Vol.2009-NL-193 No.13
2009/9/29

候補の接続関係を考慮した複合語用語抽出

小 山 照 夫[†1]　竹 内 孔 一[†2]

テキストコーパスからの複合語用語抽出においては、抽出精度を低下させることなく、出現頻度の低い候補まで抽出することが重要である。従来主として用いられてきた統計的手法では、特に低頻度の用語候補の抽出に問題があった。我々はこれまでに用語候補となる複合語を構成する形態素の細分類に応じた位置制約を設定することにより、低頻度の候補まで抽出する方法を提案して来た。今回の発表では、この手法を改善し、多くの用語は文書中に少なくとも一回は提題的な形で出現するという予測の下に、候補となる形態素並びの前後接続関係に制約を設ける方法を提案する。実際にこの方法を適用することによりさらに低頻度の候補まで、抽出精度を落とすことなく取り出せることを確認した。

Term Extraction based on the Forward and Backward Connectivities of Candidates

Teruo KOYAMA[†1] and Koichi TAKEUCHI[†2]

In composite term extraction problems, it is important to extract candidates of relatively low occurrences in the corpora, with enough precision. In previous works, we have developed a method which can extract term candidates of low occurrences, using the revised classification of Japanese morphemes. In this paper, we propose a improved method considering forward and backward connective relations of candidates. Using the method, composite term candidates of less occurrences can be extracted with high precision.

†1 国立情報学研究所
National Institute of Informatics
†2 岡山大学大学院自然科学研究科
Graduate School of Natural Science and Technology, Okayama University

1. はじめに

研究文献テキストコーパスからの用語抽出は、文献検索をはじめとして、文献を高度利用するための重要な課題であり、これまでに多くの研究が発表されてきた [1-8]。

用語候補は名詞概念を表す形態素または形態素列としてコーパス内に出現するが、その主要なものは単一形態素ないしは複合名詞の形をとる。一般に専門文書では、詳細化された概念記述が重要となる傾向があり、より詳細な概念記述を可能とする複合名詞、すなわち複合語が重要な用語候補となる。このことから用語抽出問題においても、複合語用語の抽出が重要な課題となる。

複合語用語は一般に名詞系形態素の連続としてコーパス中に出現する。しかし一方で、コーパス中に出現する名詞系形態素連接のすべてが用語候補とみなせるわけではないため、適切な候補だけを選択する基準を設定する必要がある。

複合語として出現する用語候補のコーパス内出現頻度は、多くの場合それほど高いものではない。これは、より詳細化された概念記述を行う必要のある文脈は、全体の中ではそれほどは多くないことを表していると考えられる。NTCIR-I に収録された情報処理学会研究発表抄録コーパスを例にとるなら、自然言語処理に関連した、「品詞接続強度」、「対訳例文」、「形態素解析器」などのコーパス内出現頻度はいずれも 3 であり、決して高頻度で生起しているわけではない。

このことから、複合語用語抽出にあたっては、十分な抽出精度を確保しつつ、コーパス内生起頻度の低い候補まで抽出可能であることが要請される。

従来、用語抽出に関する多くの研究では、名詞系形態素連接についてその用語らしさを評価する指標として、候補のコーパス内生起頻度と関連した、様々な統計的尺度が用いられてきた [1],[3],[4]。これらの手法は用語性の高い候補を選び出す上で実際に有用性が高い。しかし一方で、頻度に大きく左右される統計的尺度のみを基準に判定する限り、生起頻度の低い候補を網羅的に抽出することは困難であると考えられる。

この問題を緩和するために、候補となる形態素並びについて、形態素間の接続関係に関する傾向を考慮したり、その出現する文脈を考慮したりする手法も提案されてきているが [4-6]、これらの手法の多くでも候補のコーパス内生起頻度を併せた評価尺度が用いられており、結果として、生起頻度の極端に低い候補を抽出することには限界があったと言える。

我々はこの問題に対して、複合語を構成する形態素の特性から、複合語内の形態素の位置に関する制約を設定し、制約条件を満たさない候補を排除することにより、コーパス内生起

