深層距離学習を用いた動詞の意味フレーム推定

山田 康輔 [1]　笹野 遼平 [1,2]　武田 浩一 [1]
[1] 名古屋大学　[2] 理化学研究所
yamada.kosuke.v1@s.mail.nagoya-u.ac.jp　{sasano,takedasu}@i.nagoya-u.ac.jp

概要

意味フレーム推定において、事前学習済み文脈化単語埋め込みを用いる手法が主流になっている。しかし、汎用的な埋め込み空間は、必ずしもフレームに関する人間の直観と一致しているわけではない。そこで、本研究では、コーパス内の一部の述語についてのラベル付きデータの存在を仮定し、深層距離学習に基づき文脈化単語埋め込みをファインチューニングすることで、高精度な意味フレーム推定を実現する手法を提案する。実験を通し、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示す。さらに、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す。

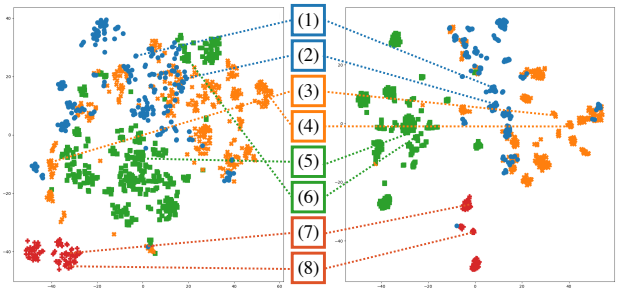
1　はじめに

動詞の意味フレーム推定は、テキスト中の動詞を、その動詞が喚起する意味フレームごとにまとめるタスクである。たとえば、表 1 に示される FrameNet [1, 2] の 8 つの用例の場合、各動詞が喚起するフレームごとにグループ化し、4 つのクラスタを形成することが目標となる。

意味フレーム推定において、ELMo [3] や BERT [4] などの文脈化単語埋め込みの有用性が報告されている [5, 6, 7]。図 1 (a) は FrameNet に含まれる用例中の動詞の事前学習済み BERT (Vanilla BERT) による埋め込みを t-SNE [8] で 2 次元にマッピングした結果である。動詞「cover」の用例 (1) と (7) は空間上で離れている一方、同じ Topic フレームを喚起する動詞の用例 (7) と (8) は近くに位置しており、ある程度、意味フレームの違いを反映した埋め込み空間であるといえる。しかし、同じフレームを喚起する動詞が離れた位置に存在するケースも散見される。たとえば、同じ Removing フレームを喚起する動詞の用例 (5) と (6) は互いに離れた位置に存在している。これは Vanilla BERT の埋め込み空間が、意味的に似た事例が近い位置に、異なる事例が離れた位置になるという人の直観と常に一致しているわけではないことを意味している。

表 1　FrameNet 内のフレームを喚起する動詞の用例

フレーム	用例
Filling	(1) She covered her mouth with her hand.
	(2) I filled a notebook with my name.
Placing	(3) You can embed graphs in your worksheet.
	(4) He parked the car at the hotel.
Removing	(5) Volunteers removed grass from the marsh.
	(6) They'd drained the drop from the teapot.
Topic	(7) Each database will cover a specific topic.
	(8) Chapter 8 treats the educational advantages.



(a) Vanilla BERT　(b) Fine-tuned BERT w/ AdaCos

図 1　Vanilla BERT と AdaCos を用いてファインチューニングされた BERT による動詞の埋め込みの 2 次元マッピング。 各色 (形) は Filling (●)、Placing (×)、Removing (■)、Topic (+) フレームを示し、数字は表 1 と対応する。

