Vol.2015-NL-220 No.6
2015/1/19

形態素解析の系統的誤りと用語抽出

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となるが、実際に専門分野の文書を既存の形態素解析器と形態素辞書を用いて解析した場合、解析精度の制約により、用語抽出性能を低下させる傾向がある。一方で解析誤りの中には、系統的な誤りと考えられるものがあり、さらにはその本来の結果がどのようなものであるのかを推定できる場合もある。これらの誤りについて解析結果を事後的に修正した上で、その結果から用語抽出を行うことにより、抽出性能を向上させることが期待できる。今回の報告では、情報処理分野の文書を解析する際に発生する系統的な誤りパタンがいくつか存在することをを明かにした上で、誤りを修正した結果から用語抽出を行うことにより、用語抽出性能が向上することを報告する。

キーワード：日本語用語抽出、形態素解析、形態素辞書、形態素解析誤り

1. はじめに

筆者らは現在用語管理システムの構築を行っている [1][2]。このシステムでは用語管理支援のために、これまでに開発してきた日本語専門文書からの用語候補抽出機能 [1][3] を組み込んでいる。用語抽出にあたっては形態素解析器と形態素辞書を利用することになるが、既存の多くの形態素解析器および形態素辞書は、一般的な日本語文書の解析を目的としており、専門文書の解析を行うためには、辞書内容や解析アルゴリズムが必ずしも最適な物となっていない可能性がある。

筆者らの現在のシステムでは、形態素解析器として chasen[4] を、また、形態素辞書として ipadic2.7.0[5] を利用している。解析誤りを改善する手段の一つとしてこれまでに、元々の用語抽出結果の中から代表的な用語と判定されるものを、仮に形態素として登録することを試みており、結果として一定程度の性能向上が実現できることを報告した [6]。

この実験の結果について、NTCIR-I[7] の学会発表データベースに含まれる情報処理分野の文書で、解析結果が新規形態素登録の前後でどのように変わるかを精査した結果、解析誤りの中には系統的に発生するものが相当数存在し、かつ、その多くについて正しい結果が推定可能であることが明かとなってきた。

本発表ではこれらの点を考慮して、元の解析器と辞書による解析結果において、系統的誤りである可能性が高く、かつ正解パタンが推定可能な場合について、当該部分を正解と考えられる形態素に書き換える修正を行った上で、修正結果から用語抽出を行うことにより、情報処理分野を対象にして用語抽出性能を改善できることを示す。

2. 形態素解析誤りの傾向と対処方法

先に報告した結果 [6] から、新規形態素追加によって情報処理分野の文書に対する解析結果がどのように変化するかを調べることにより、

- 分野に固有の基本的な形態素のいくつかについては辞書に追加することが適切である
- 元々の解析器と辞書を用いた場合、特定の形態素の直後で解析結果に系統的な誤りが発生する傾向がある
- 元の辞書に含まれる「機上」など、いくつかの形態素のについてはより慎重な扱いを必要とする

ことが明かとなった。今回はこれらの問題を修正する試みと、その修正が用語抽出結果に及ぼす影響について述べる。

2.1 分野形態素の追加

専門文書では、一般的な日本語文書には稀にしか現れない分野固有の形態素が数多く出現する傾向があるが、前回報告した通り [6]、情報処理分野では他分野と比較して分野

[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

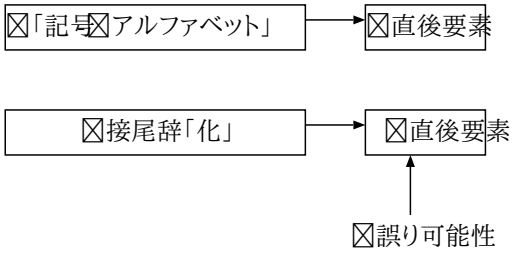

図 1 解析誤りを起こしやすい位置

固有の形態素は相対的に少ないと考えられる。今回の実験では誤り訂正の効果を評価することが主目的であるから、分野固有の「粒度」と、分野固有ではないが、元の辞書に欠けていた接頭詞「細」を新規に登録することに留めている。

