大規模 MWE データベースを組み込んだ形態素解析システム*

田　辺　利　文**
福　島　元　志***
吉　村　賢　治****
首　藤　公　昭****

A Morphological Analyzer Built in Large-Scale MWE Database

Toshifumi TANABE, Motoyuki FUKUSHIMA,
Kenji YOSHIMURA and Kosho SHUDO

The first stage of Japanese sentence analysis is to segment the input sentence on the basis of the morphological analysis. While in the ordinary linguistic framework, a word sometimes has no completed meaning but a "fragment"of some semantic constituent, the unitary expression in our model has the unitary, definite meaning by itself. Our research on this subject started in '70s by extracting manually multiword expressions as MWEs from large-scale Japanese linguistic data in the general domain. We have extracted MWEs which have at least one of the following three features; idiomaticity (semantic non-decomposability), lexical rigidity (non-separability), and statistical boundness.

In this paper, first, we present an overview of our ongoing development of Japanese MWE resources and describe our morphological analyzer which incorporates functional MWEs. Next, we introduce how to build a morphological analyzer which also incorporates conceptual MWEs, which is designed to be a base of forthcoming semantical analysis of Japanese sentences.

Key Words: Natural Language Processing, Multi-Word Expression (MWE), Morphological Analysis

1. はじめに

世界規模のインターネットの飛躍的な普及に伴い，今日我々は世界中の情報を瞬時に，かつ容易に得る事ができるようになった．さらに，企業などの社会的組織のみにとどまらず，ここ数年，ブログなど個人レベルでの情報発信も盛んになり，インターネット上の情報量も格段に増大した結果，現代は“情報爆発時代”とも呼ばれるようになった(i-explosion 情報爆発 2006)．インターネット上の情報に対して，google や yahoo，MSN などに挙げられるような検索エンジンを用いることができるが，検索結果にはユーザは探したい情報があるとは限らず，見つからない場合にはキーワードを変えて再度検索を行う必要が出てくる．このような背景から，検索精度の向上のため，“セマンティック Web”などのキーワードが登場するなど，意味を考慮したシステムの需要は急

* 平成20年１月10日受付
** 電子情報工学科
*** 工学研究科電子情報工学専攻
**** 工学研究科情報・制御システム工学専攻

激に高まっているものと思われる．

自然言語処理においても，複数の単語からなる慣用的，成句的な表現（複単語表現：Multi-Word Expression, MWE）に対処することが不可欠であることが広く認識されるようになっており（Sag et al., 2002），この事実は，意味を考慮したシステムの需要の高まりに呼応するもので，今後さらに重要性は高まるものと考えている．筆者らは日本語に関して機械処理で問題となるであろう連語候補を収集，整理する作業を従来から行っており（shudo et al., 1980, 首藤ら，1988, 首藤，1989, 安武ら，1997），連語候補による考察や予備的実験は（koyama et al., 1998, 岩瀬ら，2000, shudo et al., 2004）等に既に報告している．連語候補の収集は，確率的束縛性（要素単語相互の確率的な共起しやすさ），語彙的一体性（要素単語の間への他の単語の割り込みにくさ），熟語性（構成性原理の成り立ちにくさ）の３つの性質に注目して行っており，自立語性連語候補について，これらの性質の辞書記載について簡単に報告している（渡辺ら，2007）．

本論文の構成は以下のとおりである．第２章では拡張文節形態素解析システムの現状，次の第３章では連語候補の現状についてそれぞれ説明する．第４章では拡張文節形態素解析システムに連語を組み込む方法について説明し，最後に第５章で全体の考察と今後の課題について述べる．

2. 拡張文節形態素解析システムの現状

現在の日本語形態素解析システムの問題点の１つは，文の意味を考慮した処理を行なうための最適な分割をしていないことである．例えば，「晴れるかもしれない」と「晴れないともかぎらない」という２文において，意味が類似していることを正しく認識するためには，少なくとも「かもしれない」と「ないともかぎらない」の２つの付属語的表現，および，それらの意味を辞書に記載しておく必要がある．これまで，日本語文の構成単位を従来の形態素ではなく，意味の上から捉えなおした，拡張文節による日本語の形態素解析を行なうシステムを構築している（添島，2002）（和田，2004）．拡張文節とは，連語（単語間の結合力が比較的強いもの）を単語とみなして文節の概念を拡張したものであり，その概形は次のとおりである．

