

令和 6 年 6 月 3 日現在

機関番号：13901

研究種目：基盤研究(C)（一般）

研究期間：2021～2023

課題番号：21K12012

研究課題名（和文）文脈化単語埋め込みを用いた意味フレームの自動構築

研究課題名（英文）Building Semantic Frames using Contextualized Word Embeddings

研究代表者

笹野 遼平（Sasano, Ryohei）

名古屋大学・情報学研究科・准教授

研究者番号：70603918

交付決定額（研究期間全体）：（直接経費） 3,200,000円

研究成果の概要（和文）：本研究課題では、主に文脈化単語埋め込みを利用した意味フレーム知識の自動構築に取り組んだ。具体的には、各動詞が喚起するフレームごとに動詞をクラスタリングするフレーム推定タスクにおいては、マスクされた単語埋め込みと2段階クラスタリングを用いる手法、および、深層距離学習に基づき文脈化単語埋め込みモデルを fine-tuning した後に動詞のクラスタリングを行う手法を提案し、既存手法を超える性能を達成した。また、フレーム要素推定においても同様に、深層距離学習に基づき文脈化単語埋め込みモデルを fine-tuning した後に頂のクラスタリングを行う手法を提案し、既存手法を超える性能を達成した。

研究成果の学術的意義や社会的意義

本研究の成果は、これまで人手で行われてきた意味フレーム資源の開発コストの低減する可能性がある。具体的には、これまで大規模な意味フレーム資源が整備されてこなかった言語を対象に意味フレームを自動構築したり、人手によるフレーム資源の整備の補助に用いることが可能である。また、大規模なコーパスから学習された大規模言語モデルが、人が言語理解の前提として持つ経験的知識をどの程度含んでいるかの分析に貢献する可能性がある。

研究成果の概要（英文）：In this research project, we worked mainly on automatically constructing semantic frame knowledge using contextualized word embeddings. Specifically, we proposed a method that uses masked word embedding and two-stage clustering for the frame induction task, in which verbs are clustered according to the frames they evoke, and a method that performs fine-tuning of the contextualized word embedding model based on deep metric learning followed by clustering of the verbs. The proposed method achieves a higher performance than the existing methods. In frame element knowledge acquisition task, we proposed a method that performs fine-tuning of the contextualized word embedding model based on deep metric learning followed by clustering of arguments of frame evoking words, and achieved a higher performance than the existing methods.

研究分野：自然言語処理

キーワード：意味フレーム 文脈化単語埋め込み

1. 研究開始当初の背景

語の意味を固定長の密ベクトルで表現する手法が多く提案され、大規模なコーパスから高精度な単語ベクトル表現の構築が可能となっていた。特に BERT などの文脈を考慮した単語ベクトル表現(文脈化単語埋め込み)構築手法は、多くのタスクに応用され、人に迫る高い精度を実現していた。このことから、これらの文脈化単語埋め込みは、大量の注釈付きデータが存在しているタスクに対しては、人の言語理解を模倣するような出力や判断を行うことが可能となってきたと言えるが、これらの意味表現は大規模なコーパスのみから学習された知識であり、人が言語理解の前提として持つ経験的知識をどの程度含んでいるかは明らかではなかった。

また、大規模なコーパスから自動的にフレーム知識を構築する取り組みも存在していたが、その多くは、文法的な特徴を手掛かりとして構築された表層的なフレーム知識であり、深層的な意味を扱ってはいなかった。さらに、大規模なコーパスから述語とその項を収集し、それらの文脈から独立した意味に基づきクラスタリングを行うことでフレームを構築しているため、文脈による述語や項の意味の違いを十分に考慮できているとは言えないものであった。また、フレーム要素には人が理解しやすい意味役割は付与されてはならず、自動構築されたフレーム知識に基づく意味解析は人にとって理解しやすい形式になっていなかった。

2. 研究の目的

文脈化単語埋め込みを用いた意味フレームの自動構築に取り組むことで、BERT などの文脈化単語埋め込みが、人が常識として持つ経験的知識をどの程度捉えているかを明らかにすることを目的とする。また、自動構築したフレーム知識を、人手で整備された知識フレームと対応付けることで、人にとって理解しやすく、かつ、単語埋め込み技術と親和性の高いフレーム知識の構築を目指す。

