専門用語抽出における形態素辞書変更の効果

小山 照夫[1,a)]　竹内 孔一[2,b)]

概要：日本語用語抽出にあたっては形態素解析器および形態素辞書が必要であるが、多くのシステムで用いられている形態素辞書は、日常用いられる日本語文の解析を想定したものであり、対象分野の特殊性を考慮したものとはなっていない。本研究では抽出された用語の一部を利用して形態素辞書を編集（追加／削除）することにより、対象分野での抽出性能が向上することを示す。例えば単純な形態素追加でも情報処理学会、土木学会の二つの分野で 93.7%、88.9%の適合率で用語を増加させ、同時に 100%、90%の精度で非用語を削除できることを示す。これらの分析結果を元に、形態素解析の性能向上と辞書の関係について知見を述べる。

キーワード：日本語用語抽出、形態素解析、形態素辞書、分野固有性

1. はじめに

筆者らは現在用語管理支援システムを開発しているが [1][2]、このシステムでは、用語登録を支援する機能の一つとして用語抽出機能を利用した用語候補の提案を行っている。用語候補抽出にあたっては形態素解析機と形態素辞書が必要となるが、既存の形態素辞書は、一般的な日本語文書の解析を想定しており、対象分野の特殊性を考慮したものとはなっていない。このことを考慮し、開発中のシステムでは形態素辞書を編集する機能を用意している。

形態素辞書を対象分野に適合させて変更することにより、用語抽出の適合度や再現度を向上させることができるかどうか、また、どのような形態素の追加ないしは削除が有効であるかが興味の持たれる所である。

本稿では、既存の形態素辞書を用いて抽出された用語候補の中から、影響が大きいと期待できる用語を選択し、辞書に形態素として追加することにより、用語候補抽出性能がどのように変化するかについて調べるとともに、変化の要因についても考察する。

2. 追加形態素の選択

現在開発中の用語管理支援システムでは、用語抽出のために既に発表済みのアルゴリズムを使用しているが [3] このアルゴリズムでは現在のところ、chasen[4] を形態素解析器として使用し、また形態素辞書としては ipadic-2.7.0[5] を用いている。ipadic は一般的な日本語文書の解析を想定した形態素辞書となっているが、特定領域の文書を想定する場合、領域に適合した形態素辞書を用いることにより、解析精度を向上させ、ひいては用語抽出の性能を向上させる可能性がある。

どのような形態素を追加ないし削除すべきかは必ずしも自明ではないが、用語として確定しているものは、それ以上分割する必要性がないことから、抽出された代表的な用語を形態素として辞書に追加する方法が有効ではないかと考えられる。問題はどのような用語を代表的なものとして選択すべきであるが、今回は元の辞書による用語抽出結果のうち、抽出の元となったコーパスにおいて、ある程度以上の出現頻度を持ち、さらには多くの複合語構成に影響を及ぼすと期待できるものを選択することとした。

具体的には、コーパス内での候補の前後の接続関係から、候補が他の複合語の一部になっていないと考えられる形での出現頻度を F_i、前後関係に関わらず文字列として出現する総数を F_t とする時、F_i の大きいもの 300 の内から、用語でないと判定されたものを除いた上で、F_t/F_i の大きい順に 30 候補を取り上げることとする。

3. 用語候補抽出結果

上記で選択された用語を形態素辞書に追加して、新しく用語候補抽出を行う。chasen の場合、形態素には文法的分類と形態素コストを付与する必要があるが、とりあえず追加する形態素の分類は一般的な「名詞－一般」とし、コ

[1] 国立情報学研究所
NII, Chiyoda, Tokyo 101–8430, Japan
[2] 岡山大学大学院自然科学研究科
Okayama University, Okayama 700-8530, Japan
[a)] t_koyama@nii.ac.jp
[b)] koichi@cl.cs.okayama-u.ac.jp

ストは他の形態素より小さい「2999」を割り当てることとした。

適用する分野としては NTCIR-I[6] に収録された学会データベースのうち、情報処理学会を選択した。また、参考のため土木学会についても同様の実験を行った。形態素追加後の抽出結果と、元々の抽出結果との差分を取り、新しく抽出されたもの（出現）と、抽出されなくなったもの（消失）について、それぞれ用語であるか非用語であるかを判定した結果を表 1 に示す。

いずれの分野でも、用語の抽出数が増加すると同時に、非用語の抽出数が減少しており、一定の性能改善が見られることが確認された。非用語で新たに抽出されるものがいくつか出現しているが、その多くは「比べ－計算機」などのように、動詞連用形が関係している。

