2003－NL－155 (1)
2003／5／26

形態素解析での効率的な複合語処理

青木 和夫　中山 章弘　松崎 剛士
日本アイ･ビー･エム株式会社
ソフトウェア開発研究所
E-mail: {aokik, nakaaki, matsuza}@jp.ibm.com

形態素解析に求められる特性は、自然言語アプリケーションによって幾分異なっており、そのひとつに複合語の扱いがある。例えば複合語「情報処理学会」を形態素解析の出力で1語とするか、「情報」「処理」「学会」と3語にするかは、アプリケーションの用途により、また求められる処理速度や精度によっても異なってくる。筆者らは形態素解析の中で効率的な複合語処理の開発を行った。本稿では辞書の見出し語が複合語であるかどうかの判定を半自動的に行う手法と、形態素解析の中で複合語処理を効果的に行う手法の有効性について述べる。

キーワード:複合語、形態素解析、遺伝的アルゴリズム

Effective Decomposition Method on Morphological Analysis

Kazuo Aoki, Akihiro Nakayama, Tsuyoshi Matsuzaki
Software Development Laboratory - Yamato (YSL)
IBM Japan, Ltd.
E-mail: {aokik, nakaaki, matsuza}@jp.ibm.com

NLP (Natural Language Processing) applications are necessary to get a little bit different characteristics of a morphological analysis, and the treatment of compound words is just one of these characteristics. For example, NLP applications want to get Japanese compound noun word “情報処理学会” (“information processing society” in English) as one word or three words (“情報”+”処理”+”学会”), and it depends on not only usage of applications but also the processing speed and accuracy. We have developed the effective processing method of compound noun words in a morphological analysis, and we will report the method of checking whether all entries in a dictionary are compound words semi-automatically and the effectiveness of compound words processing in a morphological analysis.

Key words: Compound words, Japanese morphological analysis, Genetic algorithms

1. はじめに

分かち書きされていない日本文を、品詞付きの形態素（トークン、単語）に分割する形態素解析エンジンは、検索やテキスト・マイニングや機械翻訳などの自然言語処理アプリケーションの前処理として広く使用されている。しかし、形態素解析に求められる特性はアプリケーションの用途により幾分異なり、それぞれに対応することが求められる。そのひとつに、複合語の扱いがある。例えば、「情報処理学会」を 1 語として、または「情報処理」「学会」と２語として、または「情報」「処理」「学会」と 3 語として解析してほしいかは、アプリケーションの用途により違ってくる。

日本語は他の言語に比べて、比較的容易に単語を組み合わせて新しい複合語を作る事ができるた

め、自然言語処理アプリケーションにとっては複合語の処理が必ず必要であり、その処理の仕方でアプリケーションの精度やパフォーマンスにも影響を及ぼす。またメンテナンス時の、未知語の複合語を単語辞書などへの追加の保守作業が容易にできるようにする事も大変重要である。

2. 従来技術

従来の複合語処理の殆どは、形態素解析で日本文を品詞付きの単語に分かち書きした後で行っている[1][2]。例えば、形態素解析で最小の単語に分割した後で、共起情報を用いて複合語を復元する方法がある。逆に、形態素解析で複合語の単位で分割した後で、複合語を構成する単語（単位語）に分割する方法がある。どちらの方法にするかは、辞書にどのような見出し語を登録し、文法ルールをどのように定義し、複合語をどう処理するかによって決まる。また、複合語処理を形態素解析の中で行う手法として、形態素解析の最適経路探索処理で行う手法がある[3]。これは、最適経路探索処理の途中で、それぞれの接続コストの上位からいくつかの複合語や単位語の候補を保持するやり方である。

複合語処理を形態素解析で分かち書きした後で行うと、その分余分に処理時間がかかる。また複合語などを新たに登録する際に、複合語とその単位語の関係を充分調べてから辞書に登録する必要がある。

一般に形態素解析の辞書は、国語辞典、専門用語辞典、新聞記事等のコーパス等を基にして作成されるが、辞典の見出し語をそのまま登録すると、複合語とその単位語の両方が登録される。筆者らは、このような辞書に対して、見出し語が複合語か否かを半自動的に判定し複合語に分割可能フラグを付ける手法（第３節）と、形態素解析の中でこの分割可能フラグを使用して効率的に複合語処理を行う手法（第４節）の開発を行った。

