

形容詞、複合動詞を扱う述語項構造解析システム

岸邊 賢太 横山 晶一 井上 雅史
(山形大学工学部)

1. 研究概要

自然言語の解析は、「形態素解析」、「構文解析」、「意味解析」に大別できる。その中でも「形態素解析」、「構文解析」は、多くの研究がなされており、ツールも公開されている。

しかし、文字変換、情報検索の高度化に伴って、さらに意味の部分にまで踏み込んだ解析が必要とされてきている。その分野を「意味解析」と呼ぶ。

「意味解析」は、「形態素解析」や「構文解析」に比べて問題があり、研究が進んでいないのが現状である。

本研究では、その問題の一つである「語義の曖昧性」について考える。

そこで、語義の曖昧性を解消する手法の一つとして述語項構造が挙げられる[1,2]。述語項構造とは、述語と述語の表す意味を補う働きをする項を同定するものであり、これによって文の意味の骨格を表すことができる。また、格フレーム、結合価といったこれまでの手法を包括する関係にある。そのため、述語項構造の項の同定に格フレームと結合価を用いる。文の構造を明確にする構文解析として係り受け解析、格フレームを用いた格解析を行う。それぞれで述語項構造を作成し結合価のデータを用いて述語項構造の誤りを修正できると考えられる。

本研究では、先行研究[3]のシステムの述語項構造の精度を向上させ、より広範囲の述語を対象とするシステムの構築を行う。

一つ目の改良として、動詞における述語項構造で精度低下の一因であった複合動詞を正しく出力させる。そして二つ目に、新たな述語の対象として形容詞の述語出力を行えるシステムへの拡張を図る。

評価結果は、従来システムを上回るものであったが、解決すべき問題を残している。

2. 研究背景

本研究では、意味解析における語義の曖昧性の解消を目的として、述語項構造を用いる。先行研究の動詞に限った述語項構造を解析するシステムを次の二点について改良、拡張を行う。

- ・複合動詞であるが、誤って二つの動詞として出力されているものを修正する処理
- ・形容詞を出力する解析システムの構築

2.1 述語項構造

述語項構造は、意味役割付与とも呼ばれ、述語と項(述語と格関係にある名詞句の関係)を同定するものである[3,4]。動詞・形容詞などの「述語」は、文の中心で動作・状態を表す要素である。そして「項」(名詞+格助詞)は、述語が表す事態に関係する人、ものを表現する要素である。述語項構造を用いることで文中の各述語が表す意味を意味を補う働きをする項を同定し、文の意味の骨格を表すことが可能となる。

図 1 は、「私が電車に乗って帰る。」という文の述語項構造の例である。この例文には、「乗る」と「帰る」の二つの述語がある。その述語に対して、「私」と「電車」の関係する述語を矢印で表記したものである。「乗る」という述語は、ガ格に「私」、二格に「電車」が相当する。一方、「帰る」という述語は、ガ格に「私」が相当する。

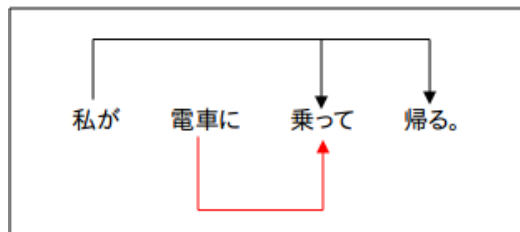


図 1：述語項構造の例

表 1：述語項構造の表現例

動詞	ガ格	ヲ格	二格	カラ格	ヘ格	ト格	ヨリ格	マデ格	デ格
乗る	私		電車						
帰る	私								

また、本研究では述語項構造を先行研究に基づいて表 1 のように表記する。表 1 は、図 1 と同様の例文「私が電車に乗って帰る。」の述語項構造を表している。

本研究では、述語のうち解析対象範囲を動詞、形容詞とする。また、項は先行研究に則って格助詞「ガ・ヲ・ニ・カラ・ヘ・ト・ヨリ・マデ・デ」に限定する。

2.1.1 述語項構造の利点、欠点

述語項構造を用いることで文の意味の骨格を表すことができる。そのため、次のような点で利用が可能である。

(1) 情報検索

現在の情報検索では、単語の共起によって検索を行っている。しかし、文単位で入力させることでよりユーザの必要な情報に近いものを結果として出力できると考えられる。

(2) 機械翻訳

機械翻訳に述語項構造を導入することで述語と項となる単語の関係が明確になる。このことを利用して機械翻訳を行うことで、翻訳結果の向上につながると考えられる。

一方で、述語を用いない、体言止めの文などを解析することができないという欠点もある。

2.2 先行研究

2.2.1 SynCha

奈良先端科学技術大学院大学が開発したシステムである。SynCha[5]の解析は、入力された文または文章をCaboChaと呼ばれる解析システムを通して構文・係り受け解析をし、NAIST テキストコーパスなどを利用した学習を利用することで述語項構造を出力している。また、SynCha は述語のみでなく事態性名詞についても述語項構造を構築でき、タグ付けをすることで文章全体から項を補うことも可能である。

