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

島森瑛貴
横浜国立大学
simamori-eiki-vj@ynu.jp

阪本 浩太郎
横浜国立大学
sakamoto@forest.eis.ynu.ac.jp

渋木 英潔
国立情報学研究所
shib@nii.ac.jp

森 辰則
横浜国立大学
mori@forest.eis.ynu.ac.jp

1 はじめに

オンライン化が進む現代では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。そのような話し言葉の書き起こしは、編集されていない状態では「文章を冗長にする語の多用」、「倒置」、「文章のねじれ」などの話し言葉に特有の現象が出現する。その中でも、話し言葉の自発性に伴って出現する「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。我々はそれらの現象を適切に修正するシステムを作成し可読性の向上に貢献することを目標とする。

その手始めに本論文では日本語の自発的な独話に出現する言い直しを検出し修正するシステムの作成を目的として、言い直し現象を定義し直し機械が処理しやすい形での分類を行う。

2 関連研究

話し言葉の自己修復を扱う代表的なモデルとして RIM(Repair Interval Model)[1] が存在する。RIM は自己修復部を被修復部、言い淀み区間、修復部の 3 つの部分に分割し，これらがこの順に出現することで一つの自己修復を行っていると仮定している。

丸山ら [2] は RIM に基づき『日本語話し言葉コーパス』[3]（Corpus of Spontaneous Japanese 以下，CSJ とする）に出現する言い直し表現について言語学的な観点から分類した。分類結果は、発音エラーに伴う言い直し (R1)、単純な繰り返し (R2)、語彙的な誤りに伴う言い直し (R3)、情報不足に伴う言い直し (R4)、別表現への言い換え (R5) の 5 つに分類される。以下に示す分類例は (**[分類] 被修復部 | 言い淀み区間 | 修復部**) の形式となっている。

(R1)コンテキスト (R1 (D いぞ) | | 依存) モデルを
(R2)その年の (R2 七月| |七月) から
(R3)独特の (R3 旋律の|(F え)|旋律を) 形作っている
(R4)海外に (R4 興味| |元々興味) がありまして
(R5)(R5 こちら側のグラフ| |右側のグラフ) は何を表わしているかと言いますと

この分類は処理の対象が形態素未満、形態素以上単語未満、一文節、二文節以上など様々であるため機械的な処理にそのままこの基準を用いることは難しい。そのため工学的な観点から、処理の方法と対応する分類を行う必要がある。

話し言葉における言い直しの検出の先行研究は [4][5] などがある。

下岡らの手法 [4] は [6] で定義された言い直しのタグを正解として言い直しの検出を行っている。また推定された箇所について、係り受け情報を用いて削除する範囲の同定について検討している。しかし、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であるかを判定しており、特定のモデルを作成していない。また、[6] は「同一の内容を指し示している対等な文節」をのみを言い直しと定義して CSJ 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ていない。

藤井らの手法 [5] は自己修復部を分割したモデルを作成し、言い直しの検出修正を行っている。しかし、そのモデルはフィラーや言い淀みの存在を仮定しており、それらが存在しない言い直し表現については検出ができない。

本研究では初めに我々の考える言い直し現象を定義し、それに基づいた検出修正に取り組むための分析を行う。

3　定義

本研究において、我々は「言い直し」を「同一の対象に複数回参照する表現のうち、同一の面を参照しているもの」であると定義する。

この定義では [6] で定義された「言い直し」に加え以下のものも対象として扱う。言い直し表現の表記は以下 **(言い直し元/言い直し先)** とする。

1. 言い淀み
・聞き分けられていると (い/言われています)
2. 修正の意図を持つ発言を含むもの
・(苦手というか/苦手意識が) ある。
3. 間に挿入節を含むもの
・(ママが/{ママと言っても短パンとTシャツを着て髭を生やした凄いごつい男の人でどう見ても男なんですけれども}その人が)
4. 対象を一回以上指示語で参照しているもの[1)]
・タイミングと振幅比の (関係/こちらを)

以下の文のように「みかん」と「りんご」のような異なる対象を参照する場合や、同じ対象であっても「実績」と「グループ名」のように異なる面から参照し直している物は言い直しではないと捉え本研究の対象としない。

・色鮮やかな (織物や/工芸品が) 並んでいます。
・(トレーニング八十パーセント達成組/つまり先程のAグループ)