＜拡張文節＞::＝
　＜接頭語＞*＜自立語｜自立語性連語＞
　＜接尾語＞*＜付属語｜付属語性連語＞

拡張文節の詳細は（首藤ら，1979）を参照されたい．拡張文節を採用する最大のメリットは意味を正しく取り扱えることである．拡張文節形態素解析システムには，動詞辞書や名詞辞書のような，見出し語数が数万規模になる辞書を必要とせず，付属語，付属語性連語，副詞，連体詞，接続詞の必要最小限の辞書を組み込んでいる．これらの品詞の辞書には，新語の登録が動詞や名詞などに比べて少なく，日本語の枠組を規定する品詞であるということができる．実際の動作は，日本語文を chasen を用いて形態素解析を行い，次に chasen の出力結果を用いて，chasen で定義される品詞体系を，拡張文節モデルで定義されている品詞体系に変換，長単位の付属語，副詞，連体詞，接続詞に変換した後，ビタビアルゴリズムによる最小コスト法に基づいた優先処理を行い出力結果を得る．単語間の接続制約としては，品詞の細分類に相当するカテゴリコード間の接続制約，および，拡張文節内部での活用形に関する制約として活用接続制約を設けている．拡張文節形態素解析システムは，付属語においては意味を出力できる．付属語の意味は，直前の単語の活用形に応じて変わることがある．例えば付属語「そうだ」においては，「行く/そうだ」は「伝聞」，「行き/そうだ」は「様態」の意味を表わす．付属語辞書には，直前の単語の活用形によって意味が変わる場合，同一表記であっても別見出しとして記載している．chasen はユーザ側で定義した辞書を組み込むことができるが，単に組み込んだ場合には正しく意味を出力することができなくなる．また，辞書の見出しはローマ字による記載を行っている．例えば，サ行変格活用である「する」は，かな記載の場合には語幹が存在しないが，「する」を「SURU」とみなし，語幹と活用語尾の境界を _ で示し「S_URU」と記載することで語幹と活用語尾と分離できるため，活用形に応じて「URU」を適切なローマ字の活用語尾に付け替えるだけでよく，機械処理を行う上で取り扱いが容易になる．また，「飲める」「飛べる」「動ける」「消せる」のような動詞に含まれる可能の意味を抽出するため，可能の意味をもつ特殊な形態素「ERU」を設けており，この場合も「E_RU」のように語幹と活用語尾を分離できるよう記載している．「E_RU」のほかには使役を表す「SE_RU」も記載している．付属語辞書作成のコンセプトは正しく意味を抽出することを最重要課題としており，これらの理由により，現状の拡張文節形態素解析システムにおいては，付属語辞書を chasen に組み込むアプローチは採用せず，chasen の出力に基づき，辞書を用いて拡張文節化するアプローチを採用している．

これまで拡張文節形態素解析システムによる日本語文末表現の意味を抽出する実験を行っており（shudo et al., 2004）適合率は約40％程度であるが再現率は約90％を超えており，意味の網羅性は保証されているが今後の課題として曖昧さの解消があげられる．

3. 連語候補の現状

我々は広範な領域の大規模日本語データに基づき1970年代から人手によって意味上の単位と考えるべき表現の収集・整理を行ってきた(shudo et al., 1980, 首藤ら, 1988, 首藤, 1989, 安武ら, 1997, 田辺ら, 2005). 我々が収集してきた表現（連語候補）は確率的束縛性，語彙的一体性，熟語性の３つの性質のうち少なくとも１つを持つと考えられる長単位表現（単語列）と言うことが出来る．確率的束縛性とは，要素単語相互の確率的な共起しやすさを意味する．語彙的一体性とは，分離しにくさ(要素単語の間への他の単語の割り込みにくさ）を，熟語性とは，構成性原理の成り立ちにくさを意味しており，構成している単語の通常の意味から全体の意味を構成するのが難しいことを指す．収集した各表現は基本的にこれらの性質の有無や程度を表す３つ組によって性格付けされるが，これらの性質の有無の判断は収集者の内省によっている．連語候補は，自立語性連語と付属語性連語に大別することができる．

連語候補のうちの自立語性連語の＜名詞・格助詞「に」・動詞＞型の表現に対しては，コーパスと連語候補とのマッチングを行い，確率的束縛性と語彙的一体性の機械的な付与を試みている(渡辺ら, 2007)．機械的な付与結果を基に，確率的束縛性および語彙的一体性をもたないと推定された連語候補に対しては，表現の収集の観点を考えると必然的に熟語性を持つ表現となりえるが，(1)コーパスサイズが小さく，確率的束縛性または語彙的一体性をもたないことが決定的でないこと，また，(2)コーパスサイズが大規模になったとしても，確率的束縛性と語彙的一体性をもたない表現の割合はかなり小さくなること，などが考えられるため，熟語性の付与は，確率的束縛性や語彙的一体性とは独立して行うべきであると考えている．(渡辺ら, 2007)では，熟語性の有無を人手で確認，連語候補に付与している．現在，自立語性連語として，約73,000個もの連語候補の表記と文法属性の整理を終えている．今後は連語候補に対し，見出しの妥当性や，確率的束縛性，語彙的一体性，熟語性を数値化しある一定値を超える表現を連語として認定することなどを考えている．しかし第２章で述べた拡張文節形態素解析システムは，付属語性連語，副詞，連体詞，接続詞のみを考慮し，自立語性連語は取り扱っておらず，どのように自立語性連語を拡張文節形態素解析システムに組み込むかが問題であった．

