様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ６ 月　 ３ 日現在

機関番号：２２６０４
研究種目：若手研究(B)
研究期間：2016～2018
課題番号：１６Ｋ１６１１７

研究課題名（和文）頑健な単語表現の学習と深層ニューラルネットワークを用いた誤り訂正

研究課題名（英文）Grammatical Error Correction using Robust Word Representation Learning and Deep Neural Networks

研究代表者
小町　守（Komachi, Mamoru）

首都大学東京・システムデザイン研究科・准教授

研究者番号：６０５８１３２９

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：本研究は、文法誤り検出・訂正に深層学習を適用するアプローチを提案しました。言語学習者の書いた文章はネイティブが書いた文章と異なり、単語自体を間違えたり文脈を間違えたりしますが、これらの文脈の違いを考慮して単語を数理的にモデル化し、深層学習を用いて文法誤り検出・訂正を行いました。
また、ネイティブの書いた大規模なテキストデータから学習した文脈つきの言語表現モデルで誤り検出をする手法も提案し、英語文法誤り検出で世界最高精度を達成しました。一方、日本語や中国語については漢字を部首に分解して単語を再構成するモデルを提案し、日中のニューラル機械翻訳で有効性を示しました。

研究成果の学術的意義や社会的意義
本研究では英語学習者の文法誤り検出について、学習者の文章の誤り方を考慮して単語をモデル化することと、ネイティブが書いた大規模な文章データから獲得した文脈つきの言語表現モデルを用いることが、それぞれ有効であることを世界で初めて示し、いずれの研究においても当時の世界最高精度の精度を達成することができました。本研究は世界を代表する英語学習者の文法誤り検出の研究の一つです。

研究成果の概要（英文）：This study proposed an approach to apply deep learning to grammatical error detection and correction. The sentences written by language learners differ from those written by native speakers in that they may make mistakes in the words themselves or in the context of which they are used. Taking these differences in context into account, we built a mathematical model for representing words and used deep learning to detect and correct grammatical errors. We also proposed a method for error detection using a contextualized language representation model learned from a large amount of text data written by native speakers, and achieved the state-of-the-art accuracy in English grammatical error detection.
On the other hand, for Japanese and Chinese, we proposed a model that reconstructs word sequence by decomposing kanji into radicals, and demonstrated its effectiveness in Japanese-Chinese neural machine translation.

研究分野：自然言語処理

キーワード：深層学習　単語分散表現　文法誤り訂正

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

本研究を提案した 2015 年後半は、ちょうど自然言語処理分野で深層学習が流行り出した時期でした。2012 年に word2vec (Mikolov et al., 2013) と呼ばれる単語ベクトルを訓練する手法が提案され、その有効性が知られていました。また 2013 年から 2014 年にかけて、エンコーダ・デコーダと呼ばれる 2 つのニューラルネットワークを用いて機械翻訳を行う手法が提案され (Sutskever et al., 2014) 、シンプルなモデルで高い翻訳精度を出すことができることが判明し、自然言語処理の研究者を驚かせていました。しかしながら、まだ文法誤り検出・訂正の分野では、word2vec のような単語ベクトルの訓練手法が効果あることは分かっておらず、ニューラルネットワークを用いたアプローチの有効性も不明で、文法誤り訂正においてはニューラルネットワークは有効ではない、と思われていました。特に、2016 年当時は統計的機械翻訳を用いた手法が文法誤り訂正ではもっとも効果的である、と言われるくらいでした (Grundkiewicz and Junczys-Dowmunt, 2016)。

参考文献

(1) Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. International Conference on Learning Representations (ICLR). 2013.
(2) Ilya Sutskever, Oriol Vinyals and Quoc V Le. Sequence to Sequence Learning with Neural Networks. Advances in Neural Information Processing Systems (NIPS). 2014.
(3) Roman Grundkiewicz and Marcin Junczys-Dowmunt. Phrase-based Machine Translation is State-of-the-Art for Automatic Grammatical Error Correction. Conference on Empirical Methods in Natural Language Processing (EMNLP). 2016.

２．研究の目的

そこで、本研究は、①文法誤り検出・訂正に適した単語ベクトルの訓練方法の提案と、②深層ニューラルネットワークを用いた文法誤り検出･訂正の高精度化に取り組むことにしました。特に言語学習者が書く文章はネイティブが書く文章とは異なりさまざまな誤りを含むため、単純にネイティブが書いた文章で訓練したモデルではうまくいかないことが予想され、言語学習者の書いたテキストに深層学習を適用するためにどのようなアプローチを取ればいいのか、ということを解明することが目的です。

また、最初は分析がしやすく使うことのできるデータも多い英語学習者の書いたテキストを対象に研究をスタートしますが、徐々に日本語学習者の書いたテキストも対象に研究を発展させる計画を立てました。日本語学習者の書いたテキストは、単語分割に失敗してしまうことが多いという問題があり、英語学習者の書いたテキストよりも格段に難しいというチャレンジがあります。

３．研究の方法

文法誤り検出・訂正に適した単語ベクトルの訓練に当たっては、word2vec の訓練では単語の正しい文脈と間違った文脈を用いて単語ベクトルを推定するのですが、間違った文脈は単語をランダムに別の単語に置き換えたものを用いる擬似負例という手法を用いて訓練していました。この擬似負例は、ネイティブが書いた文章の単語の出現頻度を用いてランダムにサンプリングされるものだったので、私たちは言語学習者の誤り方を考慮したサンプリングを行うことで、擬似負例をより現実的な負例に近づける手法を提案しました。また、言語学習者の書いた文章は大量に入手することが難しい、という問題を解決するために、大規模な文章データを用いて訓練された言語表現モデルを用い、言語学習者のテキストで微調整 (fine-tuning) を行う手法も提案しました。

