社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2006－NL－176 (8)
2006／11／22

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

小山　照夫*　　影浦　峡**　　竹内　孔一***

*国立情報学研究所　**東京大学大学院教育学研究科
***岡山大学大学院自然科学研究科

テキストコーパスからの用語抽出は、自然言語処理技術の重要な応用である。従来テキストコーパスから用語候補を抽出する方法として、主として候補出現に関わる統計的指標を用いて用語性を判定する方法が採用されて来たが、統計的手法では出現頻度の低い候補についての判定が困難であった。今回の発表では、複合語に注目し、用語性を損なう形態素出現パターンを排除する形での用語候補抽出を行うことにより、高い精度で複合語用語抽出が可能となることを示す。

A Method for Extracting Composite Terms from Japanese Domain Corpora

Teruo KOYAMA*,　Kyo KAGEURA**,　and Koichi TAKEUCHI***

*National Institute of Informatics
**Graduate School of Education, University of Tokyo
***Graduate School of Natural Science and Technology, Okayama University

Term extraction is one of the most important application of natural language processing technologies. Statistic criteria are widely adopted to evaluate the termhood of the extracted candidates. However, it is difficult to evaluate the termhood of less frequent candidates. In this study we propose a method for Japanese composite term extraction in which unproper morpheme patterns are eliminated. Using the new method, high precision of term extraction can be attained for Japanese composite terms.

1. はじめに

テキストコーパスからの用語抽出は、自然言語処理技術の重要な応用として注目を集めて来ている。この問題に対して、従来の手法では、テキストデータを形態素解析にかけた後、用語候補となる形態素／形態素列を取り出した上で、候補の出現にかかわる統計値に基づいて用語性の判定が行われて来た[1-5]。統計的指標の利用は有力なものであり、一定の成果をあげて来たと言えるであろう。

しかしながら、一方で、統計的指標を利用した用語性判定では、用語候補の出現頻度が判定結果に大きく影響することから、相対的に頻度の低い候補については用語性判定が困難であり、抽出可能な用語は、単一形態素かせいぜい比較的要素数の少ない形態素列に限定される傾向があった。結果として、抽出された用語間の関係を、入れ子関係等に基づいて解析し、用語を体系的に整理する試みなどを行うためのデータとしては、必ずしも適切なものが抽出されて来たとは言いがたい。

筆者らはこれまでに、体系化された形での用語抽出を想定した上で、いくつかの日本語／外国語

の専門分野テキストコーパス解析と用語抽出を試みて来た[6,7]。
　これらの経験を通じて、日本語専門文書において複合語が用いられる場合、その多くが用語にかかわるものであるという感触を得ている。しかし一方で、従来提案されて来た、名詞系の要素からなる形態素列全てをそのまま単純に用語候補とする手法を適用した場合、多くの非語ないしは非用語が抽出されてしまい、実用に耐える抽出精度を実現することは困難であった。
　我々はこの問題に関して、単純な名詞的要素の連接をそのまま用いた場合に、非語／非用語が候補として抽出される原因を検討し、その結果として、特定の形態素配置パターンが、不適切な候補を抽出する原因の主要な部分を占めているという結論に達した。
　そこで、非語／非用語を抽出する原因となるパターンを排除した形で用語候補を抽出することができれば、より高い精度で複合語としての用語を抽出出来る可能性があると考えられる。
　今回の発表では、これまでの検討から明らかとなった、非語／非用語を抽出してしまう、有害な形態素配置パターンを可能な限り排除した形で用語候補を抽出することにより、複合語の形態をとる分野用語について、出現頻度が低いものまで含めて、高い精度で用語抽出が可能となる事を示す。

