日本語処理を容易にする活用形態素の提案

Functional Morpheme for Easier Japanese Processing

長岡技術科学大学准教授 **山本　和英**

豊橋技術科学大学大学院工学研究科博士後期課程システム情報工学専攻修了。博士（工学）。1996 年～2005 年（株）国際電気通信基礎技術研究所（ATR）、2002 年～現在まで長岡技術科学大学、現在准教授。自然言語処理、及び日本語教育（ツール作成）の研究に従事。

yamamoto@jnlp.org

1 はじめに

本稿では活用形態素という概念について紹介し、議論する。活用形態素とは我々が提案している概念で、用言から活用部分のみを切り離して独立した形態素とするシンプルな考え方である。我々は現在開発を行っている日本語解析器「雪だるま」[Yamamoto15] に対してこの概念を導入した。本稿では、まず 2 節において雪だるまについて概略を紹介する。3 節では、我々が導入する活用形態素という概念に関連する予備実験として、動詞の活用形がどの程度復元できるのかという実験を行ったのでこれについて述べる。4 節では、我々の提案する活用形態素について説明を行い、特徴を述べる。5 節では活用形態素のまとめを行う。

2 日本語解析器「雪だるま」

雪だるまプロジェクトは 2015 年 4 月から開始したプロジェクトで、日本語解析環境を構築することを目指している。いわゆる形態素解析器や構文解析器を構築することが目標であるが、ツール（モデル）の提案が主眼ではなく、単語体系や品詞体系などの日本語の知識表現そのものの再検討、及び再構築に大きな特徴がある。

2.1 UniDic 辞書の採用

雪だるまは UniDic 辞書を採用し、UniDic の単語体系を言語解析の最小単位、すなわち「形態素」としている。UniDic の単語体系は解析誤りを起こさないようにあえて細かな単位として定義されているため、一般にこの出力を言語処理の対象として利用する場合は単語分割単位が細かすぎて使いにくい。一方、我々のように最初から形態素結合などの後処理を前提にするのであれば、UniDic 体系での出力結果をさらに分割する可能性を考慮しなくてよいため、すなわち形態素の結合のみに集中できるため非常に使いやすい。

我々は UniDic の単語項目すべてに対してアルファベット 5 桁の ID を独自に付与し、後に述べる表記統制（表記ゆれ解消）、形態素結合などの情報をすべてデータベース上で ID 管理している。これによって辞書管理作業が効率化されたと同時に、ほとんどの処理も ID の照合や変換のみの簡単な処理で可能となった。

UniDic 辞書での解析には MeCab を用いている。MeCab-UniDic をそのままの形で用いることにより最高の解析精度が期待でき、我々はその後処理に注力することが可能となる。独自の単語体系を持った単語解析器を構築したいのであれば、すべてを再設計、再構築するのでなく、我々のように MeCab-UniDic を用いるのが最も現実的である。

2.2 曖昧さの保持と精度優先の原則

単語解析のレベルにおいては、下記のような曖昧さが問題になり、これが単語解析の精度を劣化させている。

- 読みの曖昧さ（例：辛い：「つらい」と「からい」）
- 名詞再分類の曖昧さ（例：高三：普通名詞と人名）
- その他の品詞曖昧さ（例：今日：普通名詞と副詞）

これらに対し、既存の形態素解析器は概ね決定的に出力する（つまり曖昧性を解消する）が、雪だるまでは無理な曖昧性解消を行わないという方針を採用している。精度 100% とまではいかなくとも、高精度での曖昧性解消が期待できない場合は両方の可能性を併記したほうが後続の処理にとって都合がよいからである。以前に Project Next NLP で議論になった、N-best 出力を行うという考え方と方向性は同一だが、組み合わせ的に増加することを考えると我々の提案する曖昧性保持のほうが都合がよい。また、JUMAN で曖昧性を保持して KNP で解消しているのと同じ考え方である。