3. 複合語の半自動判定

見出し語に複合語とその単位語が混在する辞書に対して、見出し語が複合語か否かを判定し、その情報を辞書に追加する作業は、人手で行うと多大な作業量になる。かといって完全な自動化は難しい。理由は、見出し語を複合語と判定する基準が大きく判定者の主観に依存することにある。例えば、名詞と名詞が結びついて複合名詞が作られた場合は判断が易しいが、動詞と名詞が結びついた「祝い酒」などの複合名詞の場合、人によっては複合語とみなさない単語であり、一意に複合語として「祝い」「酒」と分割すると、精度の悪い形態素解析が行われたと判断される。

以下に、人間の主観に即した結果を反映できる複合語の自動判定の手法とその精度について述べる。

3.1. 概要

複合語判定の半自動化の手法の概要を**図１**に示す。既存の辞書に登録されている全見出し語を、以下に述べる計算式によって処理し、その結果、単語を「複合語単語群」「非複合語単語群」「曖昧単語群」の３つのグループに分類する。曖昧単語群は人手で判定を行い、最終的に複合語単語群と非複合語単語群に分類する。自動で判断できない単語を人手にゆだねるため、完全な自動化を図ることができないが、判断が難しい単語へ無理に判断を下すことがないため最終的な精度向上を図ることができる。

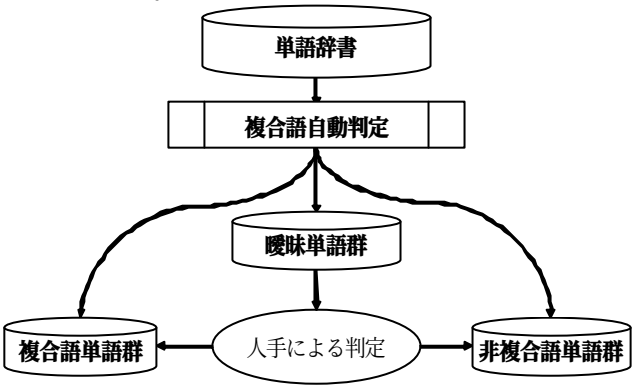


図１　概要

3.2. 複合語自動判定の詳細

「複合語自動判定」部分の詳細を図２に示す。各処理についての詳細を説明する。

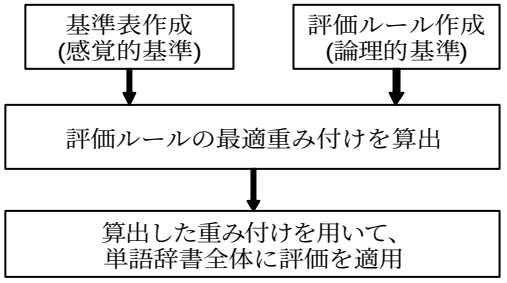
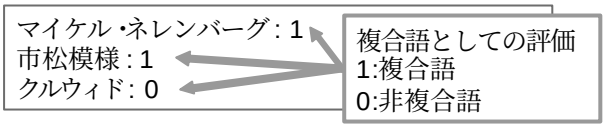


図２　複合語自動判定

3.2.1. 基準表作成

最初に、人間の主観を複合語の基準に反映させるために人間の基準表を作成する。基準表は辞書中からランダムに単語を抽出し、個々の単語を人間が複合語か否かの情報を付加して作成する。表１に例を示す。

表１　基準表の例



3.2.2. 評価ルール作成

次に、単語の特徴を評価するルールを複数用意する。このルールは単語を入力にとり、ルールに対する評価値を出力とする関数として作成する。実際に作成したルール約２０個の一部を、表２に示す。

表２　評価ルールの例

1．単語長に対し単語区切りが多かったら低スコア(四/字/熟/語)
2．単語末尾が接頭辞だったら低スコア(人事/一/新)
3．区切りの前後で文字種が変わっていたら高スコア(メリー/種)
4．単語先頭が接頭辞だったら高スコア(新/商品)
5．単語末尾が接尾辞だったら高スコア(製作/所)
:

