処理の負荷を小さくできると同時に, 結果の解釈も容易になる, 我々の試み[12]では情報処理分野について, 50 程度の語彙をクラスタリングすることにより, かなり明確な5つの部分領域を同定することができた. さらに, 部分領域に対する各抄録の関連度も計算することができるし, この抄録関連度と抄録内候補出現傾向を用いて用語候補を部分領域と関連付けることも可能である.

6 今後の課題

我々はこれまでに日本語テキストからの複合語用語抽出と, 抽出された用語の体系的整理を援助する方法について検討してきた. しかし, 実際にこれらの手法を活用して, 適切に整理された用語集などを編纂するには残された課題も多い.

抽出された用語候補をより適切に整理するためには, 複合語の中で中心的役割を果たす形態素を同定できることが望ましい. また, 複合語内部の, 形態素間の係り関係をより正確に評価する方法が必要となる.

これらの目的のためには, 現在の日本語の形態素分類は未だ十分ではなく, 少なくとも対象領域において重要なものについてはより詳しい細分類が必要であり, それに基づく複合語内係り関係の解析手法, および複合語内の主要形態素認定の方法を確立していく必要がある.

開発された技術を活用するための環境の整備は, もうひとつの重要な課題である.

現在の用語抽出と用語整理の枠組みでは, 完全自動での用語集編纂は考えにくく, どこかで人手が介入してやる必要がある. ここでの作業を円滑化するための環境整備は今後の重要な課題である. 特に関連する情報資源を統合的に管理し, 用意された各種手法を適切に利用できる環境の整備が望まれる. 今後はこれらの課題についても検討を進めて行きたい.

謝辞

本研究の一部は科学研究費補助金 19500135 の援助のもとに行われた.

参考文献

[1]Kageura, K. and. Umino, B. “Methods of Automatic Term Recognition – A Review”, Terminology Vol.3, No.2, pp.259-289, 1996.
[2] Kageura, K. and Koyama, T. eds., “Special Issue on Japanese Term Extraction”, Terminology, Vol.6, No.2, 2000.
[3]Ananiadou, S. “A Methodology for Automatic Term Recognition”, Proc. COLING-94, pp.515-521, 1994.
[4]Hisamitsu, T. et. al. Extracting Terms by a Combination of Term Frequency and a Measure of Term Representativeness”, Terminology, Vol.6, No.2, pp.211-232, 2000.
[5]Mima, H. and Ananiadou, S. “An Application and Evaluation of the C/NC- value Approach for the Automatic Term Recognition of Multi-word Units in Japanese” Terminology, Vol.6, No.2, pp.175-194, 2000.
[6] Nakagawa, H. “Automatic Term Recognition based on Statistics of Compound Nouns” Terminology, Vol.6, No.2, pp.195-210, 2000.
[7]三浦康秀，増市博，“部分文字列のパープレキシティを利用した低頻度専門用語抽出”電子情報通信学会技術研究報告, NLC2007-1～28, pp.139-144, 2007.
[8]小山照夫，竹内孔一，“候補の接続関係を考慮した複合語用語抽出”情報処理学会研究報告，SIGNL-193, pp.13/1-6, 2009.
[9]Daille, B. et. al. “Towards Automatic Extraction of Mono-lingual and Bilingual Terminology” Proc. COLING-94, pp.515-521, 1994.
[10]Kando, N. and Nozue, T. (eds) Proceedings of the first NTCIR workshop on research in Japanese text retrieval and term recognition, 1999.
[11]小山照夫，竹内孔一，“日本語複合語の入れ子関係に基づく階層的体系化” 電子情報通信学会技術研究報告，NLC2007-1～28, pp.49-54, 2007.
[12] 小山照夫，竹内孔一，“用語クラスタリングに基づく部分研究領域推定と用語分類”情報処理学会研究報告，SIGNL-183, pp.87-92, 2008.