そもそも、既存の形態素解析器で「つらい」と「からい」の曖昧性が解消できるとはとても思えない。それにも関わらず辞書中には二つの単語があり、解析時において決定的にどちらかを出力している。これは我々に言わせれば、無理な曖昧性解消を形態素解析器にやらせるのが問題で、単語表記や品詞列だけで解決できない問題は最初から形態素解析器に解かせるべきではない。我々は、この両者の単語は ID を統一し、両者の読みを持たせている[1]。このどちらの読みが正しいかは、後続の意味処理で解決できた場合に曖昧さ解消すればよい。

さらにこの考え方を進めると、十分な精度が出ない処理はそもそも行うべきではないという考えに至る。これを精度優先の原則と呼んでいる。雪だるまではすべてこの原則に従い、十分な精度を確認した処理のみを実装している。これはツールの開発にとっては重要な観点であり、すなわちツールが行った処理のほとんどは信頼できるということを意味する。無理に処理対象を広げて精度が出せないツールは使い道がない。

2.3　単語の規格化＝資源とツールの統合管理

雪だるまでは、言語資源とツールを統合管理している。すなわち、統一した単語体系、品詞体系に基づいてすべての言語資源を作成している。自然言語処理の研究開発において最大の障害はツールや資源において単語体系が統一されていないことである。

これに対し、少なくとも我々で作成する言語資源は統一された単語体系で揃えていく。既存の言語資源はすべて、我々で雪だるま体系に変換した上で管理する。これを我々は単語の規格化と呼んでいる。ねじや釘といった工業部品のように日本語の単語を規格化することによって各ツールや資源を部品化することが可能となり、ツール間や資源間の相乗効果が期待できる。

3　活用に関する議論

3.1　活用形不要仮説

次に、我々の提案する活用形態素の前に、活用と活用形そのものについて議論する。言うまでもなく、活用とは用言（動詞や形容詞）において後続する単語などによって用言の語尾が変化する現象のことで、変化した形態のことを活用形と呼ぶ。ここで、一見すると活用形は後続の単語に強く依存しており、後続する単語が決まりさえすれば活用形は容易に決めることができるように感じる。例えば、「行く」という五段活用をする動詞であれば、後続の単語によって次のように活用するのは周知の通りである。これはつまり、（下記の例においては）用言に後続する単語を見るだけで容易に活用変化できるということを意味する。

- 行か＋ない
- 行き＋ます
- 行く＋。
- 行く＋こと
- 行け＋ば
- 行こ＋う
- 行っ＋た

以上から、我々はある一つの仮説を立てた。それは、活用変化は後続の単語のために変化した形態であって、活用自身は何も情報を持たず、よって情報処理の観点からは必要ないのではないかという仮説である。言い換えれば、用言の活用をすべて削除しても容易に復元できるのではないか、ということである。もし活用が復元可能な情報であれば、削除して日本語処理したほうが処理の複雑さが軽減され、必要な時にいつでも復元すればよいということになる。一方で、この仮説が正しくなく、活用形の復元が難しいのであれば、活用形自身が何らかの情報を持つことを意味する。

1　正確には、MeCab-UniDic で解析した段階では両者は別の単語として認識され、どちらかが出力されるが、表記統制によって両単語の ID を統一している。

3.2　活用形復元実験

以上の仮説を検証するため、削減した活用形がどの程度の精度で復元可能かという検証実験を行った。

復元は、サポートベクトルマシン（SVM）によって行い、ツールとして YamCha を使用した。機械学習に使用した素性は、活用語の前後 2 単語の原形と品詞、活用語の前 2 単語の活用形である。訓練データには BCCWJ の LB レジスタ（書籍データ）を形態素解析したものを使用し、1,000、5,000、10,000、50,000、100,000 事例をそれぞれ訓練データとした。評価用データには LB レジスタから 1,000 事例を形態素解析したものを使用した。ただし評価データ中には訓練データを含まない。