2.2.2 結合価と格フレームを取り入れた述語項構造解析システム

昨年度、大澤によって開発された述語項構造解析システムである。このシステムは、KNP[6]を用いて解析を行う。解析から述語項構造を作成し、判定データにかけ、述語項構造を出力するものである。図 2 および次に処理過程を示す。

(1) 文の入力

(2) KNPを用いた文の係り受け・格解析

KNPを利用して入力文を解析を行う。

(3) 係り受け解析、格解析から述語項構造を作成

解析結果を用いて係り受け解析、格解析から個々に述語項構造を作成する。また、(5)の結合価を用いたデータとの照合のため、導き出された項はJUMAN[7]によって解析されカテゴリデータを抽出し、保存する。

(4) (3)の 2 つの述語項構造を合成

(3)で作成された 2 つの述語項構造を統合する。統合した述語項構造は、マッチングのパターンによって項を数値で表現している。

(5) (4)を結合価から作成したデータと照合、項を修正

(3)および(4)で作成した述語項構造を元にJUMAN のカテゴリを用いて結合価と比較を行い、項の修正を行う。

(6) 解析結果を表形式で出力

修正を行った述語項構造を表形式にして出力する。

本研究では、この結合価と格フレームを取り入れた述語項構造解析システム(以下、従来システム)を用いて拡張と改良を行う。

3. 研究内容

3.1 複合動詞

複合動詞は、「動詞+動詞」で構成され、2つの動詞が1つの動詞として意味・用法を持つものであり[8]、動詞全体の4割を占めている[9]。そのため、動詞の正しい判定には複合動詞であるかどうかの判断を正しく行うことが必要となる。

構文解析をルールベースで行っているJUMAN、KNPでは、複合動詞を多くの場合に、二つの個別の動詞として出力する。よって、従来システムも解析にJUMAN、KNPを使用していることから多くの場

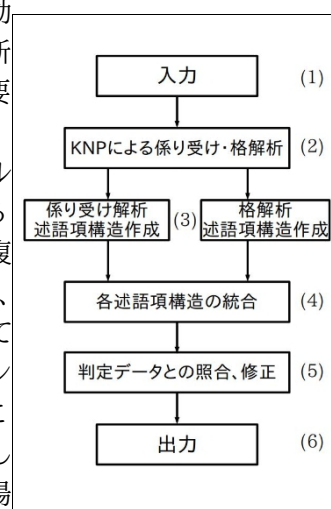


図 2 : 処理過程

合誤った解析を行っていた。表 2 は、「読者が文脈を読み取る。」の複合動詞出力である。日本語基本動詞用法辞典[10]によると「読み取る」は複合動詞として掲載されている。よってこの出力は、誤りである。

そこで本研究では、複合動詞を正しく出力するため、語順と接続の関係から複合動詞か否かを判定する。複合動詞であると判定した場合には、「動詞+動詞」を「複合動詞」として置き換えて出力する。この処理は、図 2 の「係り受け解析述語項構造作成」、「格解析述語項構造作成」で行う。図 3 は、図 2 の「係り受け解析述語項構造作成」内の複合動詞に関する処理を加えた詳細な処理過程を示す。

本研究では、(4)に複合動詞を扱う処理を追加した。導出した動詞の後の語が動詞であるかどうかを判定し、複合動詞である場合に修正して出力を行う。「格解

析述語項構造作成(図 4)」でも、同様の処理によって、複合動詞の述語項構造を出力している。

表 3 は、改良を行ったシステムで「読者が文脈を読み取る。」という文を解析した結果である。

3.2 形容詞

従来のシステムでは、述語に動詞に限定して述語項構造解析を行っている。一方で、前述の通り SynCha では、述語として形容詞も扱うことが可能である。従来のシステムで、動詞に関しては既に SynCha より精度の高い結果を得ている。そのため本研究では、述語の対象範囲を形容詞に拡張し、より多様な文の解析が行えるようにする。

