構造の不完全さに着目した等位文の述部機能表現補完

泉　朋子†　　今村　賢治†　　菊井　玄一郎†
†日本電信電話株式会社　NTT サイバースペース研究所
{izumi.tomoko, imamura.kenji, kikui.genichiro}@lab.ntt.co.jp

1　はじめに

近年，ブログやアンケートなど大量のテキストから有益な情報を自動で抽出・集計する技術が活発に研究されている．これらの情報抽出技術において，「行きたかった」など「どうした」を表す述部は欠くことのできない要素である．そこで我々は，述部が表す「出来事」の意味に着目し，同じ出来事を表す述部を同一の表層形へと言い換える述部正規化の研究を行っている(泉 et al., 2009).

この述部正規化を行う際には，述部を単純に言い換えるだけではなく，不完全な述部を完全な形に変換する必要もある．たとえば，「本当はハワイに行って，のんびりしたかった.」という場合，文中の述部(中間述部)「行って」は文末の述部(主節述部)から機能語の「たい」と「た」を補って「行きたかった」と正規化されなくては本来述部が意味する出来事と異なることを表してしまう．

そこで，本稿では，主節と対等に近いといわれる等位文中の中間述部に焦点を置き，これらを本来意味する述部に言い換える正規化を行う．

2　先行研究と本研究が対象とする述部

2.1　先行研究 (江原 et al., 2000)

中間述部を完全な形に言い換える先行研究として江原 et al. (2000)が挙げられる．江原らは，ニュース文を対象に，中間述部を終止形に言い換えることで文を短く分割した．この中間述部の変換方法として，時制とアスペクト，及び敬体を主節述部の機能語に合わせる規則を取り入れた．江原らの規則は具体的に次の通りである．第一に，主節述部が過去形であった場合に中間述部も過去形にする(時制決定規則). この時制決定規則は，中間述部が「おる」でおわっている時以外に適用する．次に，文末にアスペクト(相)の一種である継続を表す機能語「て(で)いる」があった場合に文中の述部も「て(で)いる」に変換する(相決定規則)．最後に，中間述部と主節述部の敬体を統一する(体決定規則)．この 3 つの規則によると，前述した例文における中間述部は，次のように言い換えられる(以後，必要に応じて機能語の形態素区切りを”/”で表わす).

(1)本当はハワイに行っ/て，のんびりし/たかっ/た.
　江原らの手法：行った
　本来述部が表す意味：行きたかった
　補完すべき機能語：「たい」，「た」

江原らの手法に沿うと，主節述部「のんびりしたかった」から過去を表す「た」のみを補い，中間述部と主節述部の時制を統一させる．しかし，(1)の場合は願望を表す機能語の「たい」も文末より補わなくては，対象の述部が本来表す出来事の事実関係と異なる意味に変換されてしまう．つまり，(1)の場合，実際にハワイに行ったか否かはわからないが(おそらく行けなかった)，それがハワイに「行った」という事実を表す述部に変換されてしまっている．このように，江原 et al.(2000)の規則のみでは，願望を表すモダリティ表現などの機能語は補完されず，誤った変換を行ってしまう．

2.2　本研究が対象とする中間述部

江原 et al.(2000)は，ニュース記事を対象に特定の機能語のみを補う規則を作成したため，ブログなど他の分野のテキストを扱ううえで問題がある．そこで本稿では，江原らでは扱われなかった時制や相以外の機能語をも考慮した中間述部の網羅的な言い換え規則を作成する．

一文内に複数の述部を含む文は多数あるため，本稿では等位文中の述部の言い換えに焦点を絞る．等位文とは，主節と対等に近い述部が複数連なっている文の事を指す. これら等位文における中間述部は，主節と対等という意味で，意味的に重要である．また，これらの等位文の出現頻度は新聞，ブログ双方において高い[1]．つまり，等位文中の中間述部の言い換えは，意味的な重要性という質的観点でも，出現頻度という量的観点でも重要であると言える．そこで本研究では，日本語記述文法研究会(2008)をもとに, 表 1 に記述されている 12 種類の接続形態によって接続されている中間述部を正しい形に言い換える正規化を行う．表 1 の述部の時制については後述する．以後，中間述部に必要な機能語を補う本研究の言い換えを「補完正規化」と呼ぶ.