4. 改善効果の要因

前節において、抽出された用語の一部を形態素として登録することにより、抽出性能がある程度改善されることが明かとなった。以下では情報処理学会のデータを中心に、この改善がどのような要因によるかに関して考察する。

4.1 分野固有形態素

追加された用語を調べると、二文字からなるものがかなりの程度含まれている。今回利用したアルゴリズムでは、用語として複合語のみを抽出しているから、これらは一文字の形態素が二つ複合した結果となっているが、多くの場合このような結果は形態素解析が誤ったために生じたものであると考えて良い。特に、分野に固有な形態素が、元の形態素辞書に登録されていなかった場合には、このような解析結果となる可能性が大きい。すなわち、対象分野では本来形態素とみなすべきものが、形態素登録の不備により複合語として抽出されているものがあると考えられる。

本来分野固有の形態素とすべきものがどの程度抽出されるかは、分野ごとに異なり、情報処理分野では今回の基準で 30 位以内には「粒度」のみが含まれる。ただ、もう少し下位のものまで見るなら、「尤度」、「話者」、「S 式」などが抽出されていることがわかる。一方で土木工学分野では多くの分野固有形態素が抽出されており、上位 30 位以内に「載荷」、「圧密」、「床版」など 11 個が含まれており、下位のものまで調べるなら、さらに多くのものが抽出されていることがわかる。分野ごとのこの相違は、ある程度まで分野の歴史的経緯によるものと考えられるであろう。

分野固有形態素を追加したことによる用語候補抽出結果への影響については、後に述べる形態素の追加と削除をともに行った場合とほぼ同様であるため、そちらでまとめて論じる。

	情報処理学会		土木学会	
	用語	非用語	用語	非用語
出現	60	4	80	10
消失	0	78	4	36

表 1 形態素追加による抽出結果の変化

	用語	非用語
出現	49	4
消失	0	91

表 2 不適切形態素修正による抽出結果の変化

4.2 不適切な形態素

情報処理学会コーパスについて、形態素の追加により、新しく抽出可能となった候補を調べると、2 つの新しい形態素、「計算機」と「粒度」が複数の新しい候補の抽出に関与していることがわかる。

そこでこれらの形態素に関連する形態素解析結果を調べてみると、「計算機」の場合、例えば追加以前には「ホスト－汎用－計算－機上」と解析されているものがある。これは用語「ホスト－汎用－計算機」に、相対的位置関係を示す接尾辞「上」が加わった形であり、用語とは認めにくい。ここで「計算機」を追加することにより、「上」が正しく接尾辞と認識される結果、抽出が可能になる。この例では形態素追加により形態素解析が正しく行われるようになったと言える。一方「粒度」では例えば「細－粒－度－プロセス」では、形態素「細」が形容詞と判定されるため抽出できなかったが、「細－粒度－プロセス」では「細」が固有名詞と判定されることにより、抽出されるようになっている。この場合は形態素追加後も結果が正しくなったとは言えず、抽出が可能になったのは単なる偶然であると言える。

ここで有害な効果をもたらしていると考えられるのは、「機上」および「細」という二つの形態素である。「機上」の場合は、一般的な日本語文書の場合、これを形態素とした方がトータルとしての解析精度が上がると判断されていると考えられる。しかし、相対的位置関係を表す「上」という接尾辞を無条件で複合語構造の中に組み込むことになるため、用語抽出という目的には適切とは言えない。一方「細」の場合は、この形態素に対して接頭詞としての定義がされていないことが原因と考えられる。いずれにしてもこれらの形態素については、削除ないしは追加により、相当程度幅広い範囲で用語抽出性能を向上させうると考えられる。

ここで問題は、これらの形態素データだけを修正した場合にどの程度抽出性能が向上するかである。元の辞書に追加する形態素を接頭詞「細」のみとし、「機上」、「機内」の 2 形態素を除去した形態素辞書で情報処理分野の用語候補抽出を行ったとき、元々の結果とどのように異なるかを調べた。結果を表.2 に示す。

	用語	非用語
出現	63	5
消失	0	92

表 3 形態素追加／修正による抽出結果の変化

結果は新しく出現した正解用語候補数は形態素追加のみの場合に及ばないが、消失した非正解用語候補は増加している。これは「機上」等の形態素により、相対位置を表す接尾辞が最終要素として含まれることによる誤りが「計算機－上」だけではなく、「実機－上」、「大型機－上」などでも生じており、これらの出現が抑制されたことによると考えられる。

