共通の係り先文節を持つ文節組を利用して言い直し表現を検出・修正するシステム

島森瑛貴 [1] 森辰則 [1]
[1] 横浜国立大学大学院環境情報学府
simamori-eiki-vj@ynu.jp
tmori@ynu.ac.jp

概要

本稿では日本語の自発的な独話における言い直し現象の分析を元に、言い直しを検出・修正するシステムを作成しその評価を行った。提案システムは解析部、文節内ならびに文節間に出現する言い直しの検出・修正部からなる。文節間に出現する言い直しの検出・修正は、共通の係り先を持つ文節組を言い直しの候補とする。実験の結果、クローズドテストでの適合率は 0.288、再現率は 0.277、F1 値は 0.282、オープンテストでの適合率は 0.350、再現率は 0.337、F1 値は 0.348 となり先行研究と比較し精度が改善した。

1 はじめに

新型コロナウイルスの流行により生活の様々な場面でオンライン化が進んでいる。それに伴い人の発話を書き起こしテキストとして読む場面が増えた。話し言葉の書き起こしは、未編集の状態では「言い直し表現」が頻出する。言い直し表現は書き言葉には出現しないため可読性を下げる要因となりうる。我々はそれらを適切に検出・修正するシステムを開発し可読性を向上させることを目標としている。

本稿では日本語の自発的な独話に出現する言い直しを検出・修正するシステムを提案し評価を行う。

2 関連研究

話し言葉の言い直しを修正する先行研究には [1, 2, 3] などがある。

下岡ら [1] は高梨ら [4] が定義した言い直しのタグを正解として、形態素の繰り返し情報などを素性とした SVM を用いて任意の文節が言い直しか判定するが、2 文節以上を削除するものを扱えない。また高梨ら [4] は「同一の内容を指し示している対等な文節」のみを言い直しと定義して CSJ[5] にアノテーションしており我々の扱いたい問題が含まれない。

藤井ら [2] は言い直し部を分割したモデルを作成し言い直しを検出・修正するが、そのモデルはフィラーや言い淀みが存在しない言い直し表現については検出ができず修正の対象が非常に限定的である。

島森ら [3] は島森ら [6] による言い直しの定義、分析、仮説の検証を元に日本語の自発的な独話に出現する言い直しを検出・修正するシステムを提案し評価した。島森ら [6] は、言い直しを構成する要素のうち最後のものを「言い直し先」、それ以外を「言い直し元」、これらの組を「言い直し組」と定義しており本研究もその定義に従う。このシステムは複数の文節に共通して出現する自立語に着目する。また、意図しない名詞連接の係り受け変更や言い直しを検索する範囲を節境界の間に限定することで誤検出を抑制した。一方で節境界を利用して言い直しの検索範囲を狭めると、言い直しを検出できなくなるトレードオフがある。例文 1 は「個人性に関連した音声工学には」を言い直しているが、「連体節」や「主題ハ」という節境界を利用して言い直しの検索範囲を狭めると言い直しを検出できない。一方で、例文 2 は「目標物」という自立語が共通しており「連体節」のラベルにより検索範囲を狭めないと言い直しとして誤修正される。

(1) 個人性に関連した**/連体節/**音声工学には**/主題ハ/**個人性に関連した**/連体節/**音声工学には**/主題ハ/**話者適応の技術があります

(2) コウモリが目標物に向かう**/連体節/**時は目標物の直前からパルスを放射し

例文 2 では、共通の自立語「目標物」を持つ 2 つの文節のうち「目標物に」は「向かう」に係り、「目標物の」は「直前から」に係る。このように係り先が異なるものが言い直しになるとは考えにくい。そ