3.2.3. 評価ルールの最適重み付けを算出

作成した基準表と評価ルールを結びつけるために、以下の数式を作成する。

$$F(w) = A_1 f_1(w) + A_2 f_2(w) + \cdots\cdots + A_n f_n(w)$$

F(w)：単語の複合語としての適合度
w：単語
$f_n(w)$：評価ルールの式
A_n：評価ルールの重み付け

このとき、w が複合語のときに高い値を、非複合語のときに低い値を F(w)がとるようなパラメータセット A_1～A_n を求めることができれば、F(w)を用いて単語の複合語としての適合度を求められる。

パラメータセットの計算は、遺伝的アルゴリズムを用いて行う[4]。この流れを図３に示す。

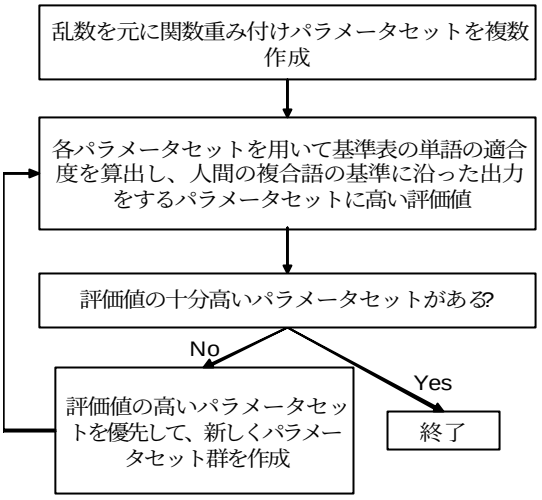


図３　パラメータセットの算出

最初に乱数を用いてパラメータセット(遺伝子)を複数作成する(初期集団)。次に、各パラメータセットを用いて、基準表の個々の単語の適合度を算出する。このとき、人間によって複合語と判断された単語に対し高い適合度をとるパラメータセットにたいし高い価値を与え、逆のものには低い評価値を与える(評価)。この処理の後、十分高い評価値を持つパラメータセットがあれば、そのパラメータセットを採用して終了する。評価値の高いものがなければ、新しくパラメータセット群を作成する。作成の方法は、評価値の高いパラメータセットを優先し、乱数で複数のパラメータセットを交互に組み合わせることによる(交叉)。また、解が局所解に陥ることがないよう、作成されたパラメータセットに対し乱数で変化を加える(突然変異)。

このようにして実際に求めたパラメータセットの一部を表３に示す。

表３　評価ルールの重み

ルール	重み(%)
1．単語長に対し単語区切りが多かったら低スコア	0.00
2．単語末尾が接頭辞だったら低スコア	0.00
3．区切りの前後で文字種が変わっていたら高スコア	0.00
4．単語先頭が接頭辞だったら高スコア	5.10
5．単語末尾が接尾辞だったら高スコア	5.46

3.3. 効果

得られたパラメータセットをもとに F(w)を構成し、辞書の全ての見出し語に対し複合語の適合

度を求める。求められた適合度の例を表４に示す。

表４　単語と適合度

高	中松山町:	2.14
	ﾏｲｹﾙ･ﾈﾚﾝﾊﾞｰｸﾞ:	1.77
	市松模様:	1.43
	ＶＰ加工:	0.99
	高砂新田:	0.79
	マル井:	0.018
	捨てぜりふ:	-0.47
	腰高:	-0.52
	アジテーター:	-0.81
低	クルウィド:	-1.56
	びわ:	-2.13

この適合度をもとにして、単語を前述の３つのグループに分類し、曖昧単語群の単語に対し人手で判断を下す。

どの程度の処理効率向上を図ることができたかを述べる。辞書約４０万語中、複合名詞を含む約３０万語の名詞類と固有名詞を処理した。３０万語を全て手作業で処理すると、１６０時間の作業が必要となる(試算値)。本手法を用いたところ、６時間(プログラムによる自動処理１０分＋手作業6時間弱)で全ての処理を終えることができた。

精度について述べる。固有名詞１１万語を処理したところ、曖昧な単語群として６０００語を得て、残りの単語は全てほぼ確実に複合語、あるいは非複合語のグループに分類でき、人手による分類作業の労力を大幅に軽減できた。一方で一般名詞は固有名詞ほどの精度が出なかった（１９万語中曖昧６万語)。これは一般名詞に造語、音便、表記揺れが多く、単語がどの部分で区切れて複合語となるかを判別できなかったためと思われる。