3　構造の不完全さに着目した補完正規化

3.1　補完正規化 2 つの問題点

等位文中の中間述部を正規化する場合，(1)の例のように後ろの述部からすべての機能語を単純に補完

[1] 毎日新聞 2000 年 1 月とブログデータより，ランダムに 500 文を抽出した結果，新聞では 34.6%，ブログでは 26.6%が等位文であった.

表 1：等位文の接続形式

接続形態	述部の時制	例
連用形，て	Untensed	ハワイに行って，のんびりしたかった．
し，だけでなく，うえに，ばかりか，ほか(に)(は) けれど，が，のに対し(て)，一方(で)，反面	Tensed	ハワイに行ったし，グアムにも行った． ハワイに行ったけれど，のんびりできなかった．

するだけでは不十分である．たとえば，下記のような例が挙げられる．

(2) 眠たい/みたい/で，早く帰り/たがっ/て/い/た．
　本来述部が表す意味：眠たいみたいだった
　補完すべき機能語：「た」

(3) 今日ではバナナは安い/が，昔は高かっ/た．
　本来述部が表す意味：安い
　補完すべき機能語：なし

(2)の場合は，主節述部の機能語から過去の「た」のみを補う必要があり，他の機能語である「たが(る)」や「て/い(る)」は補ってはならない．一方，(3)の場合は「た」も補ってはならない(*今日ではバナナは安かった)．つまり，江原らの手法のように，単純に主節述部と時制を統一させるというルールだと，誤った過剰補完をしてしまう．このように，補完正規化を行う際には，次の 2 つを判断しなくてはならない．

i.　そもそも補完が必要か否かの判断．
ii.　補完が必要な場合，何をどこに補うか．

本研究では，述部の構造そのものに着目し，機能語を文法的な性質によって 3 種類に分類することで，これらの問題を解く．

3.2　正規化を指向した述部構造と機能語

述部を，詳細な階層構造に分類した研究は言語学の分野で多数なされている(e.g., 南, 1993). 本研究では Rizzi (1999)を参考に，日本語の述部の機能語を Mod(ality)，Foc(us)，T(ense)という 3 種類に分類し，下記の補完正規化のための述部構造を想定する．

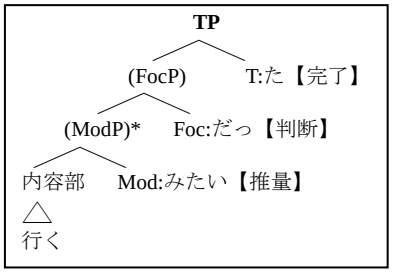

図 1：正規化を指向した述部構造と機能語[2]

ModP は願望を表す「たい」や推量を表す「みたい」のようなモダリティ機能語によって作り上げられるフレーズを意味する．FocP は助動詞「だ」など，「みたい」や「よう」などのモダリティ機能語の直後に文法的に必要な機能語が作るフレーズを意味する．TP は時制を表すフレーズを意味する．

本研究が対象とする等位文とは，本来，主節述部と中間述部が対等の関係で接続されているものである．そのため，中間述部が不完全な場合は，その不完全な要因になっている要素を主節述部から持ってくる補完という方法で完全な述部に変換できると考えられる．そこで，この述部構造を，補完処理において中間述部の不完全さをはかる指標として用いる．

また，この述部構造は，補完正規化用に独自に想定したものである．そのため，南(1993)のような詳細な意味分類に基づく階層構造とは異なり，ヴォイスを表す機能語など補完の対象にはならない機能語には独自のフレーズがない．一方，南では同一にされていた過去の「た」と判断の「だ」を区別し，文法的に必要な FocP という独自のフレーズを持たせた．これは FocP が後に説明する，不完全さをはかる処理において，重要な要素であるためだ．