こで共通の係り先を持つ文節組のみを言い直し組候補の集合として言い直しを検出する手法を島森 [3] に追加する手法を提案する。

3　システムの構成

本稿では文節間言い直し修正部において共通の係り先を持つ文節組に着目する手法を提案する。解析部と文節内言い直し修正部は島森 [3] と共通である。

提案システム全体の概要を図 1 に示す。

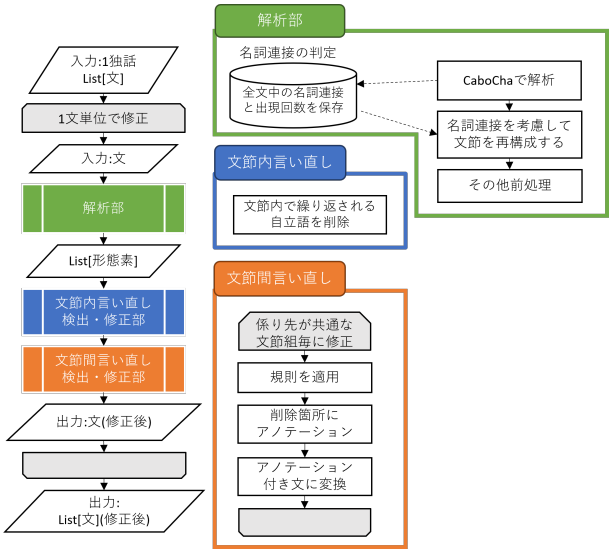

図 1　提案システムの概要

3.1　解析部の方針

解析部は、文を入力に取り CaboCha+MeCab を用いて形態素・係り受け解析を行う。その後島森 [3] の方針で、言い直しにより生じる意図しない名詞連接を係り受けを変更して解消する前処理を行う。

3.2　文節内の言い直し修正部

全ての文節を対象とし、同一文節内の部分文字列間に包含関係があるときに最初の出現を削除する。例えば「スケート$_1$ 日本 スケート$_2$ 連盟は」という文節であれば「スケート$_1$」を削除して修正する

対象とする文節 A の形態素数を N としたときに $1 \leqq n \leqq \lfloor N/2 \rfloor$ を満たす n について以下を実行する。

文節 A の最初から n 個の形態素 $a_1 \ldots a_n$ を取得し言い直し元候補 A_{bef} とする。そして $a_{n+1} \ldots a_N$ から連続する n 個の形態素の列 $a_s \ldots a_{s+n-1}$ を取得し言い直し先候補 A_{aft} とする。それぞれで要素の形態素の原型を連結した文字列を $S_{\mathrm{bef}}, S_{\mathrm{aft}}$ とする。取りうる S_{aft} と S_{bef} の組のいずれかで S_{bef} が S_{aft} の部分文字列のときに A_{bef} が言い直し元だとし削除する。

3.3　文節間の言い直し修正部

1 文節以上からなる係り受けの部分構造のうち共通の係り先を持つものを言い直し組を構成する候補の集合として絞り込み、要素の全ての対でそれぞれが言い直し組か判断する。この対が言い直しの関係かを判定する規則は対の要素の文節数の関係に着目し、「1 文節-1 文節」、「1 文節-2 文節以上」、「2 文節以上-1 文節」、「2 文節以上-2 文節以上」の 4 種類に大別される。各規則は、言い直しかどうかを判定する条件部と、言い直しである場合に削除箇所の検出と削除を行う実行部を持ち、言い直しの検出と削除を行う。また、一度削除すると判断したものは以降の組み合わせを調べる際には検索の対象外とする。

例文 2 を係り受け解析した結果の表 2 を元に本稿で提案するシステムの詳細を説明する。

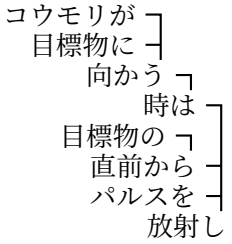

図 2　例文 2 の係り受け解析結果

表 2 から、「向かう」に係るものとして「コウモリが」、「目標物に」の 2 要素の集合、「放射し」に係るものとして「コウモリが目標物に向かう時は」、「目標物の直前から」、「パルスを」の 3 要素の集合が、共通の係先を持つ要素の集合として得られる。

共通の係り先の文節、係り元の文節の出現が早い順に上から全ての対を比較する。この例では初めに「コウモリが」と「目標物に」を比較する。「コウモリが」と「目標物に」は共に 1 文節なので「1 文節-1 文節」の規則を適用する。次に「コウモリが目標物に向かう時は」と「目標物の直前から」を比較する。これは「2 文節以上-2 文節以上」の規則を適用する。以降同様に「コウモリが目標物に向かう時は」と「目標物の直前から」、「目標物の直前から」と「パルスを」を比較し規則を適用する。この例文に言い直しはないと最終的に判断する。

3.3.1　1 文節-1 文節

言い直し元候補と言い直し先候補が共に 1 文節の場合の規則と削除箇所を以下に示す。以下「文節 A が文節 B を含む」は「文節 B の自立語が全て文節 A

に出現する」を指し、付属語の一致は考慮しない。

- 「言い直し先の文節が言い直し元の文節を含む」とき「言い直し元」を削除する

修正例を図 3 に示す。この例では「実験は$_1$」と「実験は$_2$」が共に「行いました」に係る。これは規則を満たすため言い直し元の「実験は$_1$」を削除して修正する。