2.2 特定要素直後の系統的誤り

形態素解析結果を調べてみると、図 1. に示すように形態素が「記号-アルファベット」と判定されるものおよび接尾辞「化」の直後の形態素について系統的な判定誤りが高頻度で生じていることがわかる。具体的にはこれらの要素の直後に、意味的に接続すると考えることが困難な名詞や動詞と判定された形態素が続くものがあり、これらは解析誤りの結果であると考えてよい。

そこでこのような誤りにどのような種類があるかを調べるため、それぞれのケースについて、問題となりうる形態素の直後に名詞ないし動詞が続くものについて、形態素ごとの発生頻度を数えあげた上で意味的に接続が可能であるかどうかを調べた。

結果として「記号－アルファベット」の直後で、助詞「に」が動詞「に（にる）」と判定されるものや、接尾辞「化」の直後で、助詞「から」が名詞と判定されるなど、誤って解析されたパタンで、かつ正解が容易に推定できると考えられるものが見つかった。出現頻度 5 以上のものについてこれらを調べると、誤っていて正解が推定できるものとして、「記号－アルファベット」の直後では 13 種類、接尾辞「化」の直後では 9 種類存在することが明かとなった。表 1. および表 2. にそれぞれに該当するパタンを示す。これらは助詞が誤って名詞ないしは動詞と判定されたものと考えることができるから、正しい形への置き換えを試みることとする。

2.3 注意を要する特定形態素

ipadic に登録された形態素には「機上」、「機中」などが存在するが、これらの形態素が解析誤りの原因となり、用語抽出の性能を低下させることがわかっている。これらは本来は「航空－機－上」などを意味しているが、慣用的に「航空」が陽に記述されないことが多い。結果として「機上」などが実質的に形態素として扱われることになっている。しかし、たとえば文書中に「航空機上」という文字列が

原型	誤解析結果	正解
に	にる 動詞	に 助詞
につい	につく 動詞	について 助詞
におい	におい 名詞	において 助詞
によ	にる 動詞	によって 助詞
および	およぶ 動詞	および 助詞
及び	及ぶ 動詞	及び 助詞
,	, 名詞-数	, 記号-読点
から	から 名詞	から 助詞
より	よる 動詞	より 助詞
だけ	だける 動詞	だけ 助詞
だけ	だく 動詞	だけ 助詞
かつ	かつ 名詞	かつ 助詞
も	もる 動詞	も 助詞

表 1 「記号－アルファベット」の後の誤り

原型	誤解析結果	正解
および	およぶ 動詞	および 助詞
及び	及ぶ 動詞	及び 助詞
,	, 名詞-数	, 記号-読点
から	から 名詞	から 助詞
かつ	かつ 名詞	かつ 助詞
だけ	だける 動詞	だけ 助詞
のみ	のみ 名詞	のみ 助詞
ならび	ならぶ 動詞	ならびに 助詞
より	よる 動詞	より 助詞

表 2 接尾辞「化」の後の誤り

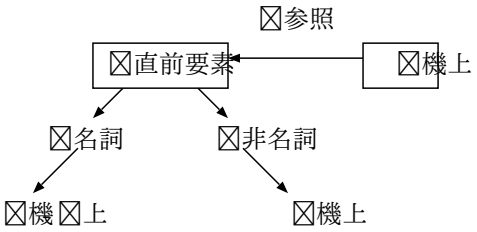

図 2 問題要素の書き換え

そのままの形で出現した場合「航空－機上」と誤って解析されてしまい、これが用語抽出の誤りにもつながっている。

このことから前回の実験 [6] ではこれらの形態素を辞書から削除する方法を試みているが、この方法では逆に「機上」そのものが出現した場合に、「機－上」と分解されることになり、必ずしも好ましい結果とは言えない。実際に問題が生じるのは、図 2. に示すように、問題形態素の直前要素が名詞系の形態素となる場合にほぼ限定できる。このパタンでは、たとえば「機上」の「機」がその直前に出現する名詞要素と先に複合していると考えられるから、これらについては直前の要素に応じて「機－上」と分解する書き換えを行うかどうか判定することを試みる。