Vol.2009-NL-193 No.13
2009/9/29

頻度 2 以上という条件ながら、精度を確保しつつ低頻度の用語候補まで抽出する方法を提案してきた [8]。しかしながらこの手法では、生起頻度 1 の候補まで含めると、抽出精度が相当程度悪化するという問題が存在した。

しかし、実際に文献 [8] の手法を適用した結果中で、これまで候補として採用しなかった生起頻度 1 のものを精査すると、そこには非常に多数の妥当な用語候補が含まれていることがわかる。さらに、生起頻度 1 の候補数は、生起頻度 2 以上の候補と比較してはるかに多数にのぼる。研究文献コーパスから可能な限り網羅的に用語候補を抽出するためには、たとえ生起頻度 1 であっても、適切な候補は抽出できる方法を確立する必要がある。

今回我々は、用語候補となりうる複合語は、コーパス内で少なくとも一回は、提題的な用いられ方をしていると考え、そのような場合に用語の前後の接続関係がどのようになっているかを考察した結果、用語候補の前後接続関係に制約を設けることにより、抽出精度を確保しながら、コーパス内生起頻度 1 の候補まで抽出する方法を考案した。

以下では 2 章で基本的な考え方と従来の手法との相違を述べたのち、3 章で詳細な用語候補抽出アルゴリズムを述べ、4 章で実際の用語抽出実験の結果を示す。最後に 5 章で考察と将来の展望を述べる。

2. 基本的な考え方

これまでに我々が提案してきた手法 [8] では、形態素解析誤りの結果として出現している可能性の高い形態素に配慮するとともに、既存の日本語形態素解析辞書に存在する形態素のいくつかについて、既存の辞書における文法的分類とは異なる特定のグループを設定することにより、用語抽出における当該形態素に関する扱いを調整するという方法を採用してきた。例えば、「内」という形態素は名詞接尾辞として分類されるが「器」などとは異なり、英語では前置詞に相当するものである。英語で前置詞から始まる並びは原則として複合語としないのと同様に、これらの形態素で終わる並びを日本語複合語候補とすることは一般に適切でないなどの制約を設けている。

同様の制約は候補並びの先頭要素についても設定されており、特定の形態素が先頭に来るものは用語候補としないとする方針を採用している。

これらの特殊な扱いを適用する形態素は、理想的にはより細かな形態素分類体系を導入することによって区別すべき問題とも考えられるが、現時点ではそれぞれのグループに属する形態素のリストを用意することによって対処している。

今回提案する手法は、基本的にはこれまでの手法を継承しながら、次の二つの点で変更を加えたものである。

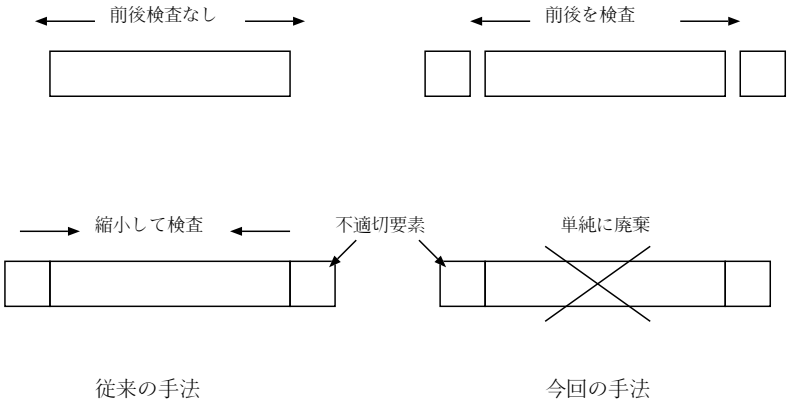

図 1 従来手法と今回手法の比較

(1) 従来手法では、候補並びの前後の形態素がどのようなものであるかについては特に制約を設けていないが、今回の手法では前後に来る形態素について相当程度強い制約を設けている。

(2) (1) の制約を設定したことと関連して、従来手法では候補となる並びの先頭ないし末尾要素が不適切要素となっている並びが得られた場合、不適切要素を削除した部分並びについて再度候補としての評価を行っているが、今回の手法ではこのような並びが得られた場合単純に候補を廃棄するという方法を採っている。