4. 自立語性連語の拡張文節形態素解析システムへの追加

自立語性連語を考慮した拡張文節形態素解析システムへの増強を考える．具体的には，chasen で用いる辞書に自立語性連語を追加する．追加すべき連語としては＜名詞・格助詞・動詞＞型の表現に限定し，動詞辞書(Verb.dic)に追加することを考える(1).

Verb.dic において，動詞「する」の辞書項目は，

(品詞（動詞 自立))((見出し語（する 0))(読み スル)(発音 スル)(活用型 サ変・スル))

と記載されている．動詞における辞書項目は基本形を見出しとして記載しており，基本形以外の活用も解析できるよう，chasen は活用形展開を行って対応しているようである(2)．例えば「口にする」のような表現を Verb.dic に追加するには次のような行を記載し追加する必要がある．

(品詞（動詞 自立))((見出し語（口にする 0))(読み クチニスル)(発音 クチニスル)(活用型 サ変・スル))

つまり，見出し語の欄に連語見出し，読み，発音の欄にはカタカナ表記を入れ，活用型の欄には「口にする」に対する chasen で定義された活用型を記載する．連語データには相当品詞情報は記載されているが，chasen での活用型体系情報は入っていないため，予め，連語を chasen で解析し，出力される活用型を Verb.dic に追加する活用型とする必要がある．例えば，「口にする」を chasen で解析した場合には次のように出力される．「口にする」の末尾の単語「する」の活用型は「サ変・スル」となり，「サ変・スル」を活用型として記載する．

```
口にする
口　　クチ　口　　名詞－一般
に　　ニ　　に　　助詞－格助詞－一般
する　スル　する　動詞－自立　サ変・スル　基本形
EOS
```

この手法で連語を追加した chasen は，実質上，連語に対応した形態素解析器となる(3)．これまでの拡張文節形態素解析システムは，入力文に対し，chasen で形態素解析された出力を基に，拡張文節化して出力していた．連語を chasen に組み込まず，分かち書きシステムに組み込む手法も考えられるが，連語データは見出し数が数万と大量であり処理速度が大幅に低下することが予想されるため，chasen の最適化された高速アルゴリズムを直接用いることができれば処理速度の低下を抑えることができることが考えられることから，自立語性連語は chasen の辞書に組み込むアプローチを考えている(4)．そのため，chasen が連語データ，つまり自立語性連語

を取り扱うことができれば，必然的に，自立語性連語および付属語性連語に対応した拡張文節形態素解析システムが構築されることになり，大規模 MWE データベースを組み込んだ拡張文節形態素解析システムとなることを示唆している．

5. おわりに

本論文では，拡張文節形態素解析システムの現状，および連語候補の現状について述べ，次に付属語性連語が組み込んである既存の拡張文節形態素解析システムに対し，chasen に自立語性連語を組み込むことで，大規模 MWE データベースを取り込んだ拡張文節形態素解析システムの構築が可能であることを述べた．これにより，付属語の意味を出力させつつ，慣用句などの自立語性連語を捉える形態素解析システムの実現が現実味を帯びたといえる．実際の日本語文では自立語性連語と付属語性連語を共に含むケースも存在しえるため，意味を正しくとらえるためには必要不可欠なシステムとなる．また，自立語性連語も付属語性連語も，文字通りの意味を取るか熟語的な意味をとるかの判断基準の導入も必要になる．応用としては，熟語性を有する表現は別の一般的な表現へと言い換えることで，既存の言語処理システムの出力の質の向上が期待できる．また，出力を電子的データとして大量に蓄積することができれば，SVM などの機械処理を行う上でもさらなる精度向上が期待される．

