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　５ 年　 ５ 月　 ８ 日現在

機関番号：１４６０３
研究種目：若手研究
研究期間：2019～2022
課題番号：１９Ｋ２０３５１

研究課題名（和文）言語解析における目標テキストへの特化技術に関する研究

研究課題名（英文）A Study of Specializing Natural Language Processing Models for Target Texts

研究代表者
大内　啓樹（Ouchi, Hiroki）

奈良先端科学技術大学院大学・先端科学技術研究科・助教

研究者番号：７０８２５４６３

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：本研究では2つの研究目的を掲げた。
一つ目が「分散表現を目標テキストに特化させる手法の開発とその効果の検証」である。これは2019年度に順調に遂行することができた。
2020年度以降は、二つ目の研究目的である「同一ラベルを持つ事例が特徴ベクトル空間上で近くに位置するように学習する手法の開発とその効果の検証」に取り組んだ。各事例を特徴ベクトル空間に写像する深層ニューラルネットワークに距離学習を適用することによって、同一ラベルを持つ事例同士が特徴ベクトル空間上で近づくような学習を実現した。その結果として、学習事例との類似度に基づいてテスト事例を分類することが可能となった。

研究成果の学術的意義や社会的意義
一つ目の研究目的遂行によって、目標テキストが所与の場合はそのテキストに単語分散表現(言語モデル)を特化させることが効果的であることを示された。実応用の文脈で言い換えると、解析したい(目標)テキスト集合を手元に保有している一般企業やユーザーは、本提案手法のように目標テキストにモデルを特化させることによってより効果的に解析可能であることが示唆された。
二つ目の研究目的遂行によって、従来の深層ニューラルネットが抱える解釈性の問題への緩和策を提示した。例えば、「この学習事例と類似しているため、このテスト事例はこのように分類します」といったように、根拠を提示しながら予測を行えるようになった。

研究成果の概要（英文）：We had two objectives in this research.
The first was to develop a method to specialize distributed representations to target texts and to test its effectiveness. This was successfully accomplished in 2019.
From 2020 onward, we worked on the second objective; developing methods to learn instances with the same label so that they are located near each other in the feature vector space and verifying the effectiveness. By applying distance learning to a deep neural network that maps each instance to a feature vector space, we achieved learning so that instances with the same label are close to each other in the feature vector space. As a result, test instances could be classified based on their similarity to the training instances.

研究分野：自然言語処理

キーワード：事例ベース学習　表現学習　構造予測

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

自然言語処理技術は,近年の深層学習手法の導入に伴ってさらなる進展を見せている.しかし,標準的なベンチマークテストにおける解析精度向上とは対象的に，実用的な問題設定における解析には課題が残されている.

自然言語処理における一般的な教師あり学習の設定では，教師ラベル付き学習データでモデルを学習し，任意の未知テキストでその汎化性能を測定する．また，分野適応の設定では，元分野と目標分野のデータからモデルを学習し，目標分野の未知テキストで汎化性能を測定する．どちらの設定においても，解析対象のテキストは未知であると仮定されてきたが，実応用上は必ずしもこの仮定を置く必要はない．例えば現在，多くの一般企業や一般ユーザーが解析したいテキストデータを独自で保有しており，そのテキストデータのみを高精度で解析したいというニーズがある．つまり，未知のあらゆるテキストの解析を目的とせず，手元にある特定のテキストのみを高精度で解析できれば良いという場合が少なくない．

本研究では，解析対象のテキストを「目標テキスト」と呼ぶ．特定の目標テキストだけが解析の対象だとすれば，従来のような汎用モデルを構築する必要はかならずしもない．むしろ，目標テキストに特化したモデルを構築することによって，精度をさらに向上させることができると期待できる．このような問題設定は「トランズダクティブ学習(Transductive Learning)」と呼ばれる．ポイントは，目標テキストが所与であり，学習中にもそれらを利用可能な点である．目標テキストを学習時に効果的に利用できれば，目標テキストに対する予測が容易になると期待できる．

以上をまとめると，従来研究が未知のテキストにモデルを汎化させること(Generalization) をめざしているのに対し，本研究では解析対象の特定テキスト集合(目標テキスト) にモデルを特化させること(Specialization) によって，言語解析タスク横断的に大幅な精度向上をめざす.