以上の比較を図 1 に示す。

このような方針を採用したのは、もし重要な用語として用いられているのなら、それは少なくとも 1 箇所以上で、提題的に使われているであろうという予測に基づいている。ここで提題的という意味は、その表す概念を、単独で取り上げる形での記述であり、典型的には文節の先頭から始まり、「は」、「が」、「を」などの助詞に接続する形を想定している。

この予測が正しければ、連接として得られる形態素列の最大のものだけを取り出し、その前後の連接関係を確認したのち、得られた並びを構成する形態素が用語候補としての条件を満たしているかどうかだけを検査することにより、低頻度のものまで、用語候補を適切に抽出できると考えられる。

次章では具体的な手法の詳細について述べる。

3. 用語抽出手法

今回の手法では用語抽出は、

- 形態素解析の実施
- 候補形態素並びの選択
- 形態素並びの前後接続関係確認
- 形態素並びの用語候補としての妥当性確認

という 4 つの手順にしたがって行われる。

3.1 形態素解析

最初にコーパス内の全文書について形態素解析を実施する。今回は形態素解析器として Chasen を用いた。形態素解析器として Chasen/Mecab を利用する場合、形態素解析に加えて、Cabocha などの構文解析器を併用することも考えられるが、今回は構文解析は行っていない。

形態素解析器の解析効率という面から考えると、Chasen よりは Mecab の方が優れていると考えられるが、用語抽出問題では将来的には形態素辞書のメンテナンスが必要となることを考慮し、Chasen の方が辞書メンテナンスを行いやすいと判断している。

構文解析を適用するかどうかに関して言えば、文節区切りを確率的に正しく判断できるという点からは、構文解析の適用が有効と期待できる。しかし反面、Cabocha など、現在利用可能な構文解析器を適用することによって問題を生じる可能性もある。特に顕著なのは、動詞連用形や副詞可能名詞が直接に一般名詞に接続する場合である。

動詞連用形が直接に一般名詞に接続する場合、構文解析器は多くの場合これを連用中止とみなして文節区切りを挿入する。同様に副詞可能名詞が直接一般名詞に接続する場合には、これを副詞であるとみなして文節区切りを挿入する。

これは、品詞とその活用形に基づいた学習結果を用いて構文解析を行う場合には当然の結果で、これらの接続関係を品詞接続頻度からみる限り、ほとんどの場合ここに文節区切りを挿入することが正しい結果となる。

しかしこの現象をもう少し詳細に観察するならば、実は種類としては少数の、しかしコーパス内出現頻度としては非常に高頻度のパターンが存在することが、文節区切りを挿入することの根拠になっていると推定される。特に、「し（する）」、「でき（できる）」などのサ変名詞を受ける動詞の連用形や、「今回」、「次回」などの時間的関係を表す副詞可能名詞は、生起頻度が高くかつ、一般名詞に接続するほとんどの場合に文節区切りと判定するのが正しいことになる。

しかし一方で、これら以外のさまざまな種類の動詞連用形や副詞可能名詞が一般名詞に直接接続するパターンを検討すると、そこには決して少なくない割合で複合語が形成されている場合のあることがわかる。たとえば「すべり—軸受」、「せり持ち—梁」、「平常—状態」、「突然—死」などはいずれも複合名詞を構成している例とみなすことができる。しかし、構文解析の結果、ここに文節区切りが挿入されてしまうと、これらの候補が抽出できなくなってしまう可能性がある。

今回はこれらの候補も抽出可能とすることを優先した結果、構文解析は行わない方式を採用している。なお、上記例に示した、複合語を構成しにくい動詞連用形や副詞可能名詞については、例外形態素を管理するリストに基づく制約を設けることにより、ノイズとして抽出しないように配慮している。

3.2 用語候補形態素並びの抽出

日本語において複合語用語は名詞系の形態素が接続した並びとして出現する。したがって用語候補を抽出するためには形態素解析を行った後に、名詞系形態素の並びを求めてやればよい。この並びを求めるにあたって、以前に我々が提案した形 [8] に準じて、複合語の要素となりにくいいくつかの形態素については、例外的な取り扱いをすることによって用語候補抽出精度の向上を図っている。なお、例外的な扱いをする形態素には、文法的分類と外形的特徴から判断できるものと、該当するものをリストの形で指定するものがある。

