学生のレポート推敲のための話しことば検出システムについての研究

Sophistication of Colloquial Writing Check System for Academic Report Writing

長谷川 哲生*1,工藤敦也*1,山下由美子*2,山川広人* 1 ,小松川浩*1
Tetsuo HASEGAWA*1, Atsuya KUDOU*1, Yumiko YAMASHITA*2,
Hiroto YAMAKAWA*1, Hiroshi KOMATSUGAWA*1
*1 千歳科学技術大学理工学部
*1 Faculty of Science and Technology, Chitose Institute of Science and Technology
*2 東京福祉大学
*2 Tokyo University of Social Welfare
Email:hasegawa215@kklab.spub.chitose.ac.jp

あらまし:本プロジェクト研究は，学生がレポートを作成する際に話しことばのチェックと学術表現に近づけるトレーニングを図れるシステムの研究を行っている．特に本稿では学生のレポート文中から話しことばを検出できるシステムを試作した上で，さらに機械的なパターンマッチングだけでは検出できない話しことばがある点に着目し，文脈を踏まえた指導を図れる機械学習アルゴリズムのためのデータベースの高度化を行う．
キーワード：自然言語処理，データ，プログラム

1. はじめに

大学の教育課程において，授業目標の到達度を測る有効な手段としてレポートがあり，それは事実や客観的な根拠に基づいて自分の意見を相手や場面に配慮した文章で書くため，「学術文章に適した文章表現（以下，学術表現）」を使用するのが一般的である．しかし，学生が提出したレポートの中には話しことばが使用されているのが多く見受けられる．そのような場合，教員が評価する点を探しづらくなり，レポートの内容面の評価に影響を与える可能性がある．しかし，Web 公開されている校正，推敲ツールは，いずれも誤字脱字や文法的誤りの箇所に指摘を行うことにとどまっており，文章をチェックし，指導まで踏み込むものにはいたっていない．そのため，本研究グループでは，学生がレポートを作成する際に話しことばのチェックと学術表現に近づけるトレーニングを図れるシステムの研究を行っている． 特に本稿では学生のレポート文中から話しことばを検出できるシステムを試作した上で，さらに機械的なパターンマッチングだけでは検出できない話しことばがある点に着目し，文脈を踏まえた指導を図れる機械学習アルゴリズムのためのデータベースの高度化を検討する．

2. 話しことば

本プロジェクト研究では，話しことばを共同研究者の山下が大学初年次における文章教育での学術表現で文章を書く際に指摘を受けやすい特に注意すべき口語や言い回しを抽出したもの」[(1)]と定義した．

2. 1 大学初年度向け話しことば事例集

学生へ提示する話しことばの学術表現への修正方法をデータとして定義するため 2 で述べた話しことばを対象に「大学初年度向け話しことば事例事例集」を作成した．これは，山下が, 2 で述べた定義を基に監修した．

2. 2 話しことばデータベース

話しことばデータベースは，前述した「大学初年度向け話しことば事例集」を基に，情報システムで扱える形にデータベース化したものである．

3. 結果

3. 1 ベースシステム

話しことばデータベースを情報システムで扱いやすくするためにデータベースの正規化を行い，これらを用いて話しことば検出システム「話しことばチェッカー」を開発した．本システムに学生のレポートを入力すると, 5 つのモジュールによる工程を経て，話しことばを検出し，黄色く表示できる．また，黄色く表示された話しことばにカーソルを合わせることで例文や解説などその話しことばの詳細をポップアップ表示にて提示できる. (図 1)

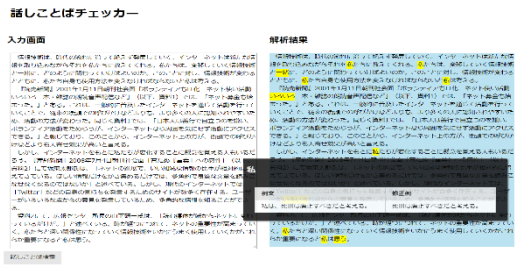

図 1 ベースシステムの画面図

3. 2 データベースの高度化

また，文脈を踏まえた指導を図れる機械学習アルゴリズムの構築に向けて， 学生のレポート推敲のための話しことば検出データベースの高度化を行った．データベースの高度化は，システムに日本語のエキスパートが作成した「大学初年度向け話しことば事例事例集」を基に，機械学習アルゴリズムで用いるための品詞や単語の共起関係を含むベクトル化の処理を行った．具体的には, (1)形態素解析器に用いる辞書の選択. (2)形態素解析を行った際の品詞などについて記載. (3)話しことばデータベースに記載されている文章を用いた Word2Vec[(2)]での各単語のベクトル化. (4)(3)で行ったベクトルを用い類似する各単語上位 20 単語を抽出の 4 つのことを行った．

5. 今後の取り組み

今後の取り組みは，「話しことば判定」の処理が一部のみ開発が完了した状態であるため，全ての話しことばを検出することができない．本研究を継続する場合，残る話しことばの処理方法を考える必要がある．また，データベースの高度化で作成したデータを基に，単語のつながりや共起関係を含む文章を取り込み，共起関係の明確化や話しことばを含む文章の取り込みによる文章の自動チェックといったことの実現が考えられる．そのために，エキスパートと相談し話しことばデータベースの文章の充実を行い．話しことばデータベースを用いた Web システムに組み込むアルゴリズムの構築を行っていく．

参考文献

(1). 山下由美子:”学生のレポートにおける話し言葉とその出現傾向”, 日本語日本文学 第 28 号, 2018 p57-p71

(2). Tomas Mikolov:”Efficient Estimation of Word Representations in Vector Space”, https: //arxiv. org/pdf/1301.3781.pdf,（2019 年 1 月アクセス）