実験結果を表 1 に示す。実験の結果、訓練規模が最大の 10 万トークンで、全体精度 96.4% という精度が得られた。活用形別では、それぞれの活用形の出現数に大きな偏りがあるが、主に出現する終止形、連体形、連用形について見ると、いずれも 3%～4%程度の誤りがあることが分かる。

以上より、当初予想していたように活用形は前後の文脈から復元できるという仮説は概ね正しいが、その一方で完全に復元できる訳ではないことも確認した。3.1 節で議論したように、もし活用形が直後の単語（機能語）にのみ依存しているのであれば 100% の精度が得られるはずだが実験結果はそうなっていない。このことから、すべての活用形が直後の単語のみに依存している訳ではない、すなわち活用形自身が何らかの意味や文法機能を持っているということを結果は示唆している。この結果から、我々は活用形を削除するのでなく、別の形で効率的に処理することを考え、次節で紹介する活用形態素を提案する。

表 1　活用形復元実験の精度

訓練規模	1K	5K	10K	50K	100K
意志推量形	0.103	0.385	0.410	0.410	0.513
仮定形	0.534	0.500	0.638	0.793	0.810
語幹	0.571	0.786	0.929	1.000	1.000
終止形	0.927	0.957	0.961	0.949	0.958
未然形	0.989	0.989	0.989	0.989	0.989
命令形	0.571	0.714	0.714	0.714	0.714
連体形	0.960	0.974	0.981	0.977	0.976
連用形	0.945	0.952	0.960	0.969	0.972
ク語法	1.000	1.000	1.000	1.000	1.000
全体	0.932	0.949	0.958	0.959	0.964

4　活用形態素の提案

今年度の新たな試みとして、単語解析の結果として活用形態素という概念を導入した [山本 16]。従来のすべての形態素解析器では、活用という概念は用言に付属される概念であって、用言の属性そのものである。例えば「動けば」という表現を単語分割すると下記のような出力結果が得られる（カッコ内は品詞、／は単語の区切り、【…】は活用形態素を表す。以降も同様）。

動けば＝動け（動詞仮定形）／ば（助詞）

これに対し、活用形態素とは用言と活用を完全に分離して、活用を単独で別の形態素と見なすという考え方である。同様に、「動けば」の例で言えば下記のようになる。

動けば＝動く（動詞）／【仮定形】／ば（助詞）

一見すると 2 単語が 3 単語になっただけで両者はほとんど同じに見える。実際に、全体として得られる情報量は増えても減ってもおらず、全く同一である。ただし、以降の処理のしやすさという観点では両者は同じではなく、我々は後者のほうが便利なのではないかと考えている。以下では、このように考えた理由について説明する。

4.1　用言への影響

まず、用言そのものに影響がある。活用を含んだ用言の場合、表面上は別の単語となってしまうので、例えば「動き」と「動く」は別の単語として様々な処理を行わざるを得ない。一例として単語の n-gram[2] を作成する際も、両者は別扱いなので同じ「動く」という 1 単語であっても活用を持たせることによって「動か」「動き」「動く」「動け」「動こ」「動い」の 6 単語になってしまう。これは明らかに誤った処理で、少なくとも直前の単語との連接を見たいだけならば「動く」1 語で十分なはずだ。

一方、単語解析を行うことでこれらの単語がすべて「動く」の活用だという情報は得られている。よって現状の

2　私の見る限り、単に「単語 n-gram」と言っても表層形で作成する場合と基本形で作成する場合の 2 種類の「流派」があるようだ。Web 日本語 N グラム第 1 版では表層形を採用している。

２単語のままでも全く問題ないのではないか、という見方もできる。しかし、基本形で単語 n-gram を作成すると、今度は n-gram の情報から活用情報が一切削除されてしまい、違う意味で問題がある。活用情報は前後の単語を文脈として用いることで完全かつ簡単に復元できるように思われるかもしれないが、３節で述べたように実際にはそうではない。