また、本研究では言い直し表現の組のうち、最後に言い終えている箇所を「言い直し先」と、それ以外の箇所を「言い直し元」と定義する。

以下の例文では「ハワイ島」「ハワイ島というか」「マウイ」の三つが言い直しの組を構成しており、「マウイ」を言い直し先、「ハワイ」、「ハワイ島というか」を言い直し元と考える。

・ただ、4泊6日で行ける (ハワイ島/ハワイ島というか/マウイ) もよかったんですけど

すべての言い直し表現の組は一つ以上の言い直し元と一つの言い直し先からなるとする。

我々は**言い直し先は言い直し元よりも後に発話されていることから重要であり残すべきである**という仮説を立てた。その仮説を元に言い直し現象の組のうち言い直し元と言い直し先のどちらを削除するか調査する。

4　言い直し表現組における削除箇所の調査

1)　CSJ では同格であると扱われている [6]

4.1　対象

以下の分析の対象として CSJ のコアのうち独話である講演、模擬講演の合計 177 講演を用いた。言い直し元が自立語を含む組は第一筆者がアノテーションしたデータを用いた。そのような言い直し表現は 1359 組得られた。前述した例のように一つの言い直し現象で二つ以上の言い直し元が存在するときはそれぞれの言い直し元と言い直し先の組を言い直し元の数だけ用意した。最終的に言い直し元と言い直し先の組は全部で 1453 件検出された。また、「助詞・助動詞・接辞・数字の言い直し」の形態素は D2 タグとして、「言い直し・言い淀み等による語断片」の形態素は D タグとして CSJ にアノテーションされている [7]。コア 177 講演中に D2 タグは 223 件、D タグは 2393 件存在する。これらは言い直し元が文節未満であるような言い直しの表現組の言い直し元であるとする。合計 4069 件について言い直し元と言い直し先のどちらを削除するかの分類を行う。削除箇所は第一著者一人が判断して頻度を調査する。表 1 に結果を示す。

4.2　結果

表 1　言い直し組と削除箇所

削除箇所	言い直し元	言い直し先	合計
文節未満	2616	0	2616
文節以上	1366	87	1453
合計	3982	87	4069

4.3　考察

言い直し元が語断片のときはそれらを削除することで修正する。また、助詞が言い直されているものは、一つの文節で最後以外の助詞を削除することで修正する。言い直し元が文節以上のものはおおよそ言い直し元を削除するが、言い直し先を削除するものが 87 件存在した。その 87 件のうち、84 件は言い直し先の自立語が指示語を含むものであった。また、言い直し元のみが指示語を含むが言い直し先を削除する例は存在しなかった。言い直し先を削除するもののうち、言い直し先が指示語でないものは以下の 3 件である。

(例 1)(百キロぐらいで/(あっちはマイルで出るんですけれども) マイルで) 飛ばすから
(例 2) この (安全保障理事会/安保理が) 最近非常に

(例 3) 傾向が我が社にも (ありました/ました)
(例 1) は文章の意味を理解する必要があり、高度な処理が求められるため今後の課題とする。(例 2) は正式名称と略語の関係であり、略語を削除することとする。(例 3) は付属語から繰り返されている場合であり、一文節中で重複する付属語を削除することで修正する。

5　システム設計方針の概要

本論文で提案するシステムは始めに言い直しの検出を行い、次に言い直しの修正を行う二つからなる。一文を入力とし、その修正を行った結果の文を出力とする。

5.1　言い直し表現候補検出の方針

検出部では、入力された文を係り受け解析して得られる文節の組について、任意の二文節が言い直し表現候補となるかを判断することで言い直し表現候補を検出する。またその時の音声認識や形態素解析の結果から、語断片や助詞の言い直しと判断されるものが見つかればそれらは全て上記の方針で削除する。下岡ら [4] は形態素の繰り返し情報が言い直しを検出する際の有効な素性であると述べている。本研究でも**言い直し元は言い直し先の自立語を含む**という仮説を立てる。

また言い直し組の検索範囲が広すぎると言い直しではない離れた文節を言い直しと誤って判断すると考えられる。そこで対象とする二文節間の最長距離を決めるために、**言い直し表現組の間に入る最長の文節数**を調査する。6 章では上記二つの調査を行う。