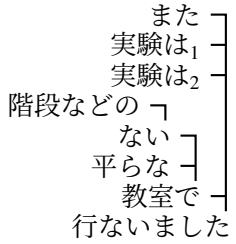

図 3　1 文節-1 文節の例

3.3.2　1 文節-2 文節以上

言い直し元候補が 1 文節、言い直し先候補が 2 文節以上の場合の規則と削除箇所を以下に示す。

1. 「言い直し先の最初の文節が言い直し元の文節を含む」とき「言い直し元」を削除する
2. 「言い直し先の最後の文節が言い直し元の文節を含む」とき「言い直し元」を削除する

修正例を図 4 に示す。この例では「演劇に$_1$」と「普通の-演劇に$_2$」が共に「行きたいなっていう」に係る。これは規則 (2) を満たすため言い直し元の「演劇に$_1$」を削除して修正する。

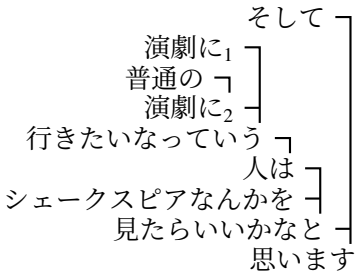

図 4　1 文節-2 文節以上の例

3.3.3　2 文節以上-1 文節

言い直し元候補が 2 文節以上、言い直し先候補が 1 文節の場合の規則と削除箇所を以下に示す。

- 「言い直し先の文節が言い直し元の最後の文節を含む」とき「言い直し元の最後の文節」を削除する

修正例を図 5 に示す。この例では「意味統合処理の-失敗時に$_1$」と「失敗時に$_2$」が共に「誘発されると」に係る。これは規則を満たすため言い直し元の最後の文節「失敗時に$_1$」を削除して修正する。

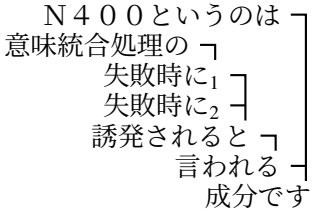

図 5　2 文節以上-1 文節の例

3.3.4　2 文節以上-2 文節以上

言い直し元候補と言い直し先候補が共に 2 文節以上の場合の規則と削除箇所を以下に示す。

- 「言い直し先が言い直し元の全ての文節を含む」とき「言い直し元の全ての文節」を削除する

修正例を図 6 に示す。この例では「その$_1$-候補$_1$ で」と「その$_2$-候補$_2$ の-パラメーターで」が共に「処理した」に係る。言い直し元候補の 2 つの自立語「その」と「候補」はいずれも言い直し先の候補に含まれる。これは規則を満たすため言い直し元の「その$_1$-候補$_1$ で」を削除して修正する。

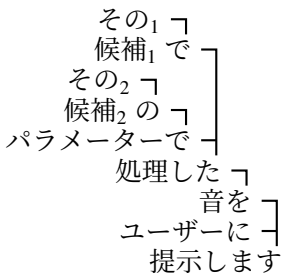

図 6　2 文節以上-2 文節以上の例

3.3.5　共通処理

上記の 4 つの規則で以下の 3 つの処理を行う。

1. 編集表現の修正
2. 言い直しを謝る表現
3. 接続詞「で」を対象外

編集表現の修正

島森ら [6] は言い直しのうち修正の意図を持つ箇所を「編集表現」と定義した。例文 3 では「かわいい」を言い直して「いい雰囲気の」という際に出現する「って言うか」が編集表現に相当する。

島森ら [3] の提案方針は、2 つの文節が共通の自立語を持つときのみを言い直しの候補とするため、

編集表現を持つもののうち言い直し元と言い直し先の自立語が異なるものを検出・修正できない。

一方で本稿の提案手法では共通の係り先を持つものに着目するため、言い直し元の削除範囲を改めて推定する必要がない。そこで、言い直し元候補が編集表現の手がかりとなる文字列を含むときに言い直し元として削除する。編集表現の文字列は島森ら [6] の分析で得られたものを使用する。

図 7 に示す解析結果より「かわいいって言うか」と「いい雰囲気の」を比較するときに、「かわいいって言うか」の箇所が言い直しであるとして削除する。以降では削除箇所を言い直しの候補から外して考えるため、例えば「港町に」と「かわいいって言うかいい雰囲気の建物が」を比較するときは「港町に」と「いい雰囲気の建物が」を比較する。それにより、並列表現を持つ箇所は最も小さい単位で言い直しを探すため誤検出を抑制できると考える。