図 2 の中で拡張を行った箇所は、複合動詞と同様に「係り受け解析述語項構造作成」、「格解析述語項

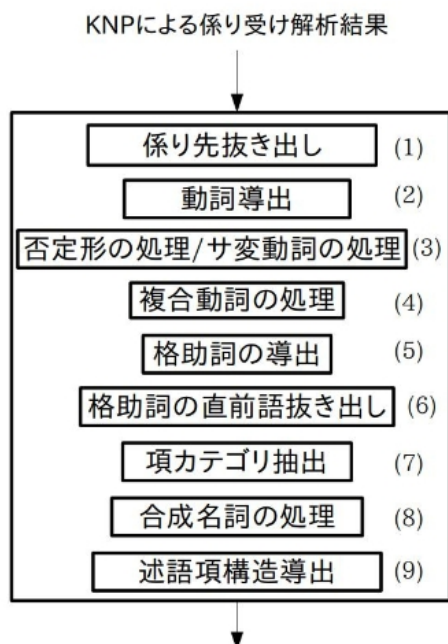


図 3：係り受け解析処理過程

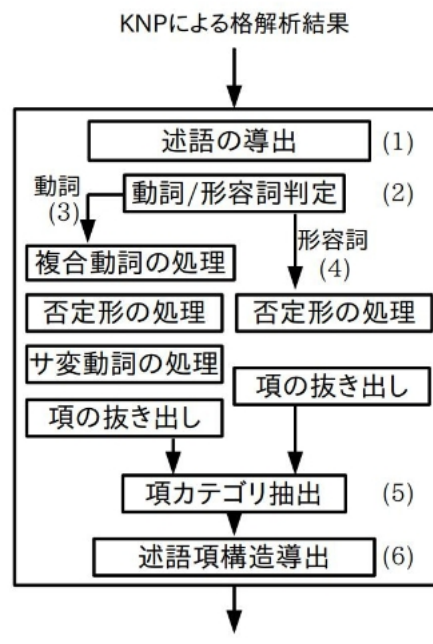


図 4：格解析処理過程

表 2：従来システムの出力例

動詞	ガ格	ヲ格	ニ格	カラ格	ヘ格	ト格	ヨリ格	マデ格	デ格
読む 取る	読者	文脈							

表 3：改良システムの出力例

動詞	ガ格	ヲ格	ニ格	カラ格	ヘ格	ト格	ヨリ格	マデ格	デ格
読み取る	読者	文脈							

構造作成」である。図 4 に前述の複合動詞の処理に加え、述語を形容詞に拡張したシステムの「格解析述語項構造作成」における処理過程を示す。

形容詞に関しては、(4)の拡張を行う。導出された述語が形容詞である場合に、否定形に対する修正をし、その形容詞に対して項の抜き出しを行うという手順で述語項構造を同定している。

また、形容詞は判定するためのデータが存在しない。そのため、計算機用日本語基本辞書 IPAL[11](以下、IPAL 形容詞)を用いる。IPAL 形容詞を動詞の判定に利用しているデータと同様のフォーマットで作成した。

3.3 評価実験

評価実験は、KNB コーパス[12]を用いた人手で行う。KNB コーパス内の 4,186 文の中から次のような点に基づいて 200 文を抽出した。

- ・ 動詞や形容詞を文に含む。
- ・ 句読点以外の記号がない。
- ・ 文語的に書かれている。

この 200 文を SynCha、従来のシステム、本研究システムのそれぞれで述語項構造解析を行う。解析結果を、KNB コーパスにある格関係を正解として評価を行った。

- ・ 正答率の算出

文単位の正答率 =

$$\frac{\text{(正しい述語項構造が出力された文の数)}}{\text{(全文数)}}$$
これを動詞、形容詞のそれぞれの場合について求める。

述語単位の正答率 =

$$\frac{\text{(正しい述語項構造が出力された述語数)}}{\text{(全述語数)}}$$
文同様、動詞、形容詞のそれぞれの場合について求める。

これらから求められた評価結果を表 4、表 5 に示す。

表 4 は、動詞の述語項構造の正答率を文(A)、動詞(B)単位で算出したものである。本研究のシステム(2)は A,B どちらにおいても従来システム(1)を越える精度である。これは、複合動詞の正しい導出によって単語間の関係が変わり他の動詞の出力も正しくなったからと考えられる。一方、SynCha(3)は短い文においても多くの項を同定しようとする傾向にあり、低い正答率になっているのではないかと考えられる。

表 4 : 動詞における比較結果

A	比較対象文数	200 文
A-1	文単位の正答率(従来)	42.0%
A-2	文単位の正答率(本研究)	46.5%
A-3	文単位の正答率(SynCha)	7.5%
B	200 文に含まれる動詞総数	625 個
B-1	動詞単位の正答率(従来)	57.1%
B-2	動詞単位の正答率(本研究)	64.4%
B-3	動詞単位の正答率(SynCha)	38.4%

