文分割による連体修飾節の言い換え

野上優 [*1]　藤田篤 [*1]　乾健太郎 [*1] [*2]

[*1] 九州工業大学情報工学部知能情報工学科

[*2] 科学技術振興事業団さきがけ研究 21

{m_nogami,a_fujita,inui}@pluto.ai.kyutech.ac.jp

1 はじめに

ある言語表現をできるだけ意味を保存したまま別の言語表現に変換する「言い換え」技術は，自然言語処理の諸分野において様々な応用が考えられる重要な要素技術である [15]．たとえば，翻訳や要約の前処理で言い換えを行う試みはすでにいくつか報告されているし [5, 7, 4]，これらと同様に日本語・手話翻訳や音声合成といったタスクの前処理に利用できる可能性もある．また，推敲支援や要約，文章読解支援のように言い換え技術が根幹をなす応用もあり，それぞれの文脈の中で言い換えの実現を目指す研究も見られるようになってきた [2, 6, 1, 17]．最近では，言い換え技術を応用からある程度独立した要素技術として捉え，その性質や実現方法を解明する試みもいくつか見られる [3, 9, 15, 10]．しかしながら，日英翻訳のような言語間翻訳が長年精力的に研究されてきた経緯と比べると，「単言語内翻訳」である言い換えの研究はまだ極めて萌芽的な段階にあると言わざるをえない．

このような背景から，我々は「要素技術としての言い換え」の事例研究として，連体修飾節に着目した言い換えの研究を進めている．連体節の不適切な使用は文の可読性を下げる要因になるので，これを取り除く言い換えは推敲支援や文章読解支援，機械翻訳の後処理，音声合成の前処理といった応用に有効な技術として期待できる．

言い換えの研究にはいくつものアプローチが考えられるが，我々は次のような立場から研究を進めている．「ある表現の言い換えを生成する」という言い回しが自然なことからわかるように，言い換えは言語生成の変形と見なせる．したがってその過程は，言語に用意された様々な選択点において所与の文脈に適合する選択肢を選ぶ過程と捉えてよい．この視点に立てば，言い換えの解明・実現には少なくとも次の 3 点を明らかにする必要がある：

(a) 選択体系： 言い換えを生成する際にどのような選択点があり，個々の選択点にどのような選択肢があるか
(b) 選択基準： どのような場合にどの選択肢を選択すべきか
(c) 必要情報： 言い換えを生成するためには，元テキストからどのような構文／意味／談話情報を抽出しておく必要があるか

本稿では，一つの連体節（ただし，非限定的修飾節に限る）とそれを含む文，およびその先行・後続文脈を入力とし，連体節を主節から切り離して主節化することによって対象文を二文に分割するタスクを取り上げ，これまでに行った事例分析の中間結果と実現への課題について論じる．以下本稿では，このタスクを「連体節主節化」と呼び，もとの文を「原文」，連体節を主節化して言い換えたものを「連体節文」，主節の残りを言い換えたものを「主節文」，連体節が原文で修飾している名詞を「被修飾名詞」，原文の直前・直後の文を「前文」・「後文」とそれぞれ便宜的に呼ぶ．

2 言い換え事例の収集と分析

今回のタスクについては，適当な既存の事例集が見つからなかったので，京大コーパス [11] をもとに人手で言い換え事例を作成した．また，事例分析による選択点や選択基準の同定作業についても，問題の性質がほとんど解明されていない現状では機械学習的なアプローチは時機尚早と考え，すべて人手で行った．

まず，京大コーパスの一部（文化・読書・芸能・特集面全部，及び総合面の一部，1,840 文）から，連体節[1]を含む文を網羅的に収集したところ，517 文集まった．各文を手作業で分類したところ，限定的修飾節を含む文が 363 文，非限定的修飾節を含む文が 213 文，内容節を含む文が 106 文あった．これらのうち非限定的修飾節は，主節化が最も容易であると考えられ，また頻度も少なくないことから，連体節の言い換えを研究する際の最初の対象として適当であると判断した．