本研究では、フレームに関する人の直観をより強く反映した意味フレーム推定手法を実現するため、コーパス内の一部の述語に対してアノテートされたデータの存在を仮定する教師あり意味フレーム推定タスクにおいて、深層距離学習に基づき文脈化単語埋め込みをファインチューニングすることで、高精度な意味フレーム推定を実現する手法を提案する。深層距離学習は、同じラベルの事例を埋め込み空間上で近づけ、異なるラベルの事例を遠ざける学習を行うものであり、教師データに基づく埋め込み空間の調整が期待できる。図 1 (b) は代表的な深層距離学習手法の 1 つである AdaCos [9] を用いてファインチューニングした BERT による埋め込みの 2 次元マッピングである。Vanilla BERT において同じ意

味フレームを喚起する動詞の用例であるにも関わらず、距離が離れていた用例 (3) と (4)、用例 (5) と (6) が、AdaCos を用いてファインチューニングした BERT では、互いに近い位置に存在していることが確認できる。これは、深層距離学習によって、意味フレームに関する人の直観をより反映させた埋め込み空間が得られたことを示している。

2 教師あり意味フレーム推定

動詞の教師あり意味フレーム推定は、コーパス内の一部の動詞への教師データの存在を仮定した意味フレーム推定タスクであり、より性能の高い意味フレーム推定手法を実現することが目的となる。教師なし意味フレーム推定では、文脈化単語埋め込みなどの特徴量ベクトルを用いたクラスタリングベースの手法が一般的であるが、教師あり意味フレーム推定においても同様の枠組みが適用できる。異なる点は、教師データを用いることで特徴量ベクトルを学習できるかどうかである。本稿では、教師データである学習セットに存在しない動詞の意味フレーム推定に対応するため、学習セットとクラスタリング対象のテストセット内の動詞が重複しない設定を採用する。ただし、異なる動詞が同一のフレームを喚起するケースは存在することから、テストセットに含まれるフレームの一部は学習セットにも出現する。

2.1 ベースライン手法

シンプルなベースライン手法として、文脈化単語埋め込みによる 1 段階クラスタリングを用いる。1 段階クラスタリングでは、ユークリッド距離に基づく群平均法による階層型クラスタリングを利用する。また、Yamada ら [10] によるマスクされた単語埋め込みと 2 段階クラスタリングを活用した手法を導入する。前者に関して、クラスタリングに用いる埋め込みは動詞の文脈化埋め込み (v_{word}) とその動詞を “[MASK]” に置き換えたときの文脈化埋め込み (v_{mask}) の加重平均 (v_{w+m}) とする。これは α を重みとして、式 (1) で定義される。

$$v_{w+m} = (1-\alpha)\cdot v_{word} + \alpha \cdot v_{mask} \tag{1}$$

後者の 2 段階クラスタリングは、1 段階目に動詞ごとの用例クラスタリング[1)]、2 段階目に動詞横断クラスタリングを行う手法である。1 段階目に X-means [11]、2 段階目に群平均法を用いる。その他の設定は、Yamada ら [10] と同様とする。

2.2 深層距離学習の適用

教師あり意味フレーム推定において、文脈化単語埋め込みのファインチューニングとして深層距離学習を適用する。これにより、同じフレームの事例が近づき、異なるフレームの事例が遠ざかることが期待される。本稿では、距離ベースと分類ベースの 2 つのアプローチを採用する。

距離ベースのアプローチ　一般に複数エンコーダを用いて事例ペア間の距離を学習するアプローチである。事例ペアはある事例をアンカーとし、同じフレームの事例を正例、異なるフレームの事例を負例として作成する。損失には以下の 2 つを導入する。

Contrastive 損失 [12] は、正例ペアを近づけ、負例ペアを一定のマージン以上遠ざける学習を実現する。これは、クラス i の事例の埋め込みを $\boldsymbol{x}_i$、マージンを m、平方ユークリッド距離に基づく距離関数を D とした式 (2) で定義される。

$$L_{\text{cont}} = \begin{cases} D\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) & i = j \\ \max\left(m - D\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right), 0\right) & i \neq j \end{cases} \tag{2}$$