2．実験で用いたテキストコーパス

　今回の解析に利用したテキストコーパスは、NTICIR-I で公開されたデータの内、学会発表データベースに含まれる、土木学会の日本語テキストデータである。
　学会発表データベースは、参加各学会の研究会／全国大会研究報告の抄録を集めたものであり、NTICIR-I に収録された土木学会の場合、タイトルも含めて平均約 247 文字の文書 20,475 件からなる。
　解析にあたっては、このデータを juman を用いて形態素解析を行った結果を利用する。なお、juman を用いた事から、以下で述べる形態素分類は、基本的に juman／益岡文法に従うものとする。
　今回は、複合語の形式をとる用語のみを対象とし、単一形態素からなる用語や、接続助詞等で結ばれる名詞／名詞句、活用形容詞で修飾される名詞／名詞句は対象としていない。

3．非用語抽出原因の考察

　日本語文書から複合語候補の抽出を行う場合、名詞的要素、すなわち、各種名詞、形容詞語幹、動詞連用形、接頭辞、名詞性接尾辞、形容詞性接尾辞語幹、未定義語のいずれかの分類に属する形態素が連続する並びを候補として考えることができる。
　ところで、これらの形態素が連続する全ての形態素列を抽出してみると、結果として非常に多くの非語ないし非用語が含まれることとなる。単純にこのパターンに相当する形態素列を全て抽出したのでは、用語抽出として十分な精度が達成できるとは言えない。実際には形態素解析の誤りが加わることにより、抽出精度はさらに制限されることになる。
　単純に名詞系形態素の連接を用語候補として取り出した結果のうち、非語／非用語となっているものを調べて、語ないしは用語とは考えにくい列を抽出してしまう原因を調査した結果、形態素解析の誤りに起因する場合と、特定の形態素の存在が用語性を低下させている場合とがあるという結論に達した。

3.1．形態素解析誤り

　形態素解析の誤りについては、さまざまな専門分野では実際に、一般の言い回しとは異なる形態素が相当程度用いられていることから、ある程度は避けられないものと考えられる。ただし、本来単一の名詞的形態素と判断されるべきものが、二つ以上の名詞的形態素に分解されて判断されるケースは、複合を考える際に結局は正しい形に戻る場合がある。例えば「載荷（普通名詞）」は、形態素辞書登録を行わない場合、「載（未定義語）」＋「荷（普通名詞）」と分解されてしまうが、複合を考える際には結合されて元の形に戻る。
　解析過程が誤っていても、結果が合えば良いとすることには問題もあるが、現時点では、このパターンに相当するものは一応許容できる候補抽出であると考えている。
　これに対して明らかに問題を生じる形態素解析の誤りも存在しており、たとえば「せん断（普通名詞）」が「せ（する、サ変動詞）」＋「ん（ぬ、助動詞）」＋「断（普通名詞）」に分解されてしまうと、正しい候補を抽出することはできなくなる。このような場合、本来は形態素辞書項目の

追加を行うべきであるが、辞書追加の問題は改めて検討することとし、現時点では juman の標準の辞書と結合強度に従った解析結果を用いている。

3.2. 非語／非用語を構成しやすい形態素

非語および非用語を構成する原因となる形態素列について検討した結果、形態素解析の誤りも含め、問題の多くををいくつかの形態素出現パターンに帰着できることが明らかとなった。逆に言えば、問題となるパターンを排除した形で用語候補を抽出することが出来るなら、用語抽出としての精度を大幅に改善できることが期待される。

