外来語の扱いを考慮した日本語専門文書からの用語抽出

小山　照夫
国立情報学研究所
t_koyama@nii.ac.jp

竹内　孔一
岡山大学工学部
koichi@cl.cs.okayama-u.ac.jp

1　はじめに

筆者らは用語管理システムの構築 [1][2] に関連して、日本語専門文書からの用語抽出に関する研究を進めている。これまでに日本語由来の用語については、形態素解析結果から、各形態素の分類と形態素間の連接関係に基づき、複数形態素のどのような連接パタンが複合語用語となりうるかを判断することが有効であることを明らかにしてきた [1][3]。一方で、日本語専門文書には、日本語由来の用語だけではなく、外来語由来の用語や、外来語を複合要素として含む複合語用語も数多く出現する。用語抽出の実用的観点からは、これらの用語についても抽出の対象とする必要がある。しかしながら外来語の場合、一般的な日本語形態素辞書には単語ないし形態素が登録されていないことが多いし、また、複合語としての用語を構成する形態素連接パタンも日本語とは異なっているため、日本語文書に出現する、外来語または外来語を複合要素として含む用語を抽出するためには、いくつかの注意を必要とする。本発表では、外来語を含む日本語文書からの用語抽出において、外来語を取り扱う一つの方法を提案する。

2　外来語を含む日本語文書からの用語抽出

専門文書からの用語抽出を行う際には、文書の形態素解析が必要となるが、このためには、形態素解析器と形態素辞書が必要である。現在日本語文書のための形態素解析器と形態素辞書として一般に利用可能なものがいくつか利用可能な形で公開されているが [4][5]、いずれも一般的な日本語文書の解析を行うために最適化されており、そのままの形で専門文書の解析を行う際にも適切なものになっているとは限らない。また、専門文書では多くの外来語が用いられるが、外来語を含む日本語文書を解析する際には、外来語部分の解析にいくつかの問題が生じる可能性がある。

```
Ｃ          記号-アルファベット
ＡＤ        名詞-一般
など        助詞-副助詞
の          助詞-連体化

実現        名詞-サ変接続
する        動詞-自立　サ変・スル　基本形
ハイ        名詞-一般
パー        名詞-一般
テキスト    名詞-一般
の          助詞-連体化
```

図 1: 外来語が分割される例

2.1　既存の日本語形態素解析器と外来語

筆者らは従来から形態素解析器及び形態素辞書として、chasen[4] 及び ipadic-2.7.0[6] を用いているが、ipadic では、外来語由来の形態素定義が決定的に不足している。しかしながら、専門文書に出現する可能性のある外来語由来の形態素のほとんどを網羅できるように辞書に追加することは多大な労力を要することが予想され、現実的とは言えない。

また、chasen による解析では形態素辞書に登録されている形態素の判定を重視しているために、一つの外来語単語を構成する文字列の部分列が辞書に定義されている場合、一つの単語が複数部分に分かれてしまうという問題がある。日本語文書中に出現する外来語の単語は、アルファベット文字列またはカタカナ文字列として出現し、同種文字が連続している限り、ほとんどの場合には全体で一つの単語として扱われるべきであり、途中で切断されることは原則としてありえない。例外的に複数単語を連続させて一つの単語のように扱う場合があるが、この場合むしろ用語としての複合語を一つの単語とみなしていると考えることができるであろう。しかし、chasen の場合、辞書に登録された文字列が出現すると、その部分だけを辞書に登録された形態素として切り出す処理を行う。結果として図

1. に示すような、本来はありえない形態素分解を行う場合がある。

2.2 日本語文中の外来語複合語

外来語の語境界とは別の問題として、日本語文中の外来語は、単語の形で出現するだけではなく、複合語の形態をとって出現する場合がある。アルファベットからなる複合語では、単語の間は空白文字かハイフンで区切られるのが一般的であるし、カタカナからなる複合語は、中黒で区切られているものが主となる。chasen などを利用する日本語形態素解析では、このような文字列並びは、単に区切り記号で区切られた独立した形態素という扱いになるが、実際には外国語文の一部であって、その中の並びに用語、または複合語要素となるものがあるかを判定する必要がある。

純粋な外国語文の場合、用語抽出を考える際には個別の単語の文法的分類と隣接する単語の接続関係から、複数の単語からなる並びが本当に複合語を形成しているかどうか確認する必要がある。このためには外来語単語の分類と、複合語の合成規則とを定義することが要求される。これは大変に労力のかかる問題となる。