また、Triplet 損失 [13] は、事例の 3 つ組に対して、アンカー $\boldsymbol{x}_a$ と負例 $\boldsymbol{x}_n$ の距離を、アンカー $\boldsymbol{x}_a$ と正例 $\boldsymbol{x}_p$ の距離より一定のマージン以上遠ざける学習を行うものであり、式 (3) で定義される。

$$L_{\text{tri}} = \max\left(D\left(\boldsymbol{x}_a, \boldsymbol{x}_p\right) - D\left(\boldsymbol{x}_a, \boldsymbol{x}_n\right) + m, 0\right) \tag{3}$$

分類ベースのアプローチ　近年、顔認識タスクを中心に広く利用されているアプローチである。このアプローチのモデルの多くは、単一エンコーダと線形層を持つネットワークを利用し、式 (4) の softmax 損失がベースとなっている。式 (4) の $\boldsymbol{w}_i$ と b_i は線形層の重みとバイアス、n はクラス数を示す。

$$L_{\text{soft}} = -\log \frac{e^{\boldsymbol{w}_i^\top \boldsymbol{x}_i + b_i}}{\sum_{j=1}^{n} e^{\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j}} \tag{4}$$

重み $\boldsymbol{w}_i$ をクラス i の埋め込みと捉え、事例とクラスの埋め込みの距離を学習するための損失がいくつか提案されている [14, 15, 16]。その中でも ArcFace 損失 [16] は幾何的な解釈に優れることから広く利用されている。ArcFace 損失は、式 (5) のように、softmax 損失をベースとして、b_i を除き、$\boldsymbol{w}_i$ と $\boldsymbol{x}_i$ に l_2 正規化を適用することで $\boldsymbol{w}_i^\top \boldsymbol{x}_i$ を $\cos\theta_i$ と表現し、

1) FrameNet では、フレームとその喚起語を結び付けたものを Lexical Unit (LU) と呼び、1 段階目で得られたクラスタは各動詞の用例をその動詞が喚起するフレームごとにまとめたものであるため、本稿では擬似 LU (pseudo-LU; pLU) と呼ぶ。

表 2　3 分割交差検証による教師あり意味フレーム推定実験結果。#pLU と#C はそれぞれ pLU 数とクラスタ数を示す。

クラスタリング	モデル	α	#pLU	#C	Pu / iPu / PiF	BcP / BcR / BcF
1 段階クラスタリング (群平均法)	Vanilla	0.00	–	429	53.0 / 57.0 / 54.9	40.8 / 44.6 / 42.6
	Contrastive	0.13	–	443	56.9 / 70.0 / 62.8	45.1 / 58.6 / 51.0
	Triplet	0.23	–	425	70.0 / 77.0 / 73.3	**60.3** / 68.1 / **63.9**
	Softmax	0.23	–	440	65.1 / 78.0 / 71.0	53.3 / 68.6 / 59.9
	ArcFace	0.37	–	436	**70.3** / 76.2 / 73.1	59.7 / 67.4 / 63.3
	AdaCos	0.30	–	446	69.0 / **78.7** / **73.5**	57.5 / **69.5** / 62.9
2 段階クラスタリング (X-means & 群平均法)	Vanilla	0.67	877	444	60.6 / 74.9 / 66.9	49.7 / 65.8 / 56.5
	Contrastive	0.23	1,904	689	69.2 / 62.5 / 65.7	59.5 / 50.9 / 54.8
	Triplet	0.50	1,014	454	73.4 / **76.7** / 74.8	64.6 / **68.0** / 66.0
	Softmax	0.43	1,428	919	**84.7** / 62.5 / 71.9	**78.4** / 50.4 / 61.4
	ArcFace	0.47	955	452	70.5 / 76.5 / 73.3	60.8 / 67.7 / 63.8
	AdaCos	0.50	1,128	656	80.8 / 71.3 / **75.6**	73.2 / 60.9 / **66.2**