以上は単語 n-gram を作成する状況を例に説明した。以上をまとめると、従来の活用情報を含めた表記（＝表層形）と活用情報を含めない表記（＝基本形）のどちらであっても異なった問題がある。これに対し、我々の提案する活用形態素という概念の導入によってこれらの問題を一挙に解決する。以上の議論から、少なくとも単語 n-gram を作成するのであれば活用形態素を導入したほうが優位であることは明らかである。用言と活用形態素を分離することによる情報の損失は一切なく、後続の処理で必要に応じて用言を活用させることは非常に容易である。また、用言は活用形情報を持っていないので、直前の単語との n-gram や共起情報などが、従来よりも意味を持ったものになる。

4.2　後続単語との関係性

活用形の多くは後続する単語に強く依存している。例えば、動詞（例：動く）に「ば」を接続する時は仮定形（動けば）となるし、「う」を伴うときは自動的に連用形（動こう）となる。これの意味するところは、活用情報の多くは後続の単語との関係性が非常に強いということである。すなわち、従来のような形態素分割では、

動け　＋　ば

このような二つの要素（チャンク）としてしか捉えることができなかったが、活用形態素を導入することによって、

動く　＋　（【仮定形】／ば）

このように（【仮定形】／ば）を一つの要素として捉えることが容易になる。

このような考え方はすでに日本語教育で利用されている。例えば文法項目において下記のような表記を目にする[3]。

- 【テ形】／てください
- 【マス形】／たいです
- 【辞書形】／ところです

以上のような記述における【テ形】【マス形】【辞書形】は、ちょうどここで提案している活用形態素そのものであり、少なくとも日本語教育で取り扱うような文法項目を記述する上では非常に都合がよい。これも、従来の形態素記述で取り扱うことが不可能ではないが、活用形態素を導入するほうが処理がはるかに容易である。

4.3　構文的境界

前節とは逆に、後続する単語を見ただけでは活用を復元できない（＝複数の可能性がある）場合が存在する。最も典型的なのは、下記のような例である。

- 走る時間＝走る＋【連体形】＋時間
- 走り時間＝走る＋【連用中止形】＋時間[4]

この例では、活用形の直前直後の単語はどちらも同じ「走る」「時間」である。従って、少なくとも直前と直後の単語や品詞を見ただけでは活用形を復元できない[5]。しかし、この事実は逆に言えば、このような場合の活用形は直後の単語に依存して変化しているのでないので、何らかの（構文的に）重要な情報なのではないかと我々は推察した。

従って、このような復元が難しい活用形態素は構文解析の重要な情報となり得るのではないかと我々は予言する。具体的には、動詞連体形は直後の名詞（句）と結合してより大きな名詞句を形成するが、連用中止形の場合はここが文節境界となる可能性がある。

従来のように、活用形が用言に埋め込まれている場合

3　この例に限り、活用形の名称は実際に日本語教育で使われている名称で記述している。

4　例えば「速く走り時間に間に合わせた」という文の「走り時間」という部分。

5　さらに文脈を広げても、例えば単語と品詞を素性とするSVM 等の方法ではやや性能向上する程度だろう。すなわち本文中で議論するように文脈が長いか短いかという問題ではないと思う。

には、この連用中止形を活用して構文解析を行うことが（不可能ではないにせよ）容易ではなかった。すなわち、規則やパタンを用いる場合は記述のしにくさ、統計情報を用いる場合は統計の取りにくさが明らかにあって、これが構文解析の精度を劣化させていた可能性がある。したがって、明示的に連用中止形を表出させることによって、文節境界の解析や構文解析の精度を向上させることができる可能性がある。

以上は単に仮説の段階であってまだ検証は行っていない。活用形態素の導入が構文解析精度にどのように影響するかは興味深い課題であるので、今後取り組んでいきたいと考えている。

4.4 関連研究

次に、活用形態素という概念に関連した研究を紹介する。

「活用形態素」という概念は我々が今回初めて提案するもので、類似する研究は存在しない。しかし、自然言語処理において単語そのものではない何らかの情報を独立した形態素と認定するという発想は古瀬ら [古瀬 99] ですでに提案されている。古瀬らは、「構成素境界」という概念を導入することで音声翻訳（話し言葉に対する機械翻訳）において有効性を議論、検証している。

