日本語話し言葉における自己修復の統計モデル

藤井なつ音　岡本紘幸　斎藤博昭
慶應義塾大学大学院 理工学研究科 開放環境科学専攻
Email: {fujii,motch,hxs}@nak.ics.keio.ac.jp

1　はじめに

「話し言葉」研究が近年まで少なかった理由の 1 つに，その文法的な不適格性 (ill-formedness) がある．中でも，発話者が自身の発話誤りを訂正/補足する際に発生する冗長表現「自己修復 (self-correction)」は特に問題とされ，幾つかの先行研究がある [2]-[5]．これらの研究には次のような問題があった．

- 自己修復捕捉モデルに起因した，文法適格性回復処理 (以下，回復処理) での情報欠落
- 品詞や語句の大幅な制限
- ルールに基づくアドホックな回復処理

本稿では自己修復の捕捉とその回復処理のための新しい手法を提案し，CSJ コーパス [1](モニタ版 2002) を用いた評価実験の結果を示す．

2　自己修復の定義と先行研究

2.1　対象とする自己修復

自己修復の多くは，繰り返し語とその間に未知語・孤立語を伴うことが指摘されている [2]．一方，船越らは自己修復をその機能から次の 3 種に分類した [3]．

1. 言い足し (繰り返しも含む)
2. 言い直し
3. リスタート

本稿では次の条件を満たすものを自己修復あるいはその候補として扱う．

- 繰り返し表現とその間の言い淀みを伴う
- 言い足し/言い直しが目的

ここで，言い淀みとは未知語・孤立語・フィラー・編集表現 (「すみません」/「いえ」等) を指す．

2.2　先行研究

先行研究の多くは RIM(Repair Interval Model)[5] に基づく．RIM は自己修復部を ReParanDum Interval(以下 RPD)，DisFluency Interval(以下 DF)，RePair Interval (以下 RP) の 3 単位に分割し，これらが下記のように順次連続して出現すると仮定する．

. . . RPD DF RP. . .

ここで，RPD は被修復部，RP は修復部，DF は言い淀み区間に相当する．この RIM ではまず DF を決定し，その後，DF の始端を RPD の終端に，DF の終端を RP の始端として各単位を決定する．捕捉例を以下に示す．

[そばが]$_{\mathrm{RPD}}$ [す えーと う]$_{\mathrm{DF}}$ [うどんが]$_{\mathrm{RP}}$ 好きで

捕捉後の解消処理は RPD を RP で置換し，同時に DF を削除する．この結果，最終的には RP のみが残される．この手法の問題点を以下に示す．

1. 自己修復の始終端の自動検出アルゴリズムが未提示．
2. RPD 内部に必要な情報がある文に対応不可能．

2 の文例を以下に示す．ここでは，必要な情報である「食べた」や「彼と」が削除されてしまう．

a. [そばを食べた]$_{\mathrm{RPD}}$ [いや]$_{\mathrm{DF}}$ [うどんを]$_{\mathrm{RP}}$

b. [そばを食べて 彼と]$_{\mathrm{RPD}}$ [え]$_{\mathrm{DF}}$ [そばを食べて]$_{\mathrm{RP}}$

これらの問題に対し，船越らの手法 [3] は自己修復範囲を漸進的構文解析によって決定する．しかし，RPD 内部の保持すべき語句は動詞のみと仮定するため，文 **b** のタイプは扱えない．また，ドメインを限定した上で全単語情報を予め辞書登録しているという問題点もある．

3　提案手法

提案手法は次の特徴を持つ．

1. 自己修復部の語句を文節を基本とする 6 単位に分割．品詞制限は無い．
2. 回復処理は各単位の削除に加え 4 パターンの移動も考慮．

1 によって，従来法が画一的に捕捉していた被修復部分 RPD はより細かい粒度で捕捉され，必要な語句が適切に残される．2 は被修復部削除後の可読性や係り受け整合性を維持するため，適宜語順変更を許すものである．以下，各特徴について更に詳しく説明する．

3.1　提案する捕捉モデル

提案モデルは自己修復部を図 1 の手順で以下のように捕捉する．

. . . **A R1 X D B R2**. . .

手順 1 言い淀み **D** を検出.
手順 2 **D** 前後の繰り返し形態素の第 n 組 $(\mathrm{r1},\mathrm{r2})_{\mathrm{n}}$ を検出.
手順 3 $(\mathrm{r1},\mathrm{r2})_{\mathrm{n}}$ をそれぞれ起点とする文節 (**R1**，**R2**)$_{\mathrm{n}}$ を取得.
手順 4 **R1**-**D** 間の全形態素 (0 個以上) で **X** 形成.
手順 5 **D**-**R2** 間の全形態素 (0 個以上) で **B** 形成.
手順 6 **R1** 以前にある **R1** と共に削除すべき形態素 (0 個以上) で **A** 形成.