クラス内の集約性とクラス間の分散性を強化するためにマージン m とスケール s を導入している[2]。

$$L_{\text{arc}} = -\log \frac{e^{s\cdot\cos(\theta_i+m)}}{e^{s\cdot\cos(\theta_i+m)} + \sum_{j=1,j\neq i}^{n} e^{s\cdot\cos\theta_j}} \tag{5}$$

Zhang ら [9] はこれらの損失の性能がハイパーパラメータ依存な点を指摘し、それらの値を調査している。結果として、マージンを除き、動的なスケール $\tilde{s}$ を用いた式 (6) の AdaCos 損失を提案している。

$$L_{\text{ada}} = -\log \frac{e^{\tilde{s}\cdot\cos\theta_i}}{\sum_{j=1}^{n} e^{\tilde{s}\cdot\cos\theta_j}} \tag{6}$$

3　実験

教師あり意味フレーム推定における深層距離学習によるファインチューニングの有用性を評価する。また、少量の学習事例における性能を検証する。

3.1　実験設定

データセット　FrameNet 1.7 [2] から、フレームを喚起する動詞の用例を抽出し、データセットを作成した。3 分割交差検証を行うため、それらを動詞単位で分割し[3]、3 つのサブセットを作成した。動詞数、LU 数、フレーム数、事例数の平均は、それぞれ 831、1,273、434、27,537 である。学習セットは文脈化単語埋め込みのファインチューニングに使用し、開発セットは埋め込み v_{w+m} の重み α[4]、クラスタ数、マージン[5]の決定に使用する。

2）　マージンとスケールの働きが類似すること [9] から、本実験では、スケールを 64 に固定し、マージンのみを探索する。
3）　多義動詞の割合は一定とする。
4）　0 から 1 まで 0.1 刻みで探索している。
5）　contrastive 損失と triplet 損失では、0.1、0.2、0.5、1.0、ArcFace 損失では、0.01、0.02、0.05、0.1 の範囲で探索している。

比較手法　文脈化埋め込みモデルとして、事前学習済み BERT (bert-base-uncased)[6] を使用する。ファインチューニングをしないモデル (Vanilla) と 5 つのファインチューニングされたモデル (Contrastive、Triplet、Softmax、ArcFace、AdaCos) に対して、1 段階クラスタリングと 2 段階クラスタリングを用いた全部で 12 の手法を比較する。埋め込みは全て l_2 正規化を適用する。また、バッチサイズは 32、学習率は 1e-5、エポック数は 5 とし、最適化アルゴリズムは AdamW [17] を使用する。

評価指標　評価指標として、B-cubed Precision (BcP)、Recall (BcR)、およびその調和平均である F 値 (BcF) [18] と、Purity (Pu)、Inverse Purity (iPu)、およびその調和平均である F 値 (PiF) [19] を使用する。

3.2　実験結果

表 2 に実験結果を示す。Vanilla モデルと比較して、ファインチューニングされたモデル、特に Triplet、ArcFace、AdaCos モデル、は全体的に高い BcF および PiF を実現しており、その有用性が確認できる。一方、Contrastive モデルでは相対的に低いスコアとなった。これは Contrastive モデルのマージンが適切な粒度のクラスタ構築との親和性が低いことが要因であると考えられる。クラスタリングに関して、2 段階クラスタリングが 1 段階クラスタリングに比べて全体的に高いスコアを達成している。しかし、Vanilla モデルの場合では BcF と PiF 共に 12 ポイント以上差があったが、ファインチューニングされたモデルの場合ではその差が縮まっているこ