２．研究の目的

本研究では，以下の 2 つの目的を掲げた。

(1) 目標テキストに特化した単語分散表現の学習：単語分散表現は，言語解析タスクの精度向上に大きく貢献することが知られている．従来の単語分散表現学習は，あらゆるテキストをできるだけカバーするように大量の生テキストを用いて行われてきた．本研究では，目標テキストにモデルを特化させるため，目標テキスト自体を単語分散表現学習に利用する．また，目標テキストに加え，目標テキストと類似した生テキストをウェブ上から大量に収集し，それらを利用することによって，目標テキストの解析に有用な分散表現を学習する手法を考案する．

(2) 事例間の類似性に基づく特徴ベクトル空間の学習：「同一ラベルを持つ事例が特徴ベクトル空間上で近くに位置するように学習する手法の開発とその効果の検証」に取り組む．これによって，従来の深層ニューラルネットワークが抱える解釈性の問題への一つの緩和策として機能する．例えば，「この学習事例と類似しているため，このテスト事例はこのように分類します」といったように，根拠を提示しながら予測を行えるようなることが期待される．

３．研究の方法

【目的(1)に関する方法】

ふたつのサブモデルからなるモデルを仮定し，段階的に学習していく手法を提案した．サブモデルのひとつめの言語モデルであり，テキストの各単語をベクトル表現に変換する．もうひとつのサブモデルはタスク依存モデルであり，変換された単語表現を入力として受け取り，各解析タスクで求められる構造を出力する.これらふたつのサブモデルを三段階に分けて学習する.まず,大規模コーパスから言語モデルを学習し，任意のテキストに汎化するようにした．次に，学習済み言語モデルを目標テキストで再学習し，目標テキストの単語分布に特化するようにした．最後に，学習済み言語モデルから単語ベクトルを入力として受け取るタスク依存モデルを，各解析タスクの教師信号によって学習した．統語・意味解析タスクにおける評価実験の結果として，言語モデルを目標テキストに特化させることによって，そうでない場合よりも性能を改善できることがわかった．特に，学習データと異なる分野の目標テキストを解析対象とする際により大きな効果が見られた.

【目的(2)に関する方法】

各事例を特徴ベクトル空間に写像する深層ニューラルネットワークに距離学習を適用することによって，同一ラベルを持つ事例同士が特徴ベクトル空間上で近づくような学習を実現した。固有表現抽出，統語チャンキング，関係抽出，文書分類タスクなどを通じてその有効性を確認した.

４．研究成果

【目的(1)に関する成果】

実験結果から，申請時に期待していた通り，目標テキストが所与の場合はそのテキストに単語分散表現(言語モデル)を特化させることが効果的であることを示された．実応用の文脈で言い換えると，解析したい(目標)テキスト集合を手元に保有している一般企業やユーザーは，本提案手法のように目標テキストにモデルを特化させることによってより効果的に解析可能であることが示唆された．

【目的(2)に関する成果】

同一ラベルを持つ事例同士が特徴ベクトル空間上で近づくような学習を実現した。その結果として、学習事例との類似度に基づいてテスト事例を分類することが可能となった。つまり，根拠を提示しながらテストデータの予測を行えるようになった．特に関係抽出タスクでは，Few-Shot 学習設定や Zero-shot 学習設定などの極めて教師データの少ない設定を想定して手法を拡張し，従来法を大幅に上回る性能を記録できることを確認した．

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 Ouchi Hiroki、Suzuki Jun、Kobayashi Sosuke、Yokoi Sho、Kuribayashi Tatsuki、Yoshikawa Masashi、Inui Kentaro	４．巻 9
２．論文標題 Instance-Based Neural Dependency Parsing	５．発行年 2021年
３．雑誌名 Transactions of the Association for Computational Linguistics	６．最初と最後の頁 1493～1507
掲載論文のDOI（デジタルオブジェクト識別子） 10.1162/tacl_a_00439	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 Van-Hien Tran, Hiroki Ouchi, Hiroyuki Shindo, Yuji Matsumoto, Taro Watanabe	４．巻 30
２．論文標題 Enhancing Semantic Correlation between Instances and Relations for Zero-Shot Relation Extraction	５．発行年 2023年
３．雑誌名 Journal of Natural Language 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 －

〔学会発表〕　計15件（うち招待講演　1件／うち国際学会　7件）

１．発表者名 Ryuto Konno, Shun Kiyono, Yuichiroh Matsubayashi, Hiroki Ouchi, Kentaro Inui
２．発表標題 Pseudo Zero Pronoun Resolution Improves Zero Anaphora Resolution
３．学会等名 The 2021 Conference on Empirical Methods in Natural Language Processing（国際学会）
４．発表年 2021年