4. 形態素解析内での複合語処理

形態素解析は図４で示すように大きく 2 つの処理に分けられる。

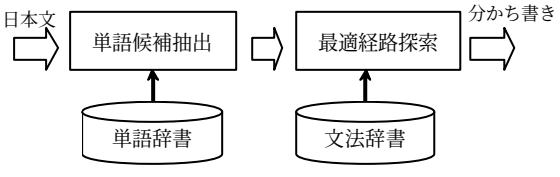


図４　形態素解析の処理

最初の単語候補抽出は、単語になり得る候補を全て見つけ出す処理で、入力されたテキストの頭から順番に、辞書をルックアップして全ての単語候補を見つけ出して、トークン候補リストに見出し語とその形態素情報を登録していく。

次の最適経路探索は、最適な経路を見つけ出す処理で、トークン候補リストの各単語の全ての可能性のある組み合わせ経路を文法ルールに則って見つけ出し、それぞれの経路の接続コストを計算して最小のコストを持つ経路を選択する（接続コスト最小法）[5]。

従来の複合語処理の殆どは、この最適経路探索の処理で選ばれた最適な経路の単語に対して行っていた。

4.1. 概要

筆者らが開発した複合語処理は、単語候補抽出の処理の中の辞書ルックアップで見つかった単語を、トークン候補リストに追加する時に行う。3 節で述べた方法で作成された単語辞書には、分割可能な複合語には分割可能フラグが立っている（１になっている)。今までは、辞書ルックアップで見つかった単語は全て候補リストに入れていた。しかし筆者らの手法は、分割可能フラグを判定して条件によってはトークン候補リストに入れない方法で複合語処理を行う。この様子を図５に示す。

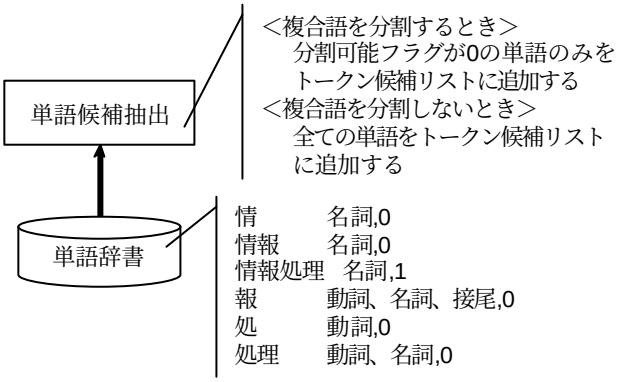


図５　複合語処理の概要

4.2. 精度について

この手法で最適解が選ばれることを説明する。

（1）複合語を分割するとき：

辞書ルックアップで見つかった全ての単語を判定して、分割可能フラグが立っている場合は候補リストに入れない。この結果、複合語の入っていない候補リストに対して、最適経路探索の処理が行われ最適解が選ばれる。この最適解には複合語は含まれないで、それを分割した単位語が含まれる。

（２）複合語を分割しないとき：
辞書ルックアップで見つかった全ての見出し語を候補リストに入れる。この結果、複合語も単位語も混在して入っている候補リストに対して、最適経路探索の処理が行われ最適解が選ばれる。この最適解には複合語を分割した単位語でなく複合語が含まれる事を説明する。最適経路探索の処理では、複合語の候補とその単位語の候補の両方が存在している。この単語までの最適経路の最小コスト値が分かっていたと仮定して、それを g(xi)とする。また、以下の品詞間の接続コストを、
“名詞”＋“名詞”= α (>0)
“名詞”＋“他の品詞”= β (>0)
とすると､この単語処理にかかる総コスト f(xi)は、
複合語の経路：f1(xi)=g(xi)+β
単位語の経路：f2(xi)=g(xi)+α(+α+…)+β
となり、明らかに f1(xi) < f2(xi)であり複合語を含む経路が選ばれる。
表５に実際に形態素解析した結果の一例を示す。左側が「複合語を分割しない」を選んだときで、右側が「複合語を分割する」を選んだときの出力結果である。

表５　形態素解析の結果

中央	名詞	中央	名詞
防災会議	名詞	防災	名詞
（	開き括弧	会議	名詞
会長	名詞	（	開き括弧
・	記号|	会長	名詞
村山富市首相	名詞	・	記号|
）	閉じ括弧	村山	名詞
専門委員会	名詞	富市	名詞
の	助詞	首相	名詞
		）	閉じ括弧
		専門	名詞
		委員	名詞
		会	接尾辞
		の	助詞