つぎに，上で収集した非限定的修飾節（229 箇所）の各々について，まずは先行・後続文脈を無視して，可能な言い換えを作成した．ただし，一文に複数の非限定的修飾節が存在する場合は，それぞれの言い換えを別々の事例として扱った．また，一つの連体節について複数の言い換えが可能な場合も異なる事例とみなした．この作業の結果，のべ 473 件の言い換え事例が得られた．つぎに，個々の言い換えについて先行・後続文脈との「つながりの良さ（結束性）」を評価したところ，不自然なものが 48 件見つかったので，それらに「結束性不良」の注釈（タグ）をつけた．なお，言い換えの可不可の判断や結束性の評価はすべて作業者の主観で行った．

事例の分析はおおよそ次のような手順で行った．まず，473 件の言い換え事例について，原文と主節文・連体節文を構文的に比較し，連体節主節化における選択点と選択肢をあらい出した．詳細は 3 節で述べるが，たとえば主節文と連体節文の順序の選択や連体節文におけるテンス・アスペクトの選択などの選択点が明らかになった．

つぎに，個々の言い換え事例がこれらの選択点でそれぞれどの選択肢を選択した結果であるか調べ，その情報を RDBMS を用いてデータベース化した．データベースのレコードの例を図 1 に示す．

上のデータベース化の作業が十分に進んだ時点で，選択基準の分析にもとりかかった．この作業ではまず，いくつかの主要な選択点に着目し，その選択に影響を与える可能性がある語彙／構文／意味／文脈情報を列挙する．

