用語集からの要素語推定の試み

小山　照夫

Extracting Elementary Terms from Dictionaries

Teruo KOYAMA

Abstract

Because many of Japanese terms in variety fields can be regarded as composite terms, decomposing terms into elementary terms seems to be helpful to understand about terms. Decomposing terms, it is neccessary to decide a basic set of elementary terms. In this paper, the author discuss about a method to extract elementary terms from analyses of dictionaries. A method selecting candidates of elementary terms referring the result of decomposition is proposed.

1. はじめに

日本語専門用語に関しては、その多くが合成語となっていることから、専門用語をより基本的な「要素語」に分解することにより、用語の分類や他の用語との関連を明らかにできるのではないかという指摘がされてきている。

これまでに筆者らは、医学分野の専門用語について、このようなアプローチからその構造を明らかにすることを試みてきた。[1]しかし、従来の筆者らの検討では、a. 要素語として何を選択するかをあらかじめ指定しておき、b. それらの要素語がどのような分類に属する語であるかもあらかじめ指定しておく、という仮定を設けていた。ここでは、どのような語を要素語として選択し、また、それらにどのような属性を付与するかについて、客観的な基準が存在していたわけではない。

今回の発表では、専門用語集の中に出現する、比較的長さの短い（副）文字列について、その頻度を手がかりとして要素語の候補を選び出した場合についての考察を述べる。また、専門分野として、複数の分野を取り上げることにより、このような手法を適用する場合の、分野依存性についても考察する。

2. 専門用語集

今回検討の対象としたのは、文部省で編集を進めている学術用語集の内から、電気工学分野および機械工学分野の用語集を選択し、また、筆者らが以前から検討を進めている医学分野の用語集として、日本内科学会編集の内科学用語集（１９８４）を選択した。

予め、これらの用語集を整理することにより、それぞれの用語集に収録されている以下の数の用語を得た。なお、これらの内には、さまざまな形での省略形や同義語等も含まれている。

電気工学　:14,525　(8,016)
機械工学　:11,314　(5,741)
内科学　　:20,000　(16,227)

() 内は漢字のみからなる用語

3. 基本語の候補

基本となる用語集が与えられたとして、第一に問題となるのは、この中から、人間の判断を挟むことなく、機械的な一定の基準で選び出すことのできる基本語の候補にどのようなものがあるかである。

筆者が当面興味があるのは、漢字のみからなる専門用語である。このため、非漢字文字を含む専門用語については文字種を、漢字、カタカナ、ひらがな、英字、記号に分類した上で、同一文字種の連続する部分に切り分け、非漢字の文字列については全て、そのままの形で基本語であると仮定している。

例えば「H型突合せ継手」は、「H／型突合／せ／継手」と分解した上で、「H」および「せ」については、無条件で基本語に組み入れている。実際には、例えば「階てい」などのように、異なる文字種の混在する表現となっている基本語も見受けられるため、このことは必ずしも完全に適切ではないが、今回の検討では、便宜上上記の取り扱いを行って

いる。実際には非漢字の基本語に加えて、漢字文字列からなる基本語の候補を考える必要がある。

漢字文字列の内、基本語となりうる候補として第一に考えられるのは、漢字のみからなる用語の内で、比較的長さの短いものである。ここで考える用語は、最初から用語集に含まれていたという意味で、（それ以上分解できるかどうかは別として）独立した語としての意味を保持していると考えられる。今回は2文字以下の長さの、元の用語集に存在する漢字のみからなる語は、無条件に全て基本語であると仮定している。

基本語の候補として次に考えられるのは、漢字・非漢字の混在する用語に含まれる、漢字のみからなる部分文字列である。先の、「H／型突合／せ／継手」では、このような文字列として、「型突合」、「継手」の二つが考えられる。ここでも、それほど長い文字列は基本語とは考えないという観点から、長さ2までのものを、基本語の候補と考えることとする。

漢字・非漢字混在用語の中の、漢字のみからなる副文字列は、必ずしもそれ自体で独立した語としての意味を保持していることが保証されるわけではない。先の「型突合」はそのひとつの例であるし、また「輻／そう」から「輻」だけを取り出して、独立した語とみなすことにも疑問が残る。ただし、今回は一文字の語については特別な評価を行っておらず、所定の手続きの後に残っていれば、それなりに意味のある基本語として取り扱っている。