4.3. パフォーマンスについて

新聞記事から 5 万文字分の文章を抽出して（５万文字＝約２５頁分、１頁＝２千文字＝５０文字X４０行とした場合)、これを基にした 20 個のサイズが異なるデータを作成して実験した。基になった 5 万文字分のデータには、単語の総数（同じ見出し語でも品詞が違う場合は違う単語として数えている）は 11 万 2511 個で、その中で分割可能フラグが立っている複合語は 6734 個であった。総単語数に占める複合語の数は約６％である。例えば、「情報処理です。」は、「情」、「情報」、「情報処理」、「報」(5 個)、「処」(3 個)、「処理」(２個)、「理」(２個)、「です」、「。」の全部で１７個の単語があり、その中の「情報処理」だけが分割可能フラグが立っている複合語である。

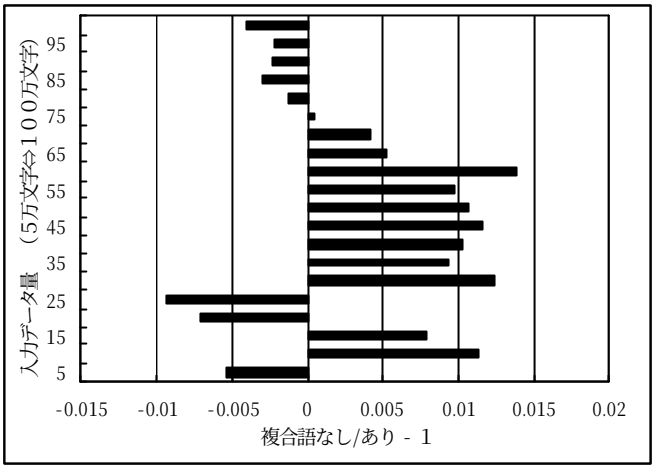


図６　複合語処理の実測結果

実験結果について述べる。大きさの異なる２０個のデータに対して、それぞれ「複合語を分割しない時の処理時間」と「分割する時の処理時間」を計測したところ、図６に示したように複合語を分割する時としない時の処理時間の差が１．５％未満であった。これは、複合語分割の処理を加えたにもかかわらず、分割を行わないときの処理時間と比較して１．５％以内の差異で処理できていることを示しており、筆者らの手法が非常に効率的であることがわかる。差異の要因として以下の要素が挙げられる。
（１）増加した時間：
単語候補抽出の処理で、全単語の分割可能フラグの判定に要する時間
（２）減少した時間：
最適経路探索の処理で、複合語の接続経路が無くなりその分の最小コスト計算に要する時間

5. まとめ

遺伝的アルゴリズムを用いて複合語の判定を行う手法については、効果的であることが実証された。既存の単語辞書の見出し語の複合語判定に費やす作業時間を大幅に減少させることができ、複合語フラグを適切に付けることができた。
この単語辞書を使用した形態素解析内での複合語処理の筆者らの手法が、複合語を分割するときと分割しないときの両方で最適解が保証され、かつパフォーマンスが複合語を分割するときと分割

しないときで殆ど同じであり、本手法が非常に有効であることが実証できた。

**　また、メンテナンス時の未知語の単語辞書への追加登録は、登録する単語が分割可能か否かを考慮するだけで良いので簡単に登録・削除が可能になり、コーパスから単語の統計情報や共起情報を収集する多大な労力が不要になった。**

参考文献

[1]㈱日立製作所、特開平9-237277、複合名詞解析方法

[2]日本電信電話㈱、特開2001-249921、複合語解析方法、装置、および複合語解析プログラムを記録した記録媒体

[3]日本IBM㈱、特開平5-46590、複数の最適解を求めるグラフ最短経路探索方法及び装置

[4]遺伝的アルゴリズムに関する情報源は多数存在する。以下のサイトは一例。
http://mikilab.doshisha.ac.jp/dia/research/pdga/index.html

[5]田中穂積 :自然言語処理-基礎と応用-、pp.2-15、電子情報通信学会発行 (㈱コロナ社販売)、平成11年3月25日