[1] 連体節の定義は文献 [12] に従った．

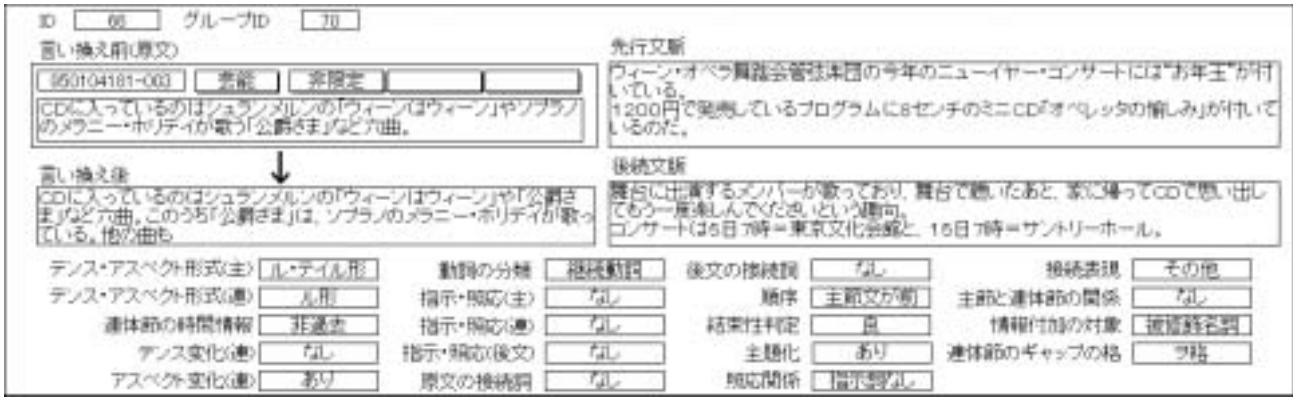

図 1: 言い換えデータベースのレコードの例

そして，それぞれの情報についてデータベース中の各レコードにタグづけし，選択肢との相関関係を調査した．

3 連体節主節化における選択点

原文と主節文・連体節文の構文的な比較から，連体節主節化には少なくとも以下のような選択点および選択肢があることがわかった．

3.1 主節文と連体節文の順序の選択

主節文を前に置くという選択肢と連体節文を前に置くという選択肢がある．

(前) 一七四二年に創立された **コスタ**は，スウェーデン最古の工場だ．

(後1) コスタは，スウェーデン最古の工場だ．一七四二年に創立された．

(後2) コスタは一七四二年に創立された．スウェーデン最古の工場だ．

3.2 主節文と連体節文の間の接続表現

主節文・連体節文のうち後続する方の文頭に接続詞（接続表現）を挿入した方がよい場合がある．挿入すべき接続表現には，少なくとも「しかし」「だから」「なぜなら」「そして」「ちなみに」の 5 種類の接続詞がある．

(前) 昨年トップだった **「変革」**は，今年は６社と３位に後退した．

(後) 「変革」は昨年トップだった．**しかし**今年は６社と３位に後退した．

3.3 文末表現の修正

主節文，連体節文それぞれについて文末表現の修正を必要とする場合がある．これには少なくとも以下の 3 つの選択点がある．

3.3.1 連体節文のテンス・アスペクト表現の修正

原文における連体節と言い換え後の連体節文とでは，テンス・アスペクト表現が必ずしも一致しない．この修正は，「ル形／タ形」の選択と，アスペクト表現（「ティル／テアル」など）の有無の選択の組み合わせからなる．次の例では，テンス・アスペクトともに変化している．

(前) 米国の新聞社が昨年から，パソコン通信を利用**した「電子新聞」サービス**を次々本格スタートさせた．

(後) 米国の新聞社が昨年から，「電子新聞」サービスを次々本格スタートさせた．このサービスはパソコン通信を利用**している**．

3.3.2 接続表現との呼応にともなう修正

たとえば，主節文と連体節文の間に接続詞「なぜなら」が入ると，それと呼応する文末表現に修正が必要である．

(前) 週休二日制があまり普及していない **韓国**では，貴重な連休であることは変わりはない．

(後) 韓国では，貴重な連休であることは変わりはない．**なぜなら**，週休二日制があまり普及していない**からだ**．

3.3.3 その他の文末表現の修正

この他，常体 / 敬体の修正などいくつかの選択点がある．また，連体節文が被修飾名詞に対する情報付加であることを明示するために，次の例のように「（〈被修飾名詞〉は . . . する）○○である」といった表現を補った方がよいといった場合もある．

(前) 政党助成法は，今月一日に 公費助成を受ける政党に法人格を与える **政党法人格付与法**や改正政治資金規正法とともに施行された．

(後) 政党助成法は，今月一日に政党法人格付与法や改正政治資金規正法とともに施行された．（ちなみに，）政党法人格付与法は，公費助成を受ける政党に法人格を与える**制度**である．

3.4 連体節文におけるギャップの復元

(a) 被修飾名詞を主題化するか，ギャップの格として補完するか，(b) 連体節のどの場所に挿入するか，に関する選択がある．

(前) 米国では医療用具として一般の小売店で販売している **コンタクトレンズ溶液**は，日本では医薬品．

(後1) **コンタクトレンズ溶液は**，米国では医療用具として一般の小売店で販売されている．しかし日本では医薬品．

(後2) 米国では**コンタクトレンズ溶液を**医療用具として一般の小売店で販売している．しかし日本では医薬品．

(後3) 米国では医療用具として一般の小売店で**コンタクトレンズ溶液を**販売している．しかし日本では医薬品．

3.5 照応表現の選択

主節文内の被修飾名詞と連体節文に挿入された被修飾名詞の共参照関係を明示するために照応表現の選択が必要になる．指示連体詞の挿入が主要な選択肢である．

(前) これだけ話題になった **ハンチントン論文**を素通りできるわけはない.

(後) ハンチントン論文はこれだけ話題になった. **そのような**論文を素通りできるわけはない.

3.6 前文や後文の修正

文の結束性を保持するために，前文や後文の修正が必要になる場合がある.