基本語の候補となる可能性があると考えられる最後のものは、用語集全体を通して出現頻度の高い副文字列の内で、比較的長さが短いものである。今回は、漢字のみからなる副文字列の内で、2文字ないしは3文字からなるものを全て取り出し、用語集全体での出現頻度を数え上げ、出現頻度が5以上のものを、基本語の候補として考えることとした。

このような基本語の候補を生成するため、用語集に含まれる、漢字のみからなる長さ3文字以内の副文字列を生成し、文字列順にソートをかけた後、頻度を数え上げた。このような副文字列は、n文字からなる漢字の文字列があった場合、

長さ2の文字列：n－1種（ただしn＞1）
長さ3の文字列：n－2種（ただしn＞2）

生成されることとなり、元の文字列が長くなるほど多数の副文字列が生成されることとなる。

このような方法で生成された基本語の候補の中には、真の意味で基本語となるものももちろん含まれているが、一方で、それ自身では独立した意味を持ち得ない文字列も数多く含まれる可能性がある。

以上のものを基本語の要素と考え、いくつかの実験を行った。

4. 実験とその結果

3. で挙げた基本語となりうる候補の中から、さまざまな基準に従って文字列集合を選び出し、それがどの程度の大きさになるか、また、これを基本語として採用した場合、用語集に含まれる語の内、どれくらいの部分が、選ばれた基本語に分解できるかについて実験を行った。

一つの用語が、予め選び出された基本語の組み合わせとしてどのように分解できるか、また、可能性のある分解方法が複数個存在する場合に、どの分解結果を採用するかについては、以前の発表でも用いたアルゴリズムを用いている。[1]すなわち、

アルゴリズム1

a. 語を構成する全ての文字について、その文字を先頭に引き続く文字列のうちで辞書（基本語辞書）に登録された全てのものを選び出す。
b. 最初に文字列の先頭を探索位置として設定する。
c. 探索位置について求められた辞書登録文字列の全てについて、その文字列の終了位置を求める。
d. 終了位置が文字列全体の最後であれば、有効な分解が一つ見つかったとする。
e. 終了位置が文字列の途中であれば、その次の位置を探索位置としてc. 以降を繰り返す。

このアルゴリズムにより、任意の文字列が、もしも与えられた辞書に基づいて分解可能であれば、可能性のある全ての分解が得られることとなる。

さらに、可能性のある分解が複数通りある時には、

アルゴリズム 2
a. 分解される要素数が最も小さい物を選択する。
b. 最小要素数となる分解が複数存在する場合には、分解後の副文字列の最大長が最小になるものを選ぶ。
c. b. の条件を満たすものが複数個存在する場合には、先のアルゴリズムで最初に見つかったものを選ぶ。

ことにより、最適分解を決定している。
このアルゴリズムを適用することにより、次の実験を行った。

4. 1. 実験 1 －元々用語集に存在する基本語－
先に述べたように、元の用語集に存在する語の内、漢字のみからなる長さ 2 以下の語は、全て無条件で要素語となると考えている。このような語が、それぞれの用語集にどのくらいの数存在し、これらのみを漢字基本用語とみなして基本用語分解を行った結果を次に示す。

	用語総数	基本語数	分解結果（成功数）
電気工学	14527 (8016)	4021 (955)	5851 (2198)
機械工学	11314 (5741)	3724 (1130)	4770 (1907)
内科学	20000 (16227)	4094 (2819)	8257 (6565)

() 内は漢字のみからなる文字列に関する結果（以下同様）

この場合には、基本用語辞書に、本来基本語と認めがたい物が含まれる可能性は少ない。しかし、結果から容易に読みとれるように、特に漢字のみからなる文字列についての分解成功率はせいぜい1/3から1/4程度となっており、余りよい結果は得られていない。