このような形で一部の名詞系形態素を例外扱いすることにより、場合によっては用語候補として容認可能な並びまでを排除してしまう可能性がないとは言い切れない。どの要素を例外とした場合どのような並びが排除されるかについて、分野に依存した判断に基づいて、適宜判断を行う必要があろう。

今回は Chasen の形態素解析結果から、次のものについては有効な名詞系形態素とは考えないこととしている。

- 平仮名一文字の形態素
- 平仮名だけからなる一般名詞
- 代名詞、助動詞語幹、および「問題」を除くナイ形容詞語幹
- 特殊な文字を含む未知語
- リストとして指定する、連用中止と判断する動詞連用形
- リストとして例外指定する以外の、平仮名を含む名詞接尾辞

- リストで指定する特定の接頭詞

3.3 候補形態素並びの前後接続関係の確認

3.2. で述べた要素を排除した上で得られる名詞系形態素並びについて、その前後の接続関係を確認する。先にも述べたように、ここでの基本的な考え方は、用語としての価値を持つ形態素並びは、名詞的なまとまりとして少なくとも一度は独立した提題的な形で出現していると期待されることである。

用語候補となりうる形態素並びは、たとえば文節先頭から出現して、助詞に接続するか、あるいは文末や句読点などの適切な文区切りで終了するなど、単独でまとまった形で出現する部分があると考えられる。

提題的な記述に現れる候補並びがその前の要素と妥当な接続関係にあることは、基本的にはその並びが文節先頭から始まることである。これは文節境界を確認することと同じ問題になる。文節境界を確認する有力な方法の一つは構文解析を適用することであるが、先に述べた理由から今回は構文解析を行っていない。ただし、このことはあらゆる場合に構文解析を行わないことが妥当であることを意味するものではなく、対象とするコーパスの特性に合わせて判断すべき問題であろう。

構文解析を行わない場合にも、名詞的要素からなる形態素並びの場合、原則的にはその並びの先頭は文節先頭になっていると期待できる。ただし、今回の手法では一部の名詞性形態素を並びに含めないという処理を行っていることから、別途配慮を必要とする。また、各種記号は普通、名詞的形態素とはみなさないが、「・」、「／」などはその前後を合わせて一種の複合語を形成する可能性があるため、正しい文節区切りとならない可能性もあることに注意する。

一方、候補並びの後に対する接続が適切かどうかは、明確にそこで並びが終了するパターンを考える。このような条件としては、文末であるか、上記のような特殊なものではない区切り記号であるか、後続形態素が助詞や接続詞である場合などが考えられる。また、並びの最終要素が動詞連用形でない場合、助動詞が接続している場合も適切な末尾であると考えることができるであろう。

これらをまとめるならば、

- 並びの先頭については、並びが文頭から始まるか、あるいは直前の形態素が
  - 名詞
  - 動詞連用形
  - 「・」または「／」

  のいずれでもないこと。
- 並びの末尾については、並びが文末であるか、あるいは直後の形態素が
  - 助詞
  - 接続詞
  - 「・」または「／」以外の区切り記号
  - 並び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

を満足している場合に当該並びを候補並びと考えることとする。

3.4 形態素並びの用語候補としての妥当性確認

これまでの手順で得られた候補並びについて、それを用語候補と判断して矛盾がないかどうかについて検査を行う。この検査は基本的には文献 [8] で行っているものと同じものである。検査は次の 5 項目について行う。

- 先頭要素の確認：並びの先頭に不適切な形態素がないかどうかを調べる。不適切な要素としては
  - 用意されたリストにより連用中止となると推定される動詞連用形
  - 用意されたリストにより副詞として利用されていると推定される副詞可能名詞
  - 接尾辞

  を考える。
- 最終要素の確認：並びの末尾に不適切な形態素がないかどうかを調べる。不適切な要素としては
  - 数詞および数接尾辞
  - 用意されたリストにより、英語での前置詞等に相当すると判断される接尾辞
  - 接頭詞

  を考える
- 末尾要素が動詞連用形の場合の制約：末尾要素が動詞連用形である場合、その直前の要素がサ変名詞となっているものは候補としない。
- 数で始まる並びの制約：数で始まる並びの場合、長さ 2 のものは用語候補としない。この形の候補は、例えば「2 −式」に見られるように、第二要素が明示的に数接尾であると判断されていなくても、ほとんどの場合に実質上数接尾とみなせることによる。一方で、長さ 3 以上のものでは、後ろのすべての要素の連接が数接尾を構成することは少ないと判断している。

