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

島森瑛貴 [1]　阪本浩太郎 [2] 渋木英潔 [2] 森辰則 [1]
[1] 横浜国立大学大学院 [2] 株式会社 BESNA 研究所
simamori-eiki-vj@ynu.jp sakamoto@besna.institute
shib@besna.institute 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、言い直しの検出と修正を行うシステムを構成した。はじめに分析に準拠したベースラインシステムを作成し、その評価を行った。実験の結果、適合率は 0.046、再現率は 0.300、F1 値は 0.080 であった。考察の結果、言い直しの範囲推定の方針と名詞連接の判定が不十分であったことが明らかとなった。これらの課題を改善し再び評価を行った。実験の結果、クローズドテストでの適合率は 0.276、再現率は 0.443、F1 値は 0.340、オープンテストでの適合率は 0.235、再現率は 0.400、F1 値は 0.296、と大幅に改善した。

1　はじめに

新型コロナウイルスの感染に伴い現在は生活の様々な場面でオンライン化が進んでいる。そのような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。話し言葉の書き起こしは、編集されていない状態では「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。我々はそれらの現象を適切に検出・修正するシステムを開発し、可読性の向上に貢献す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扱う代表的なモデルとして RIM(Repair Interval Model)[1] が存在する。RIM は言い直し部を被修復部、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言い直し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。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を検出する先行研究は [4][5] などがある。下岡らの手法 [4] は高梨ら [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、言い直しを捉える特定のモデルを作成していない。また、高梨ら [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] は言い直し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。

以上から日本語の自発的な発話に出現する言い直し全般について分析し、検出・修正するような研究はこれまで行われていない。そこで我々は島森ら [7] で言い直しを再定義し、機械的に扱える単位に対して分析を行った。

島森ら [7] では「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」であると定義した。また、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」と定義した。この 1 つ以上の言い直し元と 1 つの言い直し先から 1 つの言い直し組を構成する。本稿でもこの用語を使用する。さらに、この言い直しの定義に基づいて、日本語話し言葉コーパス CSJ[3] を用いて以下の 3 つの仮説を検証し、1 つの調査が行われている。

仮説 1. 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

仮説 2. 言い直し元は言い直し先の自立語を含む

仮説 3. 削除の対象となる箇所は一文節以下

調査 1. 言い直し表現組の間に入る最長の文節数

3 システムの構成

本稿では島森ら [7] の分析に基づき日本語話し言葉中に出現する言い直しを検出・修正するシステムを提案する。本研究における言い直しの検出・修正とは、話し言葉の書き起こし中に出現する言い直し表現の組を検出し、適切に削除することを指す。検出・修正の方針をそれぞれ以下に示す。

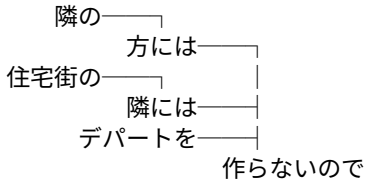
3.1 検出の方針

仮説 2 の検証結果によれば、言い直し元と言い直し先は共通の自立語を持つことが多いことが明らかとなった。そこで、ある自立語 A を含む文節 S において、同様の自立語 A を持つ文節 T が検索範囲内に存在すれば (S,T) の組を言い直し組とする。調査 1 の結果によれば言い直し組は 15 文節以上離れないため、検索範囲を 15 文節とする。

3.2 修正の方針

言い直しとして検出した組は、言い直し元か言い直し先のいずれか、もしくはその両方を削除することで修正する。仮説 1 の検証結果によれば基本的には言い直し元を削除することで修正ができると明らかになったため、それを修正の基本方針とする。しかし仮説 3 から 2 文節以上を削除することもあると判明した。従って、修正にあたっては削除範囲 (多くの場合では言い直し元の範囲) を正確に推定することが必要である。そこで**言い直し組 (S,T) の候補において、言い直し元とされた文節 S を起点として、実際の言い直し元の範囲を推定する。すなわち、係り受け木の部分木のうち、i) 文節 S を含み、かつ、ii) 言い直し先とされた文節 T 以降の文節に係るものを言い直し元とする**という方針で修正できるという仮説を立てて検証する。「隣の方には住宅街の隣にはデパートを作らないので」という例を用いてこの方針について説明する。

係り受け解析器 CaboCha を用いてこの文を係り受け解析すると以下の結果が得られる。



この例では「隣の方には」が言い直し元、「住宅街の隣には」が言い直し先とすることが正解であり、「住宅街の隣にはデパートを作らないので」と修正されるべきである。上記の方針を適用すると、「隣の」が S、「隣には」が T となる。S を含む係り受けの部分木のうち、T 以降の文節に係るものは「隣の方には」となる。よって、これが範囲の推定を行った後の言い直し元となるので、これを削除すると正しく修正できる。なお、この例では「住宅街の隣の方にはデパートを作らないので」のように、言い直し組の文節を融合して修正することもできるが本研究ではそのような修正は行わない。