5.2　言い直し修正の方針

検出部では、検出部で検出された言い直し表現候補の組のうち、言い直し元または言い直し先のいずれかを削除することで修正を行う。言い直し元が形態素未満の語断片であるときは検出部で削除するため、修正部に入力される言い直し表現候補の組はどちらも文節単位である。4 章の分析から言い直し先の文節のみが指示語を含んでいるときは言い直し先を、それ以外の場合は言い直し元を削除することで修正できることが明らかになったため、そのまま言い直し修正の方針として用いる。また、削除箇所については**削除の対象となる箇所は一文節以下である**という仮説を立てて 7 章で調査する。

6　言い直し表現組の検出基準の調査

前章で立てた**言い直し元は言い直し先の自立語を含む**という仮説と、**言い直し表現は最大いくつの文節を間に含むか**の二つを調査する。使用するデータは言い直し元が文節以上であるような言い直し元と言い直し先の組 1453 組である。言い直し元、言い直し先、その間の挿入表現をそれぞれ形態素解析、係り受け解析し共通の自立語を持つか、また挿入節の文節数を調査した。それぞれの結果は表 2、表 3 に示す。

表 2　言い直し元が言い直し先の自立語を含むか

	含む	含まない
合計	1090	363

表 3　言い直し元と言い直し先の間に入る挿入節の文節数

文節数	1	2	3	4	5	6	7	8
出現回数	2	16	19	17	17	10	11	4
文節数	9	10	11	12	13	14	15	合計
出現回数	5	3	1	1	2	2	1	111

6.1　言い直し表現が間に含む文節数

言い直し元と言い直し先のの間には最大で 15 文節までが含まれることが判明した。該当する文は「ママが (ママと/言っても/短パンと/T シャツを/着て/髭を/生やした/凄い/ごつい/男の/人で/どう/見ても/男なんです/けれども) その人が」
このことから我々のシステムにおいては注目している文節から 15 文節以内を検索の対象とする。

6.2　共通の自立語を含むか

言い直し表現組 1453 件のうち 1090 件は言い直し元が言い直し先の自立語を含むことが判明した。言い直し先と言い直し元が共通の自立語を含まない 363 件の分類は以下のとおりである。

(1)類義語 87 例
　(見紛うような/見間違うような),(人の/人間の)
(2)指示語を含む 80 例
　(加盟国/これは),(これは/この数値は)
(3)編集表現を含む 97 例
　(ハワイと言うか/マウイ島),(日本じゃない/東京の)
(4)部分文字列が共通 20 例
　(サザン/サザンテラス),(半/半年)

(6)アルファベット、数字 3 例
(ＸＹ/Ｘという単語Ｙという単語が),(１９９９/１９９０年の)
(7)分類誤り 67 例

各分類ごとの処理方針は以下のとおりである。

(1)(4)(6) は表層系以外での要因が類似している。そこで、意味が似ている、音が似ている、部分的な文字列が似ているなどの自立語間の類似度が高いものが存在したときに言い直しとして検出する。これらは、分散表現から語の類似度を調べる手法や、文字単位に分解し共通の文字を含むかといった手法で類似度を調べる。特に数字や、アルファベットは形態素として扱うことが難しいため文字単位で扱うことが有効であると考える。

(2) の指示語を含むもの 80 例を分析した結果、(加盟国/これは) や (これ/この図は) のように付属語を持たない名詞 (指示語) の文節と指示語 (名詞) の文節が連接するとき、または (これは/この数値は) のように指示語と自立語が同じ付属語を持つときのどちらかであることが分かった。上の二つを指示語を含む言い直し検出の規則とする。

(3) は我々が編集表現であると判断した表現のリスト[2)]を作成し、当てはまるものを検出する。調査に用いたコーパス内に出現する編集表現は「と言うか」「じゃなくて」のどちらかを基本形に持つ物が多い。

(7) は形態素解析の誤りや言い直し元が複数ある場合で、実際には上記のいずれかの手法で検出できる。(ハワイ/ハワイと言うか/マウイ島) のように言い直し元が複数あるとき、調査では (ハワイ/マウイ島) の組を調べたため自立語を含まない。しかし入力の先頭から調べると (ハワイ/ハワイと言うか) が共通の自立語を持ち、(ハワイと言うか/マウイ島) は編集表現を持つため上記の規則に従って、言い直し表現の検出ができる。

上記の考察を元に本システムでは 15 文節以内にある任意の 2 つの文節のうち次のような条件を満たすもので最も距離が近い物を言い直しの候補とする。検出規則を優先度が高い順に次のようにする。

