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　４ 年　 ６ 月 １４ 日現在

機関番号：２２６０４
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ１２０９９

研究課題名（和文）単語分散表現の学習における構成性とその解釈

研究課題名（英文）Compositionality and Interpretation of Word Embeddings

研究代表者
小町　守（Komachi, Mamoru）

東京都立大学・システムデザイン研究科・教授

研究者番号：６０５８１３２９

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：本研究では、自然言語処理における単語をベクトル化して表現する分野である単語分散表現の学習において、単語より小さい単位での分散表現から、より大きな単位の分散表現を計算する手法について研究を行いました。具体的には、機械翻訳を題材にして日中翻訳における単語分散表現学習における最適な入力の粒度を探究し、文法誤り訂正においても日本語・英語・ドイツ語・ロシア語など複数の言語でどのような知識が転移可能であるかを明らかにしました。また、単語分散表現の解釈についても取り組み、通時的な意味変化を捉えるための単語分散表現の学習において情報理論的な背景を持つアプローチを採用し、解釈性の高い手法を提案しました。

研究成果の学術的意義や社会的意義
本研究の成果は、日本語や中国語のような表意文字を用いる言語は、文字よりも細かい単位で意味を捉える方が適切であるという可能性を示唆している点にあります。世界的には英語に代表されるような少数のアルファベットを用いる言語が広く研究されていますが、そのような言語で提案されている手法が日本語や中国語では必ずしも最適な手法ではない、ということを意味します。深層学習の登場により多言語を同時に扱うことのできる手法がさまざま提案されていますが、それぞれの言語の特徴も考慮することの重要性を改めて示しています。

研究成果の概要（英文）：In this research, we studied methods for composing distributed representation of words from smaller units in word representation learning in natural language processing. Specifically, focusing on machine translation, we explored the optimal granularity of input for learning distributed representation of words in Japanese-Chinese translation. We also clarified what kind of knowledge is transferable across languages such as Japanese, English, German, and Russian for grammatical error correction. In addition, we addressed the interpretation of word representations, and proposed a highly interpretable method for learning word representations to capture diachronic semantic change, employing an approach with an information-theoretic background.

研究分野：自然言語処理

キーワード：単語分散表現　構成性　機械翻訳　文法誤り訂正　意味変化

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

研究代表者は単語や文の分散表現の学習と応用（Kaneko et al., 2017; Shimanaka et al., 2018)、深層ニューラルネットワークを用いた機械翻訳（Zhang and Komachi, 2018; Kurosawa et al., 2018; Katsumata et al., 2018）のような自然言語処理における深層学習の研究に着手しており、英語の文法誤り検出においても、研究代表者が提案した単語分散表現の学習手法を用いることで、当時の世界最高精度を達成することができた（Kaneko et al., 2017)。また、Zhang and Komachi（2018）では文字より小さい篇や旁からの分散表現の学習が日本語と中国語のニューラル機械翻訳に有効であることを発見し、単語が必ずしも分散表現の学習に最適な単位ではないことが明らかになりつつあった。

２．研究の目的

そこで、本研究は文の分散表現を学習するために適した単語より細かい粒度の分散表現学習の研究を行った。多くの自然言語処理アルゴリズムは入力の最小単位を単語としているが、接尾辞や接頭辞、語根のように、単語は文字から意味を構成することができるため、文字の分散表現を用いて単語の分散表現を補完することで、頑健性が高められることが知られている（dos Santos and Zadrozny, ICML 2014;　Bojanowsky et al., TACL 2017)。特に中国語では文字も偏（へん）や旁（つくり）といった部首に分解し、部首に対応する分散表現を学習する手法も提案され（Shi et al., ACL 2015)、小さな単位の分散表現を組み合わせて用いることで、さらにデータスパースネスを解消する方法が登場していた。

３．研究の方法

以上の背景を元に、単語より細かい単位での表現学習と、それを用いた言語理解の研究に取り組んだ。日本語は漢字以外にひらがなやカタカナという表音文字も用いる複雑な書字体系を持っており、表音文字は音によってさらに分解することができるという特徴がある。Zhang and Komachi（2018）では、日本語では書き順を用いて分解した分散表現が、中国語では部首を用いて分解した分散表現がそれぞれ有効であることを示した。本研究ではその研究をさらに進め、日中機械翻訳で日本語に適した単語表現を学習する手法を提案した。また、多言語の文法誤り訂正タスクにおいても、どのような文法知識が転移可能であるかを検証した。

４．研究成果

論文誌

1. 嶋中宏希, 梶原智之, 小町守. **事前学習された文の分散表現を用いた機械翻訳の自動評価**. 自然言語処理. 26 巻 3 号, pp.613-634. September, 2019.
2. Longtu Zhang and Mamoru Komachi. **Using Sub-Character Level Information for Neural Machine Translation of Logographic Languages**. ACM Transaction on Asian and Low-Resource Language Information Processing. Vol.20, No.2, Article No.31, pp.1-15. April 15, 2021.
3. 山下郁海, 金子正弘（東工大）, 三田雅人（理研）, 勝又智, Imankulova Aizhan, 小町守. **言語間での転移学習のための事前学習モデルと多言語の学習者データを用いた文法誤り訂正**. 自然言語処理, 29 巻 2 号. 2022.