4.3 形態素修正と追加をともに行う効果

比較的影響の大きい問題形態素を追加／削除した結果は、不適切な候補を排除する上では効果が大きいが、一方で正解候補の抽出を増やす効果は、用語を形態素として追加した場合に比較して低くなっている。そこで問題形態素の追加／削除を行うとともに、抽出された用語を形態素として追加することにより、さらに抽出性能を向上できると期待される。

情報処理学会コーパスについて 3. で述べた形態素追加と、4.2. で述べた形態素修正をともに適用した場合の、用語候補抽出結果と ipadic のままの結果との比較を表 3 に示す。

この結果では期待通り、排除される非用語の数は 4.2. と同程度で、用語の新規抽出数は 3. における結果と比較してもわずかながら増加している。

形態素追加で抽出可能となった正解用語が、どのような要因によるものかについて、形態素解析結果を調べると、元々の形態素辞書（ipadic-2.7.0）に基づく解析では誤った結果となっていたものが、特定の用語を形態素として追加することにより、正しい解析結果となる場合が存在することがわかる。図 1. に解析結果が正しくなった二つの例を示す。

いずれも先に示したものが辞書変更前、後のものが変更後の結果である。最初の例では、形態素辞書変更前は「に（にる）－対－する」と解析されているものが、変更後では「に対する」と正しく解析されている。この相違は後続する要素が「時－系列」から「時系列」に変化したことによる。また、次の例では、変更前には誤って動詞と判定された「及び（及ぶ）」が、変更後は正しく接続詞と解釈されている。これは直前の要素が「可視－化」から「可視化」に変わったことによる。

他のいくつかの例についても調べてみると、一般的な名詞（名詞－一般、名詞－サ変接続など）の前後と比較して、それ以外の品詞の形態素（接頭詞、接尾辞、副詞など）の前後に位置するある種の助詞や接続詞などでは、誤って動詞や名詞と判断されるものが多いことがわかる。抽出された複合語用語を形態素として登録することにより、接頭詞、接尾辞、未知語／記号などを含めた形態素列を、一つの一般的な名詞として登録することになる。結果として登録した形態素の前後で形態素解析結果がより信頼できるものになるため、新しく追加した形態素の前後での用語抽出の性能が向上していると考えることができよう。

```
□一次元CAに対する時系列解析

□一次元□イチジゲン□一次元□名詞□一般
□CA□シーエー□CA□記号□アルファベット
□に□ニ□に□動詞□自立□一段□連用形
□対□タイ□対□名詞□サ変接続
□する□スル□する□動詞□自立□サ変・スル□基本形
□時□トキ□時□名詞□非自立□副詞可能
□系列□ケイレツ□系列□名詞□一般
□解析□カイセキ□解析□名詞□サ変接続
EOS

□一次元□イチジゲン□一次元□名詞□一般
□CA□シーエー□CA□記号□アルファベット
□に対する□ニタイスル□に対する□助詞□格助詞□連語
□時系列□ジケイレツ□時系列□名詞□一般
□解析□カイセキ□解析□名詞□サ変接続
EOS

□可視化及び乱れ強度

□可視□カシ□可視□名詞□一般
□化□カ□化□名詞□接尾□サ変接続
□及び□オヨビ□及ぶ□動詞□自立□五段・バ行□連用形
□乱れ□ミダレ□乱れ□名詞□一般
□強度□キョウド□強度□名詞□一般
EOS

□可視化□カシカ□可視化□名詞□サ変接続
□及び□オヨビ□及び□接続詞
□乱れ□ミダレ□乱れ□名詞□一般
□強度□キョウド□強度□名詞□一般
EOS
```

図 1 形態素解析結果の変化

5. 考察

日本語用語抽出において、分野コーパスから抽出された用語のうち、頻度が大きく、他の用語に入れ子となって含まれる可能性の大きいものを形態素辞書に追加してやることにより、用語抽出の性能を向上させることができることを示した。性能が向上する理由としては、新規に登録した形態素の前後で形態素解析の精度が向上すること、および、有害な効果をもたらしていた形態素の影響を軽減できることという二点を挙げることができる。一方で、形態素登録をすることにより生じる副作用もある程度生じるが、今回の結果からはそれほど重大なものではないと推定できる。

形態素を追加した後の用語抽出結果を追加前の結果と比較することにより、元の形態素辞書に用語抽出の視点から

は有害な形態素が含まれていることが判明する可能性がある。用語抽出の視点から有害な形態素については、もしそれらが同定できるなら、削除や品詞変更など、適切な対処を行うことにより、かなり大きな効果が期待できる。

