論文内容の要旨

博士論文題目　Conversion of Noisy or Long Sentences into Readable Sentences
(崩れたテキストや長いテキストの読みやすいテキストへの変換)

氏名　　高橋いつみ

(論文内容の要旨)

本論文は，文章の単語的な表記揺れを統制するための変換技術や，文章全体をシンプルで短い文に変換する技術など，崩れた文章や長い文章をわかりやすい文に変換する技術を提案している．本研究は大きく3つの成果を含んでいる．

一つ目の成果は，Twitterのような辞書に存在しない崩れた語が多く含まれる口語体の文章について，単語レベルの正規化と形態素解析を同時に行う技術である．従来の形態素解析技術では，辞書にない単語が入力されると解析が失敗し，後段の処理に意味の異なる形態素が渡されてしまうことがあった．本研究では，辞書に存在しない崩れた口語を辞書に存在する正規の語に紐付ける正規化と形態素解析を同時に行うことで，形態素解析の精度向上と正規化による読みやすい文への変換の両方を達成した．また，崩れた表記と正規表記のペアを教師なしで獲得する手法についても提案した．音的な類似性と表記的な類似性によって教師なしで獲得した表記変換パタンを形態素解析に導入し，効果を確認した．

二つ目の成果は，ニューラルネットワークを用いて，方言を標準語に文レベルで変換する技術である．ニューラルネットを用いた文変換は，変換元のテキストと変換先のテキストペアの学習データが大量に存在する場合には高い精度を達成できることが知られている．ただし，方言から標準語への変換タスクにおいては学習データが少なく，大量のデータを人手で作成することも高コストであり現実的ではない．本研究では，少量の人手データから文字レベルの変換パタンを抽出し，大量のラベルなしデータに変換パタンを適用することによって疑似的なペアデータを大量に生成する手法を提案した．提案手法により，従来型の統計的機械翻訳を用いた手法に比べ複数方言ドメインでの精度向上を達成した．

三つ目の成果は，長い文書を短く読みやすい要約文へと変換する技術である．現在最も精度が良い要約システムにおいては，要約元となるテキストに対して一つの要約テキストしか出力することができない．提案手法は，単語レベルの抽出型要約と生成型要約を効果的に組み合わせることによって，生成型要約の精度向上と長さの制御性の双方を達成した．本手法は，出力したい要約長に合わせて本文から重要な語を抽出し，要約生成の補助情報として用いる．実験の結果，要約で標準的に用いられるデータセットCNNDMとNewsroomにおいて従来手法を上回る精度を達成した．

これら一連の研究は，単語レベル，文レベル，文書レベルのそれぞれにおいて崩れた文や長い文章を読みやすい文に変換する技術の提案である．

氏　名	高橋いつみ

（論文審査結果の要旨）

令和元年12月12日に開催した公聴会の結果を参考に令和2年2月7日に本博士論文の審査を行った．以下のとおり，本博士論文は，提案者が独立した研究者として，研究活動を続けていくための十分な素養を備えていることを示すものと認める．

高橋いつみは，本博士論文において，日本語の崩れたテキストや長いテキストを読みやすいテキストあるいは短いテキストに変換する新しい技術を提案した．

本論文の貢献は以下のようにまとめることができる．

1. ツイッター等のソーシャルメディアにおいては，通常の書き言葉とは異なり，口語的な崩れた表現が多く含まれる．本論文では，辞書に存在しない崩れた表現を辞書に存在する正規の語に紐付ける正規化と形態素解析を同時に行い，形態素解析の精度向上と読みやすい文への変換を同時に達成した．

2. ソーシャルメディア上の短い文を対象に，標準的な表現とその変形の対を自動的に収集する方法を提案し，人手によるアノテーションなしに大規模な崩れた表現と正規の表現との対応データの構築を可能にした。

3. 文字レベルおよび形態素レベルの変換を用いて大量のラベルなしデータから疑似的なペアデータを大量に生成する手法を提案した．方言を標準語に文レベルで変換する問題にその技術を適用し，精度向上を達成した．

4. 長い文章を短く文章に変換する文書要約では，出力となる要約テキストの長さを柔軟にコントロールすることが容易ではなかったが，本論文では，単語レベルの抽出型要約と生成型要約を効果的に組み合わせることによって，生成型要約の精度向上と長さの制御性の双方を達成することに成功した．

文章変換技術を様々な視点から取り組み，いくつかの異なる問題において言語表現の正規化や平易化の実現法を提案した本研究は，独創性が高く，しかも実用的であり，自然言語処理の分野において高い貢献があると評価する．

よって，本論文は，博士（工学）の学位論文として価値あるものと認める．