図 1: 6 単位の取得手順

以下，一部の手順について追記する．

手順 1

D の構成要素は 2.1 で述べたとおりである．本稿では単語断片やフィラーの検出自体は議論せず，CSJ に付与された (D/D2/F) タグの情報をそのまま利用する．ただし，編集表現は，CSJ では (F) 語に最も近いと考えられるが付与例は無い．本稿ではコーパスの観察から，言い淀みに隣接するものは **D** に含めるのが妥当と判断した．

手順 2

繰り返し形態素の組は複数存在する場合がある．手順 2 ではそれら $(\mathrm{r1},\mathrm{r2})_{\mathrm{n}}$ $(\mathrm{n}=1,2,\dots)$ を全て求める．ただし，通常の発話でも出現回数が多い助動詞/助詞/非自立語としての名詞 (「の」「もの」など) の 3 品詞は除く．

手順 3

$\mathrm{r1_n}$，$\mathrm{r2_n}$ をそれぞれ文節 $\mathbf{R1_n}$，$\mathbf{R2_n}$ の起点とし，次のいずれかの語の直前をその終端とする．

- 繰り返し形態素 $\mathrm{r1_n}$，$\mathrm{r2_n}$ でない自立語
- 自立語に後続する $\mathrm{r1_n}$，$\mathrm{r2_n}$ でない非自立語 (助詞/助動詞/非自立語としての名詞)

$\mathbf{R_n}$ と $\mathbf{R_{n+1}}$ が隣接する場合は，これらを連結して新たに $\mathbf{R_n}$ とする．ただし，$\mathbf{R1_n\,R1_{n+1}}$ と $\mathbf{R2_n\,R2_{n+1}}$ の一方の組のみが隣接の場合，および r の順序が異なる場合はどちらも連結しない．連結処理後に $n \geq 2$ となる場合，最も **D** に近いものとそれに対応するもののみを，自己修復に起因する繰り返し **R1**，**R2** の候補とする．なお，文節探索には次のような例外処置を設けた．

- $\mathrm{r_n}$ が接頭辞および数字の場合は，以降で初出の非自立語まで探索.
- サ変接続の名詞に後続する，サ変動詞「する」/サ行五段動詞「致す」 は非自立語扱い.
- $\mathrm{r_n}$ が，接尾語「的」/サ変動詞「する」/接尾動詞「(ら) れる」の場合は，以降初出の自立語の直前まで探索.

手順 6

R1 は繰り返し形態素を起点とする．しかし，被修復部の起点が必ずしもこれに一致するとは限らない．**A** はこのような場合のための単位で，修復部の **B** に概ね相当する．以上の手順で求めた 6 単位の例を図 2 に示す.

1. 大きさなど [色んな]$_{\mathrm{R1}}$ [だいて]$_{\mathrm{D}}$ [色んな点において]$_{\mathrm{R2}}$
2. [理事長が]$_{\mathrm{R1}}$ [えー]$_{\mathrm{D}}$ [初代の]$_{\mathrm{B}}$ [理事長が]$_{\mathrm{R2}}$
3. [お客]$_{\mathrm{R1}}$ [たくさんの]$_{\mathrm{X}}$ [えーん]$_{\mathrm{D}}$ [見物の]$_{\mathrm{B}}$ [お客さんの中で]$_{\mathrm{R2}}$
4. [時間的変動を]$_{\mathrm{R1}}$ [受けにくい]$_{\mathrm{X}}$[あ]$_{\mathrm{D}}$[時間的変動の]$_{\mathrm{R2}}$ 影響を受けにくい
5. [このような]$_{\mathrm{A}}$ [HMM]$_{\mathrm{R1}}$ [このスカラー]$_{\mathrm{X}}$ [あのーあ]$_{\mathrm{D}}$ [非同期遷移型]$_{\mathrm{B}}$ [HMM の]$_{\mathrm{R2}}$
6. [個数が]$_{\mathrm{R1}}$ [約]$_{\mathrm{X}}$ [えー]$_{\mathrm{D}}$ [個数が]$_{\mathrm{R2}}$ 六万

図 2: 6 単位の取得例

3.2 構成単位の処理方法

回復処理では，**A**，**R1**，**D** を被修復部および不要語として必ず削除し，**B**，**R2** は修復部として必ず留保する．**X** については詳説を要する．従来法は次のいずれかであった.