今後の課題として，まずは自立語性連語を組み込んだ chasen の評価が挙げられる．chasen の動詞辞書には「気に入る」などのような自立語性連語が既に組み込まれている．組み込みに際し，単語コストの設定も必要になるが，例えば，連語表現を優先すべく，連語には一律にコスト値として０を付与することが考えられるが，「３リットルの油を売る」などのような，文字通りの解釈の余地を残す必要がある．そのため，組み込みの際の単語コストをどう設定するかが出力性能の向上のために考察すべき事項になると考えている．また，大規模 MWE データベースを取り込んだ拡張文節形態素解析システムは，最初に自立語性連語を組み込み，後のステップで付属語性連語を組み込むことになるため，例えば，文中に自立語性連語と付属語性連語がオーバラップして現れた場合には，自立語性連語を優先した処理になることが考えられるため，自立語と付属語の重要度のウェイトをどのように決定するかも問題となりうる．

謝辞

本研究に対して，chasen の開発者の方々に対して心より感謝する．

注

(1) chasen には unix 版と windows 版が共にリリースされており，本論文での chasen は windows 版について言及している．

(2) chasen は，動詞が基本形以外の活用をした場合に対しても正しく読みを出力するため，見出し語の活用により，読みの欄も自動的に活用させている．読みの欄の自動的な活用は，単語の活用型を観点にし，活用形に応じて，適切な活用語尾を推定，読みの欄を書き換えたものを出力しているようである．読みの欄と活用型の整合が取れていない場合には，辞書のコンパイル時に適切な活用語尾が与えられず，その結果コンパイル時にエラーが発生する．但し，読みの欄の活用語尾より前方の文字は活用による影響がないようである．この事実から，例えば，読みの欄の活用語尾より前方の文字部分として言い換え情報などを記載しておけば，chasen で言い換えが可能になることも考えられる．

(3) 辞書を更新しただけでは chasen の実行結果には反映されない．chasen の実行結果に反映させるためには，辞書のコンパイルを行う必要がある．辞書のコンパイルには，辞書フォルダ(dic)内で make を行う．それにより連語データが実行結果に反映されることになる．

(4) 動詞辞書 Verb.dic のサイズを10倍に増やして辞書のコンパイルを行い，実行させてもエラーは発生しなかった．また実行時間は１秒以内であり処理速度の低下は全く感じられなかった．そのため，10万オーダの大量の見出しを有する辞書を組み込んでも全く問題はないと結論づけられる．

参　考　文　献

chasen　http://chasen.naist.jp/hiki/ChaSen

google　http://www.google.co.jp/

i-explosion 情報爆発．2006．情報爆発時代に向けた新しい IT 基盤技術の研究（文部科学省科学研究費補助金「特定領域研究」），http://i-explosion.ex.nii.ac.jp/i-explosion/index.php

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. The Proc. of the 3rd CICLING: pp.1-15.

岩瀬修，森元逞，首藤公昭．2000．連語を組み込んだ統計言語モデル．電子情報通信学会第34回音声言語情報処理研究会: SP2000-113: pp.109-114.

Kosho Shudo, Toshifumi Tanabe, Masahito Takahashi and Kenji Yoshimura. 2004. MWEs as

Non-propositional Content Indicators. The Proc. of the ACL2004 Workshop on Multiword Expressions: Integrating Processing: pp.32-39.

Kosho Shudo, Toshiko Narahara and Sho Yoshida. 1980. Morphological Aspect of Japanese Language Processing. The Proc. of the 8th COLING: pp.1-8.

MSN　http://jp.msn.com/

首藤公昭，楢原斗志子，吉田将．1979．日本語の機械処理のための文節構造モデル．電子通信学会論文誌，J 62-D，12，pp.872-879.

首藤公昭．1989．日本語における固定的複合表現．文部省科学研究費補助金特定研究（Ⅰ），課題番号 63101005.

首藤公昭，吉村賢治，武内美津乃，津田健蔵．1988．日本語の慣用的表現について－語の非標準的用法からのアプローチ－　自然言語処理研究会 NL-66-1: pp.1-7.

添島創．2002．日本語の拡張文節分かち書きに関する研究．平成14年度福岡大学大学院修士論文

田辺利文，高橋雅仁，吉村賢治，首藤公昭．2005．日本語の複単語表現データ，言語処理学会第11回年次大会.

和田智樹．2004．機能性 MWU を取り入れた日本語形態素解析．平成16年度福岡大学大学院修士論文

渡辺耕平，田辺利文，小山泰男，吉村賢治，首藤公昭．2007．日本語連語データの整備，言語処理学会第13回年次大会.

yahoo　http://www.yahoo.co.jp/

安武満佐子，小山泰男，吉村賢治，首藤公昭．1997．固定的共起表現とその変化形．言語処理学会第３回年次大会発表論文集: pp449-452.

Yasuo Koyama, Masako Yasutake, Kenji Yoshimura and Kosho Shudo. 1998. Large-Scale Collocation Data and Their Application to Japanese Word Processor Technology. The Proc. of the 17th COLING: pp.694-698.