幸いなことに、日本語文中にこれらのパタンが出現する場合、そのほとんどの場合に区切り記号で区切られた単語並びの全体を接合したものが、複合語用語として使用されていると考えてよい。これは、論文の中で議論の対象とされるものごと自体については、外来語や外来語由来の複合語要素を用いて指示されることはあっても、文書全体の記述構造は基本的に日本語の枠組みで構成されると考えられることによる。したがって、ほとんどの場合には、区切り記号で仕切られた外来語要素の列全体をまとめて一つの文法要素として扱っても大きな問題は生じない。図 2. に区切り記号で区切られた外来語複合語の例を示す。

3 外来語由来の要素の取扱い

前節で述べた、日本語文書に出現する外来語の特徴から、外来語ないしは外来語由来の要素を含む複合語という形をとる用語を抽出する方法として、今回次の方法を採用した。

- 対象とする文書に対して形態素解析を実施する
- 解析結果について、アルファベットまたはカタカナだけからなる形態素について、同一文字種の形態素が連続していれば一つにまとめ、分類は「未知語」とする
- アルファベットだけからなる二つの形態素が単一の空白ないしはハイフンで区切られている場合にはまとめて一つの形態素とし、分類は「未知語」とする
- カタカナだけからなる二つの形態素が中黒で区切られている場合にはまとめて一つの形態素とし、分類は「未知語」とする
- まとめられた外来語要素はそのまま用語候補とする
- 書き換えられた形態素解析結果に用語抽出アルゴリズムを適用した結果を用語候補とする

```
ｋｅｙｗｏｒｄ　ｉｎｄｅｘｉｎｇ
Ｓｏｕｎｄ　Ｓｙｓｔｅｍ
ｗｉｎｄｏｗ　ｓｙｓｔｅｍ

Ｍｕｌｔｉ－ｓｔａｇｅ
ＭＳ－ＤＯＳ
Ｓｕｐ－Ｉｎｆ　Ｍｅｔｈｏｄ

アプリケーション・プロトコル
スーパースカラ・プロセッサ
チェックポイント・ロールバック
```

図 2: 外来語複合語の例

以上の結果を図 3. に示す。

ここで結合された文字列の分類を「未知語」としているのは、前後の他の要素と複合して新しい複合語を構成する可能性を考慮していることによる。

結果として修正された形態素解析結果が得られることになるが、ここでまとめられた外来語要素は、それ自体用語としての性格を持つ可能性が高いと考えられると同時に、他の要素と結合して複合語を構成することにより、複合語用語の要素となる可能性も持つと考えられることから、外来語として結合された文字列はそれ自体を用語候補とすると同時に、修正された解析結果に対して既存の日本語用語抽出アルゴリズムを適用した結果も用語候補とする。これによって、外来語や外来語を要素として含む複合語用語抽出が可能となる。

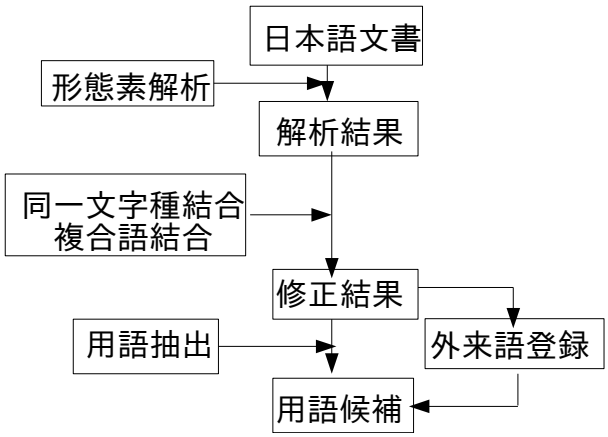

図 3: 外来語を含む文書からの用語抽出

4 用語抽出結果

前節で述べたように、外来語や外来語を構成要素として含む用語を抽出するため、元の文書を形態素解析にかけた後に、外来語の関係すると思われる箇所を修正した上で用語抽出を行った。データとしてはNTCIR-I[7] に収録された学会発表データベースから、情報処理学会データを抽出して使用している。外来語処理後の結果を、修正を行わずに用語抽出を行った結果との差分をとることによって比較した結果、新たに25,681 候補が抽出された一方、2,466 候補が抽出されなくなった。図 4. は、新規に抽出された候補のうちで用語とみなせるもの、及び消失した候補の内で非用語とみなせるものの例である。

変化の傾向を把握するため、それぞれのグループからランダムに 100 候補を選び出し、検査を行った。サンプル内のデータが用語と認められるか、非用語と考えられるかを判定した結果を表 1. に示す。