問題となるパターンについて、現時点では次のものを想定している。

グループ１：ある形態素の存在自体が、候補形態素列中の位置にかかわらず、語／用語としての性格を害しているもの

ひらかな一文字の普通名詞またはひらかなのみからなる固有名詞

多くの場合は形態素解析のあやまり、そうでない場合でも和語性が強くなりすぎ、用語として成立しにくい

ひらかな２文字の母音動詞連用形

和語性が強すぎ、用語として成立しにくい

ひらかなを含む接尾辞で、さ、づくり、作り、向き、を除くもの

和語性が強くなりすぎる

特定動詞の連用形（できる、よる、する、行う、およぶ、及ぶ）

形態素解析の誤り、または連用中止法の区切り判別の誤りであることが多い

仮に正しい解析結果であったとしても和語性が強すぎる

特定接頭辞（各、御、今、他、第、同、本、約）

用語として用いられる可能性が極めて低い

グループ２：先頭または末尾に来ることにより、語／用語としての性格を害するもの

最終要素が数詞または助数辞

数値が語の Head になる用語は特殊な場合を除いて考えにくい

最終要素が特定接尾辞（以下、以上、以内、以前、以後、以来、以降、前、後、前後、間、上、中、内、側、時、付近、程度、等、他、的、様、用、方、同士）

位置、時間、状態、関係など、基準からの相対的な変移を示すものなど、自立語として考えることに問題が残る

先頭要素がひらかなを含む形式名詞、副詞的名詞、時相名詞

何かを修飾するためには内容が乏しすぎる

グループ３：自立した語として問題があるもの

先頭要素が接尾辞または最終要素が接頭辞であるような列

普通名詞、固有名詞、サ変名詞、未定義語、形容詞語幹、動詞連用形を一つも含まない列

表現すべき概念の核になる要素が存在しない

これらの形態素出現パターンを排除した形で用語侯補抽出を行うことにより、高い精度で複合語用語の抽出が可能となることが期待される。

一方で、パターンの排除により、一部の正当な用語の抽出が不可能となることも事実で、例えば「塩化第二鉄」などは、候補として抽出できなくなる。しかし、このようなケースはあまり多くはないし、抽出できなくなるパターンもわかっていることから、この種の用語抽出が特に問題となる場合には、別途検討する方法を考えることができるであろう。

4. 用語候補抽出方法

前節で述べたパターンを考慮した上で、次の用語候補抽出アルゴリズムを採用した。

1.形態素解析を行ったテキストから、各種名詞、形容詞語幹、動詞連用形、接頭辞、名詞性接尾辞、形容詞性接尾辞語幹、未定義語以外の形態素および、上記グループ１に相当する形態素をデリミタとして、形態素解析結果から用語候補列を切り出す。

2.各候補列に対して、再帰的にグループ２の形態素を先頭ないしは末尾から削除する。

3.結果が長さ２以上の列であった場合、グループ３の判定により自立語として認められるものを用語候補とする。

以上のアルゴリズムを適用した結果、表１に示す数の用語候補が得られた。内容を検討すると、コーパス内出現頻度が１のものの中にも、興味深いものが相当数存在しているが、頻度が１しかないことは、偶然の要素に左右される可能性が多すぎると考え、今回は一応出現頻度２以上のものを最終的な用語候補として考えることとした。

	全抽出候補数	頻度２以上のもの
２要素	45887	20242
３要素	41425	13861
４要素以上	38568	9630

表 1．抽出された用語候補数

得られた結果について、用語抽出の精度を評価するため、各グループからそれぞれ３００候補をランダムサンプルし、自立語として適切か、用語性は認められるか、土木分野の用語と考えられるかについて判定を行い、表２の結果を得た。

分野がある程度広い範囲を有する場合、例えば工学全般といった場合には、用語の帰属する分野はあまり問題とはならない。しかし、土木工学というある程度狭い領域を考える場合、一般語ではないという条件に加え、当該分野の用語と考えて違和感がないかどうかも問題とする必要がある。

表２で、非語は、形態素解析の誤りまたは抽出アルゴリズムの不備により抽出された、自立語としては認めにくいものを示す。また、一般語は地名などの固有名詞も含んでいる。

	非語	一般語	他分野	当該分野
２要素	18	40	60	182
３要素	19	12	47	222
４要素以上	15	8	17	260

表 2．用語性判定結果

一般語と用語の境界、および用語の分野性判別は、それぞれ境界的な候補が存在し、完全に明確な区切りを見つけることは困難である。例えば、「バス利用者」は、普通に考えれば一般語であるが、都市交通の特性を問題とする都市工学的観点からは、わずかではあるが用語的な側面も存在している。あるいは「明石海峡大橋」は固有名詞ではあるが、最近の土木工学の重要な成果建造物という視点からは、土木分野の用語として判断することができるであろう。しかし、全ての大型土木建造物を土木工学分野の用語として良いものでもない。

