2T-02

言語学的な単位に応じた言い間違いの検出

小松 聖矢 †　　篠山 学 †
香川高等専門学校 †

1 はじめに

近年，Amazon Echo[1] や Google Home[2] などの音声をインタフェースとした対話システムが増えている．しかし，対話システムは人間と円滑なコミュニケーションを取れているとは言い難い．その理由のひとつに，対話システムは人間の言い間違いをそのまま認識しようとして，誤った応答をしてしまうことが挙げられる．そこで本研究では，対話システムに人間の言い間違いを検出し，人間の意図通りに認識したり，言い間違いを指摘するシステムの開発を目指している [3]．現在，言い間違いの検出には深層学習を用いているが，学習に必要な言い間違いのコーパスが少ないことから，実用レベルの検出精度ではない．したがって，本論文では検出システムに必要な言い間違いコーパスの拡張手法を提案する．具体的には，言語学的な単位に応じて，統計ベースのルールを用いた手法と翻訳再翻訳を用いた手法の 2 つを用いてコーパスの拡張を行う．拡張の量は，ルールを用いた手法では 10 倍，翻訳再翻訳を用いた手法では 2 倍程度の拡張を行う．また，実験を行い提案手法が有効であるか考察する．

2 言い間違いの定義

寺尾らの研究で，言い間違いには交換，代用，付加，欠落 (削除) などの種類が存在し，それらは様々な言語学的な単位で発生することが知られている [4]．言語学的な単位には素性，音素，形態素，語彙，文などがある．本研究でも同様の言い間違いの種類，言語学的な単位を用いる．また，言い間違いを「言語学単位が他の言語単位の影響を受け，変化したものであり，話者の意図と異なる発話」と定義する．言い間違いの分類については厳密な定義がされていない部分もあり，議論が行われている．本研究ではモーラや音素，音韻単位での言い間違いを「音韻単位での言い間違い」，語彙や文節，単語単位での言い間違いを「語彙単位での言い間違い」と定義し，以降この分類を用いることとする．

Detection of Speech Error depending on Linguistic Units
†National Institute of Technology, Kagawa College

3 言い間違いコーパス

本研究では寺尾らが構築した言い間違いコーパスを用いている．このコーパスの統計量を表 1 に示す．言い間違いコーパスは発話文，言い間違いが発生した箇所，言い間違いが発生した箇所の話者の意図，言い間違いの分類がカンマ区切りで構成されている．言い間違いコーパスの例としては“「映画みたらどうみる？」, どうみる, どうする, $ 語彙 $ 代用 $ 動詞 $ 文脈 ”がある．

表 1: 言い間違いコーパスのサイズと内訳

	交換	代用	付加	欠落	合計
音韻単位	117	1714	71	152	2054
語彙単位	17	636	3	18	674
合計	134	2350	74	170	2728

4 深層学習

本研究では，言い間違いの検出に深層学習を用いる．検出を行うための深層学習アルゴリズムの種類としては，Convolutional Neural Network(CNN) と Recurrent Neural Network(RNN) の一種である Long Short-Term Memory(LSTM) である．また，深層学習用のライブラリとして TensorFlow，ラッパーライブラリとして Keras を用いた．

5 データ拡張手法

提案する言い間違いコーパスにおけるデータ拡張の手法について述べる．データ拡張とは，深層学習において精度向上に有効な訓練データを増強する手法であり，主に画像や音声を訓練データとする学習でよく用いられる [5]．深層学習では，訓練データが少ない場合，過学習に陥り，汎化性能が低下しやすいので，データ拡張を行うことによって訓練データ数を増やすことは精度の向上に有効な手段である．本稿で提案するデータ拡張手法には 2 種類あり，統計ベースのルールに基づいた拡張と翻訳再翻訳を用いた手法である．ルールに基づいた拡張は音韻単位での言い間違い例を拡張するために用いる．翻訳再翻訳を用いた拡張は語彙単位での言い間違い例を拡張するために用いる．次に各手法の詳細について述べる．