これらの結果をもう少し詳細に見ると、消失した 19 の用語について実際の文中で出現する場所の文字列を調べた結果、これらの内の 16 は、例えば

消失したもの: Ｔｒｅｅ型相互結合網
文中出現文字列: Ｆａｔ―Ｔｒｅｅ型相互結合網

に見られるように、実際の文中では外来語部分が複合語となっており、その一部 (Head) と他の要素が結合して、意味的により広い用語として抽出されていることがわかる。これは、本来の用語抽出アルゴリズムから抽出されるべきものの一部が、たまたま用語に見える形で抽出されていたということで、外来語の複合語処理が正しく行われた結果、候補から完全に消失したのではなく、本来抽出されるべき形で新しく候補と

出現した用語の例
アプリケーション・プロトコル
視覚化ユーザ・インタフェース
高速化ｔｒｙ－ｍｅ－ｅｌｓｅ命令
１Ｇｂｐｓ Ｆｉｂｒｅ Ｃｈａｎｎｅｌ
Ｍｏｖｉｎｇ Ｔａｒｇｅｔ Ｓｅａｒｃｈ

消失した非用語の例
ノイ図作成アルゴリズム
情報アクセスシステムＰＯＷＥＲ
Ｐｌａｎモデル

図 4: 抽出結果の例

	用語	非用語	計
出現	79	21	100
消失	19	81	100

表 1: 外来語処理を行う前後の比較

なったために、見掛け上消失したようにみえると言うことができる。このことを考慮するなら、実際に候補から消失した用語はそれほど多くはないと言うことができるであろう。

一方で新しく候補として出現したものの中に含まれる非用語について検討すると、

- タイプミスが原因となったもの - 6
- 外来語としてまとまったものをすべてそのまま候補としたために、人名や用語になり得ない単一の単語が出現してきたもの - 6
- abbreviation として認め難いもの - 3
- その他 - 6

となっている。今回のデータは著者の手書き原稿を業者に委託して入力したものであるが、総じて外来語を記述するアルファベットやカタカナの部分での入力ミスが目立つようである。また、人名や単語がそのまま現れるのは、今回採用した方法の弱点であるが、より厳密な用語性の判定を行うためには、外来語用語の構造や形態素に関する膨大な情報を追加する必要があり、直ちに対応することは難しそうである。

今回の方法を適用した場合、外来語の関係する用語の抽出ではいくつかの原因により、日本語要素だけからなる用語の抽出と比較して適合度を上げにくい要因がある。ただし、それでもサンプル内での適合度は80%弱であり、外来語を考慮しない場合 [3] の適合度

（およそ 85%程度）と比較して、極端に悪化しているわけではない。

5　おわりに

日本語専門文書では、特に科学技術系の文書の場合、多くの外来語用語や外来語を要素に含む用語が用いられており、その割合は増加する傾向にある。したがって用語抽出の視点からは、これらの外来語由来の用語についても、できるだけ網羅的に抽出することが要請される。

文書からの用語抽出は、文書を形態素解析した結果に基づいて行われるが、既存の日本語形態素解析器と形態素辞書に基づいた形態素解析では、解析システムが一般的な日本語文書を前提としており、また外来語についてはほとんど考慮されていないため、単に解析をおこなった結果をそのまま用いてこれらの用語を候補として抽出することは必ずしも容易ではない。

用語抽出の本来のあり方からすると、外来語についても形態素定義を整備し、外来語を考慮した複合語構成パタンを明らかにしていくことが理想と言えるが、このためには多大な労力を要することが予想され、現実的な方法とは言い難い。

日本語文書に出現する外来語の記法を考えるなら、外来語相当部分はアルファベット表記をそのまま用いるか、あるいはカタカナ表記に置き換えたものが利用されており、複合語に関しては文字種ごとに特定の区切り文字が使用されている。

また、日本語文書では論述構造は日本語のままで、外来語は原則として、議論の対象とするものごとを指示するために用いられると考えてよいから、文中に出現する外来語は、ほとんどのものが用語となっていると考えられる。さらにこれら外来語用語が複合語の要素となって新しい用語を構成していく場合もある。

したがって、外来語を含む用語抽出の一つの方法として、文中での外来語や外来語並びの境界を明らかにして、外来語を用語候補とし、さらには外来語を複合語構成要素と考えて日本語用語抽出手法を適用する方法が考えられる。

実際に NTCIR-I に収録された情報処理学会データに今回提案する手法を適用した所、外来語処理を行う前と比較して、数としておよそ 20%にあたる候補を新規に抽出することが可能となり、かつ相当数の非用語候補を排除することができた。

ただし問題として、外来語関連の用語候補について、日本語用語候補と比較してやや適合度が低い可能性がある。しかしながら、適合度の低下はそれほど顕著なものではなく、場合によっては外来語を含むものと含まないものとに分けて検討することもできることから、今回提案した手法は十分に実用的なものと考えられる。

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] https://code.google.com/p/mecab/

[6] http://sourceforge.jp/projects/ipadic/

[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.