用語の帰属する分野に関する判断にもいくつか問題がある。土木工学は応用工学として、本来さまざまな隣接分野と関連を持つことになる。技術の適用対象という観点からは、交通インフラ、都市基盤整備、河川・海岸強化、上下水道整備などに関連するさまざまな概念が入り込んで来る。また、実際に適用される技術という観点からは、機械工学、材料工学、都市工学という隣接分野から、情報処理、化学、数学、生物学などで得られた知見に至るまで、必要に応じて用いられることにな

る。ただ、土木工学という分野を考える場合、化学分野の用語（例えば「アルカリ金属イオン」）や情報処理分野の用語（例えば「知識ベースシステム」）などは、別分野の用語として扱うのが適当であろう。一方で、機械工学分野の用語（例えば「ねじり振動特性」）や都市工学分野の用語（例えば「交通量配分計算」）などは、土木分野内の用語としても大きな違和感はない。したがって表２の値は、境界の取り方によって多少は出入りがありうる。もともと、サンプリングによる評価結果であることと合わせ、ここで示された値は概略値であると考えるのが適当である。

このような事情を考慮しても、表２の結果からは、抽出された候補の内およそ８５%程度が何らかの分野の用語であることが推定される。専門文書に限れば、その中に出現する複合語の大多数は用語であると考えても良いであろう。

5．考察

ここで得られた結果をさらに詳細に検討すると、ある意味で当然のことながら、要素数２の候補に対して、多くの一般語が抽出されていることがわかる。一般語であっても、２要素程度の語を形成することはまれではないことを示しているといえよう。

２要素用語候補については、用語の分野性の判断は幾分曖昧である。用語候補を構成する形態素数が少ないことは、相対的にはその用語が意味する概念の粒度が荒いことを意味している。結果として例えば「不安定挙動」などの、さまざまな分野で用語となり得るものの比率が相対的に増加する。

これらの用語では、当該分野の用語と考えることが一応はできるが、他の分野でも用語として認められる可能性があり、かつ、分野が異なれば、意味する所も変わって来る可能性を持つ語であるといえる。これらの用語について、分野性を議論することの意味は、さらに検討を加える必要があると考えられる。

逆に、構成要素の多い複合語では、概念粒度が細かくなりすぎる傾向がある。「水平方向地盤反力係数」などは、確かに土木工学分野の用語ではあるが、例えば用語集編纂という立場を想定するなら、収録する単位としては、「地盤反力係数」に留め、方向性の問題は別途考えるのが妥当な判断となるであろう。

得られた用語候補が本当に用語として認められるかどうか、さらには問題とする分野の用語であるかどうかの正確な判断については、今後の検討を必要とする。

ある程度頻度の高い候補については、例えばコーパス内あるいは他分野コーパスと比較する形で、統計的指標を利用できる可能性もある。しかし、頻度の低い候補については、必ずしも統計値に基づく判断ができない場合も存在すると考えなければならない。この場合、候補を構成する形態素について、分野性の判断を行い、その結果から候補全体の分野性を判断するなどの検討が必要であろう。

今回の方法で排除しきれなかった、自立語として問題となるパターンの中には、形態素解析の誤りが決定的な影響を及ぼしているものもあるが、その他のものとして、不適切な接辞が残ったもの（例えば「新固液分離システム」）や、同様の問題が、形態素分類の問題から生じる場合（例えば「壁面近傍」、「近傍」は名詞に分類）などがある。これらを整理して、省略すべき先頭／最終要素にすることは当然考えられるが、特に、接辞に分類されていない要素の場合、副作用もあり得るので、慎重な検討を必要とする。