(前) ＣＤに入っているのはシュランメルンの「ウィーンはウィーン」や ソプラノのメラニー・ホリデイが歌う **「公爵さま」**など六曲. (以下後文) 舞台に出演するメンバーが歌っており，舞台で聴いたあと，家に帰ってＣＤで思い出してもう一度楽しんでくださいという趣向.

(後) ＣＤに入っているのはシュランメルンの「ウィーンはウィーン」や「公爵さま」など六曲. このうち「公爵さま」は，ソプラノのメラニー・ホリデイが歌っている. **他の曲も**舞台に . . . メンバーが歌っており . . .

4 連体節主節化における選択基準

3 節で示した選択点のうち，(a) 主節文と連体節文の間の接続表現，(b) 連体節文のテンス・アスペクト表現の修正，(c) 主節文と連体節文の順序の選択，を取り上げ，それぞれの選択基準の解明を試みた. 以下に事例分析の現状を報告する.

4.1 主節文と連体節文の間の接続表現

接続表現の選択は少なくとも次の情報に依存すると考えられる.

- 連体節の表現機能 [13]
  - (a) 主節で表される事態に対する情報付加
    （付加情報の内容は，対比・逆接，原因・理由，継起，付帯状況の 4 種類）
  - (b) 被修飾名詞に対する情報付加
  - (c) 情報付加でない
- 主節文と連体節文の順序

これらと接続表現の選択との対応関係を調べたところ，次のような傾向をつかむことができた.

まず，上記 (a) の場合は，付加情報の内容を表す適当な接続詞を必要とする場合が多かった. とくに原因・理由の場合，44 事例すべてにおいて，主節文が前の場合は「なぜなら」など，連体節文が前の場合は「だから」などが必要であった. 同様に対比・逆接についても，連体節文が先行する場合は，19 事例すべてにおいて「しかし」が必要であった.

一方上記 (b) の場合は，主節文先行，連体節文先行のいずれの場合も接続詞を必要としない事例が大半を占めた（それぞれ，186 事例中 157 事例，194 事例中 185 事例）. しかしながら，やはり何らかの接続詞を必要とする場合もあり，その選択基準の解明は今後の課題である.

4.2 連体節文のテンス・アスペクト表現の修正

テンス・アスペクト表現の選択は少なくとも原文の主節，連体節における次の情報に依存すると考えられる.

- 各節の事態の時間情報： 過去／非過去
- 各節のテンス・アスペクト表現： タ形／ル形／テイル形／テイタ形
- 連体節の述語の種類： 状態動詞／継続動詞／瞬間動詞／第四種の動詞 (以上 [8]) ／判定詞

これらと連体節文のテンス・アスペクト表現の関係を調査したところ，表 1 のような結果を得た. これによると，連体節の述語が状態動詞，または判定詞の場合，テンス・アスペクトは常に変化しない. また，連体節の述語が継続動詞，瞬間動詞，第四種の動詞の場合，連体節のテンス表現は，連体節がタ形非過去の場合を除いて原則として変化しない. ただし，アスペクト表現の選択については今回の分析からは明確な基準が得られなかった. これについての分析は今後の課題である.

表 1: 連体節文におけるテンス・アスペクト表現の選択

				主節	
				ルとテイル	タとテイタ
連体節	継続動詞 瞬間動詞 第四種の動詞	ル	非過去	ル (38) テイル (54)	ル (14) テイル (7)
			過去	(0)	タ (2)
		テイル	非過去	テイル (22)	テイル (18)
		タ	非過去	テイル (30) テイル / タ (2) タ (8)	テイル (10) テイタ (3) タ (2)
			過去	タ (77) テイタ (18)	タ (46) テイタ (25)
		テイタ	過去	テイタ (10)	テイタ (10)
	状態動詞 判定詞			不変 (65)	

4.3 主節文と連体節文の順序

予備調査として，以下のような単純な仮説を立て，それぞれが分析対象の事例でどの程度成り立っているかを調べた. 仮説に適合した事例の割合を括弧内に示す.