4 言い直し元の範囲を推定する仮説についての検討

システム全体の評価前に上記仮説を検証する。

4.1 対象・結果

日本語話し言葉コーパス [3] のコア講演 177 講演に出現した言い直し組全 1453 件を分類したコーパス (島森ら [7] で作成) を用いる。その中の言い直し元または言い直し先が 2 文節以上である 617 件を目視で調査した。分析の結果、仮説に基づいて推定できたものが 303 件、推定できなかったものが 314 件だと明らかになった。

4.2 考察

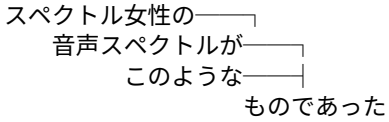
仮説の方針に従って正しく範囲を推定できなかったものの内訳は以下のとおりである。

(1)名詞が連続する　180 件
(2)範囲推定後の言い直し元の範囲以前から言い直し元が始まる 23 件
(3)編集表現を持つ 97 件
(4)言い直し先の文節に係る 14 件

上記 (1) に当てはまる文を以下に示す。

スペクトル女性の音声スペクトルがこのようなものであった

これは、既存の係り受け解析器が名詞を言いかけた言い直しを正しく解析できていないことに起因する。この文を CaboCha で解析した例を以下に示す。



この例では「スペクトル女性の」が一文節だと判断され、「スペクトル」だけ削除されるべきところ「スペクトル女性の」が削除される。そこで、これらを異なる文節として扱う必要がある。次の 5 章で評価するシステムでは対象とせず、5 章の考察で明らかとなった他の問題と共に 6 章で係り受け構造を変換する前処理を行い、7 章で評価する。

上記 (2) に当てはまる文を以下に示す。

たまに見るいつも見る犬がいる

この例では「たまに見る」が削除され「いつも見る犬がいる」と修正されるべきである。しかし言い直し元として検出されるのは初めに出現する「見る」であり、言い直し元以前の文節は推定範囲外であるため「たまに」が削除されない。仮にこの例文が「私が見るいつも見る犬がいる」だとすると、初めの「見る」だけを削除するのが正しい。このように、言い直し元より前の文節が言い直しに含まれるかを判断するタスクは、係り受けの情報だけでなく文の意味を理解する必要があるため非常に難しい。そのため、本稿ではこの例は扱わない。

上記 (3) に当てはまる文を以下に示す。

愛していますと言うか好きな

この例では編集表現「と言うか」を持つ「愛していますと言うか」を削除し「好きな」と修正するべきである。この編集表現は言い直し先の文節「好きな」に係るため仮説は成り立たない。しかし以前の調査から編集表現を含む文節はその文節を削除すると修正できることが明らかとなったため、編集表現を持つ文節があるときはその文節を削除するという規則をシステムに追加する。

上記 (4) に当てはまる文を以下に示す。

解説の意味の解説者の解説の意味の

この例では初めの「解説の意味の」を削除し「解説者の解説の意味の」と修正するべきである。しかし、言い直し元から係る一連の文節「解説の意味の」が、言い直し先である「解説の意味の」に係るため仮説が成り立たない。このように言い直し元の文節が言い直し先に係るようなときは他の規則を用いて範囲を推定する必要がある。これは今後の課題とし、本稿では扱わない。

5 システムの評価

4 章で検証した範囲推定の方針をシステムに利用し、3 章で提案したシステムを評価する。

5.1 対象・結果

分析の対象には CSJ のコア講演 177 講演を用いる。CSJ によって文末タグが付与されている箇所で文単位に分割した。得られた文数は 9261 件、言い直し組は 1705 件であった。上記の分析と同じ講演情報を使用したが言い直し組の数が増加している。これはデータを分析し直したときに、コーパス作成時に見落としていたものを追加したためで、言い直しの定義は変化していない。

システムの評価は、システムが削除すると判断した個所に IOB2 形式のラベルを文字単位で振り、削除単位ごとに一致しているかを調べ適合率、再現率、F1 値を求めた。評価の結果によると、適合率は 0.046、再現率は 0.300、F1 値は 0.080 であった。

5.2 考察

本システムを評価した結果によると適合率が非常に低いことが明らかとなった。適合率低下の大きな要因には不適切な検索範囲と名詞連接の 2 つがあった。以下にそれぞれ修正を誤った例を示し考察する。例では、システムが削除した個所を<>で囲う。

5.2.1 不適切な検索範囲

誤検出が増えた原因として言い直し候補の検索範囲に文節数を利用したことが挙げられる。

取り敢えず失業保険も<ありますし>七か月ぐらいは<収入が>あるんでねしばらくは親の収入に頼らなくても大丈夫です。

上記の例には本来言い直しと判断されるべきものはない。しかし、「ある」と「収入」といった自立語が共通している文節が言い直しとして判断されている。削除された文節はそれぞれ「失業保険もある」「収入がある」「収入に頼る」と異なる節を構成しているため、言い直しでないと人が判断することは容易である。そこでシステムが同様の構造を理解することで誤検出を減らせると考えた。そこで 6 章で、検索範囲として節境界を利用することを提案する。