また、今回は・（中黒）や、-（ハイフン）による並列関係を無視したことから、これに起因するエラーもわずかながら見られた。例えば「Ｂ-Δ 法」は土木用語として適切なものであるが、これがハイフンで切断されて「Δ 法」のみとなると、適切な用語とは言えなくなる。

今回の研究では、形態素辞書に対する形態素の追加は考慮していないが、実際には対象分野毎の適切な形態素追加は重要な課題である。この問題に関して、今回の用語抽出結果をある程度利用することが考えられる。

２要素からなる複合語候補について、二つの要素のいずれもが一文字の漢字要素である場合、一般的な日本語における漢語要素としてはかなり特異なものであると考えて良い。これらの候補は、実は単一の形態素となるべきものが、形態素登録がされていないために二つの要素に分解されてしまった場合を多く含む可能性がある。実際にこの条件に該当する用語候補を調べてみると、その多くが本来単一の形態素として扱われるべきものであることが分かっている。

同様に、カタカナ文字列が二つの形態素に分解されている場合も、多くの場合本来一つのまとまりとなるはずのものが、辞書登録の不備により二つ以上の要素に分かれてしまったものであると考えられる。

これらの情報は、より高い精度での用語抽出を行う際に、追加すべき形態素情報の候補を与えるものとして利用できる可能性が大きいと言えるであろう。

6. 結論

日本語専門文書から用語を抽出する際、単純な名詞系形態素連接をそのまま複合語用語候補とした場合に、非語／非用語を多く抽出してしまう原因について検討し、特定の形態素の関与が大きく影響していることを明らかにした。また、問題となるパターンを排除した形で形態素連接としての複合語候補を抽出するアルゴリズムを提案することにより、コーパス内出現頻度の低いものまで、高い精度で用語抽出が可能となることを示した。

今後は、残された非語／非用語の抽出に関連する形態素出現パターンを検討することにより、更に用語抽出の精度を高めるとともに、用語候補出現に関する統計値や、用語候補を構成する形態素を手がかりに、用語と一般語の識別や、用語の分野性に関する推定を行う方法について検討を進める予定である。

今回提案した手法により、出現頻度の低い、多数の形態素からなる複合用語を抽出する可能性が示されたと考えている。今回の結果に基づいて、複合度の大きい用語候補を含めて抽出された用語候補の集合に対して、入れ子関係を解析することを通じての、用語間の階層関係推定に基づく用語の体系化や、用語候補を構成する要素としての形態素の分野別分類等に基づく用語候補の帰属する分野の推定等を通じて、体系化された形での用語抽出の方法論を明らかにして行きたい。

参考文献

1.Daille, B., Gaussier, E. and Lange, M., Towerds automatic extraction of monolingual and bilingual terminology, Proc. COLING-94, pp.515-521, (1994).

2.Ananiadou, S., A Methodology for Automatic Term Recognition, Proc. COLING-94, pp.1034-1038, (1994).

3.KAGEURA, K., and KOYAMA, T., Special Issue on Japanese Term Extraction, Terminology, vol.6 no.2, (2000).

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 Tsujii, J., Measuring Term Representativeness, in Information Extraction in the Web Era (Ed. by Pazienza, M. T.), pp.45-76, Springer, (2003).

6.Teruo KOYAMA and Kyo KAGEURA, Term Extraction Using Verb Co-occurrence, Proc. 3rd International Workshop on Computational Terminology, pp79-82, (2004).

7.Koichi TAKEUCHI, Kyo KAGEURA, Teruo KOYAMA, Beatrice DAILLE, and Laurent ROMANY., Construction of Grammar-Based Term Extraction Model for Japanese, Proc. 3rd International Workshop on Computational Terminology, pp91-94, (2004).