3. 研究の方法

(1) マスク単語埋め込みと 2 段階クラスタリングを用いた動詞の意味フレーム推定

意味フレームの自動構築の第一段階として、文脈化単語埋め込みを用いた動詞の意味フレーム推定に取り組んだ。既存の文脈化埋め込みに基づくフレーム推定手法には、動詞の表層的な情報を過度に考慮してしまうこと、および、同じ動詞の用例が多くのクラスタに分かれすぎてしまうという問題があることを明らかにし、マスクされた単語埋め込みと 2 段階クラスタリングを用いる手法を構築した。

(2) 自動構築した日本語格フレームと英語 FrameNet の対応付け

日本語を対象に自動構築された表層的フレーム知識である京大格フレームと、人手で整備された英語の意味フレーム知識である FrameNet の自動的な対応付けに取り組み、重要性の高い意味役割(core)だけでなく、付属的な意味役割(non-core)も考慮した対応付け手法を構築した。

(3) 深層距離学習を用いた動詞の意味フレーム推定

文脈化単語埋め込みを用いた動詞の意味フレーム推定において、汎用的な埋め込み空間は、意味的に類似したフレームの事例が近くに位置しているという人間の直観と必ずしも一致しているわけではないため、事前学習のみに基づく文脈化単語埋め込みを用いる手法の性能には限界があることに着目し、深層距離学習に基づき文脈化単語埋め込みモデルを fine-tuning することで高精度な意味フレーム推定手法を構築した。

(4) 深層距離学習を用いたフレーム要素知識の自動獲得

意味フレームのフレーム要素推定において、埋め込み空間において類似した意味役割を担うフレーム要素の埋め込みを近づけ、異なる意味役割を担うフレーム要素の埋め込みを遠ざけるように、深層距離学習に基づく事前学習済み言語モデルを最適化し、得られたモデルの埋め込みを利用して項をクラスタリングする手法を構築した。

4. 研究成果

(1) マスク単語埋め込みと2段階クラスタリングを用いた動詞の意味フレーム推定

マスクされた単語埋め込みと2段階クラスタリングを用いた動詞の意味フレーム推定手法について、FrameNetを用いた実験を実施し、マスクされた単語埋め込みを活用することが動詞の表層的な情報に強く依存したクラスタの構築を抑制し、また、2段階のクラスタリングを行うことで各動詞の用例が属するクラスタの異なり数を適正化できることを確認した。

(2) 自動構築した日本語格フレームと英語 FrameNet の対応付け

京大格フレームとFrameNetの対応付けにおいて、重要性の高い意味役割だけでなく、付随的な意味役割も考慮して対応付けを行うことで、より高精度なフレームの対応付けが可能となることを示した。

(3) 深層距離学習を用いた動詞の意味フレーム推定

動詞の意味フレーム推定において、FrameNetを用いた評価実験を実施し、深層距離学習を適用することで8ポイント以上スコアが向上することを確認した。さらに、深層距離学習に利用する教師データの量を変化させた実験を通し、教師データが極めて少量である場合でも、提案手法が有効であることを示した。さらに、深層距離学習を実施する前と後の文脈化埋め込みを、喚起するフレームごとに色で区別し、2次元にマッピングした結果(図1)から、深層距離学習を実施することで、埋め込み表現が人のフレームに関する直観に近いものとなることを確認した。