一方で新規形態素周辺での用語抽出性能の向上は、確かに存在するものの、ある程度限定的なものに留まるように思われる。抽出という視点からは、用語の多くはコーパス内に複数出現するものであり、どこか一ヶ所に抽出可能なパタンが存在すれば、少なくとも文字列としては抽出できることになる。特に漢語系要素を中心とする日本語用語では、形態素解析に完全に失敗しているパタンからでも、表面上は妥当な文字列が抽出されることが多い。これは、抽出された結果に、本来分野固有の形態素とすべきものが相当程度含まれていることからも推定される結果である。

したがって形態素解析結果がより正確になったことにより、新たに抽出が可能となる用語があるとすれば、それが出現するすべての場所において解析結果の誤りが抽出を妨げるパタンになっていて、それがたまたま新しく追加した形態素による解析結果改善によって抽出が可能になる場合に限られる。出現頻度の高い用語の場合、そのすべての出現箇所でこのような状況になっていることは考えにくいから、新しく抽出可能になるものの出現頻度はかなり小さいと考えられる。つまり、新しく抽出可能になる用語は出現頻度が小さく、たまたまその周囲で生じていた解析結果が新規形態素の追加により、抽出不能なものから抽出可能なものに変化する場合に限られる。このような用語は一定程度はあるものの、それほど多いものではないと考えられるであろう。

一方で不適切な非用語を誤って抽出することに関して言えば、形態素解析が改善されたことにより、抽出されなくなるものもあれば、新しく抽出されてしまうものもある。辞書変更後の非用語の抽出に関しては、動詞連用形がかなりの場合に関与している。動詞連用形が名詞的要素として複合語の一部になれる条件は、分野によって様々に状況が異なり、また、前後の形態素との意味的関係によっても適切な場合と不適切な場合とがある。これらの条件を判断して複合語に含めるかどうかを判断するのは、現在のアルゴリズムでは不可能である。ただし、全般的に見るなら、この要因で出現／消失する非用語はそれほど多くはない。

この他、不適切な候補を新しく抽出する原因としては、形態素区切りが変化することにより、形態素解析誤りを起こしやすい形態素（接頭詞、接尾辞など）が新たに出現することによる解析誤りも可能性としては考えられる。

以上の結果を見るなら、一般的な形態素辞書に基づいて抽出された用語のうち、頻度が高く、他の複合語の部分となる割合が高いと推定されるものを形態素として登録することには、限定的ながら一定の効果があり、またそれほど重大な副作用も生じないと考えられることから、とりあえず実施してみる価値はあると言えよう。

また、今回の「機上」や「細」の例にも見られるように、形態素を追加してその効果を調べることにより、一般的辞書に含まれる不適切形態素を見つけることができる可能性もある。このような形態素が見つかった場合には、それらを修正することにより有効な改善が可能となる場合もあると考えられる。

用語抽出という観点を離れて形態素解析そのものの精度という視点からは、全体的に解析精度が向上している可能性があり、これは用語抽出以外の処理においても有用であると考えられる。この点については今後さらに検討する必要があろう。

6. まとめと今後の課題

本稿では chasen の ipadic-2.7.0 に対して情報処理学会、土木学会における用語抽出を例に、辞書に登録された形態素を追加／削除することで、用語抽出の性能が向上できることを示した。具体的には単純な形態素追加だけでも、情報処理学会、土木学会の二つの分野で、93.7%、88.9%の適合率で用語を増加させ、同時に 100%、90%の精度で非用語を削除できた。情報処理学会についてはさらに、辞書中で問題となる形態素を追加／削除することにより、92.6%の適合率で用語を新たに抽出し、100%の精度で非用語を削除できた。このように、注意深い形態素辞書の再構成は、総数は少ないにしても、高い精度で用語抽出の精度に影響を与えることが分かった。

今後はどのような複合語を形態素として追加することが有効であるかをより詳しく検討するとともに、追加形態素をさらに増やした場合に、用語抽出の性能がどのように変化するか、また、本当に有用な形態素解析精度の向上が達成できるか等について検討を進めたい。

謝辞　本研究は科学研究助成事業、基盤（C）24500303 の援助の下に行われている。

参考文献

[1] 小山照夫，竹内孔一：用語管理システムの開発，情報処理学会自然言語処理研究会報告，NL-212-2(2013).
[2] 濱田宏平，竹内孔一、小山照夫：用語間関係を一貫して登録できる用語管理システム，言語処理学会第 20 回年次大会, pp.35-38, 2014.3.18.
[3] 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，情報処理学会自然言語処理研究会報告，NL-193-13(2009).
[4] http://chasen-legacy.sourceforge.jp/
[5] http://sourceforge.jp/projects/ipadic/
[6] KANDO, N., and NOZUE, T. eds.: Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, Proc. NTCIR Workshop I, 1999.