(a) 言い換え文の先頭に接続詞があると連体節文は先行できない（14/14）

(b) 後文の先頭に接続詞があると主節文は先行できない（8/15）

(c) 前文と主節との間に共参照関係があると連体節文は先行できない（14/18）

(d) 前文と連体節との間に共参照関係があると主節文は先行できない（6/10）

(e) 後文と主節の間に共参照関係があると主節文は先行できない（11/23）

この結果から，(a) と (c) が比較的顕著な傾向を示しているのに対し，(b)，(d)，(e) は順序の制約としては強すぎることがわかる. すなわち，主節文に先行する連体節文は主節文と前文との結束性をほとんどの場合に破壊するのに対し，主節文に続く連体節文は主節文と後文の結束性を破壊しない場合も少なくない. 後者については，連体節文の文頭の接続表現や文末表現，主題化の選択などの情報と併せた詳細な分析が必要である.

5 言い換え知識の実装

本研究では，上述のような言い換えの選択点・選択基準の分析とともに，それらを実装するためのプラットフォームとして言い換えエンジンの開発を進めている.

5.1 言い換え規則と言い換えエンジン

選択点と選択基準からなる言い換え知識を言い換え規則の集合として記述する．言い換えエンジンは，この言い換え規則を解釈し，入力に対応する可能な言い換えを複数生成する．入力には GDA タグ [14] つきのテキストを仮定し，出力も GDA テキストである．個々の言い換え規則は，選択点，選択肢，選択条件，変数束縛，処理本体からなる．処理本体は，子の選択点あるいはプリミティブな変換処理の列で定義する．プリミティブな変換処理は，必要に応じて順次手続き的に実装し，ライブラリ化してある．

選択点:	連体節の言い換え (RelCls)
選択肢:	文分割
選択条件:	非限定的連体節
変数束縛:	H = RelCls の被修飾名詞
	Gap = RelCls における H の格
処理本体:	連体節の切り離し (S, RelCls, S1)
	文の順序決定 (S, S1)
	ギャップの復元 (H, S1, Gap)
	接続表現の修正 (S, S1)
	文末表現の修正 (S1)
	指示・照応表現の修正 (S, S1)

図 2: 言い換え規則 (連体節の言い換え)

選択点:	ギャップの復元 (H, S1, Gap)
選択肢:	主題化
選択条件:	ギャップの格がガ格
処理本体:	格助詞の設定 (H)
	対象節の先頭に挿入 (H, S1, Gap)

図 3: 言い換え規則 (ギャップの復元)

例として，非限定的修飾節を主節化するためのトップレベルの言い換え規則を図 2 に示す．図中の「連体節の切り離し」はプリミティブな変換処理の例で，たとえば次の例文 (1) を (2) に変換する処理を受け持つ．

(1) この間までは受験で悩んでいた **満男**がもう転職で悩んだりしています．

(2) この間までは受験で悩んでいた．満男がもう転職で悩んだりしています．

一方，「ギャップの復元」は複数の選択肢を持つ選択点の例であり，ここから図 3 に示すような言い換え規則が呼び出される．言い換え規則の呼び出しの際は，規則の選択条件を評価し，適合する場合にのみ処理本体を実行する．図 3 の規則を実行すると，たとえば上の (2) の状態を (3) の状態に変換することになる．

(3) 満男はこの間までは受験で悩んでいた．満男がもう転職で悩んだりしています．

選択基準には，個々の言い換え基準として局所的に記述できるものと，複数の選択の組み合わせの結果を大域的に評価する基準として記述すべきものがあると考えられる．今回の実装は前者をカバーするもので，後者の実装については今後の課題である．