5.1 音韻単位におけるデータ拡張

音韻単位のデータ拡張では bi-gram を用いる．既存の言い間違いコーパスをモーラに変換し，言い間違いが発生している部分 (話者の意図と発話が異なる部分) の bi-gram の出現頻度をカウントする．この出現頻度が一定数以上のものをルールとして用いる．具体的には，言い間違いの発生していないデータを用意し，モーラへ変換する．そのモーラ内でルールに一致しているパターンが存在すればルールに基づいてモーラを書き換えることによって行う．本手法によって生成されたルールには，“re/ta → re/ka ” や “ma/N → ni/N ” 等がある．左辺が書き換え前のモーラ，右辺が書き換え後のモーラを表している．また，発話文を拡張する際にはルールは 1 つずつ用いる．

5.2 語彙単位におけるデータ拡張

語彙単位のデータ拡張では翻訳再翻訳を用いる．翻訳再翻訳とは日本語の文章を翻訳ソフトを用いて他言語に翻訳し，その結果を日本語に再翻訳することによって元の文章を文意を保ちつつ，表現を変化させることで文章を生成する手法である．本手法では，翻訳ソフトには Google 翻訳，翻訳先の言語にはトルコ語を用いた．翻訳再翻訳の結果としては，“縁がいい (意図:縁起がいい) → いいエッジ” や “くたびれ損のなんとやら (意図:骨折り損) → あなたは疲労損失で何をしますか ” という結果を得た．

6 評価実験

本手法を用いた際の効果を調査するため，評価実験を行う．実験に用いる訓練データは言い間違いコーパスを用いる．提案手法であるデータ拡張の手法を用いて拡張を行った結果を表 2 に示す．実験では表 2 に示すうち訓練データに 9 割，テストデータに 1 割を用いた．テストデータに訓練データとして用いた言い間違いは含まれていない．

表 2: 言い間違いコーパスの統計量 (データ拡張後)

	データ拡張前	データ拡張後
音韻単位	2054	16000
語彙単位	674	1000
合計	2728	17000

6.1 実験条件と内容

言い間違いの検出には CNN と RNN の一種である LSTM を用いて行う．モデルは言語学的な単位に応じて 2 つ用意し，それぞれで実験を行う．訓練用のデータにはデータ拡張前の既存のコーパスとデータ拡張後のコーパスを用いる．

6.2 実験結果

評価実験の実験結果を表 3, 4 に示す．

表 3: 音韻単位での実験結果 (F 値)

	CNN	LSTM	平均
既存のコーパス	60.6	65.8	63.2
拡張したコーパス	64.8	69.7	67.3

表 4: 語彙単位での実験結果 (F 値)

	CNN	LSTM	平均
既存のコーパス	58.2	59.0	58.6
拡張したコーパス	60.1	62.3	61.2

6.3 考察

音韻単位，語彙単位ともに拡張の結果，検出精度が向上している事が確認できた．音韻単位での実験結果がより精度が向上しているが，これはデータ拡張の結果，より多い訓練データを使用したためだと考えられる．データ拡張を行ったが検出が出来なかった例として，音韻単位では “いちばんあや、足の速い人 ”，語彙単位では “29 行目から 57 ページの 1 行目から (意図:1 行目まで)” がある．

7 おわりに

本論文では言い間違いの検出精度向上のために言語学的な単位に応じたデータ拡張の手法を検討した．その結果，音韻単位では精度向上に有効にはたらくことが確認できた．語彙単位においても精度は向上したが，モデルの構成などの改善が必要と考えられる．

参考文献

[1] Amazon Echo, https://www.amazon.co.jp/dp/B071ZF5KCM/ref=cm_sw_r_cp_ep_dp_qwsMBb1HY9N14

[2] Google Home, https://store.google.com/jp/product/google_home

[3] 小松 聖矢, 篠山 学,“言い間違いの検出のためのデータ拡張”, 平成 30 年度電気関係学会四国支部連合大会予稿集，pp.77-78

[4] 寺尾 康,“言い間違いはどうして起こる？”, 岩波書店，2012

[5] 西本 慎之介, “データ拡張による感情分析のアスペクト推定”, 言語処理学会第 23 回年次大会発表論文集, pp. 581-584, 2017