- **R1** と共に無条件で削除.
- 品詞を動詞に限定した上で，**R1** 除去後，係り受け整合性を保持するために文末へ移動.

これらの方針の問題点は，2.2 の文例 **b** や，図 2 の文例 3，5 のタイプに対応できないことである．これに対して，提案手法は **X** の品詞を制限せず，さらに 4 パターンの移動を考慮する．6 単位の処理法を図 3 にまとめる.

原文		**AR1X**	**D**	**B**		**R2**		
処理後	自己修復でない	**AR1X**		**B**		**R2**		(**D** 消去のみ)
	パターン 1	**X**		**B**		**R2**		(**X** 移動無し)
	パターン 2			**B**	**X**	**R2**		(**X** が **B** の後へ)
	パターン 3			**B**		**R2**	**X**	(**X** が **R2** の後へ)
	パターン 4			**B**		**R2**		(**X** 削除)

図 3: 6 単位の処理法

図 2 の例文 (1)-(3)，(4) がパターン 4 に，(5) がパターン 1 に，(6) がパターン 3 に，それぞれ該当する.

本稿では自己修復か否か，また自己修復である場合には図 3 のどのパターンに該当するかを，データスパースネスに頑健で最も高性能な分類器の 1 つである SUPPORT VECTOR MACHINE(以下，SVM) で判定する．したがって，全過程では合計 5 つの SVM を用いる.

4 実験と考察

4.1 実験データ準備

書き起こしテキストには CSJ の「基本形部」を用いた．タグ情報の利用については表 1 を参照されたい．なお，**(?)** の候補は，その後の形態素解析において明らかな誤りを引き起こす場合には削除する．これは，**(?)** が発音にのみ忠実であり，言語的整合性を考慮しないためである．すなわち，候補を残すことによって本来必要な

語が補われる場合もあるが，逆に，明瞭に発音されていれば **(D/D2)** が付与されていたと判断できる場合もある．前述のとおり，本稿では **(D/D2)** の検出法を議論しないため，後者の場合には候補を削除する．タグ処理後は，文献 [6][7] に基づいて擬似文単位を作成し，茶筅 [8] による形態素解析後，提案モデルにより自己修復候補文を抽出した．これを人手で分類した結果を表 2 に示す．また，候補文の始終端は言い淀み/文末とした．

表 1: 本研究で使用した CSJ のタグ一覧

CSJ 内の表記	主な意味	実験データのための処理
(F)	フィラー	全て「あのー」に置換
(D)，(D2)	単語断片	「未知語 D」として残す
(?)	聞き取りに自信無し	候補を残す (複数の場合は 1 つ目)
(M)	引用	「 」で始終端を囲む
(R)	伏せ字対象語	人名 → 鈴木，タイトル →「タイトル」
(O)	外国語や古語など	原語のみ残す
(A)	基本形に漢字仮名以外の文字	その文字 (「;」の右側) を残す
(K)	漢字表記不能	その漢字 (「;」の右側) を残す
(S)	未登録の口語表現	未出
<FV>	母音不確定音	タグ除去
<C>	短単位が複数の転記基本単位に跨る	1 短単位に修正

表 2: 文例パターン一覧

自己修復でない	3,281
パターン 1	1,055 (内 **X** 有り 68)
パターン 2	1
パターン 3	9
パターン 4	137
合計	4,473

これによると，パターン 2/3 の文例は極めて少なく有意な実験が期待できないことから，本稿では次の 2 件について実験する．

1. 自己修復であるか判定する SVM**-1**
2. パターン 1 か 4 か (**X** を削除するか) を判定する SVM**-2**

本来書き言葉を想定して構築された茶筅は，話し言葉の解析に高い精度を持たない．そのため，本実験ではデータ準備の段階で茶筅解析結果を修正した．主な修正点を以下に示す．

- 文末に置かれる引用の格助詞「と」← 接続助詞「と」
- 接頭辞，接尾辞，非自立語 (今，風/点/的/語/波/度/場，の/もの)
- 種々の解析誤り (場合 ← 場+合/ですね ← で+すね/ 並立助詞「とか」← と+か/助動詞「ます」← 名詞「ます」/ 助動詞「ない」← 形容詞「ない」)

また，表 3 の口語表現については修正候補が無いためデータから除外した．

以上の手順で得た候補文に 4.2 で述べる素性値を与えて素性ベクトル群を作成した．次いで，正/負事例ベクトルに +1/−1 とラベル付けし，それぞれ学習用とテスト用に分割して実験データを作成した．