Vol.2009-NL-193 No.13
2009/9/29

表 1 用語抽出実験結果の比較（NTCIR-I 情報処理学会）

	抽出候補数	精度	分野外複合語	非語
従来手法	46,609	85.8% (429/500)	7.2% (36/500)	7.0% (35/500)
今回手法	130,876	84.6% (423/500)	8.0% (40/500)	7.4% (37/500)

- 並びの中に有意味な形態素が存在する：並びに含まれる要素のうち少なくとも一つが次のいずれかに分類されていること
  - 名詞の内、一般名詞、固有名詞、サ変名詞、形容動詞語幹
  - 自立動詞
  - 自立形容詞
  - アルファベット記号
  - 未知語

これは少なくとも一つの形態素が明確な意味を担う可能性を持つことを要請している。

以上、3.1. から 3.4. までの 4 段階で抽出および検査を行った結果、残った形態素列を用語候補として抽出されたものと考える。

4. 用語抽出結果とその評価

3 章で示す手順に従って、実際に研究論文テキストコーパスからの用語抽出実験を行い、結果を従来の我々の手法 [8] と比較した。

用語抽出に用いたコーパスは、NTCIR-I 学会発表コーパスに収録されたものの内、情報処理学会研究発表抄録 26,803 件から成っている。各抄録はタイトルを含めて平均文字数約 290 文字、標準偏差 74.7 文字という規模のものである。これは文献 [9] で用いたものと同じコーパスである。

実際に用語候補抽出を行った結果の比較を表 1. に示す。

従来の結果と大きく異なる点として、抽出候補数が約 3 倍弱に増加している点がまず注目される。これは、頻度 1 の候補までを抽出したことによると考えられ、より低頻度の候補まで抽出するという所期の目的が達成された結果と考えられる。

用語抽出の精度を評価するために、それぞれの抽出結果から 500 サンプルをランダムに選び出し、その内容について評価を行った。判定にあたっては、広く考えれば情報処理分野に関連する概念を表していると判断できるという、やや甘めの評価となっているが、従来手法で抽出精度 85.8%(429/500) に対して今回提案する手法では 84.6%(423/500) という結果が得られている。数値自体はやや低下しているが、サンプリングの精度を考慮するなら、ほぼ同等の精度が得られていると考えられるであろう。

情報処理分野の用語ではないと判断されたものの中には、複合語としては成立しているが、情報処理分野に関連するとは言いにくく、他分野ないしは一般の複合語と考えられるものがある。これら分野外複合語の割合はそれぞれ 7.2%(36/500) および 8.0%(40/500) となっている。これも有意の差があるとは考えにくいが、頻度を 1 まで下げることによって、分野とは関連の薄い複合語が抽出される可能性が高まることには、特に矛盾がないと考えてよい。

その他、非語は形態素誤りの影響や、リストに収録していない例外要素の影響などにより、複合語とみなすことが不適切なものである。この割合はそれぞれ 7.0%(35/500) および 7.4%(37/500) で、これも両者で大差はない。ただし、今回の手法では TYPO による（また、その結果としての形態素誤りによる）と推定される非語が 0.8%(4/500) 出現していた。コーパス内出現頻度を 2 以上とする従来手法では、別の場所に出現する同一の候補に対して同じ TYPO が生じることはほとんど考えられないため、これまではこの問題が現れることはなかったが、今回頻度 1 の候補まで対象とすることにより問題が顕在化したと考えられる。

全体として見れば、今回採用した手法は、従来の我々の手法と比較して、ほぼ同等の精度を確保しながら、抽出用語候補数を大幅に増加させることができたと考えている。

5. 考　　察

今回、対象とする分野において、用語として成立する程度に重要な概念を表現する複合語は、少なくとも一度はコーパス中に独立して堤題的に出現しているという予測の下に、候補形態素列の前後接続関係に制約を設ける形での用語抽出を試みた結果、コーパス内生起頻度が 1 のものまで対象とした場合に、従来の我々の手法と同等の抽出精度を保ちながら、抽出候補数を 3 倍弱にまで増加させることが可能となった。