(3) 港町にかわいいって言うかいい雰囲気の建物がありました

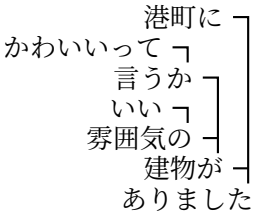

図 7 例文 3 の係り受け解析結果

言い直しを謝る表現

本研究が対象とする言い直しの中には、言い直し元を発話した直後に「すみません」などの言い誤ったことを謝る意図の発話を含むものが存在する。これらは編集表現と同様に修正の対象とする。具体的な処理としては、言い直しとして削除する文節の直後に品詞が「感動詞」の形態素が存在したとき、この感動詞を上記の表現であるとして削除する。

接続詞「で」の除外

話し言葉では接続詞「で」が 1 文中に複数回出現することがある。このような場合には「で」が言い直しとして誤修正される。そこで、言い直し元候補の文節として接続詞「で」が検出されたときには言い直しではないと判断し、以降の処理を行わない。

4 システムの評価実験

島森 [3] のベースラインシステム (ベースライン)、島森 [3] の 6 章の提案システム (BL+前処理+節境界)、本稿 3 章の提案方針によるシステム (BL+前処理+共通の係り先) の合計 3 つを比較し評価する。

評価の対象には日本語話し言葉コーパス (CSJ)[5] のノンコア講演 140 講演を用いる。このうち 40 講演をシステム改良のための分析に利用する。また、開発する際に参照した事例に対する評価ということでクローズドテストとして評価する。残りの 100 講演は未参照の事例に対する評価を行うときのデータとして使用し、オープンテストとして評価する。言い直しの正解は島森ら [3] が付与したものを利用する。講演を文単位に分割したものをシステムへの入力とした。クローズドテストの 40 講演は 1959 文で言い直しは 486 個、オープンテストの 100 講演は 5201 文で言い直しの個数は 1139 個であった。

システムの評価は、システムが削除すると判断した箇所に IOB2 形式のラベルを文字単位で振り、削除単位毎に一致するかを調べ適合率、再現率、F1 値を求めた。島森ら [3] が提案した方針に基づくシステムと精度を比較する。クローズドテストの評価を表 1 に、オープンテストの評価を表 2 に示す。

表 1 クローズドテスト

	適合率	再現率	F1 値
ベースライン (BL)	0.048	0.325	0.083
BL+前処理+節境界	0.197	0.319	0.243
BL+前処理+共通の係り先	0.288	0.277	0.282

表 2 オープンテスト

	適合率	再現率	F1 値
ベースライン (BL)	0.048	0.315	0.084
BL+前処理+節境界	0.200	0.343	0.253
BL+前処理+共通の係り先	0.359	0.337	0.348

5 考察とまとめ

本稿で提案した方針によりシステムの精度が向上した。また、クローズドテストの精度よりオープンテストの方が精度が良かった。これはオープンテストはクローズドテストより文節間言い直しの割合が低いことが原因と考えられる。対象の文によっては係り受け解析器が解析誤りを起こすことがわかっている。文節間言い直しは文節内言い直しより解析誤りの影響を受けやすいため修正を誤るものがある。一方で、現在のシステムでもうまく修正できないような例が複数存在する。具体的な例は以下の付録 A に示す。現時点で修正できていない誤りの修正は今後の課題である。

参考文献

[1] 下岡和也, 河原達也, 内元清貴, 井佐原均. 『日本語話し言葉コーパス』 における自己修復部 (d タグ) の自動検出および修正に関する検討. 情報処理学会研究報告音声言語情報処理（SLP）, Vol. 2005, No. 50, pp. 95–100, 2005.

[2] 藤井はつ音, 岡本紘幸, 斎藤博昭. 日本語話し言葉における自己修復の統計モデル. 言語処理学会第 10 回年次大会発表論文集, pp. 2–7, 2004.

[3] 島森瑛貴, 阪本浩太郎, 渋木英潔, 森辰則. 自発的な独話における可読性向上のための言い直し表現を検出・修正するシステム. 言語処理学会第 28 回年次大会発表論文集, pp. 1739–1743, 2022.

[4] 内元清貴, 丸山岳彦, 高梨克也, 井佐原均. 『日本語話し言葉コーパス』における係り受け構造付与. 国立国語研究所平成 15 年度公開研究発表会予稿集, 2003.

[5] 国立国語研究所. 『日本語話し言葉コーパスの構築法』. 国立国語研究所報告 No.124, pp. 1–552, 2006.

[6] 島森瑛貴, 阪本浩太郎, 渋木英潔, 森辰則. 自発的な独話における可読性向上のための言い直し表現の定義と分析. 言語処理学会第 27 回年次大会発表論文集, pp. 365–370, 2021.