3. 形態素解析結果の書き換え

前節で述べたように、chasen/ipadic による情報処理分野

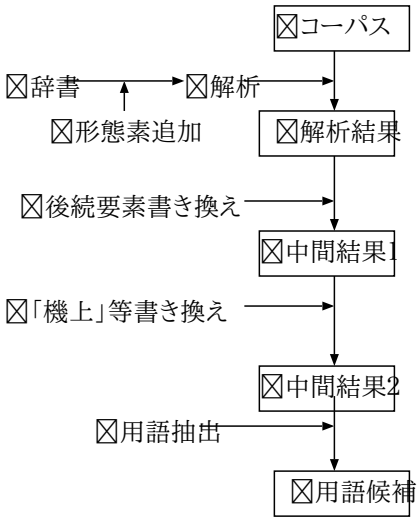

図 3 解析結果書き変えと用語抽出の概要

の文書の形態素解析では、いくつかの系統的解析誤りが生じる。これらを正しいと考えられる形態素に書き換えることにより、用語抽出性能がどのように変化するかを調べる。

まず、元々の辞書に、「粒度」および接頭詞「細」を追加した形で形態素解析を実施し、その結果に対して指定したパタンが出現する場所を特定して、当該部分を書き換える。書き変えの概要は次の通りである。

- 「記号ーアルファベット」と判定された形態素の直後要素が、予め用意した 13 通りのパタンであった場合、その要素を対応する正解要素で置き換える
- 同様に接尾辞「化」の直後の要素が用意された 9 通りのパタンであった場合、対応する正解要素で置き換える
- 「機上」、「機中」、「機内」については、その直前の形態素が名詞系の形態素かどうかを調べ、名詞系形態素の場合例えば「機－上」のように分解された形態素列で置き換える。

以上の概要を図 3. に示す。

置き換えにあたって一点注意すべき問題が存在する。元の判定結果が「に／動詞－にる」「につい／動詞－につく」、「におい／名詞」、「によ／動詞－にる」、「ならび／動詞－ならぶ」となるものでは、本来の形態素はより長い要素であると考えられる場合がある。例えば「につい」は複合的慣用助詞「についー て」の部分となっていると考えてよい。これらの場合については、本来はもう一つ後の形態素まで調べて、正しい形態素区切りと形態素分類に書き換えることが望ましい。ただし、用語抽出問題に限定して考えたときには、これらの位置に来るものが助詞であるということだけが判定できるなら、抽出結果は変化しないと考えてよい。従って今回は形態素区切りまでは修正せず、分類を書き換えることに留めている。

その他の置き換えパタンでは、単純に形態素分類を修正するだけで正しい解析結果になると考えてよい。

4. 用語抽出結果

前節で述べた形態素解析および結果の書き変えを行ったデータに基づいて、用語抽出を行った結果に対して、何も操作を加えない chasen/ipadic による解析結果から用語抽出を行った結果との差分を取ることにより、新規に 791 候補が出現すると同時に、これまで抽出されていた候補のうち 225 候補が抽出結果から消えていることがわかった。

抽出結果から消失した 225 候補を調べると、そのすべてが非用語であると判定できる。一方新たに出現したものについてはいくつか検討すべき問題がある。

新たに抽出された 791 候補の中の 542 候補は末尾の文字種別がアルファベットとなっている。このこと自体は、「記号ーアルファベット」と分類された形態素の直後が助詞という複合語区切りとなるものに変更された箇所が多数あることからある程度予想できる結果である。

問題は、これらアルファベットで終わる候補中 247 候補とほぼ半数が、例えば「アニメーションシステムＭＯＶＥ」に見られるように、システム（方式）種別に開発者の付けたシステム名称を付与した形となっていることである。これらは、実際に作成された実体を表しているという点で広義の用語に含めることも考えられる一方、長期にわたって広く参照される可能性が低いものも多いことから、用語として認める価値が低いと考えることもできる。これらの候補の用語性については、システム自体の重要性に加えて、作者の名づけたシステム名がどの程度幅広く受け入れられているかにも依存する。例えば Lisp など、当初は特定のシステム名であったものが、システムの重要性および名称が広く受け入れられることによって現在では用語として完全に確立していると考えて良いものもある。

この種の候補の用語性判定についてはさらに検討を行う必要があるが、当面は判断を保留することとして検討の対象とはしないこととする。すると、アルファベットで終わる候補の残りの 295 候補については、193 が、例えば「シングルタスクＯＳ」等の用語、102 が、例えば「データｐ」などの非用語となる。末尾がアルファベットで終わる候補をさらに調べると、末尾のアルファベット列の長さが 1 のものが 60 出現していることがわかるが、これらは例えば「データｐ」に見られるような変数参照であるか、あるいは数量単位（数接尾辞）であるものがほとんどであって、これらは非用語とみなしてよい。実際には「２Ｄ」のみが用語とみなせるものであって、残りの 59 は非用語と判定できる。