今回の方法により、複合語用語に関する限り、相当程度網羅的な抽出を可能とする方法が確立できたと考えている。

今回提案する手法を実際にさまざまな分野に適用する場合に予想される問題として、いくつかの形態素分類について、例外とすべき要素をどのように決定すればよいかという問題がある。「前」、「後」などの接尾辞は、多くの場合英語の前置詞に相当する使われ方がされ、ほとんどの場合には最終要素として不適切と判断できるが、しかし、これも絶対的なものとは言い切れない。実際には対象とする分野に応じて、抽出精度と抽出候補の範囲を勘案

しながら、ある程度は試行錯誤的に例外リストを調整する必要があると考えられる。この、例外扱いをする形態素を効率的に決定する方法を検討する必要があろう。

抽出候補数が増加し、抽出漏れが少なくできることは望ましい特徴ではあるが、抽出候補数が多数になると、何らかの体系的整理が必要となることが考えられる。既に我々はその最初の試みとして、用語候補間の入れ子関係を用いた候補相互間の関係整理を行う方法 [9] や、対象研究分野の部分研究テーマに関連付けて用語を整理する方法 [10] などを進めている。これらに加えて、候補の重要度を評価する指標についても検討を進める必要がある。

複合語の表す意味を、それを構成する形態素の意味とどのように関連付けられるかもまた、用語候補を体系的に整理する重要な視点を提供するものと考えられる。用語を構成する形態素分類から、形態素間にどのような意味的関係が成立する可能性があるかについても検討を進める必要がある [9]。

用語抽出にあたって、現在の形態素辞書は決して十分なものではなく、対象分野に応じての形態素辞書のメンテナンスが今後必要となることが考えられる。実際には辞書のメンテナンス環境と用語抽出システムを統合的に運用できる環境を構築していく必要がある。

実際に抽出された用語候補を、たとえば情報検索や文献の分類などに利用するためには、抽出された候補に重みづけをして、用語集の形で整理していく必要がある。この問題に関しては、ある程度人手で対処せざるを得ない面も出てくると考えられるが、ここで必要となる作業を支援する環境についても整備を進める必要があろう。最終的には、用語集そのもの、コーパス群、形態素解析辞書、用語間関係などを統一的にメンテナンスできる形での環境整備が望まれる。

今後はこれらの課題を検討し、実際に有効な用語抽出と利用を可能とするシステムの実現を目指したい。

謝辞 本研究の一部は、科学研究費補助金 19500135 の援助の下に行われた。

参 考 文 献

1) Kageura, K. and Koyama T. eds., Special Issue on Japanese Term Extraction, Terminology, vol.6, no.2, (2000).
2) Daille, B., Gaussier, E., and Lange, M., Towards automatic extraction of monolingual and bilingual terminology, Proc. COLING-94, pp.515-521, (1994).
3) Ananiadou,S., A Methodology for Automatic Term Recognition, PROC. COLING-94, pp.1034-1038, (1994).
4) 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, pp27-45, (2003).
5) Hisamitsu, T. and Tshujii, J., Measuring Term Representativeness, in Information Extraction in the Web Era (Ed. By Pazienza, M. T.), pp45-76, Springer, (2003).
6) Koyama, T. and Kageura, K., Term Extraction Using Verb Co-occurrence, Proc. 3rd International Workshop on Computational Terminology, pp79-82, (2004).
7) Takeuchi, K., Kageura, K., Koyama, T., Daille, B. and Romany, L., Construction on Grammer-Based Term Extraction Model for Japanese, Proc. 3rd International Workshop on Computational Terminology, pp91-94, (2004).
8) 小山照夫、影浦峡、竹内孔一、日本語専門分野テキストコーパスからの複合語用語の抽出、情報処理学会自然言語処理研究報告、2006-NL-176, pp55-60, (2006).
9) 小山照夫、竹内孔一、日本語複合語用語の入れ子関係に基づく階層的体系化、信学技報 NL2007-7, pp49-54, (2007).
10) 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、情報処理学会自然言語処理研究会報告、2008-NL-183, pp87-92, (2008).