A　現在のシステムで修正できない具体的な例文

A.1　別語彙を用いた言い直し

別語彙で言い直すものを修正することができておらず、どのように類似性を判定するかを検討しているところである。

(1)　その日の朝食は**全て**パンを全部手作りで焼いたりとか
(2)　またスタイログロッサスは茎状突起より出て**せんじょうに至る**舌尖に至る細長い筋肉です
(3)　私はアイスホッケーを今年で**九年目**九年間やっておりまして
(4)　人間の**言語生活**言語の活動のうちの基本的なものをなすと思われます
(5)　やはり**ロシアが**ソ連邦が崩壊したからと言って

A.2　意図的な繰り返し

アクセントやモダリティなどの差異に言及するために意図的に繰り返すものと言い直しを区別することができない。

(1)　このような研究におきましては韻律の差異として「**行った**」「行った」のような基本周波数の変化が取り上げられており分節的特徴としては「行った」「行って」のような音節の変化が取り上げられてきました

A.3　言い直しではないが冗長な表現

冗長な表現や話題の提示を行う箇所を言い直しと区別することができていない。

(1)　機械を用いて可聴域に変換した後**パソコンで**分解能が十六ビットでパソコンに取り込みました
(2)　**それぞれの境界は**おとがい棘と舌骨上部を結ぶ線上をそれぞれの境界としました
(3)　**ハイオグロッサスについてですけれども**ハイオグロッサスは膜状の筋肉なのでハイオグロッサスの前部アンテリアとハイオグロッサスポステリアの二か所に分けて計測いたしました

A.4　2 文節以上-2 文節以上の規則における言い直し箇所の推定

2 つの要素で共通する部分を A、異なる部分をそれぞれ BC としたときに、言い直しの組が $\mathrm{BA}_{(1)}-\mathrm{CA}_{(2)}$ となるときに、$\mathrm{BA}_{(1)}$ を削除するもの (1)、$\mathrm{A}_{(1)}$ のみを削除するもの (2)、何も削除しないもの (3) を区別することができない。

(1)　例えば**よく**$_{(B)}$**見られる**$_{(A)}$日常的に一番頻繁に$_{(C)}$見られる$_{(A)}$ 距離がパーソナルディスタンスなんですけれども
(2)　これまでに行なってきました$_{(B)}$**カグラコウモリの超音波パルスの**$_{(A)}$
静止状態における$_{(C)}$カグラコウモリの超音波パルスについて$_{(A)}$ 説明いたします
(3)　経験的損失の値が最小となる$_{(B)}$ワーピングを$_{(A)}$最適な$_{(C)}$ワーピングとして$_{(A)}$ 選びます

A.5　削除後に語順の並び替え等の後処理が必要なもの

言い直し元を削除した後に語順を並び替えないと意図が変わるものが存在する。語順の変更等の後処理はまだ実現していない。

(1)　私の所属している神奈川県下では現在五チームの**エントリー**トーナメントなどがある時にはエントリーがあります
(削除)私の所属している神奈川県下では現在五チームのトーナメントなどがある時にはエントリーがあります
(並び替え)私の所属している神奈川県下では現在トーナメントなどがある時には五チームのエントリーがあります
(2)　五年生の時の担任の**先生が**生徒の筆順ていうのがあまりにもでたらめなので先生が黒板に大きく書いて
(削除)五年生の時の担任の生徒の筆順ていうのがあまりにもでたらめなので先生が黒板に大きく書いて
(並び替え)生徒の筆順ていうのがあまりにもでたらめなので五年生の時の担任の先生が黒板に大きく書いて

A.6　編集表現

編集表現を持つものには言い直し元を削除することで修正できるもの (1,2) や、編集表現を含む箇所のうち更に実際に言い直されている箇所を推定する必要があるもの (3)、言い直しではなく例示と説明と関係にあるもの (4)、比喩表現に対する対する具体的な心情の説明が続いているもの (5) などがあり、これらを適切に区別して言い直しかどうかを判定する必要がある。

(1)　でとにかく**信仰心て言うんですか**宗教に対しての信仰心が強くて
(2)　何百メートルも**交番じゃない**交差点がなくて
(3)　コンクリートとか木でできた**工場って言うか**港の風景みたいのが多いですね
(4)　**オフィス街と言うか**何か目立って大きな建物とか楽しめるところというのはなくて
(5)　ハワイはとてもいいとこで過ごし易いし南国だしでも**パラダイスじゃないなって言うか**向こうでは随分色んな人に会ってずっとお喋りして笑ってたりしたんですけれどやっぱ飛行機の中で一人になって何か自分で持ってた静かな音楽とか聞いてるとやっぱりあたしは何かちょっと寂しいようなとこの方が合うのかなっていう風に思って