古瀬らが導入している構成素境界は、単語解析の結果内容語が連続する状況において、前後の内容語の品詞情報を連結したもの（古瀬らはこれを品詞バイグラムマーカと呼んでいる）を挿入している。例えば、「こちら観光局」という入力に対しては下記のようになる。例では、「こちら」が代名詞、「観光局」が名詞で内容語が連続しているので、【代名詞＋名詞】がバイグラムマーカとして挿入される。

こちら観光局＝こちら／【代名詞＋名詞】／観光局

このような構成素境界を挿入することによって、これをキーにして構文解析を行うことを容易にする。すなわち、構成素境界が【名詞＋名詞】であれば（多くの場合）複合名詞処理すればいいし、【代名詞＋名詞】ではあれば複合名詞には決してならず、代名詞を主語とみなして処理するのが適当であろう。これはちょうど、【代名詞＋名詞】という構成素境界が係助詞「は」と全く同様の扱いをしていることを意味し、この意味では「は」に相当する形態素を補完していると解釈することもできる[6]。話し言葉においてはこのような助詞の欠落が多いので、このような処理を行うことで有効な構文解析を実現している。

以上のように、我々の活用形態素と古瀬らの構成素境界とは導入の背景も挿入される情報も異なっているが、構文解析を容易にするためという目的は部分的に共有している。さらにこの発想を一般化すると、ここで取り扱っているような何らかの構文的情報だけでなく、意味的な情報も疑似的な形態素として文中に挿入することで自然言語処理の各種タスクの精度向上が見込める場合がありそうだ。これについても、今後検討を進めていきたい。

5 まとめ

以上の議論をまとめると、活用形態素の導入によって下記に示す 3 点の特長を生む。

1. 活用情報を損失することなく用言をまとめあげることができ、単語 n-gram や共起表現などの処理を行う上でより有用な情報を獲得できる
2. 日本語教育で行われているように、後続の単語と共に文法項目を記述することが容易にできる
3. 活用形によって構文的境界を表現している場合に、構文解析を容易にする

一方、活用形態素の導入による問題点が全くない訳ではない。現在憂慮している最大の問題点は、活用形態素を用言から独立して表出させることによって文中の単語数が（表面上）増えることによって何らかの弊害をもたらすのではないかという点である。同じ表層表記であっても用言を含む場合はそれだけ単語 n-gram が長くなって n-gram 情報が希薄化するのではないかといった懸念や、例えば統計的機械翻訳において形式的に入力が長くなることによって翻訳精度が逆に劣化するのではないかといった懸念がある。これらの様々なタスクへの影響

6 さらに古瀬ら [古瀬 99] では用例利用型 (example-based) の構文解析を行っているのでパタン記述によって各構成素境界も細かな解析の制御が可能であるが、ここでは詳細を省略する。

については まだ調査が始まったばかりであるので、順次検証していく。

参考文献

[古瀬 99] 古瀬 蔵、山本 和英、山田 節夫. 構成素境界解析を用いた多言語話し言葉翻訳. 自然言語処理、Vol.6, No.5, pp.63-91, 言語処理学会 (1999.7)

[Yamamoto15] Kazuhide Yamamoto, Yuki Miyanishi, Kanji Takahashi, Yoshiki Inomata, Yuki Mikami and Yuta Sudo. What We Need is Word, Not Morpheme; Constructing Word Analyzer for Japanese. Proceedings of the International Conference on Asian Language Processing (IALP 2015), pp.49-52 (2015.10)

[山本 16] 山本 和英、高橋 寛治、椙澤 優希、西山 浩気. 日本語解析システム「雪だるま」第２報～進捗報告と活用形態素の導入～. 電子情報通信学会 テキストマイニングシンポジウム、信学技報、Vol.116, No.213, pp.63-68 (2016.9)