末尾がシステム名と考えられるものについて判断を保留し、末尾のアルファベット並びの長さが 1 のものを除外すると、新規出現候補数は 484 となり、そのうち 402 が用語、82 が非用語と判定される。この結果を表の形でまとめると表 3. のようになる。

	用語	非用語
出現	402	82
消失	0	225

表 3 解析結果修正による用語抽出結果の変化

	用語	非用語
出現	63	5
消失	0	92

表 4 形態素追加／修正による抽出結果の変化

以前に報告した、一部形態素を追加・削除し、用語候補の内主要なもの 30 語を形態素として登録した形で用語抽出を行った結果と、形態素辞書や解析結果を何も変更しない場合との抽出結果の差は表 4. の通りであったから、今回の結果は新規出現数も消失数も大幅に増加していると言えるが、一方で新規に出現した候補の中に非用語が含まれる割合がやや増加している。

5. 考察

情報処理分野の文書を既存の chasen/ipadic によって形態素解析を行った結果を精査すると、特定の状況の下で系統的な解析誤りが発生しており、その部分に対する正しい解析結果が推定できるものがある。

これらの誤りを修正する本来の方法は、chasen の連接コストを変更することであると考えられるが、連接コストの変更は影響する範囲が大きく、慎重な検討を必要とする。

誤り部分について正解パタンが高い確度で推定できるのであれば、むしろその部分を直接書き換えることにより、解析精度が向上した結果が得られると期待できるのであり、この書き換え後の結果に対して用語抽出手法を適用することにより、抽出性能を向上させることが可能となる。

実際に置き換え後のデータに対して用語抽出実験を行った結果、30 程度の用語を形態素として登録した結果と比較して、より多くの候補について抽出結果に変化が見られた。これは、形態素の追加では、その効果は追加された形態素の近傍に限定されるのに対して、問題パタンの書き換えでは、該当するパタンを網羅的に修正できることによると考えてよい。

書き換えられるパタンはそのほとんどが名詞ないしは動詞連用形を助詞に置き換えるものである。この結果、修正前の要素が複合語の一部となる形で誤って抽出されていた候補を排除することが可能となっている。結果として多くの非用語候補が排除されていると考えられる。

書き変えを行う部分の多くは、「記号－アルファベット」と判定された形態素の直後であることから、書き換えによって新規に抽出可能となる候補の中には、末尾がアルファベットで終わるものが数多く出現する。全般的に言って、末尾がアルファベットで終わるものには「システム種別－開発者のつけたシステム名」という形を取るものがあり、用語か非用語かの判定が難しくなる。また、それ以外にも末尾がアルファベットで終わるものには様々な種類のものが含まれており、全般的には精度を低下させる傾向がある。ただ、システム名の関連するものを保留して評価した結果では、新規に抽出された候補内での精度は 80%強となっており、用語抽出システム作成当初に評価した抽出精度（85%程度）[3] と比較してそれほど劣っているわけではない。さらには、新しく出現した非用語候補以上に、多くの非用語候補が抽出されなくなる効果もある。このことから、今回の手法は用語抽出性能を向上させる上で効果的であり、十分実用性があると考えられる。

今回は用語抽出を中心に議論を進めてきたが、系統的誤りを置き換えによって修正することは、形態素解析結果をより正しい形に近づけることになる。このことは用語抽出以外の自然言語処理に対しても新しい展望を開く可能性があると言えるであろう。

今回の結果は、chasen/ipadic を用いた形態素解析に限定されたものであり、他の形態素解析器や形態素辞書を用いた場合についてはどの程度有効なものであるかは明確ではない。ただ、形態素解析機と形態素辞書が、一般的日本語文書に対して最適化されている場合に、専門文書の形態素解析が系統的な誤りを生じ易いという傾向を持つ可能性はどのシステムにも存在しうると考えられることから、それぞれのシステムについて調査を行ってみる価値はあると考えられる。

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た。

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] 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] 小山照夫，竹内孔一：専門用語抽出における形態素辞書変更の効果，情報処理学会自然言語処理研究会報告，NL-218-4(2014).
[7] 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.