- 同一の自立語を含む
- 編集表現を持つ
- 指示語を含むときの規則に当てはまる
- 文節間の類似度で判断するときの規則に当てはまる

2) 実際の検出例は音便変化が多いため付録に掲載する

7 修正における削除箇所の文節長の調査

削除箇所が一文節であるという仮説を調査する。使用するデータは言い直し元が文節以上であるような言い直し元と言い直し先の組 1453 組である。調査結果を表 4 に示す。

表 4 削除箇所の長さ

削除箇所	1 文節以下	2 文節以上
言い直し元	1004	362
言い直し先	58	29
合計	1062	391

言い直し表現 1453 組のうち削除箇所が一文節のものは 1062 組存在した。これらは言い直し候補として検出されたのちに削除規則に従って文節単位で削除することで修正を実現できる。また削除箇所が二文節以上であるようなものが 391 件存在した。

以下の例文では共通の自立語「隣り」を持つ文節である「隣の」だけではなく「方」も削除する必要がある。係り受けの情報を用いると「方」は「隣の」に係られているが、「隣には」には係っていないため「隣の方」を削除するというように削除箇所を推定できると考えている。

・(隣りの方/隣りには) デパートは作らない

係り受け情報を用いた削除範囲の推定は今後の課題とする。

8 まとめ

本研究では話し言葉の可読性を向上させることを目的として、「同一の概念に複数回参照しているものの内、異なる面から参照しているもの」と言い直しを新たに定義した。その定義のもとで言い直しの検出修正の規則を作成するため CSJ の独話講演情報を分析した。今後の課題は自立語間の類似度を測る基準を設けること、また削除箇所が二文節以上であるときに削除箇所を適切に推定すること、これらの方針で実際にシステムを作成し性能を評価することなどがある。

参考文献

[1]Christine Nakatani and Julia B Hirschberg. A speech-first model for repair detection and correction. *31st Annual Meeting of the Association for Computational*

Linguistics, 13(1):46–53, 1993.

[2]佐野 信一郎 丸山 岳彦. 自発的な話言葉に現れる言い直し表現の機能的分析. **言語処理学会第 *13* 回年次大会**, 13:1026–1029.

[3]国立国語研究所. 『日本語話し言葉コーパスの構築法』. http://pj.ninjal.ac.jp/corpus_center/csj/, 2006.

[4]内元 清貴 井佐原 均 下岡 和也, 河原 達也. 『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*）**, 2005(50):95–100, 2005.

[5]斎藤 博昭 藤井 はつ音, 岡本 紘幸. 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**, pages 2–7, 2004.

[6]高梨 克也 井佐原 均 内元 清貴, 丸山 岳彦. 『日本語話し言葉コーパス』における係り受け構造付与. https://pj.ninjal.ac.jp/corpus_center/csj/manu-f/dependency.pdf, 13(1):1, 2006.

[7]間 淵 洋 子 小 磯 花 絵, 西 川 賢 哉. 『日 本 語 話 し 言 葉 コ ー パ ス の 構 築 法』2 章 転 記 テ キ ス ト. http://pj.ninjal.ac.jp/corpus_center/csj/k-report-f/02.pdf, 2006.

A　付録

A.1　編集表現のリスト

　言い直し表現に出現する編集表現について
出現形:出現回数の形で分類した表を以下に示す。

A.1.1　じゃない系統

　「じゃない」の派生は 5 種 14 回出現している。
じゃありませんね: 1
じゃない: 7
じゃないですね: 1
じゃないな: 1
じゃないや: 1
ではないですね: 1
ではないですが: 1
ってことはないですね: 1

A.1.2　と言うか系統

　「と言うか」の派生は 25 種 136 回出現している
言うか: 2
っちゅうか: 2
っつうんですか: 1
って: 1
って言いますか: 1
って言うか: 33
って言うかね: 1
って言うんすか:1
って言うんでしょうか: 2
って言うんですか: 9
って言うんですかね: 2
ってえか: 1
ってか: 3
て言うか:4
て言っていいんでしょうか: 1
て言うんですか: 1
て言うんでしょうかね: 1
と言いましたけど: 1
と言いますか: 15
と言うか: 45
というが: 1
と言うのか: 1
と言うよりは: 1
と言うんでしょうか: 5
と言うんですか: 1

A.1.3　その他

　その他は 2 種 4 回出現している
でしたかね: 1
ですか: 3