6）　https://huggingface.co/bert-base-uncased

表 3　学習事例数を変化させたときの実験結果。各列は学習セット内の各 LU の使用された最大事例数を示す。

クラスタリング	モデル	PiF					BcF				
		1 /	2 /	5 /	10 /	all	1 /	2 /	5 /	10 /	all
1 段階クラスタリング (群平均法)	Vanilla	54.9	54.9	54.9	54.9	54.9	42.6	42.6	42.6	42.6	42.6
	Triplet	**68.2**	**70.9**	**71.7**	**72.9**	73.3	**57.4**	**60.6**	**61.8**	**63.0**	**63.9**
	AdaCos	57.5	59.7	66.5	70.9	**73.5**	44.7	47.1	54.6	60.0	62.9
2 段階クラスタリング (X-means & 群平均法)	Vanilla	66.9	66.9	66.9	66.9	66.9	56.5	56.5	56.5	56.5	56.5
	Triplet	**71.7**	**72.5**	**73.9**	74.0	74.8	**62.4**	**63.1**	**64.8**	64.9	66.0
	AdaCos	67.3	69.4	73.4	**74.3**	**75.6**	57.6	59.7	64.5	**65.3**	**66.2**

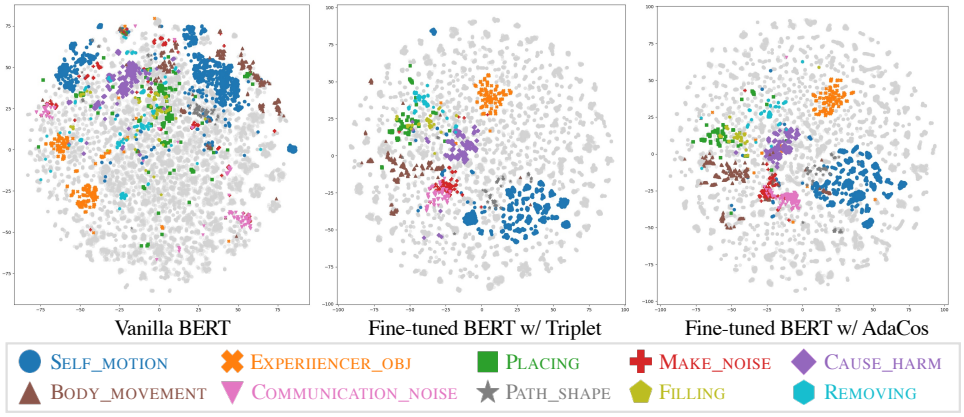


図 2　3 つのモデルにおける v_{word} の 2 次元マッピング。各色および形状は事例数の多い上位 10 フレームを示す。

とが確認できる。このため、ファインチューニングを行う場合、1 段階クラスタリングも有望な選択肢になり得る。また、重み α に関して、スコアの高い AdaCos や Triplet モデルを用いた 2 段階クラスタリングにおいて値が 0.5 であり、マスクされた単語埋め込みを用いることの有効性が確認できる。

以上の実験結果から、約 30,000 という大規模な学習事例数が存在する場合については、学習事例に基づきファインチューニングを行うことで高い性能を実現できることが確認できた。しかし、FrameNet のような大規模リソースが存在しない言語への適用を考えた場合、少量の学習事例に対しても高い性能が期待できることが重要となる。そこで、LU ごとの最大学習事例数の条件を '1'、'2'、'5'、'10'、'all' として実験を行った。3 セットの平均学習事例数はそれぞれ 1,273、2,445、5,680、10,053、27,537 である。表 3 に結果を示す。Triplet モデルは少ない学習事例数においても有用であることが確認できる。特に、2 段階クラスタリングでは、'1' と 'all' で学習事例数が 20 倍以上異なるにも関わらず、スコアは 3 ポイント程度しか変わらないことから、少量の教師データが存在するのであれば、Triplet モデルを用いた 2 段階クラスタリングを適用することで、高精度な意味フレーム推定が実現可能と考えられる。一方、AdaCos モデルでは学習事例が少ない場合は大きな性能の改善が確認できない。これは線形層の重みが十分に学習されないためであると考えられる。

図 2 に Vanilla、Triplet、AdaCos モデルの v_{word} における t-SNE による 2 次元マッピングを示す。Vanilla モデルでは、フレームごとに事例がまとまる傾向はあるが、SELF_MOTION フレームの事例は大きく 2 つのクラスタに分かれており、REMOVING フレームの事例は散らばっている。一方、Triplet や AdaCos モデルでは、Vanilla モデルと比較して、より意味フレームごとにまとまっていることが確認できる。