4. 2. 実験 2 －候補となる全ての物－
実験 1 の文字列分解の結果があまり芳しくなかったことから、基本語候補を最大限まで増やしてみることを考える。3. で述べた、候補となりうる物を全て基本用語に組み入れた場合の結果を次に示す。

	用語総数	基本語数	分解結果（成功数）
電気工学	14527 (8016)	5619 (2553)	12581 (6324)
機械工学	11314 (5741)	4995 (2401)	9997 (4562)
内科学	20000 (16227)	6771 (5496)	17490 (13821)

実験 1 の結果と比較して、分解成功率は飛躍的に向上し、80%程度にまで達している。しかしながら一方で基本語数も実験 1 と比較するといずれも場合の約倍（漢字のみからなる文字列の場合）となっている。また、基本語を部分的に調べてみても、本来の意味で基本語とすることに問題があると思われる物が多く含まれていることがわかる。すなわち、基本語の集合としての精度に問題があると思われる。

4. 3. 実験 3 －用語分解結果の参照－
実験 1、実験 2 の結果が、ある意味で不満な物に終わったことから、基本語の集合としてより高い精度の期待できるものを選びながら、しかも分解成功率を落とさないような、基本語選出の基準がないかどうかが問題となる。
実験 2 で問題となるのは、基本語として不適切な語を排除しきれない点にある。このよ

うな語を排除する一つの方法として、このような語は、本来の基本語と比較すると、それが実際の分解に関与することが少ないであろうということである。今、3.で挙げた基本語の候補を、

a.元の用語集に存在する、長さの短い漢字文字列
b.漢字・非漢字混在語を分解した際に得られる、長さの短い漢字文字列
c.漢字文字列の内、長さ2または3の物で、用語集全体で出現頻度が5以上の物

と分類し、次の操作によって基本語候補を選び出す。

1.a.+b.を基本語として、漢字用語を分解する。
2.分解成功例（最適分解）に含まれる文字列のみを取り出し、これにc.を加えた物を新しい基本語集合とし、再度漢字用語の分解を行う。
3.2.の分解成功例に含まれる文字列のみを最終的な基本用語とする。

以上の方法で求めた基本語を用いた結果は次の通りである。

	用語総数	基本語数	分解結果（成功数）
電気工学	14527 (8016)	4987 (1921)	12197 (6318)
機械工学	11314 (5741)	4416 (1822)	9365 (4552)
内科学	20000 (16227)	5779 (4504)	17328 (13821)

ここでは、特に漢字基本語数が大幅に減少しているにも関わらず、漢字文字列についての分解成功率はほとんど変化していないことがわかる。

5.考察

4.3.では、実際に得られた基本語集合にどの程度不適切な物が含まれているかが問題となる。実際に機械工学分野について、結果として得られた漢字基本語について、基本語として適切かどうかを検討してみた結果、不適切と判定された物が68語含まれていた。

例えば4.1.と4.3.との比較では、漢字基本語が 692増加した内の68語が不適切な物であったことになる。

不適切な語が含まれることとなる原因としては、機械工学の場合、接頭辞あるいは接尾辞に相当する語を含めて切り出してしまった場合（「掘削機」－＞「削機」、「放射率」－＞「射率」など）や、2文字の基本語2語からなる文字列の内の3文字を切り出してしまった場合（「蒸気噴射」－＞「気噴射」、「吸気余裕」－＞「気余裕」など）が多く見られる。

電気工学分野については、機械工学分野とほぼ同じ傾向が認められる。一方、内科学分野では、接頭辞、接尾辞に関わる不適切な基本語候補が含まれる頻度が高いようである。これは、医学分野の用語の性格からして、接頭辞、接尾辞を数多く用いるところからきていると考えられる。

分野にも依存するが、確実性の高い基本語を初期集合として、用語分解結果を参照しながら、ある程度可能性の高い要素語集合を段階的に追加することにより、ある程度有効な基本語集合が得られることを確認した。今後はこのようにして求められた基本語に対して、その属性や意味素性を付与する問題についても考察を進めたい。

参考文献

1.小山照夫、大江和彦、日本語医学専門用語の構造解析、情報知識学会第2回研究報告会講演論文集、p17-20、1994

学術情報センター研究開発部
National Center for Science Information Systems