本研究では，この 3 種類に分類される機能語に焦点を置く．そこで，機能語の表層形を一度抽象的な意味ラベルに変換し，そのラベルから 3 種類に属するものを選ぶ．今回は，多種多様な機能語に対応した「日本語機能表現辞書つつじ(松吉 et al., 2007)」を使用し，つつじのエントリーに付与されている意味ラベルから T，Mod，Foc の 3 種類に属するものを選出した．これによると，Foc には【判断】もしくは【名詞化】の意味ラベルをもつ機能語が，T には【完了】かつ表層形が「た(だ)」の機能語が，Mod には【願望】や【推量】といったモダリティを表す意味ラベルの機能語が選出された[3]．

この意味ラベルと述部構造を手がかりに，「述部のどの要素が欠けているか」という不完全さをはかり，補完正規化の問題を解く．具体的な補完正規化ルールは次の通りである．

・補完が必要か否かの決定

中間述部に何らかの補完が必要ということは，その述部は独立した文として成り立たないという事である．つまり，文として成り立てば，その述部は完全であり補完は必要ないと言える．そこで，生成文法などでとられている，「文 = 時制をもつ句(TP)」という考えをもとに(e.g., Adger, 2003, Ch.5)，時制の有無により補完が必要か否かを決定する．具体的には

[2] ModP と FocP は必ず現れるわけではなくオプショナルなものであると考える．さらにモダリティを表す ModP は，1 つ以上つくことができる．

[3] Mod に属する意味ラベルは多数あるためここではすべてを列挙しない．

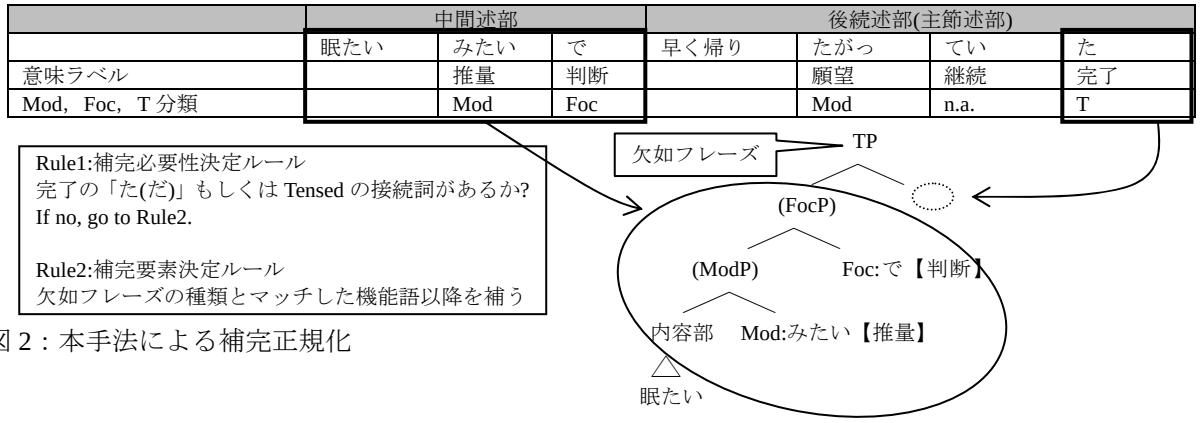

図 2：本手法による補完正規化

次のように解く．まず，中間述部に【完了】の意味ラベルをもつ「た(だ)」があるかを見て，あれば補完は行わない．これは，日本語が過去の時制を「た(だ)」でマークするため，「た(だ)」があれば，述部は時制を持つ(すなわち TP を持つ)と言えるからである．しかし，時制には過去以外に，現在・未来時制があるが，日本語にはこれらをマークする形態素は存在しない(e.g.,行く＋【未来】=行く Ø). そこで，等位文の中間述部につく接続詞を手がかりに時制を判定する．具体的には，あらかじめ過去の時制を持つ述部と接続できる接続詞を選出しておく. そして，これらの接続詞がついた場合には述部に「見えない時制(現在・未来)がある」と考える．これらの接続形態は，表 1 において「Tensed」に分類されている接続である[4].