4　おわりに

本稿では、文脈化単語埋め込みを深層距離学習に基づきファインチューニングすることで高性能な意味フレーム推定手法が可能となることを示した。特に、Triplet、ArcFace、AdaCos モデルは全体的に高いスコアを獲得しており、その有用性が確認できた。また、Triplet モデルにおいては、少量の学習事例数でも有用であることを示した。今後の方針として、多言語の動詞や名詞の意味フレーム推定における提案手法の有用性を検証したいと考えている。

謝辞

本研究は、JST 創発的研究支援事業 JPMJFR216N、および JSPS 科研費 22J14993 の支援を受けたものである。

参考文献

[1] Collin F Baker, Charles J Fillmore, and John B Lowe. The Berkeley FrameNet project. In **ACL-COLING**, pp. 86–90, 1998.

[2] Josef Ruppenhofer, Michael Ellsworth, Myriam Schwarzer-Petruck, Christopher R Johnson, and Jan Scheffczyk. **FrameNet II: Extended theory and practice**. International Computer Science Institute, 2016.

[3] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In **NAACL-HLT**, pp. 2227–2237, 2018.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **NAACL**, pp. 4171–4186, 2019.

[5] Nikolay Arefyev, Boris Sheludko, Adis Davletov, Dmitry Kharchev, Alex Nevidomsky, and Alexander Panchenko. Neural GRANNy at SemEval-2019 task 2: A combined approach for better modeling of semantic relationships in semantic frame induction. In **SemEval**, pp. 31–38, 2019.

[6] Saba Anwar, Dmitry Ustalov, Nikolay Arefyev, Simone Paolo Ponzetto, Chris Biemann, and Alexander Panchenko. HHMM at SemEval-2019 task 2: Unsupervised frame induction using contextualized word embeddings. In **SemEval**, pp. 125–129, 2019.

[7] Eugénio Ribeiro, Vânia Mendonça, Ricardo Ribeiro, David Martins de Matos, Alberto Sardinha, Ana Lúcia Santos, and Luísa Coheur. L2F/INESC-ID at SemEval-2019 task 2: Unsupervised lexical semantic frame induction using contextualized word representations. In **SemEval**, pp. 130–136, 2019.

[8] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. **Journal of Machine Learning Research**, Vol. 9, pp. 2579–2605, 2008.

[9] Xiao Zhang, Rui Zhao, Yu Qiao, Xiaogang Wang, and Hongsheng Li. AdaCos: Adaptively scaling cosine logits for effectively learning deep face representations. In **CVPR**, pp. 10823–10832, 2019.

[10] Kosuke Yamada, Ryohei Sasano, and Koichi Takeda. Semantic frame induction using masked word embeddings and two-step clustering. In **ACL-IJCNLP**, pp. 811–816, 2021.

[11] Dan Pelleg and Andrew Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In **ICML**, pp. 727–734, 2000.

[12] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In **CVPR**, Vol. 2, pp. 1735–1742, 2006.

[13] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. **Journal of Machine Learning Research**, Vol. 10, No. 2, 2009.

[14] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. SphereFace: Deep hypersphere embedding for face recognition. In **CVPR**, pp. 212–220, 2017.

[15] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. CosFace: Large margin cosine loss for deep face recognition. In **CVPR**, pp. 5265–5274, 2018.

[16] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In **CVPR**, pp. 4690–4699, 2019.

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **ICLR**, 2017.

[18] Amit Bagga and Breck Baldwin. Entity-based cross-document coreferencing using the vector space model. In **ACL-COLING**, pp. 79–85, 1998.

[19] Ying Zhao and George Karypis. Criterion functions for document clustering: Experiments and analysis. Technical report, Retrieved from the University of Minnesota Digital Conservancy, 2001.