5.2 予備実験

分析に用いた事例のうちギャップの格がガ格である 151 箇所の非限定的修飾節に対して，言い換え規則を適用した結果と人手で作成した言い換え文を比較するという実験を進めている．入力には，京大コーパス [11] から取り出した形態素・構文情報を GDA タグ [14] に変換し，これに言い換えに必要な情報（たとえば，ギャップの格の情報）を人手で付加したものを用いる．これまでのところ，連体節の切り離しとギャップの復元に関する変換処理を実装し，上記 151 件中 139 件に対して人手による言い換えと同じ変換を実現できることが確認できており，現在の方針で言い換え知識を実装できる見通しを得ることができた．今後は，接続表現の選択や文末表現の修正，照応表現の生成に実装対象を広げていく予定である．

6 おわりに

本稿では，言い換えの事例研究として文分割により連体修飾節の言い換えについて論じた．事例分析の結果，言い換えを構成する選択点と選択肢，および一部の選択基準で有力な手がかりを得ることができた．今後は，不明な選択基準に関する詳細な分析と，それらの実装方法のより高度な検討が必要である．また，オープンテストによる評価が必要なことは言うまでもない．

参 考 文 献

[1] Carroll, J., et al. Simplifying text for language-impaired readers. *Proc. of EACL*, pp. 271–272, 1999.

[2] Dras, M. Reluctant paraphrase: Textual Restructuring under optimization model. *Proc. of PACLING*, pp. 98–104, 1997.

[3] Dras, M. Representing paraphrases using STAGs. *Proc. of ACL-EACL'97*, pp. 516–518, 1997.

[4] 福島孝博, 江原暉将, 白井克彦. 短文分割の自動要約への効果. 自然言語処理 Vol. 6, No. 6, pp. 131–147, 1999.

[5] 金淵培, 江原暉将. 日英機械翻訳のための日本語ニュース文自動短文分割と主語補完. 自然言語処理研究会報告書, NL-93-3, pp. 15–22, 1993.

[6] 片岡明, 増山繁, 山本和英. 要約のための連体修飾節の“A の B”への言い換え. 自然言語処理研究会報告書, NL-133-7, pp. 37–44, 1999.

[7] 木村真理子, 野村浩一, 平川秀樹. 日英機械翻訳前編集における日本語文分割処理について. 自然言語処理研究会報告書, NL-96-8, pp. 57–64, 1993.

[8] 金田一春彦. 国語動詞の一分類. 1945.

[9] 近藤恵子, 佐藤理史, 奥村学. 「サ変名詞＋する」から動詞相当句への言い換え. 情報処理学会論文誌 Vol. 40, No. 11, pp. 4064–4074, 1999.

[10] 近藤恵子, 佐藤理史, 奥村学. 格変換による単文の言い換え. 自然言語処理研究会報告書, NL-135-16, pp. 119–126, 2000.

[11] Kurohashi, S. and Nagao, M. Building a Japanese parsed corpus while improving the parsing system. In *Proc. of NLPRS*, 1997.

[12] 益岡隆志, 田窪行則. 基礎日本語文法 (改訂版). くろしお出版, 1994.

[13] 益岡隆志. 複文. くろしお出版, 1997.

[14] Neumann, C. and Hashida, K. Enhance of Machine Translation with GDA-Tags. 自然言語処理研究会報告書, NL-126-3, pp. 17–24, 1998.

[15] 佐藤理史. 論文表題を言い換える. 情報処理学会論文誌, Vol. 40, No. 7, pp. 2937–2945, 1999.

[16] 白井諭, 池原悟, 河岡司, 中村行宏. 日英機械翻訳における原文自動書き換え型翻訳方式とその効果. 情報処理学会論文誌, Vol. 36, No. 1, pp. 12–21, 1995.

[17] 山本聡美, 乾健太郎, 野上優, 藤田篤, 乾裕子. 聾者向け文章読解支援のための文可読性基準の調査. 自然言語処理研究会報告書, NL-135-17, pp. 127–134, 2000.