このように，本技術では補完必要性決定ルール(Rule1)を導入し，時制がない述部のみを「不完全」とし，後の補完処理に進む.

・何をどこに補うかの決定

「不完全」と判断された中間述部に対し，何をどこに補うかを決定する．具体的には次のように解く．まず，当該の中間述部と，後続する述部(複数ある場合は一番近い述部)の意味ラベルを比較し，足りない意味ラベルを決定する．これらを，補完候補の機能語とする.

次に，中間述部が保持している機能語を手がかりに，述部構造と照合し，中間述部のどの部分が不完全かを決める(欠如フレーズ). 次に，補完候補の機能語の中から，**欠如フレーズにマッチする機能語以降**で Foc，Mod，T に属する機能語を補う．この処理を図 2 に示す. 図 2 が示すように当該の中間述部は，【判断】の意味ラベルを持っているために，述部構造では FocP まで完全に出来上がっている．そのため，補完候補機能語内でも，Mod に属する【願望】の「たい」は補完しない．【継続】の「ている」は Mod, Foc, T のいずれにも属さないため補完しない．唯一，TP が欠如しているので，T に属する「た」のみを補う(補完要素決定 Rule2)．最後に，すべての要素を適切に活用させる.

4　評価実験

本研究で提案した補完ルールの精度を測るため人手による評価データを用いてシステムの実装を行った.

4.1　評価用データ

ブログと新聞を対象に，あらかじめ等位文の述部を抽出し，人手で中間述部の言い換えを付与した．1 名の作業者が，(1)のような例文を参考に，後続の述部から機能語を補う必要がある場合には**すべて**補い，中間述部の意味を保った言い換えを作成した．もう 1 名が評価者となり，言い換えが作業指針と合っているか確認し，合っていない場合は，作業者と相談し，2 人が合意する言い換えを付与した．合計 1,217 の中間述部(ブログ 742 述部，新聞 475 述部)に正解を付与した．また，言い換え付与作業評価者とは別の作業者が，述部の機能語に「つつじ」の意味ラベルを付与した[5]．この正解をもとに，本システムの補完ルールを評価する.

4.2　比較に用いた手法

本技術の比較対象として，次の 3 つを使用する.

1. 補完なし：何も補完せずに終止形に変換
2. 単純補完 : 後続述部から中間述部が保持していない機能語をすべて補う
3. 江原(2000) :「(て)おり」の時以外に完了の「た」と継続の「ている」のみ補う

精度は，人手の正解との表層マッチで比べる．結果を表 2 に示す.

5　考察

本手法の述部の不完全さに着目した補完ルールを用いることで，従来法や補完をしない正規化よりも，高い精度で中間述部を正しい形に正規化することが

[4] 南(1993, pp.74-103)でも類似の議論がされており，接続形態と述部の階層レベルの関係が述べられている.

[5] 対応する意味ラベルがない場合は，「NULL」のラベルを付与した.

表 2:評価結果

	本手法	江原(2000)	補完なし	単純補完
精度	**73.1% (890/1217)**	58.9% (728/1217)	58.0% (707/1217)	26.1% (318/1217)

$$\text{精度} = \frac{\text{正解と表層マッチしたシステムの出力}}{\text{人手での正解(1,217 述部)}}$$

できた．特に，下記のような例が本手法によって可能となった正規化である．

(4) このような文明の利器が今から六十年前の世界に存在していれば，故江藤淳氏が指摘した「閉ざされた言語空間」が占領下の日本には存在せず… 全く異なる世界が開けていたのかもしれない．

本手法：存在しなかったの**かもしれない**

江原(2000)：存在しなかった

従来法ではできなかった「推量」の意味を補うことができ，述部が表す出来事の事実関係に沿った正規化が可能となった．また，時制の有無をもとに，補完が必要か否かを判断するルール(Rule1)を加えることで，過剰な補完処理を防ぐことができた．この Rule1 に沿わない例(すなわち，時制を持っているにもかかわらず補完処理を必要とした例)は全体で 8 例(0.7%)しかなかったため，時制の有無と補完の必要性との相関関係は，実際の言語現象とも一致していると考えてよい．