表 3: 修正候補が無い口語表現

「易い」(CSJ の書き起こし上の問題)
自立語としての「上」
自立語としての「以上」
接続詞「から」
接続詞「そいで」
接続詞「そいから」
接続詞的「ですんで」
接続詞「んで」
接続詞「あと」
接続詞「で」
助詞-格助詞-連語「っていう」の変形「つう」
およびその変形「ったら」「って」「っつのは」
格助詞の連語「とかって」
終助詞「かしら」は直後が「?」でないと解析不能
その他表現「ですけども」「だけども」は品詞分解不可能

4.2 実験で使用した素性

素性の種類を図 4 に示す．素性値は各素性の有無を表す 1/0 を与えた．一部の素性について以下に追記する．

- a 各単位の品詞 (茶筅の解析において深さ 3 までの品詞分類) の並び
- b 各単位の末尾の品詞 (深さ 3)
- c 各単位の長さ．最大長で [0，10] に正規化．**A**，**X**，**B** は NULL をとり得る．
- d 各単位の直前の品詞 (深さ 3)
- e **D** 以外の 5 要素の **D** からの距離
- f-k 各単位の文字種割合 (平仮名，カタカナ，アルファベット，数字，記号，漢字の 6 種)
- l **R1**，**X**，**B** の何れかの内部の文末表現の有無
- m **R1**，**X**，**B** の何れかの内部の提題表現の有無
- n **X**，**B** の何れかの末尾の文末表現の有無
- o **X** の有無
- p **R1** と **R2** で異なる形態素が品詞深さ 1(緩) でなら一致する割合
- q **R1** と **R2** の隣接形態素の組が基本形・品詞深さ 1(厳) で一致する割合
- r **R1** と **R2** の直前同士が品詞深さ 1(緩) で一致するか
- s **R1** と **R2** の末尾同士が品詞深さ 1(緩) で一致するか
- t **R1** と **R2** の直後同士が品詞深さ 1(緩) で一致するか
- u **R1** の末尾と **R2** の直後が品詞深さ 1(緩) で一致するか
- v **R1** の直後と **R2** の末尾が品詞深さ 1(緩) で一致するか
- w **R1** と **R2** の長さの差
- x **R1** と **R2** の **D** からの距離の差
- y 長さについて **R1**<**R2**∩**R1**<(**R2**+**B**) を満たすか
- z **R2** と一致しない **R1** 内の形態素の並びが **R2** 後方に出現するか

図 4: 本研究で用いた素性一覧

素性 l

文末表現は，(接続助詞/引用の格助詞「と」/名詞/活用語連用形) +(接続詞/話し言葉における接続詞「後」) および終助詞の 2 種である．

素性 m

提題表現は，名詞 +(「は」/「って」/「ったら」/「なら」) である．

素性 n

文末表現は，接続助詞/引用の格助詞「と」/名詞 (直後が接尾語でない)/連用形/終助詞の 5 種とする．

表中の「(厳)」は品詞深さ 1 での細分類を含むこと，「(緩)」は含まないことを表す．

4.3 実験結果

実験結果を表 5，6 に示す．ただし，この実験は表 4 に示す条件で行い，SVM の実装は TinySVM[9] を用いた．

表 4: 実験条件

検定数	3 点交叉
素性閾値	2
多項式カーネルの次数 d	$d = 1, 2, 3, 4$ のうち実験による最適値
エラーへのペナルティ C	$C = 0.001, 0.01, 0.1, 1.0$ のうち実験による最適値
SVM-**1**	全 4,473 文例 全素性 **a-z**(計 9,711 次元) 正事例: 自己修復文 / 負事例: 非自己修復文
SVM-**2**	205 文例 (**X** を含む自己修復文の数) 素性 **a-k(計 1,270 次元)** 正事例: **X** 削除文 / 負事例: **X** 非削除文

表中の Acc は全体の精度，Pre は再現率，Rec は適合率であり，F は次式で求めた．

$$F = \frac{2 \times Pre \times Rec}{Pre + Rec} \tag{1}$$

また，表 5 には，本手法と同様に繰り返しと言い淀みから自己修復を検出し，定量的評価を行った唯一の先行研究である佐川らの結果 [2] を併記した．

表 5: SVM-**1** の実験結果 ($d = 2$, $C = 0.001$)