表 5 : 形容詞における比較結果

C	比較対象文数	81 文
C-2	文単位の正答率(本研究)	55.5%
C-3	文単位の正答率(SynCha)	35.8%
D	200 文に含まれる形容詞総数	111 個
D-2	形容詞単位の正答率(本研究)	63.9%
D-3	形容詞単位の正答率(SynCha)	44.1%

表 5 は、評価対象 81 文で形容詞においての述語項構造解析の正答率を算出したものである。本研究システム(2)と SynCha(3)の比較であるが、本研究システムが良い精度を出している。しかしながら、形容詞の抜き出しに関しては SynCha が正しい場合も見受けられており、述語の抜き出しは改善すべきであると考えられる。

4. 問題点と今後の課題

評価実験から従来システムより結果の向上が見られた。しかしながら次のような問題点がある。

(1) 記号への対応

本研究では、記号に対する処理を行っていない。入力対象の拡大には必要となると考えられる。

(2) 長い項の出力

本研究では、複合名詞以上の長さの項を項として扱うことはできない。そのため文を項とする場合に出力がミスになってしまう。

(3) 固有名詞への対応

JUMAN で対応していない固有名詞があり正

しく項が出力されなかった。

また、これらの問題点に加えて、今後の精度向上に必要な課題として以下が挙げられる。

(1) 機械学習の利用

新語、固有名詞などに対応するためには学習での精度向上が必要であると考えられる。

(2) 判定データの追加

広範囲の判定を行うためには更なる判定データの追加の検討が考えられる。

(3) 述語の意味の同定

本研究システムでは、述語の意味を同定することまでは行っていない。しかしながら、判定データは各述語が意味ごとに区別されており、判定データとの一致による意味を同定は可能である。

(4) 複数文からの述語項構造

評価に用いた KNB コーパスにおいて同一文外からの格関係も付与されていた。そのため、文外からの項の同定も考えて行く必要がある。

これらの問題点、課題の解決によって精度の向上ができ、実用的な述語項構造解析システムの構築が可能であると考えられる。

5. まとめ

本研究では、先行研究である結合価と格フレームを取り入れた述語項構造解析システムを「複合動詞」、「形容詞」について改良、拡張し、より精度の高い正しい述語項構造解析が行えるようにシステムの構築を行ってきた。

結果としては、「複合動詞」を含めた動詞の述語項構造、「形容詞」の述語項構造ともに先行研究を上回る正答率を出している。しかしながら、問題点は数多く存在する。また、研究の進んでいない分野であるため、今後も新たな課題が出てくることが考えられる。今回の評価は、それら問題を除外した形で評価にとどまっている。よって、実用性という面においては改善の余地を残すシステムとなった。

だが、第4章で述べたような問題・今後の課題を1つずつ解決していくことにより、精度の高いシステム、またその他の分野に組み込んで使用できるシステムに一步步つながり近づいていくことが期待できる。

参考文献

- [1]飯田龍, 小町守, 乾健太郎, 松本裕治: 日本語書き言葉を対象とした述語項構造と共参照関係のアノテーション: NAIST テキストコーパス開発の経験から, 言語処理学会第13回年次大会, pp282-285 (2007)
- [2]鈴木勘平: 述語項構造と共参照関係, 山形大学情報科学専攻文献調査資料(2008)
- [3]大澤有美: 結合価と格フレームを取り入れた述語項構造解析システム, 山形大学卒業論文 (2009)
- [4]平博順, 藤田早苗, 永田昌明: 決定リストを用いた述語項構造解析, 言語処理学会第15回年次大会, pp368-371 (2009)
- [5]SynCha: (<http://cl.naist.jp/syncha/>)
- [6]KNP: (<http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/knp.html>)
- [7]JUMAN: (<http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html>)
- [8]姫野昌子: 複合動詞の構造と意味用法, ひつじ書房 ひつじ研究叢書第16巻 (1999)
- [9]森田良行: 日本語の複合動詞について, 早稲田大学語学教育研究所 (1987)
- [10]小泉保, 船城道雄, 本田晶治, 仁田義雄, 塚本秀樹: 日本語基本動詞用法辞典, 大修館書店 (1993)
- [11]計算機用日本語基本辞書 IPAL : (<http://www.gsk.or.jp/catalog/GSK2007-D/catalog.html>)
- [12]KNB コーパス(解析済みプログコーパス): (<http://nlp.kuee.kyoto-u.ac.jp/kuntt/>)