深層ニューラルネットワークを用いた文法誤り検出・訂正については、上記のような単語ベクトルの訓練を行なったものを、先行研究 (Rei and Yannakoudakis, 2016) に基づき、Long Short-Term Memory (LSTM) と呼ばれる再帰的なニューラルネットワーク（recurrent neural networks）の入力として用い、各単語が正しいか誤っているかの 2 値分類を行う系列ラベリングというタスクとして定式化して解く手法を提案しました。また、自然言語処理分野でも深層学習が盛んに研究され、Google ニューラル機械翻訳で採用された Transformer と呼ばれる深層ニューラルネットワーク (Vaswani et al., 2017) が最も性能が高く、かつ深層ニューラルネットワークの層ごとの分析もでき、単語の表層的な特徴から深い意味的な特徴も考慮できることが知られているため、Transformer を文法誤り検出に適用しました。

一方、日本語学習者の文法誤り検出・訂正については、単語分割の問題を扱う前に、単語単位の処理ではなく文字単位、そして文字をさらに細かくした単位でベクトルを訓練し、そこから文字そして単語そして文へとベクトルを組み上げていくという方法で、単語の分割が難しいという問題を解決する手法を検討しました。まず日本語でも中国語でも使われる漢字を対象にこのアプローチを試し、日本語と中国語の間のニューラル機械翻訳で有効性を示すことにしました。

参考文献

(1) Marek Rei and Helen Yannakoudakis. Compositional Sequence Labeling Models for Error Detection in Learner Writing. Annual Meeting of Association for Computational Linguistics (ACL). 2016.
(2) Ashish Vaswani, Noan Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. Attention is All you Need. Advances in Neural Information Processing Systems (NIPS). 2017.

４．研究成果

英語学習者の誤用を考慮して単語ベクトルを訓練する手法では、LSTM を用いて文法誤り検出を行うことで、2017 年当時の世界最高精度を達成し、国際会議で発表するとともに、論文誌でも発表しました (Kaneko et al., 2017; 金子ら, 2018)。また、ネイティブの書いた大規模なテキストから言語表現モデルを学習し、Transformer を用いて文法誤り検出を行う手法は、2018 年当時の世界最高精度を達成し、国際会議で発表しました (Kaneko et al., 2019)。

漢字を文字より小さい単位に分解して深層ニューラルネットワークの学習を行う手法では、日本語と中国語で部首単位に分解してニューラル機械翻訳の学習を行い、いずれの言語においても統計的に有意な性能向上が得られることを示し、機械翻訳の世界最高峰の国際会議で発表しました (Zhang and Komachi, 2018)。また、この手法は教師なしニューラル機械翻訳という 2018 年に提案された新しい対訳データを使わずに翻訳を行う手法でも有効性を示し、国際会議で発表しています (Zhang et al., 2018)。

５．主な発表論文等

〔雑誌論文〕（計　1　件）

1. 金子正弘, 堺澤勇也, 小町守. 正誤情報と文法誤りパターンを考慮した単語分散表現を用いた文法誤り検出. 自然言語処理, Vol.25, No.4, pp.421-440. September, 2018.

〔学会発表〕（計　5　件）

1. Masahiro Kaneko and Mamoru Komachi. Multi-Head Multi-Layer Attention to Deep Language Representations for Grammatical Error Detection. In 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing). April, 2019. (poster)
2. Longtu Zhang and Mamoru Komachi. Neural Machine Translation of Logographic Language Using Sub-character Level Information. In Proceedings of the Third Conference on Machine Translation, pp.17-25. Brussels, Belgium. October 31, 2018.(poster)
3. Longtu Zhang, Yuting Zhao, Mamoru Komachi. TMU Japanese-Chinese Unsupervised NMT System for WAT 2018 Translation Task. Proceedings of the Third Workshop on Asian Translation (WAT 2018). Hong Kong. December 3, 2018. (poster)
4. Masahiro Kaneko, Tomoyuki Kajiwara and Mamoru Komachi. TMU System for SLAM-2018. In Proceedings of The 13th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2018): Shared Task on Second Language Acquisition Modeling, pp.365-369. New Orleans, Louisiana, USA. June 5, 2018. (poster)
5. Masahiro Kaneko, Yuya Sakaizawa, Mamoru Komachi. Grammatical Error Detection Using Error- and Grammaticality-Specific Word Embeddings. In Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP 2017), pp.40-48. Taipei, Taiwan. November 28, 2017. (oral)

〔図書〕（計　1　件）

1. 小町守．自然言語処理の教育応用．人工知能学会編，人工知能学大事典，pp.688-689．2017 年 7 月．共立出版．

〔産業財産権〕

○出願状況（計　0　件）

○取得状況（計　0　件）

〔その他〕

ホームページ等

http://kanekomasahiro.sakura.ne.jp/revision_support/

６．研究組織

(1) 研究分担者
なし

(2)研究協力者
研究協力者氏名：金子正弘
ローマ字氏名：Masahiro Kaneko

研究協力者氏名：张龙图
ローマ字氏名：Longtu Zhang