5.2.2 名詞連接

名詞連接とは名詞が続くことを指す。言い直しが発生すると、上述した「スペクトル女性の」のように意図しない名詞連接が生じ、正しく構文解析が行われないことが多い。また以下のように、複合語として扱いたいものとの区別も難しい。

<被験者情報は>被験者から個別で聴取した

上記の例には本来言い直しと判断されるべきものはない。しかし「被験者情報は」と「被験者から」が言い直しとして判断されている。「被験者情報」を複合語と判断できれば誤検出の減少を期待できる。そこで 6 章では、名詞連接が、言い直しによる意図しないものか、複合語かを正しく判断し、言い直しであれば正しい係り受け構造に変更する前処理について検討する。

6 システムの改善

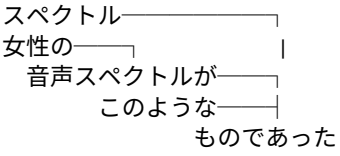
5 章の考察をもとに提案システムを改善する。

6.1 節境界の利用

同一節内のみを対象とするように言い直し組の検索範囲を変更する。節境界の推定には日本語節境界推定プログラム CBAP[8] を利用する。

6.2 係り受け構造の修正を行う前処理

名詞連接を正しく扱うために、システムが言い直しを検出・修正する前に係り受け構造を変更するような前処理を行う。文中に出現する名詞連接は、複合名詞の場合と、「スペクトル女性」のように名詞のみの言い直しの後に名詞から始まる文節が続く場合がある。本研究では後者を積極的に言い直しとして認識したい。そこで、あらかじめ複合名詞の候補を見つけ言い直しの対象から外し、それ以外の名詞連接について言い直しの可能性を確認する。具体的には名詞連接のうち、同一講演で複数登場するものは、複合名詞として不可分な一つの名詞として扱う。その後、複合名詞でない名詞連接の係り受けを変更する。具体的には名詞 A(A1) と名詞 B の連接があり、AB と同一節内で A を持つ他の文節 (A2) が存在したとき、A1 の係り先を A2 の係り先に合わせる。上記の「スペクトル女性の音声スペクトルがこのようなものであった」の例では「スペクトル (A1) 女性 (B)」という連接に対して「音声スペクトル (A2) が」という文節が同一節内に存在するため以下のように係り受けを変更する。



7 改良したシステムの評価

7.1 対象・結果

改良したシステムを評価する。評価方法はベースラインシステムの時と同様である。クローズドテストではベースラインシステムの評価と同じデータを用いる。オープンテストでは非コア講演 40 講演を用いる。5 章と同様の文分割を行い、文数は 1996 件、言い直し組は 297 件であった。

クローズドテストの結果は表 1 に、オープンテストの結果は表 2 に示す。

表 1 クローズドテスト

	precision	racall	f1
ベースライン (BL)	0.046	0.300	0.080
BL+節境界+前処理	0.276	0.443	0.340

表 2 オープンテスト

範囲推定の手法	precision	racall	f1
ベースライン (BL)	0.035	0.275	0.063
BL+節境界	0.181	0.386	0.246
BL+前処理	0.044	0.325	0.077
BL+節境界+前処理	0.235	0.400	0.296

7.2 考察

節境界と前処理を利用することで精度が向上した。しかし改善後でもうまく修正できなかったものも存在する。例えば CBAP が口語的な節境界を正しく判断できていないものがあった。新たに節境界の規則を追加することを検討している。他にも「リンゴを食べ食べました」のような動詞の言い直しは修正できていなかった。動詞の後には何らかの節境界が付与されることが多く、検索範囲から外れるためである。また、講演中に一度しか出現しない複合語も多数存在しており、それらは言い直しによる意図しない名詞連接と同列に扱われていた。そのため、名詞連接の判定も不十分で今後の課題である。

8 まとめ

本稿では日本語話し言葉の言い直し現象を検出、修正するシステムを提案し、その評価を行った。上記の課題に加えて、異なる語彙を用いた言い直しを判定するために類似性を判定することや、言い直しの範囲をより正確に推定するために並列構造を組み込むことが今後の課題である。

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**, Vol. 13, No. 1, pp. 46–53, 1993.

[2] 丸山岳彦, 佐野信一郎. 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 回年次大会発表論文集, pp. 1026–1029, 2007.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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, 2020.

[8] 丸山岳彦, 柏岡秀紀, 熊野正, 田中英輝. 節境界自動検出ルールの作成と評価. 言語処理学会第 9 回年次大会発表論文集, pp. 517–520, 2003.