	Acc	Pre	Rec1	F
提案手法	87.61	78.00	74.58	76.23
佐川らの手法	-	71.30	-	-

表 6: SVM-**2** の実験結果 ($d = 4$, C は全ての場合)

Acc	Pre	Rec	F
70.27	75.84	81.77	78.63

提案法の検出処理の有効性は表 5 より明らかである．一方，表 6 については着眼点を明らかにする．提案法の重要な特徴は，従来法では欠落した情報 (**X**) が保持できることである．その有効性は Pre 値で検証される．Pre は，本手法が **X** を削除すべきと判定した文例の内，実際に削除すべきであったものの割合を示す．すなわち，表 6 は従来法では全て削除していた **X** の 75% 以上を保持できたことを示す．以上より，提案法の有効性が確認された．

4.4 対象外の文例

本実験で，提案法による対処が困難な以下の事例が明らかになった．

例 1 左右聴覚野 えー 左右 え 左右聴覚野 え に

例 2 全ての場合に 母音の「あ」を 全てのモーラに「あ」を 持っ 母音の「あ」を 持っている訳です

例 3 矩形波に近いような [音]$_{R1}$[ん]$_{D}$[を]$_{B}$[音を]$_{R2}$

例 4[ホールの]$_{R1}$[せお]$_{D}$[音響]$_{B}$[ホールの]$_{R2}$ 音響を 考える

例 5[分散の]$_{R1}$[えー]$_{D}$[最大のものを]$_{B(消)}$
[出現頻度の]$_{B(残)}$[分散]$_{R1}$ 最大のものを取りまして

例 1 は言い淀みと前後の繰り返し表現を 2 箇所以上含むものである．これは先行研究の知見にはなく，本稿でもそれと同様に 1 箇所を含む場合しか想定していない．対処法としては繰り返し語の探索範囲の拡大が考えられる．しかし，コーパスの観察では **D-R**/**R1-R2** の距離が広いほど非自己修復文となることが多い．提案法のパターン別にみた文長 (形態素数) でも，全文例の平均 12.9 を上回ったのは非自己修復文 (14.5) のみである．したがって，単純な範囲拡大ではなく何らかの停止条件を設けた漸進的探索が有効と考えられる．また，例 2 のような複雑な構造は特殊なモデルを要する．さらに，**B** を削除すべき例 3-5 が明らかになった．例 3 は **B** が非自立語の場合のルールを与えれば良い．例 4 は **B-R2** の構文解析が有効と考えられる．例 5 は本稿対象外の「言い淀みを持たない自己修復」を含んでおり，そちらの枠組みと併せた処理が適切と考えられる．

5 むすび

本稿では，日本語話し言葉における自己修復の捕捉と，その文法適格性回復処理のための新しい手法を提案した．評価実験ではその有効性を示し，これまで指摘されなかった自己修復の新しい構造も明らかにした．今後は，より多くのデータによる検証や，新たな構造に対するモデルの検討を目指す．

謝辞

CSJ，茶筌，TinySVM の構築にあたられた方々に感謝いたします．

参考文献

[1] K. Maekawa, H. Koiso, S. Furui and H. Isahara;“Spontaneous speech corpus of Japanese,” In Proceedings of the Second International Conference of Language Resources and Evaluation (LREC2000), Athens, 947-952, 2000.

[2] 佐川雄二, 大西昇, 杉江昇; “大規模コーパスに基づいた日本語自己修復文の分析,” 情報処理学会研究会報告, NL-100-10, pp.73-80, 1994.

[3] 船越孝太郎, 徳永健伸, 田中穂積; “音声対話システムにおける日本語自己修復の処理,” 自然言語処理, Vol.10, No.4, pp.33-53, 2003-7.

[4] W. Levelt; “Monitoring and self-repair in speech,” Cognition, Vol.14, pp.41-104, 1983.

[5] C. Nakatani and J. Hirschberg; “A speech-first model for repair identification and correction,” in Proceedings of 31st Annual Meeting of ACL, pp. 200-207, 1993.

[6] 高梨克哉, 丸山岳彦, 内元清貴, 井佐原均; “話し言葉の文境界-CSJ コーパスにおける文境界の定義と半自動認定,” 言語処理学会第 9 回年次大会, 2003-3.

[7] 太田公子, 高梨克哉, 井佐原均; “話し言葉における接続詞「で」の特徴分析,” 言語処理学会第 9 回年次大会, 2003-3.

[8] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原 正幸; “日本語形態素解析システム『茶筌』 version 2.3.2 使用説明書,” 2003.

[9] Taku Kudo; “TinySVM,” http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM, 2001.