１．発表者名 Tatsuki Kuribayashi, Hiroki Ouchi, Naoya Inoue, Paul Reisert, Toshinori Miyoshi, Jun Suzuki, Kentaro Inui
２．発表標題 An Empirical Study of Span Representations in Argumentation Structure Parsing
３．学会等名 言語処理学会第27回年次大会（招待講演）
４．発表年 2021年

１．発表者名
大内啓樹，鈴木潤，小林颯介，横井祥，栗林樹生，吉川将司，乾健太郎

２．発表標題
事例ベース依存構造解析のための依存関係表現学習

３．学会等名
言語処理学会第27回年次大会

４．発表年
2021年

１．発表者名
佐藤俊，大内啓樹，塙一晃，佐々木翔大，乾健太郎

２．発表標題
事例ベース推論を行うニューラルモデルの説明性とハブ現象の関係

３．学会等名
情報処理学会自然言語処理研究会

４．発表年
2021年

１．発表者名
佐藤俊，大内啓樹，佐々木翔大，塙一晃，乾健太郎

２．発表標題
説明性の高いニューラルモデルの予測確信度に関する分析

３．学会等名
言語処理学会第27回年次大会

４．発表年
2021年

１．発表者名
Hiroki Ouchi, Jun Suzuki, Sosuke Kobayashi, Sho Yokoi, Tatsuki Kuribayashi, Ryuto Konno, Kentaro Inui

２．発表標題
Instance-Based Learning of Span Representations: A Case Study through Named Entity Recognition

３．学会等名
Association for Computational Linguistics（国際学会）

４．発表年
2020年

１．発表者名
Hiroki Ouchi, Jun Suzuki

２．発表標題
Transductive Learning of Neural Language Models for Syntactic and Semantic Analysis

３．学会等名
The 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)（国際学会）

４．発表年
2019年

１．発表者名
大内啓樹，鈴木潤，小林颯介，横井祥，栗林樹生，乾健太郎

２．発表標題
スパン間の類似性に基づく事例ベース構造予測

３．学会等名
言語処理学会第26回年次大会

４．発表年
2020年

１．発表者名
佐々木翔大，大内啓樹，鈴木潤，Ana Brassard，乾 健太郎

２．発表標題
単一評価サンプルのためのトランズダクティブ学習

３．学会等名
言語処理学会第26回年次大会

４．発表年
2020年

１．発表者名
佐藤俊，佐々木翔大，大内啓樹，鈴木潤，乾健太郎

２．発表標題
評価データのクラスタリングを用いた記述式答案自動採点のためのトランズダクティブ学習

３．学会等名
言語処理学会第26回年次大会

４．発表年
2020年

１．発表者名
Van-Hien Tran, Hiroki Ouchi, Taro Watanabe, Yuji Matsumoto

２．発表標題
Improving Discriminative Learning for Zero-Shot Relation Extraction

３．学会等名
1st Workshop on Semiparametric Methods in NLP: Decoupling Logic from Knowledge（国際学会）

４．発表年
2022年

１．発表者名
Jungmin Choi, Ukyo Honda, Taro Watanabe, Hiroki Ouchi, Kentaro Inui

２．発表標題
Law Retrieval with Supervised Contrastive Learning Using the Hierarchical Structure of Law

３．学会等名
36th Pacific Asia Conference on Language, Information and Computation（国際学会）

４．発表年
2022年

１．発表者名
Akio Hayakawa, Tomoyuki Kajiwara, Hiroki Ouchi, Taro Watanabe

２．発表標題
JADES: New Text Simplification Dataset in Japanese Targeted at Non-Native Speakers

３．学会等名
Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)（国際学会）

４．発表年
2022年

１．発表者名
Shuhei Kurita, Hiroki Ouchi, Kentaro Inui, Satoshi Sekine

２．発表標題
Iterative Span Selection: Self-Emergence of Resolving Orders in Semantic Role Labeling

３．学会等名
29th International Conference on Computational Linguistics（国際学会）

４．発表年
2022年

１．発表者名 大羽未悠，栗林樹生，大内 啓樹，渡辺 太郎
２．発表標題 言語モデルの第二言語獲得効率
３．学会等名 情報処理学会自然言語処理研究会
４．発表年 2022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