国際会議

1. Longtu Zhang and Mamoru Komachi. **Chinese--Japanese Unsupervised Neural Machine Translation Using Sub-character Level Information**. *The 33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33)*. September, 2019.
2. Hwichan Kim, Tosho Hirasawa and Mamoru Komachi. **Korean to Japanese Neural Machine Translation System Using Hanja Information**. In *Proceedings of the 7th Workshop on Asian Translation (WAT)*, pp. 127-134. December, 2020.
3. Hwichan Kim, Tosho Hirasawa and Mamoru Komachi. **Zero-shot North Korean to English Neural Machine Translation by Character Tokenization and Phoneme Decomposition**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop (ACL 2020 SRW)*, pp.72-78. Seattle, WA, USA. July, 2020.
4. Ikumi Yamashita, Satoru Katsumata, Masahiro Kaneko, Aizhan Imankulova and Mamoru Komachi. **Cross-lingual Transfer Learning for Grammatical Error Correction**. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING), pp.*4704-4715. December 11, 2020.
5. Taichi Aida, Mamoru Komachi, Toshinobu Ogiso (National Institute for Japanese Language and Linguistics), Hiroya Takamura (National Institute of Advanced Science and Technology), Daichi Mochihashi (The Institute of Statistical Mathematics). **A Comprehensive Analysis of PMI-based Models for Measuring Semantic Differences**. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation (PACLIC 2021), pp. 21-31*. November 7, 2021.
6. Kazuma Kobayashi, Taichi Aida and Mamoru Komachi. **Analyzing Semantic Changes in Japanese Words Using BERT**. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation (PACLIC 2021), pp. 273-283*. November 7, 2021.

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　5件／うち国際共著　0件／うちオープンアクセス　5件）

１．著者名 Longtu Zhang and Mamoru Komachi	４．巻 -
２．論文標題 Using Sub-Character Level Information for Neural Machine Translation of Logographic Languages	５．発行年 2021年
３．雑誌名 ACM Transaction on Asian and Low-Resource Language Information Processing	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 嶋中宏希，梶原智之，小町守	４．巻 26
２．論文標題 事前学習された文の分散表現を用いた機械翻訳の自動評価	５．発行年 2019年
３．雑誌名 自然言語処理	６．最初と最後の頁 613-634
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 山下郁海 ，　金子正弘，三田雅人，勝又智，Imankulova Aizhan，小町守	４．巻 -
２．論文標題 言語間での転移学習のための事前学習モデルと多言語の学習者データを用いた文法誤り訂正	５．発行年 2022年
３．雑誌名 自然言語処理	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計17件（うち招待講演　0件／うち国際学会　17件）

１．発表者名 Ikumi Yamashita, Satoru Katsumata, Masahiro Kaneko, Aizhan Imankulova and Mamoru Komachi
２．発表標題 Cross-lingual Transfer Learning for Grammatical Error Correction
３．学会等名 28th International Conference on Computational Linguistics (COLING)（国際学会）
４．発表年 2020年

１．発表者名
Hwichan Kim, Tosho Hirasawa and Mamoru Komachi

２．発表標題
Zero-shot North Korean to English Neural Machine Translation by Character Tokenization and Phoneme Decomposition

３．学会等名
58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop (ACL 2020 SRW)（国際学会）

４．発表年
2020年

１．発表者名
Hwichan Kim, Tosho Hirasawa and Mamoru Komachi

２．発表標題
Korean to Japanese Neural Machine Translation System Using Hanja Information

３．学会等名
7th Workshop on Asian Translation (WAT)（国際学会）

４．発表年
2020年

１．発表者名
Longtu Zhang and Mamoru Komachi

２．発表標題
Chinese--Japanese Unsupervised Neural Machine Translation Using Sub-character Level Information

３．学会等名
The 33rd Pacific Asia Conference on Language, Information and Computation（国際学会）

４．発表年
2019年

１．発表者名
Taichi Aida, Mamoru Komachi, Toshinobu Ogiso, Hiroya Takamura, Daichi Mochihashi

２．発表標題
Comprehensive Analysis of PMI - based Model s for Measuring Semantic Differences

３．学会等名
35th Pacific Asia Conference on Language, Information and Computation (PACLIC 2021),（国際学会）

４．発表年
2021年

１．発表者名 Kazuma Kobayashi, Taichi Aida and Mamoru Komachi
２．発表標題 Analyzing Semantic Changes i n Japanese Words Using BERT
３．学会等名 35th Pacific Asia Conference on Language, Information and Computation (PACLIC 2021) (国際学会)
４．発表年 2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関