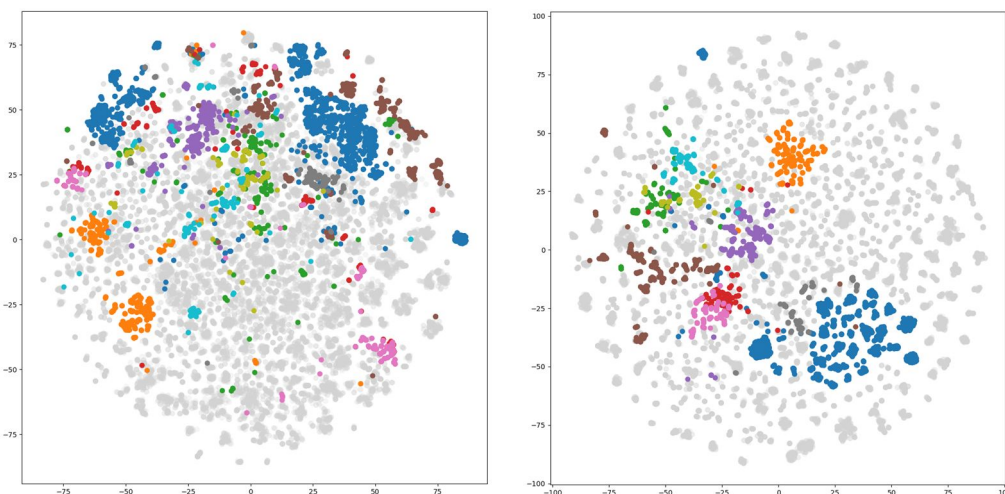


図1 深層距離学習前(左)と後(右)のフレーム喚起語の埋め込みの2次元マッピング。

(4) 深層距離学習を用いたフレーム要素知識の自動獲得

フレーム要素推定において、FrameNetを用いた評価実験を実施し、事前学習済み言語モデルを最適化し、得られたモデルの埋め込みを利用して項をクラスタリングすることで、高精度にフレーム要素推定が可能であることを確認した。特にフレーム要素が各フレーム固有に定義されているという特徴を利用したフレーム別の項クラスタリングアプローチによって、従来手法に比べて高性能なフレーム要素知識獲得手法を実現した。

5. 主な発表論文等

〔雑誌論文〕 計3件（うち査読付論文 3件/うち国際共著 0件/うちオープンアクセス 3件）

1. 著者名 山田康輔, 笹野遼平, 武田浩一	4. 巻 29
2. 論文標題 マスクされた単語埋め込みと2段階クラスタリングを用いた動詞の意味フレーム推定	5. 発行年 2022年
3. 雑誌名 自然言語処理	6. 最初と最後の頁 395-415
掲載論文のDOI (デジタルオブジェクト識別子) 10.5715/jnlp.29.395	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 山田康輔, 笹野遼平, 武田浩一	4. 巻 30
2. 論文標題 深層距離学習を用いた動詞の意味フレーム推定	5. 発行年 2023年
3. 雑誌名 山田康輔, 笹野遼平, 武田浩一	6. 最初と最後の頁 1130-1150
掲載論文のDOI (デジタルオブジェクト識別子) 10.5715/jnlp.30.1130	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 山田康輔, 笹野遼平, 武田浩一	4. 巻 65
2. 論文標題 深層距離学習を用いた意味フレーム構築におけるフレーム要素知識の自動獲得	5. 発行年 2024年
3. 雑誌名 情報処理学会論文誌	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6件（うち招待講演 0件/うち国際学会 6件）

1. 発表者名 Kosuke Yamada, Ryohei Sasano, Koichi Takeda
2. 発表標題 Semantic Frame Induction using Masked Word Embeddings and Two-Step Clustering
3. 学会等名 The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (国際学会)
4. 発表年 2021年

1. 発表者名 Kosuke Yamada, Ryohei Sasano, Koichi Takeda
2. 発表標題 Verb Sense Clustering Using Contextualized Word Representations for Semantic Frame Induction
3. 学会等名 Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (国際学会)
4. 発表年 2021年

1. 発表者名 Ryohei Sasano
2. 発表標題 Cross-lingual Linking of Automatically Constructed Frames and FrameNet
3. 学会等名 The 13th International Conference on Language Resources and Evaluation (国際学会)
4. 発表年 2022年

1. 発表者名 Kosuke Yamada, Ryohei Sasano, Koichi Takeda
2. 発表標題 Semantic Frame Induction with Deep Metric Learning
3. 学会等名 The 17th Conference of the European Chapter of the Association for Computational Linguistics (国際学会)
4. 発表年 2023年

1. 発表者名 Kosuke Yamada, Ryohei Sasano, Koichi Takeda
2. 発表標題 Argument Clustering with Deep Metric Learning for Semantic Frame Induction
3. 学会等名 Findings of the Association for Computational Linguistics: ACL 2023 (国際学会)
4. 発表年 2023年

1. 発表者名 Masashi Oshika, Kosuke Yamada, Ryohei Sasano, Koichi Takeda
2. 発表標題 Transformer-based Live Update Generation for Soccer Matches from Microblog Posts
3. 学会等名 The 2023 Conference on Empirical Methods in Natural Language Processing (国際学会)
4. 発表年 2023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