次に，正解との表層マッチが取れなかった 327 例(27.0%)のエラーを 3 種類に分けて考察する．ひとつが，正解と表層マッチはしなかったものの，述部が表している出来事の事実は正しく表現できているエラーである(Factuality preserved; F-PRESERVED)．これらは，下記のような例である．

(5) 拓ちゃんに会って，握手してもらったんだってぇ

人手の正解：会ったんだってぇ

本手法：会った

人手の正解は，「ん/だっ/てぇ」という終助詞的な機能語も補完しているが，これらは本システムが想定する補完対象の機能語(Mod, Foc, T)には属さないために補完できなかった[6]．このようなエラーは，正しい言い換えという意味では無視はできないかもしれないが，我々が目的とする「事態の事実関係を変えない述部正規化」という観点からは重要ではない．これらがエラーの半分以上を占めている．

一方，正しく機能語が補完できなかったために述部が表す出来事が異なってしまったエラー(Factuality changed; F-CHANGED)には下記のような例があげられる．

(6) スワロフスキーがキラキラで，今日も使い/まし/た.

[6]「だ」は，モダリティ機能語の直前直後に現れるもののみが FocP に属し，補完の対象であるとしている．

表 3:エラー分析

エラーの種類	エラー内の内訳
F-PRESERVED	52.0% (170/327)
F-CHANGED	42.8% (140/327)
Others	5.2% (17/327)

人手の正解：キラキラだ

本手法：キラキラだった

エラー原因：過去時制「た」の誤った補完

これらのエラーは，述部の内容語の意味や文脈までをも考慮しなくては，判断できない．つまり，次のような文脈だと本手法の補完で正解になる．

(7) 今はくすんでしまったが，買った当初のスワロフスキーはキラキラで，素晴らしかった．

本来述部が表す意味：キラキラだった

これらは，機能語の構造のみに着目した手法だと正しく補完することができない．文脈をも考慮した正規化ということで，今後の課題として検討したい．

6 結論

本稿では，述部の正規化として等位文中の中間述部に正しい機能語を補う補完正規化を行った．述部を 3 段階の階層構造に想定し，意味ラベルを手がかりに不完全さを決定することで，網羅的かつ簡潔なルールで中間述部の正規化を行うことができた. 特に，我々が目的とする「事実関係に着目した正規化」という観点からみると，本手法は極めて高い精度で中間述部の正規化を行うことができる．今後は，構造だけではなく文脈をも考慮に入れ，述部の正規化研究を進めていきたい．

謝辞

本研究を進めるにあたり，貴重なご意見を下さった名古屋大学大学院工学研究科の佐藤 理史教授，及び桝田 達也氏に深く御礼申し上げます.

References

Adger, D. (2003). *Core Syntax: A minimalist approach.* New York: Oxford University Press.

Rizzi., L. (1999). *On the position "Int(errogative)" in the left periphery of the clause.* Ms., Università di Siena.

泉朋子・今村賢治・菊井玄一郎・藤田篤・佐藤理史(2009). 正規化を指向した機能動詞表現の述部正規化. *第15回言語処理学会年次大会発表論文集*

江原暉将・福島孝博・和田裕二・白井克彦(2000). 聴覚障害者向け字幕放送のためのニュース文自動短文分割. *電子情報通信学会技術研究報告. NLC2000-12*, 17-22.

神田慎哉・藤田篤・乾健太郎 (2001). 連用節主節化に関する規則の追試と洗練. *第15回人工知能学会全国大会 1A1-06.*

日本語記述文法研究会(2008). *現代日本語文法6. 第11 部複文.* 東京:くろしお出版.

松吉俊・佐藤理史・宇津呂武仁(2007). 日本語機能表現辞書の編纂. *自然言語処理, 14, 5*, 123-146.

南不二男(1993). *現代日本語文法の輪郭.* 東京:大修館書店